MATHEMATICS OF COMPUTATION
Volume 87, Number 310, March 2018, Pages 803–819
http://dx.doi.org/10.1090/mcom/3234
Article electronically published on July 7, 2017

ON RELATIVE ERRORS OF FLOATING-POINT OPERATIONS: OPTIMAL BOUNDS AND APPLICATIONS

CLAUDE-PIERRE JEANNEROD AND SIEGFRIED M. RUMP

Abstract. Rounding error analyses of numerical algorithms are most often carried out via repeated applications of the so-called standard models of floating-point arithmetic. Given a round-to-nearest function fl and barring underflow and overflow, such models bound the relative errors $E_1(t) = |t - \mathrm{fl}(t)|/|t|$ and $E_2(t) = |t - \mathrm{fl}(t)|/|\mathrm{fl}(t)|$ by the unit roundoff u. This paper investigates the possibility and the usefulness of refining these bounds, both in the case of an arbitrary real t and in the case where t is the exact result of an arithmetic operation on some floating-point numbers. We show that $E_1(t)$ and $E_2(t)$ are optimally bounded by $u/(1+u)$ and u, respectively, when t is real or, under mild assumptions on the base and the precision, when $t = x \pm y$ or $t = xy$ with x, y two floating-point numbers. We prove that while this remains true for division in base $\beta > 2$, smaller, attainable bounds can be derived for both division in base $\beta = 2$ and square root. This set of optimal bounds is then applied to the rounding error analysis of various numerical algorithms: in all cases, we obtain significantly shorter proofs of the best-known error bounds for such algorithms, and/or improvements on these bounds themselves.

1. Introduction

Given two integers $\beta, p \geqslant 2$, let $\mathbb{F}$ be the associated set of floating-point numbers having base β, precision p, and no restriction on the exponent range:

$$\mathbb{F} = \{0\} \cup \{M \cdot \beta^e \,:\, M, e \in \mathbb{Z},\ \beta^{p-1} \leqslant |M| < \beta^p\}.$$

Also, let $\mathrm{fl} : \mathbb{R} \to \mathbb{F}$ denote any round-to-nearest function, such that

$$|t - \mathrm{fl}(t)| = \min_{f \in \mathbb{F}} |t - f|, \qquad t \in \mathbb{R}. \tag{1}$$

In particular, no specific tie-breaking strategy is assumed for the function fl. Two *relative errors* can then be defined, depending on whether the exact value or the rounded value is used to divide the absolute error in (1): the error relative to t is

$$E_1(t) = \frac{|t - \mathrm{fl}(t)|}{|t|} \qquad \text{if } t \neq 0,$$

while the error relative to $\mathrm{fl}(t)$ is

$$E_2(t) = \frac{|t - \mathrm{fl}(t)|}{|\mathrm{fl}(t)|} \qquad \text{if } \mathrm{fl}(t) \neq 0.$$

(In each case the relative error may be defined to be zero if the denominator is zero: the exponent range being unbounded, $\mathrm{fl}(t) = 0$ implies $t = 0$, so in both cases a zero denominator means that no error occurs when rounding t.)

Received by the editor April 20, 2016 and, in revised form, October 26, 2016.
2010 *Mathematics Subject Classification.* Primary 65G50.

For rounding error analysis purposes, the most commonly used bounds are $E_1(t) \leqslant u$ and $E_2(t) \leqslant u$, where

$$u = \frac{1}{2}\beta^{1-p}$$

is the *unit roundoff* associated with fl and $\mathbb{F}$, and such bounds are typically handled via the so-called *standard models* $\mathrm{fl}(t) = t(1+\delta_1) = t/(1+\delta_2)$, $|\delta_1|, |\delta_2| \leqslant u$; see [5, pp. 38–39]. It is also known that the bound on the first relative error can be refined slightly [12, p. 232],[1] so that

$$E_1(t) \leqslant \frac{u}{1+u} \qquad \text{and} \qquad E_2(t) \leqslant u. \tag{2}$$

These worst-case bounds hold for any real number t and any round-to-nearest function fl. Furthermore, they can be regarded as *optimal* in the sense that each of them is attained for some pair (t, fl) with t expressed in terms of β and p: since $1+u$ is exactly halfway between the two consecutive elements 1 and $1+2u$ of $\mathbb{F}$, we have $E_1(1+u) = u/(1+u)$ and, assuming further that fl rounds ties "to nearest even", $E_2(1+u) = u$.

In this paper, we investigate the possibility and the usefulness of refining the bounds in (2) when t is not just an arbitrary real number but the exact result of an operation on some floating-point number(s), that is, when

$$t = x \,\mathrm{op}\, y, \qquad x, y \in \mathbb{F}, \qquad \mathrm{op} \in \{+, -, \times, /\} : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$$

as well as $t = \sqrt{x}$.

Our first contribution is to establish optimal bounds on both E_1 and E_2 for each of these five basic operations, as shown in Table 1.

TABLE 1. Optimal relative error bounds for various inputs t.

t	bound on $E_1(t)$	bound on $E_2(t)$
real number	$\frac{u}{1+u}$	u
$x \pm y$	$\frac{u}{1+u}$	u
xy	$\frac{u}{1+u}$	u
x/y	$\begin{cases} u - 2u^2 & \text{if } \beta = 2, \\ \frac{u}{1+u} & \text{if } \beta > 2 \end{cases}$	$\begin{cases} \frac{u-2u^2}{1+u-2u^2} & \text{if } \beta = 2, \\ u & \text{if } \beta > 2 \end{cases}$
$\sqrt{x}$	$1 - \frac{1}{\sqrt{1+2u}}$	$\sqrt{1+2u} - 1$

As we shall see later in the paper, each of these bounds is attained for some explicit input values in $\mathbb{F}$ and rounding functions fl, possibly under some mild (necessary and sufficient) conditions on β and p. Specifically, for addition, subtraction, and multiplication the condition for optimality is that β is *even*, and in the case of multiplication in base 2 it is that $2^p + 1$ is *not* a Fermat prime. In most practical situations such conditions are satisfied and thus the general bounds in (2) remain

[1] The refined bound $E_1(t) \leqslant u/(1+u)$ already appears in [18, p. 74] and, in some special cases, in [2] for $\beta = 2$ and [6] for β even.

the best ones for floating-point addition, subtraction, and multiplication; as Table 1 shows, this is also the case of division in any base larger than 2. In contrast, for division in base 2 and for square root, the general bound $u/(1+u) \approx u-u^2$ on E_1 can be decreased further to $u-2u^2$ and to $1-(1+2u)^{-1/2} \approx u-\frac{3}{2}u^2$, respectively; likewise, the general bound u on E_2 can be decreased to $(u-2u^2)/(1+u-2u^2) \approx u-3u^2$ and $(1+2u)^{1/2}-1 \approx u-\frac{1}{2}u^2$.

Our second contribution is to show that in the context of rounding error analysis of numerical algorithms, applying these optimal bounds in a systematic way leads to simpler and sharper bounds, and/or to more direct proofs of existing ones.

In particular, this allows us to establish the following three new error bounds:

- For the summation of n floating-point numbers $x_1, \ldots, x_n$ (using $n-1$ floating-point additions and any ordering), we show that the resulting floating-point approximation $\widehat{s}$ satisfies

$$\left|\widehat{s} - \sum_{i=1}^{n} x_i\right| \leqslant \sum_{i=1}^{n-1} |e_i| \leqslant \frac{(n-1)u}{1+u} \sum_{i=1}^{n} |x_i|, \tag{3}$$

  where $|e_i|$ denotes the absolute error of the ith floating-point addition.
- For the summation of n real numbers $x_1, \ldots, x_n$, we show that by first rounding each x_i into $\mathrm{fl}(x_i)$ and then summing the $\mathrm{fl}(x_i)$ in any order, the resulting approximation $\widehat{s}$ satisfies

$$\left|\widehat{s} - \sum_{i=1}^{n} x_i\right| \leqslant \sum_{i=1}^{n} |d_i| + \sum_{i=1}^{n-1} |e_i| \leqslant \zeta_n \sum_{i=1}^{n} |x_i|, \qquad \zeta_n < nu, \tag{4}$$

  where the d_i are given by $d_i = x_i - \mathrm{fl}(x_i)$ and the e_i are as before.
- For the Euclidean norm of a vector of n floating-point numbers $x_1, \ldots, x_n$ we show that summing the squares in any order and then taking the square root of the result yields a floating-point number $\widehat{r}$ such that

$$\widehat{r} = \Big(\sum_{i=1}^{n} x_i^2\Big)^{1/2} \cdot (1+\epsilon), \qquad |\epsilon| \leqslant (n/2+1)u. \tag{5}$$

Note that each of these bounds holds without any restriction on n. The bounds in (3) and (4) improve upon the best previous ones, from [10], in two ways: they are sharper and apply not only to the absolute error of $\widehat{s}$ but also to the sum of the absolute *local* errors. Furthermore, we will see that the new bound in (3) implies the one in (4) almost immediately; in other words, using (3) allows us to recover the constant nu established in [10, Proposition 4.1] for sums of n reals (and thus n-dimensional inner products as well), but in a much more direct way. Finally, the bound in (5) nicely replaces the expression $(n/2+1)u + \mathcal{O}(u^2)$ that would result from using the suboptimal bound $E_1(\sqrt{x}) \leqslant u$.

Besides the evaluation of sums and norms, and to illustrate further the benefits of applying the refined bounds in Table 1, we provide four other typical examples of rounding error analysis. These examples deal with small arithmetic expressions, minors of tridiagonal matrices, Cholesky factorization, and complex floating-point multiplication. We shall see that in each case the existing error bound can be either replaced by a simpler and sharper one, or recovered via a significantly shorter proof.

Notation and assumptions. All our results hold under the customary assumption that $\beta, p \geqslant 2$. Furthermore, following [5] (and unless stated otherwise) we shall ignore underflow and overflow by assuming that the exponent range of $\mathbb{F}$ is unbounded. Finally, the common tool used to establish all the error bounds in Table 1 is the function $\mathrm{ufp} : \mathbb{R} \to \mathbb{F}_{\geqslant 0}$ from [16], called *unit in the first place* (ufp) and defined as follows: $\mathrm{ufp}(0) = 0$ and, if $t \in \mathbb{R}\backslash\{0\}$, $\mathrm{ufp}(t)$ is the largest integer power of β such that $\mathrm{ufp}(t) \leqslant |t|$. Hence, in particular, for t nonzero,

$$\mathrm{ufp}(t) \leqslant |t| < \beta\mathrm{ufp}(t), \tag{6a}$$

and for any t,

$$\mathrm{ufp}(t) \leqslant |\mathrm{fl}(t)| \leqslant \beta\mathrm{ufp}(t). \tag{6b}$$

Outline. This paper is organized as follows. We begin in section 2 by recalling how to derive the bounds in (2) and by completely characterizing their attainability. Section 3 then gives proofs for all the bounds announced in Table 1 together with explicit expressions of input values at which these bounds are attained. A first application of these results is described in section 4, where we establish the new bounds for summation shown in (3) and (4). We conclude in section 5 with the derivation of the bound (5) together with the analysis of four other examples.

2. Optimal error bounds when rounding a real number

Using the function ufp, the bounds $E_i(t) \leqslant u$ for $i = 1, 2$ are easily derived as follows. First, recall from (6) that $|t| \neq 0$ belongs to the right-open interval $[\mathrm{ufp}(t), \beta\,\mathrm{ufp}(t))$, which contains $(\beta - 1)\beta^{p-1}$ equally-spaced elements of $\mathbb{F}$. The distance between two such consecutive elements is thus $\frac{\beta\,\mathrm{ufp}(t) - \mathrm{ufp}(t)}{(\beta-1)\beta^{p-1}} = 2u\,\mathrm{ufp}(t)$, and rounding to nearest implies that the absolute error is bounded as

$$|t - \mathrm{fl}(t)| \leqslant u\,\mathrm{ufp}(t). \tag{7}$$

This bound is sharp and the values at which it is attained are the *midpoints* of $\mathbb{F}$, that is, the rational numbers lying exactly halfway between two consecutive elements of $\mathbb{F}$.

Dividing both sides of (7) by either $|t|$ or $|\mathrm{fl}(t)|$ gives

$$E_1(t) \leqslant u\frac{\mathrm{ufp}(t)}{|t|} \qquad \text{and} \qquad E_2(t) \leqslant u\frac{\mathrm{ufp}(t)}{|\mathrm{fl}(t)|}, \tag{8}$$

and by using the lower bounds in (6) we arrive at the classical bounds $E_1(t) \leqslant u$ and $E_2(t) \leqslant u$. As noted in [5, Theorem 2.2], we have in fact the strict inequality $E_1(t) < u$, since t cannot be at the same time a midpoint and equal to its ufp; on the other hand, the derivation above shows that the bound on $E_2(t)$ is attained if and only if t is a midpoint such that $|\mathrm{fl}(t)| = \mathrm{ufp}(t)$.

Let us now refine the bound $E_1(t) < u$. As shown in [12, p. 232], all we need for this is a lower bound on $|t|$ slightly sharper than the one in (6): by definition of rounding to nearest, $|t - \mathrm{fl}(t)| \leqslant |t - f|$ for all $f \in \mathbb{F}$, so taking in particular $f = \mathrm{sign}(t)\mathrm{ufp}(t)$ gives $|t - \mathrm{fl}(t)| \leqslant |t| - \mathrm{ufp}(t)$; in other words,

$$\mathrm{ufp}(t) + |t - \mathrm{fl}(t)| \leqslant |t| \tag{9}$$

and, because of (7), equality occurs if and only if $|t| \leqslant (1+u)\mathrm{ufp}(t)$. Thus, by applying (9) and then (7) to the definition of E_1, we find that

$$E_1(t) \leqslant \frac{|t - \mathrm{fl}(t)|}{\mathrm{ufp}(t) + |t - \mathrm{fl}(t)|} \leqslant \frac{u}{1+u}.$$

Furthermore, due to the conditions of attainability of (7) and (9) given above, this bound on $E_1(t)$ is attained if and only if t is a midpoint such that $|t| \leqslant (1+u)\mathrm{ufp}(t)$, that is, if and only if $|t| = (1+u)\mathrm{ufp}(t)$.

We summarize our discussion in the theorem below. Although the bounds given there already appear in [18, p. 74] and [12, p. 232] for E_1, and in [5, Theorem 2.3] for E_2, the characterization of their attainability does not seem to have been reported elsewhere.

Theorem 2.1. *If $t \in \mathbb{R}\backslash\{0\}$, then*

$$E_1(t) \leqslant \frac{u}{1+u} \qquad \text{and} \qquad E_2(t) \leqslant u.$$

Furthermore, the bound on E_1 is attained if and only if $|t| = (1+u)\mathrm{ufp}(t)$, and the bound on E_2 is attained if and only if $|t| = (1+u)\mathrm{ufp}(t)$ and $|\mathrm{fl}(t)| = \mathrm{ufp}(t)$.

3. Optimal error bounds for floating-point operations

We establish here optimal bounds on both E_1 and E_2 for the operations of addition, subtraction, multiplication, fused multiply-add, division, and square root.

3.1. Addition, subtraction, and fused multiply-add.

When t has the form $t = x + y$ with x, y in $\mathbb{F}$, we show in the theorem below that the general bounds $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ given in (2) remain optimal unless the basis β is odd. This extends the analysis done by Holm [6], who considers only the first relative error and assumes implicitly that the basis is even.

Theorem 3.1. *Let $\beta, p \geqslant 2$ and $t = x + y$ with $x, y \in \mathbb{F}$. The bounds in (2) are optimal if and only if β is even. Furthermore, when β is even, they are attained for $(x, y) = (1, u)$ and rounding "to nearest even".*

Proof. If β is odd, then $u = \frac{1}{2}\beta^{1-p} = \sum_{i=0}^{\infty} \frac{\beta-1}{2}\beta^{-i-p}$ with $\frac{\beta-1}{2} \in \{1, \ldots, \beta-1\}$, so u and $1+u$ have infinite expansions in base β. Hence $x + y$ cannot have the form $\pm(1+u)\beta^e$ with $e \in \mathbb{Z}$ and thus, by Theorem 2.1, equality never occurs in the bounds $E_1(x+y) \leqslant u/(1+u)$ and $E_2(x+y) \leqslant u$.

If β is even, then u is in $\mathbb{F}$. Consequently, we can take $(x, y) = (1, u)$, which gives $E_1(x+y) = u/(1+u)$ and, for ties rounded "to nearest even", $E_2(x+y) = u$. □

Since $1 + 2u$ belongs to $\mathbb{F}$ for any β and since $1 + u = (1+2u) - u = 1 \times 1 + u$, the theorem above extends immediately to subtraction as well as to higher-level operations encompassing addition, like the fused multiply-add $(x, y, z) \mapsto \mathrm{fl}(xy+z)$.

3.2. Multiplication.

When $t = xy$ with $x, y \in \mathbb{F}$, the theorem below shows that the situation is more subtle than for addition: although the condition that β is even remains necessary for the optimality of the bounds $E_1(t) \leqslant u/(1+u)$ and $E_2(t) \leqslant u$, it is sufficient only when $\beta \neq 2$; when $\beta = 2$, optimality turns out to be equivalent to $2^p + 1$ being composite.

Theorem 3.2. *Let $\beta, p \geqslant 2$ and $t = xy$ with $x, y \in \mathbb{F}$. We then have the following, depending on the value of the base β:*

- *If $\beta = 2$, then the bounds in (2) are optimal if and only if $2^p + 1$ is not prime; furthermore, if D is a nontrivial divisor of 2^p+1, then these bounds are attained for*
$$(x, y) = \left(\frac{(2 + 2u)\mathrm{ufp}(D)}{D}, \frac{D}{\mathrm{ufp}(D)} \right)$$
and rounding "to nearest even".
- *If $\beta > 2$, then the bounds in (2) are optimal if and only if β is even, and when the latter is true these bounds are attained for*
$$(x, y) = (2 + 2u, 2^{-1})$$
and rounding "to nearest even".

Before proving this result, note that the attainability of the bound $E_1(xy) \leqslant u/(1 + u)$ has been observed in [6, p. 10] in the particular case $(\beta, p) = (2, 5)$, by taking x and y in the form shown above with $D = 11$.

Proof. By Theorem 2.1, the optimality of the bounds in (2) when $t = xy$ is equivalent to the existence of a pair $(x, y) \in \mathbb{F} \times \mathbb{F}$ such that $|xy| = (1 + u)\mathrm{ufp}(xy)$.

Assume first that $\beta > 2$. Similarly to Theorem 3.1, a necessary condition for optimality is that β be even. Furthermore, if β is even, then both $2 + 2u$ and 2^{-1} are in $\mathbb{F}$, and since their product equals $1+u$, it suffices to take $(x, y) = (2+2u, 2^{-1})$ to show that the bounds in (2) are optimal.

Let us now consider the case $\beta = 2$. Since $\mathrm{ufp}(t \cdot 2^e) = \mathrm{ufp}(t) \cdot 2^e$ for all $(t, e) \in \mathbb{R} \times \mathbb{Z}$, we can assume with no loss of generality that $1 \leqslant x, y < 2$. This implies that $\mathrm{ufp}(xy) \in \{1, 2\}$ and, since $x, y \in \{1, 1 + 2u, 1 + 4u, \ldots\}$ and $u > 0$, that the product xy cannot be equal to $1 + u$. Hence, optimality is equivalent to the existence of $x, y \in \mathbb{F} \cap [1, 2)$ such that $xy = 2 + 2u$, that is, equivalent to the existence of integers X, Y such that

$$XY = (2^p + 1) \cdot 2^{p-1} \qquad \text{and} \qquad 2^{p-1} \leqslant X, Y < 2^p. \tag{10}$$

If $2^p + 1$ is prime, then either X or Y must be larger than 2^p, so (10) has no solution.

If 2^p+1 is composite, one can construct a solution (X_0, Y_0) to (10) as follows. Let D denote a nontrivial divisor of 2^p+1, and let $X_0 = \frac{2^p+1}{D}\mathrm{ufp}(D)$ and $Y_0 = D\frac{2^{p-1}}{\mathrm{ufp}(D)}$. Clearly, X_0 is an integer and the product X_0Y_0 has the desired shape. Thus, it remains to check that $Y_0 \in \mathbb{Z}$ and that both X_0 and Y_0 are in the range $[2^{p-1}, 2^p)$. Since $2^p + 1$ is odd, D must be odd too, which implies that

$$\mathrm{ufp}(D) + 1 \leqslant D < 2\mathrm{ufp}(D) \qquad \text{and} \qquad D < 2^p. \tag{11}$$

Consequently, $\mathrm{ufp}(D) \leqslant 2^{p-1}$, so that $\mathrm{ufp}(D)$ divides 2^{p-1} and Y_0 is an integer. Furthermore, (11) leads to $2^{p-1} < \frac{2^p+1}{2} < X_0 \leqslant (2^p + 1)(1 - \frac{1}{D}) < 2^p$ and $2^{p-1} < Y_0 < 2^p$, so that X_0 and Y_0 satisfy the range constraint in (10).

Finally, multiplying X_0 and Y_0 by $2u = 2^{1-p}$ gives $x_0 = (2 + 2u)\mathrm{ufp}(D)/D$ and $y_0 = D/\mathrm{ufp}(D)$ in $\mathbb{F} \cap [1, 2)$ and such that $x_0y_0 = 2 + 2u = (1 + u)\mathrm{ufp}(x_0y_0)$. □

In the rest of this section, we show that the optimality condition "$2^p + 1$ is not prime" arising in Theorem 3.2 for radix two is in fact satisfied in most practical situations.

First of all, if p is *not* a power of two, this condition is well known to hold [3, Exercise 18, §4], since p can be factored as $p = mq$ with q odd and then

$$2^p + 1 = (2^m + 1)\big(2^{p-m} - 2^{p-2m} + \cdots - 2^m + 1\big). \tag{12}$$

The binary *basic* formats specified by the IEEE 754-2008 standard being such that $p \in \{24, 53, 113\}$, they fall into this category. More precisely, since here p is either odd or a multiple of three, we deduce from (12) explicit divisors of $2^p + 1$ and thus explicit pairs $(x, y) \in \mathbb{F}^2$ for which the bounds in Theorem 3.2 are attained:

- If p is odd, then 3 divides $2^p + 1$ and we can take
$$(x, y) = \big(\tfrac{4+4u}{3}, \tfrac{3}{2}\big);$$
- If $p \equiv 0 \bmod 3$, then $2^p + 1$ can be factored as $(2^{p/3} + 1)(2^{2p/3} - 2^{p/3} + 1)$, so we can take
$$(x, y) = (2 - 2u^{1/3} + 2u^{2/3}, 1 + u^{1/3}).$$

In fact, the sufficient condition "p is not a power of two" is satisfied not only by those basic formats but also by all the binary *interchange* formats of IEEE 754-2008, for which either $p \in \{11, 24, 53, 113\}$ or

$$p = k - d + 13 \quad \text{with} \quad d = \lfloor 4 \log_2 k \rceil, \quad k = 32j, \quad j \in \mathbb{N}_{\geqslant 5}, \tag{13}$$

and where $\lfloor \cdot \rceil$ denotes rounding to a nearest integer. (A proof of the fact that (13) implies that p is not a power of two is deferred to Appendix A.)

Assume now that p is a power of two. In this case $2^p + 1$ is not prime if and only if it is not a prime number of the form $F_\ell = 2^{2^\ell} + 1$, called a Fermat prime. Currently, the only known Fermat primes are F_0, F_1, F_2, F_3, F_4 and, on the other hand, F_ℓ is known to be composite for all $5 \leqslant \ell \leqslant 32$; see for example [11, 17].

To summarize, the only values of p for which the optimality condition "$2^p + 1$ is not prime" can fail to be satisfied are $2, 4, 8, 16$ and $p = 2^\ell \geqslant 2^{33} \approx 8.6 \times 10^9$, and none of these values corresponds to an IEEE format.

3.3. **Division.** This section focuses on the largest possible relative errors committed when rounding x/y with x, y nonzero elements of $\mathbb{F}$. As the theorem below shows, the general bounds $E_1(t) \leqslant \frac{u}{1+u}$ and $E_2(t) \leqslant u$ given in (2) can be refined further for base 2, but remain optimal for all other bases. Note that unlike for addition or multiplication, optimality is achieved without any extra assumption on the parity of β or the number theoretic properties of p.

Theorem 3.3. *Let $\beta, p \geqslant 2$ and let $x, y \in \mathbb{F}$ be nonzero. Then*

$$E_1(x/y) \leqslant \begin{cases} u - 2u^2 & \text{if } \beta = 2, \\ \frac{u}{1+u} & \text{if } \beta > 2, \end{cases}$$

and

$$E_2(x/y) \leqslant \begin{cases} \frac{u-2u^2}{1+u-2u^2} & \text{if } \beta = 2, \\ u & \text{if } \beta > 2. \end{cases}$$

The bounds for $\beta = 2$ are attained at $(x, y) = (1, 1 - u)$ and, assuming ties are rounded "to even", the bounds for $\beta > 2$ are attained at $(x, y) = (2 + 2u, 2)$.

Proof. When $\beta > 2$ the bounds are the general ones given in (2), and the fact they are attained for division follows immediately from $2+2u \in \mathbb{F}$. The rest of the proof is thus devoted to the case $\beta = 2$.

Let $t = x/y$. Since t cannot be a midpoint [9, 13], we have $\mathrm{fl}(-t) = -\mathrm{fl}(t)$ and $\mathrm{fl}(t \cdot 2^e) = \mathrm{fl}(t) \cdot 2^e$, $e \in \mathbb{Z}$, regardless of the tie-breaking rule of fl. Consequently, we can assume $1 \leqslant t < 2$ and $x, y > 0$. When $t = 1$ both E_1 and E_2 are zero, so we are left with handling t such that $1 < t < 2$.

The lower bound on t implies $x > y$, which for x and y in $\mathbb{F}$ is equivalent to $x \geqslant y + 2u\,\mathrm{ufp}(y)$. Hence, using $y \leqslant (2-2u)\mathrm{ufp}(y)$,

$$t \geqslant 1 + 2u\frac{\mathrm{ufp}(y)}{y} \geqslant 1 + \frac{u}{1-u} = \frac{1}{1-u}. \tag{14}$$

Since $1/(1-u)$ is strictly larger than the midpoint $1+u$, it follows that

$$\mathrm{fl}(t) \in \{1+2u, 1+4u, \ldots\}. \tag{15}$$

■ Assume for now that $p \geqslant 3$. (For simplicity, the case $\beta = p = 2$ is handled separately at the end of the proof.)

If $\mathrm{fl}(t) \geqslant 1+4u$, then $t \geqslant 1+3u$, so that $E_1(t) \leqslant u\,\mathrm{ufp}(t)/t \leqslant u/(1+3u) < u-2u^2$ for $p \geqslant 3$. Similarly, $E_2(t) \leqslant u/(1+4u) < (u-2u^2)/(1+u-2u^2)$ for $p \geqslant 3$.

If $\mathrm{fl}(t) = 1+2u$, then, recalling (14) and the fact that t is not a midpoint,

$$\frac{1}{1-u} \leqslant t < 1+3u. \tag{16}$$

We now distinguish between the following two subcases, depending on how t compares to $\mathrm{fl}(t) = 1+2u$:

- If $t \leqslant 1+2u$, then $E_1(t) = (1+2u)/t - 1$ and $E_2(t) = 1 - t/(1+2u)$, so that the first inequality in (16) gives immediately the desired bounds, which are attained only when $t = 1/(1-u)$; since both 1 and $1-u$ are in $\mathbb{F}$ when $\beta = 2$, this value of t is obtained for $(x, y) = (1, 1-u)$.
- If $t > 1+2u$, then $E_1(t) = 1 - (1+2u)/t$ and $E_2(t) = t/(1+2u) - 1$. The bound $t < 1+3u$ from (16) then gives immediately $E_1(t) < u/(1+3u)$, which is less than $u - 2u^2$ for $p \geqslant 3$.

  For $E_2(t)$, however, using $t < 1+3u$ is not enough and we show first how to replace this bound by the slightly sharper one

$$t \leqslant \frac{2+7u}{2+u} = 1 + 3u - \tfrac{3}{2}u^2 + \mathcal{O}(u^3). \tag{17}$$

  The range of t implies $y + 2u\,\mathrm{ufp}(y) < x < y + 6u\,\mathrm{ufp}(y)$. Note that y cannot be equal to $(2-2u)\mathrm{ufp}(y)$, for otherwise $2\mathrm{ufp}(y) < x < (2+4u)\mathrm{ufp}(y)$, thus contradicting the fact that $x \in \mathbb{F}$. Therefore, $y \leqslant (2-4u)\mathrm{ufp}(y)$ and

$$x = y + 4u\,\mathrm{ufp}(y).$$

  Writing $y = (1+2ku)\mathrm{ufp}(y)$ with k a nonnegative integer, we deduce that

$$t = 1 + \frac{4u}{1+2ku},$$

  so $t < 1+3u$ is equivalent to $k > 2^{p-1}/3$, that is, $k \geqslant (2^{p-1}+1)/3$ for k is an integer. Hence $2ku \geqslant (1+2u)/3$ and (17) follows.

  Recalling that $E_2(t) = t/(1+2u) - 1$ and applying (17), we arrive at $E_2(t) \leqslant \frac{2u(1-u)}{(2+u)(1+2u)}$, which is less than $\frac{u-2u^2}{1+u-2u^2}$ for $p \geqslant 3$.

■ Assume now that $p = 2$. We have $u = 1/4$ and, assuming with no loss of generality that $\mathrm{ufp}(x) = 1$, we see that $x \in \{1, 3/2\}$ and that y is either $\mathrm{ufp}(y)$ or $3/2 \cdot \mathrm{ufp}(y)$. This yields four possibilities for $t = x/y$, all leading to $E_1(t) = E_2(t) = 0$ except for $x = 1$ and $y = 3/2 \cdot \mathrm{ufp}(y)$. In this case, the constraint $1 < t < 2$ implies further $y = 3/4 = 1 - u$. It follows that $t = 1/(1-u)$, from which we deduce $\mathrm{fl}(t) = 1 + 2u$.

The announced bounds thus hold also in the case $\beta = p = 2$, and since 1 and $1 - u$ are in $\mathbb{F}$, they are attained for $(x, y) = (1, 1 - u)$. □

3.4. **Square root.** Finally, we show how to refine further the bounds $E_1(t) \leqslant u/(1 + u)$ and $E_2(t) \leqslant u$ in the special case where $t = \sqrt{x}$ for some positive floating-point number x, thereby establishing the optimal bounds in the last row of Table 1. This result is independent of any specific property of the base and the precision, and holds for any tie-breaking strategy.

Theorem 3.4. *Let $\beta, p \geqslant 2$ and let $x \in \mathbb{F}$ be positive. Then*

$$E_1(\sqrt{x}) \leqslant 1 - \frac{1}{\sqrt{1+2u}} \qquad \text{and} \qquad E_2(\sqrt{x}) \leqslant \sqrt{1+2u} - 1,$$

and these bounds are attained only for $x = (1 + 2u)\beta^{2e}$ with $e \in \mathbb{Z}$.

Proof. Let $t = \sqrt{x}$. Writing $x = \mu\beta^{2e}$ with $e \in \mathbb{Z}$ and $\mu \in \mathbb{F} \cap [1, \beta^2)$, we see that $t = \sqrt{\mu}\,\beta^e$ with $\mu \in \{1, 1 + 2u, 1 + 4u, \ldots\}$. If $\mu = 1$, then $E_1(t) = E_2(t) = 0$, so we are left with the following two cases.

If $\mu = 1 + 2u$, then we deduce from $1 < \sqrt{1+2u} < 1 + u$ that $\mathrm{fl}(t) = \beta^e$ and, consequently, that $E_1(t) = 1 - 1/\sqrt{1+2u}$ and $E_2(t) = \sqrt{1+2u} - 1$.

If $\mu \geqslant 1 + 4u$, then, recalling that $\mu < \beta^2$, we have $\sqrt{1+4u}\,\beta^e \leqslant t < \beta^{e+1}$ and $\mathrm{ufp}(t) = \beta^e$. This implies that $E_1(t) \leqslant u\,\mathrm{ufp}(t)/t \leqslant u/\sqrt{1+4u} =: \varphi$ and it can be checked that $\varphi < 1 - 1/\sqrt{1+2u}$ for $u \leqslant 1/2$. Furthermore, using $\sqrt{1+4u} > 1 + u$ gives $\mathrm{fl}(t) \geqslant (1 + 2u)\beta^e$ and then $E_2(t) \leqslant u/(1 + 2u) < \sqrt{1+2u} - 1$. □

4. Application to summation

4.1. **Sums of floating-point numbers.** Consider first the evaluation of the sum of n floating-point numbers. Here $x_1, \ldots, x_n \in \mathbb{F}$ are given and we assume that an approximation $\widehat{s} \in \mathbb{F}$ to the exact sum

$$s = x_1 + \cdots + x_n$$

is produced after $n - 1$ floating-point additions, using any evaluation order.[2] Each of these additions takes a pair of floating-point numbers, say $(a, b) \in \mathbb{F}^2$, and returns $\mathrm{fl}(a+b)$, thus committing the local error $e := a + b - \mathrm{fl}(a+b)$. When evaluating the sum as above, $n-1$ such errors can occur and, denoting their set as $\{e_1, \ldots, e_{n-1}\}$, it is easy to see that

$$s = e_1 + \cdots + e_{n-1} + \widehat{s}; \tag{18}$$

see for example [5, p. 81].

The theorem below shows how to bound the sum of the $|e_i|$ and, therefore, $|\widehat{s} - s|$ as well. This bound is slightly sharper than the one given in [10, Proposition 3.1] and can be established in the same way, using $u/(1+u)$ instead of u to bound the relative rounding error committed by each floating-point addition. (For completeness a

[2]For example, for $n = 4$ possible evaluation orders include $((x_1 + x_2) + x_3) + x_4$ and $((x_4 + x_3) + x_2) + x_1$ and $(x_1 + x_2) + (x_3 + x_4)$.

detailed proof is presented in Appendix B; note that as in [10] this bound holds even if underflow occurs.)

Theorem 4.1. *For $x_1, \dots, x_n \in \mathbb{F}$, any order of evaluation of the sum $s = \sum_{i=1}^{n} x_i$ produces an approximation $\widehat{s}$ such that the rounding errors $e_1, \dots, e_{n-1}$ satisfy*

$$|\widehat{s} - s| \leqslant \sum_{i=1}^{n-1} |e_i| \leqslant \sigma_{n-1} \sum_{i=1}^{n} |x_i|, \qquad \sigma_n := \frac{nu}{1+u}.$$

A direct consequence of this result is a sharper and simpler bound for the following compensated summation scheme (see [14] and the references therein):

$$\widehat{s}_{\mathrm{comp}} = \mathrm{fl}(\widehat{s} + \widehat{e}), \qquad \widehat{e} := \text{a floating-point evaluation of } e_1 + \cdots + e_{n-1}.$$

Since each e_i belongs to $\mathbb{F}$ and can be computed exactly—using for example Knuth's *TwoSum* algorithm [12, p. 236], we can apply Theorem 4.1 twice and, writing $e = \sum_{i=1}^{n-1} e_i$, we deduce that

$$|\widehat{e} - e| \leqslant \sigma_{n-1}\sigma_{n-2} \sum_{i=1}^{n} |x_i|. \tag{19}$$

Since $s = \widehat{s} + e$, we have also

$$\begin{aligned} |\widehat{s}_{\mathrm{comp}} - s| &\leqslant |\mathrm{fl}(\widehat{s} + \widehat{e}) - (\widehat{s} + \widehat{e})| + |\widehat{e} - e| \\ &\leqslant \frac{u}{1+u}|\widehat{s} + \widehat{e}| + |\widehat{e} - e| \\ &\leqslant \frac{u}{1+u}|s| + \Big(1 + \frac{u}{1+u}\Big)|\widehat{e} - e|. \end{aligned} \tag{20}$$

Then, combining (19) and (20) and using the fact that $(1 + \frac{u}{1+u})(\frac{u}{1+u})^2 \leqslant \frac{u^2}{1+u^2}$ we arrive at

$$|\widehat{s}_{\mathrm{comp}} - s| \leqslant \frac{u}{1+u}|s| + (n-1)(n-2)\frac{u^2}{1+u^2} \sum_{i=1}^{n} |x_i|. \tag{21}$$

The bound in (21) holds without any restriction on n and regardless of the orderings used for adding the x_i and adding the e_i. Furthermore, it is slightly sharper than the bound $u|s| + (n-1)^2 \frac{u^2}{(1-(n-1)u)^2} \sum_{i=1}^{n} |x_i|$ given in [14, Proposition 4.5] in the special case of recursive compensated summation.

4.2. **Sums of real numbers.** We now turn to the case where $x_1, \dots, x_n$ are in $\mathbb{R}$ instead of $\mathbb{F}$. An approximation $\widehat{s} \in \mathbb{F}$ to $s = x_1 + \cdots + x_n \in \mathbb{R}$ is obtained in two steps, by first rounding each x_i into $\mathrm{fl}(x_i)$ and then evaluating $\mathrm{fl}(x_1) + \cdots + \mathrm{fl}(x_n)$ as above. A typical example is the computation of inner products, where each x_i is the exact product of two given floating-point numbers.

Writing $d_i = x_i - \mathrm{fl}(x_i)$ for the errors due to rounding the data and, as before, e_i for the errors due to the $n-1$ additions, we deduce that the exact and computed sums are now related as

$$s = d_1 + \cdots + d_n + e_1 + \cdots + e_{n-1} + \widehat{s}. \tag{22}$$

By combining (2) and Theorem 4.1 we obtain almost immediately the following bound on the sum of the $|d_i|$ and the $|e_i|$.

Theorem 4.2. *For $x_1, \dots, x_n \in \mathbb{R}$, any order of evaluation of the sum $\sum_{i=1}^n \mathrm{fl}(x_i)$ produces an approximation $\widehat{s}$ to the sum $s = \sum_{i=1}^n x_i$ such that the rounding errors $d_1, \dots, d_n$ and $e_1, \dots, e_{n-1}$ satisfy*

$$|\widehat{s} - s| \leqslant \sum_{i=1}^{n} |d_i| + \sum_{i=1}^{n-1} |e_i| \leqslant \zeta_n \sum_{i=1}^{n} |x_i|, \qquad \zeta_n := \frac{(1+2u)nu - u^2}{(1+u)^2} < nu.$$

Proof. The lower bound follows from (22). To establish the upper bound, let $v = u/(1+u)$. Theorem 4.1 gives $\sum_{i<n} |e_i| \leqslant (n-1)v \sum_{i \leqslant n} |\mathrm{fl}(x_i)|$ and, on the other hand, (2) implies that $|d_i| \leqslant v|x_i|$ and $|\mathrm{fl}(x_i)| \leqslant (1+v)|x_i|$. Hence $\sum_{i \leqslant n} |d_i| + \sum_{i<n} |e_i| \leqslant (v + (n-1)v(1+v)) \sum_{i \leqslant n} |x_i|$, and it is easily checked that $v + (n-1)v(1+v)$ simplifies to $((1+2u)nu - u^2)/(1+u)^2$, which is less than nu. □

Thus, the theorem above gives in particular the bound

$$\Big|\widehat{s} - \sum_{i=1}^{n} x_i\Big| \leqslant \zeta_n \sum_{i=1}^{n} |x_i|.$$

This bound has a slightly smaller constant than the one given in [10, Proposition 4.1] and, perhaps more importantly, its proof is significantly shorter and avoids a tedious induction and ufp-based case distinctions.

Furthermore, the applications mentioned in [10] obviously benefit directly from this new bound: for inner products, matrix-vector products, and matrix-matrix products, the constants nu obtained in [10, Theorem 4.2 and p. 343] can all be replaced by ζ_n.

5. Other application examples

5.1. **Example 1: Small arithmetic expressions.** Rounding error analyses as those done in [5] typically involve bounds on $|\theta_n|$, where θ_n is an expression of the form $\theta_n = \prod_{i=1}^n (1+\delta_i) - 1$ with δ_i a relative error term associated with a single floating-point operation. Using the classical bound $|\delta_i| \leqslant u$ it is easily checked that for all n,

$$-nu \leqslant \theta_n \leqslant (1+u)^n - 1. \tag{23}$$

Note that only the upper bound has the form $nu + \mathcal{O}(u^2)$, that is, contains terms nonlinear in u. From (23) it follows that

$$|\theta_n| \leqslant (1+u)^n - 1$$

and, assuming $nu < 1$, this bound itself is usually bounded above by the classical fraction $\gamma_n = nu/(1-nu)$.

Using the refined bound $E_1(t) \leqslant u/(1+u)$ from (2), we can replace $|\delta_i| \leqslant u$ by $|\delta_i| \leqslant u/(1+u)$, which immediately leads to the refined enclosure

$$-\frac{nu}{1+u} \leqslant \theta_n \leqslant \Big(1 + \frac{u}{1+u}\Big)^n - 1. \tag{24}$$

Although the upper bound in (24) still has $\mathcal{O}(u^2)$ terms in general, it is bounded by nu as long as $n \leqslant 3$. Consequently, by just systematically using $E_1(t) \leqslant u/(1+u)$ instead of $E_1(t) \leqslant u$, we can replace $\gamma_n = nu + \mathcal{O}(u^2)$ by nu in every error analysis where θ_n appears with $n \leqslant 3$.

For example, when evaluating $ab + cd$ in the usual way as $\widehat{r} = \mathrm{fl}(\mathrm{fl}(ab) + \mathrm{fl}(cd))$, we have $\widehat{r} = (ab(1+\delta_1) + cd(1+\delta_2))(1+\delta_3)$ with $|\delta_i| \leqslant u/(1+u)$. Hence

$$\widehat{r} = ab(1+\theta_2) + cd(1+\theta_2'), \qquad |\theta_2|, |\theta_2'| \leqslant 2u,$$

so the usual γ_2 is indeed replaced by $2u$. Note that once we have replaced $E_1(t) \leqslant u$ by $E_1(t) \leqslant u/(1+u)$, we obtain that term $2u$ immediately, without having to resort to a sophisticated ufp-based argument as the one introduced in [1, pp. 1470–1471].

Similar examples include the evaluation of small arithmetic expressions like the product $x_1x_2x_3x_4$ (using any parenthesization) or the sums $((x_1+x_2)+x_3)+x_4$ and $((x_1+x_2)+(x_3+x_4))+((x_5+x_6)+(x_7+x_8))$ (using these specific parenthesizations); in each case the forward error bound classically involves γ_3, which we now replace by $3u$.

5.2. **Example 2: Leading principal minors of a tridiagonal matrix.** Consider the tridiagonal matrix

$$A = \begin{bmatrix} d_1 & e_1 & & & \\ c_2 & d_2 & e_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & e_{n-1} \\ & & & c_n & d_n \end{bmatrix} \in \mathbb{F}^{n\times n},$$

and let $\mu_1, \mu_2, \ldots, \mu_n$ be the sequence of its n leading principal minors. Writing $\mu_{-1} = 0$ and $\mu_0 = 1$, those minors are thus defined by the linear recurrence

$$\mu_k = d_k\mu_{k-1} - c_ke_{k-1}\mu_{k-2}, \qquad 1 \leqslant k \leqslant n.$$

Using the usual bound $E_1(t) \leqslant u$ and barring underflow and overflow, Wilkinson shows in [19, §3] that the evaluation of this recurrence produces floating-point numbers $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$ such that

$$\widehat{\mu}_k = d_k(1+\epsilon_k)\widehat{\mu}_{k-1} - c_k(1+\epsilon_k')e_{k-1}(1+\epsilon_k'')\widehat{\mu}_{k-2},$$

where $(1-u)^2 - 1 \leqslant \epsilon_k \leqslant (1+u)^2 - 1$ and $(1-u)^{3/2} - 1 \leqslant \epsilon_k', \epsilon_k'' \leqslant (1+u)^{3/2} - 1$. In other words, the computed $\widehat{\mu}_k$ are the leading principal minors of a nearby tridiagonal matrix $A + \Delta A = [a_{ij}(1+\delta_{ij})]$ that satisfies

$$-2u < \delta_{ii} \leqslant 2u + u^2 \quad \text{and} \quad -\tfrac{3}{2}u < \delta_{ij} \leqslant \tfrac{3}{2}u + \mathcal{O}(u^2) \quad \text{if } i \neq j. \tag{25}$$

Notice that the terms u^2 and $\mathcal{O}(u^2)$ come exclusively from the upper bounds on ϵ_k, ϵ_k', ϵ_k''. By using the refined bound $E_1(t) \leqslant u/(1+u)$ from (2) instead of just $E_1(t) \leqslant u$, these upper bounds are straightforwardly improved to

$$\epsilon_k \leqslant (1 + \tfrac{u}{1+u})^2 - 1 < 2u \qquad \text{and} \qquad \epsilon_k', \epsilon_k'' \leqslant (1 + \tfrac{u}{1+u})^{3/2} - 1 < \tfrac{3}{2}u.$$

Consequently, Wilkinson's bounds in (25) can be replaced by the following more concise and slightly sharper ones:

$$|\delta_{ii}| < 2u \quad \text{and} \quad |\delta_{ij}| < \tfrac{3}{2}u \quad \text{if } i \neq j.$$

5.3. **Equation 3: Euclidean norm of an n-dimensional vector.** Given a vector $[x_1, \ldots, x_n]^T \in \mathbb{F}^n$, let its norm

$$r = \sqrt{x_1^2 + \cdots + x_n^2}$$

be evaluated in floating point in the usual way: form the squares $\mathrm{fl}(x_i^2)$, sum them up in any order into $\widehat{s}$, and return $\widehat{r} = \mathrm{fl}\big(\sqrt{\widehat{s}}\big)$.

By applying the usual bound $E_1(t) \leqslant u$, all we can say is $\widehat{s} = (\sum_{i=1}^n x_i^2)(1+\theta_n)$ with θ_n as in (23), and $\widehat{r} = \sqrt{\widehat{s}}\,(1+\delta)$ with $|\delta| \leqslant u$. Consequently,

$$\widehat{r} = r(1+\epsilon),$$

where $\epsilon = \sqrt{1+\theta_n} \cdot (1+\delta) - 1$ satisfies $(1-u)^{n/2+1} - 1 \leqslant \epsilon \leqslant (1+u)^{n/2+1} - 1$. Although the lower bound has absolute value at most $(n/2+1)u$ (see for example [4, p. 42]), the upper bound is strictly larger than this, so that

$$-(n/2+1)u \leqslant \epsilon \leqslant (n/2+1)u + \mathcal{O}(u^2). \tag{26}$$

To avoid the $\mathcal{O}(u^2)$ term above, we can use the refined bound $E_1(t) \leqslant u/(1+u)$, which says $|\delta| \leqslant u/(1+u)$, together with the improved bound for inner products from [10], which says $|\theta_n| \leqslant nu$. Indeed, from these two bounds we deduce that ϵ is upper bounded by $\sqrt{1+nu} \cdot (1 + u/(1+u)) - 1$, and the latter quantity is easily checked to be at most $(n/2+1)u$. Thus, recalling the lower bound in (26), we conclude that

$$|\epsilon| \leqslant (n/2+1)u. \tag{27}$$

In particular, evaluating the hypotenuse $\sqrt{x_1^2 + x_2^2}$ in floating-point produces a relative error of at most $2u$; this improves over the classical bound $2u + \mathcal{O}(u^2)$, stated for example in [7, p. 225].

Of course, the bound in (27) also applies when scaling by integer powers of the base is introduced to avoid underflow and overflow.

5.4. **Example 4: Cholesky factorization.** We consider $A \in \mathbb{F}^{n\times n}$ symmetric and its triangularization in floating-point arithmetic using the classical Cholesky algorithm. If the algorithm runs to completion, then by using the bounds $E_i(t) \leqslant u$, $i = 1, 2$, the traditional rounding error analysis concludes that the computed factor $\widehat{R}$ satisfies $\widehat{R}^T\widehat{R} = A + \Delta A$ with

$$|\Delta A| \leqslant \gamma_{n+1} |\widehat{R}^T||\widehat{R}|;$$

see for example [5, Theorem 10.3]. Here $\gamma_{n+1} = \frac{(n+1)u}{1-(n+1)u}$ has the form $(n+1)u + \mathcal{O}(u^2)$ and requires $n+1 < u^{-1}$. It was shown in [15] that both the quadratic term in u and the restriction on n can be removed, resulting in the improved backward error bound

$$|\Delta A| \leqslant (n+1)u |\widehat{R}^T||\widehat{R}|.$$

In the proof of [15, Theorem 4.4], one of the ingredients used to suppress the $\mathcal{O}(u^2)$ term is the following property:

$$\Big(a \in \mathbb{F}_{\geqslant 0} \quad \text{and} \quad b = \mathrm{fl}\big(\sqrt{a}\,\big)\Big) \quad \Rightarrow \quad |b^2 - a| \leqslant 2ub^2. \tag{28}$$

In [15] it is shown that this property may not hold if only the bound $E_2(t) \leqslant u$ is assumed, and that in this case all we can say is $-(2u+u^2)b^2 \leqslant b^2 - a \leqslant 2ub^2$. Furthermore, a proof of (28) is given, which is about 10 lines long and based on a ufp-based case analysis; see [15, p. 692].

Instead, our optimal bound on E_2 for square root provides a direct proof: Theorem 3.4 gives $b(1+\delta)=\sqrt{a}$ with $|\delta| \leqslant \sqrt{1+2u}-1$; hence $|b^2-a| = b^2|2\delta+\delta^2|$ and $|2\delta+\delta^2| \leqslant (2+|\delta|)|\delta| \leqslant (\sqrt{1+2u}+1)(\sqrt{1+2u}-1) = 2u$, from which (28) follows immediately.

5.5. **Example 5: Complex multiplication with an FMA.** Given $a, b, c, d \in \mathbb{F}$, consider the complex product

$$z = (a+ib)(c+id).$$

Various approximations $\widehat{z} = \widehat{R} + i\,\widehat{I}$ to z can be obtained, depending on how $R = ac-bd$ and $I = ad+bc$ are evaluated in floating-point. It was shown in [1] that the conventional way, which uses 4 multiplications and 2 additions, gives $\widehat{z} = z(1+\epsilon)$ with $\epsilon \in \mathbb{C}$ such that $|\epsilon| < \sqrt{5}u$, and that the constant $\sqrt{5}$ is, at least in base 2, best possible. Assume now that an FMA is available, so that we compute, say,

$$\widehat{R} = \mathrm{fl}(ac - \mathrm{fl}(bd)) \quad \text{and} \quad \widehat{I} = \mathrm{fl}(ad + \mathrm{fl}(bc)).$$

For this algorithm and its variants[3] it was shown in [8] that the bound $\sqrt{5}u$ can be reduced further to $2u$, and that the latter is essentially optimal. The fact that $2u$ is an *upper* bound is established in [8, Theorem 3.1] with a rather long proof. As we shall see in the paragraph below, a much more direct proof follows from simply applying the refined bound $E_1(t) \leqslant u/(1+u)$ in a systematic way.

Denoting by $\delta_1, \ldots, \delta_4$ the four rounding errors involved, we have

$$\begin{aligned}\widehat{R} &= (ac - bd(1+\delta_1))(1+\delta_2)\\ &= R + R\delta_2 - bd\delta_1(1+\delta_2)\end{aligned}$$

and, similarly, $\widehat{I} = I + I\delta_4 + bc\delta_3(1+\delta_4)$. Now let $\lambda, \mu \in \mathbb{R}_{\geqslant 0}$ be such that

$$|\delta_2|, |\delta_4| \leqslant \lambda \qquad \text{and} \qquad |\delta_1(1+\delta_2)|, |\delta_3(1+\delta_4)| \leqslant \mu.$$

This implies that $|R-\widehat{R}| \leqslant \lambda|R| + \mu|bd|$ and $|I-\widehat{I}| \leqslant \lambda|I| + \mu|bc|$, from which we deduce

$$\begin{aligned}|z-\widehat{z}|^2 &= (R-\widehat{R})^2 + (I-\widehat{I})^2\\ &\leqslant \lambda^2|z|^2 + 2\lambda\mu A + \mu^2 B,\end{aligned} \tag{29}$$

where $A = |R||bd| + |I||bc|$ and $B = (bd)^2 + (bc)^2$. It turns out that

$$A, B \leqslant |z|^2. \tag{30}$$

For B, this bound simply follows from the equality $|z|^2 = (ac)^2+(bd)^2+(ad)^2+(bc)^2$. For A, define $\pi = abcd$ and notice that $A = |\pi - (bd)^2| + |\pi + (bc)^2|$ is equal to either B or $\pm(2\pi + (bc)^2 - (bd)^2)$; furthermore, in the latter case we have

$$\begin{aligned}A &\leqslant 2|\pi| + \left|(bc)^2 - (bd)^2\right|\\ &\leqslant \min\left\{(ac)^2 + (bd)^2, (ad)^2 + (bc)^2\right\} + \max\left\{(bc)^2, (bd)^2\right\}\\ &\leqslant |z|^2.\end{aligned}$$

Thus, combining (29) and (30), $|z-\widehat{z}| \leqslant (\lambda+\mu)|z|$. Since the refined bound $E_1(t) \leqslant u/(1+u)$ implies $|\delta_i| \leqslant u/(1+u)$ for all i, we can take $\lambda = u/(1+u)$

[3]There are three other ways to insert the innermost rounding fl, all giving the same error as the one developed here.

and $\mu = u/(1+u)\cdot(1+u/(1+u))$, which are both less than u. Hence, barring underflow and overflow and since $z=0$ implies $\widehat{z}=0$, we conclude that

$$\widehat{z} = z(1+\epsilon), \qquad |\epsilon| \leqslant \tfrac{2u+3u^2}{(1+u)^2} < 2u.$$

Note that $\frac{2u+3u^2}{(1+u)^2}$ has the form $2u - u^2 + \mathcal{O}(u^4)$ as $u \to 0$. Thus, our approach not only yields a shorter and more direct proof of the bound $2u$ of [8], but it also improves on that bound.

Appendix A. Proof that p as in (13) is not a power of two

If $j \in \{5,6,7\}$, then $p \in \{144, 175, 206\}$ and is not a power of two. Assume now that $j \geqslant 8$. Writing $d = 4m+i$ for integers m, i with $0 \leqslant i \leqslant 3$, we have

$$4m + i - \frac{1}{2} \leqslant 4\log_2 k \leqslant 4m + i + \frac{1}{2}.$$

If $i \neq 0$, this implies $2^{m+1/8} \leqslant k \leqslant 2^{m+7/8}$ and then

$$2^{m+1/8} - 4m + 10 \leqslant p \leqslant 2^{m+7/8} - 4m + 12.$$

Since the assumption $j \geqslant 8$ implies $k \geqslant 2^8$ and thus $m \geqslant 8$, it follows that $2^m < p < 2^{m+1}$. Consequently, p cannot be a power of two when $i \neq 0$. On the other hand, when $i = 0$, we see that $p = 32j - 4m + 13$ must be odd, and thus cannot be a power of two.

Appendix B. Proof of Theorem 4.1

The lower bound follows from (18) and the triangle inequality. For the upper bound, the proof is by induction on n, the case $n=1$ being trivial (since then there is no rounding error at all). For $n \geqslant 2$, we assume that the result is true up to $n-1$, and we fix one evaluation order in dimension n. The approximation $\widehat{s}$ obtained with this order has the form $\widehat{s} = \mathrm{fl}(\widehat{s}_1 + \widehat{s}_2)$, where $\widehat{s}_j$ is the result of a floating-point evaluation of $s_j = \sum_{i\in I_j} x_i$ for $j = 1,2$ and with $\{I_1, I_2\}$ a partition of the set $I = \{1, 2, \ldots, n\}$. For $j = 1, 2$ let n_j be the cardinality of I_j and let $e_j^{(1)}, \ldots, e_j^{(n_j-1)}$ be the rounding errors committed when evaluating s_j, so that $s_j = \sum_{i<n_j} e_j^{(i)} + \widehat{s}_j$. Consequently,

$$\sum_{i<n} e_i = \delta + \sum_{i<n_1} e_1^{(i)} + \sum_{i<n_2} e_2^{(i)}, \qquad \delta = \widehat{s}_1 + \widehat{s}_2 - \mathrm{fl}(\widehat{s}_1 + \widehat{s}_2).$$

Since $1 \leqslant n_j < n$, the inductive assumption leads to

$$\sum_{i<n} |e_i| \leqslant |\delta| + \frac{u}{1+u}\Big((n_1-1)\widetilde{s}_1 + (n_2-1)\widetilde{s}_2\Big),$$

where $\widetilde{s}_j = \sum_{i\in I_j} |x_i|$ for $j = 1, 2$. Since $n = n_1 + n_2$ and $\sum_{i=1}^n |x_i| = \widetilde{s}_1 + \widetilde{s}_2$, it remains to check that

$$|\delta| \leqslant \frac{u}{1+u}(n_2\widetilde{s}_1 + n_1\widetilde{s}_2). \tag{31}$$

To do so, we note that (2) and [10, Lemma 2.2] imply that

$$|\delta| \leqslant \min\Big\{|\widehat{s}_1|, |\widehat{s}_2|, \frac{u}{1+u}|\widehat{s}_1 + \widehat{s}_2|\Big\},$$

and we consider the following three cases. Assume first that $\widetilde{s}_2 \leqslant u/(1+u)\cdot\widetilde{s}_1$. Then $\widetilde{s}_2 \leqslant \widetilde{s}_1$ and, using $|\delta| \leqslant |\widehat{s}_2|$, we obtain

$$|\delta| \leqslant |\widehat{s}_2 - s_2| + \widetilde{s}_2 \leqslant \sum_{i<n_2} |e_2^{(i)}| + \widetilde{s}_2 \leqslant (n_2-1)\frac{u}{1+u}\widetilde{s}_2 + \frac{u}{1+u}\widetilde{s}_1 \leqslant n_2\frac{u}{1+u}\widetilde{s}_1$$

and (31) thus follows. Second, when $\widetilde{s}_1 \leqslant u/(1+u)\cdot\widetilde{s}_2$ we proceed similarly, simply swapping the indices 1 and 2. Third, when $u/(1+u)\cdot\widetilde{s}_1 < \widetilde{s}_2$ and $u/(1+u)\cdot\widetilde{s}_2 < \widetilde{s}_1$, we have $|\delta| \leqslant u/(1+u)\cdot|\widehat{s}_1 + \widehat{s}_2|$ with

$$|\widehat{s}_1 + \widehat{s}_2| \leqslant |\widehat{s}_1 - s_1| + \widetilde{s}_1 + |\widehat{s}_2 - s_2| + \widetilde{s}_2$$

and $|\widehat{s}_j - s_j| \leqslant (n_j-1)u/(1+u)\cdot\widetilde{s}_j \leqslant (n_j-1)\widetilde{s}_k$ for $(j,k) \in \{(1,2),(2,1)\}$. Hence (31) follows in this third case as well, thus completing the proof of Theorem 4.1.

References

[1] R. Brent, C. Percival, and P. Zimmermann, *Error bounds on complex floating-point multiplication*, Math. Comp. **76** (2007), no. 259, 1469–1481 (electronic), DOI 10.1090/S0025-5718-07-01931-X. MR2299783

[2] T. J. Dekker, *A floating-point technique for extending the available precision*, Numer. Math. **18** (1971/72), 224–242. MR0299007

[3] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*: *A Foundation for Computer Science*, 2nd ed., Addison-Wesley Publishing Company, Reading, MA, 1994. MR1397498

[4] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge, at the University Press, 1952. 2nd ed. MR0046395

[5] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. MR1927606

[6] J. E. Holm, *Floating-Point Arithmetic and Program Correctness Proofs*, Ph.D. thesis, Cornell University, Ithaca, NY, August 1980, pp. vii+133.

[7] T. E. Hull, T. F. Fairgrieve, and P. T. P. Tang, *Implementing complex elementary functions using exception handling*, ACM Trans. Math. Software **20** (1994), no. 2, 215–244.

[8] C.-P. Jeannerod, P. Kornerup, N. Louvet, and J.-M. Muller, *Error bounds on complex floating-point multiplication with an FMA*, Math. Comp. **86** (2017), no. 304, 881–898, DOI 10.1090/mcom/3123. MR3584553

[9] C.-P. Jeannerod, N. Louvet, J.-M. Muller, and A. Panhaleux, *Midpoints and exact points of some algebraic functions in floating-point arithmetic*, IEEE Trans. Comput. **60** (2011), no. 2, 228–241, DOI 10.1109/TC.2010.144. MR2815427

[10] C.-P. Jeannerod and S. M. Rump, *Improved error bounds for inner products in floating-point arithmetic*, SIAM J. Matrix Anal. Appl. **34** (2013), no. 2, 338–344, DOI 10.1137/120894488. MR3038111

[11] W. Keller, *Prime factors $k \cdot 2^n + 1$ of Fermat numbers F_m and complete factoring status*, August 2016, web page available at http://www.prothsearch.net/fermat.html.

[12] D. E. Knuth, *The Art of Computer Programming. Vol. 2*: *Seminumerical Algorithms; Third edition [of MR0286318]*, Addison-Wesley, Reading, MA, 1998. MR3077153

[13] P. W. Markstein, *Computation of elementary functions on the IBM RISC System/6000 processor*, IBM J. Res. Develop. **34** (1990), no. 1, 111–119, DOI 10.1147/rd.341.0111. MR1057659

[14] T. Ogita, S. M. Rump, and S. Oishi, *Accurate sum and dot product*, SIAM J. Sci. Comput. **26** (2005), no. 6, 1955–1988, DOI 10.1137/030601818. MR2196584

[15] S. M. Rump and C.-P. Jeannerod, *Improved backward error bounds for LU and Cholesky factorizations*, SIAM J. Matrix Anal. Appl. **35** (2014), no. 2, 684–698, DOI 10.1137/130927231. MR3211033

[16] S. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation. I. Faithful rounding*, SIAM J. Sci. Comput. **31** (2008), no. 1, 189–224, DOI 10.1137/050645671. MR2460776

[17] N. J. A. Sloane and D. W. Wilson, *Sequence A019434*, The On-Line Encyclopedia of Integer Sequences, published electronically at http://oeis.org.

[18] P. H. Sterbenz, *Floating-Point Computation*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974. Prentice-Hall Series in Automatic Computation. MR0349062

[19] J. H. Wilkinson, *Error analysis of floating-point computation*, Numer. Math. **2** (1960), 319–340. MR0116477

Inria and Université de Lyon, laboratoire LIP (CNRS, ENS de Lyon, Inria, UCBL), 46 allée d'Italie 69364 Lyon cedex 07, France

E-mail address: claude-pierre.jeannerod@inria.fr

Hamburg University of Technology, Schwarzenbergstrasse 95, Hamburg 21071, Germany — and — Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

E-mail address: rump@tuhh.de