Types of Errors

Solving a problem on a computer generally involves errors. There are errors in the data, the mathematical model, and the numerical algorithm used to approximate solutions. As numerical analysts we are primarily concerned with the numerical algorithm.

Errors in a numerical algorithm stem from the fact that both the computer memory and speed of calculation is finite. Finiteness is the cause of all difficulties. The problems in continuous mathematics involve infinite dimensional spaces, differential and integral operators, and continuous evolution in time. Finiteness leads us to represent the continuum of the real numbers by a finite set of floating point numbers. Similarly it forces us approximate the space of continuous functions $f : \mathbf{R} \to \mathbf{R}$ by some finite dimensional subspace. Finite speed of calculation encourages us to employ coarser approximations that can be solved more quickly. Often the time constraint is more important than available memory.

Consider a mathematical model M which takes v for its input and returns $M(v)$ as its output. Let v^* represent an approximation of the input v and let M^* represent a numerical approximation of M. Then, we may classify the errors as

approximation	approximand	type of error
v^*	v	initial error
$M(v^*)$	$M(v)$	propagated error
$M^*(v^*)$	$M(v^*)$	generated error
$M^*(v^*)$	$M(v)$	total cumulative error

To measure the error in an approximation we need a notion of how far v^* is from v. In the case that v is complicated, like the current state of the atmosphere, this can be a difficult problem in itself. Generally, we desire some sort of metric or norm to measure the errors with. In this section we simplify things by only studying errors that arise directly from the approximation of the real numbers and the operations of addition and multiplication.

Given $x \in \mathbf{R}$ we denote by x^* the best (correctly rounded) approximation of x to within whatever constraints the specific floating point representation under consideration allows. To determine how far x^* is from x, define the absolute and relative errors as

$$e = x^* - x \qquad \text{and} \qquad \tilde{e} = \frac{x^* - x}{x}.$$

Note that the the size of the absolute error depends on the dimensional units in which x is expressed, whereas the relative error is dimensionless ratio.

We first consider mathematical addition $M(x, y) = x + y$. Let $x, y \in \mathbf{R}$. Note that x and y must be expressed in the same dimensional units for the sum $x + y$ to make sense. Therefore the absolute errors measured for x and y are dimensionally compatible. Let

$$e_x = x^* - x \qquad \text{and} \qquad e_y = y^* - y.$$

Then the absolute propagated error

$$e_{\text{prop}} = (x^* + y^*) - (x + y) = (x^* - x) + (y^* - y) = e_x + e_y.$$

Let $+^*$ be the approximation of $+$ discussed in Lecture 02. By definition, the absolute generated error is given by

$$e_{\text{gen}} = (x^* +^* y^*) - (x^* + y^*).$$

Therefore, the absolute total cumulative error

$$\begin{aligned} e_{\text{tot}} &= (x^* +^* y^*) - (x + y) = (x^* +^* y^*) - (x^* + y^*) + (x^* + y^*) - (x + y) \\ &= e_{\text{gen}} + e_{\text{prop}}. \end{aligned}$$

Next we consider mathematical multiplication $M(x, y) = xy$ with the numerical approximation $x * y = (xy)^*$. Since multiplication makes sense when x and y are expressed using different dimensional units, then we shall consider relative errors.

$$\tilde{e}_x = \frac{x^* - x}{x} \qquad \text{and} \qquad \tilde{e}_y = \frac{y^* - y}{y}.$$

The relative propagation error is

$$\begin{aligned} \tilde{e}_{\text{prop}} &= \frac{x^*y^* - xy}{xy} = \frac{x^*y^* - x^*y}{xy} + \frac{x^*y - xy}{xy} = \frac{x^*}{x}\tilde{e}_y + \tilde{e}_x \\ &= \Big(\frac{x^* - x}{x} + 1\Big)\tilde{e}_y + \tilde{e}_x = \tilde{e}_x\tilde{e}_y + \tilde{e}_x + \tilde{e}_y. \end{aligned}$$

By definition, the relative generated error is

$$\tilde{e}_{\text{gen}} = \frac{x^* * y^* - x^*y^*}{x^*y^*}.$$

Therefore, the relative total cumulative error

$$\begin{aligned} \tilde{e}_{\text{tot}} &= \frac{x^* * y^* - xy}{xy} = \frac{x^* * y^* - x^*y^*}{xy} + \frac{x^*y^* - xy}{xy} = \frac{x^*y^*}{xy}\tilde{e}_{\text{gen}} + \tilde{e}_{\text{prop}} \\ &= \frac{x^*}{x}\Big(\frac{y^* - y}{y} + 1\Big)\tilde{e}_{\text{gen}} + \tilde{e}_{\text{prop}} = (\tilde{e}_x + 1)(\tilde{e}_y + 1)\tilde{e}_{\text{gen}} + \tilde{e}_{\text{prop}} \\ &= (\tilde{e}_{\text{prop}} + 1)\tilde{e}_{\text{gen}} + \tilde{e}_{\text{prop}} = \tilde{e}_{\text{prop}}\tilde{e}_{\text{gen}} + \tilde{e}_{\text{gen}} + \tilde{e}_{\text{prop}}. \end{aligned}$$

We end this lecture noting that similar results hold for subtraction and division. However, as Trefethen stated,

Numerical analysis is [not] *the study of rounding errors.*

Have a good labor day weekend!