

Error

Numerical methods are intrinsically limited by error. The question is how large an error. There are two sources of error involved in the calculation process itself and three further sources in the accompanying activities.

Round-off Error

Computers can represent only quantities with a finite number of digits, the number of which depending on how many bits (possible 1's or 0's) used to describe each quantity. This has the obvious result of limiting the size of number possible (in the same way one runs out of space for 9's on a calculator screen), but also that only a finite number of subdivisions of the range is possible. This is of far greater consequence as it results in quantising error due to numbers having to be chopped or rounded to the nearest division. An advantage of using floating-point (scientific) numbers is that this quantising error is proportional to the magnitude of the number, resulting in a constant relative error – the machine epsilon or machine limit.

Increasing the number of bits used to define each number (by using Double Precision) allows larger numbers and smaller subdivision giving smaller quantising error. There is of course a penalty in increased memory usage and reduced computation speed. There are particular operations which make this penalty worth paying and which even with double precision give cause for concern:

Interdependent Computations: If many calculations are performed sequentially using the result of the previous calculation as input to the present calculation, quantising error can accumulate to give errors very much larger than the machine epsilon.

Adding a Large and a Small Number: In the process of adding two numbers the smaller number is modified so that its exponent matches that of the larger number, this can result in moving the significant figures of the smaller number out of the range of significant figures carried by the computer. As a result the contribution of the small number is lost. This kind of error is often encountered when summing series, it can be avoided by summing the series in reverse so that the cumulative total grows with the magnitude of the terms.

Subtractive Cancellation: Subtraction is performed in the same manner as addition, hence when two very similar numbers are subtracted the difference can be lost.

Smearing: Occurs whenever individual terms in a summation are larger than the summation itself.

Inner Products (Dot Products): This is a series widely used in matrix based numerical methods. There is little to be done about rounding error other than using double precision.

Truncation Error

Truncation error is the discrepancy introduced by the fact that numerical methods may use approximations to represent exact mathematical functions and quantities.

Other Sources

Blunders, formulation or model errors and data uncertainty contribute to error. These must be considered in the design phase. For example in modelling gas flows use of the ideal gas approximation introduces a model error and uncertainty in gas property data introduces data uncertainty, both of which must be evaluated. Mistakes in measurement of physical dimensions, data entry, etc contribute to blunder error.

Accuracy and Precision

In considering error we are in fact amalgamating two types of error – inaccuracy and imprecision. To use an analogy, accuracy can be considered to constitute the distance of an arrow from the centre of a target and precision the scatter of a series of arrows across the target. The objective is to repeatedly get the arrow close to the centre of the target.

Definition of Error

For comparison of methods we are interested in relative error rather than magnitude:

$$\varepsilon = \frac{Value - Approximation}{Value} \times 100\%$$

This error can be relative to the actual value if it is known or to an estimated value if it is not. Choice of the estimated value is clearly vital to error definition, but is not straightforward. For iterative methods error is often considered to be the difference between successive values:

$$\varepsilon = \frac{Current \cdot Approximation - Previous \cdot Approximation}{Current \cdot Approximation} \times 100\%$$

Iterative methods require a convergence criterion at which to stop the code once the required error reduction has been achieved. If the permissible error is expressed by specifying a number of significant figures* to which the answer is to be correct, it can be shown (Scarborough, 1966) that this accuracy is achieved when:

$$|\varepsilon| < (0.5 \times 10^{2-n})\%$$

*Significant Figures:

Significant Digits - When reading a scale the number of significant digits is the number of digits explicitly marked on the scale plus one figure to indicate the position between the finest markings. This final figure is traditionally zero or five. For example on a speedometer with markings for each mile per hour the number of significant figures is the value of the nearest marking plus a zero to indicate that the needle is higher than it or a five to indicate that the needle is lower than it.

Significant figures: The number of significant figures given for a quantity is used as an indicator of the confidence in the value, in accordance with the understanding of significant digits given above. There is a complication however involving zeros. Zeros may be added before or after a number to position the decimal point: 0.00567, 0.000567 & 0.567 all have 3 significant figures. Similarly 567000, 56.7 & 56700 also have 3 significant figures. Use of scientific notation eliminates this uncertainty by eliminating the need for zeros to position the decimal point: 5.67×10^4 clearly has 3 significant figures, as would any similar number with different exponents.

© Ben Todd 2002
Last Updated: August 19, 2003