# On relative errors of floating-point operations: optimal bounds and applications

Claude-Pierre Jeannerod, Siegfried M. Rump

# 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 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 provide explicit and attainable bounds on $E_1(t)$, which are all less than or equal to $u/(1+u)$ and, therefore, smaller than $u$. For $E_2(t)$ the bound $u$ is attainable whenever $t = x \pm y$ or $t = xy$ or, in base $\beta > 2$, $t = x/y$ with $x, y$ two floating-point numbers. However, for division in base 2 as well as for square root, smaller bounds are derived, which are also shown to be attainable. This set of sharp bounds is then applied to the rounding error analysis of various numerical algorithms: in all cases, we obtain either much shorter proofs of the best-known error bounds for such algorithms, or improvements on these bounds themselves.

**Key words.** floating-point arithmetic, rounding to nearest, relative error, unit in the first place

**AMS subject classifications.** 65G50

## 1. Introduction.

Let $\mathbb{F}$ be a standard set of finite floating-point numbers defined by a radix $\beta$, a precision $p$, and two extremal exponents $e_{\min}$ and $e_{\max}$. Let also $\mathrm{fl} : \mathbb{R} \to \mathbb{F} \cup \{\pm\infty\}$ denote any round-to-nearest function, such that for all $t \in \mathbb{R}$ and if no overflow occurs,

$$|t - \mathrm{fl}(t)| = \min_{f \in \mathbb{F}} |t - f|. \tag{1.1}$$

In particular, no specific tie-breaking strategy is assumed for the function fl.

Our first goal is to provide optimal bounds on the relative errors produced when applying this rounding function, 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. Here and hereafter, *optimal* means that each of our bounds is attained for at least one value of $t$, which we shall give explicitly.

Our second goal is to illustrate the interest of such bounds for the rounding error analysis of numerical algorithms: we provide several application examples for which these bounds yield analyses that are shorter or sharper.

Classically, two *relative errors* can be defined, depending on whether the exact value or the rounded value is used to divide the absolute error in (1.1): 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

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

*INRIA, Laboratoire LIP (CNRS, ENS de Lyon, INRIA, UCBL), Université de Lyon, 46 allée d'Italie 69364 Lyon cedex 07, France (claude-pierre.jeannerod@ens-lyon.fr).

†Institute for Reliable Computing, 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 (rump@tu-harburg.de).

In each case the relative error is undefined if the denominator is zero.

Our bounds on $E_1(t)$ and $E_2(t)$ are summarized in Table 1 below, where $x$ and $y$ denote elements of $\mathbb{F}$, and where $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff associated with fl and $\mathbb{F}$. All these bounds are attained for specific input values, which we shall describe later on in the paper.

TABLE 1.1
*Optimal relative error bounds for various kinds of input $t$.*

| $t$ | bound on $E_1(t)$ | bound on $E_2(t)$ |
|---|---|---|
| real number | $\frac{u}{1+u}$ | $u$ |
| $x \pm y$ | $\frac{u}{1+u}$ | $u$ |
| $xy$ | $\frac{u}{1+u}$ | $u$ |
| $x/y$ | $\begin{cases} u-2u^2 & \text{for } \beta = 2 \text{ and } p \geqslant 3, \\ \frac{u}{1+u} & \text{for even } \beta \neq 2 \end{cases}$ | $\begin{cases} \frac{u-2u^2}{1+u-2u^2} & \text{for } \beta = 2 \text{ and } p \geqslant 3, \\ u & \text{for even } \beta \neq 2 \end{cases}$ |
| $\sqrt{x}$ | $1 - \frac{1}{\sqrt{1+2u}}$ | $\sqrt{1+2u} - 1$ |

It is easily verified that the bounds appearing Table 1 satisfy the ordering

$$\frac{u-2u^2}{1+u-2u^2} \leqslant u - 2u^2 < 1 - \frac{1}{\sqrt{1+2u}} < \frac{u}{1+u} < \sqrt{1+2u} - 1 < u.$$

The only differences between all these expressions are in the constants of their $\mathcal{O}(u^2)$ terms, as can be seen from the following expansions at $u = 0$:

$$\frac{u-2u^2}{1+u-2u^2} = u - 3u^2 + \mathcal{O}(u^3),$$

$$1 - \frac{1}{\sqrt{1+2u}} = u - \tfrac{3}{2}u^2 + \mathcal{O}(u^3),$$

$$\frac{u}{1+u} = u - u^2 + \mathcal{O}(u^3),$$

$$\sqrt{1+2u} - 1 = u - \tfrac{1}{2}u^2 + \mathcal{O}(u^3).$$

In particular, the first row in Table 1 shows that an optimal bound for $E_1(t)$ is generally $u/(1+u)$, thus refining the so-called first standard model of floating-point arithmetic. In contrast, it turns out that the second model, giving $E_2(t) \leqslant u$, cannot be improved without further assumptions on $t$. These results about the standard models will be explored in Section 2. In Section 3 we establish the other bounds in Table 1 and provide in each case a proof of optimality. Finally, Section 4 illustrates with some examples the two main benefits of such refined bounds: in some cases, more concise and slightly sharper bounds can be obtained and, in other ones, shorter proofs of existing bounds can be given.

**Notation and assumptions.** Throughout this paper, $t \in \mathbb{R}$ and $x, y \in \mathbb{F}$, and we assume $t$ belongs to the normal range of $\mathbb{F}$, that is,

$$\beta^{e_{\min}} \leqslant |t| \leqslant (\beta - \beta^{1-p})\beta^{e_{\max}}.$$

Also, all our results hold under the customary assumptions that $\beta$ is even and $p \geqslant 2$, and that underflow and overflow do not occur when applying the rounding function fl. Finally, the common tool used to establish all the error bounds in Table 1 is the function $\mathrm{ufp} : \mathbb{R} \to \mathbb{Z}$ from [10], called *unit in the first place* and defined as follows: $\mathrm{ufp}(0) = 0$ and, if $r \in \mathbb{R}\backslash\{0\}$, $\mathrm{ufp}(r)$ is the largest integer power of $\beta$ such that $\mathrm{ufp}(r) \leqslant |r|$. Hence, in particular,

$$\mathrm{ufp}(t) \leqslant |t| < \beta\mathrm{ufp}(t) \qquad \text{and} \qquad \mathrm{ufp}(t) \leqslant |\mathrm{fl}(t)| \leqslant \beta\mathrm{ufp}(t). \tag{1.2}$$

## 2. Preliminaries: floating-point arithmetic models.

**2.1. The two standard models.** In numerical analysis, the most common way of modeling IEEE floating-point arithmetic is to bound both relative errors $E_1$ and $E_2$ by the unit roundoff $u$. This is usually expressed by means of the following two relations, called the *standard models* of floating-point arithmetic [4, p. 40]:

$$\begin{aligned} \mathrm{fl}(t) &= t \cdot (1+\delta_1), \qquad |\delta_1| \leqslant u \qquad \text{(first standard model)}, \\ &= t/(1+\delta_2), \qquad |\delta_2| \leqslant u \qquad \text{(second standard model)}. \end{aligned}$$

Using the ufp-function, these two models are easily derived as follows. First, recall from (1.2) that $|t|$ belongs to the 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

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

Then, dividing by either $|t|$ or $|\mathrm{fl}(t)|$ gives immediately

$$E_1(t) \leqslant u\frac{\mathrm{ufp}(t)}{|t|} \qquad \text{and} \qquad E_2(t) \leqslant u\frac{\mathrm{ufp}(t)}{|\mathrm{fl}(t)|}. \tag{2.1}$$

Finally, the lower bounds in (1.2) imply that the two ratios in (2.1) are at most 1. Hence $E_1(t) \leqslant u$ and $E_2(t) \leqslant u$, and the two standard models follow.

From (2.1) we can also deduce the classical phenomenon of *wobbling precision* [4, p. 39]: when $|t|$ and $|\mathrm{fl}(t)|$ come close to $\beta\mathrm{ufp}(t)$, then the relative errors $E_1(t)$ and $E_2(t)$ can be almost as small as $u/\beta$.

The bound $u$ given by the *second* standard model is best possible. Indeed, if the function fl is such that ties are rounded 'to even', then this bound is attained for $t = 1 + u$, since in this case $\mathrm{fl}(1+u) = 1$. If ties are rounded 'to away', then it is easily checked that the strict inequality $|\delta_2| < u$ holds; but the value $u$ is best possible in the sense that for every $\epsilon \in \mathbb{R}$ such that $0 < \epsilon \leqslant u$, setting $t = 1 + u - \epsilon$ implies $\mathrm{fl}(t) = 1$ and $|\delta_2| = u - \epsilon$.

In contrast, the bound given by the *first* standard model is never attained: no matter what the tie-breaking strategy, we have in fact the strict inequality

$$|\delta_1| < u.$$

This was already remarked in [4, p. 38] and, to see this, it suffices to note that $|t| \neq \mathrm{ufp}(t)$ implies that the first inequality in (2.1) becomes strict, and that otherwise $\mathrm{fl}(t) = t$ and thus $\delta_1 = 0$. An attainable bound for $|\delta_1|$ is described in the next paragraph, which refines [4, Theorem 2.2].

**2.2. A refinement of the first standard model.** The following result refines the first standard model by giving a bound that can be attained no matter what the tie-breaking strategy. This bound has most probably been known since a long time, as it already appears in Dekker's 1971 paper [3, pp. 233-234] in the special case of floating-point addition and multiplication in radix 2. But a general version and its optimality feature do not seem to have been reported elsewhere.

THEOREM 2.1. *If $t$ is a real number in the normal range of $\mathbb{F}$ then*

$$\mathrm{fl}(t) = t \cdot (1+\delta_1), \qquad |\delta_1| \leqslant \frac{u}{1+u}.$$

*Furthermore, this bound is attained if and only if $0 \neq |t| = (1+u)\mathrm{ufp}(t)$.*

*Proof.* If $|t| = (1+u)\mathrm{ufp}(t)$ then $\mathrm{fl}(|t|)$ is either $\mathrm{ufp}(t)$ or $(1+2u)\mathrm{ufp}(t)$. Recalling (for example from [8, Lemma 2.1]) that

$$|t - \mathrm{fl}(t)| = \big||t| - \mathrm{fl}(|t|)\big|,$$

we deduce that in both cases $|t - \mathrm{fl}(t)| = u\,\mathrm{ufp}(t)$, from which the equality $E_1(t) = u/(1+u)$ follows.

If $|t| < (1+u)\mathrm{ufp}(t)$, then $|\mathrm{fl}(t)| = \mathrm{ufp}(t) \leqslant |t|$ and thus $|t - \mathrm{fl}(t)| = |t| - \mathrm{ufp}(t)$. Consequently, $E_1(t) = 1 - \mathrm{ufp}(t)/|t|$, which is strictly less than $1 - 1/(1+u) = u/(1+u)$.

If $|t| > (1+u)\mathrm{ufp}(t)$, the strict inequality $E_1(t) < u/(1+u)$ follows from the first bound in (2.1). □

In the same way as in the proof above, we can check that when rounding ties 'to even', the bound in the *second* standard model is attained if and only if the nonzero real $t$ has the form $t = (1+u)\mathrm{ufp}(t)$.

To summarize, we have seen how to derive the bounds appearing in the first row of Table 1, that is,

$$E_1(t) \leqslant \frac{u}{1+u} \qquad \text{and} \qquad E_2(t) \leqslant u, \tag{2.2}$$

and characterized their optimality when $t$ can be any *real* number in the normal range of $\mathbb{F}$. In the next section we examine whether these bounds can or cannot be refined further when $t$ is the exact result of a basic arithmetic operation on *floating-point* numbers.

**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. In some of our proofs, it will be convenient to use the following straightforward property about the invariance of relative errors under scaling by an integer power of the radix.

PROPERTY 3.1. *Let $t \in \mathbb{R}$ and $e \in \mathbb{Z}$ be such that both $t$ and $t\beta^e$ lie in the normal range of $\mathbb{F}$. Then $E_1(t) = E_1(t\beta^e)$ and, if $t$ is not a midpoint, $E_2(t) = E_2(t\beta^e)$.*

*Proof.* If $t$ is not a midpoint then its scaled counterpart $t\beta^e$ cannot be a midpoint either, so $\mathrm{fl}(t\beta^e) = \mathrm{fl}(t)\beta^e$ independent of the way the rounding function fl breaks ties. Hence the two claimed equalities. Let us now show that the first equality still holds if $t$ is a midpoint. In this case, $|t - \mathrm{fl}(t)| = u\,\mathrm{ufp}(t)$ and, $t\beta^e$ being also halfway between two consecutive floating-point numbers, $|t\beta^e - \mathrm{fl}(t\beta^e)| = u\,\mathrm{ufp}(t\beta^e)$. Now, $\mathrm{ufp}(t\beta^e) = \mathrm{ufp}(t)\beta^e$, and we conclude that $E_1(t\beta^e) = u\,\mathrm{ufp}(t)/|t| = E_1(t)$. □

In case of midpoints, scaling invariance is lost for $E_2$ when the rounding function fl breaks ties according to the ufp of such midpoints. Fortunately, this does not occur for any of the IEEE 754-2008 roundings: their tie-breaking strategies being independent of $\mathrm{ufp}(t)$, we always have $\mathrm{fl}(t\beta^e) = \mathrm{fl}(t)\beta^e$.

**3.1. Addition, subtraction, and fused multiply-add.** When $t = x + y$ for some $x, y$ in $\mathbb{F}$, the bounds in (2.2) remain optimal, as they can be attained for $(x, y) = (1, u) \in \mathbb{F}^2$. The same holds for subtraction as well as for higher-level operations encompassing addition, like the fused multiply-add operation.

**3.2. Multiplication.** When $t = xy$, the theorem below shows that the bounds in (2.2) remain optimal unless the base $\beta$ and precision $p$ are such that $\beta = 2$ and $2^p + 1$ is a Fermat prime.

THEOREM 3.2. *Let $\beta \geqslant 2$ be even, $p \geqslant 2$, and $x, y \in \mathbb{F}$ be such that $xy$ lies in the normal range of $\mathbb{F}$. We have the following, depending on the value of the base $\beta$:*

- *If $\beta = 2$ then the bounds in (2.2) are optimal if and only if $2^p + 1$ is not prime; furthermore, if $D$ is a non-trivial divisor of $2^p + 1$, then these bounds are attained for $t = xy$ with*
$$(x, y) = \left( \frac{(2 + 2u)\mathrm{ufp}(D)}{D}, \frac{D}{\mathrm{ufp}(D)} \right).$$
- *If $\beta > 2$ then the bounds in (2.2) are optimal, and attained for $t = xy$ with*
$$(x, y) = (2 + 2u, 2^{-1}).$$

*Proof.* Recalling Theorem 2.1, we see that for $t = xy$ the optimality of the bounds in (2.2) is equivalent to the existence of $x, y \in \mathbb{F}$ such that $xy = (1 + u)\mathrm{ufp}(xy)$.

When $\beta > 2$, taking $x = 2 + 2u$ and $y = 2^{-1}$ gives $xy = 1 + u$ and $\mathrm{ufp}(xy) = 1$, from which $xy = (1 + u)\mathrm{ufp}(xy)$ follows. Furthermore, it is easily checked that such values of $x$ and $y$ are in $\mathbb{F}$. (Note however that $2 + 2u \notin \mathbb{F}$ when $\beta = 2$.)

Let us now consider the case $\beta = 2$. With no loss of generality, we can assume $1 \leqslant x, y < 2$. This implies $\mathrm{ufp}(xy) \in \{1, 2\}$ and, since $x, y \in \{1, 1 + 2u, 1 + 4u, \ldots\}$ and $u > 0$, 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

$$XY = (2^p + 1) \cdot 2^{p-1} \qquad \text{and} \qquad 2^{p-1} \leqslant X, Y < 2^p. \tag{3.1}$$

If $2^p + 1$ is prime, then either $X$ or $Y$ must be larger than $2^p$, so (3.1) has no solution.

If $2^p + 1$ is composite, one can construct a solution $(X_0, Y_0)$ to (3.1) as follows. Let $D$ denote a non-trivial 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 $X_0 Y_0$ 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 Hence

$$\mathrm{ufp}(D) + 1 \leqslant D < 2\mathrm{ufp}(D) \qquad \text{and} \qquad D < 2^p. \tag{3.2}$$

Consequently, $\mathrm{ufp}(D) \leqslant 2^{p-1}$, so that $\mathrm{ufp}(D)$ divides $2^{p-1}$ and $Y_0$ is an integer. Furthermore, (3.2) 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 (3.1).

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 $x_0 y_0 = 2 + 2u = (1 + u)\mathrm{ufp}(x_0 y_0)$. □

For the so-called *basic* binary formats of the IEEE 754-2008 standard [5], we have

$$p \in \{24, 53, 113\}.$$

These are special cases of the following two situations, where an explicit value of a non-trivial divisor of $2^p + 1$ can be obtained:

- If $\beta = 2$ and $p$ is odd (for example, $p = 53$ or $p = 113$), then 3 divides $2^p + 1$ and thus

$$(x, y) = (\tfrac{4+4u}{3}, \tfrac{3}{2})$$

  is in $\mathbb{F}^2$ and such that $xy = 2 + 2u$.
- If $\beta = 2$ and $p \equiv 0 \bmod 3$ (for example, $p = 24$), then $2^p + 1$ can be factored as $(2^{p/3} + 1)(2^{2p/3} - 2^{p/3} + 1)$, so that

$$(x, y) = (2 - 2u^{1/3} + 2u^{2/3}, 1 + u^{1/3})$$

  is in $\mathbb{F}^2$ and satisfies $xy = 2 + 2u$.

Currently, the only known Fermat primes are $2^{2^\ell} + 1$ with $\ell \in \{0, 1, 2, 3, 4\}$, so that besides $p \in \{2, 4, 8, 16\}$, no precision is known for which the bounds are not sharp. However, for IEEE floating-point arithmetic, we can go beyond Theorem 3.2 by proving that optimality is guaranteed for any precision $p$. Indeed, for all the binary formats other than the basic ones seen above, the IEEE 754-2008 standard assumes $p$ has a special form, and the lemma below shows that in this case $2^p + 1$ cannot be a Fermat prime.

LEMMA 3.3. *Let $j$ and $k$ be two integers such that $j \geqslant 4$ and $k = 32j$, and let*

$$p = k - d + 13 \quad \textit{with } d \textit{ an integer nearest to } 4\log_2 k.$$

*Then $2^p + 1$ is not prime.*

*Proof.* If $2^p + 1$ is prime then $p$ must be an integer power of two, so that it suffices to check that the latter never occurs for our particular values of $p$.

If $j \in \{4, 5, 6, 7\}$ then $p \in \{113, 144, 175, 206\}$. Hence in this case $p$ is not an integer 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 an integer 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 an integer power of two neither. □

COROLLARY 3.4. *When $t = xy$, the bounds in (2.2) are optimal for all the floating-point formats of the IEEE 754-2008 standard.*

**3.3. Division.** This section focuses on the largest possible relative errors committed when rounding $x/y$ with $x, y$ in $\mathbb{F}$. As the two theorems below show, the bounds in (2.2) can be refined further in base 2, but remain optimal for other bases.

THEOREM 3.5. *Let $\beta \geqslant 2$ be even, $p \geqslant 3$, and $x, y \in \mathbb{F}$ be such that $x/y$ lies in the normal range of $\mathbb{F}$. Then*

$$E_1(x/y) \leqslant u - 2u^2 \qquad \textit{if } \beta = 2,$$

*and*

$$E_1(x/y) \leqslant \frac{u}{1+u} \qquad \text{if } \beta > 2.$$

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

*Proof.* Consider first the case where $\beta = 2$. Writing $t = x/y$, we can assume that $t > 0$ with $x, y > 0$; applying Property 3.1, we can assume further that $1 \leqslant t < 2$, and since $E_1(t)$ is zero for $t = 1$, 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 $y \leqslant (2-2u)\mathrm{ufp}(y)$,

$$t \geqslant 1 + 2u\frac{\mathrm{ufp}(y)}{y} \geqslant 1 + \frac{u}{1-u} = \frac{1}{1-u}. \tag{3.3}$$

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

$$\mathrm{fl}(t) \in \{1+2u, 1+4u, \ldots\}. \tag{3.4}$$

If $\mathrm{fl}(t) \geqslant 1+4u$ then $t \geqslant 1+3u$ and, applying (2.1) together with $\mathrm{ufp}(t) = 1$, we obtain $E_1(t) \leqslant \frac{u}{1+3u}$; this bound is at most $u - 2u^2$ as soon as $u \leqslant \frac{1}{6}$, which holds for $\beta = 2$ and $p \geqslant 3$.

If $\mathrm{fl}(t) = 1+2u$ then either $\mathrm{fl}(t) \leqslant t \leqslant 1+3u$, in which case $E_1(t) = 1 - \frac{1+2u}{t} \leqslant \frac{u}{1+3u}$, or $\mathrm{fl}(t) > t$, in which case (3.3) gives $E_1(t) = \frac{1+2u}{t} - 1 \leqslant u - 2u^2$. This concludes the proof of the bound on $E_1(t)$ when $\beta = 2$. This bound is attained at $x = 1$ and $y = 1-u$, since we then have $x, y \in \mathbb{F}$ as well as $x/y = 1/(1-u)$ and $\mathrm{fl}(x/y) = 1+2u$.

Consider now $\beta > 2$. In this case, the upper bound on $E_1(t)$ is the one already established for any real $t$ in the normal range of $\mathbb{F}$. To prove that this bound is attained when dividing $x = 2+2u$ by $y = 2$, it suffices to check that $x, y \in \mathbb{F}$ and that $x/y = 1+u$. □

THEOREM 3.6. *Let $\beta \geqslant 2$ be even, $p \geqslant 3$, and $x, y \in \mathbb{F}$ be such that $x/y$ lies in the normal range of $\mathbb{F}$. Then*

$$E_2(x/y) \leqslant \frac{u - 2u^2}{1 + u - 2u^2} \qquad \text{if } \beta = 2,$$

*and*

$$E_2(x/y) \leqslant u \qquad \text{if } \beta > 2.$$

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

*Proof.* The attainability of the two bounds can be shown in exactly the same way as for Theorem 3.5. Furthermore, when $\beta > 2$ the upper bound $u$ on $E_2$ is the general one already established earlier. The rest of the proof thus establishes the bound $\frac{u-2u^2}{1+u-2u^2}$ when $\beta = 2$.

We can assume $x, y > 0$ and let $t = x/y$. Since $t$ is not a midpoint [7], we can apply Property 3.1 and proceed as in the proof of Theorem 3.5 to restrict to the situation where $1 < t < 2$, for which (3.3) and (3.4) are true.

If $\mathrm{fl}(t) \geqslant 1+4u$ then, applying (2.1) together with $\mathrm{ufp}(t) = 1$, we obtain $E_2(t) \leqslant \frac{u}{1+4u}$. This bound at most $\frac{u-2u^2}{1+u-2u^2}$ as soon as $u \leqslant \frac{1}{6}$, which holds for $\beta = 2$ and $p \geqslant 3$.

Assume now that $\mathrm{fl}(t) = 1+2u$. If $t \leqslant \mathrm{fl}(t)$ then $E_2(t) = 1 - \frac{t}{1+2u}$, and we deduce from (3.3) that $E_2(t) \leqslant 1 - \frac{1}{(1-u)(1+2u)} = \frac{u-2u^2}{1+u-2u^2}$, as wanted. If $t > \mathrm{fl}(t)$ then

$$E_2(t) = \frac{t}{1+2u} - 1 \tag{3.5}$$

with $1+2u < t < 1+3u$. This upper bound on $t$, which we have used to bound $E_1(t)$ in Theorem 3.5 is now not enough to bound $E_2(t)$ as wanted. We can improve it as follows. The range of $t$ implies $y + 2u\,\mathrm{ufp}(y) < x < y + 6u\,\mathrm{ufp}(y)$ and, since $x$ is in $\mathbb{F}$, we deduce that

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

On the other hand, the floating-point number $y$ can be written

$$y = (1+2ku)\mathrm{ufp}(y), \qquad k \in \{0, 1, \ldots, 2^{p-1}-1\}.$$

Consequently, $t = 1 + \frac{4u}{1+2ku}$ and the condition $t < 1+3u$ is equivalent to $k > \frac{2^{p-1}}{3}$. Since $k$ is an integer, the latter inequality is equivalent to $k \geqslant \frac{2^{p-1}+1}{3}$. Hence $2ku \geqslant \frac{1+2u}{3}$ and we arrive at the upper bound

$$t \leqslant \frac{2+7u}{2+u} = u + 3u - \tfrac{3}{2}u^2 + \mathcal{O}(u^3). \tag{3.6}$$

From (3.5) and (3.6) it follows that $E_2(t) \leqslant \frac{2u(1-u)}{(2+u)(1+2u)}$, which is less than $\frac{u-2u^2}{1+u-2u^2}$ for $\beta = 2$ and $p \geqslant 3$. This concludes the proof of the bound on $E_2(t)$ when $\beta = 2$. □

**3.4. Square root.** Finally, we show how to refine further the bounds (2.2) on $E_1(t)$ and $E_2(t)$ in the special case where $t = \sqrt{x}$ for some positive floating-point number $x$, thereby establishing the bounds in the last row of Table 1.

THEOREM 3.7. *Let $\beta \geqslant 2$ be even, $p \geqslant 2$, and $x \in \mathbb{F}_{>0}$. Then*

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

*and this bound is attained for $x = 1+2u$.*

*Proof.* Defining $t = \sqrt{x}$, we distinguish between the following two cases:

- If $t \geqslant (1+u)\mathrm{ufp}(t)$ then $t^2 = x \geqslant (1+4u)\mathrm{ufp}(t)$, so that $t \geqslant \sqrt{1+4u}\,\mathrm{ufp}(t)$. This implies $E_1(t) \leqslant \varphi$ with $\varphi = \frac{u}{\sqrt{1+4u}}$, and it can be checked that $\varphi \leqslant 1 - \frac{1}{\sqrt{1+2u}}$ when $u \leqslant 1/2$.
- If $t < (1+u)\mathrm{ufp}(t)$ then $\mathrm{fl}(t) = \mathrm{ufp}(t)$ and due to the fact that $t^2$ is a floating-point number, $t \leqslant \sqrt{1+2u}\,\mathrm{ufp}(t)$. Therefore, $E_1(t) = 1 - \mathrm{ufp}(t)/t \leqslant 1 - \frac{1}{\sqrt{1+2u}}$.

Furthermore, $x = 1+2u$ implies $\mathrm{fl}(\sqrt{x}) = 1$, and thus the bound is attained for this value of $x$. □

THEOREM 3.8. *Let $\beta \geqslant 2$ be even, $p \geqslant 2$, and $x \in \mathbb{F}_{>0}$. Then*

$$E_2(\sqrt{x}) \leqslant \sqrt{1+2u} - 1,$$

*and this bound is attained for $x = 1+2u$.*

*Proof.* Writing $t = \sqrt{x}$ we have $t > 0$ and $\mathrm{fl}(t) \geqslant \mathrm{ufp}(t) > 0$ and we consider the following two subcases:

- If $\mathrm{fl}(t) = \mathrm{ufp}(t)$ then $t$ is at most $(1+u)\mathrm{ufp}(t)$, which implies $t \leqslant \sqrt{1+2u}\,\mathrm{ufp}(t)$ since $t^2 = x$ is in $\mathbb{F}$. Hence $E_2(t) = t/\mathrm{fl}(t) - 1 \leqslant \sqrt{1+2u} - 1$.
- If $\mathrm{fl}(t) > \mathrm{ufp}(t)$ then $\mathrm{fl}(t) \geqslant (1+2u)\mathrm{ufp}(t)$. On the other hand, $|t - \mathrm{fl}(t)| \leqslant u\mathrm{ufp}(t)$, so that $E_2(t) \leqslant \frac{u}{1+2u} < \sqrt{1+2u} - 1$ for $u$ positive.

Furthermore, $x = 1 + 2u$ implies $\mathrm{fl}(\sqrt{x}) = 1$, and thus the bound is attained for this value of $x$. □

**4. Application examples.** We now illustrate with a few examples the two main benefits of the refinements presented in the previous section: in some cases, more concise and slightly sharper bounds can be obtained and, in other ones, much shorter proofs of existing bounds can be given. In both cases, the derivations are (relatively) straightforward and do not involve ufp-based discussions.

**4.1. Example 1: Small arithmetic expressions.** Rounding error analyses as those done in [4] typically involve bounds on $|\theta_n|$, where $\theta_n$ is an expression of the form $\theta_n = \prod_{i=1}^{n}(1+\delta_i) - 1$ with $\delta_i$ a relative error term associated with a single floating-point operation. Using the classical bound $|\delta_i| \leqslant u$ given by the first standard model, it is easily checked that for all $n$,

$$-nu \leqslant \theta_n \leqslant (1+u)^n - 1. \tag{4.1}$$

Note that only the upper bound has the form $nu + \mathcal{O}(u^2)$, that is, contains terms nonlinear in $u$. From (4.1) 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 $\gamma_n = nu/(1-nu)$.

By applying Theorem 2.1, we can replace $|\delta_i| \leqslant u$ by $|\delta_i| \leqslant u/(1+u)$, which immediately leads to the refined enclosure

$$-\frac{nu}{1+u} \leqslant \theta_n \leqslant \left(1 + \frac{u}{1+u}\right)^n - 1. \tag{4.2}$$

Although the upper bound in (4.2) still has $\mathcal{O}(u^2)$ terms in general, it is bounded by $nu$ as long as $n \leqslant 3$. Consequently, by just applying the refinement of the first standard model, we can replace $\gamma_n = nu + \mathcal{O}(u^2)$ by $nu$ in every error analysis where $\theta_n$ appears with $n \leqslant 3$. This is the case when evaluating 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 $\gamma_3$, which we now replace by $3u$.

**4.2. 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

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

Assuming the (first) standard model of floating-point arithmetic and barring underflow and overflow, Wilkinson shows in [11, §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

$$-2u < \delta_{ii} \leqslant 2u + u^2 \quad \text{and} \quad -\tfrac{3}{2}u < \delta_{ij} \leqslant \tfrac{3}{2}u + \mathcal{O}(u^2) \quad \text{if } i \neq j. \tag{4.3}$$

Notice that the terms $u^2$ and $\mathcal{O}(u^2)$ come exclusively from the upper bounds on $\epsilon_k$, $\epsilon'_k$, $\epsilon''_k$. By using Theorem 2.1, which says $|\delta| \leqslant u/(1+u)$ instead of just $|\delta| \leqslant u$, 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 (4.3) can be replaced by the following more concise and slightly sharper ones:

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

**4.3. Example 3: Euclidean norm of an $n$-dimensional vector.** Given $x_1, \ldots, x_n$ in $\mathbb{F}$, let the 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 first standard model, all we can say is $\widehat{s} = (\sum_{i=1}^n x_i^2)(1+\theta)$ with $(1-u)^n - 1 \leqslant \theta \leqslant (1+u)^n - 1$, and $\widehat{r} = \sqrt{\widehat{s}}\,(1+\delta)$ with $|\delta| \leqslant u$. Consequently,

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

where $\epsilon = \sqrt{1+\theta} \cdot (1+\delta) - 1$ 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 Lemma A.1), the upper bound is strictly larger than this, so the standard model gives only

$$-(n/2+1)u \leqslant \epsilon \leqslant (n/2+1)u + \mathcal{O}(u^2). \tag{4.4}$$

To avoid the $\mathcal{O}(u^2)$ term above, we can simply apply Lemma 2.1, which says $|\delta| \leqslant u/(1+u)$, together with the improved bound for inner products from [8], which says $|\theta| \leqslant nu$. Indeed, from these two bounds we deduce that $\epsilon$ 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 (4.4), we conclude that

$$|\epsilon| \leqslant (n/2+1)u. \tag{4.5}$$

In particular, evaluating the hypotenuse $\sqrt{x_1^2 + x_2^2}$ in floating-point produces a relative error of at most $2u$.

Of course, the bound in (4.5) also applies when scaling by integer powers of the base is introduced to avoid underflow and overflow.

**4.4. 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 two standard models 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 [4, 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 [9] 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 [9, Theorem 4.4], one of the ingredients used to suppress the $\mathcal{O}(u^2)$ term is the following property:

$$\Big(a \in \mathbb{F}_{\geqslant 0} \quad \text{and} \quad b = \mathrm{fl}\big(\sqrt{a}\big)\Big) \quad \Rightarrow \quad |b^2 - a| \leqslant 2ub^2. \tag{4.6}$$

In [9] it is shown that this property may not hold if only the (second) standard model 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 (4.6) is given, which is about 10 lines long and based on a ufp-based case analysis.

Instead, our refinement of the second standard model for square root provides a much shorter argument: using Theorem 3.8, we see that $b(1+\delta) = \sqrt{a}$ with $|\delta| \leqslant \sqrt{1+2u} - 1$; hence $b^2 - a = -(2+\delta)\delta \cdot b^2$ and the range of $\delta$ leads to

$$(2+\delta)\delta \in \big[2u + 4 - 4\sqrt{1+2u}, 2u\big] \subset [-2u, 2u],$$

from which (4.6) follows immediately.

**4.5. 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[1] it was shown in [6] 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 [6, Theorem 3.1], whose proof is rather long. As we shall see, a much more concise proof can be obtained by applying directly the refined version of the first standard model.

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

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

$$\begin{aligned}|z - \widehat{z}|^2 &= (R - \widehat{R})^2 + (I - \widehat{I})^2 \\ &\leqslant \lambda^2|z|^2 + 2\lambda\mu A + \mu^2 B,\end{aligned} \tag{4.7}$$

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

$$A, B \leqslant |z|^2. \tag{4.8}$$

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

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

Thus, combining (4.7) and (4.8), $|z - \widehat{z}| \leqslant (\lambda + \mu)|z|$. Since our refined model gives $|\delta_i| \leqslant u/(1+u)$ for all $i$, we can take $\lambda = u/(1+u)$ 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 $2u - u^2 + O(u^4)$ as $u \to 0$. Thus, our approach not only yields a much shorter proof of the bound $2u$ of [6], but it also improves slightly on that bound.

**Appendix A. A useful inequality.**

We give below a slight variant of the generalized Bernoulli inequality [2, p. 10, Exercise 18(d)]. A detailed proof is given for the sake of completeness.

LEMMA A.1. *Let $u, x \in \mathbb{R}$ be such that $0 \leqslant u < 1$ and $x \geqslant 1$. Then*

$$(1-u)^x \geqslant 1 - xu.$$

*Proof.* We show that for each $u$, the difference $f_u(x) := (1-u)^x - 1 + xu$ is nonnegative over $[1, +\infty)$. Since $f_u(1) = 0$, it suffices to check that $f_u'(x) \geqslant 0$ for all $x \geqslant 1$. We have $f_u'(x) = \ln(1-u)\cdot(1-u)^x + u$, where $\ln(1-u) \leqslant 0$ and $x \mapsto (1-u)^x$ is decreasing over $[1, +\infty)$. Therefore, $f_u'$ is increasing over $[1, +\infty)$, and all we need is $f_u'(1) \geqslant 0$ for all $u \in [0, 1)$. This is true because $g(u) := f_u'(1) = \ln(1-u)\cdot(1-u) + u$ satisfies $g(0) = 0$ and $g'(u) = -\ln(1-u) \geqslant 0$. ∎

## REFERENCES


[1] R. P. Brent, C. Percival, and P. Zimmermann, *Error bounds on complex floating-point multiplication*, Math. Comp., 76 (2007), pp. 1469–1481.

[2] N. L. Carothers, *Real analysis*, Cambridge University Press, 2000.

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

[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, second ed., 2002.

[5] IEEE Computer Society, *IEEE Standard for Floating-Point Arithmetic*, IEEE Standard 754-2008, Aug. 2008. available at `http://ieeexplore.ieee.org/servlet/opac?punumber=4610933`.

[6] C.-P. Jeannerod, P. Kornerup, N. Louvet, and J.-M. Muller, *Error bounds on complex floating-point multiplication with an FMA*, Sept. 2013. Submitted to Math. Comp.

[7] 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), pp. 228–241.

[8] C.-P. Jeannerod and S. M. Rump, *Improved error bounds for inner products in floating-point arithmetic*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 338–344.

[9] S. M. Rump and C.-P. Jeannerod, *Improved backward error bounds for LU and Cholesky factorizations*, July 2013. Submitted to SIAM J. Matrix Anal. Appl.

[10] S. M. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation, Part I: Faithful rounding*, SIAM J. Sci. Comput., 31 (2008), pp. 189–224.

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