Course: Numerical Computing I
Supervisor: Ms. Shaista Rais
Department of Computer Science
University of Karachi

Chapter # 2

Error Analysis:

True error: difference between true value and observed or measured value is known as true error.

Error= True value – approximation

Types of error:

1) Absolute error: difference between true value and approximation or measured value that must be absolute is known as absolute error.

Eabs=| **True value – approximation |**

2) Relative error: is the ratio between absolute error and the true value

Er = | **True value – approximation | / true value**

3) Absolute relative approximate error: (or Percentage relative error)

Relative error when expressed in terms of percentage is known as percentage relative error or Absolute relative approximate error.

$$|Ea| = \frac{|True\ value - Observed\ Value|}{|True\ value|} * 100$$

Why do we measure errors :

1) with the help of measuring error we can find out the accuracy and efficiency of numerical methods.

2) by measuring errors we can use these errors as conditional criteria to restrict the number of iterations for iterative procedures.

Sources of Errors in Numerical Calculations

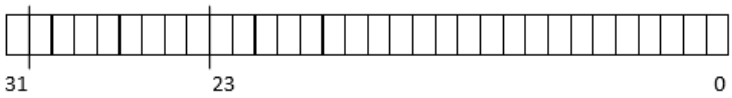
There are at least two sources of errors in numerical calculations:

1. Rounding Errors
2. Truncation Errors

Rounding errors originate from the fact that computers can only represent numbers using a fixed and limited number of significant figures. Thus, numbers such as π or $\sqrt{2}$ cannot be represented exactly in computer memory. The discrepancy introduced by this limitation is call **round-off error.** Even simple addition can result in round-off error. Often computers have the capacity to represent numbers in two different precisions, called **single** and **double** precision.

In **single precision**, 23 bits out of a total of 32 bits are used to represent the significant digits in the number. Of the remaining bits, 8 are used to store the exponent and one bit is used to store the sign.

Bit Organization in a Single Precision Number:



With 23 bits to represent the significant figures, a single precision number can represent data to about seven decimal places. The 8 bit exponent allows scaling in the range 10^{38} and 10^{-38}. Taken together, single precision numbers can range from $\pm 1.175494351 \times 10^{-38}$ to $\pm 3.4028235 \times 10^{38}$. In single precision, π can be represented as 3.141593.

Due to the limited number of significant digits in single precision, most modern computers use **double precision** which uses **64 bits** with **52 digits** used to represent the significant figures. This allows π, for example, to be represented as 3.141592653589793, that is, 16 digits. The full range for number representation using double precision is $\pm 2.2250738585072020 \times 10^{-308}$ to $\pm 1.7976931348623157 \times 10^{308}$.

The use of double precision reduces the effects of rounding error and should be used whenever possible in numerical calculations.

Truncation errors in numerical analysis arise when approximations are used to estimate some quantity. Often a Taylor series is used to approximate a solution which is then truncated. The figure below shows a function $f(x_i)$ being approximated by a Taylor series that has been truncated at different levels. The more terms that are retained in the Taylor series the better the approximation and the smaller the truncation error.

Course: Numerical Computing I
Supervisor: Ms. Shaista Rais
Department of Computer Science
University of Karachi

Taylor Series:

$$f(x_{i+1}) = f(x_i) + f'(x_i)\,h + \frac{f''(x_i)}{2!}\,h^2 + \ldots.$$