A COMPARISON OF PRE-ROUNDING AND POST-ROUNDING FLOATING POINT DIVIDE ORDERS

PAUL C. KETTLER

ABSTRACT. This paper addresses the advantages and disadvantages of pre-rounding *vs.* post-rounding in a floating-point register. In this day when high speed computers with long registers are the norm, one may think that a rounding scheme is nearly irrelevant, because rounding cannot improve over the addition of a single bit of increased accuracy. However, the scheme is important precisely because of the speed of computation, and the extended computations that speed allows. The propagation of error then becomes a significant issue, and the scheme of rounding is the starting point for making any such subsequent analysis.

1. INTRODUCTION

Some manufacturers of scientifically oriented digital computers have in recent years introduced shift-in pre-rounding as a means of reducing or eliminating truncation error bias in floating point divide orders. The appeal of this procedure is that no time is needed for post operative rounding and possible subsequent renormalizing. The prices for convenience, however, appear in the form of increased error bound and increased error variance.

This paper addresses certain probabilistic aspects of the pre-rounding paradigm. Focus concentrates on a standard floating scheme, without special attention to differing modes of computation, *e.g.*, sign-magnitude, two's complement, nor to special cases, *e.g.*, overflow, underflow, unnormalization, nor to variants in keeping low-order accuracy, *e.g.*, guard digits, sticky bits. Such cases are treated in several of the references cited next.

For background reading on this subject see these books (Wilkinson 1963; Feller 1967; Feller 1971), these papers (Ashenhurst 1965; Gregory 1966; Nickel 1966; Cody 1967; Urabe 1968; Knödel 1968; Knödel 1968; Matula 1969; Reinsch 1979; Harrison 2006), these Ph.D. theses (Jacobi 2002; Boldo 2004), and these standards (IEEE 1985; IEEE 1987).

2. STATEMENT OF THE PROBLEM

Let

$$2^a\left(f_H + 2^{-k} f_L\right) \quad \text{and} \quad 2^b\left(g_H\right)$$

be normalized floating point dividend and divisor, respectively. The subscripts H and L signify high and low order word portions and k represents high order register length. In single precision calculations the fraction f_L is absent, but may be supplied nonzero by a

Date: 3 March 2008.

2000 *Mathematics Subject Classification.* Primary: 60G50. Secondary: 65Y20.

1991 *Journal of Economic Literature Subject Classification.* C15, C63.

Key words and phrases. roundoff error, error bounds, error propagation.

The author wishes to thank those many colleagues and friends over the years from Princeton University, the University of Chicago, the University of California at Berkeley, and the University of Oslo, who contributed to his valuable computing experiences.

processor to induce modification of the quotient. This process, if successful in reducing error bias, may properly be called rounding.

The exact quotient is in one of the following two forms depending on the relationship between f_H and g_H.

$$Q = \begin{cases} 2^{a-b}\left(\dfrac{f_H}{g_H} + \dfrac{2^{-k}f_L}{g_H}\right) & \text{if} \quad \dfrac{1}{2} \leqslant \dfrac{f_H}{g_H} < 1 \\ 2^{a-b+1}\left(\dfrac{f_H}{2g_H} + \dfrac{2^{-k}f_L}{2g_H}\right) & \text{if} \quad 1 \leqslant \dfrac{f_H}{g_H} < 2 \end{cases}$$

Assume f_H and g_H are uniformly and independently distributed in their logarithms, base 2, *e.g.*,

$$\begin{aligned} \Pr\{f_H \leqslant x\} &= \Pr\{\log_2 f_H \leqslant \log_2 x\} \\ &= \log_2 x + 1 \end{aligned}$$

Ample theoretical and empirical evidence exists for this assumption, although at least one author has reported otherwise, probably erroneously. Setting

$$\alpha = \log_2 f_H \quad \text{and} \quad \beta = \log_2 g_H,$$

the mean and variance of the increment to the quotient as a result of introducing f_L are, respectively,

$$\begin{aligned} \mu &= \left[\int_{-1}^{0}\int_{-1}^{\beta} 2^{-\beta}\,d\alpha\,d\beta + \int_{-1}^{0}\int_{\beta}^{0} 2^{-(\beta+1)}\,d\alpha\,d\beta\right] f_L \\ &= \frac{1}{2\log_e^2 2}\, f_L \approx 1.040684\, f_L \end{aligned}$$

$$\begin{aligned} \sigma_2 &= \left\{\left[\int_{-1}^{0}\int_{-1}^{\beta} 2^{-2\beta}\,d\alpha\,d\beta + \int_{-1}^{0}\int_{\beta}^{0} 2^{-2(\beta+1)}\,d\alpha\,d\beta\right] - \mu^2\right\} f_L \\ &= \frac{9\log_e^2 2 - 4}{16\log_e^4 2}\, f_L^2 \approx 0.087746\, f_L^2 \end{aligned}$$

See Figure 1 for a diagrammatic interpretation.

Note also that the posterior distribution of f_H/g_H is uniform in the logarithm, base 2. Specifically,

$$\begin{aligned} \Pr\{f_H/g_H \leqslant x\} &= \Pr\{\log_2 f_H \leqslant \log_2 g_H + \log_2 x\} \\ &= \Pr\{\alpha \leqslant \beta + \log_2 x\} \\ &= \int_{-(\log_2 x+1)}^{0}\int_{-1}^{\beta+\log_2 x} d\alpha\,d\beta + \left[\frac{1}{2} - \int_{-1}^{-(\log_2 x+1)}\int_{\beta+\log_2 x+1}^{0} d\alpha\,d\beta\right] \\ &= \log_2 x + 1 \end{aligned}$$

See Figure 2 for a diagrammatic interpretation.

3. Analysis of the results

For rounding without bias it is desirable to set $\mu = 1/2$, the expected loss from truncation, given long registers. The immediate consequence is that $f_L = \log_e^2 2 \approx 0.480453$. Also, $\sigma^2 = (9 \log_e^2 2 - 4)/16 \approx 0.020255$.

For comparison with the case of post-rounding, it is necessary to look at the statistics in that event. The mean error is zero; the variance is that of the uniform distribution on $(-1/2, 1/2]$, *i.e.*, $1/12 \approx 0.083333$, and the absolute error bound, both lower and upper, is $1/2$. Observe that the entire variance is introduced by truncation, following an add of $1/2$.

We may now easily compute the composite statistics for shift-in rounding followed by terminal truncation. The mean error is zero; the variance has increased to $\sigma^2 + 1/12 = (27 \log_e^2 2 - 8)/48 \approx 0.103588$; the absolute error bound has increased to $|f_L/2 - 1| \approx 0.759773$ (lower) and $|2f_L| \approx 0.960906$ (upper). The rationale for adding the variances above is not that the distributions are independent. In fact, the amount of truncation is completely determined given the arguments of the division. Rather, the explicit assumption is that the covariance be zero.

4. Conclusion

Nothing is gained but perhaps time by dividend pre-rounding. Post-rounding has both (a) lower error variance, and (b) lower error bounds, than pre-rounding, under the chosen distribution assumptions.

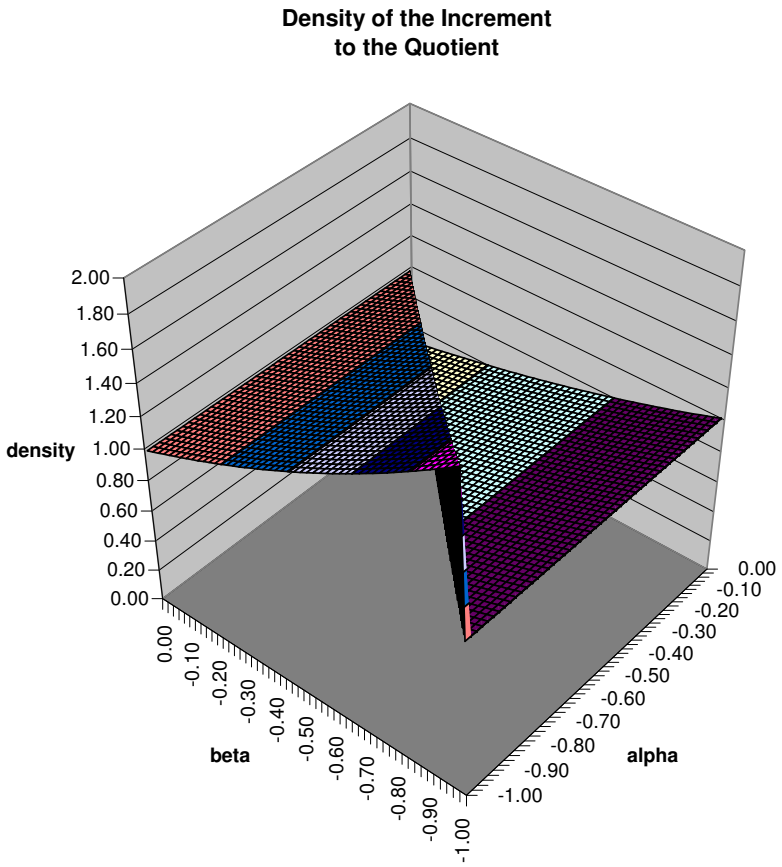

FIGURE 1. Density of the increment to the quotient as a result of introducing f_L

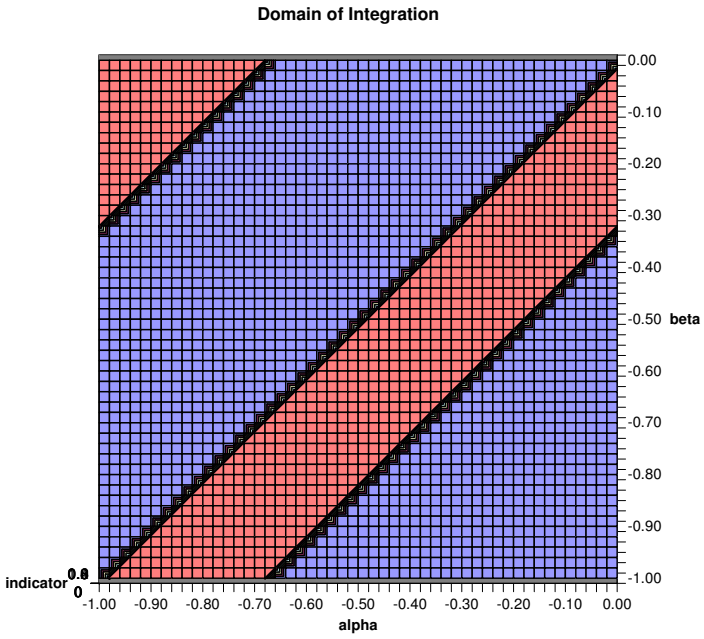

FIGURE 2. Domain of integration of $\Pr\{\alpha \leqslant \beta + \log_2 x\}$, for $\log_2 x = -\frac{2}{3}$

REFERENCES

Ashenhurst, R. L. (1965). Experimental investigation of unnormalized arithmetic. *Error in Digital Computation 2*, 3–37. Mathematics Research Center, U. S. Army, Madison, Wisconsin.

Boldo, S. (2004, Nov.). *Preuves formelles en arithmétiques à virgule flottante*. Ph. D. thesis, École Normale Supérieure de Lyon. Directeur: Marc Daumas.

Cody, W. J. (1967). The influence of machine design on numerical algorithms. In *Proceedings of the Spring Joint Computer Conference*, Volume 30, pp. 305–309. American Federation of Information Processing Societies.

Feller, W. K. (1967). *Introduction to Probability Theory and Its Applications* (3rd ed.), Volume I. New York: Wiley.

Feller, W. K. (1971). *Introduction to Probability Theory and Its Applications* (2nd ed.), Volume II. New York: Wiley.

Gregory, R. T. (1966, April). On the design of the arithmetic unit of a fixed-word-length computer from the standpoint of computational accuracy. In *Transactions on Electronic Computers*, Number 2, pp. 255–257. Institute of Electrical and Electronics Engineers.

Harrison, J. (2006). *Floating-point verification using theorem proving*, pp. 212–242. Lecture Notes in Computer Science: Formal Methods for Hardware Verification. Berlin/Heidelberg: Springer.

IEEE (1985). Standard for binary floating-point arithmetic. ANSI/IEEE standard 754-1985. Technical report, The Institute of Electrical and Electronic Engineers, 345 East 47th Street, New York, NY 10017.

IEEE (1987). Standard for radix-independent floating-point arithmetic. ANSI/IEEE standard 854-1987. Technical report, The Institute of Electrical and Electronic Engineers, 345 East 47th Street, New York, NY 10017.

Jacobi, C. (2002, Apr.). *Formal verification of a fully IEEE compliant floating point unit*. Ph. D. thesis, University of the Saarland. Dekan: Philipp Slusallek.

Knödel, W. (1968). Über die Verteilung der Binarziffern einer gemessenen Grösse. *Computing 3*(4), 354–361.

Matula, D. W. (1969). Towards an abstract mathematical theory of floating-point arithmetic. In *Proceedings of the Spring Joint Computer Conference*, Volume 34, pp. 765–772. American Federation of Information Processing Societies.

Nickel, K. (1966, Nov.). Über die Notwendigkeit einer Fehlerschranken-Arithmetik für Rechenautomaten. *Numer. Math. 9*(1), 69–79.

Reinsch, C. H. (1979). Principles and preferences for computer arithmetic. *SIGNUM Newsl. 14*(1), 12–27.

Urabe, M. (1968, Jun.). Roundoff error distribution in fixed-point multiplication and a remark about the rounding rule. *SIAM J. Numer. Anal. 5*(2), 202–210.

Wilkinson, J. H. (1963). *Rounding Errors in Algebraic Processes*. Englewood Cliffs: Prentice-Hall.

(Paul C. Kettler)
CENTRE OF MATHEMATICS FOR APPLICATIONS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OSLO
P.O. BOX 1053, BLINDERN
N–0316 OSLO
NORWAY
E-mail address: paulck@math.uio.no
URL: http://www.math.uio.no/~paulck/