PROPAGATION OF ERROR

Suppose we are evaluating a function $f(x)$ in the machine. Then the result is generally not $f(x)$, but rather an approximate of it which we denote by $\widetilde{f}(x)$. Now suppose that we have a number $x_A \approx x_T$. We want to calculate $f(x_T)$, but instead we evaluate $\widetilde{f}(x_A)$. What can we say about the error in this latter computed quantity?

$$f(x_T) - \widetilde{f}(x_A) = [f(x_T) - f(x_A)] + \left[f(x_A) - \widetilde{f}(x_A)\right]$$

The quantity $f(x_A) - \widetilde{f}(x_A)$ is the "noise" in the evaluation of $f(x_A)$ in the computer, and we will return later to some discussion of it. The quantity $f(x_T) - f(x_A)$ is called the <u>propagated error</u>; and it is the error that results from using perfect arithmetic in the evaluation of the function.

If the function $f(x)$ is differentiable, then we can use the "mean-value theorem" to write

$$f(x_T) - f(x_A) = f'(\zeta)\,(x_T - x_A)$$

for some ζ between x_T and x_A.

Since usually x_T and x_A are close together, we can say ζ is close to either of them, and

$$f(x_T) - f(x_A) \approx f'(x_T)\,(x_T - x_A)\,, \qquad \text{(*)}$$

Example. Define

$$f(x) = b^x$$

where b is a positive real number. Then (*) yields

$$b^{x_T} - b^{x_A} \approx (\log b)\, b^{x_T}\,(x_T - x_A)$$

$$\begin{aligned} \mathrm{Rel}\,(b^{x_A}) &\approx x_T\,(\log b)\,(x_T - x_A)\,/x_T \\ &= x_T\,(\log b)\,\mathrm{Rel}\,(x_A) \\ &= K \cdot \mathrm{Rel}\,(x_A) \end{aligned}$$

with $K = x_T\,(\log b)$. Note that $K = 10^4$ and $\mathrm{Rel}(x_A) = 10^{-7}$, then $\mathrm{Rel}(b^{x_A}) \approx 10^{-3}$. This is a large decrease in accuracy; and it is independent of how we actually calculate b^x. The number K is called a condition number for the computation.

PROPAGATION IN ARITHMETIC OPERATIONS

Let ω denote arithmetic operation such as $+$, $-$, $*$, or $/$. Let ω^* denote the same arithmetic operation as it is actually carried out in the computer, including rounding or chopping error. Let $x_A \approx x_T$ and $y_A \approx y_T$. We want to obtain $x_T \omega y_T$, but we actually obtain $x_A \omega^* y_A$. The error in $x_A \omega^* y_A$ is given by

$$\begin{aligned} x_T \omega y_T - x_A \omega^* y_A = \; & [x_T \omega y_T - x_A \omega y_A] \\ & + [x_A \omega y_A - x_A \omega^* y_A] \end{aligned}$$

The final term is the error is introduced by the inexactness of the machine arithmetic. For it, we usually assume

$$x_A \omega^* y_A = fl\,(x_A \omega y_A)$$

This means that the quantity $x_A \omega y_A$ is computed exactly and is then rounded or chopped to fit the answer into the floating point representation of the machine.

The formula

$$x_A \omega^* y_A = fl\left(x_A \omega y_A\right)$$

implies

$$x_A \omega^* y_A = \left(x_A \omega y_A\right)(1+\varepsilon) \qquad \text{(**)}$$

with limits given earlier for ε. Manipulating (**), we have

$$\mathsf{Rel}\left(x_A \omega^* y_A\right) = -\varepsilon$$

With rounded binary arithmetic having n digits in the mantissa,

$$-2^{-n} \leq \varepsilon \leq 2^{-n}$$

The term

$$x_T \omega y_T - x_A \omega y_A$$

is the <u>propagated error</u>; and we now examine it for particular cases.

Consider first $\omega = *$. Then for the relative error in $x_A * y_A \equiv x_A y_A$,

$$\text{Rel}\,(x_A y_A) = \frac{x_T y_T - x_A y_A}{x_T y_T}$$

Write

$$x_T = x_A + \xi, \qquad y_T = y_A + \eta$$

Then

$$\begin{aligned}
\text{Rel}\,(x_A y_A) &= \frac{x_T y_T - x_A y_A}{x_T y_T} \\
&= \frac{x_T y_T - (x_T - \xi)(y_T - \eta)}{x_T y_T} \\
&= \frac{x_T \eta + y_T \xi - \xi\eta}{x_T y_T} \\
&= \frac{\xi}{x_T} + \frac{\eta}{y_T} - \frac{\xi}{x_T} \cdot \frac{\eta}{y_T} \\
&= \text{Rel}\,(x_A) + \text{Rel}\,(y_A) - \text{Rel}\,(x_A) \cdot \text{Rel}\,(y_A)
\end{aligned}$$

Since we usually have

$$|\mathsf{Rel}\,(x_A)|\,,\ |\mathsf{Rel}\,(y_A)| \ll 1$$

the relation

$$\mathsf{Rel}\,(x_A y_A) = \mathsf{Rel}\,(x_A) + \mathsf{Rel}\,(y_A) - \mathsf{Rel}\,(x_A) \cdot \mathsf{Rel}\,(y_A)$$

says

$$\mathsf{Rel}\,(x_A y_A) \approx \mathsf{Rel}\,(x_A) + \mathsf{Rel}\,(y_A)$$

Thus small relative errors in the arguments x_A and y_A leads to a small relative error in the product $x_A y_A$. Also, note that there is some cancellation if these relative errors are of opposite sign.

There is a similar result for division:

$$\mathsf{Rel}\left(\frac{x_A}{y_A}\right) \approx \mathsf{Rel}\,(x_A) - \mathsf{Rel}\,(y_A)$$

provided

$$|\mathsf{Rel}\,(y_A)| \ll 1$$

ADDITION AND SUBTRACTION

For ω equal to $-$ or $+$, we have

$$[x_T \pm y_T] - [x_A \pm y_A] = [x_T - x_A] \pm [y_T - y_A]$$

Thus the error in a sum is the sum of the errors in the original arguments, and similarly for subtraction. However, there is a more subtle error occurring here.

Suppose you are solving

$$x^2 - 26x + 1 = 0$$

Using the quadratic formula, we have the true answers

$$r_T^{(1)} = 13 + \text{sqrt}(168), \qquad r_T^{(2)} = 13 - \text{sqrt}(168)$$

From a table of square roots, we take

$$\text{sqrt}(168) \doteq 12.961$$

Since this is correctly rounded to 5 digits, we have

$$|\text{sqrt}(168) - 12.961| \leq .0005$$

Then define

$$r_A^{(1)} = 13+12.961 = 25.961, \quad r_A^{(2)} = 13-12.961 = .039$$

Then for both roots,

$$|r_T - r_A| \leq .0005$$

For the relative errors, however,

$$\text{Rel}\left(r_A^{(1)}\right) \leq \frac{.0005}{25.9605} \doteq 3.13 \times 10^{-5}$$

$$\text{Rel}\left(r_A^{(2)}\right) \leq \frac{.0005}{.0385} \doteq .0130$$

Why does $r_A^{(2)}$ have such poor accuracy in comparison to $r_A^{(1)}$?

The answer is due to the loss of significance error involved in the formula for calculating $r_A^{(2)}$. Instead, use the mathematically equivalent formula

$$r_T^{(2)} = \frac{1}{13 + \text{sqrt}(168)} \doteq \frac{1}{25.961}$$

This results in a much more accurate answer, at the expense of an additional division.