The relative accuracy of $(x+y)*(x-y)$

Claude-Pierre Jeannerod

THE RELATIVE ACCURACY OF (x+y)*(x-y)

CLAUDE-PIERRE JEANNEROD

Abstract. We consider the relative accuracy of evaluating $(x+y)(x-y)$ in IEEE floating-point arithmetic, when x, y are two floating-point numbers and rounding is to nearest. This expression can be used, for example, as an efficient cancellation-free alternative to $x^2 - y^2$ and (at least in the absence of underflow and overflow) is well known to have low relative error, namely, at most about $3u$ with u the unit roundoff. In this paper we propose to complement this traditional analysis with a finer-grained one, aimed at improving and assessing the quality of that bound. Specifically, we show that if the tie-breaking rule is *to away* then the bound $3u$ is asymptotically optimal. In contrast, if the tie-breaking rule is *to even*, we show that asymptotically optimal bounds are now $2.25u$ for base two and $2u$ for larger bases (such as base ten). In each case, asymptotic optimality is obtained by the explicit construction of a certificate, that is, some floating-point input (x, y) parametrized by u and for which the error of the result is equivalent to the error bound as u tends to zero.

1. Introduction

In IEEE floating-point arithmetic, evaluating $x^2 - y^2$ in the most natural way as the difference of two squares is well known to be prone to damaging cancellation: if the floating-point numbers x and y are close enough to each other then the subtraction mostly reveals the rounding error(s) due to squaring and this can yield a totally wrong result (equal to zero, say, although $x \neq y$ and underflow has not occurred). A classical workaround consists in evaluating the factored form $(x + y)(x - y)$ instead, as suggested by Sterbenz [13, p. 118] and Kahan and Thomas [8]. This second formula retains the simplicity of the first one and, barring underflow and overflow, now ensures high relative accuracy. Specifically, if each of the three operations $+$, $-$, $\times$ is performed with relative error at most the unit roundoff u then the computed result $\widehat{r}$ has the form $\widehat{r} = (x^2-y^2)(1+\theta)$ with $(1-u)^3 \leqslant 1+\theta \leqslant (1+u)^3$ and, therefore, has its relative error $|\theta|$ bounded as

$$|\theta| \leqslant (1 + u)^3 - 1.$$

For simplicity and assuming u is small enough, the expression $(1 + u)^3 - 1$ can then be approximated by $3u$ or rewritten as $3u + O(u^2)$ or bounded further by $3u/(1-3u) =: \gamma_3$ or $3.03u$. This kind of analysis is typical of Wilkinson's traditional approach [16, 17] and has been done by various authors, including Stoer [15] (see also Stoer and Bulirsch [14]), Goldberg [2], and Higham [3].

In this paper we propose to complement this traditional analysis with a finer-grained one, aimed at improving and assessing the quality of such error bounds in the context of IEEE floating-point arithmetic. To do this, the implementation of

April 16, 2019.
2010 *Mathematics Subject Classification.* Primary 65G50.

the formula $(x+y)(x-y)$ that we shall study can be described as follows:

$$x, y \in \mathbb{F}: \qquad \widehat{r}_1 := \mathrm{fl}(x+y), \qquad \widehat{r}_2 := \mathrm{fl}(x-y), \qquad \widehat{r} := \mathrm{fl}(\widehat{r}_1\widehat{r}_2), \tag{1}$$

where $\mathbb{F}$ is a set of floating-point numbers in base β and precision p, defined as

$$\mathbb{F} = \{0\} \cup \left\{ M \cdot \beta^E \,:\, M, E \in \mathbb{Z},\ \beta^{p-1} \leqslant |M| < \beta^p \right\}, \tag{2}$$

and where fl denotes a round-to-nearest function from $\mathbb{R}$ to $\mathbb{F}$, such that

$$|\mathrm{fl}(t) - t| = \min_{x\in\mathbb{F}} |x - t| \qquad \text{for all } t \in \mathbb{R}. \tag{3}$$

We shall assume that β is even, $p \geqslant 2$, and the tie-breaking rule for fl is either *to even* or *to away*: if fl breaks ties to even, then every real number lying halfway between two consecutive elements of $\mathbb{F}$ in (2) is rounded to the one whose integral significand M is even; if fl breaks ties to away, then it is rounded to the one for which $|M|$ is largest. In practice these assumptions are very mild and will be enough to cover simultaneously the possibilities offered by the IEEE 754 standard [4], where $\beta \in \{2, 10\}$ and $\mathrm{fl} \in \{\mathrm{roundTiesToEven}, \mathrm{roundTiesToAway}\}$. Furthermore, since the definition of $\mathbb{F}$ imposes no restriction on the exponent range, our results will hold as long as underflows and overflows do not occur.

As a first and easy step towards a fine-grained accuracy analysis of (1), we can exploit the main consequence of (2) and (3), that says that the relative error due to rounding is bounded as follows [9, p. 232]:

$$\text{for all } t \in \mathbb{R}, \qquad \mathrm{fl}(t) = t \cdot (1+\delta), \qquad |\delta| \leqslant \frac{u}{1+u}, \qquad u := \frac{1}{2}\beta^{1-p}. \tag{4}$$

Applying this inequality three times to (1), we deduce that

$$\widehat{r}_1 = (x+y)(1+\delta_1), \qquad \widehat{r}_2 = (x-y)(1+\delta_2), \qquad \widehat{r} = \widehat{r}_1\widehat{r}_2(1+\delta_3) \tag{5}$$

for some rational numbers δ_i such that $|\delta_i| \leqslant u/(1+u)$ and, writing

$$\widehat{r} = (x^2 - y^2)(1+\theta), \qquad \theta := (1+\delta_1)(1+\delta_2)(1+\delta_3) - 1,$$

we can then easily check using $u > 0$ that the relative error of $\widehat{r}$ satisfies $|\theta| < 3u$. This simple analysis already refines the traditional bound $(1+u)^3 - 1$ slightly, showing that the $O(u^2)$ term can be removed and that the commonly used alternative forms $3u + O(u^2)$ and γ_3 mentioned before are in fact not needed. However, this says nothing about the quality of all these bounds and, therefore, raises the following question: *Can the leading constant* 3 *be reduced further, and if so by which value should it be replaced?*

We show with Theorem 1.1 below that the answer actually depends on the tie-breaking rule and the base: if ties are broken to away, then 3 is indeed best possible; otherwise, this constant can be decreased down to 2.25 for binary arithmetic and 2 for larger bases, these new constants now being best possible as well. Here "best possible" means that we have constructed an input $(x_0, y_0) \in \mathbb{F}^2$ that is parametrized by the unit roundoff u and such that the ratio (relative error of $\widehat{r}$ for this input)/(relative error bound) tends to 1 as $u \to 0$. We will say that such a bound is *asymptotically optimal* and may call the input (x_0, y_0) a *certificate of asymptotic optimality* for that bound.

Theorem 1.1. *Let $x, y \in \mathbb{F}$ and let* fl *denote a round-to-nearest map from $\mathbb{R}$ to $\mathbb{F}$. Then, when evaluating $x^2 - y^2$ as $\widehat{r} = \mathrm{fl}(\mathrm{fl}(x+y)\mathrm{fl}(x-y))$, the returned value $\widehat{r}$ satisfies*

$$\widehat{r} = (x^2 - y^2)(1+\theta), \qquad |\theta| < \begin{cases} 3u & \text{if fl breaks ties to away,} \\ \frac{9}{4}u & \text{if fl breaks ties to even and } \beta = 2, \\ 2u & \text{if fl breaks ties to even and } \beta \neq 2. \end{cases}$$

Furthermore, each of these bounds on the relative error $|\theta|$ is asymptotically optimal.

In practice this result implies that for default IEEE floating-point arithmetic—which has base 2 and ties broken to even, the overall relative error of evaluating $(x+y)(x-y)$ can never get close to the traditional bound $3u$ and will at worst approach $2.25u$. Our analysis can also be seen as a typical example of *fine-grained accuracy analysis*, as surveyed in [5] and whose goal is to provide not only a priori, worst-case error bounds but also certificates of the quality of such bounds. Other examples include optimal bounds for the five basic operations [7] and for summation in high dimension [11, 10], as well as asymptotically optimal bounds in the context of complex arithmetic [1, 6].

1.1. **Ingredients for the proof.** To establish Theorem 1.1 we shall exploit (4) as well as several other, lower level properties of IEEE floating-point arithmetic that are all straightforward consequences of (2) and (3) and which we briefly review in this subsection. It turns out that several of these properties are conveniently expressed in terms of the unit roundoff $u = \frac{1}{2}\beta^{1-p}$ and also via the real functions ufp (*unit in the first place*, introduced in [12]) and ulp (*unit in the last place*), defined by $\mathrm{ufp}(0) = \mathrm{ulp}(0) = 0$ and

$$\text{for } t \in \mathbb{R}_{\neq 0}, \quad \mathrm{ufp}(t) = \beta^{\lfloor \log_\beta |t| \rfloor} \quad \text{and} \quad \mathrm{ulp}(t) = 2u\,\mathrm{ufp}(t).$$

Clearly, these two functions are even (that is, independent of the sign of t) and non-decreasing over $\mathbb{R}_{>0}$: if $|t| \leqslant |t'|$ then $\mathrm{ufp}(t) \leqslant \mathrm{ufp}(t')$ and $\mathrm{ulp}(t) \leqslant \mathrm{ulp}(t')$.

Some properties of $\mathbb{F}$. Note first that if $x \in \mathbb{F}$ then $-x \in \mathbb{F}$ (symmetry) and $x\beta^k \in \mathbb{F}$ for all $k \in \mathbb{Z}$ (auto-similarity).

Furthermore, since $2u = \beta^{1-p}$, any nonzero $x \in \mathbb{F}$ can be rewritten as

$$x = \pm m\beta^e, \qquad m = 1 + j \cdot 2u, \qquad j \in \{0, 1, 2, \ldots, (\beta-1)\beta^{p-1} - 1\}, \qquad e \in \mathbb{Z}.$$

Here, $\beta^e = \mathrm{ufp}(x)$ and, for example, the subset for which $\mathrm{ufp}(x) = 1$ is

$$\mathbb{F} \cap [1, \beta) = \Big\{1, 1 + 2u, 1 + 4u, \ldots, \beta - 2u\Big\}.$$

It is worth noting that the *midpoints* associated with $\mathbb{F}$, that is, the rational numbers lying halfway between two consecutive elements of $\mathbb{F}$, can be expressed in a similar way as $\pm\big(1 + j\cdot 2u + u\big)\beta^e$. Their set will be written $\mathbb{M}$ and, in particular, $\mathbb{M} \cap [1, \beta) = \Big\{1 + u, 1 + 3u, 1 + 5u, \ldots, \beta - u\Big\}$.

From the definition of ufp, it follows that $\mathrm{ufp}(t) \leqslant |t| < \beta\mathrm{ufp}(t)$ for $t \in \mathbb{R}_{\neq 0}$. Combining this with the structure of $\mathbb{F}$ just described, we deduce that over $\mathbb{F}$ the strict inequality can be refined: for $x \in \mathbb{F}$, $\mathrm{ufp}(x) \leqslant |x| \leqslant (\beta - 2u)\mathrm{ufp}(x)$.

Finally, it will be useful to exploit the fact that floating-point numbers are integral multiples of their ulp: if $x \in \mathbb{F}$ then $x \in \mathrm{ulp}(x)\mathbb{Z}$. Conversely, if a nonzero real number t satisfies $|t| \in \mathrm{ulp}(t)\mathbb{Z}$ and $|t|/\mathrm{ulp}(t) \leqslant \beta^p$, then $t \in \mathbb{F}$.

Some properties of fl**.** A first important property of rounding to nearest is that it is a non-decreasing function over $\mathbb{R}$: for any $t, t' \in \mathbb{R}$ such that $t \leqslant t'$, we have $\mathrm{fl}(t) \leqslant \mathrm{fl}(t')$.

Another property, made possible by the fact that our tie-breaking rules are independent of both the sign and the order of magnitude of the number to be rounded, is that

$$\mathrm{fl}(\pm t\beta^k) = \pm \mathrm{fl}(t)\beta^k, \qquad t \in \mathbb{R}, \qquad k \in \mathbb{Z}. \tag{6}$$

Third, the relative error due to rounding a real number can be bounded by means of the ufp function as follows:

$$\text{for all } t \in \mathbb{R}_{\neq 0}, \qquad \mathrm{fl}(t) = t \cdot (1+\delta), \qquad |\delta| \leqslant u\frac{\mathrm{ufp}(t)}{|t|}.$$

This bound, which improves upon the bound $u/(1+u)$ given in (4) as soon as $|t| > (1+u)\mathrm{ufp}(t)$, can be as small as about u/β when $|t| \approx \beta\mathrm{ufp}(t)$.

Finally, $\mathrm{fl}(t) = t$ whenever $t \in \mathbb{F}$. This obvious fact, which is not implied by (4), may be used under the form $t \notin \mathbb{F} \Rightarrow \delta \neq 0$.

Sufficient conditions to ensure $|\theta| < 2u$**.** In addition to the low-level features that we have just recalled, we will use the following two facts:

$$\delta_i\delta_j \leqslant 0 \qquad \Rightarrow \qquad |\theta| < 2u \tag{7}$$

and

$$x, y \in \mathbb{F} \,:\, 0 \leqslant y \leqslant x \leqslant y + \beta\mathrm{ufp}(y) \qquad \Rightarrow \qquad x - y \in \mathbb{F}. \tag{8}$$

The implication in (7) follows immediately from $\theta = (1+\delta_1)(1+\delta_2)(1+\delta_3) - 1$ and the bounds $|\delta_i| \leqslant u/(1+u)$ for $i = 1, 2, 3$.

On the other hand, the implication in (8) can be referred to as the Sterbenz–Ziv property [13, 18] and says that if two floating-point numbers are close enough to each other, then their exact difference is itself a floating-point number.

Proving this result is particularly simple thanks to the properties of $\mathbb{F}$, ufp, and ulp seen above: when $y = 0$ it is obvious; otherwise, $\mathrm{ulp}(y)$ divides $\mathrm{ulp}(x)$ because $y \leqslant x$, so that x and y and $x - y$ are integral multiples of $\mathrm{ulp}(y)$. Hence $x - y = M\mathrm{ulp}(y)$ for some integer M that by assumption satisfies $0 \leqslant M \leqslant \beta\mathrm{ufp}(y)/\mathrm{ulp}(y) = \beta^p$. Therefore, $x - y \in \mathbb{F}$.

Note also that the condition in (8) is essentially best possible in the sense that if $x > y + \beta\mathrm{ufp}(y)$, then $x - y$ need not be in $\mathbb{F}$ anymore. For example, taking $x = \beta + 1 + \beta \cdot 2u$ and $y = 1 + 2u$ gives $x > \beta + 1 + 2u = y + \beta\mathrm{ufp}(y)$ and $x - y = \beta + (\beta - 1) \cdot 2u \notin \mathbb{F}$.

Finally, remark that $x - y \in \mathbb{F}$ is equivalent to $\delta_2 = 0$, which by (7) implies $|\theta| < 2u$. Consequently, both (7) and (8) will be used as ways to filter out various easy sub-cases that occur during the proof of the upper bounds $2.25u$ and $2u$ in Theorem 1.1.

1.2. **Outline.** The rest of the paper is devoted to the proof of Theorem 1.1. We begin, in section 2, by setting up a certificate showing that if ties are broken to *away* then the bound $3u$ is asymptotically optimal. We go on to consider round to nearest *even* in the next two sections. First, we show in section 3 that for this choice of tie-breaking rule smaller upper bounds are possible, namely, $2.25u$ when $\beta = 2$ and $2u$ for larger bases: after a preliminary range reduction on the input x and y, we focus on the resulting three sub-cases, which are $x \pm y \in [1, \beta)$, $x + y > \beta$,

and $x - y < 1$, and provide for each of them a detailed analysis. We conclude in section 4 with the construction of two certificates of asymptotic optimality, one for the bound $2.25u$ in the case where $\beta = 2$, and one for the bound $2u$.

2. Asymptotic optimality of the error bound $3u$ when rounding ties to away

Lemma 2.1. *Assume that $p \geqslant 4$, and let*

$$j = \left\lceil \frac{1}{2\sqrt{u}} \right\rceil, \qquad x = 1 + j \cdot 2u, \qquad y = u.$$

Then $x, y \in \mathbb{F}$ and, for rounding ties to away, $\theta = 3u - \epsilon$ with $\epsilon = O(u^{3/2})$.

Proof. Note first that $y \in \mathbb{F}$ for β even. On the other hand, for $p \geqslant 4$, we have $u \leqslant 1/16$, which together with $1 \leqslant j < \frac{1}{2\sqrt{u}} + 1$ implies that both $x = 1 + j \cdot 2u$ and the expression $1 + (2j + 1) \cdot 2u$ are in $(1, \beta)$ and, thus, in $\mathbb{F} \cap (1, \beta)$.

We deduce from $x \in \mathbb{F} \cap (1, \beta)$ and $y = u$ that

$$x + y = 1 + j \cdot 2u + u, \qquad x - y = 1 + j \cdot 2u - u$$

are midpoints in $(1, \beta)$; furthermore, since ties are broken to away, they are rounded up to $\widehat{r}_1 = 1+(j+1)\cdot 2u$ and $\widehat{r}_2 = 1+j\cdot 2u$, respectively. Now, using $j = \frac{1}{2\sqrt{u}}+O(1)$, we see that $x \pm y = 1+O(\sqrt{u})$ and that the associated relative errors have the form

$$\delta_i = u - \epsilon_i, \qquad \epsilon_i = O(u^{3/2}), \qquad i = 1, 2.$$

Consider now the relative error δ_3 that occurs when rounding $\widehat{r}_1\widehat{r}_2$. We have

$$\widehat{r}_1\widehat{r}_2 = 1 + (2j + 1) \cdot 2u + j(j + 1) \cdot 4u^2,$$

where, as noted above, $1+(2j+1)\cdot 2u$ belongs to $\mathbb{F}\cap[1, \beta)$ and where it can be checked that $j(j+1)\cdot 4u^2 \in (u, 2u)$ for all $p \geqslant 4$: the lower bound follows from $j \geqslant \frac{1}{2\sqrt{u}}$; for the upper bound we use $j < \frac{1}{2\sqrt{u}}+1$ to deduce that $j(j+1)\cdot 4u^2 < u+6u^{3/2}+8u^2$, which is at most $2u$ if $p \geqslant 6$; if $p \in \{4, 5\}$, then $\varphi_{\beta,p} := j(j+1)\cdot 4u^2$ satisfies for all $\beta \geqslant 3$, $\varphi_{\beta,p} < \varphi_{2,p} = \frac{3}{4} \cdot 2u < 2u$. Consequently, the exact product $\widehat{r}_1\widehat{r}_2$ is rounded up to $\widehat{r} = 1 + (2j + 2) \cdot 2u$; since it also has the form $\widehat{r}_1\widehat{r}_2 = 1 + O(\sqrt{u})$ and since $j(j + 1) \cdot 4u^2 = u + O(u^{3/2})$, we deduce that

$$\delta_3 = u - \epsilon_3, \qquad \epsilon_3 = O(u^{3/2}).$$

The conclusion then follows from $\widehat{r}/(x^2 - y^2) = 1 + \delta_1 + \delta_2 + \delta_3 + O(u^2)$. □

3. New error bounds when rounding ties to even

The goal of this section is to establish the following theorem, which says that when the rounding map breaks ties to even the traditional bound $3u$ can be reduced further, depending on the value of the base.

Theorem 3.1. *If* fl *breaks ties to even, then*

$$|\theta| < \begin{cases} 2.25u & \text{if } \beta = 2, \\ 2u & \text{otherwise.} \end{cases}$$

In the rest of this section, we first reduce the range of both x and y. This yields the following three regimes: $x \pm y \in [1, \beta)$, $x + y > \beta$, and $x - y < 1$. Then, for each of them, we eliminate the easy subcases corresponding to (7) and (8), and deal finally with the remaining, nontrivial subcases with Lemmas 3.1, 3.2, 3.3, respectively.

3.1. **Range reduction.** Letting $\varphi : (x, y) \mapsto \mathrm{fl}(\mathrm{fl}(x+y)\mathrm{fl}(x-y))$, we deduce from the property of rounding in (6) that $\varphi(x, y) = \varphi(-x, y) = \varphi(x, -y) = -\varphi(y, x)$ and that $\varphi(x\beta^k, y\beta^k) = \varphi(x, y)\beta^{2k}$ for any $k \in \mathbb{Z}$. Since these equalities hold for the exact expression $x^2 - y^2$ as well and since $\varphi(0, 0) = 0$ is the exact result when $x = y = 0$, we can restrict our error analysis to the pairs (x, y) such that

$$0 \leqslant y \leqslant x \qquad \text{and} \qquad 1 \leqslant x < \beta.$$

Two further restrictions can be made. First, if $y \geqslant 1$ then $\mathrm{ufp}(y) = 1$ and, by the Sterbenz–Ziv property in (8), we obtain $\delta_2 = 0$; hence in this case $|\theta| < 2u$ and Theorem 3.1 is proved. Second, if $y < u$ then $(\widehat{r}_1, \delta_3) = (1, 0)$ for $x = 1$ and $\delta_1 \leqslant 0 \leqslant \delta_2$ for $x \geqslant 1 + 2u$. Thus, $y < u$ implies $|\theta| < 2u$ as well. Consequently, we can reduce the range of y accordingly and assume from now on that

$$u \leqslant y < 1 \leqslant x < \beta. \tag{9}$$

This reduced range for x and y has several implications. First, $\mathrm{ufp}(x) = 1$ and the range of $\mathrm{ufp}(y)$ is itself reduced: using $u \geqslant \beta^{-p}$ for $\beta \geqslant 2$, we see that $\beta^{-p} \leqslant \mathrm{ufp}(y) \leqslant \beta^{-1}$, or, equivalently,

$$1 \leqslant e \leqslant p, \qquad \mathrm{ufp}(y) =: \beta^{-e} \leqslant y < \beta^{1-e}.$$

Second, since $y \in \mathbb{F}_{<1}$ implies $y \leqslant 1 - 2u/\beta = 1 - \beta^{-p}$, we deduce from (9) that

$$x + y \in [1, \beta) \cup [\beta, \beta^2) \qquad \text{and} \qquad x - y \in [\beta^{-p}, 1) \cup [1, \beta).$$

If $x+y = \beta$, then $\delta_1 = 0$ and so $|\theta| < 2u$. Consequently, we are left with the following three nontrivial cases, which we shall analyze separately in the next subsections:

$$x \pm y \in [1, \beta) \qquad \text{or} \qquad x + y > \beta \qquad \text{or} \qquad x - y < 1.$$

3.2. **Analysis when $x \pm y \in [1, \beta)$.** Since x belongs to $\mathbb{F} \cap [1, \beta)$, it has the form $x = 1 + j \cdot 2u$ for some integer j and, on the other hand, setting $k = \lfloor y/(2u) \rfloor$ gives $y = k \cdot 2u + \epsilon$ for some $\epsilon \in [0, 2u)$. Thus

$$x \pm y = 1 + (j \pm k) \cdot 2u \pm \epsilon, \qquad j \pm k \in \mathbb{Z}_{\geqslant 0}, \qquad 0 \leqslant \epsilon < 2u.$$

We now consider two cases. If $\epsilon < u$ or if $\epsilon = u$ with $j + k$ even, then $x + y$ is rounded down, while $x - y$ is rounded up, that is, $\delta_1 \leqslant 0 \leqslant \delta_2$. Likewise, if $\epsilon > u$ or if $\epsilon = u$ with $j + k$ odd, then $\delta_2 \leqslant 0 \leqslant \delta_1$. Hence $\delta_1\delta_2 \leqslant 0$ in all cases and, recalling (7), we arrive at the following result.

Lemma 3.1. *Let $x, y \in \mathbb{F}$ be as in (9) and such that $x \pm y \in [1, \beta)$. If* fl *breaks ties to even, then*

$$|\theta| < 2u.$$

3.3. **Analysis when $x+y>\beta$.** If $x-y\leqslant 1$ then $y=\frac{1}{2}(x+y-(x-y))\geqslant\frac{\beta-1}{2}$. Together with $y<1$, this requires $\beta=2$ and $\mathrm{ufp}(y)=1/2$. We can thus rewrite the assumption $x-y\leqslant 1$ as $x\leqslant y+2\mathrm{ufp}(y)$ and, using the Sterbenz–Ziv property in (8), we deduce that in this case $\delta_2=0$ and $|\theta|<2u$.

Let us now assume that $x-y>1$, so $x\pm y$ are in two consecutive open intervals:

$$x-y\in(1,\beta),\qquad x+y\in(\beta,\beta^2). \tag{10}$$

In this case, the next lemma shows how to bound the overall relative error $|\theta|$ depending on the value of the base β.

Lemma 3.2. *Let $x,y\in\mathbb{F}$ be as in (9) and (10). If* fl *breaks ties to even, then*

$$|\theta|<\begin{cases}2.25u & \text{if } \beta=2,\\ 2u & \text{otherwise.}\end{cases}$$

We give a detailed proof of this result in the rest of this subsection. We focus only on the case where all the δ_i have the same sign, for otherwise the result is clearly true thanks to (7).

3.3.1. *Preliminaries.* Note that $0<y<x$ implies that $x\pm y$ are integral multiples of $\mathrm{ulp}(y)=2u\,\mathrm{ufp}(y)$. Defining, for simplicity,

$$\eta:=\mathrm{ufp}(y)$$

and using the strict inequality $x+y>\beta$ together with $x-y=(x+y)-2y$ and $y\leqslant(\beta-2u)\eta$, we deduce that

$$x+y\geqslant\beta+2u\eta,\qquad x-y\geqslant\beta-2\beta\eta+6u\eta. \tag{11}$$

(We shall use the latter bound only when $\beta>2$ or $e\geqslant 2$, that is, only in the cases where it is larger than the lower bound $1+2u\eta$ resulting from $x-y>1$.)

Let $x=1+j\cdot 2u$ and, to handle the fact that $x+y\geqslant\beta$, let us also consider the decompositions $y=k\cdot 2u+\epsilon$ and $j+k=k_1\beta+k_0$, where

$$k:=\lfloor y/(2u)\rfloor,\quad k_1:=\lfloor(j+k)/\beta\rfloor,\quad k_0\in\{0,1,\dots,\beta-1\},\quad 0\leqslant\epsilon<2u. \tag{12}$$

It follows from (10) and (12) that

$$x+y=\underbrace{1+k_1\cdot\beta\cdot 2u}_{\in\mathbb{F}\cap[\beta,\beta^2)}+\underbrace{k_0\cdot 2u+\epsilon}_{\in[0,\beta\cdot 2u)},\qquad x-y=\underbrace{1+(j-k)\cdot 2u}_{\in\mathbb{F}\cap(1,\beta]}-\epsilon. \tag{13}$$

We will also rely on the following specific properties. The first one will be useful for large bases, while the second one holds only for base 2.

Property 3.1. *Assume that $\beta>2$. If $\delta_1\leqslant 0$ or $\delta_3>0$, then*

$$|\delta_3|\leqslant\frac{u}{\beta-2}.$$

Proof. Since $0<\eta\leqslant\beta^{-1}$ and $\beta,\beta-2\in\mathbb{F}$, the lower bounds in (11) imply $\widehat{r}_1\geqslant\beta$ and $\widehat{r}_2\geqslant\beta-2$. Hence $|\delta_3|\leqslant u\,\mathrm{ufp}(\widehat{r}_1\widehat{r}_2)/(\beta(\beta-2))$. To conclude, it suffices to prove that $\mathrm{ufp}(\widehat{r}_1\widehat{r}_2)\leqslant\beta$, that is, $\widehat{r}_1\widehat{r}_2<\beta^2$, which can be done as follows. The ranges of $x\pm y$ in (10) imply that $\widehat{r}_1\leqslant x+y+\beta u$ and $\widehat{r}_2\leqslant x-y+u$. Using $x\leqslant\beta-2u$ then gives $\widehat{r}_1\widehat{r}_2\leqslant\big(\beta(1+u)-2u+y\big)(\beta-u-y)$, which for $\beta\geqslant 2$ and $u,y>0$ implies $\widehat{r}_1\widehat{r}_2<(1+u)\beta^2$. If $\delta_3>0$, then the latter bound suffices to ensure $\widehat{r}_1\widehat{r}_2<\beta^2$ (for otherwise that product is rounded *down* to β^2, a contradiction). If $\delta_1\leqslant 0$, then we start instead with $\widehat{r}_1\leqslant x+y$; this leads to $\widehat{r}_1\widehat{r}_2\leqslant(\beta-2u+y)(\beta-u-y)=(\beta-2u)(\beta-u)+y(u-y)<\beta^2$ for $u\leqslant y$. □

Property 3.2. *Assume that $\beta = 2$ and that* fl *breaks ties to even. If $\delta_1\delta_2 \geqslant 0$ then $\epsilon \neq u$ and, more precisely, either*

$$\epsilon \leqslant (1-2\eta)u \quad \text{or} \quad \epsilon \geqslant (1+2\eta)u.$$

Proof. Assume for contradiction that $\epsilon = u$. This means $x - y$ is halfway between the floating-point numbers $1 + (j-k-1)\cdot 2u$ and $1 + (j-k)\cdot 2u$. If $j-k$ is even, then the choice of tie-breaking rule implies $\delta_2 > 0$; furthermore, $k_0 = (j+k) \bmod 2$ is then equal to zero, which implies $k_0 \cdot 2u + \epsilon < 2u$ and thus $\delta_1 < 0$. If $j-k$ is odd then, using the same reasoning, we deduce that $\delta_2 < 0$ and $\delta_1 > 0$. This shows that if δ_1 and δ_2 have the same sign, then $\epsilon \neq u$. Now, from $y = k \cdot 2u + \epsilon$ and $\mathrm{ufp}(y) = \eta = 2^{-e} < 1$ for $e \geqslant 1$, we deduce that $\epsilon = \ell \cdot 2u\eta$ for some $\ell \in \mathbb{Z}_{\geqslant 0}$. Hence $\epsilon \neq u$ is equivalent to $\ell \neq 1/(2\eta)$. Since $e \geqslant 1$ implies that $1/(2\eta) = 2^{e-1}$ is an integer, we conclude that either $\ell \leqslant 1/(2\eta) - 1$ or $\ell \geqslant 1/(2\eta) + 1$. □

3.3.2. *Case where $\delta_i > 0$ for all i.* In this case $x + y$ and $x - y$ are rounded up in $\mathbb{F}$. Because of (13), this implies that

$$k_0 \cdot 2u + \epsilon \geqslant \beta u, \qquad \epsilon \leqslant u,$$

and that the associated relative errors can be expressed exactly as

$$\delta_1 = \frac{\beta \cdot 2u - (k_0 \cdot 2u + \epsilon)}{x+y}, \qquad \delta_2 = \frac{\epsilon}{x-y}.$$

Since $\epsilon \leqslant u$, we must have $2k_0 \geqslant \beta - 1$, which for β even is equivalent to $k_0 \geqslant \beta/2$. Hence the overall error θ is bounded as

$$1 + \theta \leqslant \left(1 + \frac{\beta u - \epsilon}{x+y}\right)\left(1 + \frac{\epsilon}{x-y}\right)(1+\delta_3) =: F(x, y, \epsilon, \delta_3). \tag{14}$$

Note that F increases with ϵ, since $\partial F/\partial \epsilon = (2y + \beta u - 2\epsilon)(1+\delta_3)/(x^2 - y^2)$ is positive for all $\delta_3 > -1$, $x > y > 0$, $\beta \geqslant 2$, and $\epsilon \leqslant u$.

■ Assume first that $\beta \geqslant 4$. In this case, because of $\eta \leqslant \beta^{-1}$ and (11) and, on the other hand, because of $\delta_3 > 0$ and Property 3.1, we have the bounds

$$\epsilon \leqslant u, \qquad x + y > \beta, \qquad x - y > \beta - 2, \qquad \delta_3 \leqslant \frac{u}{\beta - 2},$$

which, when applied to (14), lead to

$$1 + \theta \leqslant \left(1 + \left(1 - \frac{1}{\beta}\right)u\right)\left(1 + \frac{u}{\beta - 2}\right)^2.$$

It can be checked that the latter bound decreases with $\beta \geqslant 4$, and we conclude that $1 + \theta \leqslant (1 + \frac{3}{4}u)(1 + \frac{1}{2}u)^2 = 1 + \frac{7}{4}u + u^2 + \frac{3}{16}u^3$, which is less than $1 + 2u$. This completes the proof of Lemma 3.2 in the case where $\delta_i > 0$ for all i and $\beta \geqslant 4$.

■ Assume now that $\beta = 2$. Using (11) and Property 3.2, we see that we can take

$$\epsilon \leqslant (1-2\eta)u, \qquad x + y \geqslant 2 + 2u\eta, \qquad x - y \geqslant 2 - 4\eta + 6u\eta, \qquad \delta_3 \leqslant u_1,$$

and, applying these bounds to (14), we arrive at

$$1 + \theta \leqslant \left(1 + \frac{1+2\eta}{2+2u\eta}u\right)\left(1 + \frac{1-2\eta}{2-4\eta+6u\eta}u\right)(1+u_1) =: G(\eta).$$

Now, $\delta_2 > 0$ implies that ϵ cannot be zero, so that $0 < \epsilon \leqslant (1-2\eta)u$. Since $\eta = 2^{-e}$, this forces $e \geqslant 2$ or, equivalently,

$$\eta \leqslant 1/4.$$

If $u = 1/4$, then $G(\eta)$ reaches its maximum at $\eta^* := \frac{121-\sqrt{8113}}{136} = 0.227...$ and so $1+\theta \leqslant G(\eta^*) < 1 + 2.17u < 1 + \frac{9}{4}u$.

If $u \leqslant 1/8$, then one can check that the derivative of G has the form $G'(\eta) = (1+u_1)\cdot H(\eta)/(\cdots)^2$, where H is a degree-2 polynomial in η (whose coefficients are polynomials in u), and that H is positive for all $\eta \leqslant 1/4$. (It is positive at 0 and has two positive roots, the smallest one being larger than $1/4$.) Hence G increases with η over $(0, 1/4]$ and, consequently, $1+\theta \leqslant G(1/4)$, that is,

$$\begin{aligned} 1+\theta &\leqslant \left(1+\frac{\frac{3}{4}u}{1+\frac{u}{4}}\right)\left(1+\frac{\frac{u}{2}}{1+\frac{3}{2}u}\right)(1+u_1) = 1 + \frac{9}{4}u - \frac{5}{16}u^2 + O(u^3) \\ &< 1 + \frac{9}{4}u. \end{aligned}$$

Thus we have proved Lemma 3.2 in the case where $\delta_i > 0$ for all i and $\beta = 2$.

3.3.3. *Case where $\delta_i < 0$ for all i.* In this case, we consider $\widetilde{\theta}$ defined by

$$1 + \widetilde{\theta} := \big(1+|\delta_1|\big)\big(1+|\delta_2|\big)\big(1+|\delta_3|\big) \tag{15}$$

and, using the fact that $-\widetilde{\theta} < \theta < 0$, we focus on determining an upper bound on $\widetilde{\theta}$.

Since δ_1 and δ_2 are negative, $x+y$ and $x-y$ are rounded down in $\mathbb{F}$ and, recalling (13), we see that this implies

$$k_0 \cdot 2u + \epsilon \leqslant \beta u, \qquad \epsilon \geqslant u,$$

and that the (negative) relative errors δ_1 and δ_2 satisfy

$$|\delta_1| = \frac{k_0 \cdot 2u + \epsilon}{x+y}, \qquad |\delta_2| = \frac{2u-\epsilon}{x-y}.$$

Since $\delta_2 \neq 0$, we must have $\epsilon > 0$ and thus, for β even, the constraint $k_0 \cdot 2u + \epsilon \leqslant \beta u$ seen above implies that the integer k_0 satisfies $k_0 \leqslant \beta/2 - 1$. Consequently, $\widetilde{\theta}$ can be bounded in terms of the function F in (14) as follows:

$$1 + \widetilde{\theta} \leqslant F\big(x, y, \widetilde{\epsilon}, |\delta_3|\big), \qquad \widetilde{\epsilon} := 2u - \epsilon.$$

Note that $\epsilon \geqslant u$ implies $\widetilde{\epsilon} \leqslant u$, and recall that F increases with $\widetilde{\epsilon}$ in $[0, u]$.

We can then conclude using the same analysis as in §3.3.2—where all the δ_i were assumed to be positive, and arrive at exactly the same bounds. Simply note that when $\beta \geqslant 4$, Property 3.1 can still be applied because now $\delta_1 < 0$. When $\beta = 2$, Property 3.2 now implies $\epsilon \geqslant (1+2\eta)u$ and thus $\widetilde{\epsilon} \leqslant (1-2\eta)u$; also, it is the fact that $\epsilon < 2u$ (by definition) which implies $\widetilde{\epsilon} > 0$ and thus forces η to satisfy $\eta \leqslant 1/4$.

This terminates the analysis of the case where $\delta_i < 0$ for all i and, therefore, concludes the proof of Lemma 3.2.

3.4. **Analysis when $x - y < 1$.** Here we will show that $|\theta|$ is always less than $2u$. Assume first that the integer e such that $\mathrm{ufp}(y) = \beta^{-e}$ satisfies $e = 1$. In this case $x - y < 1$ is equivalent to $x < y + \beta\mathrm{ufp}(y)$ and we deduce from the Sterbenz–Ziv property in (8) that $\delta_2 = 0$ and thus $|\theta| < 2u$.

Let now consider the situation where

$$e \geqslant 2.$$

The inequalities in (9) can then be replaced by

$$u \leqslant y < \beta^{-1}, \qquad 1 \leqslant x < \beta, \tag{16}$$

which together with $\beta \geqslant 2$ and $x-y<1$ lead to the lower bound $x-y>1-\beta^{-1}\geqslant \beta^{-1}$ and to the upper bound $x+y=(x-y)+2y<1+2\beta^{-1}\leqslant\beta$. We are thus in a situation where $x-y$ and $x+y$ belong to the following consecutive open intervals:

$$x-y\in(\beta^{-1},1),\qquad x+y\in(1,\beta). \tag{17}$$

The next lemma tells us that when ties are broken to even then the bound $2u$ holds in this case too.

Lemma 3.3. *Let $x,y\in\mathbb{F}$ be as in (16) and (17). If* fl *breaks ties to even, then*

$$|\theta|<2u.$$

The rest of this subsection is devoted to the proof of this result. As before, the only nontrivial cases are those where the δ_i are either all positive or all negative.

3.4.1. *Preliminaries.* We write $x=1+j\cdot 2u$ as in the previous subsections, but since $x-y$ is now below 1, we decompose y as $y=k_1\cdot 2u+k_0\cdot 2u/\beta+\epsilon$, where

$$k_1:=\lfloor y/(2u)\rfloor,\qquad k_0\in\{0,1,\ldots,\beta-1\},\qquad 0\leqslant\epsilon<2u/\beta. \tag{18}$$

(The values of j,k_1,k_0,ϵ are determined uniquely by those of x and y.) Then, using (17) and (18), we can check that the exact sum and difference have the form

$$x+y=\underbrace{1+(j+k_1)\cdot 2u}_{\in\mathbb{F}\cap[1,\beta)}+\underbrace{k_0\cdot 2u/\beta+\epsilon}_{\in[0,2u)} \tag{19a}$$

and

$$x-y=\underbrace{1+(j-k_1)\cdot 2u-k_0\cdot 2u/\beta}_{\in\mathbb{F}\cap(\beta^{-1},1]}-\epsilon. \tag{19b}$$

Recall that $\eta=\operatorname{ufp}(y)=\beta^{-e}$. When $\beta=2$, the following two properties will turn out to be useful.

Property 3.3. *Assume that $\beta=2$ and that* fl *breaks ties to even. If $\delta_1\delta_2\geqslant 0$ then either*

$$\epsilon\leqslant(1-4\eta)\frac{u}{2}\quad\text{or}\quad\epsilon\geqslant(1+4\eta)\frac{u}{2}.$$

Proof. We can proceed in the same way as for Property 3.2. First, if $\epsilon=u/2$ then $x-y\in\mathbb{M}$, which due to (19) and the tie-breaking rule implies that δ_1 and δ_2 are nonzero and of opposite signs; this contradicts the assumption $\delta_1\delta_2\geqslant 0$, and so $\epsilon\neq u/2$. Then, since y is an integral multiple of $\operatorname{ulp}(y)=2u\eta=2^{1-e}u$, so is ϵ, and the conclusion follows from applying this fact together with $\eta\leqslant 1/4$ to each of the strict inequalities $\epsilon<u/2$ and $\epsilon>u/2$. □

Property 3.4. *Assume that $\beta=2$ and $x=1$. If $\delta_1\delta_3>0$ then*

$$|\delta_3|\leqslant\frac{u}{(1+2u)^2}.$$

Proof. It follows from (17) and the monotonicity of rounding that $\widehat{r}_1\geqslant 1$ and $\widehat{r}_2\geqslant 1/2$. Furthermore, $\delta_3\neq 0$ implies that $\widehat{r}_1\neq 1$ and $\widehat{r}_2\neq 1/2$. Hence, recalling that the successor of 1 in $\mathbb{F}$ is $1+2u$, we must have $\widehat{r}_1\widehat{r}_2\geqslant(1+2u)^2/2$ and, consequently, $|\delta_3|\leqslant 2u\operatorname{ufp}(\widehat{r}_1\widehat{r}_2)/(1+2u)^2$.

Let now check that $\operatorname{ufp}(\widehat{r}_1\widehat{r}_2)\leqslant 1/2$, that is, $\widehat{r}_1\widehat{r}_2<1$ by specializing (19) to $j=0$ and $\beta=2$. If $\delta_1>0$ then $\widehat{r}_1=1+(k_1+1)\cdot 2u$ and $k_0=1$. Therefore,

$\widehat{r}_2 \leqslant 1-(2k_1+1)u$ and $\widehat{r}_1\widehat{r}_2 \leqslant 1+u-(k_1+1)(2k_1+1)\cdot 2u^2 < 1+u$ for $k_1 \geqslant 0$. It then follows from $\delta_3 > 0$ that $\widehat{r}_1\widehat{r}_2 < 1$.

If $\delta_1 < 0$ then $\widehat{r}_1 = 1+k_1\cdot 2u$. Furthermore, $x-y \leqslant 1-k_1\cdot 2u \in \mathbb{F}$, so that $\widehat{r}_2 \leqslant 1-k_1\cdot 2u$ as well. Hence $\widehat{r}_1\widehat{r}_2 \leqslant 1-k_1^2\cdot 4u^2 \leqslant 1$ and, using $\delta_3 \neq 0$, we must have $\widehat{r}_1\widehat{r}_2 < 1$. □

3.4.2. *Case where $\delta_i > 0$ for all i.* Here $x \pm y$ are rounded up, which by (19) gives

$$k_0\cdot 2u/\beta+\epsilon \geqslant u, \qquad \epsilon \leqslant u/\beta, \qquad \delta_1 = \frac{2u-(k_0\cdot 2u/\beta+\epsilon)}{x+y}, \qquad \delta_2 = \frac{\epsilon}{x-y}.$$

For β even, it follows that $k_0 \geqslant \beta/2$ and, therefore,

$$1+\theta \leqslant \left(1+\frac{u-\epsilon}{x+y}\right)\left(1+\frac{\epsilon}{x-y}\right)(1+\delta_3) =: f(x,y,\epsilon,\delta_3). \tag{20}$$

Note that f increases with ϵ, since $\partial f/\partial\epsilon = (2y+u-2\epsilon)(1+\delta_3)/(x^2-y^2)$ is positive for all $\delta_3 > -1$, $x > y > 0$, and $\epsilon \leqslant u/\beta \leqslant u/2$.

■ If $\beta \geqslant 4$, then by combining (20) with $\epsilon \leqslant u/\beta$, $x \geqslant 1$, and $\delta_3 \leqslant u_1$, we obtain

$$1+\theta \leqslant \left(1+\frac{1-\beta^{-1}}{1+y}u\right)\left(1+\frac{\beta^{-1}}{1-y}u\right)(1+u_1) =: g(y). \tag{21}$$

Now, the derivative of g has the form $g'(y) = -u(1+u_1)/\big(\beta(1-y^2)\big)^2\cdot h(y)$, where $h(y)$ is the following quadratic polynomial:

$$h(y) = h_0(1+y^2)-h_1 y, \qquad h_0 = \beta(\beta-2), \qquad h_1 = 2\beta^2+(\beta-1)\cdot 2u.$$

Applying $\beta-2 \geqslant 2$ and $y \geqslant u$ to $h_0(1+y^2)$, and $y \leqslant (\beta-2u)\beta^{-2}$ to $h_1 y$, one can check that $h(y) \geqslant (2+2\beta^{-1})u+(2\beta+4\beta^{-1}-4\beta^{-2})u^2$, which is positive for $\beta \geqslant 2$. Hence g decreases with $y \geqslant u$ and thus $1+\theta \leqslant g(u)$. Finally, it can be checked that $\beta \geqslant 4$ and $p \geqslant 2$ imply $g(u) < 1+2u$.

■ Assume now that $\beta = 2$. In this case the function $g(y)$ introduced in (21) does not suffice, as it can now be larger than $1+2u$ (namely, as large as about $1+\frac{7}{3}u$ when $y = (1-u)/2$—and thus even larger than the uniform bound $1+\frac{9}{4}u$ we target). Hence, instead of $\epsilon \leqslant u/2$, we shall apply to (20) the refined bound $\epsilon \leqslant (1-4\eta)u/2$ from Property 3.3:

$$1+\theta \leqslant \left(1+\frac{\frac{1}{2}+2\eta}{x+y}u\right)\left(1+\frac{\frac{1}{2}-2\eta}{x-y}u\right)(1+\delta_3) =: g_2(x,y,\eta,\delta_3).$$

One can check that $\partial g_2/\partial y$ has the form $(1+\delta_3)u/(x^2-y^2)^2\cdot P(x,y,\eta)$, where

$$P(x,y,\eta) = -4(x^2+y^2)\eta+2xy+(\tfrac{1}{2}-8\eta^2)uy.$$

Using $x^2 \geqslant x \geqslant 1$, $y^2 \geqslant u^2$, $y \leqslant (1-u)\cdot 2\eta < 2\eta$, and $-8\eta^2 < 0$, we deduce that

$$P(x,y,\eta) < -4(x+u^2)\eta+2x\cdot(1-u)\cdot 2\eta+\tfrac{u}{2}\cdot 2\eta = (1-4x)u\eta-4u^2\eta,$$

which is negative for $x \geqslant 1$. Hence g_2 decreases with $y \geqslant \eta$, and thus $1+\theta \leqslant g_2(x,\eta,\eta,\delta_3) =: h(x,\eta,\delta_3)$. Now, it can be checked that $\partial h/\partial\eta$ equals $-u(1+\delta_3)/(x^2-\eta^2)^2\cdot(2x-1/2)(4x+u+4xu)\eta$ and, therefore, is negative for $x \geqslant 1$. Consequently, h decreases with $\eta \geqslant u$ and thus $1+\theta \leqslant h(x,u,\delta_3)$ or, equivalently,

$$1+\theta \leqslant \left(1+\frac{\frac{1}{2}+2u}{x+u}u\right)\left(1+\frac{\frac{1}{2}-2u}{x-u}u\right)(1+\delta_3). \tag{22}$$

From $x \geqslant 1$ and $\delta_3 < u$, it then follows immediately that $1+\theta \leqslant 1+2u+O(u^2)$.

If $x \geqslant 1+2u$, then applying the general bound $\delta_3 \leqslant u/(1+u)$ to (22) suffices to conclude that $1+\theta < 1+2u$. If $x = 1$, this yields only $1+2u+\frac{1}{4}u^2$, but then the term $\frac{1}{4}u^2$ can be removed using the refined bound $\delta_3 \leqslant u/(1+2u)$ from Property 3.4.

This concludes the proof of Lemma 3.3 in the case where $\delta_i > 0$ for all i.

3.4.3. *Case where $\delta_i < 0$ for all i.* Using again $\widetilde{\theta}$ as in (15) together with the fact that $-\widetilde{\theta} < \theta < 0$, it suffices to check that $\widetilde{\theta} < 2u$.

From $\delta_1 < 0$ and $\delta_2 < 0$ we deduce that

$$k_0 \leqslant \beta/2 - 1, \qquad |\delta_1| = \frac{k_0 \cdot 2u/\beta + \epsilon}{x+y}, \qquad \epsilon \geqslant u/\beta, \qquad |\delta_2| = \frac{2u/\beta - \epsilon}{x-y}.$$

Consequently,

$$|\delta_1| \leqslant \frac{u - \widetilde{\epsilon}}{x+y}, \qquad |\delta_2| = \frac{\widetilde{\epsilon}}{x-y}, \qquad \widetilde{\epsilon} := 2u/\beta - \epsilon \in (0, u/\beta].$$

Thus, for f as in (20), we arrive at

$$1 + \widetilde{\theta} \leqslant f\big(x, y, \widetilde{\epsilon}, |\delta_3|\big).$$

For $\beta \geqslant 4$, using $x \geqslant 1$, $y \geqslant u$, $\widetilde{\epsilon} \in (0, u/\beta]$, and $|\delta_3| \leqslant u/(1+u)$, we can show as in the previous section that $1 + \widetilde{\theta} \leqslant g(y) \leqslant g(u) < 1+2u$.

For $\beta = 2$, Property 3.3 tells us that $\epsilon \geqslant (1+4\eta)u/2$, which means $\widetilde{\epsilon} \leqslant (1-4\eta)u/2$, and we obtain as before $1 + \widetilde{\theta} \leqslant g_2(x, y, \eta, |\delta_3|) \leqslant g_2(x, \eta, \eta, |\delta_3|) \leqslant g_2(x, u, u, \delta_3)$. We conclude in the same way, using the fact (given by Property 3.4) that $|\delta_3| \leqslant u/(1+2u)^2$ in the special case where $x = 1$.

4. Asymptotic optimality of the new error bounds when rounding ties to even

We conclude with two lemmas showing that the upper bounds established in the previous section (Theorem 3.1) are asymptotically optimal; this completes the proof of Theorem 1.1.

Lemma 4.1. *Assume that $\beta = 2$ and $p \geqslant 5$, and let*

$$j = \left\lceil 1/\sqrt{8u} \right\rceil, \qquad x = \frac{3}{2} + (2j+1) \cdot 2u, \qquad y = \frac{1}{2} - \frac{7}{2}u.$$

Then $x, y \in \mathbb{F}$ and $\theta = \frac{9}{4}u - \epsilon$ with $\epsilon = O(u^{3/2})$.

Proof. Since $u = 2^{-p}$, we have $x = X \cdot 2^{1-p}$ and $y = Y \cdot 2^{-p-1}$, where $X = \frac{3}{4u} + 2j + 1$ and $Y = 2^p - 7$ are integers. Using $1 \leqslant j < \frac{1}{\sqrt{8u}} + 1$ gives $0 < X < \frac{3}{4u} + \frac{1}{\sqrt{2u}} + 3$, and it can be checked that for $p \geqslant 5$ the upper bound on X is less than $1/u = 2^p$; on the other hand, $p \geqslant 5$ clearly implies $0 < Y < 2^p$, and we conclude that $x, y \in \mathbb{F}$.

To show that θ is asymptotically equivalent to $\frac{9}{4}u$, note first that because of $1 \leqslant j < 1/\sqrt{8u} + 1$ and $p \geqslant 5$ each of the three quantities

$$1 + (j-1) \cdot 2u, \qquad 1 + (2j+2) \cdot 2u, \qquad 1 + (3j+3) \cdot 2u$$

belongs to $[1, 2)$ and, therefore, to $\mathbb{F} \cap [1, 2)$.

Consequently,

$$x + y = \underbrace{2 + (j-1) \cdot 4u}_{\in \mathbb{F} \cap [2,4)} + \frac{5}{2}u, \qquad x - y = \underbrace{1 + (2j+2) \cdot 2u}_{\in \mathbb{F} \cap [1,2)} + \frac{3}{2}u,$$

and, using $\frac{5}{2}u \in (2u, 4u)$ and $\frac{3}{2}u \in (u, 2u)$, we deduce that $x + y$ and $x - y$ are rounded up to $\widehat{r}_1 = 2 + j \cdot 4u = x + y + \frac{3}{2}u$ and $\widehat{r}_2 = 1 + (2j + 3) \cdot 2u = x - y + \frac{1}{2}u$, respectively. On the other hand, $j = 1/\sqrt{8u} + O(1)$ implies $x + y = 2 + O(u^{1/2})$ and $x - y = 1 + O(u^{1/2})$, and so the associated relative errors δ_1 and δ_2 satisfy

$$\delta_1 = \frac{3}{4}u - \epsilon_1, \qquad \delta_2 = \frac{1}{2}u - \epsilon_2, \qquad \epsilon_1, \epsilon_2 = O\big(u^{3/2}\big).$$

It remains to estimate the third relative error, δ_3, that occurs when rounding the product $\widehat{r}_1\widehat{r}_2$. The expressions given above for $\widehat{r}_1$ and $\widehat{r}_2$ lead to

$$\widehat{r}_1\widehat{r}_2 = \underbrace{2 + (3j + 3) \cdot 4u}_{\in\, \mathbb{F} \cap\, [2,4)} + j(2j + 3) \cdot 8u^2.$$

Furthermore, one can check that $j(2j + 3) \cdot 8u^2$ is in $(2u, 4u)$: the lower bound follows from $j \geqslant 1/\sqrt{8u}$; for the upper bound, using $j < 1/\sqrt{8u} + 1$ leads to $j(2j+3)\cdot 8u^2 < 2u + 7\sqrt{8}u^{3/2} + 40u^2$, which is at most $4u$ for all $p \geqslant 8$; if $p \in \{5, 6, 7\}$, then the ratio $\big(j(2j + 3) \cdot 8u^2\big)/(4u)$ is in $\{7/8, 27/32, 11/16\}$ and thus less than 1, as wanted. Therefore, $\widehat{r}_1\widehat{r}_2$ is rounded up to $\widehat{r} = 2 + (3j + 4) \cdot 4u$ and, using again $j = 1/\sqrt{8u} + O(1)$, we deduce that $\widehat{r}_1\widehat{r}_2 - \widehat{r} = j(2j + 3) \cdot 8u^2 - 4u = -2u + O(u^{3/2})$ and $\widehat{r}_1\widehat{r}_2 = 2 + O(u^{1/2})$. Hence

$$\delta_3 = u - \epsilon_3, \qquad \epsilon_3 = O\big(u^{3/2}\big).$$

Using $\theta = \delta_1 + \delta_2 + \delta_3 + O(u^2)$, we obtain $\theta = \big(\frac{3}{4} + \frac{1}{2} + 1\big)u - \epsilon$ with $\epsilon = O(u^{3/2})$. □

Lemma 4.2. *Assume that $p \geqslant 4$ and let*

$$x = 1 + 2u, \qquad y = 3u - 4u^2.$$

Then $x, y \in \mathbb{F}$ and $\theta = -2u + 13u^2 + O(u^3)$.

Proof. The fact that $x \in \mathbb{F}$ is clear and, on the other hand, it is easily checked that $y \in \mathbb{F}$ for β even. Now, the exact sum has the form $x + y = (1 + 4u) + (u - 4u^2)$ with $1 + 4u \in \mathbb{F} \cap [1, \beta)$ and, for $p > 2$, $u - 4u^2 \in (0, u)$, so it must be rounded down to $\widehat{r}_1 = 1 + 4u$. Similarly, $x - y = 1 - u + 4u^2$ with $1 - u \in \mathbb{F} \cap [\beta^{-1}, 1)$ for β even and, since $p \geqslant 4$, $4u^2 \in (0, \beta^{-1}u)$; consequently, $x - y$ is rounded down to $\widehat{r}_2 = 1 - u$. Hence $\widehat{r}_1\widehat{r}_2 = (1 + 2u) + (u - 4u^2)$ and thus $\widehat{r} = 1 + 2u$. The conclusion follows from $\theta = \widehat{r}/(x^2 - y^2) - 1$ and $x^2 - y^2 = 1 + 4u - 5u^2 + 24u^3 - 16u^4$. □

References

[1] Richard Brent, Colin Percival, and Paul Zimmermann. Error bounds on complex floating-point multiplication. *Math. Comp.*, 76:1469–1481, 2007.

[2] David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991.

[3] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. xxx+680 pp. ISBN 0-89871-521-0.

[4] IEEE Computer Society. *IEEE Standard for Floating-Point Arithmetic, IEEE Standard 754-2008*. IEEE Computer Society, New York, August 2008. x+58 pp. ISBN 978-0-7381-5752-8.

[5] Claude-Pierre Jeannerod. Recent results in fine-grained rounding error analysis. In *Book of Abstracts of the 18th International Symposium on Scientific Computing, Computer Arithmetic, and Verified Numerical Computations (SCAN 2018)*, Tokyo, Japan, September 2018, page 20.

[6] Claude-Pierre Jeannerod, Peter Kornerup, Nicolas Louvet, and Jean-Michel Muller. Error bounds on complex floating-point multiplication with an FMA. *Math. Comp.*, 86:881–898, 2017.

[7] Claude-Pierre Jeannerod and Siegfried M. Rump. On relative errors of floating-point operations: optimal bounds and applications. *Math. Comp.*, 87(310):803–819, 2018.

[8] W. Kahan and J. W. Thomas. Augmenting a programming language with complex arithmetic. Technical Report UCB/CSD-92-667, EECS Department, University of California, Berkeley, 1991. Available at https://www2.eecs.berkeley.edu/Pubs/TechRpts/1992/6127.html.

[9] Donald E. Knuth. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*. Third edition, Addison-Wesley, Reading, MA, USA, 1998. xiii+762 pp. ISBN 0-201-89684-2.

[10] Marko Lange and Siegfried M. Rump. Sharp estimates for perturbation errors in summations. *Math. Comp.*, 88(315):349–368, 2019.

[11] Walter F. Mascarenhas. Floating point numbers are real numbers, May 2016. arXiv report 1605.09202. 57 pp.

[12] Siegfried M. Rump, Takeshi Ogita, and Shin'ichi Oishi. Accurate floating-point summation, Part I: Faithful rounding. *SIAM J. Sci. Comput.*, 31(1):189–224, 2008.

[13] Pat H. Sterbenz. *Floating-Point Computation*. Prentice-Hall, 1974. xiv+316 pp. ISBN 0-13-322495-3.

[14] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, New York, 1980. ix+609 pp. ISBN 0-387-90420-4.

[15] Josef Stoer. *Einführung in die Numerische Mathematik I*. Springer-Verlag, Berlin, 1972. ix+250 pp. ISBN 978-3-540-05750-5 (print), 978-3-662-06865-6 (online).

[16] J. H. Wilkinson. Error analysis of floating-point computation. *Numer. Math.*, 2:319–340, 1960.

[17] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, 1965. xviii+662 pp. ISBN 0-19-853403-5 (hardback), 0-19-853418-3 (paperback).

[18] Abraham Ziv. Fast evaluation of elementary mathematical functions with correctly rounded last bit. *ACM Trans. Math. Software*, 17(3):410–423, 1991.

Univ Lyon, Inria, CNRS, ENS de Lyon, Université Claude Bernard Lyon 1, LIP UMR 5668, F-69007 LYON, France

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