Numerical and Statistical Methods

Dr. Jolly Puri
School of Mathematics
TIET, Patiala, Punjab, India

Introduction to Numerical Analysis

Numerical analysis is the study of algorithms for the problems of continuous mathematics.

(L. N. Trefethen)

- **NUMERICAL ANALYSIS** is the branch of mathematics that provides tools and methods for solving mathematical problems in numerical form.

- A major advantage of using numerical method is that a numerical answer can be obtained even when a problem has no analytical solution.

  1. Solve $\int_0^1 e^{-x^2}\, dx$
  2. Solve $x - e^{-x} = 0$
  3. Find zero(s) of $y = e^{\cos(x)} - 1.5$ on the interval [0, 2].
  4. For a matrix A of order *n*, find the largest eigen value in magnitude.

- In numerical analysis, we are mainly interested in implementation and analysis of numerical algorithms for finding an approximate solution to a mathematical problem.

- **NUMERICAL ALGORITHM** is a set of procedures which gives an approximate solution of a mathematical problem to a specified degree of accuracy.

- Criteria for a **GOOD** algorithm: 1) Applicable to a class of problems, 2) Speed of convergence, 3) Error management, and 4) Stability (how a numerical scheme propagate error).

- **NUMERICAL ITERATIVE METHOD** is the process of finding successive approximations, i.e., it generates a sequence of improved approximate solution for a class of problems.

- Besides the study of the numerical methods, the study of ***error analysis*** is equally important.

Error = Exact value – Approximate value

Types of Error

Let x = Exact value and $fl(x)$ = Approximate value,

then types of error is classified as:

1. Absolute Error = $|x - fl(x)|$
2. Relative Error = $\frac{|x - fl(x)|}{|x|}$
3. Percentage Error = $\frac{|x - fl(x)|}{|x|} \times 100$

Note: The relative and percentage errors are independent of units.

4. **Inherent Error**: It is that quantity which is already present in the statement of the problem before its solution. It arises either due to the simplified assumptions in the mathematical formulation of the problem or due to errors in the physical measurements of the parameters of the problem.

5. **Round-off Error**: It is the quantity which arises from the process of rounding off numbers. It sometimes also called numerical error.

6. **Truncation Error**: It occurs when an infinite process is replaced by a finite one, i.e., when a function is evaluated after truncating it at certain stage.

Find the largest interval in which $fl(x)$ must lie to approximate $\sqrt{2}$ with relative error at most 10^{-5} for each value of x.

Sol. We have

$$\left|\frac{\sqrt{2} - fl(x)}{\sqrt{2}}\right| \leq 10^{-5}.$$

Therefore

$$\begin{aligned}
|\sqrt{2} - fl(x)| &\leq \sqrt{2} \cdot 10^{-5}, \\
-\sqrt{2} \cdot 10^{-5} &\leq \sqrt{2} - fl(x) \leq \sqrt{2} \cdot 10^{-5} \\
-\sqrt{2} - \sqrt{2} \cdot 10^{-5} &\leq -fl(x) \leq -\sqrt{2} + \sqrt{2} \cdot 10^{-5} \\
\sqrt{2} + \sqrt{2} \cdot 10^{-5} &\geq fl(x) \geq \sqrt{2} - \sqrt{2} \cdot 10^{-5}.
\end{aligned}$$

Hence interval (in decimals) is $[1.4141994\cdots, 1.4142277\cdots]$.

Floating-point representation of numbers

Any real number is represented by an infinite sequence of digits. For example

$$\frac{8}{3} = 2.66666\cdots = \left(\frac{2}{10^1} + \frac{6}{10^2} + \frac{6}{10^3} + \ldots\right) \times 10^1.$$

This is an infinite series, but computer use an finite amount of memory to represent numbers. Thus only a finite number of digits may be used to represent any number, no matter by what representation method.

For example, we can chop the infinite decimal representation of $\frac{8}{3}$ after 4 digits,

$$\frac{8}{3} = \left(\frac{2}{10^1} + \frac{6}{10^2} + \frac{6}{10^3} + \frac{6}{10^4}\right) \times 10^1 = 0.2666 \times 10^1.$$

Generalizing this, we say that number has n decimal digits and call this n as precision.

For each real number x, we associate a floating point representation denoted by $fl(x)$, given by

$$fl(x) = \pm(0.a_1a_2\ldots a_n)_\beta \times \beta^e,$$

$$
\begin{aligned}
42.965 &= 4 \times 10^1 + 2 \times 10^0 + 9 \times 10^{-1} + 6 \times 10^{-2} + 5 \times 10^{-3} \\
&= 0.42965 \times 10^2. \\
-0.00234 &= -0.234 \times 10^{-2}.
\end{aligned}
$$

The above representation is not unique.

For example, $0.2666 \times 10^1 = 0.02666 \times 10^2$ *etc.*

Normal form

A non-zero floating-point number is in normal form if the values of mantissa lies in $(-1, -0.1]$ *or* $[0.1, 1)$.

Therefore, we normalize the representation by $a_1 \neq 0$. Not only the precision is limited to a finite number of digits, but also the range of exponent is also restricted. Thus there are integers m and M such that $-m \leq e \leq M$.

Rounding and chopping

Let x be any real number and $fl(x)$ be its machine approximation.

There are two ways to do the "cutting" to store a real number

$$x = \pm(0.a_1a_2\dots a_na_{n+1}\dots) \times 10^e, \quad a_1 \neq 0.$$

(1) Chopping: We ignore digits after a_n and write the number as following in chopping

$$fl(x) = (0.a_1a_2\dots a_n) \times 10^e.$$

(2) Rounding: Rounding is defined as following

$$fl(x) = \begin{cases} \pm(0.a_1a_2\dots a_n) \times 10^e, \ 0 \leq a_{n+1} < 5 & \text{(rounding down)} \\ \pm[(0.a_1a_2\dots a_n) + (0.00\dots01)] \times 10^e, \ 5 \leq a_{n+1} < 10 & \text{(rounding up).} \end{cases}$$

Example *The error in the measurement of area of a circle is not allowed to exceed* 0.5%. *How accurately the radius should be measured.*

Sol. Area of the circle is $A = \pi r^2$ (say).

$$\therefore \frac{dA}{dr} = 2\pi r.$$

Relative error (in percentage) in area is

$$\begin{aligned} \frac{dA}{A} \times 100 &\leq 0.5 \\ \implies dA &\leq \frac{0.5 \times A}{100} = \frac{0.5\pi r^2}{100} = \frac{1}{200}\pi r^2. \end{aligned}$$

Relative error (in percentage) in radius is therefore

$$\begin{aligned} \frac{dr}{r} \times 100 &= \frac{100}{r}\frac{dA}{\frac{dA}{dr}} \\ &\leq \frac{100}{r} \times \frac{\pi r^2}{200 \times 2\pi r} = 0.25. \end{aligned}$$