Chapter 3

NUMERICAL ANALYSIS

3.1 Introduction

3.1 Introduction. The various areas of mathematics provide theoretical models. Numerical analysis deals with performing the necessary arithmetical manipulations to obtain numerical results to the theoretically formulated problems. Because irrational numbers can't be written in decimal form, computations always carry errors. Using correct algorithms would provide acceptable results, that is, results where the error is negligible. Another point of using numerical analysis is to build a model that can be performed with the possible minimal number of computations. A few decades ago, before the computer era, computations had to be performed by hand. Long, tedious, never-ending calculations had to be performed in the fulfillment of projects. Nowadays, fast, accurate ways to find results are performed by computers. Yet the challenge remains: the new state of the art technology requires endless calculations that sometimes even the fastest supercomputer have trouble to conclude. An entirely new area of numerical analysis was then born; the finding of algorithms and shortcuts for computer programming. In this course we will see just an introduction to the old and the new technology.

3.2 Numbers

Numbers exist since ever. It is known that some animals, including birds, can count. Archeological research has shown that as far as twenty thousand years ago humans (or whatever they were at that time) have already developed some kind of numeral systems. They used only the natural numbers: 1, 2, 3, …. Rational numbers appeared later. The Egyptians and the Babylonians have already introduced the rational numbers. The Greek were the first to discover the existence of numbers that cannot be written as a fraction or quotient or two natural numbers. These are the irrational numbers.

3.2.1 Types of numbers.

Rational numbers. A rational number is a quotient of two natural numbers, they are best known by the word “fractions”, so a rational number is of the form p/q, where p and q are integer numbers. Rational numbers have either finite or infinite periodic decimal expansion.

$$\frac{7}{4}=1.75, \quad \frac{7}{11}=0.636363.....$$

Irrational numbers. There are numbers that cannot be represented as a fraction of two integers. They are called *irrational numbers*. The first irrational number discovered was $\sqrt{2}$. Now it is known that for any positive integer *n* either *n* is a perfect square numbers (*n* m^2, such as 4, 9, 16, etc.) or it is irrational. Irrational numbers have infinite non-periodic decimal expansions.

$$\pi = 3.141592654\cdots \quad \sqrt{2} = 1.414213562\cdots$$

This means that an irrational number cannot be written in decimal form. But there is a theorem in mathematics that states that any irrational number can be approximated as much as we wish by a sequence or rational numbers. So, in practice, each time that you need to perform calculations with irrational numbers, we will substitute it by a rational number "close enough" to it, so the error in the final calculation would be negligible. For machine computation, every number must be replaced by a finite sequence of digits. Therefore, in writing numbers, computers may generate errors.

3.3 Representation of Numbers

Computers can't represent infinite sequences of numbers. Because irrational numbers have infinite decimal expansion, computers cannot calculate irrational numbers.

Number of significant digits of a number. Is the number of digits, excluding zeros on the left.

EXAMPLE. The numbers 2537, 1.200, 0.01325, all have four significant digits.

Computers can represent numbers in two ways

Fixed point. All numbers have a fixed number of decimal places (impractical, inefficient).

EXAMPLE. 121.3514, 0.0012, 3.0000, all these numbers are represented with four decimals.

Floating point. The number of significant digits, and not the number of decimals, is kept fixed.

$65.3451 = 0.653451\times10^{2} = 0.653451E02$, 6 significant digits

$0.001312 = 0.1312\times10^{-2} = 0.1312E-02$, 4 significant digits

$-3 = -0.3\times10^{1} = -0.3E01$, 1 significant digit

$0.00000000013 = 0.13\times10^{-9} = 0.13E-09$, 2 significant digits

Any number *a* is of the form

$$a = \pm m\times10^{e}, \quad 0.1 \le m < 1, \quad e \text{ integer.}$$

But computers do not have unlimited capabilities, numbers must be rounded, so they can only represent numbers as

. $$\overline{a} = \pm \overline{m}\times10^{e}, \quad \overline{m} = 0.d_1 d_2 ... d_t, \quad d_i > 0 \quad |e| < M$$

With $\overline{m}$ being limited to a fixed number of t significant digits, and e a bounded integer. We call $\overline{m}$ the mantissa and e is called the exponent. There are two ways of approximate (round) a number a by $\overline{a}$.

Chopping. Discard all decimals from some decimal on (imprecise). 0.34518 $\longrightarrow$ 0.3451 when chopped to 4 decimals.

Rounding-off. Round to the nearest unit. 0.34518 $\longrightarrow$ 0.3452; 3.45 $\longrightarrow$ 3.4; 3.452 $\longrightarrow$ 3.5

Rounding errors. There are two types, the true error, ε, and the relative error, ε_r, defined by

$$\varepsilon = |\, a - \overline{a}\,|, \quad \varepsilon_r = \left|\frac{a-\overline{a}}{a}\right| \tag{1}$$

Substituting in the definition above a and $\overline{a}$ by their expressions using mantissa,

$$\varepsilon = |\, m - \overline{m}\,| \times 10^{e}, \quad \varepsilon_r = \left|\frac{m-\overline{m}}{m}\right| \tag{2}$$

EXAMPLE. Round (to the nearest unit) the number $a = 53.013407 = 0.53013407\times 10^2$, to 5 significant digits and calculate the error and the relative error

$$\overline{a} = 53.013 = 0.53013\times 10^2, \quad \varepsilon = 0.000407, \quad \varepsilon_r = \frac{0.000407}{53.013407} = 0.0000076773$$

Notice in the example above that

$$\varepsilon < 0.0005 = \frac{1}{2}\times 10^{2-5} = \frac{1}{2}\times 10^{-3} = \beta$$

and

$$|\, m - \overline{m}\,| = |\, .53013407 - .53013\,| = 0.00000407 \le 0.000005 = 5\times 10^{-6} = \frac{5}{10}\times 10^{-6}\times 10 = \frac{1}{2}\times 10^{-5} = \beta$$

ε is **the error** and β is the **error bound**. In general, if a number is rounded to t significant digits, then

$$\boxed{|\, m - \overline{m}\,| \le \frac{1}{2}\times 10^{-t}} \tag{3}$$

Dividing by m and using the fact that $m \ge 0.1 = 10^{-1}$, so $1/m \le 10$, we have

$$\varepsilon_r = \left|\frac{a-\overline{a}}{a}\right| = \left|\frac{m-\overline{m}}{m}\right| \le \frac{1}{2}\times 10^{-t}\times 10 = \frac{1}{2}\times 10^{1-t}$$

Therefore $\boxed{\varepsilon_r \le \frac{1}{2}\times 10^{1-t} = \beta_r}$ (relative error bound) (4)

Combining (2) and (3) we obtain $\boxed{\varepsilon \le \frac{1}{2}\times 10^{e-t} = \beta}$ (error bound) (5)

The quantities on the right side of (4) and (5) are called error bounds, since it is known that the error is at most these values. In the case of the example above, with $t = 5$ and $e = 2$, we have,

$$\varepsilon = 0.000407 \le 0.0005 = 5\times 10^{-4} = \frac{5}{10}\times 10^{-3} = \frac{1}{2}\times 10^{-3} = \frac{1}{2}\times 10^{2-5}$$

as in (5), and also

$$\varepsilon_r = \frac{0.000407}{53.013407} = \frac{0.407\times10^{-3}}{0.53013407\times10^{2}} \le \frac{0.5\times10^{-5}}{0.1} = \frac{(1/2)\times10^{-5}}{10^{-1}} = \frac{1}{2}\times10^{1-5}$$

as in (4). Thus, in our example, the quantities $\frac{1}{2}\times10^{-3}$ and $\frac{1}{2}\times10^{-4}$ are the total error bound and the relative error bound respectively. We denote the total error bound by β and the relative error bound by β_r.

(include examples and problems)

3.3 Error Propagation

Undesirable outcomes may occur in solving problems that require large amounts of calculations. A negligible error repeated a great number of times can become a large final error. In a problem that requires a large amount of computations, if a consistent small error is carried out in each computation, the final result may have an unacceptable large error.

THEOREM. (a) In addition and subtraction, an error bound for the results is given by the sum of the error bounds for the terms.

(b) In multiplication and division, a bound for the relative error is given (approximately) by the sum of the bounds for the relative errors of the factors.

Let $S = a_1$ …..

EXAMPLE. First, calculate the sum S = 1.234560 + 0.01398999 + 3.312000 + 1.199999. Second, write the same numbers with 5 significant digits precision and sum again. Third, calculate the error. Then, find an error bound for the sum and verify that the error is less than or equal to that error bound

S = 5.7605490

The sum after rounding to 5 significant digits is

S_1 = 0.12345E01 + 0.13989E-01 + 0.33120E01 + 0.11999E01 = 0.576039E01 = 5.76039

The error is $\varepsilon = |S - S_1| = 0.000159$

The error bounds are 0.5×10^{-4} for the first, third and fourth numbers, and 0.5×10^{-6} for the second number. The error bound for the sum is $2\times10^{-4} + 0.5\times10^{-6} = 0.0002005$. The total error ε, therefore, is less than the error bound for the sum.

HOMEWORK

1. Let $a = 123.40176$.
(a) Write a in scientific notation and find the mantissa m and the exponent e.
(b) Round a to 4 significant digits and find $\overline{a}$
(c) Find error bound ε and relative error bound ε_r.

2. (a) Write the following numbers in scientific notation rounding to 4 significant digits.
1425.023100 0.00019800 80000001

(b) Sum the numbers and find the error bound of the sum

SOLUTIONS

1. Let $a = 123.40176$.

(a) Write a in scientific notation and find the mantissa m and the exponent e.

$a = 0.12340176\text{E}03 \quad m = 0.12340176, \quad e = 3$

(b) Round a to 4 significant digits and find $\overline{a}$

$\overline{a} = .1234\text{E}03 \quad \overline{m} = .1234 \quad e = 3$

(c) Find error bound ε and relative error bound ε_r.

$\beta = \frac{1}{2}10^{-1}, \quad \beta_r = \frac{1}{2}10^{-3}$