Odd but Error-Free FastTwoSum

Sehyeok Park
Rutgers University
New Brunswick, New Jersey, U.S.A
sp2044@cs.rutgers.edu

Jay P. Lim
University of California, Riverside
Riverside, California, U.S.A
jlim@ucr.edu

Santosh Nagarakatte
Rutgers University
New Brunswick, New Jersey, U.S.A
santosh.nagarakatte@cs.rutgers.edu

***Abstract*—This paper proposes sufficient, yet more general conditions for applying FastTwoSum as an error-free transformation (EFT) under all faithful rounding modes. Additionally, it also identifies guarantees tailored to round-to-odd for establishing FastTwoSum as an EFT. This paper also describes a floating-point splitting tailored for round-to-odd that is an EFT where the distribution of bits is configurable (*i.e.*, ExtractScalar for round-to-odd). Our sufficient conditions are more general than those previously known in the literature (*i.e.*, it applies to a wider operand domain).**

***Index Terms*—FastTwoSum, EFT, round-to-odd**

I. Introduction

Under select circumstances, the rounding error of a finite-precision, floating-point addition is itself a floating-point number. FastTwoSum is the most prominent method for computing floating-point rounding errors using floating-point operations. Given a floating-point set $\mathbb{F}$ and a rounding operation $\circ$ mapping $\mathbb{R}$ to $\mathbb{F}$, FastTwoSum takes inputs $a, b \in \mathbb{F}$ and produces outputs $x, y \in \mathbb{F}$ as shown below.

FastTwoSum(a, b) :
$x = \circ(a + b)$
$z = \circ(x - a)$
$y = \circ(b - z)$
return x, y

FastTwoSum is the foundation for various algorithms designed to improve the accuracy of finite-precision addition. Through the operations shown above, FastTwoSum produces x, the floating-point sum of its inputs using the rounding rule imposed by $\circ$. The second output y is an estimate of $\delta = a + b - x$: the rounding error incurred in computing x. Algorithms based on FastTwoSum utilize this error estimate to either compensate the initial floating-point sum or extend the precision past what is available in $\mathbb{F}$ [1] [2] [3]. Compensation and extended-precision algorithms can maximize the accuracy of their results by applying FastTwoSum as an *error-free transformation (EFT)*, through which the original inputs are transformed to an equivalent pair of outputs (*i.e.*, $x+y = a+b$).

Dekker's analysis of FastTwoSum [1] demonstrates that y is exactly equal to the rounding error δ given the following conditions: $\mathbb{F}$'s base is $\beta \leq 3$, $|a| \geq |b|$, and $\circ$ applies round-to-nearest (RN). The equivalence between y and $\delta = a+b-x$ implies $x + y = a + b$, which ensures FastTwoSum is an EFT.

Later analyses of FastTwoSum have explored EFT guarantees beyond the setting proposed by Dekker. In particular, several works have analyzed FastTwoSum's exactness and rounding properties in the absence of RN [4] [5] [6] [7]. The rounding error induced by floating-point addition under RN is guaranteed to be a floating-point number barring any overflow [8] [9] [10]. However, the guarantee does not extend to other faithful rounding modes without the RN properties [10].

The IEEE-754 standard [11] supports multiple rounding modes: round-to-nearest (ties-to-even (RNE) or ties-to-away (RNA)), round-down (RD), round-up (RU), and round-toward-zero (RZ). The latter three, which are collectively referred to as directed rounding modes, do not guarantee that numbers are rounded to their closest values in $\mathbb{F}$. While IEEE-754 enforces RNE as the default rounding mode, the directed rounding modes have multiple use cases including interval arithmetic. Due to the high cost of switching rounding modes on modern machines and the emergence of custom floating-point units without RNE support, generalizing EFT guarantees for FastTwoSum to other rounding modes is of great interest.

Round-to-odd (RO) is a faithful rounding mode that will be supported by the upcoming P3109 standard for machine learning arithmetic formats. Akin to RN, RO is an unbiased rounding mode for which the probabilistic mean value of the error for a single instance of rounding is zero. Unlike RN, RO makes double rounding innocuous. Rounding to some intermediate, higher precision format via RO prior to the final rounding produces the same result as directly rounding to the destination format [12] [13]. An interesting application of RO is the development of multi-precision math libraries [14]. Given the anticipated increase in RO adoption due to upcoming standards, establishing EFT guarantees for FastTwoSum under RO will be of practical use in the design of future algorithms.

FastTwoSum is an EFT when the operation $z = \circ(x - a)$ is exact and the rounding error $\delta = a + b - x$ is representable in floating-point. This paper establishes that both properties are guaranteed under all faithful rounding modes if (i) a is an integer multiple of $\text{ulp}(b)$ and (ii) b is an integer multiple of $2u^2 \cdot \text{ufp}(a)$, where u, $\text{ufp}(\cdot)$, and $\text{ulp}(\cdot)$ denote the unit round-off, unit in the first place, and unit in the last place, respectively. Given p-bits of available precision, these conditions can potentially tolerate up to a $2p - 1$ exponent difference between a and b. For operands that satisfy our conditions, the permitted exponent difference nearly doubles

that of previously known EFT conditions for faithfully rounded FastTwoSum, making our conditions applicable to a wider input domain. In addition to these conditions, which suffice for all faithful rounding modes including RO, we present new EFT guarantees specifically tailored to RO. We apply these conditions to the ExtractScalar algorithm - a FastTwoSum-based EFT for RN that splits a single floating-point input across two numbers while maintaining the original value [15]. Specifically, we design a new variant of ExtractScalar that preserves its original properties under RO.

II. Background

A. Definitions

We denote by $\mathbb{F}$ the set of floating-point numbers of base $\beta = 2$, precision $p \geq 2$, and extremal exponents E_{min} and E_{max}. We additionally include infinities in $\mathbb{F}$. For all reals $x \in \mathbb{F}$, $x = M_x \times 2^{E_x - p+1}$ for a unique pair of $M_x, E_x \in \mathbb{Z}$ that meet the following conditions.

$$E_{min} \leq E_x \leq E_{max} \qquad |M_x| < 2^p$$

$$E_x > E_{min} \Rightarrow |M_x| \geq 2^{p-1}$$

A nonzero real $x \in \mathbb{F}$ is normal if $|x| \geq 2^{E_{min}}$ and subnormal otherwise. We denote the largest finite number in $\mathbb{F}$ by $\Omega = (2^p - 1) \times 2^{E_{max}-p+1}$. Similarly, $\omega = 2^{E_{min}-p+1}$ represents the smallest nonzero magnitude.

We express through $\circ \in$ RN that $\circ$ applies RN with any arbitrary tie breaking rule (*e.g.*, ties-to-nearest, ties-to-away). We apply the notation $\circ \in$ FR from [7] to indicate that $\circ$ performs faithful rounding (*i.e.*, $\forall r \in \mathbb{R}, \circ(r) \in \{\mathrm{RD}(r), \mathrm{RU}(r)\}$). We use $\circ(a \pm b)$ to denote floating-point addition or subtraction between $a, b \in \mathbb{F}$ under the rounding mode specified by $\circ$. We denote by $u = 2^{-p}$ the unit round-off, which is the distance between 1 and its closest number in $\mathbb{F}$. For a given $r \in \mathbb{R}$, we define its exponent e_r, unit in the first place ($\mathrm{ufp}(r)$), and unit in the last place ($\mathrm{ulp}(r)$) as shown below.

Definition 1. *For $r \in \mathbb{R}$,*

$$e_r = \begin{cases} -\infty, & \textit{if } r = 0 \\ \lfloor \log_2(|r|) \rfloor & \textit{otherwise} \end{cases} \qquad \mathrm{ufp}(r) = 2^{e_r}$$

$$\mathrm{ulp}(r) = \begin{cases} 2u \cdot \mathrm{ufp}(r), & \textit{if } |r| \geq 2^{E_{min}} \\ \omega & \textit{otherwise} \end{cases}$$

Definition 1 ensures for all finite $x \in \mathbb{F}$ that $\mathrm{ulp}(x) \geq 2u \cdot \mathrm{ufp}(x)$. Moreover, for all finite $x, y \in \mathbb{F}$, $\mathrm{ufp}(x) \geq \mathrm{ufp}(y)$ implies $\mathrm{ulp}(x) \geq \mathrm{ulp}(y)$.

We denote by $\mathrm{pred}(r)$ the *floating-point predecessor* of $r \in \mathbb{R}$. Conversely, $\mathrm{succ}(r)$ represents the *floating-point successor*.

Definition 2. *For $r \in \mathbb{R}$ and $r \notin \mathbb{F}$,*

$$\mathrm{pred}(r) = \max\{x \in \mathbb{F} \mid x < r\}$$
$$\mathrm{succ}(r) = \min\{x \in \mathbb{F} \mid x > r\}$$

We collectively refer to $\mathrm{pred}(r)$ and $\mathrm{succ}(r)$ as r's floating-point neighbors. For finite numbers in $\mathbb{F}$, we define their floating-point neighbors as follows.

Definition 3. *For finite $x \in \mathbb{F}$,*

$$\mathrm{pred}(x) = \begin{cases} -\omega, & \textit{if } x = 0 \\ -\infty, & \textit{if } x = -\Omega \\ x - \frac{1}{2}\mathrm{ulp}(x), & \textit{if } x = \mathrm{ufp}(x) \textit{ and } x > 2^{E_{min}} \\ x - \mathrm{ulp}(x) & \textit{otherwise} \end{cases}$$

$$\mathrm{succ}(x) = \begin{cases} \omega, & \textit{if } x = 0 \\ \infty, & \textit{if } x = \Omega \\ x + \frac{1}{2}\mathrm{ulp}(x), & \textit{if } x = -\mathrm{ufp}(x) \textit{ and } x < -2^{E_{min}} \\ x + \mathrm{ulp}(x) & \textit{otherwise} \end{cases}$$

B. Basic Properties of Floating-Point

We present the basic floating-point properties most relevant to our theorems. Henceforth, the notation $x \in y\mathbb{Z}$ for $x, y \in \mathbb{R}$ expresses that x is an integer multiple of y. Every finite, nonzero $x \in \mathbb{F}$ satisfies $\mathrm{ufp}(x) \leq |x| < 2\mathrm{ufp}(x)$. Every finite $x \in \mathbb{F}$ also satisfies $x \in \omega\mathbb{Z}$ and $x \in \mathrm{ulp}(x)\mathbb{Z}$. If $x \in 2^k\mathbb{Z}$ for $k \in \mathbb{Z}$, then $x \in 2^i\mathbb{Z}$ for any $i \in \mathbb{Z}$ less than k. Because $2u \cdot \mathrm{ufp}(x)$ is an integer power of 2 and Definition 1 guarantees $\mathrm{ulp}(x) \geq 2u \cdot \mathrm{ufp}(x)$, it follows that $x \in 2u \cdot \mathrm{ufp}(x)\mathbb{Z}$. By extension, x is an integer multiple of any smaller integer power of 2 (*e.g.*, $u \cdot \mathrm{ufp}(x)$, $2u^2 \cdot \mathrm{ufp}(x)$, *etc.*).

We assume throughout the remainder of the paper that the operands of FastTwoSum are finite (*i.e.*, $|a|, |b| \leq \Omega$). All finite $a, b \in \mathbb{F}$ satisfy $a, b \in \min(\mathrm{ulp}(a), \mathrm{ulp}(b))\mathbb{Z}$. Because addition and subtraction preserve common factors, the real sum $a + b$ and its floating-point counterpart $\circ(a+b)$ satisfy $a + b, \circ(a + b) \in \min(\mathrm{ulp}(a), \mathrm{ulp}(b))\mathbb{Z}$ for all $\circ \in$ FR. Consequently, the resulting rounding error $\delta = a + b - \circ(a + b)$ adheres to $\delta \in \min(\mathrm{ulp}(a), \mathrm{ulp}(b))\mathbb{Z}$. Because $a, b \in \omega\mathbb{Z}$, it also follows that $a+b, \circ(a+b), \delta \in \omega\mathbb{Z}$. This property implies neither the sum of numbers in $\mathbb{F}$ nor the associated rounding error is subject to underflow, meaning these values are integer multiples of ω.

A number $r \in \mathbb{R}$ is in $\mathbb{F}$ if there exist $M_r, E_r \in \mathbb{Z}$ such that $r = M_r \times 2^{E_r - p+1}$, $|M_r| < 2^p$, and $E_{min} \leq E_r \leq E_{max}$. We provide the conditions sufficient to meet these constraints in the following lemma.

Lemma 1. *Let $r \in \mathbb{R}$, $k \in \mathbb{Z}$, and $\sigma = 2^k$. If $|r| \leq \min(\sigma, \Omega)$ and $r \in \max(u \cdot \sigma, \omega)\mathbb{Z}$, then $r \in \mathbb{F}$.*

If $a, b \in \mathbb{F} \setminus \{\pm\infty\}$, then $\delta = a + b - \circ(a + b) \in \omega\mathbb{Z}$ for all $\circ \in$ FR. Therefore, $\delta \in u \cdot 2^k\mathbb{Z}$ for $k \in \mathbb{Z}$ implies $\delta \in \max(u \cdot 2^k, \omega)\mathbb{Z}$.

Corollary 1. *Let $a, b \in \mathbb{F} \setminus \{\pm\infty\}$ and $k \in \mathbb{Z}$. Let $\delta = a+b-\circ(a+b)$ and $\sigma = 2^k$. If $|\delta| \leq \min(\sigma, \Omega)$ and $\delta \in u \cdot \sigma\mathbb{Z}$, then $\delta \in \mathbb{F}$.*

Lastly, we summarize the key properties of rounding. For all $r \in \mathbb{R}$ such that $|r| \leq \Omega$, faithful rounding guarantees $|\circ(r) - r| \leq \frac{1}{2}\mathrm{ulp}(r)$ when $\circ \in$ RN and $|\circ(r) - r| < \mathrm{ulp}(r)$ otherwise. Faithful rounding also enforces sign preservation: $\circ(r) \cdot r \geq 0$ for all $\circ \in$ FR and $r \in \mathbb{R}$.

C. Properties of Round-to-Odd

For base $\beta = 2$, round-to-odd (RO) is a faithful rounding mode that maps all numbers $r \notin \mathbb{F}$ to a floating-point neighbor $x \in \{\text{pred}(r), \text{succ}(r)\}$ such that x's binary encoding forms an *odd integer*. For all numbers $r \in \mathbb{R}$ and $\mathbb{F}$ such that $p \geq 2$, we define $\text{RO}(r)$ as follows.

Definition 4. *For $r \in \mathbb{R}$ and $\mathbb{F}$ such that $p \geq 2$,*

$$\text{RO}(r) = \begin{cases} r, & \text{if } r \in \mathbb{F} \\ \text{succ}(r), & \text{if } r \notin \mathbb{F} \text{ and } \text{succ}(r) \text{ has odd encoding} \\ \text{pred}(r) & \text{if } r \notin \mathbb{F} \text{ and } \text{pred}(r) \text{ has odd encoding} \end{cases}$$

Due to Definition 4, $x = \text{RO}(r) = M_x \times 2^{E_x - p + 1}$ has an *even significand* (*i.e.*, M_x is an even integer) only if $r \in \mathbb{F}$. If $p \geq 2$, all finite $x \in \mathbb{F}$ such that $|x| = \text{ufp}(x) > \omega$ (*i.e.*, $|x|$ is an integer power of 2 greater than ω) have even significands. Therefore, RO does not renormalize within the dynamic range: for all $r \in \mathbb{R}$ such that $\omega \leq |r| \leq \Omega$, $x = \text{RO}(r) \in \mathbb{F}$ satisfies $e_x = e_r$. Definition 4 also implies RO $\notin$ RN; hence, $|\text{RO}(r) - r| < \text{ulp}(r)$ for all reals r such that $|r| \leq \Omega$.

D. Exactness Conditions for FastTwoSum

Given finite inputs $a, b \in \mathbb{F}$ such that $|a + b| \leq \Omega$ and the operations $x = \circ_1(a + b)$, $z = \circ_2(x - a)$, and $y = \circ_3(b - z)$ where $\circ_1, \circ_2, \circ_3 \in$ FR, we present below the properties that ensure $x + y = a + b$ (*i.e.*, FastTwoSum is an EFT). While previous works on the topic denote the rounding error induced by floating-point addition (*i.e.*, $a + b - x$) as e [7], we refer to said error throughout the rest of the paper using δ.

Property 1. $x - a \in \mathbb{F}$
Property 2. $\delta = a + b - x \in \mathbb{F}$

If Property 1 holds, then $z = \circ_2(x - a) = x - a$ for all $\circ_2 \in$ FR. The equality $z = x - a$ implies $y = \circ_3(b - z) = \circ_3(a + b - x)$, the latter of which is the *correct rounding* of the rounding error $\delta = a + b - x$ under $\circ_3$. Properties 1 and 2 thus jointly induce $y = a + b - x$, thereby ensuring the desired equality $x + y = a + b$. We present below an overview of previously established sufficient conditions for each property.

Conditions for Property 1. Jeannerod and Zimmermann prove that $a \in \text{ulp}(b)\mathbb{Z}$ is sufficient for all faithfully rounded sums $\circ_1(a + b)$ [7, Lemma 2]. We recall henceforth that $a \in \text{ulp}(b)\mathbb{Z}$ ensures $z = x - a$ for all $\circ_1, \circ_2 \in$ FR.

Conditions for Property 2. If $\circ \in$ RN, then $\delta = a + b - \circ(a + b)$ must be an element of $\mathbb{F}$. This guarantee, however, could fail if $a + b$ is not rounded to its closest neighbor (*i.e.*, $|\delta| > \frac{1}{2}\text{ulp}(a + b)$). In particular, δ may not be in $\mathbb{F}$ when the exponent ranges occupied by the operands *do not overlap* (*e.g.*, $\text{ulp}(a) > |b|$). Suppose $|b| < \frac{1}{2}\text{ulp}(a)$. Under RN, $\circ(a+b) = a$ and $\delta = b$. Because $b \in \mathbb{F}$, the rounding error is by default an element of $\mathbb{F}$. If $\circ(a + b)$ is not the nearest neighbor of $a + b$ in $\mathbb{F}$, however, δ could be $b \pm \frac{1}{2}\text{ulp}(a)$ or $b \pm \text{ulp}(a)$ (see Definition 3). In such cases, the final rounding error may not be exactly representable in $\mathbb{F}$. For example, suppose $a = 2^p$ and $b = 2^{-p}$. In this example, $|b| < \frac{1}{2}\text{ulp}(a) = 2^0$. If $\circ =$ RU, $\circ(a + b) = 2^p + 2 = a + \text{ulp}(a)$. The resulting rounding error is $a + b - \circ(a + b) = 2^{-p} - 2$, and thus $\delta \notin \mathbb{F}$.

For nonzero $a, b \in \mathbb{F}$, Boldo *et al.* present $|e_a - e_b| \leq p - 1$ as a sufficient condition for Property 2 under all $\circ \in$ FR [6, Lemma 2.6]. Jeannerod and Zimmermann prove in [7, Lemma 1] that the less restrictive conditions $a \in \text{ulp}(b)\mathbb{Z}$ and $e_a - e_b \leq p$ suffice. We recall henceforth that $a \in \text{ulp}(b)\mathbb{Z}$ and $e_a - e_b \leq p$ ensure $\delta \in \mathbb{F}$ for all $\circ \in$ FR. In Section III, we establish new conditions that correctly ensure $\delta \in \mathbb{F}$ for all faithful rounding modes, *even when* $|e_a - e_b|$ *exceeds* p.

For directed rounding modes, the anticipated rounding direction (*i.e.* the expected sign of δ) can influence δ's representability. As before, consider operands $a = 2^p$ and $b = 2^{-p}$. Under RD, δ *must be non-negative* (*i.e.* $a + b \geq \text{RD}(a + b)$). Subsequently, $b \geq 0$ would imply $\text{RD}(a + b) = a$ and $\delta = a + b - a = b \in \mathbb{F}$. Likewise, $b \leq 0$ is sufficient to guarantee $\delta \in \mathbb{F}$ under RU. While assuming $a \in \text{ulp}(b)\mathbb{Z}$, Jeannerod and Zimmermann prove $b \geq 0$, $b \leq 0$, and $a \times b \geq 0$ are sufficient for RD, RU, and RZ respectively [7]. The condition $a \times b \geq 0$ for RZ also appears in [4], which assumes $|a| \geq |b|$. In Section IV, we introduce conditions tailored to RO that *do not restrict the signs of the operands*.

III. EFT Conditions for Faithful Rounding

To establish EFT guarantees for FastTwoSum, we first identify conditions that ensure $\delta = a + b - \circ(a + b) \in \mathbb{F}$ for all $\circ \in$ FR. For this purpose, we present a method of determining δ's membership based on the sum's magnitude.

Lemma 2. *Let $p \geq 2$ and $x = \circ(a + b)$. If both conditions*
(i) $\circ \in$ FR
(ii) $|a + b| \leq 2^{E_{min}+1}$
are satisfied, then $\delta = a + b - x = 0$.

Proof. All finite numbers in $\mathbb{F}$ are integer multiples of ω. Addition and subtraction preserve common factors, so $a+b \in \omega\mathbb{Z}$. Given $\omega = u \cdot 2^{E_{min}+1}$, the bound $|a + b| \leq 2^{E_{min}+1}$ and $a + b \in \omega\mathbb{Z}$ collectively imply $a + b \in \mathbb{F}$ due to Lemma 1. Since x is a faithful rounding of $a+b$, it follows that $x = a+b$ and $\delta = 0$. □

Lemma 2 signifies that the addends' magnitudes can induce implicit bounds on their sum that guarantee $\delta = 0$, which then ensures $\delta \in \mathbb{F}$. We proceed with exploring sufficient conditions for $\delta \in \mathbb{F}$ by leveraging the *relative* magnitudes of a and b.

Lemma 3. *Let $p \geq 2$ and $x = \circ(a + b)$. If all conditions*
(i) $\circ \in$ FR
(ii) $|a + b| \leq \Omega$
(iii) $|a| \geq u \cdot \text{ufp}(b)$
(iv) $|b| \geq u \cdot \text{ufp}(a)$
are satisfied, then $\delta = a + b - x \in \mathbb{F}$.

Proof. Condition (ii) ensures that overflow does not occur for any $\circ \in$ FR (*i.e.*, $|x| \leq \Omega$). Lemma 2 addresses all cases in which $|a + b| \leq 2^{E_{min}+1}$. We thus assume for the remainder of the proof that $|a+b| > 2^{E_{min}+1}$. Because $\delta \in \mathbb{F}$ is trivially true when $a = 0$ or $b = 0$, we also disregard such cases.

Suppose $|a| \geq |b|$. By the definition of ufp, $|a| \geq |b|$ implies $|a| \geq \text{ufp}(b)$. The bound $|a| \geq \text{ufp}(b)$ subsequently induces $|a| \geq u \cdot \text{ufp}(b)$, thereby satisfying Condition (iii). The alternative assumption $|b| > |a|$ would have similarly induced $|b| > u \cdot \text{ufp}(a)$, which signifies *either Condition (iii) or (iv) trivially holds* for all operands.

For all finite $a \in \mathbb{F}$, $a \in \text{ulp}(a)\mathbb{Z}$. If $|a| \geq |b|$, then $\text{ulp}(a) \geq \text{ulp}(b)$ and $a \in \text{ulp}(b)\mathbb{Z}$. Condition (iv) induces $\text{ufp}(b) \geq u \cdot \text{ufp}(a)$, which subsequently implies $e_a - e_b \leq p$. It then follows from [7, Lemma 1] that $\delta \in \mathbb{F}$. Alternatively, $|b| > |a|$ and Condition (iii) would imply $b \in \text{ulp}(a)\mathbb{Z}$ and $e_b - e_a \leq p$ respectively. We can thus swap a and b and apply [7, Lemma 1] to complete the proof. □

Conditions (iii) and (iv) each ensure $\text{ufp}(a) \geq u \cdot \text{ufp}(b)$ and $\text{ufp}(b) \geq u \cdot \text{ufp}(a)$. If $|a| > 0$, Condition (iii) is equivalent to $e_b - e_a \leq p$. Similarly, Condition (iv) is equivalent to $e_a - e_b \leq p$ when $|b| > 0$. Assuming both operands are nonzero, the conjunction of Conditions (iii) and (iv) is subsequently equivalent to $|e_a - e_b| \leq p$. The bound $|e_a - e_b| \leq p$ is comparable to the conditions $a \in \text{ulp}(b)\mathbb{Z}$ and $e_a - e_b \leq p$ proposed in [7]. For any finite $x \in \mathbb{F}$ such that $|x| \geq 2^{E_{min}}$, $\text{ulp}(x) = 2u \cdot \text{ufp}(x)$ due to Definition 1. As such, $|e_a - e_b|$ could be p *or greater* when $\text{ulp}(\max(|a|, |b|)) > \min(|a|, |b|)$. Effectively, Lemma 3 requires that the exponent difference is at most p *when the operands' significands do not overlap*.

Although the conditions in Lemma 3 are *sufficient* to ensure $\delta \in \mathbb{F}$, adherence to the bound $|e_a - e_b| \leq p$ is *not necessary*. For example, consider operands $a = 2^p$ and $b = 2^{-1}$ for which $|a + b| \leq \Omega$ and $e_a - e_b = p + 1$. If $\circ = \text{RU}$, the floating-point sum $x = \circ(a + b)$ produces $2^p + 2$. The associated rounding error $\delta = (2^p + 2^{-1}) - (2^p + 2) = 2^{-1} - 2$ thus satisfies $\delta \in \mathbb{F}$ for precision $p \geq 2$. For $\circ = \text{RD}$, $x = 2^p = a$. The resulting rounding error is $\delta = b$, which is by default in $\mathbb{F}$. If $\circ \in \text{FR}$, then $x \in \{\text{RD}(a + b), \text{RU}(a + b)\}$, and thus $\delta \in \mathbb{F}$ under all faithful rounding modes in this example.

The previous example exhibits that δ can be an element of $\mathbb{F}$ for all $\circ \in \text{FR}$ even when $|e_a - e_b| > p$. We present an additional set of less restrictive conditions that guarantee $\delta \in \mathbb{F}$ for operands that would otherwise be excluded by Lemma 3.

Theorem 1. *Let $p \geq 2$ and $x = \circ(a + b)$. If all conditions*

(i) $\circ \in \text{FR}$
(ii) $|a + b| \leq \Omega$
(iii) $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$
(iv) $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$

are satisfied, then $\delta = a + b - x \in \mathbb{F}$.

Proof. Given Condition (ii), $x = \circ(a + b)$ is finite for all $\circ \in \text{FR}$. As before, we assume $a \neq 0$ and $b \neq 0$. We also assume $|a + b| > 2^{E_{min}+1}$ as Lemma 2 is sufficient otherwise. Because $a \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ for all $a \in \mathbb{F}$, $|a| \geq |b|$ ensures $\text{ufp}(a) \geq \text{ufp}(b)$ and $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$. Similarly, $|b| > |a|$ guarantees $\text{ufp}(b) \geq \text{ufp}(a)$ and $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. Hence, all finite, nonzero $a, b \in \mathbb{F}$ trivially satisfy at least one of Conditions (iii) and (iv). Without loss of generality, we henceforth assume $|a| \geq |b|$. We proceed with the proof by decomposing Condition (iv) into two cases: $|b| \geq u \cdot \text{ufp}(a)$ and $|b| < u \cdot \text{ufp}(a)$.

Case 1: $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ and $|b| \geq u \cdot \text{ufp}(a)$. Note that Conditions (i) and (ii) are identical to their counterparts in Lemma 3. Because $|b| \geq u \cdot \text{ufp}(b)$ for all finite $b \in \mathbb{F}$, $|a| \geq |b|$ guarantees $|a| \geq u \cdot \text{ufp}(b)$. The assumptions $|a| \geq |b|$ and $|b| \geq u \cdot \text{ufp}(a)$ therefore jointly satisfy the conditions in Lemma 3, which subsequently ensures $\delta \in \mathbb{F}$.

Case 2: $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ and $|b| < u \cdot \text{ufp}(a)$. Since $|a| \geq |b|$ and $|b| < u \cdot \text{ufp}(a)$, $|a + b| < \text{succ}(|a|)$ due to Definition 3. From the bound on $|a+b|$, we derive $\text{ufp}(a+b) \leq \text{ufp}(a)$. Given the expected error bound $|\delta| < \text{ulp}(a+b) = 2u \cdot \text{ufp}(a+b)$ under faithful rounding, $|\delta|$ is thus strictly less than $2u \cdot \text{ufp}(a)$. Since a immediately satisfies $a \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$, we derive $\delta \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ through Condition (iv). Through Corollary 1, the bound $|\delta| < 2u \cdot \text{ufp}(a)$ and $\delta \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ collectively ensure $\delta \in \mathbb{F}$.

If $|b| > |a|$, a and b immediately satisfy Condition (iv). We can prove $\delta \in \mathbb{F}$ for $|b| > |a|$ by decomposing Condition (iii) into $|a| \geq u \cdot \text{ufp}(b)$ and $|a| < u \cdot \text{ufp}(b)$ as has been shown for Condition (iv).

□

Any operands that satisfy the conditions in Lemma 3 immediately satisfy those in Theorem 1. While we omit the proof for brevity, one can easily deduce $|a| \geq u \cdot \text{ufp}(b)$ and $|b| \geq u \cdot \text{ufp}(a)$ each imply $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$ and $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. It then follows that Theorem 1 guarantees $\delta \in \mathbb{F}$ for all nonzero $a, b \in \mathbb{F}$ such that $|e_a - e_b| \leq p$.

Due to Definition 1, the ulp of a given number x is bounded by $\text{ulp}(x) \geq 2u \cdot \text{ufp}(x)$. For any $x \in \mathbb{F}$ such that $2u^2 \cdot \text{ufp}(x) \geq \omega$, the term $2u^2 \cdot \text{ufp}(x)$ as used in Conditions (iii) and (iv) is effectively the ulp of x *if $\mathbb{F}$ had twice the available precision*. Consequently, the exponent difference for operands that satisfy these conditions could potentially exceed p. We highlight the implications of Theorem 1 through the following example.

Suppose b is exactly equal to $2u^2 \cdot \text{ufp}(a)$, thereby satisfying $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. If $b = 2u^2 \cdot \text{ufp}(a)$, then $\text{ufp}(b) = 2u^2 \cdot \text{ufp}(a)$. The latter equality ensures $2u^2 \cdot \text{ufp}(b) < 2u^2 \cdot \text{ufp}(a)$. Since a immediately satisfies $a \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$, $b = 2u^2 \cdot \text{ufp}(a)$ indicates $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$. Assuming $|a + b| \leq \Omega$, the operands in this example meet all conditions in Theorem 1. Since $u = 2^{-p}$, $\text{ufp}(b) = 2u^2 \cdot \text{ufp}(a)$ implies $\text{ufp}(b) = 2^{1-2p} \cdot \text{ufp}(a)$. The exponent difference $e_a - e_b$ in this example is thus $2p-1$, which is nearly double the bound enforced by Lemma 3. Theorem 1's conditions are thus significantly less restrictive than those in Lemma 3 and are applicable to a wider domain.

Theorem 1's conditions also ensure that the real sum $a+b$ is exactly representable in $2p$-bits of precision. Suppose $|a| > |b|$ and $e_a - e_b = p + k$, where $0 \leq k \leq p - 1$. In this example, Condition (iv) would ensure that b has at least k trailing zeroes. Consequently, b will have at most $p - k$ effective bits (*i.e.*, available precision minus trailing zeroes). Hence, $a + b$ would require at most $e_a - e_b + (p - k) = 2p$ bits to represent.

We now explore conditions that guarantee FastTwoSum is an EFT. As described in Section II, $x - a \in \mathbb{F}$ ensures $z =$

$\circ_2(x-a) = x-a$ and $y = \circ_3(a+b-x) = \circ_3(\delta)$. Therefore, $\delta \in \mathbb{F}$ is sufficient for $x+y = a+b$ when $x-a \in \mathbb{F}$. We establish properties that guarantee $x+y = a+b$ by combining the respective requirements for $x-a \in \mathbb{F}$ and $\delta \in \mathbb{F}$.

Theorem 2. *Let $p \geq 2$, $x = \circ_1(a+b)$, $z = \circ_2(x-a)$, and $y = \circ_3(b-z)$. If all conditions*

(i) $\circ_1, \circ_2, \circ_3 \in$ FR
(ii) $|a+b| \leq \Omega$
(iii) $a \in \text{ulp}(b)\mathbb{Z}$
(iv) $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$

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

Proof. As mentioned in Section II, Condition (iii) guarantees $x-a \in \mathbb{F}$ and $z = \circ_2(x-a) = x-a$ for all $\circ_1, \circ_2 \in$ FR. Because the definitions of ufp and ulp imply $\text{ulp}(b) \geq 2u^2 \cdot$ ufp(b), any operands that meet Condition (iii) implicitly satisfy $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$ (*i.e.*, Condition (iii) in Theorem 1). All $a, b \in \mathbb{F}$ that meet Conditions (i) through (iv) therefore satisfy the conditions in Theorem 1, which confirms $\delta \in \mathbb{F}$ is valid for all $\circ_1 \in$ FR. If $z = x-a$ and $\delta = a+b-x \in \mathbb{F}$, then $y = \circ_3(a+b-x) = a+b-x$ for all $\circ_3 \in$ FR. Conditions (i) through (iv) thus suffice to conclude $x+y = a+b$. □

Conditions (iii) and (iv) each impose a lower bound on the magnitude of a and b respectively when one is less than the other. If $|a| \geq |b|$, a and b immediately satisfy Condition (iii). When $|a| < |b|$, Condition (iii) requires any nonzero $|a|$ to be no less than $\text{ulp}(b)$. Since $\text{ulp}(b) \geq 2u \cdot \text{ufp}(b)$, any nonzero operands that meet Condition (iii) are subject to the bound $e_b - e_a \leq p-1$. Similarly, Condition (iv) requires that $|b|$ satisfies $|b| \geq 2u^2 \cdot \text{ufp}(a)$ when $|a| > |b|$. The nonzero operands that meet Condition (iv) thus satisfy $e_a - e_b \leq 2p-1$. Theorem 2 thus maintains the *maximum possible exponent difference* permissible under Theorem 1 while imposing a stricter upper bound on $e_b - e_a$ to ensure $x-a \in \mathbb{F}$.

IV. EFT Conditions for Round-to-Odd

For any $r \in \mathbb{R}$, Definition 4 (see Section II) ensures that $x = \text{RO}(r)$ has an even significand only if $r \in \mathbb{F}$. Assuming $p \geq 2$, the definition $\Omega = (2^p - 1) \times 2^{E_{max}-p+1}$ implies that the maximum magnitude finite elements in $\mathbb{F}$ have odd significands. For all finite $a, b \in \mathbb{F}$ such that $|a+b| > \Omega$, $a+b$ is not an element of $\mathbb{F}$. RO would thus round such sums to a neighbor $x \in \mathbb{F}$ for which $|x| = \Omega$. RO effectively enforces *saturation*, thereby preventing overflow for all sums $x = \text{RO}(a+b)$. Leveraging this property, we present a method to determine $\delta = a+b-\text{RO}(a+b) \in \mathbb{F}$ based on $|a+b|$.

Lemma 4. *Let $p \geq 2$, $a, b \in \mathbb{F} \setminus \{\pm\infty\}$, and $x = \text{RO}(a+b)$. If $|a+b| > \Omega$, then $\delta = a+b-x \in \mathbb{F}$.*

Proof. If $|a+b| > \Omega$, then $|x| = \Omega$ due to saturation. Without loss of generality, we assume $|a| \geq |b|$. The assumptions $|a+b| > \Omega$ and $|a| \geq |b|$ imply $a \times b > 0$ and $\text{ufp}(a) = 2^{E_{max}}$ as $|a+b| \leq \Omega$ otherwise. Since $|a| \geq |b|$ implies $\text{ufp}(a) \geq \text{ufp}(b)$, we derive $a, b \in 2u \cdot \text{ufp}(b)\mathbb{Z}$. It then follows that $\delta \in 2u \cdot \text{ufp}(b)\mathbb{Z}$.

Since RO $\in$ FR and $|x|, |a+b| > 0$, it immediately follows that $x \cdot (a+b) > 0$. Given $|a| \leq \Omega$ (*i.e.*, a is finite), $|a+b| > \Omega$, $|x| = \Omega$, and $x \cdot (a+b) > 0$, it then follows that $|\delta| = |a+b-x| \leq |a+b| - |x| \leq |a| + |b| - \Omega \leq |b|$. From $|\delta| \leq |b|$, we subsequently derive $|\delta| < 2\text{ufp}(b)$. Due to Corollary 1, $\delta \in 2u \cdot \text{ufp}(b)\mathbb{Z}$ and $|\delta| < 2\text{ufp}(b)$ certify $\delta \in \mathbb{F}$. □

Lemma 4 signifies that the rounding error of $\circ(a+b)$ is in $\mathbb{F}$ when $|a+b| > \Omega$ and the rounding mode enforces saturation. The bound $|a+b| > \Omega$ is thus also sufficient for $\delta \in \mathbb{F}$ when $x = \text{RZ}(a+b)$. We note, however, saturation only guarantees x and δ are finite; it does not preserve all properties of faithful rounding expected for $|a+b| \leq \Omega$. Specifically, saturation does not guarantee $|\delta| < \text{ulp}(a+b)$.

Since RO $\in$ FR, the conditions presented in Section III directly apply to RO. Theorem 1 provides the relevant conditions for determining $\delta \in \mathbb{F}$ when $|a+b| \leq \Omega$. Combining Theorem 1 and Lemma 4, we derive the following conditions for RO that are independent of the magnitude of $a+b$.

Corollary 2. *Let $p \geq 2$, $a, b \in \mathbb{F} \setminus \{\pm\infty\}$, and $x = \text{RO}(a+b)$. If both conditions*

(i) $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$
(ii) $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$

are satisfied, then $\delta = a+b-x \in \mathbb{F}$.

For directed rounding modes, the rounding direction for a given $r \in \mathbb{R}$ is either fixed (*e.g.*, RD and RU) or determined by the sign of r (*e.g.*, RZ). RO, on the other hand, determines the rounding direction for a number based on *the parity of its neighbors' integral significands*. RO's dependence on the parity of elements in $\mathbb{F}$ induces the following property.

Lemma 5. *Let $p \geq 2$, $r \in \mathbb{R}$, and $x = M_x \times 2^{E_x-p+1} \in \mathbb{F} \setminus \{\pm\infty\}$. If both conditions*

(i) M_x is an odd integer
(ii) $\text{pred}(x) < r < \text{succ}(x)$

are satisfied, then $x = \text{RO}(r)$.

Proof. We assume $r \neq x$, as the lemma is trivially true otherwise. We decompose the proof into two cases: $r < x$ and $r > x$. If $r < x$, Condition (ii) ensures $\text{pred}(x) < r < x$. Because no other numbers in $\mathbb{F}$ exist between $\text{pred}(x)$ and x, $\text{pred}(x)$ and x are equal to $\text{pred}(r)$ and $\text{succ}(r)$ respectively. Given Condition (i), $x = \text{RO}(r)$ immediately follows from Definition 4. Similarly, $r > x$ implies $x = \text{pred}(r) < r < \text{succ}(x) = \text{succ}(r)$. Condition (i) and Definition 4 also induce $x = \text{RO}(r)$ for such cases. □

We apply Lemma 5 to identify the rounding direction for RO sums when the addends' significands do not overlap (*i.e.*, $\min(|a|, |b|) < \text{ulp}(\max(|a|, |b|))$). Under such circumstances, the larger magnitude operand could be a floating-point neighbor of $a+b$. It would then follow that $r = a+b$ and $|x| = \max(|a|, |b|)$ satisfy Condition (ii). Based on these observations, we present a condition for $\delta \in \mathbb{F}$ that leverages the parity of the operands.

Lemma 6. *Let* $p \geq 2$, $a = M_a \times 2^{E_a - p+1} \in \mathbb{F} \setminus \{\pm\infty\}$, $b = M_b \times 2^{E_b - p+1} \in \mathbb{F} \setminus \{\pm\infty\}$, *and* $x = \text{RO}(a+b)$. *If* $M_{\max(|a|,|b|)}$ *is an odd integer, then* $\delta = a + b - x \in \mathbb{F}$.

Proof. If $|a+b| > \Omega$, $\delta \in \mathbb{F}$ holds due to Lemma 4. We thus only consider cases where $|a+b| \leq \Omega$. Without loss of generality, we assume $|a| \geq |b|$. Our assumptions thus imply $M_{\max(|a|,|b|)} = M_{|a|}$ and that M_a is an odd integer. We proceed with the proof using two cases: $|b| \geq \text{ulp}(a)$ and $|b| < \text{ulp}(a)$.

Case 1: $|b| \geq \text{ulp}(a)$. Since $\text{ulp}(a)$ is an integer power of 2, b satisfies $\text{ufp}(b) \geq \text{ulp}(a)$. Given $b \in u \cdot \text{ufp}(b)\mathbb{Z}$, it follows that $b \in u \cdot \text{ulp}(a)\mathbb{Z}$. Because $\text{ulp}(a) \geq 2u \cdot \text{ufp}(a)$, $b \in u \cdot \text{ulp}(a)\mathbb{Z}$ induces $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. Similarly, $|a| \geq |b|$ guarantees $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$. Therefore, a and b satisfy the conditions of Theorem 1, which confirms $\delta \in \mathbb{F}$.

Case 2: $|b| < \text{ulp}(a)$. If $p \geq 2$, all integer powers of 2 in $\mathbb{F}$ greater than ω have even integral significands. The same applies to their negatives. If M_a is an odd integer, either $|a| \neq \text{ufp}(a)$ or $|a| = \omega$ must hold. Due to Definition 3, the distance between a and its neighbors is thus $|\text{pred}(a) - a| = |\text{succ}(a) - a| = \text{ulp}(a)$. The assumption $|b| < \text{ulp}(a)$ would then indicate $\text{pred}(a) < a + b < \text{succ}(a)$. M_a being an odd integer and $\text{pred}(a) < a + b < \text{succ}(a)$ match the conditions of Lemma 5, which affirms $x = a$. If $x = a$, then $\delta = a + b - x = b$. It immediately follows that δ is in $\mathbb{F}$. □

Provided that the operand with the larger magnitude has an odd significand, Lemma 6 does not restrict the magnitude of the other operand. As such, the exponent difference of operands that adhere to Lemma 6 could span the entire dynamic range of $\mathbb{F}$. The lemma thereby deviates from its counterparts in Section II (see Lemma 3 and Theorem 1), which entail bounds on the operands' relative magnitudes. Additionally, Lemma 6 does not restrict the sign of the smaller magnitude operand as required for directed rounding modes.

Lemmas 4 and 6 each leverage RO's saturation and dependence on significand parity to ensure $\delta \in \mathbb{F}$. As a result, the lemmas provide guarantees for operands that may not satisfy the sufficient conditions for all faithful rounding modes. By applying Lemmas 4 and 6 in conjunction with exactness conditions for FastTwoSum's second operation (*i.e.*, $z = \circ_2(x-a)$), we produce the following criteria that ensure FastTwoSum is an EFT under RO.

Theorem 3. *Let* $p \geq 2$, $a = M_a \times 2^{E_a - p+1} \in \mathbb{F} \setminus \{\pm\infty\}$, *and* $b = M_b \times 2^{E_b - p+1} \in \mathbb{F} \setminus \{\pm\infty\}$. *Let* $x = \text{RO}(a+b)$, $z = \circ_2(x-a)$, *and* $y = \circ_3(b-z)$. *If all conditions*

(i) $\circ_2, \circ_3 \in \text{FR}$
(ii) $a \in \text{ulp}(b)\mathbb{Z}$
(iii) M_a *is an odd integer*

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

Proof. Condition (ii) ensures $x - a \in \mathbb{F}$, which then guarantees $z = x - a$ for all $\circ_2 \in \text{FR}$. It therefore suffices to prove $\delta = a+b-x \in \mathbb{F}$ as it guarantees $y = a+b-x$ for all $\circ_3 \in \text{FR}$. We thus prove δ is in $\mathbb{F}$ for all operands that meet Conditions (ii) and (iii) by separately analyzing the cases $|a| \geq |b|$ and $|a| < |b|$.

Case 1: $|a| \geq |b|$. If $|a| \geq |b|$, Condition (iii) implies $\max(|a|, |b|)$ has an odd significand. The operands thus satisfy the condition in Lemma 6, which affirms $\delta \in \mathbb{F}$.

Case 2: $|a| < |b|$. The assumption $|a| < |b|$ indicates $\text{ufp}(b) \geq \text{ufp}(a)$. From $\text{ufp}(b) \geq \text{ufp}(a)$, we derive $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. Because $\text{ulp}(b) \geq 2u^2 \cdot \text{ufp}(b)$, Condition (ii) induces $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$. The operands that satisfy both $|a| < |b|$ and Condition (ii) thus meet the conditions in Corollary 2, which confirms $\delta \in \mathbb{F}$. □

Theorem 3 enables for RO the use of FastTwoSum on operands such that an EFT is not guaranteed under directed rounding modes. For example, suppose $a = 2^p + 2$ and $b = -2^{-p}$. In this example, a is an integer multiple of $\text{ulp}(b) = 2^{1-2p}$ and has an odd significand (*i.e.*, $a = 2^p + 2 = (2^{p-1} + 1) \times 2$). The operands therefore satisfy the conditions in Theorem 3. However, b is not an integer multiple of $2u^2 \cdot \text{ufp}(a) = 2^{1-p}$ as required by Theorem 2. If $\circ_1 = \circ_2 = \circ_3 = \text{RZ}$, $x = \text{RZ}(a+b) = 2^p$ and $z = \text{RZ}(x-a) = -2$. As a result, $y = \text{RZ}(b-z) = 2 - 2^{1-p}$, and $x + y \neq a + b$. Under RO, FastTwoSum produces $x = 2^p + 2$, $z = 0$, and $y = -2^{-p}$. Hence, FastTwoSum under RO achieves $x + y = a + b$ as guaranteed by Theorem 3.

V. Applications of Round to Odd FastTwoSum

A common application of FastTwoSum is floating-point splitting: an EFT that splits $x \in \mathbb{F}$ across two numbers $x_h, x_\ell \in \mathbb{F}$ such that $x = x_h + x_\ell$. Dekker's splitting [1], for example, splits a p-precision number across two numbers that each have up to $\lfloor \frac{p}{2} \rfloor$ effective bits of precision. In [15], the authors present a more general splitting known as ExtractScalar, for which the distribution of bits is configurable.

ExtractScalar(σ, x) :
$s = \circ_1(\sigma + x)$
$x_h = \circ_2(s - \sigma)$
$x_\ell = \circ_3(x - x_h)$
return x_h, x_ℓ

ExtractScalar applies FastTwoSum on $x, \sigma \in \mathbb{F}$ such that $\sigma = 2^k$ for some $k \in \mathbb{Z}$ (*i.e.*, σ is an integer power of 2). It is shown in [15] that if $|x| \leq 2^k$ and $\circ_1, \circ_2, \circ_3 \in \text{RN}$, then $x_h \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ and $x = x_h + x_\ell$. If $x_h \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$, the number of effective bits in x_h is determined by the exponent difference between σ and x. Hence, ExtractScalar can adjustably distribute bits across x_h and x_ℓ by assigning an appropriate value of σ relative to x. Given a vector $v \in \mathbb{F}^n$, ExtractScalar can also produce an EFT for *all elements* v_i if σ is configured relative to $\max_i |v_i|$.

In [15], ExtractScalar performs all operations under RN while assuming overflow does not occur. Furthermore, the assumptions $\sigma = 2^k$ and $|x| \leq 2^k$ do not immediately satisfy the EFT conditions for FastTwoSum detailed in Section III. As

such, ExtractScalar may not be an EFT when $\circ_1 \notin$ RN. Specifically, an EFT is not guaranteed when the exponent difference between σ and x exceeds $2p-1$. Suppose $\circ_1, \circ_2, \circ_3 =$ RO and $x = 2^{k-2p}$. In this example, $s = \text{RO}(\sigma + x) = 2^k + 2^{k-p+1}$ and $x_h = \text{RO}(2^k + 2^{k-p+1} - 2^k) = 2^{k-p+1}$. Subsequently, $x_\ell = \text{RO}(2^{k-2p} - 2^{k-p+1}) = 2^{k-2p+1} - 2^{k-p+1}$ and $x \neq x_h + x_\ell$. Similar examples can be constructed for directed rounding modes (*e.g.*, $x = 2^{k-2p}$ for RU and $x = -2^{k-2p-1}$ for RD and RZ). Given these limitations, we present new conditions that ensure ExtractScalar is an EFT under RO.

Theorem 4. *Let $p \geq 2$, $x, \sigma \in \mathbb{F} \setminus \{\pm\infty\}$, and $k \in \mathbb{Z}$. Let $s = \text{RO}(\sigma + x)$, $x_h = \circ_2(s - \sigma)$, and $x_\ell = \circ_3(x - x_h)$. If all conditions*

(i) $\circ_2, \circ_3 \in$ FR
(ii) $\sigma = 2^k + \text{ulp}(2^k)$ *and* $2^k \geq 2\omega$
(iii) $|x| \leq 2^k$

are satisfied, then $x_h \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ and $x = x_h + x_\ell$.

Proof. Given $p \geq 2$ and Condition (ii), we derive $\text{ufp}(\sigma) = 2^k$ and $\text{ulp}(\sigma) = \text{ulp}(2^k)$. Through Conditions (ii) and (iii), we also derive $|x| \leq \text{ufp}(\sigma) < \sigma$. We first prove $x_h \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$. If $s, \sigma \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$, it must be the case that $x_h = \circ_2(s - \sigma) \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$. As $\sigma \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ by default, it suffices to show $s \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$. We split our proof into three cases: $x \geq 0$, $-2^{k-1} \leq x < 0$, and $x < -2^{k-1}$.

We further decompose $x \geq 0$ into two cases: $\sigma + x > \Omega$ and $\sigma + x \leq \Omega$. If $\sigma + x > \Omega$, then $s = \text{RO}(\sigma + x) = \Omega = (2^p - 1) \times 2^{E_{max} - p + 1}$ due to RO's saturation property. Because $|x| \leq \text{ufp}(\sigma)$, $\sigma + x > \Omega$ implies $\text{ufp}(\sigma) = 2^{E_{max}}$ and $\text{ulp}(\sigma) = 2^{E_{max} - p + 1}$. It then follows that $\text{ulp}(s) = \text{ulp}(\sigma)$. Since $s \in \text{ulp}(s)\mathbb{Z}$ and $\text{ulp}(s) = \text{ulp}(\sigma) > \frac{1}{2}\text{ulp}(\sigma)$, we conclude $s \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ when $\sigma + x > \Omega$. If $\sigma + x \leq \Omega$, $\text{ufp}(s) = \text{ufp}(\sigma + x)$ (*i.e.*, $e_s = e_{\sigma + x}$) for $s = \text{RO}(\sigma + x)$. Furthermore, $x \geq 0$ indicates $\sigma + x \geq \sigma$ and $\text{ufp}(\sigma + x) \geq \text{ufp}(\sigma)$. Given $\text{ufp}(s) = \text{ufp}(\sigma + x) \geq \text{ufp}(\sigma)$, Definition 1 ensures $\text{ulp}(s) = \text{ulp}(\sigma + x) \geq \text{ulp}(\sigma) > \frac{1}{2}\text{ulp}(\sigma)$. As $s \in \text{ulp}(s)\mathbb{Z}$, s thus satisfies $s \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ when $x + \sigma \leq \Omega$.

If $-2^{k-1} \leq x < 0$ and $s = \text{RO}(\sigma + x)$, then $\sigma + x \geq 2^{k-1} + \text{ulp}(2^k)$ and $\text{ufp}(s) = \text{ufp}(\sigma + x) \geq 2^{k-1} = \frac{1}{2}\text{ufp}(\sigma)$. As such, $\text{ulp}(s) \geq \frac{1}{2}\text{ulp}(\sigma)$ and $s \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$. Lastly, $x < -2^{k-1}$ implies $\text{ufp}(x) \geq 2^{k-1} = \frac{1}{2}\text{ufp}(\sigma)$ and $\text{ulp}(x) \geq \frac{1}{2}\text{ulp}(\sigma)$. Since $s = \text{RO}(\sigma + x)$ and $\sigma + x, s \in \min(\text{ulp}(\sigma), \text{ulp}(x))\mathbb{Z}$, it follows that $s \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$. Hence, $x_h \in \frac{1}{2}\text{ulp}(\sigma)\mathbb{Z}$ for all cases considered.

Due to Definition 3, $\sigma = 2^k + \text{ulp}(2^k)$ in Condition (ii) is equal to $\text{succ}(2^k)$. All integer powers of 2 in $\mathbb{F}$ greater than ω have even significands. Consequently, any σ that satisfies Condition (ii) must have an odd significand. Because $|x| < \sigma$, $\sigma \in \text{ulp}(x)\mathbb{Z}$. All $\sigma, x \in \mathbb{F}$ that satisfy these conditions thus meet the requirements of Theorem 3. Through Theorem 3, we derive $\sigma + x = s + x_\ell$. Because $\sigma \in \text{ulp}(x)\mathbb{Z}$ ensures $x_h = s - \sigma$ (see Section II), $\sigma + x = s + x_\ell$ induces $x = s - \sigma + x_\ell = x_h + x_\ell$ as was to be shown. □

If the requirements of Theorem 4 are met, ExtractScalar ensures error-free splitting under RO even when σ and x are non-overlapping (*i.e.*, $\text{ulp}(\sigma) > |x|$). Given a σ that satisfies the conditions in Theorem 4 for $x = \max_i |v_i|$ in a vector $v \in \mathbb{F}^n$, ExtractScalar can thus produce an EFT for all elements v_i under RO without requiring a lower bound on $\min_i |v_i|$.

VI. Related Work

In this section, we discuss prior work in the literature that have also explored EFT guarantees for FastTwoSum under various faithful rounding modes. Jeannerod and Zimmermann propose $a \in \text{ulp}(b)\mathbb{Z}$ and $e_a - e_b \leq p$ as sufficient conditions for ensuring FastTwoSum is an EFT under all faithful rounding modes [7, Theorem 2]. Operands that satisfy $e_a - e_b \leq p$ must adhere to the condition $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$. Accordingly, Theorem 2 of this paper ensures $x + y = a + b$ for any operands that meet the conditions from [7, Theorem 2]. Theorem 2 of this paper also guarantees EFTs for operands that would be excluded by the bound $e_a - e_b \leq p$ (*e.g.*, $a = 2^p$ and $b = 2^{-1}$). As such, our conditions are applicable to a larger set of operands than those addressed by previous works.

In [7, Theorems 9, 10, 11], Jeannerod and Zimmermann establish that when $a \in \text{ulp}(b)\mathbb{Z}$ but $e_a - e_b > p$, FastTwoSum under directed rounding modes produces a faithful rounding of $a + b$ for a format with twice the current available precision (*e.g.*, $x + y = \text{RZ}_{2p}(a + b)$). These theorems signify that FastTwoSum under directed rounding modes is an EFT if $a \in \text{ulp}(b)\mathbb{Z}$ and the real sum is exactly representable in $2p$-bits of precision. Our conditions in Theorem 2 ensure that $a + b$ is exactly representable in $2p$-bits and thus guarantee EFTs for directed rounding modes under the same circumstances as [7, Theorems 9, 10, 11].

For operands that adhere to Theorem 1, the exponent difference could be as large as $2p - 1$. Conditions with similar implications appear in [4]. In [4], Linnainmaa presents conditions for $\delta \in \mathbb{F}$ under RZ and RO using the notation h_y: the exponent of the *least significant nonzero bit* of $y \in \mathbb{F} \setminus \{0\}$. Alternatively, h_y is the *largest* integer k such that $y \in 2^k\mathbb{Z}$. Given this definition, all finite, nonzero $y \in \mathbb{F}$ satisfy $y \in 2^{h_y}\mathbb{Z}$ and $\text{ulp}(y) \leq 2^{h_y} \leq \text{ufp}(y)$. Moreover, if $|y| = \text{ufp}(y)$ (*i.e.*, $|y|$ is an integer power of 2), then $h_y = e_y$. Linnainmaa proves that if $x = \circ(a + b)$ for $\circ \in \{\text{RZ}, \text{RO}\}$, $e_x - h_{\min(|a|,|b|)} < 2p$ guarantees $\delta = a + b - x \in \mathbb{F}$. If $\text{ulp}(a) > |b|$ (*i.e.*, $|a| > |b|$ and the operands' significands do not overlap) and $a \times b > 0$, $e_x = e_a$ must hold under $\circ \in \{\text{RZ}, \text{RO}\}$ due to Definitions 1 and 3. The condition $e_x - h_{\min(|a|,|b|)} < 2p$ for such cases is thus equivalent to $e_a - h_{|b|} < 2p$. Since $h_{|b|} = h_b = e_b$ when $|b| = \text{ufp}(b)$, the example indicates that Linnainmaa's condition can tolerate up to a $2p - 1$ difference in exponents.

While Linnainmaa's condition provides comparable flexibility under RZ and RO, it lacks precision under RD and RU. For example, consider operands $a = 2^p - 1$ and $b = 2^{-p}$. Given $|a| \geq |b|$ and $2u^2 \cdot \text{ufp}(a) = 2^{-p}$, the operands satisfy $a \in 2u^2 \cdot \text{ufp}(b)\mathbb{Z}$ and $b \in 2u^2 \cdot \text{ufp}(a)\mathbb{Z}$ as required by Theorem 1. If $\circ = \text{RU}$, $x = \circ(a + b) = 2^p$ and $\delta = 2^{-p} - 1 \in \mathbb{F}$. In this example, $e_x = p$, $h_{\min(|a|,|b|)} = -p$, and $e_x - h_{\min(|a|,|b|)} = 2p$. As such, the example operands do not satisfy Linnainmaa's condition despite δ being representable in

$\mathbb{F}$. One can construct a similar example for RD by reversing the signs of a and b. Hence, Theorem 1 is applicable to a broader range of operands under RD and RU while addressing all faithful rounding modes including RZ and RO.

Akin to Lemma 6, Linnainmaa proposes $M_{\max(|a|,|b|)}$ having an odd significand as a sufficient condition for $\delta \in \mathbb{F}$ under RO. However, Linnainmaa's analysis assumes an unbounded model of $\mathbb{F}$ and does not address guarantees due to RO's saturation property (see Lemma 4). Furthermore, Linnainmaa proposes $|a| \geq |b|$ and a having an odd significand as EFT guarantees for FastTwoSum under RO. Instead of $|a| \geq |b|$, Theorem 3 enforces the less restrictive $a \in \text{ulp}(b)\mathbb{Z}$ to ensure the second operation of FastTwoSum is exact.

VII. Conclusion and Future Work

This paper identifies more general conditions than previously known in literature that ensure FastTwoSum is an EFT for all faithful rounding modes. Specifically, we identify new properties of operands that ensure the rounding error $\delta = a + b - \circ(a + b)$ is in $\mathbb{F}$. Our conditions enable EFTs for operands with large magnitude differences for all faithful rounding modes (even when the operands' exponent difference is nearly double the available precision). Hence, our EFT guarantees for FastTwoSum are applicable to a wide range of inputs, thereby offering improved precision over previously established conditions.

This paper also presents EFT conditions tailored to RO. We leverage RO's parity-based rounding behavior to identify sufficient requirements for both $\delta \in \mathbb{F}$ and $x + y = a + b$. We observe that when the larger operand has an odd significand, FastTwoSum can serve as an EFT under RO without restricting the magnitude or the sign of the smaller operand, which highlights RO's versatility over directed rounding modes. By analyzing a bounded floating-point model, we also examine RO's saturation property and its enabling of EFTs. We leverage these findings to identify conditions for error-free floating-point splittings under RO.

FastTwoSum-based splitting facilitates more sophisticated EFTs such as integer rounding [16], correctly rounded summation [15] [17], and accurate dot-products [18]. To leverage EFT guarantees, most algorithms based on FastTwoSum require RN as the default rounding mode. We expect our findings to aid the development of advanced EFTs for other standard rounding modes as well as emerging alternatives such as RO. Applying our conditions to EFTs for floating-point multiplication is also of interest. In a similar vein, extending our observations on rounding errors induced by floating-point addition to fused multiply-add operations is a potential future direction.

Acknowledgments

We thank the ARITH reviewers, Bill Zorn, and members of the Rutgers Architecture and Programming Languages (RAPL) lab for their feedback on this paper. This material is based upon work supported in part by the research gifts from the Intel corporation and the National Science Foundation with grants: 2110861 and 2312220. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Intel corporation or the National Science Foundation.

References

[1] T. J. Dekker, "A floating-point technique for extending the available precision," *Numer. Math.*, vol. 18, no. 3, p. 224–242, Jun. 1971. [Online]. Available: https://doi.org/10.1007/BF01397083

[2] D. Priest, "Algorithms for arbitrary precision floating point arithmetic," in *ARITH 1991*, 1991, pp. 132–143. [Online]. Available: https://doi.org/10.1109/ARITH.1991.145549

[3] Hida, Y. and Li, X. S. and Bailey, D. H., "Algorithms for quad-double precision floating point arithmetic," in *ARITH 2001*, 2001, pp. 155–162. [Online]. Available: https://doi.org/10.1109/ARITH.2001.930115

[4] S. Linnainmaa, "Analysis of some known methods of improving the accuracy of floating-point sums," *BIT*, vol. 14, no. 2, p. 167–202, Jun. 1974. [Online]. Available: https://doi.org/10.1007/BF01932946

[5] J. Demmel and H. D. Nguyen, "Fast reproducible floating-point summation," in *ARITH 2013*, ser. ARITH '13. USA: IEEE Computer Society, 2013, p. 163–172. [Online]. Available: https://doi.org/10.1109/ARITH.2013.9

[6] S. Boldo, S. Graillat, and J.-M. Muller, "On the robustness of the 2sum and fast2sum algorithms," *ACM Trans. Math. Softw.*, vol. 44, no. 1, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3054947

[7] C.-P. Jeannerod and P. Zimmermann, "FastTwoSum revisited," in *2025 IEEE 32nd Symposium on Computer Arithmetic (ARITH)*, 2025, pp. 141–148. [Online]. Available: https://doi.org/10.1109/ARITH64983.2025.00030

[8] G. Bohlender, W. Walter, P. Kornerup, and D. Matula, "Semantics for exact floating point operations," in *ARITH 1991*, 1991, pp. 22–26. [Online]. Available: https://doi.org/10.1109/ARITH.1991.145529

[9] M. Daumas, L. Rideau, and L. Théry, "A generic library for floating-point numbers and its application to exact computing," in *TPHOL 2001*, ser. TPHOLs '01. Berlin, Heidelberg: Springer-Verlag, 2001, p. 169–184. [Online]. Available: https://doi.org/10.1007/3-540-44755-5_13

[10] S. Boldo and M. Daumas, "Representable correcting terms for possibly underflowing floating point operations," in *ARITH 2003*, ser. ARITH '03. USA: IEEE Computer Society, 2003, p. 79. [Online]. Available: https://doi.org/10.1109/ARITH.2003.1207663

[11] IEEE, "754-2019 - IEEE Standard for Floating-Point Arithmetic," pp. 1–84, Jul. 2019. [Online]. Available: https://doi.org/10.1109/IEEESTD.2019.8766229

[12] S. Boldo and G. Melquiond, "When double rounding is odd," in *17th IMACS World Congress*, Paris, France, Jul. 2005, p. 11. [Online]. Available: https://inria.hal.science/inria-00070603

[13] Russinoff, D. M, *Formal Verification of Floating-Point Hardware Design: A Mathematical Approach*, 2nd ed. Springer, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-87181-9

[14] J. P. Lim and S. Nagarakatte, "One polynomial approximation to produce correctly rounded results of an elementary function for multiple representations and rounding modes," *Proc. ACM Program. Lang.*, vol. 6, no. POPL, Jan. 2022. [Online]. Available: https://doi.org/10.1145/3498664

[15] S. Rump, T. Ogita, and S. Oishi, "Accurate Floating-Point Summation Part I: Faithful Rounding," *SIAM Journal on Scientific Computing*, vol. 31, no. 1, pp. 189–224, 2008. [Online]. Available: https://doi.org/10.1137/050645671

[16] C.-P. Jeannerod, J.-M. Muller, and P. Zimmermann, "On various ways to split a floating-point number," in *ARITH 2018*, 2018, pp. 53–60. [Online]. Available: https://doi.org/10.1109/ARITH.2018.8464793

[17] S. Rump, T. Ogita, and S. Oishi, "Accurate Floating-Point Summation Part II: Sign, K-Fold Faithful and Rounding to Nearest," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 1269–1302, 2009. [Online]. Available: https://doi.org/10.1137/07068816X

[18] K. Ozaki, T. Ogita, S. Oishi, and S. Rump, "Error-free transformations of matrix multiplication by using fast routines of matrix multiplication and its applications," *Numer. Algorithms*, vol. 59, no. 1, p. 95–118, Jan. 2012. [Online]. Available: https://doi.org/10.1007/s11075-011-9478-1