ON THE PRECISION ATTAINABLE WITH VARIOUS FLOATING-POINT NUMBER SYSTEMS

RICHARD P. BRENT

Abstract

For scientific computations on a digital computer the set of real number is usually approximated by a finite set F of "floating-point" numbers. We compare the numerical accuracy possible with difference choices of F having approximately the same range and requiring the same word length. In particular, we compare different choices of base (or radix) in the usual floating-point systems. The emphasis is on the choice of F, not on the details of the number representation or the arithmetic, but both rounded and truncated arithmetic are considered. Theoretical results are given, and some simulations of typical floating-point computations (forming sums, solving systems of linear equations, finding eigenvalues) are described. If the leading fraction bit of a normalized base-2 number is not stored explicitly (saving a bit), and the criterion is to minimize the mean square roundoff error, then base 2 is best. If unnormalized numbers are allowed, so the first bit must be stored explicitly, then base 4 (or sometimes base 8) is the best of the usual systems.

Comments

Only the Abstract is given here. The full paper appeared as [1].

References

[1] R. P. Brent, "On the precision attainable with various floating-point number systems", *IEEE Transactions on Computers* C–22 (1973), 601–607. CR 14#25960, Zbl 261.65036. Also appeared as Report TR RC 3751, IBM Research, Yorktown Heights, New York (February 1972), 28 pp. rpb017.

Mathematical Sciences Department, IBM T. J. Watson Research Center, Yorktown Heights, New York 10598

Current address: Computer Centre, Australian National University, Canberra, ACT, Australia

1991 *Mathematics Subject Classification.* Primary 65G05; Secondary 65Y99, 68M07, 68P20.

Key words and phrases. Base, floating-point arithmetic, radix, representation error, rounding error, rms error, simulation.

Manuscript received May 15, 1972.

rpb017a typeset using $\mathcal{AMS}$-LaTeX.