01 Intro Lecture 1 Instructor

Errors in Numerical Calculations

1.1 INTRODUCTION

In practical applications, an engineer would finally obtain results in a numerical form. For example, from a set of tabulated data derived from an experiment, inferences may have to be drawn; or, a system of linear algebraic equations is to be solved. The aim of numerical analysis is to provide efficient methods for obtaining numerical answers to such problems.

Recall Basic Definitions

(a) *Algebraic and transcendental equations:* The problem of solving nonlinear equations of the type $f(x) = 0$ is frequently encountered in engineering. For example, the equation

$$\frac{M_0}{M_0 - u_f t} = e^{(u+gt)/u_0} \qquad (1.1)$$

is a nonlinear equation for t when M_0, g, u, u_0 and u_f are given. Equations of this type occur in rocket studies.

(b) *Interpolation:* Given a set of data values (x_i, y_i), $i = 0, 1, 2, \ldots, n$, of a function $y = f(x)$, where the explicit nature of $f(x)$ is not known, it is often required to find the value of y for a given value of x, where $x_0 < x < x_n$. This process is called *interpolation*. If this process is carried out for functions of several variables, it is called *multivariate* interpolation.

(c) *Curve fitting:* This is a special case where the data points are subject to errors, both round off and systematic. In such a case, interpolation formulae yield unsatisfactory solutions. Experimental results are often subject to errors and, in such cases, the method is to fit a curve which passes through the data points and then use the curve to predict the intermediate values. This problem is usually referred to as *data smoothing.*

(d) *Numerical differentiation and integration:* It is often required to determine the numerical values of

(i) $\frac{dy}{dx}, \frac{d^2y}{dx^2}, \ldots,$ for a certain value of x in $x_0 \leq x \leq x_n$, and

(ii) $I = \int_{x_0}^{x_n} y\,dx,$

where the set of data values (x_i, y_i), $i = 0, 1, \ldots, n$ is given, but the explicit nature of $y(x)$ is not known. For example, if the data consist of the angle θ (in radians) of a rotating rod for values of time t (in seconds), then its angular velocity and angular acceleration at any time can be computed by numerical differentiation formulae.

(e) *Matrices and linear systems:* The problem of solving systems of linear algebraic equations and the determination of eigenvalues and eigenvectors of matrices are major problems of disciplines such as differential equations, fluid mechanics, theory of structures, etc.

(f) *Ordinary and partial differential equations:* Engineering problems are often formulated in terms of an ordinary or a partial differential equation. For example, the mathematical formulation of a falling body involves an ordinary differential equation and the problem of determining the steady-state distribution of temperature on a heated plate is formulated in terms of a partial differential equation. In most cases, exact solutions are not possible and a numerical method has to be adopted. In addition to the finite difference methods, this book also presents a brief introduction to the cubic spline method for solving certain partial differential equations.

(g) *Integral equations:* An equation in which the unknown function appears under the integral sign is known as an *integral equation.* Equations of this type occur in several areas of higher mathematics such as aerodynamics, elasticity, electrostatics, etc. A short account of some well-known methods is given.

In the numerical solution of problems, we usually start with some initial data and then compute, after some intermediate steps, the final results. The given numerical data are only approximate because they may be true to two, three or more figures. In addition, the methods used may also be approximate and therefore the error in a computed result may be due to the errors in the data, or the errors in the method, or both. In Section 1.3, we discuss some basic ideas

1.1.1 Computer and Numerical Software

It is well known that computers and mathematics are two important tools of numerical methods. Prior to 1950, numerical methods could only be implemented by manual computations, but the rapid technological advances resulted in the production of computing machines which are faster, economical and smaller in size. Today's engineers have access to several types of computing systems, viz., mainframe computers, personal computers and super computers. Of these, the personal computer is a smaller machine which is useful, less expensive and, as the name implies, can easily be possessed and used by individuals. Nevertheless, mere possession of a computer is not of great consequence; it can be used effectively only by providing suitable instructions to it. These instructions are known as *software*. It is therefore imperative that we develop suitable software for an effective implementation of numerical methods on computers.

1.1.2 Computer Languages

Several computer languages have so far been developed and there are limitations on every language. The question of preferring a particular language over others depends on the problem and its requirements. We list below some important problem-solving languages, which are currently in use:

(a) *FORTRAN:* Standing for FORmula TRANslation, FORTRAN was introduced by IBM in 1957. Since then, it has undergone many changes and the present version, called FORTRAN 90, is favoured by most scientists and engineers. It is readily available on almost all computers and one of its important features is that it allows a programmer to express the mathematical algorithm more precisely. It has special features like extended double precision, special mathematical functions and complex variables. Besides, FORTRAN is the language used in numerically oriented subprograms developed by many software libraries. For example, (IMSL) (International Mathematical and Statistical Library, Inc.) consists of FORTRAN subroutines and functions in applied mathematics, statistics and special functions. FORTRAN programs are also available in the book, *Numerical Recipes*, published by the Cambridge University Press, for most of the standard numerical methods.

(b) *C:* This is a high-level programming language developed by Bell Telephone Laboratories in 1972. Presently, it is being taught at several engineering colleges as the first computer language and is therefore used by a large number of engineers and scientists. Computer programs in C for standard numerical methods are available in the book, *Numerical Recipes in C*, published by the Cambridge University Press.

(c) *BASIC:* Originally developed by John Kemeny and Thomas Kurtz in 1960, BASIC was used in the first few years only for instruction purposes. Over the years, it has grown tremendously and the present version is called Visual Basic. One of its important applications is in the development of software on personal computers. It is easy to use.

1.1.3 Software Packages

It is well known that the programming effort is considerably reduced by using standard *functions* and *subroutines*. Several software packages for numerical methods are available in the form of 'functions' and these are being extensively used by engineering students. One such package is MATLAB, standing for MATrices LABoratory. It was developed by Cleve Moler and John N. Little. As the name implies, it was originally founded to develop a matrix package but now it incorporates several numerical methods such as root-finding of polynomials, cubic spline interpolation, discrete Fourier transforms, numerical differentiation and integration, ordinary differential equations and eigenvalue problems. Besides, MATLAB has excellent display capabilities which can be used in the case of two-dimensional problems. Using the

MATLAB functions, it is possible to implement most of the numerical methods on personal computers and hence it has become one of the most popular packages in most laboratories and technical colleges. MATLAB has its own programming language and this is described in detail in the text by Stephen J. Chapman.*

Relevant Theorems

Theorem 1.1 If $f(x)$ is continuous in $a \leq x \leq b$, and if $f(a)$ and $f(b)$ are of opposite signs, then $f(\xi) = 0$ for at least one number ξ such that $a < \xi < b$.

Theorem 1.2 (*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 ξ, such that $f'(\xi) = 0$, $a < \xi < b$.

Theorem 1.3 (*Generalized Rolle's theorem*) Let $f(x)$ be a function which is n times differentiable on $[a, b]$. If $f(x)$ vanishes at the $(n+1)$ distinct points $x_0, x_1, \ldots, x_n$ in (a, b), then there exists a number ξ in (a, b) such that $f^{(n)}(\xi) = 0$.

Theorem 1.4 *(Intermediate value theorem)* Let $f(x)$ be continuous in $[a, b]$ and let k be any number between $f(a)$ and $f(b)$. Then there exists a number ξ in (a, b) such that $f(\xi) = k$ (see Fig. 1.1).

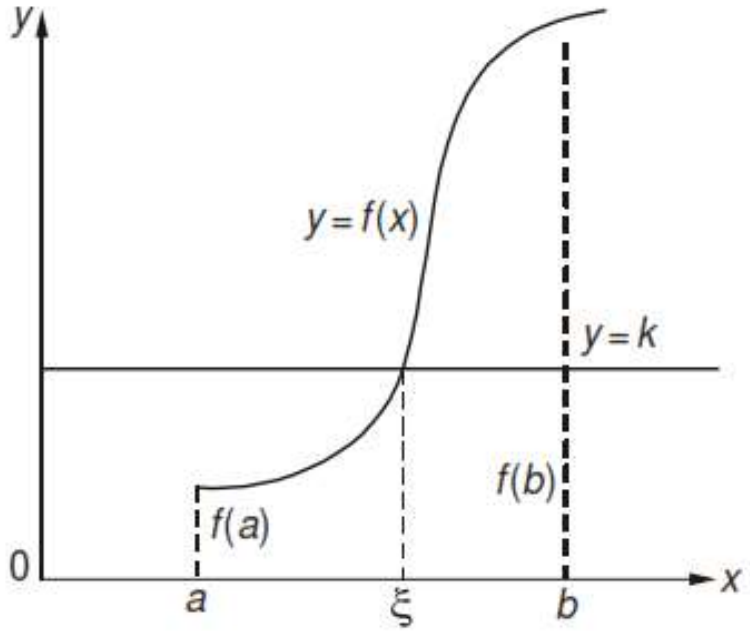

Figure 1.1 Intermediate value theorem.

Theorem 1.5 (*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 ξ, between a and b such that

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}, \qquad a < \xi < b.$$

Setting $b = a + h$, this theorem takes the form

$$f(a + h) = f(a) + hf'(a + \theta h), \qquad 0 < \theta < 1.$$

Theorem 1.6 (*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

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

where $R_n(x)$, the *remainder term*, can be expressed in the form

$$R_n(x) = \frac{(x-a)^n}{n!}f^{(n)}(\xi), \qquad a < \xi < x.$$

Theorem 1.7 *(Maclaurin's expansion)* It states

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

Theorem 1.8 (*Taylor's series for a function of two variables*) It states

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

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

This can easily be generalized.

Theorem 1.9 *(Taylor's series for a function of several variables)*

$$f(x_1+\Delta x_1, x_2+\Delta x_2, \ldots, x_n+\Delta x_n)$$

$$= f(x_1, x_2, \ldots, x_n) + \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n$$

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

$$\left. + 2\frac{\partial^2 f}{\partial x_{n-1}\partial x_n}\Delta x_{n-1}\,\Delta x_n\right] + \cdots$$

1.3 ERRORS AND THEIR COMPUTATIONS

There are two kinds of numbers, *exact* and *approximate* numbers. Examples of exact numbers are 1, 2, 3, ..., 1/2, 3/2, ..., $\sqrt{2}$, π, e, etc., written in this manner. Approximate numbers are those that represent the numbers to a certain degree of accuracy. Thus, an approximate value of π is 3.1416, or if we desire a better approximation, it is 3.14159265. But we cannot write the *exact* value of π.

The digits that are used to express a number are called *significant digits* or *significant figures*. Thus, the numbers 3.1416, 0.66667 and 4.0687 contain five significant digits each. The number 0.00023 has, however, only two significant digits, viz., 2 and 3, since the zeros serve only to fix the position of the decimal point. Similarly, the numbers 0.00145, 0.000145 and 0.0000145 all have three significant digits. In case of ambiguity, the scientific notation should be used. For example, in the number 25,600, the number of significant figures is uncertain, whereas the numbers 2.56×10^4, 2.560×10^4 and 2.5600×10^4 have three, four and five significant digits, respectively.

Rules for Rounding off Errors

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

(a) less than half a unit in the nth place, leave the nth digit unaltered;

(b) greater than half a unit in the nth place, increase the nth digit by unity;

(c) exactly half a unit in the nth place, increase the nth digit by unity if it is odd; otherwise, leave it unchanged.

The number thus rounded-off is said to be correct to n significant figures.

Example 1.1
figures:

Absolute, relative and percentage errors

Absolute error is the numerical difference between the true value of a quantity and its approximate value. Thus, if X is the true value of a quantity and X_1 is its approximate value, then the absolute error E_A is given by

$$E_A = X - X_1 = \delta X. \qquad (1.2)$$

The relative error E_R is defined by

$$E_R = \frac{E_A}{X} = \frac{\delta X}{X}, \qquad (1.3)$$

and the percentage error (E_P) by

$$E_P = 100\, E_R. \tag{1.4}$$

Let ΔX be a number such that

$$|X_1 - X| \leq \Delta X. \tag{1.5}$$

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

$$\frac{\Delta X}{|X|} \approx \frac{\Delta X}{|X_1|}$$

measures the *relative accuracy*.

It is easy to deduce that if two numbers are added or subtracted, then the magnitude of the absolute error in the result is the sum of the magnitudes of the absolute errors in the two numbers. More generally, if $E_{\mathrm{A}}^{1}, E_{\mathrm{A}}^{2}, \ldots, E_{\mathrm{A}}^{n}$ are the absolute errors in n numbers, then the magnitude of the absolute error in their sum is given by

$$|E_{\mathrm{A}}^{1}| + |E_{\mathrm{A}}^{2}| + \cdots + |E_{\mathrm{A}}^{n}| .$$

Note: While adding up several numbers of different absolute accuracies, the following procedure may be adopted:

(i) Isolate the number with the greatest absolute error,

(ii) Round-off all other numbers retaining in them one digit more than in the isolated number,

(iii) Add up, and

(iv) Round-off the sum by discarding one digit.

To find the absolute error, E_A, in a product of two numbers a and b, we write $E_A = (a + E_A^1)(b + E_A^2) - ab$, where E_A^1 and E_A^2 are the absolute errors in a and b respectively. Thus,

$$\begin{aligned} E_A &= aE_A^2 + bE_A^1 + E_A^1 E_A^2 \\ &= bE_A^1 + aE_A^2, \text{ approximately} \end{aligned} \tag{1.6}$$

Similarly, the absolute error in the quotient a/b is given by

$$\frac{a+E_{\mathrm{A}}^{1}}{b+E_{\mathrm{A}}^{2}}-\frac{a}{b}=\frac{bE_{\mathrm{A}}^{1}-aE_{\mathrm{A}}^{2}}{b(b+E_{\mathrm{A}}^{2})}$$

$$=\frac{bE_{\mathrm{A}}^{1}-aE_{\mathrm{A}}^{2}}{b^{2}(1+E_{\mathrm{A}}^{2}/b)}$$

$$=\frac{bE_{\mathrm{A}}^{1}-aE_{\mathrm{A}}^{2}}{b^{2}}, \text{ assuming that } E_{\mathrm{A}}^{2}/b \text{ is small in comparison with 1}$$

$$=\frac{a}{b}\left(\frac{E_{\mathrm{A}}^{1}}{a}-\frac{E_{\mathrm{A}}^{2}}{b}\right). \qquad (1.7)$$

Example 1.2

Example 1.3

a

Example 1.4

Example 1.5

Example 1.6

Example 1.7

Example 1.8

Example 1.9

1.4 A GENERAL ERROR FORMULA

We now derive a general formula for the error committed in using a certain formula or a functional relation. Let

$$u = f(x, y, z) \tag{1.8}$$

and let the errors in x, y, z be Δx, Δy and Δz, respectively. Then the error Δu in u is given by

$$u + \Delta u = f(x + \Delta x, y + \Delta y, z + \Delta z) \tag{1.9}$$

Expanding the right-side of Eq. (1.9) by Taylor's series, we obtain

$$u + \Delta u = f(x, y, z) + \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z$$

$$+ \text{ terms involving higher powers of } \Delta x, \Delta y \text{ and } \Delta z \tag{1.10}$$

Assuming that the errors Δx, Δy, Δz are small, their higher powers can be neglected and Eq. (1.10) becomes

$$\Delta u = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + \frac{\partial u}{\partial z}\Delta z \tag{1.11}$$

The relative error in u is then given by

$$E_R = \frac{\Delta u}{u} = \frac{\partial u}{\partial x}\frac{\Delta x}{u} + \frac{\partial u}{\partial y}\frac{\Delta y}{u} + \frac{\partial u}{\partial z}\frac{\Delta z}{u} \tag{1.12}$$

Example 1.10

Example 1.11

1.5 ERROR IN A SERIES APPROXIMATION

The truncated error committed in a series approximation can be evaluated by using Taylor's series stated in Theorem 1.6. If x_i and x_{i+1} are two successive values of x, then we have

$$f(x_{i+1}) = f(x_i) + (x_{i+1} - x_i) f'(x_i) + \cdots + \frac{(x_{i+1} - x_i)^n}{n!} f^{(n)}(x_i) + R_{n+1}(x_{i+1}), \quad (1.13)$$

where

$$R_{n+1}(x_{i+1}) = \frac{(x_{i+1} - x_i)^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \qquad x_i < \xi < x_{i+1} \qquad (1.14)$$

In Eq. (1.13), the last term, $R_{n+1}(x_{i+1})$, is called the *remainder term* which, for a convergent series, tends to zero as $n \to \infty$. Thus, if $f(x_{i+1})$ is approximated by the first-n terms of the series given in Eq. (1.13), then the maximum error committed by using this approximation (called the *nth order approximation*) is given by the remainder term $R_{n+1}(x_{i+1})$. Conversely, if the accuracy required is specified in advance, then it would be possible to find n, the number of terms, such that the finite series yields the required accuracy.

Defining the interval length,

$$x_{i+1} - x_i = h, \tag{1.15}$$

Equation (1.13) may be written as

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2!} f''(x_i) + \cdots + \frac{h^n}{n!} f^{(n)}(x_i) + \mathrm{O}(h^{n+1}), \quad (1.16)$$

where $\mathrm{O}(h^{n+1})$ means that the truncation error is of the order of h^{n+1}, i.e., it is proportional to h^{n+1}. The meaning of this statement will be made clearer now.

Let the series be truncated after the first term. This gives the *zero-order* approximation:

$$f(x_{i+1}) = f(x_i) + \mathrm{O}(h), \quad (1.17)$$

which means that halving the interval length h will also halve the error in the approximate solution. Similarly, the *first-order* Taylor series approximation is given by

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \mathrm{O}(h^2), \tag{1.18}$$

which means that halving *the interval length*, h will quarter the error in the approximation. In such a case we say that approximation has a

second-order of convergence. We illustrate these facts through numerical examples.

Example 1.12

Example 1.13

EXERCISES

1.1 Explain the term 'round-off error' and round-off the following numbers to two decimal places:

48.21416, 2.3742, 52.275, 2.375, 2.385, 81.255

1.2 Round-off the following numbers to four significant figures:

38.46235, 0.70029, 0.0022218, 19.235101, 2.36425

1.3 Calculate the value of $\sqrt{102} - \sqrt{101}$ correct to four significant figures.

1.4 If $p = 3c^6 - 6c^2$, find the percentage error in p at $c = 1$, if the error in c is 0.05.

1.5 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 and 0.0467, where each number is correct to the digits given.

1.6 If $z = \frac{1}{8}xy^3$, find the percentage error in z when $x = 3.14 \pm 0.0016$ and $y = 4.5 \pm 0.05$.

1.7 Find the absolute error in the product uv if $u = 56.54 \pm 0.005$ and $v = 12.4 \pm 0.05$.

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

1.9 Find the relative error in the quotient 4.536/1.32, the numbers being correct to the digits given.

1.10 The exponential series is given by

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Find the number of terms of the above series such that their sum gives the value of e correct to five decimal places.

1.12 Write down the Taylor's series expansion of $f(x) = \cos x$ at $x = \frac{\pi}{3}$ in terms of $f(x)$, and its derivatives at $x = \frac{\pi}{4}$. Compute the approximations from the zeroth order to the fifth order and also state the absolute error in each case.

1.13 The Maclaurin expansion of $\sin x$ is given by

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

where x is in radians. Use the series to compute the value of $\sin 25°$ to an accuracy of 0.001.