Errors in Numerical Processes

Calculation errors

Numerical methods involve computations. Computations introduce calculation errors.

Error

If the true value of a number is x and this is approximated by x^* then the error is defined as

$$\varepsilon = x - x*$$

The absolute error is the size of the error – that is, its modulus

$$|\varepsilon| = |x - x*|$$

The relative error is

$$\text{relative error} = \frac{\text{error}}{\text{true value}} = \frac{\varepsilon}{x}$$

This means that relative error has a sign (+ or $-$), so the absolute relative error is

$$\left|\frac{\varepsilon}{x}\right|$$

Recording numbers

Numbers recorded with fixed point notation have their decimal point fixed. For example

9876. 54

In floating point notation, which is also known as scientific notation, the number takes the form

$$a \times 10^b$$

where $|a|$ lies between 1 and 10 and b is an integer. For example

9.87654×10^{3}

It is usual to express very large and very small numbers using the floating point notation. In floating point notation only the significant digits are included.

Rounding errors

In decimal notation the number

$$\frac{2}{3} = 0.\dot{6}$$

The dot indicates that the 6 recurs infinitely

$$\frac{2}{3} = 0.\dot{6} = 0.66666.......$$

which can also be shown by dots following the number. Calculators and machines cannot hold more than a fixed number of significant digits, or, if fixed point notation is used, decimal places. Therefore, at some point the number will be "chopped off". For example

$$\frac{2}{3} = 0.666667$$

to 6 decimal places. This is called "rounding off" and introduces a maximum rounding error. Here, when $2/3$ is approximated by 0.666667 the maximum rounding error is 0.0000005 or 5×10^{-7}. Rounding errors also occur when floating point notation is used. For example, when

9.87654×10^{3}

is rounded to 3 significant figures

9.88×10^{3}

the maximum rounding error is $0.5 \times 10^{-3} \times 10^{3} = 5$

Propagation of errors

When operations are performed on pairs of numbers (for example, adding pairs of numbers, multiplying pairs of numbers) the errors in the existing numbers translate into an error in the resultant number. Errors grow as a result of calculations. As a result, we need some rules to prevent numbers being quoted to a higher degree of accuracy than is strictly justified.

Addition and subtraction of numbers

When two numbers in fixed point notation are added, the number of decimal places in the sum should be no more than the number of decimal places in the original number with the least number of decimal places. For example

$$52.379 + 63.1 = 116.5\left(1.D.P.\right)$$

That is, $52.379\left(3.D.P.\right) + 63.1\left(1.D.P.\right)$ gives a sum quoted to 1 decimal place - $116.5\left(1.D.P.\right)$.

Multiplication and division of numbers

The number of significant figures in the product should be no more than the number of significant figures in the number with the lower number of significant figures. For example

$$52.379 \times 63.1 = 3310\left(3.S.F.\right)$$

That is, the product of a number, 52.379 with 5 significant figures, and a number, 63.1, with 3 significant figures, is a number 3310 with 3 significant figures.

Subtraction of nearly equal quantities

This is a rule of thumb to prevent errors creeping into calculations where one might be tempted to misapply the rules for the addition and subtraction of numbers. The rule is

> If two positive numbers of nearly equal size are subtracted then you should bear in mind that there may be a significant loss of accuracy in the answer.

For example, when evaluating

$$\begin{aligned} 33.572(382.561-15.441\times 24.773) &= 33.572(382.561-382.52) && (1)\\ &= 33.572\times 0.041 && (2)\\ &= 1.4(2.S.F.) \end{aligned}$$

The answer can only be quoted to 2 significant figures. At line (1) the product (15.441×24.773) is correctly quoted to 5 significant figures (382.52). At line (2) the sum $(382.561-382.520)$ is also correctly quoted to 3 decimal places (0.041). However, this sum (0.041) is now only accurate to 2 significant figures, so the final product (1.4) can also only be given to 2 significant figures.

The point of the rule is that as the original numbers are all accurate to 5 *S.F.* one might suppose the result can also be quoted to 5 *S.F.*, but this is not the case when two nearly equal numbers are subtracted one from the other, and, as the example illustrates, this eventuality can be disguised since the product (15.441×24.773) is not obviously nearly equal to the number from which it is subtracted.

Ill-conditioned problems

Problems are ill-conditioned if a small change in the input data leads to a large change in the output data.

<u>Example</u>

By solving the two pairs of simultaneous equations

$$\begin{aligned} &(1) \quad && y = x\\ & && y = 0.9999x+1\\ &(2) && y = x\\ & && y = 1.0001x+1 \end{aligned}$$

show that the system

$$\begin{aligned} y &= x\\ y &= \alpha x+1 \end{aligned}$$

is ill-conditioned when $\alpha \to 1$

Solution

The solution to the first pair is

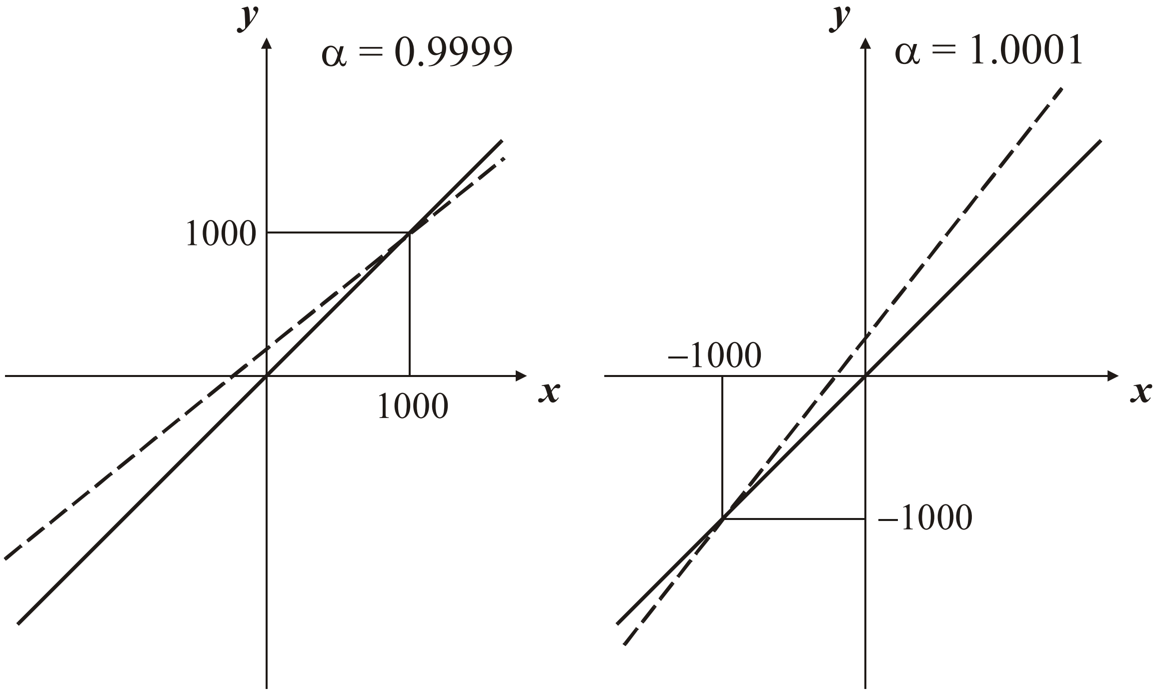
$$x = 0.9999x + 1$$
$$x = 10,000$$
$$y = 10,000$$

The solution to the second pair is

$$x = 1.0001x + 1$$
$$x = -10,000$$
$$y = -10,000$$

so a small change in the coefficient α leads to a large change in the solution.

The reason why is that the gradients of the two lines $y = x$ and $y = \alpha x + 1$ are almost the same as $\alpha \to 1$ so small changes in α lead to very large changes in the solutions.


y
α = 0.9999
1000
1000
x
y
α = 1.0001
–1000
x
–1000