FastTwoSum revisited

Claude-Pierre Jeannerod[†], Paul Zimmermann[‡]

[†] Inria, Université de Lyon, CNRS, ENSL, UCBL, LIP, F-69342 Lyon, France
[‡] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

***Abstract*—The FastTwoSum algorithm is a classical way to evaluate the rounding error that occurs when adding two numbers in finite precision arithmetic. Starting with Dekker in the early 1970s, numerous floating-point analyses have been made of this algorithm, that are aimed at identifying sufficient conditions for the error to be computed exactly and, otherwise, at quantifying the quality of the error estimate thus produced. In this paper we revisit these two aspects of FastTwoSum. We first provide new, less restrictive conditions for exactness, and show that FastTwoSum performs an error-free transform in more general situations than those found so far in the literature. Second, when exactness cannot be guaranteed we give several error analyses of the output of FastTwoSum and show that the bounds obtained are tight. In particular, this provides further insight into how the algorithm behaves when roundings other than 'to nearest' are used, or when the operands are reversed.**

***Index Terms*—FastTwoSum algorithm, floating-point arithmetic, faithful rounding, directed rounding, rounding error analysis, IEEE 754 standard.**

I. Introduction

The FastTwoSum algorithm is a classical way to evaluate the rounding error of finite-precision additions, which in the context of floating-point arithmetic is generally traced back to Kahan [1] and Møller [2], [3]. Given a and b from a floating-point number set $\mathbb{F}$ and given some rounding $\circ$ from $\mathbb{R}$ to $\mathbb{F}$, FastTwoSum consists in a sequence of three operations producing $x, y \in \mathbb{F}$ as follows:

$$\begin{aligned} x &= \circ(a+b), \\ z &= \circ(x-a), \\ y &= \circ(b-z). \end{aligned} \tag{1}$$

Once the sum x has been computed, the next two floating-point operations aim at producing a suitable estimate y of the associated rounding error

$$e = a + b - x.$$

This scheme is at the heart of many higher-level algorithms, such as compensated algorithms and algorithms for extended-precision arithmetic (see [4] and the references therein). These algorithms are used extensively to provide extra accuracy, in particular to ensure correct rounding as in the CORE-MATH project [5]. Perhaps the most known feature of FastTwoSum is its ability to exactly recover the error e under appropriate conditions. Specifically, Dekker [6] showed that $y = e$ whenever the floating-point base is $\beta \leq 3$, x is produced by rounding to nearest (denoted RN, with any tie-breaking rule), and the exponent of a is larger than or equal to that of b. In practice, this means that FastTwoSum implements an error-free transform (EFT), that is, it transforms the pair $(a, b) \in \mathbb{F}^2$ into another pair $(x, y) \in \mathbb{F}^2$ such that $a + b = x + y$ and $x = \circ(a + b)$, as soon as

$$\beta = 2, \quad \circ \in \mathrm{RN}, \quad |a| \geq |b|.$$

Since Dekker's analysis, many extensions have been proposed, aimed especially at understanding the behavior of FastTwoSum for larger bases, other roundings, and weaker restrictions on a and b: what conditions suffice to ensure that FastTwoSum is an EFT and, otherwise, what is the quality of the error estimate y?

The goal of this paper is to reexamine these two questions by reviewing existing answers and providing several improvements. Specifically, we aim at combining roundings other than 'to nearest'—such as the standard directed roundings (down, up, to zero) and faithful rounding—with conditions on a and b weaker than the ones above (such as $a \in \mathrm{ulp}(b)\mathbb{Z}$ that is, a is an integral multiple of the unit in the last place of b). In this study we shall focus on binary arithmetic with unbounded exponent range. Our contributions can be summarized as follows.

After some preliminaries in Section II, we begin in Section III by providing new, less restrictive sets of sufficient conditions for the error $e = a + b - x$ to be a floating-point number and for the last two operations in (1) to be exact.

Then in Section IV we show how to combine these sufficient conditions to deduce five EFTs depending on the rounding mode used for the first operation $x = \circ(a+b)$. We also prove via carefully chosen examples that when these conditions are not satisfied, then $x + y$ can differ from $a + b$.

Section V is devoted to the accuracy analysis of $x+y$ in the absence of EFT. We provide provably-tight error bounds in two different situations: first, for faithful rounding and $a \in \mathrm{ulp}(b)\mathbb{Z}$ we show how to bound $|x + y - (a + b)|$ in an asymptotically optimal way; second, for each of the three standard directed roundings and $|a| < |b|$ (reversed operands), we establish optimal bounds on $|x + y - (a + b)|/|x|$.

In Section VI, we study directed roundings in more detail when $a \in \mathrm{ulp}(b)\mathbb{Z}$. We show that when FastTwoSum is run with a directed rounding $\circ$, then $x + y$ is either the exact sum $a + b$ or its correctly-rounded value in doubled precision and in the same direction as $\circ$. This provides in particular further insight into the nature of the interval obtained by running FastTwoSum with both rounding down and up on the same input.

2576-2265/25/$31.00 ©2025 IEEE
DOI 10.1109/ARITH64983.2025.00030

We conclude in Section VII by showing that this result for rounding down and up does not hold anymore for the variant of (1) that is sometimes seen in the literature and that consists in computing $y' = \circ(z - b)$ and approximating $a + b$ by $x - y'$.

II. Preliminaries

A. Notation and definitions

For $\beta, p \geq 2$, we denote by $\mathbb{F}$ the set of floating-point numbers in base β and precision p, with unbounded exponent range: $\mathbb{F} = \{M \cdot \beta^E : M, E \in \mathbb{Z}, |M| < \beta^p\}$. Rounding from $\mathbb{R}$ to $\mathbb{F}$ is denoted by $\circ$, and in case we use different roundings in (1), we denote them by $\circ$ for rounding $a + b$, $\circ'$ for rounding $x - a$, and $\circ''$ for rounding $b - z$. If on the contrary the same rounding $\circ$ is used for these three operations, then we may write $[x, y] = \text{FastTwoSum}(a, b, \circ)$. We will also write $\circ_{2p}$ to indicate rounding to doubled precision.

The five roundings specified in the IEEE 754 standard [7] are written RNE, RNA, RD, RU, RZ for, respectively, rounding to nearest with ties to even, rounding to nearest with ties to away, rounding down (that is, towards $-\infty$), rounding up (that is, towards $+\infty$), and rounding to zero (or truncating). For rounding to nearest, when no assumption is made on the tie-breaking rule, we simply write RN. All these roundings, be they to nearest or directed, yield faithfully rounded values, that is, they satisfy the following property: for $r \in \mathbb{R}$, $\circ(r) \in \{\text{RD}(r), \text{RU}(r)\}$, which implies $\circ(r) = r$ when $r \in \mathbb{F}$. For simplicity we shall write $\circ \in$ FR to mean that some faithful rounding is used. In this paper, all the roundings used are at least faithful.

For all our analyses we will assume that $\beta = 2$ and rely on on the unit roundoff $u = 2^{-p}$ as well as on the usual notions of exponent, unit in the first place, and unit in the last place: for $r \in \mathbb{R} \backslash \{0\}$,

$$e_r = \lfloor \log_2 |r| \rfloor, \quad \text{ufp}(r) = 2^{e_r}, \quad \text{ulp}(r) = 2u \, \text{ufp}(r),$$

and $e_0 = -\infty$, $\text{ufp}(0) = \text{ulp}(0) = 0$.

B. Basic properties

Our proofs will exploit a few basic properties that we briefly review here (but more can be found for example in [8]).

If r is a nonzero real number, then $\text{ufp}(r) \leq |r| < 2\text{ufp}(r)$.

If f is in $\mathbb{F}$ then $\text{ufp}(f) \leq |f| \leq (2 - 2u)\text{ufp}(f)$ and, more precisely, $|f| = (1 + k \cdot 2u)\text{ufp}(f)$ for some integer k such that $0 \leq k < 1/(2u)$. Since $\text{ulp}(f) = 2u \, \text{ufp}(f)$, we see that f is an integral multiple of its ulp, which we write $f \in \text{ulp}(f)\mathbb{Z}$.

Conversely, if for $r \in \mathbb{R}$, $r \in \text{ulp}(f)\mathbb{Z}$ for some $f \in \mathbb{F}$, and $|r| \leq 2\text{ufp}(f)$, then $r \in \mathbb{F}$.

Recall also that for $r \in \mathbb{R}$, $|\circ(r) - r| \leq u \, \text{ufp}(r)$ if $\circ \in$ RN, and $|\circ(r) - r| \leq 2u \, \text{ufp}(r)$ if $\circ \in$ FR, the latter inequality being strict whenever r is nonzero.

If the same rounding $\circ$ is used for the three operations in FastTwoSum, then FastTwoSum may inherit some of its properties. Specifically, if $\circ$ satisfies $\circ(-r) = -\circ(r)$ for all $r \in \mathbb{R}$ (anti-symmetry), then

$$\text{FastTwoSum}(-a, -b, \circ) = -\text{FastTwoSum}(a, b, \circ)$$

for all $a, b \in \mathbb{F}$. This is the case when $\circ \in \{\text{RNE}, \text{RNA}, \text{RZ}\}$. When $\circ \in \{\text{RD}, \text{RU}\}$, the anti-symmetry property is lost and we have instead $\text{RD}(-r) = -\text{RU}(r)$ for any $r \in \mathbb{R}$, from which we deduce that

$$\text{FastTwoSum}(-a, -b, \text{RD}) = -\text{FastTwoSum}(a, b, \text{RU}).$$

C. Exactness properties

Recall first that by Sterbenz' theorem [9], if $a, b \in \mathbb{F}$ satisfy $b/2 \leq a \leq 2b$ then their difference $a - b$ is in $\mathbb{F}$, so that $\circ(a - b) = a - b$ is exact whenever $\circ$ is faithful.

In a similar way, when analyzing FastTwoSum we will identify in Section III some conditions on a and b that are sufficient to ensure one of the following three properties:

$$e \in \mathbb{F} \qquad \text{(P)}$$
$$z = x - a \qquad \text{(P')}$$
$$y = b - z. \qquad \text{(P")}$$

Obviously, if (P') holds, then $y = \circ(a + b - x) = \circ(e)$ is the correctly-rounded value of the exact error.

If in addition (P) holds, then $y = \circ(e) = e = a + b - x = b - z$, that is, (P") holds as well. Conversely, if (P") holds together with (P'), then $y = b - z = a + b - x = e$ and since $y \in \mathbb{F}$, we have $e \in \mathbb{F}$ as well, so that (P) holds.

We have thus shown that if (P') holds then (P) and (P") are equivalent. Expressed another way,

$$\text{(P) and (P')} \quad \Rightarrow \quad \text{(P")}$$
$$\text{(P') and (P")} \quad \Rightarrow \quad \text{(P)}.$$

Furthermore, we have seen that $y = a + b - x$ in each of these two cases, which means that FastTwoSum is then an EFT. Thus, after having studied these three properties independently in Section III, it will suffice to combine (P') with either of the other two to arrive at the EFTs described in Section IV.

We conclude these preliminaries by noting that it is possible to have (P) and (P") without having (P'). This can occur in particular when $|a| < |b|$.

Example 1 ((P) and (P") without (P'))**.** *Let $a = u$, $b = 1$ and $x = \circ(a + b)$ with $\circ \in \{\text{RNA}, \text{RU}\}$. Then $x = 1 + 2u$, $e = -u \in \mathbb{F}$ (that is, (P) is true), $x - a = 1 + u \notin \mathbb{F}$ (that is, (P') is false) and so, depending on the rounding mode used for subtracting a from x, we have $z \in \{1, 1 + 2u\}$, $b - z \in \{0, -2u\} \subset \mathbb{F}$ (that is, (P") is true). In this example, y is either 0 or $-2u$, which provides a very poor approximation to $e = -u$, the relative error $|y - e|/|e|$ being 1 in both cases.*

III. Sufficient conditions for exactness

A. Sufficient conditions to ensure (P)

We start by providing sufficient conditions for the rounding error $e = a + b - \circ(a + b)$ to be exactly representable in $\mathbb{F}$. These conditions are given for rounding to nearest with any tie-breaking rule, as well as for faithful rounding and its most common specializations RD, RU, RZ.

Lemma 1. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ with $a \in \mathrm{ulp}(b)\mathbb{Z}$ and $x = \circ(a + b)$. If one of the conditions*

(i) $\circ \in \mathrm{RN}$,
(ii) $\circ \in \mathrm{FR}$ *and* $e_a - e_b \leq p$,
(iii) $\circ = \mathrm{RD}$ *and* $b \geq 0$,
(iv) $\circ = \mathrm{RU}$ *and* $b \leq 0$,
(v) $\circ = \mathrm{RZ}$ *and* $ab \geq 0$

is satisfied, then (P) holds: the error $e = a + b - x$ is in $\mathbb{F}$.

Before giving a detailed proof we note that the value p in condition (ii) cannot in general be replaced by $p+1$. Similarly, when $b < 0$ with RD, or $b > 0$ with RU, or $ab < 0$ with RZ, it can happen that $e \notin \mathbb{F}$. The following example provides some values of a and b showing this.

Example 2. *Consider first $a = 1 + 2u$ and $b = -u/2 - u^2$. Then $a, b \in \mathbb{F}$, $e_a - e_b = p + 1$, and for $\circ = \mathrm{RZ} \in \mathrm{FR}$ we have $x = \circ(a + b) = 1$ and $e = 3u/2 - u^2 \notin \mathbb{F}$. For* RD *and* RZ*, if we take $a = 1$ and $b = -u^3$, then we have $b < 0$, $ab < 0$, $x = \mathrm{RD}(1 - u^3) = \mathrm{RZ}(1 - u^3) = 1 - u$, which gives $e = u - u^3 \notin \mathbb{F}$. For* RU*, taking $a = 1 - u$ and $b = u^3$ yields $x = \mathrm{RU}(1 - u + u^3) = 1$ and $e = -u + u^3 \notin \mathbb{F}$.*

Proof. If a or b is zero then $e = 0 \in \mathbb{F}$, so we are left with $|a|, |b| > 0$. Since $a, b \in \mathrm{ulp}(b)\mathbb{Z}$, $a + b, x, e \in \mathrm{ulp}(b)\mathbb{Z}$ and so, in particular, $e = M\mathrm{ulp}(b)$ for some $M \in \mathbb{Z}$.

If $\circ \in \mathrm{RN}$ then $|e| = \min_{f \in \mathbb{F}} |a + b - f| \leq |b|$ since $a \in \mathbb{F}$. Hence $|e| = |M|\mathrm{ulp}(b) \leq |b|$, which leads to $|M| \leq |b|/\mathrm{ulp}(b) < 2^p$ and thus $e \in \mathbb{F}$.

If $\circ \in \mathrm{FR}$ then $|e| \leq 2u\,\mathrm{ufp}(a + b)$ and thus $|M| \leq 2u\,\mathrm{ufp}(a + b)/\mathrm{ulp}(b) = \mathrm{ufp}(a + b)/\mathrm{ufp}(b)$. Now, $\mathrm{ufp}(a + b) \leq |a + b| \leq |a| + |b| \leq (2 - 2u)(\mathrm{ufp}(a) + \mathrm{ufp}(b))$. Since $e_a - e_b \leq p$ implies $\mathrm{ufp}(a) \leq 2^p\mathrm{ufp}(b)$, we obtain $\mathrm{ufp}(a+b) \leq (2-2u)(2^p+1)\mathrm{ufp}(b) < 2^{p+1}\mathrm{ufp}(b)$. The latter inequality being strict, it yields $\mathrm{ufp}(a+b) \leq 2^p\mathrm{ufp}(b)$. Hence $|M| \leq 2^p$ and thus $e \in \mathbb{F}$.

If $\circ = \mathrm{RD}$ and $b > 0$, then $a \leq \mathrm{RD}(a + b) \leq a + b$, so that $e \in [0, b]$; for $\circ = \mathrm{RU}$ and $b < 0$, we have $a + b \leq \mathrm{RU}(a+b) \leq a$, so that $e \in [b, 0]$. Hence in both cases $|e| \leq |b|$ and we conclude as for RN that $e \in \mathbb{F}$.

If $\circ = \mathrm{RZ}$ and $ab > 0$, either $a, b > 0$ and we are in the RD case, or $a, b < 0$ and we are in the RU case. □

Note that condition (i) follows from Dekker's analysis [6]. A condition similar to condition (ii) appears in [10], for any base and with one specific example in base 10 and precision 3 that shows that p cannot in general be replaced by $p + 1$; for base 2, the exponent difference is restricted to at most $p - 3$ and $p - 1$ in [11] and [12], respectively. Condition (v) appears in [13], for any base. To the best of our knowledge, conditions (iii) and (iv) are new.

Note also that with an ADD3 operation, which for $a, b, c \in \mathbb{F}$ returns $\circ(a + b + c)$, then for any of the situations described in Lemma 1 the exact error of floating-point addition could be obtained in only two steps, instead of three with FastTwoSum: compute $x = \circ(a + b)$ and then $e = \mathrm{ADD3}(a, b, -x)$. (For more about ADD3 we refer to [14], [15].) This would be further reduced to just one step if an augmentedAddition operation as the one specified in [16], [7] were available, since then condition (i) in Lemma 1 is satisfied (in augmentedAddition, $\circ \in \mathrm{RN}$ with ties rounded towards zero).

B. Sufficient conditions to ensure (P')

We now study sufficient conditions to ensure that $x - a$ is exactly representable as a floating-point number.

Lemma 2. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ and $x = \circ(a+b)$. If $\circ \in \mathrm{FR}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$ then (P') holds, that is, the difference $x - a$ is in $\mathbb{F}$.*

Proof. If $a + b \in \mathbb{F}$ then $x - a = b \in \mathbb{F}$, so we assume that $a + b \notin \mathbb{F}$, which implies in particular $b \neq 0$ and $a + b \neq 0$. Since $a, b \in \mathrm{ulp}(b)\mathbb{Z}$, the same holds for $a+b$, x, and $x-a$. On the other hand, $x \in \mathrm{ulp}(x)\mathbb{Z}$ and $a \in \mathrm{ulp}(a)\mathbb{Z}$, so $x - a \in \mu\mathbb{Z}$ with $\mu := \min\{\mathrm{ulp}(x), \mathrm{ulp}(a)\}$. Hence $x - a \in \nu\mathbb{Z}$ with $\nu := \max\{\mu, \mathrm{ulp}(b)\}$. Consequently, there exists $M \in \mathbb{Z}$ such that $x-a = M\nu$, and since $|x-a| \leq |\circ(a + b)-(a+b)|+|b| < \mathrm{ulp}(x) + |b|$, we arrive at

$$|M| < \mathrm{ulp}(x)/\nu + |b|/\nu.$$

If $\mathrm{ulp}(x) \leq \mathrm{ulp}(a)$ then $\mu = \mathrm{ulp}(x)$, from which it follows that $\nu = \max\{\mathrm{ulp}(x), \mathrm{ulp}(b)\}$. Hence $|M| < 1+|b|/\mathrm{ulp}(b) \leq 1 + (2^p - 1) = 2^p$, and so $x - a \in \mathbb{F}$.

Assume now that $\mathrm{ulp}(x) > \mathrm{ulp}(a)$. In this case $\mu = \mathrm{ulp}(a)$ and, therefore,

$$\nu = \max\{\mathrm{ulp}(a), \mathrm{ulp}(b)\}.$$

We have $|a + b| \leq 2\max\{|a|, |b|\} \in \mathbb{F}$, so $|x| = |\circ(a + b)| \leq 2\max\{|a|, |b|\}$ and, by monotonicity,

$$\mathrm{ulp}(x) \leq 2\max\{\mathrm{ulp}(a), \mathrm{ulp}(b)\} = 2\nu.$$

Using this bound together with $\nu \geq \mathrm{ulp}(b)$, we deduce that $|M| < 2+|b|/\mathrm{ulp}(b) \leq 2+(2^p-1) = 2^p+1$. Since M is an integer, it follows that $|M| \leq 2^p$ and, therefore, $x-a \in \mathbb{F}$. □

Note that this property appears in various places, but with stronger assumptions about either $\circ$ or a and b. For example, $e_a \geq e_b$ is assumed in [10] (for any base) and in [12] (for base 2); in [6] it is assumed that $\circ \in \mathrm{RN}$ and $e_a \geq e_b$ (and that the base is at most 3); in [13] it is assumed that $\circ \in \{\mathrm{RNA}, \mathrm{RZ}\}$ and $|a| \geq |b|$ (in any base), while [8] assumes $\circ = \mathrm{RNE}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$, and [11] assumes $\circ \in \{\mathrm{RD}, \mathrm{RU}, \mathrm{RZ}\}$ and $|a| \geq |b|$. For $\circ \in \mathrm{RN}$, a similar property also appears in [17].

C. Sufficient conditions to ensure (P")

We conclude this section by proving a sufficient condition for (P") to hold, that is, for the exact difference $b - z$ to be a floating-point number. Of course, (P") follows from (P) and (P'), but it turns out that (P") can also be satisfied independently of these two properties, as we shall now see.

Lemma 3. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$, $x = \circ(a + b)$, and $z = \circ'(x - a)$. If $\circ, \circ' \in \mathrm{FR}$ and $e_a - e_b \leq p$ then (P") holds, that is, the difference $b - z$ is in $\mathbb{F}$.*

Proof. The result is clear if $b = 0$, so assume $b \neq 0$ and consider the two cases $e_b < e_a$ and $e_b \geq e_a$ separately.

We consider first the case where $e_b < e_a$. This implies $a \in \mathrm{ulp}(b)\mathbb{Z}$, and since $e_a - e_b \leq p$ by assumption, condition (ii) in Lemma 1 is satisfied, which leads to $e = a + b - x \in \mathbb{F}$. Furthermore Lemma 2 applies, thus $x - a \in \mathbb{F}$. Hence $z = x - a$ for any $\circ' \in \mathrm{FR}$ and thus $b - z = b - (x - a) = e \in \mathbb{F}$.

Let us now consider the case where $e_b \geq e_a$. Assume $b > 0$ (the case $b < 0$ is similar) and $x \neq 0$ (for otherwise $a + b = 0$ and then $b - z = 0 \in \mathbb{F}$, as wanted). We have $|e| < \mathrm{ulp}(x) \leq 2\mathrm{ulp}(b)$, thus $f := b - 2\mathrm{ulp}(b) < b - e < b + 2\mathrm{ulp}(b) =: g$. The left bound f is in $\mathbb{F}$, but g might not, for example with $b = 1 - u$. Assume first $e < 0$. Then $z = \circ'(b - e) \in [b, g']$, where $g' = \mathrm{RU}(g) \leq b + 4\mathrm{ulp}(b)$, $b - z \in \mathrm{ulp}(b)\mathbb{Z}$ with $|b - z| \leq 4\mathrm{ulp}(b)$, thus $b - z \in \mathbb{F}$. Assume now $e \geq 0$. Then $z = \circ'(b - e) \in [f, b]$ and we consider two subcases as follows.

First, if $p \geq 3$ or $b = \frac{3}{2}\mathrm{ufp}(b)$ with $p = 2$, then it can be checked that $\mathrm{ulp}(f) \geq \frac{1}{2}\mathrm{ulp}(b)$. Hence $\mathrm{ulp}(z) \geq \frac{1}{2}\mathrm{ulp}(b)$ and thus $b - z \in \frac{1}{2}\mathrm{ulp}(b)\mathbb{Z}$. Since $|b - z| = b - z \leq b - f = 2\mathrm{ulp}(b)$, this implies $b - z \in \mathbb{F}$.

Second, let $p = 2$ and $b = \mathrm{ufp}(b)$. Since $0 \leq z \leq b$, $|b - z| \leq b$. Since $b - z \in \mathrm{ulp}(a)\mathbb{Z}$, if $b \leq 2\mathrm{ufp}(a)$ then $b - z \in \mathbb{F}$; otherwise, $b = \mathrm{ufp}(b) \geq 4\mathrm{ufp}(a)$ and then $|e| \leq b/2$ and $b/2 \leq z \leq 3b/2 \leq 2b$, which by Sterbenz' theorem gives $b - z \in \mathbb{F}$. □

In Lemma 3 the upper bound p improves over the bound $p - 3$ from [11] and, as the next example shows, it cannot be replaced by $p + 1$ even for $\circ \in \{\mathrm{RD}, \mathrm{RU}, \mathrm{RZ}\}$.

Example 3. *For $\circ \in \{\mathrm{RD}, \mathrm{RZ}\}$, if we consider $a = 1 + 2u$ and $b = -u/2 - u^2$ as in Example 2, we have $e_a - e_b = p + 1$, $x = 1$, $x - a = -2u = z$ for any $\circ' \in \mathrm{FR}$, and $b - z = 3u/2 - u^2 \notin \mathbb{F}$. For $\circ = \mathrm{RU}$, taking $a = 1$ and $b = u/2 + u^2$ similarly yields $e_a - e_b = p + 1$ and $b - z = -3u/2 + u^2 \notin \mathbb{F}$.*

Note also that (P”) can hold when neither (P) nor (P’) does.

Example 4. *Let $a = u^3$, $b = 1$, $\circ = \mathrm{RU}$, and $\circ' = \mathrm{RD}$. Then the conditions of Lemma 3 are satisfied and thus we know that $b - z \in \mathbb{F}$, that is, (P”) is true. However, $x = \mathrm{RU}(1 + u^3) = 1 + 2u$, $a + b - x = -2u + u^3 \notin \mathbb{F}$ (that is, (P) is false), and $z = \mathrm{RD}(1 + 2u - u^3) = 1$ (that is, (P’) is false).*

Finally, we note that Lemma 3 does not involve any lower bound on $e_a - e_b$, so $|b|$ can be arbitrarily larger than $|a|$. In particular, this means that (P”) holds when $|b| > |a|$.

IV. Error-free transforms

By combining the sufficient conditions established in the previous section we can now deduce very easily five EFTs depending on the rounding mode used for performing the operation $x = \circ(a + b)$ in (1). For each of the other two operations, any rounding is allowed as long as it is faithful, which we denote by $z = \circ'(x - a)$ and $y = \circ''(b - z)$.

A. EFT for rounding to nearest

Theorem 1. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If the conditions*

- $\circ \in \mathrm{RN}$,
- $\circ', \circ'' \in \mathrm{FR}$,
- $a \in \mathrm{ulp}(b)\mathbb{Z}$

are all satisfied then $x + y = a + b$.

Proof. Since $\circ \in \mathrm{RN}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$, Lemma 1 implies (P), that is, $a + b - x \in \mathbb{F}$, and Lemma 2 implies (P’), that is, $z = x - a$. Hence $y = \circ''(b - z) = \circ''(a + b - x) = a + b - x$. □

This result slightly extends the ones from [6] and [8], since those assume $e_a \geq e_b$ (in base at most 3) and $\circ = \circ' = \circ'' = \mathrm{RNE}$, respectively. Note also that $x + y$ can differ from $a + b$ when one of the conditions $\circ \in \mathrm{RN}$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$ is not satisfied.

Example 5 (When $\circ \notin \mathrm{RN}$ or $a \notin \mathrm{ulp}(b)\mathbb{Z}$). *If $\circ = \mathrm{RU}$, $a = 2 - 2u$, and $b = u - u^2$, then $a, b \in \mathbb{F}$ with $a \in \mathrm{ulp}(b)\mathbb{Z}$, but $x = \mathrm{RU}(a + b) = 2$ and $a + b - x = -u - u^2 \notin \mathbb{F}$. (This example is in fact a generalization to any precision of the example given by Dekker for $p = 4$ in [6, p. 229].)*

If $a = -u$ and $b = 1 + 2u$, then $\mathrm{ulp}(b) = 2u$ and so $a \notin \mathrm{ulp}(b)\mathbb{Z}$. It can then be checked that for any $\circ \in \mathrm{RN}$ (that is, whatever the tie-breaking rule), $x + y$ is either 1 or $1 + 2u$, and thus differs from $a + b = 1 + u$.

B. EFT for faithful rounding

Theorem 2. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If the conditions*

- $\circ, \circ', \circ'' \in \mathrm{FR}$,
- $a \in \mathrm{ulp}(b)\mathbb{Z}$,
- $e_a - e_b \leq p$

are all satisfied then $x + y = a + b$.

Proof. Since $a \in \mathrm{ulp}(b)\mathbb{Z}$, Lemma 2 implies (P’), and since $e_a - e_b \leq p$, Lemma 3 implies (P”). Hence $z = \circ'(x - a) = x - a$ and $y = \circ''(b - z) = b - z = a + b - x$. □

A similar result appears in [10] for any base, but with our assumption $a \in \mathrm{ulp}(b)\mathbb{Z}$ replaced by the stronger constraint $e_a \geq e_b$. Furthermore, Examples 2 and 5 can be reused to show that if one of the conditions $e_a - e_b \leq p$ and $a \in \mathrm{ulp}(b)\mathbb{Z}$ is not satisfied, then $x + y$ can differ from $a + b$.

C. EFTs for directed roundings

Theorem 3. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If the conditions*

- $\circ = \mathrm{RD}$,
- $\circ', \circ'' \in \mathrm{FR}$,
- $a \in \mathrm{ulp}(b)\mathbb{Z}$,
- $b \geq 0$ *or* $e_a - e_b \leq p$

are all satisfied then $x + y = a + b$.

Proof. Since $a \in \mathrm{ulp}(b)\mathbb{Z}$, Lemma 2 implies (P’). If $b \geq 0$ then, since $\circ = \mathrm{RD}$, Lemma 1 implies (P); otherwise $\mathrm{RD} \in \mathrm{FR}$ and $e_a - e_b \leq p$ imply (P”) by Lemma 3. Hence we have (P’) with either (P) or (P”), and we conclude as in the proofs of Theorems 1 and 2. □

We can show in the same way the following companion result for rounding up.

Theorem 4. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If the conditions*

- $\circ =$ RU*,*
- $\circ', \circ'' \in$ FR*,*
- $a \in \mathrm{ulp}(b)\mathbb{Z}$*,*
- $b \leq 0$ *or* $e_a - e_b \leq p$

are all satisfied then $x + y = a + b$.

Finally, for rounding to zero we obtain the following result.

Theorem 5. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If the conditions*

- $\circ =$ RZ*,*
- $\circ', \circ'' \in$ FR*,*
- $a \in \mathrm{ulp}(b)\mathbb{Z}$*,*
- $ab \geq 0$ *or* $e_a - e_b \leq p$

are all satisfied then $x + y = a + b$.

Proof. Since $a \in \mathrm{ulp}(b)\mathbb{Z}$, Lemma 2 implies (P'). If $ab \geq 0$ then Lemma 1 implies (P), otherwise RZ $\in$ FR and $e_a - e_b \leq p$ imply (P") by Lemma 3. We conclude as for Theorem 3. □

This EFT for RZ appears in [13], where it is given for any base but assuming $|a| \geq |b|$ instead of just $a \in \mathrm{ulp}(b)\mathbb{Z}$. To the best of our knowledge, the EFTs for RD and RU given above are new. Furthermore, we show in the next example that $x+y$ can differ from $a+b$ when one of the last two conditions listed in each of Theorems 3, 4, 5 is not satisfied.

Example 6. *For* RD*, the input $(a, b) = (1, -u^3)$ from Example 2 is such that the condition $a \in \mathrm{ulp}(b)\mathbb{Z}$ is satisfied while the condition ($b \geq 0$ or $e_a - e_b \leq p$) is not, and we have seen there that $a + b - x \notin \mathbb{F}$. Conversely, if $(a, b) = (-u^3, 1)$ then $a \notin \mathrm{ulp}(b)\mathbb{Z}$ while the second condition is satisfied, and again $a + b - x \notin \mathbb{F}$.*

Since for these inputs, $a + b$ is positive, they can be used for RZ *as well.*

For RU*, using the fact that* $\mathrm{RU}(-r) = -\mathrm{RD}(r)$ *for any $r \in \mathbb{R}$, we obtain similar examples with $(a, b) = (-1, u^3)$ and $(a, b) = (u^3, -1)$.*

V. Tight error bounds

We have seen in the previous section several EFTs as well as examples where the conditions for such EFTs are not met and, possibly, $x + y \neq a + b$. In such situations $x + y$ only approximates the exact sum $a+b$ and a natural question is thus 'how far can $x + y$ be from $a + b$?' This section answers this question by providing a variety of bounds on $|x + y - (a + b)|$ and establishing their tightness. For readability, we let

$$\Delta := x + y - (a + b)$$

and thus provide bounds on $|\Delta|$.

A. Tight error bounds when $a \in \mathrm{ulp}(b)\mathbb{Z}$

We show in Theorem 6 below how to exploit Lemma 2 in order to obtain asymptotically optimal bounds on $|\Delta|$ for faithful rounding and when $a \in \mathrm{ulp}(b)\mathbb{Z}$. The bound $2u^2\mathrm{ufp}(a + b)$ is already given in [10] for faithful rounding (and in any base) but under the stronger hypothesis $e_a \geq e_b$. Here we relax this requirement and show furthermore that this bound is asymptotically optimal as $u \to 0$, and this even when restricting faithful rounding to RD, RU, or RZ.

Several other bounds can also be found in the literature. A bound of the form $2u^2|x|$ has been established in [18] for faithful rounding, when $|a| \geq |b|$, and allowing gradual underflow; furthermore, its tightness is shown for rounding up and when $|x|$ is close to $\mathrm{ufp}(x)$, for otherwise $2u^2|x|$ can be up to about twice larger than $2u^2\mathrm{ufp}(x)$. In [11], a bound of the form $|\Delta| < 2u\,\mathrm{ulp}(x)$ is given for directed roundings and assuming $|a| \geq |b|$, which implies $|\Delta| < 4u^2\mathrm{ufp}(x)$.[1] Similar bounds also appear in [19], [20], [21], such as $|\Delta| \leq 4u^2|a+b|$ and $|\Delta| \leq 4u^2|x|$, assuming $|a| \geq |b|$ as well. All these bounds with a term of the form $4u^2$ are about twice larger than the ones we give in the theorem below.

Theorem 6. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$. If $\circ \in$ FR and $a \in \mathrm{ulp}(b)\mathbb{Z}$, then $\Delta := x + y - (a + b)$ satisfies*

$$|\Delta| \leq 2u^2\mathrm{ufp}(a + b) \leq 2u^2\mathrm{ufp}(x).$$

Proof. If $x + y = a + b$, $\Delta = 0$, and the bounds trivially hold, since $\mathrm{ufp}(0) = 0$. Assume now $x + y \neq a + b$. It follows that $a + b \neq 0$, otherwise $x = 0$, $z = -a$, $y = 0$, and thus $x + y = a + b$. Since $a \in \mathrm{ulp}(b)\mathbb{Z}$, Lemma 2 implies $x - a \in \mathbb{F}$. Hence $b - z = a + b - x = e$ and $y = \circ(e)$. Furthermore, $e \notin \mathbb{F}$ and so $e \neq 0$, otherwise $y = e = a+b-x$. Using $a + b \neq 0$, we deduce that $|e| < \mathrm{ulp}(a + b) = 2^k$, $k \in \mathbb{Z}$. Hence $\mathrm{ufp}(e) \leq 2^{k-1}$, which together with $e \neq 0$ gives $|\Delta| = |\circ(e)-e| < \mathrm{ulp}(e) \leq 2u{\cdot}2^{k-1} = u{\cdot}2^k = 2u^2\mathrm{ufp}(a+b)$.

The second bound follows immediately from the fact that $x = \circ(a + b)$ implies $\mathrm{ufp}(a + b) \leq \mathrm{ufp}(x)$. □

The bounds from Theorem 6 are asymptotically optimal as $u \to 0$, and this is true already in the special cases where $\circ \in \{\mathrm{RD}, \mathrm{RU}, \mathrm{RZ}\}$. This is shown in the following example.

Example 7 (Asymptotic optimality of the bounds)**.** *For* RD*, let $a = 1 + 2u$ and $b = -u^3$. Then $x = \mathrm{RD}(a + b) = 1$, $z = \mathrm{RD}(-2u) = -2u$, $y = \mathrm{RD}(2u - u^3) = 2u(1 - u)$, so that $a + b = 1 + 2u - u^3$ and $x + y = 1 + 2u - 2u^2$, which gives $\mathrm{ufp}(a+b) = \mathrm{ufp}(x+y) = 1$ and $|\Delta| = 2u^2 - u^3 \sim 2u^2$ as $u \to 0$. The same example can be used for* RZ*, since $a + b$ and $b - z$ are positive and $x - a$ is a floating-point number. For* RU*, it suffices to negate a and b and to exploit the fact that* $\mathrm{RU}(-r) = -\mathrm{RD}(r)$ *for any $r \in \mathbb{R}$.*

B. Tight error bounds when $|a| < |b|$

We now turn to the case of reversed operands, that is, when $|a| < |b|$ instead of the most common assumption $|a| \geq |b|$. Linnainmaa analyzes this case in [13, Theorem 5] and gives the bounds $|\Delta| \leq \frac{1}{2}\mathrm{ulp}(x)$ for RNA, and $|\Delta| < 2\mathrm{ulp}(x)$ for RZ and rounding to odd, in any base. These bounds are for the variant of FastTwoSum where $y' = \circ(z - b)$. However, since these rounding modes are anti-symmetric, they also apply to our variant in (1). Another analysis is done in [18] for $\beta = 2$ and $p \geq 2$, when neither underflow nor overflow occurs, with

[1] The authors of [11] use $u = 2^{1-p}$ when handling directed roundings, so we have adapted their bounds to our notation, where u always means 2^{-p}.

a bound $|\Delta| \leq u|x|$ for rounding to nearest, and $|\Delta| < 3u|x|$ for any rounding mode.

In this section, we slightly refine the analysis from [18] for the directed roundings RD, RU, RZ, and show that the resulting bounds are attained for some values of a and b.

Theorem 7. *Let $a, b \in \mathbb{F}$ and $[x, y] = \text{FastTwoSum}(a, b, \text{RZ})$ with $\beta = 2$ and $p \geq 5$. If $|a| < |b|$ then*

$$|\Delta| \leq \frac{3u}{1+4u}|x|.$$

Proof. We assume without loss of generality $0 < |a| < 1 \leq b < 2$ (since RZ is anti-symmetric, we can assume $b > 0$), and we denote $x = a + b - e$, and $z = b - e + \delta$. First, by Lemma 3, $y = b - z$, thus the only extra rounding error is due to the second operation $z = \circ(x - a) = \circ(b - e)$, and we have $|\Delta| = |\delta|$. We distinguish 4 cases: $a + b \geq 2$, $1 \leq a + b < 2$, $1/2 < a + b < 1$, and $a + b \leq 1/2$.

Case $a+b \geq 2$. Since $a+b < 3$, we have $|e| < \text{ulp}(2) = 4u$, $0 < b-e < 4$, thus $|\delta| < \text{ulp}(2) = 4u$. Since $a+b \geq 2$ implies $x \geq 2$, this yields $|\delta|/x \leq 2u \leq 3u/(1+4u)$ as soon as $p \geq 3$.

Case $1 \leq a + b < 2$. Since $a + b < 2$, we have $|e| < \text{ulp}(1) = 2u$, and since $b < 2$ implies $b \leq 2 - 2u$, it follows $0 < b - e \leq 2$, thus $|\delta| < \text{ulp}(1) = 2u$. Since $a + b \geq 1$ implies $x \geq 1$, this yields $|\delta|/x \leq 2u$ as above.

Case $1/2 < a+b < 1$. If $|a| \geq 1/2$, then $x \in \text{ulp}(a)\mathbb{Z}$, with $x \leq 2\text{ufp}(a)$, thus x is exact, $z = b$, and $\delta = 0$. If $|a| < 1/2$, we have $|e| < u$, $0 < b - e < 2$, thus $|\delta| < \text{ulp}(1) = 2u$. If $a \in [-1/4, 1/2)$ then $a + b \geq 3/4$, $x \geq 3/4$, thus $|\delta|/x < 8u/3 \leq 3u/(1 + 4u)$ as soon as $p \geq 5$. It remains to deal with the case $a \in (-1/2, -1/4)$. Then $\text{ulp}(a) = u/2$, and since $\delta \in \text{ulp}(a)\mathbb{Z}$, we have $|\delta| \leq 3u/2$. If $b \geq 1 + 2u$, then $a + b \geq 1/2 + 2u$, $x \geq 1/2 + 2u$, and $|\delta|/x \leq 3u/(1 + 4u)$. If $b = 1$, we consider two subcases. If $a \geq -1/2 + 2u$, then $a+b \geq 1/2+2u$ as above. It remains three values of a to check for $b = 1$, namely, $a \in \{-1/2+u/2, -1/2+u, -1/2+3u/2\}$. For $a = -1/2 + u/2$, we get $x = 1/2$, $z = 1 - u$, and $|\delta|/x = u < 3u/(1 + 4u)$. For $a = -1/2 + u$, we get $x = 1/2 + u = a + b$ and $\delta = 0$. For $a = -1/2 + 3u/2$, we get $x = 1/2+u$, $z = 1-u$, and $|\delta|/x = u/(1+2u) < 3u/(1+4u)$.

Case $a+b \leq 1/2$. This case implies $a < 0$ and thus $a+b = b-|a| \leq 1/2 \leq b/2$. Hence $b/2 \leq |a| \leq b$, which by Sterbenz' theorem yields $a + b \in \mathbb{F}$ and $\delta = 0$. □

Theorem 8. *Let $a, b \in \mathbb{F}$ and $[x, y] = \text{FastTwoSum}(a, b, \circ)$ with $\circ \in \{\text{RD}, \text{RU}\}$, $\beta = 2$, and $p \geq 5$. If $|a| < |b|$ then*

$$|\Delta| \leq \frac{3u}{1+2u}|x|.$$

Proof. We assume without loss of generality $0 < |a| < 1 \leq b < 2$ (since $\text{RU}(-r) = -\text{RD}(r)$ for $r \in \mathbb{R}$, we can assume $b > 0$, by swapping the rounding modes), and we denote $x = a+b-e$, and $z = b-e+\delta$. First, by Lemma 3, $y = b-z$, thus the only extra rounding error is due to the second operation $z = \circ(x-a) = \circ(b-e)$, and we have $|\Delta| = |\delta|$. We distinguish 4 cases: $a + b \geq 2$, $1 \leq a + b < 2$, $1/2 < a + b < 1$, and $a + b \leq 1/2$.

The cases $a + b \geq 2$, $1 \leq a + b < 2$, and $a + b \leq 1/2$ are dealt exactly as in the proof of Theorem 7, since for these cases we did not use the rounding mode, and $3u/(1 + 4u) \leq 3u/(1 + 2u)$.

It remains to deal with the case $1/2 < a + b < 1$. Again the subcases $|a| \geq 1/2$ and $a \in [-1/4, 1/2)$ are dealt as in the proof of Theorem 7, and similarly for $a \in (-1/2, -1/4)$ and $b \geq 1 + 2u$. It thus only remains to check the case $b = 1$ and $a \in \{-1/2 + u/2, -1/2 + u, -1/2 + 3u/2\}$, for both rounding modes RD and RU. Since in the proof of Theorem 7 we obtained non-negative values of x and z for these three values of a and with RZ, we will obtain the same values with RD, and thus the bound of Theorem 7 applies. For RU we obtain for $a = -1/2 + u/2$: $x = 1/2 + u$, $z = 1 + 2u$, and $|\delta|/x = 3u/(1+2u)$; for $a = -1/2+u$: $x = 1/2+u = a+b$ and $\delta = 0$; and for $a = -1/2 + 3u/2$: $x = 1/2 + 2u$, $z = 1 + 2u$, and thus $|\delta|/x = 3u/(1 + 4u) < 3u/(1 + 2u)$. □

We show in the next example that the bounds established in Theorems 7 and 8 are attained.

Example 8 (Optimality of the bounds)**.** *For RZ, let $a = -1+u$ and $b = 2 + 4u$. Then $x = 1 + 4u$, $z = 2$, $y = 4u$, and $|\Delta|/|x| = 3u/(1 + 4u)$, thus the bound from Theorem 7 is attained. For RU, let $a = -1+u$ and $b = 2$. Then $x = 1+2u$, $z = 2 + 4u$, $y = -4u$, and $|\Delta|/|x| = 3u/(1 + 2u)$. For RD, use the same example with $(-a, -b)$.*

We conclude this section by summarizing in Table I the various error bounds seen here and in Section V-A. For directed roundings, the tight bound $2u^2\text{ufp}(x)$ in the normal case is from Theorem 6 and the bounds $3u/(1 + 4u) \cdot |x|$ and $3u/(1 + 2u) \cdot |x|$ for the reversed case are from Theorems 7 and 8. Although the latter only slightly improve the bound $3u|x|$ from [18], they are attained for some values of a and b, as shown in the above example. The bound $u|x|$, that is attained for rounding to nearest with ties to even, is from [18].

TABLE I
TIGHT BOUNDS ON THE ERROR $|\Delta| = |x + y - (a + b)|$.

	RN	RZ	RD, RU
Normal: $\lvert a\rvert \geq \lvert b\rvert$	0	$2u^2\text{ufp}(x)$	$2u^2\text{ufp}(x)$
Reversed: $\lvert a\rvert < \lvert b\rvert$	$u\lvert x\rvert$	$\frac{3u}{1+4u}\lvert x\rvert$	$\frac{3u}{1+2u}\lvert x\rvert$

VI. Correctly-rounded results in doubled precision

In the case $a \in \text{ulp}(b)\mathbb{Z}$, we know from Section V-A that when FastTwoSum is not an EFT then $x + y$ is very close to $a + b$ in the sense that their difference is tightly bounded by $2u^2\text{ufp}(x)$. Here we go one step further by showing for RZ, RU, and RD that when $x + y$ differs from $a + b$ then it must be the correctly-rounded value of $a + b$ in precision $2p$ and in the same direction.

A. Rounding to zero

For RZ, such a result appears in [13, Theorem 4], for any base and assuming $|a| \geq |b|$. The version we give in Theorem 9 is for base 2 but assumes only $a \in \mathrm{ulp}(b)\mathbb{Z}$, and we show how to extend it to both RU and RD in Theorems 10 and 11.

Theorem 9. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ and $[x, y] = \mathrm{FastTwoSum}(a, b, \mathrm{RZ})$. If $a \in \mathrm{ulp}(b)\mathbb{Z}$ then*

$$x + y = \begin{cases} a + b & \text{if } ab \geq 0 \text{ or } e_a - e_b \leq p, \\ \mathrm{RZ}_{2p}(a + b) & \text{otherwise.} \end{cases}$$

Proof. We know from Theorem 5 that $x + y = a + b$ when $ab \geq 0$ or $e_a - e_b \leq p$. We are thus left with the case where

$$ab < 0 \quad \text{and} \quad e_a - e_b > p,$$

for which we want to show that $x + y = \mathrm{RZ}_{2p}(a + b)$.

If $a + b = 0$ then $x = y = 0$ and the result holds. Since $\mathrm{FastTwoSum}(-a - b, \mathrm{RZ}) = -\mathrm{FastTwoSum}(a + b, \mathrm{RZ})$ and $\mathrm{RZ}_{2p}(-a - b) = -\mathrm{RZ}_{2p}(a + b)$, we see that it suffices to prove the result in the case where $a + b > 0$.

Since $ab < 0$ implies $a \neq 0$ and $b \neq 0$, we deduce from $e_a - e_b > p$ that $\mathrm{ulp}(a) = 2^{e_a - p + 1} \geq 2^{e_b + 2} > 2|b|$. Hence a has the same sign as $a + b$, that is, $a > 0$, and thus $b < 0$.

If $a > \mathrm{ufp}(a)$ then, since $|b| < \mathrm{ulp}(a)$ and $b < 0$, we have $\mathrm{ufp}(a + b) = \mathrm{ufp}(a) \leq a + b < a$, $x = a - \mathrm{ulp}(a)$, $z = -\mathrm{ulp}(a)$, and $y = \mathrm{RZ}(e)$ with $e := b + \mathrm{ulp}(a)$. Since $|b| < \mathrm{ulp}(a)/2$ and $b < 0$, $\mathrm{ufp}(e) = \mathrm{ulp}(a)/2 = 2^{e_a - p}$. Now since $y = \mathrm{RZ}(e)$ with $e > 0$, $y = e - \varepsilon$ with $0 \leq \varepsilon < \mathrm{ulp}(e) = 2^{e_a - 2p + 1} = \mathrm{ulp}_{2p}(a) = \mathrm{ulp}_{2p}(a + b)$. We thus have $x + y = a + b - \varepsilon$ with $a + b$ not an integral power of 2 and $0 \leq \varepsilon < \mathrm{ulp}_{2p}(a + b)$; furthermore, one can check that $x + y \in [0, 2^{e_a + 1})$ and $x + y \in 2^{e_a - 2p + 1}\mathbb{Z}$, so $x + y$ is exact in precision $2p$. Therefore, $x + y = \mathrm{RZ}_{2p}(a + b)$.

If $a = \mathrm{ufp}(a)$ then $\mathrm{ufp}(a + b) = 2^{e_a - 1}$ and $a + b > a - \mathrm{ulp}(a)/2$, which is the predecessor of a in $\mathbb{F}$. Thus $x = \mathrm{RZ}(a + b) = a - \mathrm{ulp}(a)/2$, $z = -\mathrm{ulp}(a)/2$, and $y = \mathrm{RZ}(e)$ with $e = b + \mathrm{ulp}(a)/2$. Now two subcases can happen. If $e_a - e_b \geq p + 2$ then $|b| < \mathrm{ulp}(a)/4$, and $\mathrm{ufp}(e) = \mathrm{ulp}(a)/4 = 2^{e_a - p - 1}$. Now since $y = \mathrm{RZ}(e)$ with $e > 0$, $y = e - \varepsilon$ with $0 \leq \varepsilon < \mathrm{ulp}(e) = 2^{e_a - 2p}$. We thus have $x + y = a + b - \varepsilon$ with $a + b$ not an integral power of 2 and $0 \leq \varepsilon < \mathrm{ulp}_{2p}(a + b)$; furthermore, one can check that $x + y \in [0, 2^{e_a})$ and $x + y \in 2^{e_a - 2p}\mathbb{Z}$, so $x + y$ is exact in precision $2p$. Therefore, $x + y = \mathrm{RZ}_{2p}(a + b)$. If $e_a - e_b = p + 1$, then $\mathrm{ulp}(b) = 2^{e_a - 2p} = \mathrm{ulp}_{2p}(a + b)$, thus $a + b$ is exact in precision $2p$, and $x + y = a + b$. □

B. Rounding up and rounding down

Theorem 10. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ and $[x, y] = \mathrm{FastTwoSum}(a, b, \mathrm{RU})$. If $a \in \mathrm{ulp}(b)\mathbb{Z}$ then*

$$x + y = \begin{cases} a + b & \text{if } b \leq 0 \text{ or } e_a - e_b \leq p, \\ \mathrm{RU}_{2p}(a + b) & \text{otherwise.} \end{cases}$$

Proof. When $b \leq 0$ or $e_a - e_b \leq p$, we have an EFT (Theorem 4). Assume $b > 0$ and $e_a - e_b > p$. Hence $a \neq 0$, $0 < b < \mathrm{ulp}(a)/2$, and $a + b$ has the same sign as a. We distinguish the cases $a < 0$ and $a > 0$.

If $a < 0$, we have $a + b < 0$ and $x = \mathrm{RU}(a + b) \in \{a + \mathrm{ulp}(a), a + \mathrm{ulp}(a)/2\}$. Therefore $x - a \in \mathbb{F}$ and so $z = \mathrm{RU}(x - a) = x - a \in \{\mathrm{ulp}(a), \mathrm{ulp}(a)/2\}$. Finally, $y = \mathrm{RU}(e)$ with $e \leq 0$ such that $e \in \{b - \mathrm{ulp}(a), b - \mathrm{ulp}(a)/2\}$. In all cases $a + b < 0$, $x - a \in \mathbb{F}$, and $e \leq 0$, thus RU $\approx$ RZ, and the proof of Theorem 9 applies.

Assume $a > 0$. Since $0 < b < \mathrm{ulp}(a)/2$, we have $x = a + \mathrm{ulp}(a)$, $z = \mathrm{ulp}(a)$, and $y = \mathrm{RU}(e)$ with $e = b - \mathrm{ulp}(a)$. Now $\mathrm{ufp}(e) = \mathrm{ulp}(a)/2 = 2^{e_a - p}$, and so $y = e + \varepsilon$ with $0 \leq \varepsilon < \mathrm{ulp}(e) = 2^{e_a - 2p + 1}$. We also have $\mathrm{ufp}(a + b) = \mathrm{ufp}(a) = 2^{e_a}$. It follows that $x + y = a + b + \varepsilon$ with $0 \leq \varepsilon < 2^{e_a - 2p + 1} = \mathrm{ulp}_{2p}(a + b)$; furthermore, one can check that $x + y \in [0, 2^{e_a + 1}]$ and $x + y \in 2^{e_a - 2p + 1}\mathbb{Z}$, so $x + y$ is exact in precision $2p$. Consequently, $x + y = \mathrm{RU}_{2p}(a + b)$. □

Theorem 11. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ and $[x, y] = \mathrm{FastTwoSum}(a, b, \mathrm{RD})$. If $a \in \mathrm{ulp}(b)\mathbb{Z}$ then*

$$x + y = \begin{cases} a + b & \text{if } b \geq 0 \text{ or } e_a - e_b \leq p, \\ \mathrm{RD}_{2p}(a + b) & \text{otherwise.} \end{cases}$$

Proof. When $b \geq 0$ or $e_a - e_b \leq p$, we have an EFT (Theorem 3). When $b < 0$ and $e_a - e_b > p$, the result follows from Theorem 10, since the fact that $\mathrm{RD}(r) = -\mathrm{RU}(-r)$ for all $r \in \mathbb{R}$ implies $[x, y] = -\mathrm{FastTwoSum}(-a, -b, \mathrm{RU})$ and $\mathrm{RD}_{2p}(a + b) = -\mathrm{RU}_{2p}(-a - b)$. □

Theorems 9, 10, 11 assume that a is an integral multiple of $\mathrm{ulp}(b)$. The next example shows that this assumption cannot, in general, be replaced by a weaker one of the form $a \in 2^{-k}\mathrm{ulp}(b)\mathbb{Z}$ with $k \geq 1$.

Example 9 (When $a \notin \mathrm{ulp}(b)\mathbb{Z}$)**.** *Consider $a = -2u \cdot 2^{-k}$ for $k \geq 1$, and $b = 1 + 2u$. Then a and b are in $\mathbb{F}$ and such that $a + b = 1 + 2u - 2u \cdot 2^{-k}$, $\mathrm{ulp}(b) = 2u$, and $a \notin \mathrm{ulp}(b)\mathbb{Z}$. For* RZ, *this gives $x = 1$, $z = 1$, $y = 2u$, and $x + y = 1 + 2u$. Since $a + b$ and $\mathrm{RZ}_{2p}(a + b)$ are both strictly less than $1 + 2u$, $x + y$ is neither of them.*

For RU, *consider $a = 2u \cdot 2^{-k}$ with $k \geq 1$, and $b = 1$. Then $a + b = 1 + 2u \cdot 2^{-k}$, $x = 1 + 2u$, $z = 1 + 2u$, $y = -2u$, and $x + y = 1 \notin \{a + b, \mathrm{RU}_{2p}(a + b)\}$, since both $a + b$ and* $\mathrm{RU}_{2p}(a + b)$ *are strictly larger than one.*

For RD, *it is sufficient to negate the values of a and b used above for* RU *to reach the conclusion that $x + y$ is neither $a + b$ nor* $\mathrm{RD}_{2p}(a + b)$.

Remark that when $a \in \mathrm{ulp}(b)\mathbb{Z}$ and $e_a - e_b \leq p$, the exact sum $a + b$ is a floating-point number in precision $2p$ and thus $a + b = \mathrm{RD}_{2p}(a + b) = \mathrm{RU}_{2p}(a + b)$.

Hence a direct consequence of Theorems 10 and 11 is the fact that when we run FastTwoSum twice, once with RD and one with RU, then among the two resulting floating-point pairs, one gives the exact sum $a + b$ and the other gives its correctly-rounded value in precision $2p$, either upwards or downwards. This fact is stated precisely in the following corollary.

Corollary 1. *For $\beta = 2$, $p \geq 2$, let $a, b \in \mathbb{F}$ be such that $a \in \mathrm{ulp}(b)\mathbb{Z}$, let $[\underline{x}, \underline{y}] = \mathrm{FastTwoSum}(a, b, \mathrm{RD})$ and $[\overline{x}, \overline{y}] = \mathrm{FastTwoSum}(a, b, \mathrm{RU})$. If $b \geq 0$ then*

$$\underline{x} + \underline{y} = a + b \leq \overline{x} + \overline{y} = \mathrm{RU}_{2p}(a + b),$$

else

$$\underline{x} + \underline{y} = \mathrm{RD}_{2p}(a + b) \leq a + b = \overline{x} + \overline{y}.$$

In particular, this result implies that if $a \in \mathrm{ulp}(b)\mathbb{Z}$ then $a + b \in [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$, a fact shown in [21, Prop. 3.3] under the more restrictive assumption $|a| \geq |b|$. But it says more about the very nature of the interval $[\underline{x} + \underline{y}, \overline{x} + \overline{y}]$ by showing that this interval can take only two very specific forms, namely, $[a + b, \mathrm{RU}_{2p}(a + b)]$ and $[\mathrm{RD}_{2p}(a + b), a + b]$.

VII. Concluding remarks

In this paper we revisited the FastTwoSum algorithm for binary arithmetic with unbounded exponent range, by giving sufficient conditions for the following three properties to hold for various rounding modes: the error $e = a + b - x$ is a floating-point number, the middle result $z = \circ(x - a)$ is exact, and the final result $y = \circ(b - z)$ is exact. These properties enable one to obtain sufficient conditions for error-free transforms (EFTs) and tight error bounds otherwise. In addition, we identified cases where FastTwoSum yields the correct rounding in doubled precision of $a + b$.

Although we focused on FastTwoSum as described in [6, p. 228], other versions exist in the literature. A first variant consists in defining the error estimate as $y = \circ(b + \circ(a - x))$; see [1], [6, p. 230], [22], [23], [24], and [25]. The results we obtained still hold here if an anti-symmetric rounding is used or if $a \in \mathrm{ulp}(b)\mathbb{Z}$ (since then $a - x = -(x - a) \in \mathbb{F}$ by Lemma 2). But for directed roundings and $|a| < |b|$, the behavior can be different and this requires further investigation.

A second variant consists in computing $y' = \circ(z - b)$ and approximating $a + b$ by $x - y'$; see for example [13], [26], [27], and [28]. Again, an anti-symmetric rounding would preserve our results. However, and in contrast with the first variant, some properties can be lost for directed roundings even if $a \in \mathrm{ulp}(b)\mathbb{Z}$. In particular, we show in the example below that Theorems 10 and 11 do not hold anymore.

Example 10. *(Failure of the second variant with* RD *and* RU*) Let $a = 1$ and $b = u^2$. Then $x = \mathrm{RU}(a + b) = 1 + 2u$, $z = \circ(x - a) = 2u$ for any faithful rounding, and $y' = \mathrm{RU}(z - b) = 2u$. This gives $x - y' = 1$, which is neither $a + b = 1 + u^2$, nor $\mathrm{RU}_{2p}(a + b) = 1 + 2u^2$. For* RD*, a similar counter-example is obtained by negating a and b.*

In addition to these two variants of FastTwoSum, various results have been given for larger bases [6], [10], [13], [29], when underflow or overflow occurs [12], [18], and when other roundings are assumed, such as rounding to odd [13] or double rounding [30], [31]. It would be worthwhile to try to adapt the techniques presented in the paper to these settings. We also plan to explore how our work can help complement or refine existing analyses of more sophisticated algorithms such as TwoSum and its variants [2], [12], [22], [32].

Acknowledgements. We thank the three anonymous referees for their helpful comments.

References

[1] W. Kahan, “Further remarks on reducing truncation errors,” *Comm. ACM*, 1965.

[2] O. Møller, “Quasi double-precision in floating point addition,” *BIT*, 1965.

[3] ——, “Note on quasi double-precision,” *BIT*, 1965.

[4] J.-M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldes, V. Lefèvre, G. Melquiond, N. Revol, and S. Torres, *Handbook of Floating-Point Arithmetic*, 2nd ed. Birkhäuser, 2018.

[5] A. Sibidanov, P. Zimmermann, and S. Glondu, “The CORE-MATH Project,” in *ARITH Proc.*, 2022, https://core-math.gitlabpages.inria.fr/.

[6] T. J. Dekker, “A floating-point technique for extending the available precision,” *Numer. Math.*, 1971.

[7] IEEE, *IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019)*, 2019.

[8] S. M. Rump, T. Ogita, and S. Oishi, “Accurate floating-point summation, Part I: Faithful rounding,” *SIAM J. Sci. Comput.*, 2008.

[9] P. H. Sterbenz, *Floating-Point Computation*. Prentice-Hall, 1974.

[10] V.-E. J. Virkkunen, “A unified approach to floating-point rounding with applications to multiple-precision summation,” 1980, Report A-1980-1, Department of Computer Science, University of Helsinki.

[11] J. Demmel and H. D. Nguyen, “Fast reproducible floating-point summation,” in *ARITH Proc.*, 2013.

[12] S. Boldo, S. Graillat, and J.-M. Muller, “On the robustness of the 2Sum and Fast2Sum algorithms,” *ACM Trans. Math. Software*, 2017.

[13] S. Linnainmaa, “Analysis of some known methods of improving the accuracy of floating-point sums,” *BIT*, 1974.

[14] M. Dukhan, R. Vuduc, and J. Riedy, “Wanted: Floating-point add round-off error instruction,” 2016, preprint (https://arxiv.org/abs/1603.00491).

[15] C. Lauter, “An efficient software implementation of correctly rounded operations extending FMA: $a + b + c$ and $a \times b + c \times d$,” in *ACSSC Proc.*, 2017.

[16] J. Riedy and J. Demmel, “Augmented arithmetic operations proposed for IEEE-754 2018,” in *ARITH Proc.*, 2018.

[17] M. Daumas, L. Rideau, and L. Théry, “A generic library of floating-point numbers and its application to exact computing,” in *TPHOLs Proc.*, 2001.

[18] S. Corbineau and P. Zimmermann, “Note on FastTwoSum with directed roundings,” 2024, https://inria.hal.science/hal-03798376.

[19] S. Graillat, F. Jézéquel, and R. Picot, “Numerical validation of compensated summation algorithms with stochastic arithmetic,” *Electronic Notes in Theoretical Computer Science*, 2015.

[20] ——, “Numerical validation of compensated algorithms with stochastic arithmetic,” *Applied Mathematics and Computation*, 2018.

[21] S. Graillat and F. Jézéquel, “Tight interval inclusions with compensated algorithms,” *IEEE Trans. Comput.*, 2020.

[22] D. E. Knuth, *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*, 3rd ed. Addison-Wesley, 1998.

[23] D. Goldberg, “What every computer scientist should know about floating-point arithmetic,” *ACM Computing Surveys*, 1991.

[24] S. M. Rump, “Verification methods: Rigorous results using floating-point arithmetic,” *Acta Numerica*, 2010.

[25] S. Boldo and G. Melquiond, *Computer Arithmetic and Formal Proofs*. ISTE Press and Elsevier, 2017.

[26] M. Pichat, “Correction d’une somme en arithmétique à virgule flottante,” *Numer. Math.*, 1972.

[27] ——, “Contribution à l’étude des erreurs d’arrondi en arithmétique à virgule flottante,” Ph.D. dissertation, Université Scientifique et Médicale de Grenoble & Institut National Polytechnique de Grenoble, 1976, https://tel.archives-ouvertes.fr/tel-00287209/.

[28] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed. SIAM, 2002.

[29] M. Lange and S. Oishi, “A note on Dekker’s FastTwoSum algorithm,” *Numer. Math.*, 2020.

[30] S. Boldo and M. Daumas, “Representable correcting terms for possibly underflowing floating point operations,” in *ARITH Proc.*, 2003.

[31] É. Martin-Dorel, G. Melquiond, and J.-M. Muller, “Some issues related to double rounding,” *BIT*, 2013.

[32] D. M. Priest, “Algorithms for arbitrary precision floating point arithmetic,” in *ARITH Proc.*, 1991.