MATH-351 Numerical Methods

Lecture 1

Book :- Numerical Analysis (10th Edition)
by
Burden and Faries

a) Analytical Methods b) Numerical Methods

c) Errors

Analytical Methods and Analytical Solution

analytical
An method for solving a given mathematical (Engineering) problem, which is based on rigorous mathematical analysis and leads to the true (exact) solution is known as **analytical solution**

Numerical Methods and **Numerical solution**

A numerical method for solving a given mathematical problem, which is based on rigorous mathematical analysis and leads to the approximate (non-exact) solution is known as **numerical solution**.

Why Numerical Methods

1) No Analytical Methods are available

Example

$$\frac{dy}{dx} = e^{x^2}$$

$$x^{1/3} - x^2 - 8 = 0$$

2) Analytical methods are complicated.

Example

$$\int \frac{1}{8 - x^3}\, dx$$

Solution : $\frac{\sqrt{3}}{12} \tan^{-1}\left(\frac{x+1}{\sqrt{3}}\right) + \frac{1}{24} \ln\left[\frac{x^2+2x+4}{x^2-4x-4}\right] + C$

Hint : Partial fractions.

Errors

The difference between a computed or measured value and a true or theoretically correct (exact) value.

Types of errors

I) Initial errors

Example

i) $x - y = 1$

$x - 1.0000y = 0$

Solution

$x = 100,001 \quad y = 100,000$

~~Solution~~

ii) $x - y = 1$

$x - 0.99999y = 0$

Solution

$x = -99,999$

$y = -100,000$

II) Discretization Error

III) Truncation error

IV) Rounding Error.

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

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

$$\left|\frac{X - X^*}{X}\right|.$$

(d) Types and Sources of Errors

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

(i) Initial errors

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

(ii) Discretization error

Most of the literature on the subject of computational errors does not make a distinction between discretization and truncation errors, the reason being that the two types of errors are almost inseparable. In this presentation we separate the two because truncation errors are special types of discretization errors.

The true (exact) solutions of some mathematical problems are continuous functions $y = f(x)$ of their respective independent variables. In almost all cases, numerical methods for solving such problems approximate the unknown continuous solution $f(x)$ by a sequence $\{f(x_n)\}$ of approximate values of the solution at a discrete set of points $\{x_n\}$ in the domain of the solution function $f(x)$. For example, the continuous function $f(x) = x + e^{-x}$ is the solution of the initial value problem $y' + y = 1 + x$, $y(0) = 1$

A typical numerical method for solving this problem is given by the recurrence relation

$x_0 = 0, \quad y_0 = 1, \quad y_n = (1-h)y_{n-1} + (1+x_{n-1})h, \quad n = 1,2,3,\ldots,$ h being a constant distance between two consecutive discrete values of the variable x. The error resulting from such a discretization process is called **discretization error.**

(iii) Truncation error

Truncation errors are special types of discretization errors. The term truncation error refers to the error in a method, which occurs because some infinite process is stopped prematurely (truncated) to a fewer number of terms or iterations in the process.

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

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

(iv) Rounding error

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

(e) Methods of reducing errors

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

(i) How to reduce initial error

Initial errors can have a devastating effect on numerical solutions.

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

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

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

However, if the entry -1.00001 in the matrix A is changed

to -0.99999 while all other data items remain unchanged, the resulting system of equations

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

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

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

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

(ii) How to reduce discretization errors

Different numerical methods for approximating the solution of a given mathematical problem can result in numerical solutions with very different degrees of accuracy due to the magnitudes of their respective discretization errors. Consider the problem of evaluating the definite integral:

$\int_0^1 \frac{dx}{1+x^2}$. The exact (analytical) value of the integral, correct to six decimal places is $\tan^{-1}(1) = 0.785398$.

Let us apply the trapezoidal method and Simpson's rule using an interval length

$h = 0.25$. First we evaluate the integrand $f(x) = \frac{1}{1+x^2}$ at the relevant points and get

i	0	1	2	3	4
x_i	0	0.25	0.5	0.75	1
$f(x_i)$	1.000000	0.941176	0.800000	0.640000	0,500000

The trapezoidal rule $\frac{h}{2}\left\{ f(x_0) + 2\left[f(x_1) + f(x_2) + f(x_3)\right] + f(x_4)\right\}$

gives the numerical solution 0.782794 .

Simpson's rule $\frac{h}{3}\left\{f(x_0) + 4\left[f(x_1) + f(x_3)\right] + 2f(x_2) + f(x_4)\right\}$ leads to

the numerical solution 0.785392.

One observes that, while the solution obtained using the trapezoidal rule is correct to only **two decimal places**, the solution obtained using Simpson's rule is correct to **four decimal places**. This significant difference in the accuracy of the two numerical solutions is caused by the differences in the discretization errors of the two numerical methods. Simpson's rule has a smaller discretization error than the trapezoidal rule.

In general, discretization errors cannot be avoided. However, one can reduce them substantially by being careful in **selecting a numerical method whose discretization error is known** *a priori* **to be relatively small.**

(iii) How to reduce truncation errors

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

The truncation error can be reduced either by

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

Example 1.2

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

The bisection method

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

The Newton-Raphson method

$x = x - \frac{f(x_n)}{f'(x_n)}$, provided $f'(x_n) \neq 0$.

If one performs only three iterations (truncation after three iterations) with each method using the starting values $x_0 = 0$ and $x_1 = 1$ for the bisection method and $x_1 = 0$ for the Newton-Raphson method, one gets the following sequence of approximations for each method.

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

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

DO THIS

Continue applying the bisect method on the above example until the solution is correct to three decimal places. How many more iterations did this require?

(iv) How to Reduce Rounding Errors

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

• What are Figures or Digits

In computational mathematics, the words **"figure"** and **"digit"** are synonyms. They are used interchangeably to mean any one of the ten numerals in the set **{0, 1, 2, 3, 4, 5, 6, 7, 8, 9}**.

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

$N = 00073920600365.00004507000$

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

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

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

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

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

- **How many Significant Figures are in a given Number**

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

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

Example 1.3

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

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

Example 1.4

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

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

(iv) How to Reduce Rounding Errors

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

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

Example 1.5

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

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

If we neglect the second decimal digit from each term and form their sum we find the approximate sum

$$S_1 = 2.3 + 1.4 + 4.2 = 7.9$$

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

A better approximation of S under the same limitations is

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

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

This error is significantly smaller than that in S_1.

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

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

The learner will soon know how to round numbers

What it means to round a number

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

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

Example 1.6

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

Note

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

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

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

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

Rules for Rounding Numbers

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

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

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

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

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

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

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

Example1.7

Rounding a given number correct to two significant figures

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

Formative Evaluation: Students should work through this exercise carefully writing full solutions for each problem. They should check their work thoroughly using the solutions provided.

Questions

1. (a) Using the method of substitution find the exact solution of the linear system of equations

$$\begin{aligned} 5x + 7y &= 12.075 \\ 7x + 10y &= 16.905 \end{aligned}$$

(b) Round the values on the right hand side of each equation to two significant figures and then find the exact solution of the resulting system of linear equation.

(c) Use the solutions obtained from the two systems of equations to explain why initial errors need to be avoided as much as possible.

2. (a) How many significant figures are in each of the following numbers:

(i) 00001000020000
(ii) 10000200003004
(iii) 000123.0004500

(b) Round each of the following numbers correct to five significant figures.

(i) 0123.395
(ii) 0123.205
(iii) 0123.206

3. Given the quantity $X = \left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}$, perform the following calculations:

(a) Find the exact value of X correct to five significant figures.

(b) Approximate value of X using three digit **chopping arithmetic** (rounding without making any adjustments).

(c) Approximate value of X using three digit rounding arithmetic.

(d) Calculate the absolute errors and percentage errors in the approximations obtained in parts (b) and (c).

Absolute and Relative Error

Let x^* be an approximation to an exact quantity X. Then,

The absolute error 'ε' $= |X - x^*|$ and

the relative/percentage error

$$\varepsilon_r = \left| \frac{X - x^*}{X} \right| \times 100$$

Example ①

The approximation $\tan x \simeq x + \frac{1}{3}x^3$, is valid for small values of 'x' in radians. Find the relative error in calculation of $\tan x$ at $x = 0.2$

$$\text{Exact value} = X = \tan(0.2) = 0.20271$$

$$\text{Approximate value} = X^* = x + \frac{1}{3}x^3 = 0.2 + \frac{1}{3}(0.2)^3 = 0.202667$$

$$\varepsilon_{rel} = \left| \frac{X - X^*}{X} \right| \times 100 = \left| \frac{0.20271 - 0.202667}{0.20271} \right| \times 100$$

$$= 0.0212126 \quad \text{Ans.}$$

Example (2)

Two students measure the temperature of the room with the true value of 20°C. Student A measures the temperature as 10°C and the student B measures it as 60°F. Determine the accuracy of these measurements and conclude which student has a better overall accuracy.

Solution

$$\text{Exact value} = X = 20^\circ C$$

$$\text{Approximate value (student A)} = X_A^* = 10^\circ C$$

$$\text{" " (student B)} = X_B^* = 60^\circ F = 15.56^\circ C$$

$$\text{Absolute error (A)} = \varepsilon_A = |X - X_A^*| = |20 - 10| = 10^\circ C.$$

$$\text{Absolute error (B)} = \varepsilon_B = |X - X_B^*| = |20 - 15.56| = 4.44^\circ C$$

Student (B) has better accuracy. Ans.

Example 3

The width of a rectangular piece of land is measured to be 48.25 ft. If the measurement has a relative error $\mathcal{E}_r$ of atmost 2%, then what is an upper bound for the absolute error?

Solution

Given $\mathcal{E}_r = 2\%$

We know $\mathcal{E}_r = \left| \frac{x - x^*}{x} \right| = \frac{|x - x^*|}{|x|}$

$$\mathcal{E}_r = \frac{\mathcal{E}}{|x|}$$

$$\mathcal{E} = \mathcal{E}_r |x| = \frac{2}{100} \times 48.25$$

$$\boxed{\mathcal{E} \leq 0.965}$$

Relative Error: Relative error gives an indication of how good a measurement is relative to the size of the thing being measured. Let's say that two students measure two objects with a meter stick. One student measures the height of a room and gets a value of 3.215 meters $\pm 1mm$ (0.001m). Another student measures the height of a small cylinder and measures 0.075 meters $\pm 1mm$ (0.001m). Clearly, the overall accuracy of the ceiling height is much better than that of the 7.5 cm cylinder. The comparative accuracy of these measurements can be determined by looking at their relative errors.

$$relative\ error = \frac{absolute\ error}{value\ of\ \ thing\ measured}$$

or in terms common to *Error Propagation*

$$relative\ error = \frac{\Delta x}{x}$$

where x is any variable. Now, in our example,

$$relative\ error_{ceiling\ height} = \frac{0.001m}{3.125m} \bullet 100 = 0.0003\%$$

$$relativeerror_{cylinder\ height} = \frac{0.001m}{0.075m} \bullet 100 = 0.01\%$$

Clearly, the relative error in the ceiling height is considerably smaller than the relative error in the cylinder height even though the amount of absolute error is the same in each case.

Assignment # 1

Degree / Syndicate: _____ NAME: _____ REGISTRATION No:

Q. 1 The absolute error of a measurement is equal to one half the unit of measure. Find the absolute error of each measurement. Then explain its meaning.

a. $2\frac{1}{8}$**ft** **b.** **4.81 mm** **c.** **3.9 mile**

Q. 2 The absolute error of a measurement is equal to one half the unit of measure. For an art show, a photographer makes prints that are each 1 foot tall. Find the relative error of this measurement.

Q. 3 Assume the series approximation of $\boldsymbol{Log(1+x)=x+\frac{1}{3}x^2}$. Compute the relative error in the approximation of $\boldsymbol{Log(1.2)}$

Q. 4 The approximation $\boldsymbol{\sqrt{1-x}=1-\frac{x}{2}-\frac{x^2}{8}}$, is valid for small values of $\boldsymbol{x}$ in radians. Find the relative error in the approximation of $\boldsymbol{\sqrt{1-x}}$ at $\boldsymbol{x = 0.5}$

Q. 5 The curve of a function $\boldsymbol{y = f(x)}$ passes through two points whose coordinates are **(0.2, 1.183)** and **(0.4, 1.342).** Approximate value of $\boldsymbol{y}$ at $\boldsymbol{x = 0.27}$.