GENERATION OF ERRORS IN DIGITAL COMPUTATION

ALSTON S. HOUSEHOLDER

Consider the problem of computing the numerical value $f(x)$ of some function f corresponding to a particular value of x. If x is a physical quantity whose value is known only from measurement, then one has available not x itself but some approximation x^*. If x is a mathematical constant, one may again be forced to replace x by some approximate value x^*. This is to be expected at least when x is irrational, and is often true when x is rational. For these or other reasons, though $f(x)$ is desired, one may have to accept some $f(x^*)$ instead.

Unless the function f is rational, one will generally be forced to represent f approximately by a truncated Taylor series, or an orthogonal series, or in some other fashion. In general, therefore, one does not strictly compute even $f(x^*)$, but rather some $f_a(x^*)$, where f_a represents a function which approximates f over some range containing x^*.

Suppose that f_a is defined by a finite sequence of elementary arithmetic operations, and that the computations are digital. In general the result of a division is not representable exactly by a terminating decimal. The same statement holds, of course, if one were using any fixed base. The result of multiplying two numbers, each expressible by n digits, will in general require $2n$ digits for its representation and further multiplications would increase the number of digits in proportion. Hence for either a product, or a quotient, one will in general replace the true result by an approximation obtained by truncating the sequence of digits and perhaps adjusting by some rule the last one retained. Hence one does not, in general, end up even with $f_a(x^*)$, but rather with some $f^*(x^*)$, in which true products and quotients are replaced by pseudo products and pseudo quotients, obtained according to rules that are determined in part by the nature of the facilities used for the computing.

Thus one starts to compute $f(x)$ and ends by computing some $f^*(x^*)$, thereby committing an error whose amount can be expressed as a sum of three independent components:

$$\begin{aligned} f(x) - f^*(x^*) = {} & [f(x) - f(x^*)] + [f(x^*) - f_a(x^*)] \\ & + [f_a(x^*) - f^*(x^*)]. \end{aligned} \tag{1}$$

An address delivered before the Spartanburg meeting of the Society on November 28, 1953 by invitation of the Committee to Select Hour Speakers for Southeastern Sectional Meetings; received by the editors December 4, 1953.