How do we improve the solution?

- Smaller steps.

1.3 CONSERVATION LAWS

Conservation laws form the basis of a variety of complicated and powerful model and are conceptually easy to understand.

Conservation laws provide the foundation for many model functions.

— They boil down to

Change = increase – decreases

— Can be used to predict changes with respect to time by given it a special name "the time-variant (or transient)" computation

— If no change occurs, the increases and decreases must be in balance.

Change = 0 = increases – decreases

- It is given a special name, the "steady-state" calculation

Example : Fluid flow

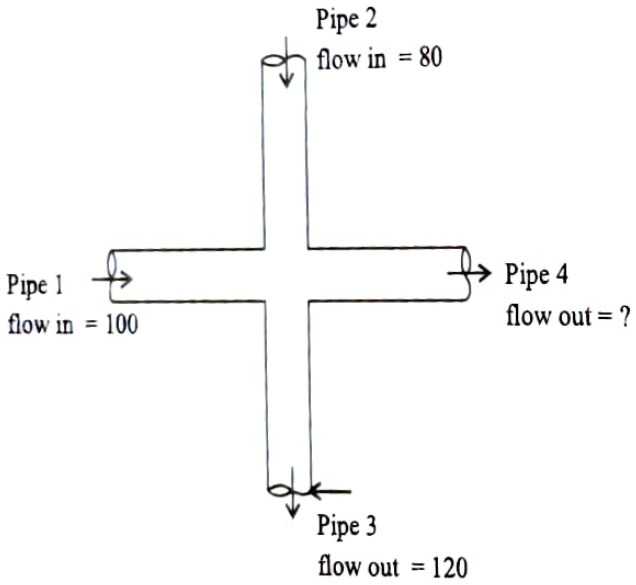

Fig : A flow balance for steady incompressible fluid flow at the junction of pipes.

— For steady-state incompressible fluid flow in pipes flow in = Flow out

- The flow out of the fourth pipe must be 60.

Table 1: Devices and types of balances that are commonly used in the four major areas of engineering.

For each case, the conservation law upon which the balance is based is specified.

●●●

2 APPROXIMATIONS AND ROUND-OFF ERRORS

STRUCTURE:

2.0. INTRODUCTION

When a calculator or digital computer is used to perform numerical calculations, an unavoidable error, called round-off error must be considered. This error arises because the arithmetic operations performed in a machine involve numbers with only a finite number of digits, with the result that many calculations are performed with approximate representation of actual numbers. The computer has the in-built capability to perform only the basic arithmetic operations of addition, subtraction, multiplication and division. While formulating algorithm all other mathematical operations are reduced to these basic operations even when solving problems involving the operations of calculus. We will discuss the importance of significant figures and the rules for significant figures we will also discuss accuracy and precision and error definitions. We will see the impact of round-off errors on the significant figures defined by computer system.

2.1 OBJECTIVES

After going through this unit, you should be able to:

- Learn about floating-point representation of numbers.
- Learn about significant figures.
- Learn about sources of errors and error definitions.
- Understand the propagation of errors in subsequent calculations.
- Understand the meaning of accuracy and precision.
- Know round of errors.

2.2 SIGNIFICANT FIGURES

The concept of a significant figure or digit has been developed to formally designate the reliability of a numerical value.

The significant digits of a number are those that can be used with confidence.

The concept of significant figures has two important implications of numerical methods.

1. Numerical methods yield approximate result. We must therefore develop criteria to specify how confident we are in our approximate result. One way to do this is in terms of significant figures. For example. we might decide that our approximation is acceptable if it is correct to four significant figures.
2. Although quantities like π, e, $\sqrt{7}$ etc. represent specific quantities, they cannot be expressed exactly by a limited number of digits.

   For example, $\pi = 3.1415926535897323\ldots\ldots$

   Because computers retain only a finite number of significant figures, such numbers can never be represented exactly.

   The omission of the remaining significant figures is called round-off errors.

Rules for significant figures

The significant figures of a number are digits that carry meaning contributing to its measurement resolution. This includes all digits except.

- All leading zeros.
- Trailing zeros when they are merely placeholders to indicate the scale of the number.
- Spurious digits introduced for example, by calculations carried out to greater precision than that of the original data or measurements reported to a greater precision than the equipment supports.

Here are the basic rules for significant digits –

1. All non-zero digits are significant.
2. Zero's between non-zero digits are significant.
3. Leading zero's are never significant
4. In a number with a decimal point, trailing zeros those to the right of the last non-zero digits are significant.
5. In a number without a decimal point, trailing zeros may or may not be significant.

Significant figures rules explain with examples.

1. All non-zero digits are significant

   Example 62 has two significant figures (6 and 2) while 123.45 has five significant figures (1, 2, 3, 4 and 5).
2. Zero's appearing anywhere between two non-zero digits are significant.

   Example 701.1204 has seven significant digits while 10001 has five significant digits.
3. Leading zero's are not significant.

Example 0.00048 has two significant figures (4 and 8)

Example 0.001048 has four significant figures.

4. Trailing zero's in a number containing a decimal point are significant.

Example 12.4800 has 6 significant figures (1, 2, 4, 8, 0 and 0)

The number 0.000124800 also has 6 significant figures (The zero's before the 1 are not significant) In addition, 240.00 has five significant figures since it has three trailing zeros. This convention clarifies the precision of such numbers; for example, if 0 measurement precise to four decimal places (0.0001) it is given as 12.23 then it might be understood that only two decimal places of precision are available.

Stating the result as 12.2300 makes clear that it is precise to four decimal places (in this case, six significant figures).

The significance of trailing zeros in a number not containing a decimal point can be ambiguous. For example, it may be clear if a number like 1400 is precise to the nearest unit (and just happens coincidentally to be an exact multiple of hundred) or if its only shown to the nearest hundred due to rounding or uncertainty.

2.3 ACCURACY AND PRECISION

The errors associated with both calculations and measurements can be characterized with regard to their accuracy and precision.

Accuracy and precision are used in context of measurement. Accuracy refers to the degree of conformity and correctness of something when compared to a true or absolute value, while precision refers to a state of strict exactness – how consistently something is strictly exact.

In other words, the precision of an experiments, object, or value is a measure of the reliability and consistency.

The accuracy of an experiment, object or value is a measurement of how closely results agree with the true or accepted value.

Precision refers to how closely individual computed or measured values agree with each other.

In the field of science, engineering and statistics, the accuracy of a measurement system is the degree of closeness of measures of a quantity to that quantity's True Value.

The precision of a measurement system, related to reproducibility and repeatability is the degree to which repeated measurements under unchanged conditions show the same results.

Difference between Accuracy and Precision

	Accuracy	Precision
(1)	The ability of a measurement to match the actual value of the quantity being measured.	The ability of a measurement to be consistently Reproduced.
(2)	The degree of conformity and correctness of something when compared to a true or absolute value.	A state of strict exactness how often something is strictly exact.

(3)	Single factor or measurement.	Multiple measurements or factors are needed.
(4)	Something can be accurate on occasion as flake. For something to be consistently and reliably accurate, it must also be precise.	Results can be precise without being accurate. Alternatively results can be precise and accurate.
(5)	Example. If in the reality it is 32°F outside and a temperature sensor reads 32.0°F then the sensor is accurate.	If on several test the temperature sensor matches the actual temperature while the actual temperature is held constant, then the temperature sensor is precise.

For example on Accuracy and Precision

(1) Good Accuracy and Good Precision

Suppose a lab refrigerator holds a constant temperature of 38.0 F. A temperature sensor is tested 10 times in the refrigerator. The temperature from the test gives as 38.0, 38.0, 37.9, 38.1, 38.0, 37.9, 38.0, 38.2, 38.0, 37.9.

This distribution does show a tendency toward a particular value (High precision) and is very near the actual temperature each time (High accuracy).

(2) Good Accuracy and Bad Precision.

Suppose a lab refrigerator holds a constant temperature of 38.0 F. A temperature sensor is tested 10 times in the refrigerator. The temperature from the test yield the temperatures of 37.8, 38.3, 38.1, 38.0, 37.6, 38.2, 38.0, 38.0, 37.4, 38.3.

This distribution shows no tendency toward a particular value (lock of precision) but each value does come close to the actual temperature (High accuracy).

(3) Bad Accuracy Good Precision.

Suppose a lab refrigerator holds a constant temperature of 38.0 F. A temperature sensor is tested 10 times in the refrigerator. The temperature from the test yield the temperature 39.2, 39.3, 39.1, 39.0, 39.1, 39.3, 39.2, 39.1, 39.2, 39.2.

This distribution does show a tendency toward a particular value (High precision) but every measurement is well off from the actual temperature (Low accuracy).

(4) Bad Accuracy and Bad Precision

Suppose a lab refrigerator holds a constant temperature of 38.0 F. A temperature sensor is tested to times in the refrigerator. The temperature from the test yield the temperatures 39.1, 38.1, 39.3, 37.5, 38.3, 39.1, 37.1, 37.8, 38.8, 39.0.

This distribution shows no tendency toward a particular value (Lack of precision) and does not acceptably match the actual temperature (Lack of accuracy).

2.4 ERROR DEFINITIONS

Numerical errors arise from the use of approximations to represent exact mathematical operations and quatities. These include **truncation errors**, which result when approximations are used to represent exact mathematical procedures and **Round-off errors**, which result when numbers having limited significant figures are used to represent exact numbers.

For both types the relationship between the exact or true results and the approximation can be formulated as.

$$\text{True value} = \text{Approximation} + \text{Error.}$$

$$\therefore \quad \text{Error } (E_t) = \text{True value} - \text{Approximation}$$

$$\text{Relation Error} = \frac{\text{True value} - \text{Approximation}}{\text{True value}}$$

$$\text{Percentage Relative Error} = \left(\frac{\text{True value} - \text{Approximation}}{\text{True value}}\right) \times 100$$

Floating Point of Numbers and Errors

Fractional quantities are typically represented in computers using floating point form. In this approach, the number is expressed as a fractional part, called a mantissa or significance, and an integer part, called an exponent or characteristic, as in

$$m \cdot b^e$$

where m = the mantissa, b = the base of the number system being used and e = the exponent for instance, the numbers 156.78 could be represented as 0.15678×10^3 in a floating point base 10 system. The first bit 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 1/34 = 0.029411765 ... was stored in a floating point base to system that allowed only four decimal places to be stored. Thus $\frac{1}{34}$ would be stored as:

$$0.0294 \times 10^{\circ}$$

However, in the process of doing this, the inclusion of the useless zero to the right of the decimal forces as to drop the digit 1 in the fifth decimal place. The number can be normalized to remove the leading zero by multiplying the mantissa by 10 and lowering the exponents by 1 to give.

$$0.2941 \times 10^{-1}$$

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

The consequence of normalization is that the absolute value of m is limited. That is,

$$\frac{1}{b} \leq m < 1$$

Where, b = the base for example for a base –10 system, m would range between 0.1 and 1, and for a base –2 system between 0.5 and 1.

Floating point representation allows both fractions and very large numbers to be exposed on the computer.

However, it has some disadvantages. For example, floating-point numbers take up more room and take longer to process than integer numbers. More significantly, however, their use introduces a source of error because the mantissa holds only a finite number of significant figures. Thus, a round-off error in introduced.

2.5 ROUND-OFF ERRORS

Round-off errors arise due to floating point representation of initial data in the mac Subsequent errors in the solution due to this is called propogated errors. Due to finite digit arithn operations, the computer generates, in the solution of a problem errors known as Rounding errors.

Errors arising due to in exact arithmetic operation is called generated error. In exact arithn operations results due to finite digit arithmetic operations in the machine. If arithmetic operati done with the infinite digit representation then this error would not appear.

During arithmetic operation of two floating point numbers of same length n, we obta floating point number of different length m (usually $m > n$). Computer cannot store the resu number exactly since it can represent numbers a length n. So only n digits are stored. This gives ri error.

Example: Let $a = 0.75632 \times 10^2$ and $b = 0.235472 \times 10^{-1}$

Now $a + b = 75.632 + 0.0235472$

$= 75.655472$ in accumulator

$a + b = 0.756555 \times 10^2$ if 6 decimal digit arithmetic is used.

$a + b = .756555\text{E-}2$

Example: Let $a = 0.23 \times 10^1$ and $b = 0.30 \times 10^2$

Now $\frac{a}{b} = \frac{2.3}{30}$

$= 0.07666666667$

$= 0.766667\text{E}{-1}$

Example: How many bits of significantly will be lost in the following subtration (4 decimal

$37.593621 - 37.584216$

Solution:

$= 0.009405$

$= 9405\text{E}{-2}$

Example: Show that $a(b - c) \neq ab - bc$ (upto 4 decimals)

$a = .5555 \times 10^1$

$b = .4545 \times 10^1$

$c = .4535 \times 10^1$

Solution:

$b - c = .4545 \times 10^1 - .4535 \times 10^1$

$= 4.545 - 4.535 = 0.010 = .10 \times 10^{-1}$

$a \times (b - c) = 0.5555 \times 10^1 \ (.10 \times 10^1)$

$= .05555 = .55 \times 10^{-1}$

$a \times b = (0.5555 \times 10^1)(.4545 \times 10^1)$

$= .25247475 \times 10^2$

$$= .2525 \times 10^2$$

$$ac = (.5555 \times 10^1) \times (.4535 \times 10^1)$$

$$= 0.25191925 \times 10^2$$

$$= .2519E{-}2$$

Now ab = .2525E2 and ac = .2579E2

$$ab - ac = (0.2525 - .2519) \times 10^2$$

$$= 0.0006. \times 10^2$$

$$= .6 \times 10^{-1}$$

$$= .6000E - 1$$

$$\therefore \quad a(b - c) \neq ab - ac$$

EXERCISE

1. Explain significant figures.
2. Explain the difference between Accuracy and Precision.
3. Explain Accuracy and Precision.
4. Explain different types of errors.
5. Explain Round-off errors with example.
6. Prove that $a + (b + c) \neq (a + b) + c$ with 4 decimal places where a, b, c, are floating point numbers.
7. Evaluate

   $y = x^3 - 7x^2 + 8x - 0.35$ at $x = 1.37$ use 3-digit and 4-digit arithmetic and find the significant digits lost. Also find the relative error after rounding-off.
8. Evaluate $y = 2x^3 - 5x^2 + 7x - 1$ at $x = 0.84$ using 4 digit arithmetic. Find the significant digits lost and find relative errors after rounding-off.
9. If $a = 5.645$ and $b = 7.821$, find $a + b$ actual and find $a + b$ upto 4 digit arithmetic by floating point numbers. Find the error due to rounding-off.
10. If $a = 42.33$ and $b = 0.0033$; find $a + b$ using floating point numbers upto 4 digits and find the error % relative error.
11. If $a = 10.34$ and $b = -10.27$ using floating point upto 4 digits and find relative error.
12. Use the floating point number with 4 digit mantissa and find the error, relative error and loss of significant figures.
    (a) 12.24 + 14.23
    (b) 9.834 + 2.450
    (c) 2.314 – 2.273
    (d) 23.45 – 23.39
    (e) $1 + x - e^x$ for $x = 0.1$

13. Find $a * b$ and find round-off error and significant digits with 6 digits.

(a) $a = 5.645$ $b = 7.820$

(b) $a = 6.123$ $b = 4.247$

(c) $a = 0.0063$ $b = 0.0079$

(d) $a = 0.0003$ $b = 0.0022$

2.6 SUMMARY

In this unit we have covered the following,

After discussing floating-point representation of numbers, we have discussing the arithmetic operations with normalized floating point numbers. This leads to a discussion on rounding errors. Also we have discussed other sources of errors ... like propogated errors loss of significant digits etc.

Very brief idea about stability or instability of a numerical algorithm is presented. We have also discussed in detail about the significant figures.

A brief overview about accuracy and precisions has been discussed

ANSWERS

Answers to the exercise are not given and it is left for students to find using actual values and floating point numbers.

●●●

3 TRUNCTION ERRORS AND THE TAYLOR SERIES

STRUCTURE:

3.0. INTRODUCTION

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

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

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

As mentioned earlier, a number of different types of errors arise during the process of numerical computing. All these errors contribute to the total error in the final result. A taxonomy of errors encountered in a numerical process is given in Fig. 3.1 which shows that every stage of the numerical computing cycle contributes to the total error.

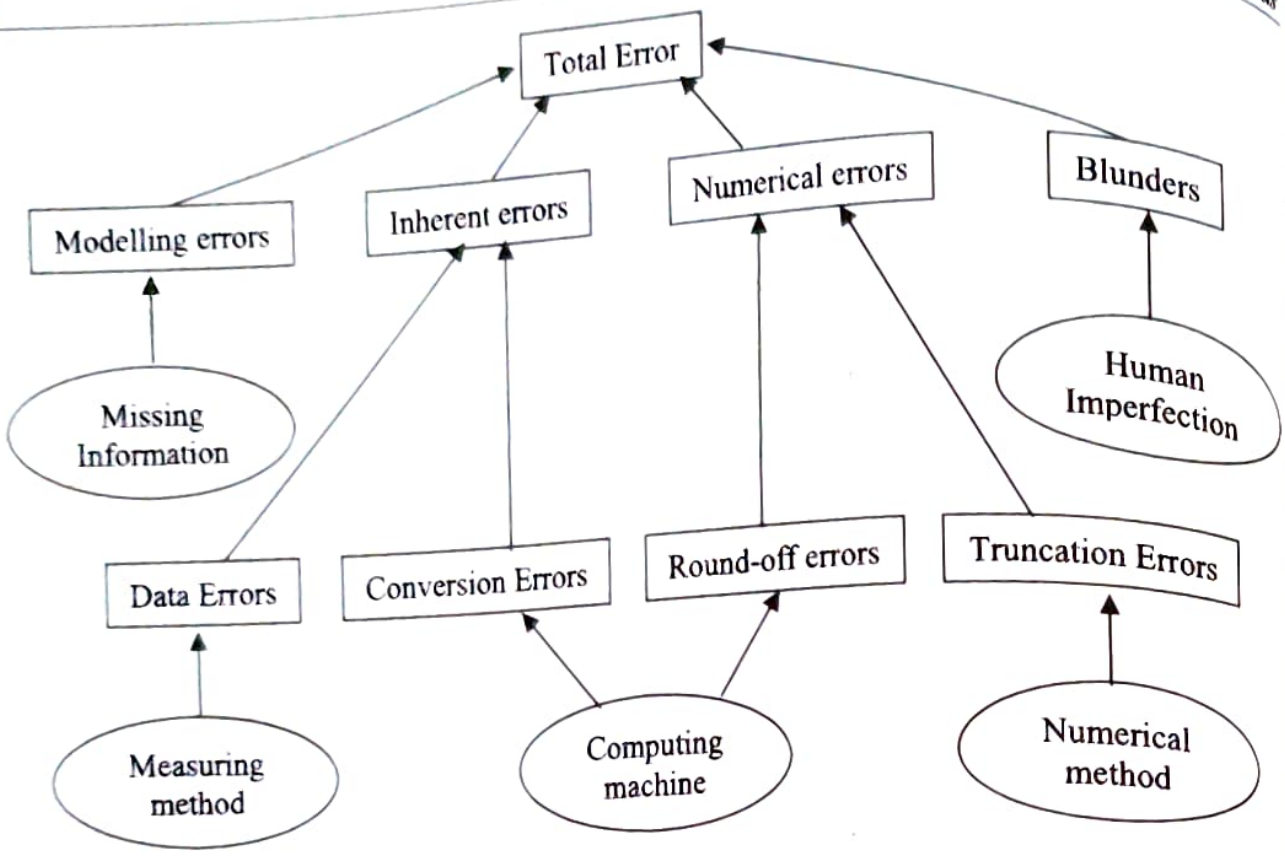

Fig. 3.1

Although we always insist for perfection but it is rarely achieved in practice due to a variety of factors. But that must not deter our attempts to achieve near perfection. Again the question is. **How much near**?

In this chapter we discuss the various forms of approximations and errors, their sources how they propogate, and how they affect the result as well as the solution process.

3.1 OBJECTIVES

After going through this chapter, you should be able to:

- Learn Taylor series and errors in calculations
- Learn propagation of errors
- Learn total numerical errors
- Learn Formulation Errors and Data Uncertainty

3.2 TAYLOR SERIES

Taylor theorem and its associated formula, the Taylor series is of great value in the study of numerical methods. In essence, the Taylor series provides a mean to predict a function value at one point in terms of the function value and its derivatives at another point. In particular, the theorem states that any smooth function can be approximated as a polynomial.

A useful way to gain insight into the Taylor series is to build it term by term. For example, the first term in the series is,

$$f(x_{i+1}) \cong f(x_i) \qquad \ldots (1)$$

This relationship called the zero-order approximation, indicates that the value of f at the new point is the same as its value at the old point. This result makes initiative because if and x_i and x_{i+1} are close to each other, it is likely that the new value probably similar to the old value.

Equation (1), provides a perfect estimate if the function being approximated is, in fact a constant. However, if the function changes at all over the interval, additional terms of the Taylor series are require to provide a better estimate.

Taylor theorem

It the function f and its first $n + 1$ derivatives are continuous on an interval containing a and x, then the value of the function at x is given by,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)(x-a)^3}{3!} + \frac{f^{(n)}(a)(x-a)^n}{n! + R_n} + \frac{r^{(n)}(a)}{n!}(x-a)^n + R_n \quad \ldots (1.1)$$

where the remainder R_n in defined as:

$$R_n = \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)dt \quad \ldots (1.2)$$

where, $t = a$ dummy variable eqn. (1.1) is called the Taylor series or Taylors formula. If the remainder is omitted, the right hand side to eqn. (1.1) is the Taylor polynomial approximation to $f(x)$. In essence, the theorem states that any smooth function can be approximated as a polynomial.

Eqn. (1.2) is but one way, called the integral form by which the remainder can be expressed, An alternative formulation can be derived on the basis of the integral mean-value theme.

First Theorem of mean for integrals:

If the function g is continuous and integrable on an interval containing a and x, then there exists a point ξ between a and x such that,

$$\int_a^x g(t)dt = g(\xi)(x-a) \quad \ldots (1.3)$$

In other words, this theorem, states that the integral can be represented by an average value for the function $g(\xi)$ times the interval length $x - a$. Because the average must occur between the minimum and maximum values for the interval, there is a point $x = \xi$ at which the function takes on the average value.

The first theorem is in fact a special case of a second mean-value theorem for integrals.

Second theorem of mean for integrals:

If the functions g and h are continuous and intergrable on an interval containing a and x and h does not change sign in the interval, then there exists a point ξ between a and x such that,

$$\int_a^x g(t)\,h(t)\,dt = g(\xi)\int_a^x h(t)\,dt \quad \ldots (1.4)$$

Thus the eqn· (1.4) is equivalent to eqn (1.4) with $h(t) = 1$

The second theorem can be applied to eqn (1.2) with

$$g(t) = f^{(n+1)}(t) \qquad h(t) = \frac{(x-t)^n}{n!}$$

As t varies from a to x $h(t)$ is continuous and does not change sign. Therefore, if $f^{(n+1)}(t)$ is continuous, then the integral mean-value theorem holds

and $R_n = \dfrac{f^{(n-1)}(\xi)}{(n+1)!}(x-a)^{n+1}$

This equation is referred to as the derivative or lagrange form of the remainder.

For the fist order approximation is developed by adding another term to yield

$$f(x_{i+1}) \equiv f(x_i) + f'(x_i)(x_{i+1} - x_i) \qquad \ldots (1.5)$$

The additional first order term consists of a slope $f'(x_i)$ multiplied by the distance between x_i and x_{i+1}. Thus the expression is now in the form of a straight line and is capable of predicting an increase or decrease of the function between x_i and x_{i+1}.

A second order term is added to the series to capture some of the curvature that the function night exhibit.

$$f(x_{i+1}) \equiv f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2$$

In a similar manner, additional terms can be included to develop the complete Taylor series expansion.

$$f(x_{i+1}) = f'(x_i) + f'(x_{i+1} - x_i) + f'' \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2$$

$$+ \frac{f^{(3)}(x_i)}{3!}(x_{i+1} - x_i)^3 + \ldots + \frac{f^{(n)}(x_i)}{n!}(x_{i+1} - x_i)^n + R_n \qquad \ldots (1.6)$$

Note that because eqn (1.6) is an infinite series, an equal sign replaces approximation sign that was used an eqn (1.6) through (1.5). A reminder term is included to account for all terms form $n + 1$ to infinity?

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x_{i+1} - x_i)^{n+1} \qquad \ldots (1.7)$$

where, the subscript n can notes that this is the remainder for the nth order approximation and ξ is a value of x that lies somewhere between x_i and x_{i+1}. The introduction of the ξ is so important that we will work into it carefully in later section.

For the time being it is sufficient to recognise that there is such a value that provides an exact determination of the error.

It is of often inconvenient to simplify the Taylor series by defining a step size $h = x_{i+1} - x_i$ and expressing eqn (1.6) as:

$$f(x_{i+1}) = f(x_i) + f'(x_i)\,h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)h^3}{3!} + \ldots + \frac{f^{(n)}(x_i)h^n}{n!} + R_n \qquad \ldots (1.8)$$

where the remainder term is now

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}$$

Example : Use of Taylor series Expansion to Approximate a function with an infinite number of derivatives.

Problem Statement: Use Taylor series expansion with $n = 0$ to 6 to approximate $f(x) = \cos x$ at $x_{i+1} = \pi/3$ on the basis of the value of $f(x)$ and its derivatives at $x_i = \pi/4$.

Note that this means that $h = \pi/3 - \pi/4 = \pi/12$.

Solution: We can determine the correct value $f(\pi/3) = 0.5$. The zero order approximation is,

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) = 0.707106731$$

which represents a percent relative error of

$$\varepsilon_t = \frac{0.5 - 0.707106781}{0.5}100\% = -41.4\%$$

For the first order approximation, we add the first derivative term where $f'(x) = -\sin x$

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) - \sin\left(\frac{\pi}{4}\right)\left(\frac{\pi}{12}\right) = 0.521986659$$

which has $\varepsilon t = -4.40$ percent

For the second order approximations, we add the second derivative term, where $f''(x) = -\cos x$

$$f\left(\frac{\pi}{3}\right) \cong \cos\left(\frac{\pi}{4}\right) - \sin\left(\frac{\pi}{4}\right)\left(\frac{\pi}{12}\right) - \frac{\cos(\pi/4)}{2}\left(\frac{\pi}{12}\right)^2 = 0.497754491$$

with $\varepsilon t = 0.449$ percent. Thus, the inclusion of additional terms results in an improved estimate.

The process can be continued and the results listed in table 1. Notice that the derivatives never go to zero as was the case with polynomial Therefore, each additional terms results in some improvement in the estimate. However, also notice that how most of the improvement comes with the initial terms. For this case, by the time we have added the third order term, the error in reduced to 2.62×10^{-2} percent which means that we have attained 99.9738 percent of the true value. Consequently, although the addition of more terms will reduce the error further, the improvement become negligible.