# A radix-independent error analysis of the Cornea-Harrison-Tang method

Claude-Pierre Jeannerod

# A radix-independent error analysis of the Cornea-Harrison-Tang method

Claude-Pierre Jeannerod*

**Abstract**

Assuming floating-point arithmetic with a fused multiply-add operation and rounding to nearest, the Cornea-Harrison-Tang method aims to evaluate expressions of the form $ab + cd$ with high relative accuracy. In this paper we provide a rounding error analysis of this method, which unlike previous studies is not restricted to binary floating-point arithmetic but holds for any radix $\beta$. We show first that an asymptotically optimal bound on the relative error of this method is $\frac{2\beta u+2u^2}{\beta-2u^2} = 2u+\frac{2}{\beta}u^2+O(u^3)$, where $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff in radix $\beta$ and precision $p$. Then we show that the possibility of removing the $O(u^2)$ term from this bound is governed by the radix parity and the tie-breaking strategy used for rounding: if $\beta$ is odd or rounding is *to nearest even*, then the simpler bound $2u$ is obtained, while if $\beta$ is even and rounding is *to nearest away*, then there exist floating-point inputs $a, b, c, d$ that lead to a relative error larger than $2u + \frac{2}{\beta}u^2 - 4u^3$. All these results hold provided underflows and overflows do not occur and under some mild assumptions on $\beta$ and $p$ satisfied by IEEE 754-2008 formats.

## 1 Introduction

Given four floating-point numbers $a, b, c, d$ the Cornea-Harrison-Tang method [1, p. 273] aims to evaluate

$$x = ab + cd$$

efficiently and accurately using the fused multiply-add operation. Writing RN to denote rounding to nearest, this method can be described as follows:

**algorithm** CHT$(a, b, c, d)$
  $p_1 := \mathrm{RN}(ab)$;  $p_2 := \mathrm{RN}(cd)$;
  $e_1 := \mathrm{RN}(ab - p_1)$;  $e_2 := \mathrm{RN}(cd - p_2)$;  // these two operations are exact.
  $r := \mathrm{RN}(p_1 + p_2)$;  $e := \mathrm{RN}(e_1 + e_2)$;
  $\widehat{x} := \mathrm{RN}(r + e)$;
  **return** $\widehat{x}$

One key feature of this algorithm is its use of the fused multiply-add operation to compute the rounding errors of the two multiplications *exactly* in the absence of underflow and overflow, so that $e_1 = ab - p_1$ and $e_2 = cd - p_2$. The rounded sum $e$ of these error terms is then added to the (possibly highly inaccurate) rounded sum $r$ of the two products in order to obtain an approximation

*Inria, LIP (CNRS, ENSL, Inria, UCBL), Université de Lyon

$\widehat{x}$ having a tiny relative error. Another attractive feature of this method is its *symmetry*, which ensures that $ab + cd$ and $cd + ab$ are approximated by the same quantity and thus makes it straightforward to provide implementations of complex floating-point multiplication that preserve commutativity.

The accuracy of the CHT algorithm has been studied extensively in radix 2: assuming $p$-bit floating-point numbers and an unbounded exponent range, Cornea, Harrison, and Tang [1, pp. 273–275] showed that the relative error $|\widehat{x} - x|/|x|$ is always in $O(u)$ with $u = 2^{-p}$ the unit roundoff; this result was refined recently by Muller [8], who derived the upper bound $2u + 7u^2 + 6u^3$ and found that $|\widehat{x} - x|/|x|$ can be as large as $2u - 7u^2 + O(u^3)$ for some values of $a, b, c, d$. In other words, in radix 2 the relative error of algorithm CHT is bounded by $2u + O(u^2)$, and this bound is *asymptotically optimal* in the sense that there are inputs for which the ratio error/(error bound) tends to one as $u$ tends to zero.

These results raise two questions, however, which we answer in this paper:

1. Does the bound $2u + O(u^2)$ hold beyond radix $\beta = 2$, that is, for $\beta > 2$ and $u$ equal to $\frac{1}{2}\beta^{1-p}$?

2. Is it possible to remove the quadratic term $O(u^2)$ and thus to bound the relative error simply by $2u$?

The first question is natural since the IEEE 754-2008 standard [3] specifies floating-point arithmetic not only for radix 2, but also for radix 10. Furthermore, although the techniques developed in [8] for $\beta = 2$ extend to $\beta > 2$, the resulting bound on $|\widehat{x} - x|/|x|$ would be larger than $\frac{3\beta+4}{\beta+4}u$ and thus larger than $2.2u$ when $\beta \geqslant 6$.

The second question is motivated by the rounding error analysis of another method for evaluating $ab + cd$ with a fused multiply-add, namely Kahan's algorithm [2, p. 60]. Kahan's algorithm computes only one product and its error term (say, $p_1$ and $e_1$), then handles the other product directly by using the fused multiply-add operation $r = \mathrm{RN}(p_1 + cd)$, and finally returns $\mathrm{RN}(r + e_1)$. This approach thus saves three floating-point operations compared with algorithm CHT and, as shown in [5], it admits $2u$ as an asymptotically optimal bound on its relative error. However, this comes at the price of a lack of symmetry, and it is therefore important to understand whether this simple $O(u^2)$-free bound $2u$ can still be achieved in the symmetrized version provided by algorithm CHT.

**Main results.** Our first contribution is to answer the first question above positively, by proving that the bound $2u + O(u^2)$ holds for $p \geqslant 6$; this extends in particular the result of [8] to the practical case $\beta = 10$. Our second contribution is to show that, perhaps surprisingly, the answer to the second question depends on the parity of $\beta$ and the way RN breaks ties: in some cases (say, when $\beta$ is odd or ties are rounded *to even*), the bound $2u + O(u^2)$ can be replaced by $2u$, while in other cases the $O(u^2)$ term cannot be removed.

More precisely, we shall work under the following customary assumptions (all of which are implicitly or explicitly used for the analyses in radix 2 given in [1, 8]), and establish Theorems 1 and 2 below. Here and hereafter $a, b, c, d$

are taken from a set $\mathbb{F}$ of finite floating-point numbers in base $\beta$ and precision $p$. We assume that

$$\beta \geqslant 2 \quad \text{and} \quad p \geqslant 2,$$

and that the exponent range of $\mathbb{F}$ is unbounded, so

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

We also assume that the exact result of every operation on some element(s) of $\mathbb{F}$ is rounded back to $\mathbb{F}$ using a round-to-nearest function RN satisfying the following properties: for all $t \in \mathbb{R}$,

$$\begin{aligned} |\mathrm{RN}(t) - t| &= \min_{s \in \mathbb{F}} |s - t|; \\ \mathrm{RN}(-t) &= -\mathrm{RN}(t); \\ \mathrm{RN}(\beta^i t) &= \beta^i \mathrm{RN}(t) \quad \text{for all } i \in \mathbb{Z}. \end{aligned}$$

The last two properties say that the way of breaking ties is independent of the sign and magnitude of the number being rounded. This assumption is in particular satisfied by roundTiesToEven and roundTiesToAway, the two specifications of rounding to nearest given in the IEEE 754-2008 standard: when $t$ is a *midpoint*, that is, a number halfway between two consecutive elements of $\mathbb{F}$, then roundTiesToEven requires that the significand $S$ of $\mathrm{RN}(t)$ is an even integer, while roundTiesToAway requires that $|S|$ is maximal. For example, writing

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

for the unit roundoff associated with $\mathbb{F}$ and RN, the midpoint $1 + u$ is rounded down to $1 \in \mathbb{F}$ by roundTiesToEven, and up to $1 + 2u \in \mathbb{F}$ by roundTiesToAway.

We can now state our main results more formally:

**Theorem 1.** *If $\beta^{p-1} \geqslant 24$ then the value $\widehat{x}$ computed by algorithm CHT satisfies*

$$\widehat{x} = x(1+\epsilon), \qquad |\epsilon| \leqslant \begin{cases} 2u & \text{if } \beta \text{ is odd or } \mathrm{RN}(1+u) = 1, \\ \frac{2\beta u + 2u^2}{\beta - 2u^2} & \text{otherwise.} \end{cases}$$

*Furthermore, these bounds on the relative error $|\epsilon|$ are asymptotically optimal.*

Since $\frac{2\beta u + 2u^2}{\beta - 2u^2} = 2u + \frac{2}{\beta}u^2 + O(u^3)$, this first result shows that the relative error of algorithm CHT is always bounded by $2u + O(u^2)$ and that the leading constant $2u$ is best possible as $u$ tends to zero. The next result shows that when $\beta$ is even and RN is so that the midpoint $1 + u$ is rounded up to $1 + 2u$, then the term $O(u^2)$ cannot, in general, be removed.

**Theorem 2.** *Assume $\beta$ is even and $\mathrm{RN}(1 + u) = 1 + 2u$. If $\beta = 2$ and $2^p + 1$ is not a Fermat prime, or if $\beta \neq 2$, then there exist $a, b, c, d$ in $\mathbb{F}$ for which the value $\widehat{x}$ computed by algorithm CHT satisfies*

$$\widehat{x} = x(1+\epsilon), \qquad |\epsilon| > 2u + \frac{2}{\beta}u^2 - 4u^3.$$

**Consequences for IEEE arithmetic.** When $\beta = 2$ Theorem 2 excludes values of $p$ such that $2^p + 1$ is a Fermat prime, that is, a prime number of the form $2^{2^q}+1$ with $q \in \mathbb{N}$. However, this is not a restriction in practice, since $2^p+1$ is known to be composite for any of the binary interchange formats specified by the IEEE 754-2008 standard; see [6]. Similarly, it is easily checked that the assumption $\beta^{p-1} \geqslant 24$ used in Theorem 1 is satisfied for all formats. Third, roundTiesToEven and roundTiesToAway imply $\mathrm{RN}(1+u) = 1$ and $\mathrm{RN}(1+u) = 1+2u$, respectively. Therefore, in the specific context of IEEE arithmetic Theorems 1 and 2 lead to the following conclusion:

**Corollary 1.** *Assume floating-point arithmetic as specified by the IEEE 754-2008 standard, with radix $\beta$ and unit roundoff $u$. Then, in the absence of underflow and overflow, algorithm CHT has a relative error bounded by $2u$ when* RN *is roundTiesToEven, and by $\frac{2\beta u+2u^2}{\beta-2u^2} = 2u + O(u^2)$ when* RN *is roundTiesToAway. Furthermore, for roundTiesToAway, the $O(u^2)$ term cannot be removed, since there exist floating-point numbers $a, b, c, d$ leading to a relative error larger than $2u + \frac{2}{\beta}u^2 - 4u^3$.*

**Outline of the paper and additional background.** The rest of this paper is devoted to the proof of Theorems 1 and 2. We begin in Section 2 by presenting the three main tools used to establish the upper bounds in Theorem 1. Then we give in Section 3 an outline of our proof of that theorem, showing that it mainly consists of analyzing separately several cases which depend on the features of the exact or rounded values of the products $ab$ and $cd$; overall, this case analysis leads to about ten different error bounds, whose detailed proofs are postponed to the appendix for readability. Finally, the lower bound given in Theorem 2 is established independently in Section 4.

A useful tool for our analyses will be the *unit in the first place* function [9], denoted by ufp and defined for $t \in \mathbb{R}$ by

$$\mathrm{ufp}(t) = \begin{cases} 0 & \text{if } t = 0, \\ \beta^{\lfloor \log_\beta |t| \rfloor} & \text{if } t \neq 0. \end{cases}$$

By definition of $\mathbb{F}$, RN, and $u$ we have the classical relations

$$|\mathrm{RN}(t) - t| \leqslant u\,\mathrm{ufp}(t) \leqslant u \min\{|t|, |\mathrm{RN}(t)|\} \qquad \text{for all } t \in \mathbb{R},$$

which lead in particular to the *standard models* $\mathrm{RN}(t) = t(1+\epsilon)$ with $|\epsilon| \leqslant u$ and $\mathrm{RN}(t) = t/(1+\epsilon')$ with $|\epsilon'| \leqslant u$. For $|\epsilon|$, the upper bound $u$ can in fact be replaced by the slightly smaller quantity

$$u_1 := \frac{u}{1+u},$$

giving the following *refined model*:

$$\mathrm{RN}(t) = t(1+\epsilon), \qquad |\epsilon| \leqslant u_1. \tag{1}$$

This refined bound appears for example in [7, p. 232] and its attainability was noted in [6] along with the attainability of the bound $|\epsilon'| \leqslant u$.

When using $|\epsilon| \leqslant u_1$ instead of $|\epsilon| \leqslant u$, the bound $2u+7u^2+O(u^3)$ obtained in [8] immediately becomes $2u + 5u^2 + O(u^3)$. However, this sharper bound is not enough for our purposes, since it assumes $\beta = 2$ and has a $O(u^2)$ term regardless of the tie-breaking strategy.

# 2 Tools for the proof of Theorem 1

## 2.1 Inequalities resulting from the refined model

Applying the refined model (1) to the operations in algorithm CHT, we deduce that there exist rational numbers $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5$ such that

$$\begin{aligned} p_1 &= ab(1+\epsilon_1), & |\epsilon_1| &\leqslant u_1, \\ p_2 &= cd(1+\epsilon_2), & |\epsilon_2| &\leqslant u_1, \\ e &= (e_1+e_2)(1+\epsilon_3), & |\epsilon_3| &\leqslant u_1, \\ r &= (p_1+p_2)(1+\epsilon_4), & |\epsilon_4| &\leqslant u_1, \\ \widehat{x} &= (r+e)(1+\epsilon_5), & |\epsilon_5| &\leqslant u_1. \end{aligned}$$

Recalling that $e_1 = ab - p_1$ and $e_2 = cd - p_2$, we have

$$x = p_1 + p_2 + e_1 + e_2$$

and, therefore,

$$\widehat{x} = x(1+\epsilon_4)(1+\epsilon_5) + (e_1+e_2)(\epsilon_3-\epsilon_4)(1+\epsilon_5). \tag{2}$$

On the other hand, the definition of $\epsilon_1$ and $\epsilon_2$ implies that

$$e_1 = -\epsilon_1 ab \qquad \text{and} \qquad e_2 = -\epsilon_2 cd.$$

Hence, for $x$ nonzero and with

$$K = \frac{|ab| + |cd|}{|ab + cd|}, \tag{3}$$

we arrive at the following inequalities:

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + \max\{|\epsilon_1|, |\epsilon_2|\} \cdot |\epsilon_3 - \epsilon_4|(1+\epsilon_5) \cdot K \\ &\leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + 2u_1^2(1+u_1) \cdot K. \end{aligned} \tag{4}$$

## 2.2 Range constraints resulting from large relative errors

In addition to the usual unit roundoff $u$ and to the quantity $u_1 = \frac{u}{1+u}$, the following generalization will prove very useful in the sequel:

$$u_k := \frac{u}{1+ku}, \qquad k \in \mathbb{R}_{\geqslant 0}.$$

In particular, for fixed $k$ we see that $u_k = u - ku^2 + O(u^3)$ as $u$ tends to zero.

Having defined $u_k$, we can now state the following two properties, which indicate the range constraints implied by large enough relative errors. Property 1 says that if rounding a real number $t$ yields a relative error that is larger than $u_k$, then $|t|$ is necessarily close enough to its ufp. This immediate property is then refined by Property 2, which exploits further the sign of the relative error in order to confine $|t|$ to unions of about $k/2$ intervals of width $O(u^2) \cdot \mathrm{ufp}(t)$.

**Property 1.** *Let $k \in \mathbb{R}_{>1}$. Then for $t \in \mathbb{R}_{\neq 0}$ we have the following implication:*

$$\frac{|\mathrm{RN}(t) - t|}{|t|} > u_k \qquad \Rightarrow \qquad 1 < \frac{|t|}{\mathrm{ufp}(t)} < 1 + ku.$$

*Proof.* The lower bound on $|t|/\mathrm{ufp}(t)$ follows from the fact that $t \notin \mathbb{F}$, and the upper bound follows from $u_k|t| < |\mathrm{RN}(t) - t| \leqslant u\,\mathrm{ufp}(t)$. □

**Property 2.** *Given $k \in \mathbb{R}_{>1}$, let $\ell = \lceil (k-1)/2 \rceil$ and define the half-open intervals*

$$I_j = \Big[1 + (2j-1)u, \frac{1+2ju}{1+u_k}\Big), \qquad j = 1, \ldots, \ell,$$

*and*

$$\widetilde{I}_j = \Big(\frac{1+2ju}{1-u_k}, 1 + (2j+1)u\Big], \qquad j = 0, \ldots, \ell - 1.$$

*Then for $t \in \mathbb{R}_{\neq 0}$ we have the following implications:*

*(i)* $$\frac{\mathrm{RN}(t) - t}{t} > u_k \qquad \Rightarrow \qquad \frac{|t|}{\mathrm{ufp}(t)} \in I_1 \cup I_2 \cup \cdots \cup I_\ell;$$

*(ii)* $$\frac{\mathrm{RN}(t) - t}{t} < -u_k \qquad \Rightarrow \qquad \frac{|t|}{\mathrm{ufp}(t)} \in \widetilde{I}_0 \cup \widetilde{I}_1 \cup \cdots \cup \widetilde{I}_{\ell-1}.$$

*Proof.* We can assume $t > 0$ and $\mathrm{ufp}(t) = 1$, so that $1 \leqslant t < \beta$ and $|\mathrm{RN}(t) - t| \leqslant u$. To prove (i), note first that since $\mathrm{RN}(t)$ is in $\mathbb{F}$ and larger than $t$, it has the form

$$\mathrm{RN}(t) = 1 + 2ju$$

for some integer $j \geqslant 1$. The assumption $\frac{\mathrm{RN}(t)-t}{t} > u_k$ is thus equivalent to

$$t < \frac{1+2ju}{1+u_k}. \tag{5a}$$

In addition, $|\mathrm{RN}(t) - t| \leqslant u$ implies $\mathrm{RN}(t) - t \leqslant u$, that is,

$$1 + (2j-1)u \leqslant t. \tag{5b}$$

Hence the expression for interval $I_j$ follows from the inequalities (5). Recalling from Property 1 that $t < 1 + ku$, we deduce from (5b) that the integer $j$ and the

real number $k$ satisfy $2j-1<k$, that is, $j \leqslant \lceil (k-1)/2 \rceil = \ell$. This concludes the proof of (i).

Let us now prove (ii). From $1 \leqslant \mathrm{RN}(t) \in \mathbb{F}$ it follows that $\mathrm{RN}(t) = 1+2ju$ for some integer $j \geqslant 0$. The assumption $\frac{\mathrm{RN}(t)-t}{t} < -u_k$ is then equivalent to

$$\frac{1+2ju}{1-u_k} < t. \tag{6a}$$

On the other hand, $|\mathrm{RN}(t)-t| \leqslant u$ implies $-u \leqslant \mathrm{RN}(t)-t$, that is,

$$t \leqslant 1+(2j+1)u. \tag{6b}$$

From the inequalities (6) we deduce the definition of interval $\widetilde{I}_j$. Finally, using again Property 1 we have $t < 1+ku$, which together with (6a) and $u_k = u/(1+ku)$ leads to $1+2ju < (1+ku)(1-u_k) = 1+(k-1)u$. The latter inequality is equivalent to $j \leqslant \ell - 1$, which concludes the proof. □

In practice, when analyzing the CHT algorithm we shall avoid using the unwieldy rational functions involved in the right endpoint of $I_j$ and the left endpoint of $\widetilde{I}_j$. Instead, it will be enough to consider the following simpler intervals $\mathcal{I}_j$ and $\widetilde{\mathcal{I}}_j$, defined by degree-2 polynomials in $u$: for $j = 1, \ldots, \ell$,

$$\mathcal{I}_j := \big[\alpha_j, \alpha_j + \epsilon_{j,k}\big), \qquad \alpha_j := 1+(2j-1)u, \qquad \epsilon_{j,k} := (k-2j+1)u^2$$

and, for $j = 0, \ldots, \ell-1$,

$$\widetilde{\mathcal{I}}_j := \big(\widetilde{\alpha}_j - \widetilde{\epsilon}_{j,k}, \widetilde{\alpha}_j\big], \qquad \widetilde{\alpha}_j := 1+(2j+1)u, \qquad \widetilde{\epsilon}_{j,k} := (k-2j-1)u^2.$$

Since $\ell$ is defined in Property 2 as $\ell = \lceil (k-1)/2 \rceil$, it is easily checked that

$$I_j \subset \mathcal{I}_j, \qquad \widetilde{I}_{j-1} \subset \widetilde{\mathcal{I}}_{j-1}, \qquad j = 1, \ldots, \ell.$$

## 2.3 Properties of floating-point products

Algorithm CHT is built upon the fact that for an unbounded exponent range, the rounding error of the product of two elements of $\mathbb{F}$ is always itself in $\mathbb{F}$. The next two properties show that if this error is nonzero then its ulp cannot be too small. Those properties (which are straightforward extensions to radix $\beta$ of those used in [8] for radix two) will be useful when dealing with the case where $p_1$ and $-p_2$ are so close to each other that $p_1 + p_2$ is computed exactly.

**Property 3.** *Let $i \in \mathbb{Z}$ and $a, b \in \mathbb{F}$ be such that $\beta^i \leqslant |ab| < \beta^{i+1}$. Then $ab - \mathrm{RN}(ab)$ is an integer multiple of $\beta^{i-2p+1}$.*

*Proof.* Writing $a = A\beta^{e_a-p+1}$ and $b = B\beta^{e_b-p+1}$ with $A, B$ two integers such that $\beta^{p-1} \leqslant |A|, |B| \leqslant \beta^p - 1$, we have $ab = AB\beta^{e_a+e_b-2p+2}$. Hence $ab$, $\mathrm{RN}(ab)$, and $ab - \mathrm{RN}(ab)$ are integer multiples of $\beta^{e_a+e_b-2p+2}$. Now, $i$ is either $e_a + e_b$ or $e_a + e_b + 1$, so $ab - \mathrm{RN}(ab)$ is always an integer multiple of

$\min\{\beta^{i-2p+2}, \beta^{i-2p+1}\} = \beta^{i-2p+1}$. □

The property above can be refined in the sense that either the error is an integer multiple of a larger quantity, or the product admits a smaller upper bound:

**Property 4.** *Let $i \in \mathbb{Z}$ and $a, b \in \mathbb{F}$ be such that $\beta^i \leqslant |ab| < \beta^{i+1}$. Then*

- *either $ab - \mathrm{RN}(ab)$ is an integer multiple of $\beta^{i-2p+2}$,*
- *or $|ab| \leqslant (1 - \frac{2u}{\beta})^2 \beta^{i+1}$.*

*Proof.* Using the same notation and reasoning as in the proof of Property 3, if $i = e_a + e_b$ then $ab - \mathrm{RN}(ab)$ is an integer multiple of $\beta^{i-2p+2}$. Else $i = e_a + e_b + 1$, but then $|ab| = |AB|\beta^{1-2p+i} \leqslant (\beta^p - 1)^2 \beta^{1-2p+i} = (1 - \frac{2u}{\beta})^2 \beta^{i+1}$. □

# 3 Proof outline for Theorem 1

The first step of the proof of the bounds in Theorem 1 consists of restricting the input set using symmetry arguments. Since $\mathrm{CHT}(a, b, c, d) = \mathrm{CHT}(c, d, a, b)$, we can exchange $ab$ and $cd$ to ensure $|cd| \leqslant |ab|$. On the other hand, using $\mathrm{RN}(-t) = -\mathrm{RN}(t)$ gives $\mathrm{CHT}(-a, b, -c, d) = -\mathrm{CHT}(a, b, c, d)$, so that we can also restrict further to $ab \geqslant 0$ and eventually assume that

$$|cd| \leqslant ab.$$

As a second step, we consider the situation where either $ab$ or $cd$ or $x$ is zero. In this case, by propagating the equality $\mathrm{RN}(0) = 0$ within the CHT algorithm, we see that $\widehat{x} = \mathrm{RN}(x)$ and, recalling the refined model in (1),

$$\widehat{x} = x(1 + \epsilon), \qquad |\epsilon| \leqslant u_1. \tag{7}$$

The third and main step of the proof is the analysis of the remaining cases, namely when $a, b, c, d$ are such that

$$ab > 0 \quad \text{and} \quad cd \neq 0 \quad \text{and} \quad -ab < cd \leqslant ab. \tag{8}$$

To derive the bounds in Theorem 1 for inputs as in (8), we shall analyze separately several subcases and, using the tools described in Section 2, obtain the nine bounds from (9) to (17) below. The rest of this section only gives an overview of those subcases and the associated bounds, the detailed proofs being deferred to Appendix.

## 3.1 Analysis when $ab$ and $cd$ have the same sign

In this case, the number $K = (|ab| + |cd|)/|ab + cd|$ introduced in (3) satisfies $K = 1$. If $\min\{\epsilon_4, \epsilon_5\} \leqslant u_3$ then, using (4), we easily obtain

$$\frac{|\widehat{x} - x|}{|x|} < 2u - u^2 + 4u^3; \tag{9}$$

else, using further Property 1, we deduce that

$$\frac{|\widehat{x} - x|}{|x|} < 2u - u^2 + u^4 \qquad \text{for } \beta^{p-1} \geqslant 3. \tag{10}$$

## 3.2 Analysis when $ab$ and $cd$ have opposite signs

### 3.2.1 When $\frac{1}{2}p_1 \leqslant |p_2|$

In this case the sum $p_1 + p_2$ is computed exactly thanks to Sterbenz's theorem [10], and a direct consequence of this will be that if $K \leqslant 1/u_1$ then the bound in (10) still applies here. If $K > 1/u_1$, then $\mathrm{ufp}(cd)$ turns out to be either $\mathrm{ufp}(ab)$ or $\beta^{-1}\mathrm{ufp}(ab)$, and we can handle those two cases separately: in the first case, by applying Property 3 to both $ab$ and $cd$, we can show that $e_1 + e_2$ is a floating-point number and then deduce that

$$\frac{|\widehat{x} - x|}{|x|} \leqslant u_1; \tag{11}$$

in the second case, applying Property 3 to $ab$ and Property 4 to $cd$ leads either to the same bound as in (11) or to the bound

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{3}{2}u. \tag{12}$$

### 3.2.2 When $\frac{1}{2}p_1 > |p_2|$

In this case, noting that $K$ is at most about 3, we shall consider separately four cases defined by the pair $(\epsilon_4, \epsilon_5)$. The first two subcases can be handled using this bound on $K$ together with (4): if $\epsilon_4$ and $\epsilon_5$ have opposite signs, then

$$\frac{|\widehat{x} - x|}{|x|} \leqslant u + 7u^2 \qquad \text{for } \beta^{p-1} \geqslant 4; \tag{13}$$

if $\min\{|\epsilon_4|, |\epsilon_5|\} \leqslant u_7$, then

$$\frac{|\widehat{x} - x|}{|x|} < 2u \qquad \text{for } \beta^{p-1} \geqslant 24. \tag{14}$$

The remaining two subcases, which correspond to $\min\{|\epsilon_4|, |\epsilon_5|\} > u_7$ with $\epsilon_4$ and $\epsilon_5$ either both positive or both negative, turn out to be more involved and their analysis relies on either part (i) or part (ii) of Property 2 with $k = 7$. When $\epsilon_4 > u_7$ and $\epsilon_5 > u_7$, we show the following:

- if $\beta$ is odd or $\mathrm{RN}(1 + u) = 1$, then

$$\frac{|\widehat{x} - x|}{|x|} < 2u \qquad \text{for } \beta^{p-1} \geqslant 10; \tag{15}$$

- else

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{2\beta u + 2u^2}{\beta - 2u^2} \qquad \text{for } \beta^{p-1} \geqslant 10. \tag{16}$$

When $\epsilon_4 < -u_7$ and $\epsilon_5 < -u_7$, we show that

$$\frac{|\widehat{x} - x|}{|x|} < 2u - u^2 \qquad \text{for } \beta^{p-1} \geqslant 10. \tag{17}$$

The two error bounds in Theorem 1 then follow directly from the intermediate bounds (7) and (9–17) shown above: these ten bounds require no more than $\beta^{p-1} \geqslant 24$; furthermore, all of them are less than $2u$, except the one in (16), which has the form $2u + \frac{2}{\beta}u^2 + O(u^3)$.

Finally, we know from [4, Corollary 4.2] that if $\beta^{p-1} \geqslant 11$, then there exist $a, b, c, d \in \mathbb{F}$ for which the CHT algorithm returns $\widehat{x}$ such that

$$\frac{|\widehat{x} - x|}{|x|} > 2u - 8u^{1.5} - 6u^2;$$

this proves the asymptotic optimality of the two error bounds in Theorem 1.

# 4 Proof of Theorem 2

The assumption on $\beta$ and $p$ implies that there exist $a, b \in \mathbb{F}$ such that

$$ab = 1 + u; \tag{18}$$

see [6, Theorem 3.2]. Hence, using the assumption that $\mathrm{RN}(1+u) = 1+2u$, we have

$$p_1 = 1 + 2u \qquad \text{and} \qquad e_1 = -u.$$

Define further

$$c = u + 2u^2 \qquad \text{and} \qquad d = -1 + \frac{\beta - 1}{\beta} \cdot 2u.$$

Recalling that $u = \frac{1}{2}\beta^{1-p}$, we have $c = C \cdot \beta^{1-2p}$ with $C = \frac{1}{2}\beta^p + \frac{\beta}{2}$ and $d = -D \cdot \beta^{-p}$ with $D = \beta^p - \beta + 1$. Since $C$ and $D$ are integers such that $\beta^{p-1} \leqslant C, D < \beta^p$, we deduce that $c$ and $d$ are in $\mathbb{F}$. In addition,

$$cd = -\Big(u + \frac{2}{\beta}u^2 - 4(1 - \frac{1}{\beta})u^3\Big), \tag{19}$$

which implies

$$u < |cd| < u + \frac{2}{\beta}u^2 = u + u\,\mathrm{ufp}(u)$$

and thus

$$p_2 = -u.$$

Consequently, $p_1 + p_2 = 1 + u$, which rounds to

$$r = 1 + 2u.$$

On the other hand, noticing that $e_1 = p_2$ we obtain $e = \mathrm{RN}(p_2 + e_2) = p_2$, that is,

$$e = -u.$$

Hence $r + e = 1 + u$, which rounds to

$$\widehat{x} = 1 + 2u.$$

Finally, we deduce from (18) and (19) that

$$x = 1 - \frac{2}{\beta}u^2 + 4(1 - \frac{1}{\beta})u^3,$$

which is such that $0 < x < \widehat{x}$. Thus, overall, $\frac{|\widehat{x}-x|}{|x|} = \frac{\widehat{x}}{x} - 1 = 2u\frac{\beta+u+(2-2\beta)u^2}{\beta-2u^2+(4\beta-4)u^3}$, and one can check that the latter quantity is larger than $2u + \frac{2}{\beta}u^2 - 4u^3$. This concludes the proof of Theorem 2.

# APPENDIX

## A Analysis when the two products have the same sign

When the products $ab$ and $cd$ have the same sign, the assumption in (8) implies that they are both positive and that $K$ in (3) is equal to one. Using (4), we deduce that

$$\frac{|\widehat{x} - x|}{|x|} \leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + 2u_1^2 + 2u_1^3.$$

Using the obvious inequality $|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant 2u_1 + u_1^2$ is not enough, as this would yield the bound $|\widehat{x} - x|/|x| \leqslant 2u_1 + 3u_1^2 + 2u_1^3 = 2u + u^2 - 2u^3 + O(u^4)$, which is slightly larger than $2u$ for any radix and tie-breaking rule. Instead, we consider two sub-cases separately, as follows.

### A.1 Case where $\epsilon_4 \leqslant u_3$ or $\epsilon_5 \leqslant u_3$

In this case $\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5 = (1 + \epsilon_4)(1 + \epsilon_5) - 1$ satisfies

$$-2u_1 + u_1^2 \leqslant \epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5 \leqslant u_1 + u_3 + u_1u_3.$$

Since $u_1 + u_3 + u_1u_3 \geqslant 2u_1 - u_1^2$, we have

$$|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant u_1 + u_3 + u_1u_3,$$

which gives

$$\begin{aligned}\frac{|\widehat{x}-x|}{|x|} &\leqslant u_1 + u_3 + u_1u_3 + 2u_1^2 + 2u_1^3\\ &< 2u - u^2 + 4u^3\end{aligned}$$

and thus proves (9).

## A.2 Case where $\epsilon_4 > u_3$ and $\epsilon_5 > u_3$

Since both $ab$ and $cd$ are positive and since the exponent range of $\mathbb{F}$ is unbounded, we have $p_1 + p_2 > 0$. Thus, applying Property 1 with $k = 3$ gives

$$\beta^i < p_1 + p_2 < (1+3u)\beta^i, \qquad \beta^i = \mathrm{ufp}(p_1 + p_2).$$

Since $\epsilon_4$ is positive, rounding to nearest coincides here with rounding up, and the rounded sum $r = \mathrm{RN}(p_1 + p_2)$ must be

$$r = (1+2u)\beta^i.$$

Let us now bound $|e|$. We have $|e| \leqslant (1+u_1)\big(|e_1| + |e_2|\big) \leqslant (u_1 + u_1^2)x$. Furthermore, it follows from $r = ab(1+\epsilon_1)(1+\epsilon_4) + cd(1+\epsilon_2)(1+\epsilon_4)$ that

$$(1-u_1)^2 x \leqslant r \leqslant (1+u_1)^2 x. \tag{20}$$

Using the lower bound in (20) thus leads to

$$\begin{aligned}|e| &\leqslant \frac{u_1 + u_1^2}{(1-u_1)^2} r = (1+2u)ur\\ &< 2u\beta^i \qquad \text{for } \beta^{p-1} \geqslant 3.\end{aligned}$$

Consequently,

$$\beta^i < r + e < (1+4u)\beta^i \leqslant \beta^{i+1}.$$

Now, the assumption $\epsilon_5 > u_3$ implies that when rounding $r + e$ to nearest then rounding up occurs and, by Property 1, that $r + e$ must be less than $(1+3u)\beta^i$. In other words,

$$(1+u)\beta^i \leqslant r + e < (1+2u)\beta^i$$

and

$$\widehat{x} = (1+2u)\beta^i.$$

Therefore, $\widehat{x} = r$ and, using (20), we conclude that

$$\begin{aligned}\frac{|\widehat{x}-x|}{|x|} = \frac{|r-x|}{|x|} &\leqslant 2u_1 + u_1^2\\ &< 2u - u^2 + u^4 \qquad \text{for } \beta^{p-1} \geqslant 3.\end{aligned} \tag{21}$$

This proves (10) and concludes the analysis of the case where the two products $ab$ and $cd$ have the same sign.

# B Analysis when the two products have opposite signs

When $ab$ and $cd$ have opposite signs, the assumption in (8) can be rewritten

$$-ab < cd < 0 < ab. \tag{22}$$

For $K$ as in (3), this implies

$$K = \frac{ab + |cd|}{ab - |cd|} \tag{23}$$

and, rounding being monotonic,

$$-p_1 \leqslant p_2 < 0 < p_1.$$

Note that neither $p_1$ nor $p_2$ can be zero (because the exponent range of $\mathbb{F}$ is unbounded) and that $p_1 + p_2 \geqslant 0$.

## B.1 When $\frac{1}{2}p_1 \leqslant |p_2|$

In this case, the two floating-point numbers $p_1$ and $-p_2$ satisfy $\frac{1}{2}p_1 \leqslant -p_2 \leqslant p_1$, so that for any radix $\beta$ the sum $p_1 + p_2$ is computed exactly by Sterbenz's theorem [10, p. 138] (see also [2, p. 45]). Hence $\epsilon_4 = 0$ and, using (2),

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant |\epsilon_5| + \frac{|(e_1 + e_2)\epsilon_3|}{|x|}(1 + \epsilon_5) \\ &\leqslant u_1 + u_1^2(1 + u_1)K. \end{aligned} \tag{24}$$

If $K \leqslant 1/u_1$ then we deduce immediately from the latter bound that the relative error on $x$ is bounded by $2u_1 + u_1^2$ and thus as in (21). Hence the rest of this section is devoted to handling the case

$$K > \frac{1}{u_1} = \frac{1}{u} + 1. \tag{25}$$

(This case of a huge value of $K$ does occur, for example when $a = 1 + 2u$, $b = 1 - u$, $c = 1$, and $d = -1$.)

From (22), (23), and (25) we deduce that

$$\frac{1}{\beta}ab \leqslant \frac{1}{1 + 2u}ab < |cd| < ab.$$

Consequently, if $ab \in [\beta^i, \beta^{i+1})$ with $i \in \mathbb{Z}$, then $|cd|$ is either in $[\beta^{i-1}, \beta^i)$ or in $[\beta^i, \beta^{i+1})$. This yields two sub-cases which we handle separately.

### B.1.1 Case $\beta^i \leqslant |cd| < ab < \beta^{i+1}$

In this case, $\mathrm{ufp}(ab) = \mathrm{ufp}(cd) = \beta^i$ and, using $|e_1| \leqslant u\,\mathrm{ufp}(ab)$ and $|e_2| \leqslant u\,\mathrm{ufp}(cd)$, we deduce that

$$|e_1 + e_2| \leqslant 2u\beta^i = \beta^{i-p+1}.$$

On the other hand, Property 3 implies the existence of integers $E_1$, $E_2$ such that

$$e_1 = E_1 \cdot \beta^{i-2p+1}, \qquad e_2 = E_2 \cdot \beta^{i-2p+1}.$$

Hence $e_1 + e_2 = (E_1 + E_2)\beta^{i-2p+1}$ and the integer $E_1 + E_2$ must satisfy $|E_1 + E_2| \leqslant \beta^p$. This means that $e_1 + e_2 \in \mathbb{F}$ or, equivalently, $\epsilon_3 = 0$. It then follows from (24) that $|\widehat{x} - x|/|x| \leqslant |\epsilon_5| \leqslant u_1$, which proves (11).

### B.1.2 Case $\beta^{i-1} \leqslant |cd| < \beta^i \leqslant ab < \beta^{i+1}$

We now have $\mathrm{ufp}(ab) = \beta^i$ and $\mathrm{ufp}(cd) = \beta^{i-1}$, so that

$$|e_1 + e_2| \leqslant u\beta^i + u\beta^{i-1} < 2u\beta^i. \tag{26}$$

Property 3 still gives

$$e_1 = E_1 \cdot \beta^{i-2p+1}$$

for some integer $E_1$, and by applying Property 4 to the product $cd$ we have either

$$e_2 = E_2 \cdot \beta^{i-2p+1}$$

for some integer $E_2$, or

$$|cd| \leqslant \left(1 - \frac{2u}{\beta}\right)^2 \beta^i.$$

We handle these two situations independently as follows.

■ If $e_2 = E_2 \cdot \beta^{i-2p+1}$, then $|e_1 + e_2| = |E_1 + E_2|\beta^{i-2p+1}$. Using (26) gives $|E_1 + E_2| < \beta^p$, from which we deduce $e_1 + e_2 \in \mathbb{F}$, that is, $\epsilon_3 = 0$. We conclude as in Section B.1.1 that $|\widehat{x} - x|/|x| \leqslant u_1$.

■ Assume now that $|cd| \leqslant \left(1 - \frac{2u}{\beta}\right)^2 \beta^i$. Then

$$|x| = ab - |cd| \geqslant \beta^i - \left(1 - \frac{2u}{\beta}\right)^2 \beta^i.$$

On the other hand, the strict inequality in (26) implies $\mathrm{ufp}(e_1 + e_2) \leqslant \beta^{i-p}$ and, since $\mathrm{RN}(e_1 + e_2) = (e_1 + e_2)(1 + \epsilon_3)$, we obtain

$$|(e_1 + e_2)\epsilon_3| \leqslant u\beta^{i-p}.$$

Applying (24) thus gives $|\widehat{x} - x|/|x| \leqslant u_1 + (1 + u_1)\varphi$ with

$$\varphi := \frac{|(e_1 + e_2)\epsilon_3|}{|x|} \leqslant \frac{u\beta^{-p}}{\frac{4u}{\beta} - \frac{4u^2}{\beta^2}}.$$

Recalling that $u = \frac{1}{2}\beta^{1-p}$, it is easily checked that $\varphi \leqslant \frac{u}{2-2u/\beta} \leqslant \frac{u}{2-u}$ for $\beta \geqslant 2$, so that

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{3}{2}u,$$

which is the bound claimed in (12).

## B.2 When $\frac{1}{2}p_1 > |p_2|$

We begin by noting that in this case the ratio $K = \frac{ab+|cd|}{ab-|cd|}$ is at most about 3: since $(1+u_1)ab \geqslant p_1$ and $|p_2| \geqslant (1-u_1)|cd|$, the assumption $\frac{1}{2}p_1 > |p_2|$ implies

$$\psi := \frac{ab}{|cd|} > \frac{1-u_1}{\frac{1}{2}(1+u_1)}$$

and, since $K = 1 + \frac{2}{\psi-1}$, this is equivalent to

$$K < \frac{3-u_1}{1-3u_1} = 3 + 8u + O(u^2). \tag{27}$$

Combining (27) and (4), we obtain

$$\frac{|\widehat{x} - x|}{|x|} \leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + 2u_1^2(1+u_1)\frac{3-u_1}{1-3u_1}. \tag{28}$$

For the same reason as in Section A, applying to (28) the straightforward inequality $|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant 2u_1 + u_1^2$ is not enough for our purpose, since the resulting relative error bound would then have the form $2u + 5u^2 + O(u^3)$. In order to achieve the sharper bounds claimed in Theorem 1, we shall refine further this analysis by examining separately four cases defined by the pair $(\epsilon_4, \epsilon_5)$.

### B.2.1 Case where $\epsilon_4$ and $\epsilon_5$ have opposite signs

In this case $|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant u_1$ and it follows from (28) that

$$\frac{|\widehat{x} - x|}{|x|} \leqslant u \cdot \frac{1 + 6u + 13u^2 + 6u^3}{(1-2u)(1+u)^3}.$$

It is then easily checked that this bound is at most $u + 7u^2$ when $u \leqslant 1/8$ or, equivalently, when $\beta^{p-1} \geqslant 4$, which proves (13).

### B.2.2 Case where $|\epsilon_4| \leqslant u_7$ or $|\epsilon_5| \leqslant u_7$

In this case $|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant u_1 + u_7 + u_1u_7$, and applying (28) then leads to

$$\begin{aligned}\frac{|\widehat{x} - x|}{|x|} &\leqslant 2u \cdot \frac{1 + \frac{15}{2}u + 26u^2 + \frac{89}{2}u^3 + 19u^4}{(1-2u)(1+7u)(1+u)^3} = 2u - u^2 + O(u^3) \\ &< 2u \quad \text{for } \beta^{p-1} \geqslant 24,\end{aligned}$$

which proves (14).

### B.2.3 Case where $\epsilon_4 > u_7$ and $\epsilon_5 > u_7$

In this case, we will derive the bounds in (15) and (16), depending on radix parity and the tie-breaking strategy of rounding to nearest. Defining the condition

$$(\mathrm{C}): \qquad \beta \text{ is odd} \qquad \text{or} \qquad \mathrm{RN}(1+u) = 1,$$

our goal in this section is thus to show that for $\beta^{p-1} \geqslant 10$,

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \begin{cases} 2u - \eta \text{ for some } \eta > 0 & \text{if (C) holds,} \\ \frac{2\beta u + 2u^2}{\beta - 2u^2} = 2u + \frac{2}{\beta}u^2 + O(u^3) & \text{otherwise.} \end{cases} \tag{29}$$

To establish (29), we shall apply Property 2 in order to obtain suitable ranges for the exact sum $p_1 + p_2$ from which we can then deduce some values for $\widehat{x}$ together with some ranges for $x$, and eventually some bounds on $|\widehat{x} - x|/|x|$.

**Preliminaries.** Since $\epsilon_4$ is nonzero, $p_1 + p_2$ is not in $\mathbb{F}$ (and thus nonzero as well), so there exists an integer $i$ such that $\beta^i < p_1 + p_2 < \beta^{i+1}$. In order to simplify the expressions used in the sequel, we shall assume that $i = 0$ (which is possible up to a scaling by an integer power of the base $\beta$ and because the exponent range of $\mathbb{F}$ is unbounded). Therefore,

$$1 < p_1 + p_2 < \beta.$$

Since $\epsilon_4 > u_7$, applying part (i) of Property 2 with $k = 7$ gives

$$p_1 + p_2 \in \underbrace{[1+u, 1+u+6u^2)}_{=:\mathcal{I}_1} \cup \underbrace{[1+3u, 1+3u+4u^2)}_{=:\mathcal{I}_2} \cup \underbrace{[1+5u, 1+5u+2u^2)}_{=:\mathcal{I}_3}.$$

Furthermore, $\epsilon_4 > 0$ implies that $r = \mathrm{RN}(p_1 + p_2)$ satisfies

$$r > p_1 + p_2.$$

From this strict inequality and the range of $p_1 + p_2$ shown above, we deduce that

$$r \in \{1 + 2u, 1 + 4u, 1 + 6u\}. \tag{30}$$

Furthermore, the right endpoint of $\mathcal{I}_3$ leads to the following bound on $|e|$: since $\frac{1}{2}p_1 > |p_2|$ by assumption, we have $p_1 + |p_2| < 3(p_1 + p_2)$ and thus

$$\begin{aligned} |e| &\leqslant (1+u_1)\big(|e_1| + |e_2|\big) \\ &\leqslant u(1+u_1)\big(p_1 + |p_2|\big) \\ &< 3u(1+u_1)(p_1+p_2) \\ &< 4u \qquad\qquad \text{when } \beta^{p-1} \geqslant 10. \end{aligned} \tag{31}$$

From (30) and (31) we deduce that $1 - 2u < r + e < 1 + 10u$. Since $\beta^{p-1} \geqslant 10$, this implies $r + e \in [\beta^{-1}, 1) \cup [1, \beta)$. On the other hand, if $1 - 2u < r + e \leqslant 1$

then the relative error $|\epsilon_5|$ is at most $u\beta^{-1}/(1-2u)$, which contradicts the assumption $\epsilon_5 > u_7$. Thus, overall, we must have

$$1 < r + e < \beta,$$

and, using $\epsilon_5 > 0$, the rounded value $\widehat{x} = \mathrm{RN}(r+e)$ must be such that

$$\widehat{x} > r + e.$$

Finally, applying Property 2 (i) with $k = 7$, we deduce from $\epsilon_5 > u_7$ that

$$r + e \in \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3. \tag{32}$$

**Analysis depending on whether $p_1 + p_2$ belongs to $\mathcal{I}_1$, $\mathcal{I}_2$, or $\mathcal{I}_3$.** We now consider each of these three cases in turn in order to deduce the possible values for $\widehat{x}$ together with the corresponding intervals for $x$.

■ If $p_1 + p_2 \in \mathcal{I}_1$ then

$$r = 1 + 2u$$

and, using (32), we deduce that

$$e \in \underbrace{[-u, -u + 6u^2)}_{=: \mathcal{I}_1^{(1)}} \cup \underbrace{[u, u + 4u^2)}_{=: \mathcal{I}_1^{(2)}} \cup \underbrace{[3u, 3u + 2u^2)}_{=: \mathcal{I}_1^{(3)}}.$$

These intervals are valid no matter what the radix $\beta$ and the tie-breaking strategy of RN. If in addition $\beta$ is odd or RN rounds $1 + u$ down to 1, as is the case when condition (C) holds, then we have further

$$-u < e \quad \text{and} \quad -u < e_1 + e_2 \tag{33}$$

(see Appendix C for a detailed proof); thus, in this special case one can in particular replace $\mathcal{I}_1^{(1)}$ by

$$\begin{aligned} \mathcal{I}_1^{(1,C)} &:= \mathcal{I}_1^{(1)} \backslash \{-u\} \\ &= (-u, -u + 6u^2). \end{aligned}$$

Then, for each of the four intervals $\mathcal{I}_1^{(1)}$, $\mathcal{I}_1^{(1,C)}$, $\mathcal{I}_1^{(2)}$, and $\mathcal{I}_1^{(3)}$ we deduce the value of $\widehat{x}$ and a range for $e_1 + e_2$ and for $x$, as shown in Table 1 below. The value of $\widehat{x}$ follows immediately from rounding the sum $1 + 2u + e$ up to the nearest floating-point number. Let us now bound $e_1 + e_2$ from below when $e \in \mathcal{I}_1^{(1)}$: since in this case $|e| \leqslant u$, we have $\mathrm{ufp}(e) \leqslant \mathrm{ufp}(u) = \beta^{-p}$; using $|e_1 + e_2 - e| \leqslant u\,\mathrm{ufp}(e)$ then gives $e_1 + e_2 \geqslant e - u\,\mathrm{ufp}(e) \geqslant -u - \frac{2}{\beta}u^2$. When $e \in \mathcal{I}_1^{(1,C)}$, we use the strict lower bound already mentioned in (33). The remaining lower bounds and upper bounds for $e_1 + e_2$ are all deduced from the fact that $e_1 + e_2 = e(1 + \delta)$ with $|\delta| \leqslant u$. Finally, since $x = p_1 + p_2 + e_1 + e_2$, the range of $x$ is obtained

Table 1: Ranges or values of $e$, $\widehat{x}$, $e_1+e_2$, $x$ in the case $p_1+p_2 \in \mathcal{I}_1$.

| $e$ | $\widehat{x}$ | $e_1+e_2$ | $x$ |
|---|---|---|---|
| $\mathcal{I}_1^{(1)}$ | $1+2u$ | $[-u-\frac{2}{\beta}u^2, -u+7u^2-6u^3)$ | $[1-\frac{2}{\beta}u^2, 1+13u^2-6u^3)$ |
| $\mathcal{I}_1^{(1,C)}$ | $1+2u$ | $(-u, -u+7u^2-6u^3)$ | $(1, 1+13u^2-6u^3)$ |
| $\mathcal{I}_1^{(2)}$ | $1+4u$ | $[u-u^2, u+5u^2+4u^3)$ | $[1+2u-u^2, 1+2u+11u^2+4u^3)$ |
| $\mathcal{I}_1^{(3)}$ | $1+6u$ | $[3u-3u^2, 3u+5u^2+2u^3)$ | $[1+4u-3u^2, 1+4u+11u^2+2u^3)$ |

simply by adding the range $\mathcal{I}_1 = [1+u, 1+u+6u^2)$ of $p_1+p_2$ to the range of $e_1+e_2$ just computed.

■ If $p_1+p_2 \in \mathcal{I}_2$ then

$$r = 1+4u.$$

Furthermore, since $0 < -p_2 < \frac{1}{2}p_1$, we also have the lower bound

$$-\frac{5}{2}u \leqslant e_1+e_2. \tag{34}$$

(See Appendix C for a detailed proof.) Recalling that $e = (e_1+e_2)(1+\epsilon_3)$ with $|\epsilon_3| \leqslant u_1$, we deduce that

$$-\frac{5}{2}u(1+u_1) \leqslant e.$$

Applying (32) with $r = 1+4u$ and using the fact that $-3u+6u^2 \leqslant -\frac{5}{2}u(1+u_1)$ for $\beta^{p-1} \geqslant 10$, we finally obtain

$$e \in \underbrace{[-u, -u+4u^2)}_{=:\mathcal{I}_2^{(1)}} \cup \underbrace{[u, u+2u^2)}_{=:\mathcal{I}_2^{(2)}}.$$

Then we proceed in the same way as in the previous case and deduce for the intervals $\mathcal{I}_2^{(1)}$ and $\mathcal{I}_2^{(2)}$ the data collected in Table 2: rounding $1+4u+e$ up to the nearest floating-point number gives the values of $\widehat{x}$, applying $e_1+e_2 = e(1+\delta)$ with $|\delta| \leqslant u$ gives the ranges of $e_1+e_2$, and adding the ranges of $e_1+e_2$ to the one of $p_1+p_2$ leads to the ranges of $x$.

Table 2: Ranges or values of $e$, $\widehat{x}$, $e_1+e_2$, $x$ in the case $p_1+p_2 \in \mathcal{I}_2$.

| $e$ | $\widehat{x}$ | $e_1+e_2$ | $x$ |
|---|---|---|---|
| $\mathcal{I}_2^{(1)}$ | $1+4u$ | $[-u-u^2, -u+5u^2-4u^3)$ | $[1+2u-u^2, 1+2u+9u^2-4u^3)$ |
| $\mathcal{I}_2^{(2)}$ | $1+6u$ | $[u-u^2, u+3u^2+2u^3)$ | $[1+4u-u^2, 1+4u+7u^2+2u^3)$ |

■ If $p_1+p_2 \in \mathcal{I}_3$ then

$$r = 1+6u$$

and, using (32) and recalling from (31) that $e$ cannot be smaller than $-4u$, we deduce that for $\beta^{p-1} \geqslant 10$

$$e \in \underbrace{[-3u, -3u + 4u^2)}_{=: \mathcal{I}_3^{(1)}} \cup \underbrace{[-u, -u + 2u^2)}_{=: \mathcal{I}_3^{(2)}}.$$

Proceeding as in the two previous cases, we then obtain the values of $\widehat{x}$ and the ranges of $e_1 + e_2$ and $x$ shown in Table 3.

Table 3: Ranges or values of $e$, $\widehat{x}$, $e_1 + e_2$, $x$ in the case $p_1 + p_2 \in \mathcal{I}_3$.

| $e$ | $\widehat{x}$ | $e_1 + e_2$ | $x$ |
|---|---|---|---|
| $\mathcal{I}_3^{(1)}$ | $1 + 4u$ | $[-3u - 3u^2, -3u + 7u^2 - 4u^3)$ | $[1 + 2u - 3u^2, 1 + 2u + 9u^2 - 4u^3)$ |
| $\mathcal{I}_3^{(2)}$ | $1 + 6u$ | $[-u - u^2, -u + 3u^2 - 2u^3)$ | $[1 + 4u - u^2, 1 + 4u + 5u^2 - 2u^3)$ |

**Conclusion.** For $\beta^{p-1} \geqslant 10$, the second and fourth columns of Tables 1–3 lead to $\widehat{x} \geqslant x > 0$ and to the following relative error bounds:

$$\frac{|\widehat{x} - x|}{|x|} = \frac{\widehat{x}}{x} - 1 \leqslant \begin{cases} \frac{2\beta u + 2u^2}{\beta - 2u^2} = 2u + \frac{2}{\beta}u^2 + O(u^3) & \text{if } e \in \mathcal{I}_1^{(1)}, \\ 2u - \eta \text{ for some } \eta > 0 & \text{if } e \in \mathcal{I}_1^{(1,C)}, \\ \frac{2u + u^2}{1 + 2u - u^2} = 2u - 3u^2 + O(u^3) & \text{if } e \in \mathcal{I}_1^{(2)}, \\ \frac{2u + 3u^2}{1 + 4u - 3u^2} = 2u - 5u^2 + O(u^3) & \text{if } e \in \mathcal{I}_1^{(3)}, \\ \frac{2u + u^2}{1 + 2u - u^2} = 2u - 3u^2 + O(u^3) & \text{if } e \in \mathcal{I}_2^{(1)}, \\ \frac{2u + u^2}{1 + 4u - u^2} = 2u - 7u^2 + O(u^3) & \text{if } e \in \mathcal{I}_2^{(2)}, \\ \frac{2u + 3u^2}{1 + 2u - 3u^2} = 2u - u^2 + O(u^3) & \text{if } e \in \mathcal{I}_3^{(1)}, \\ \frac{2u + u^2}{1 + 4u - u^2} = 2u - 7u^2 + O(u^3) & \text{if } e \in \mathcal{I}_3^{(2)}. \end{cases}$$

From these eight cases and for $\beta^{p-1} \geqslant 10$, it is easily deduced that when discarding the interval $\mathcal{I}_1^{(1)}$, the error is always less than $2u$, while it is at most $\frac{2\beta u + 2u^2}{\beta - 2u^2}$ when discarding $\mathcal{I}_1^{(1,C)}$. This shows (29) and, therefore, concludes the analysis of the case where $\epsilon_4 > u_7$ and $\epsilon_5 > u_7$.

### B.2.4 Case where $\epsilon_4 < -u_7$ and $\epsilon_5 < -u_7$

In this case our goal is to show (17), that is, $|\widehat{x} - x|/|x| < 2u - u^2$ for $\beta^{p-1} \geqslant 10$. This bound shall be obtained in the same way as in Section B.2.3, but since it is less than $2u$ independently of the tie-breaking strategy of RN, the analysis will be slightly simpler.

**Preliminaries.** We can assume as before that

$$1 < p_1 + p_2 < \beta$$

and, applying Property 2 (ii) with $k = 7$, we deduce from $\epsilon_4 < -u_7$ that

$$p_1 + p_2 \in \underbrace{(1 + u - 6u^2, 1 + u]}_{=: \widetilde{\mathcal{I}}_0} \cup \underbrace{(1 + 3u - 4u^2, 1 + 3u]}_{=: \widetilde{\mathcal{I}}_1} \cup \underbrace{(1 + 5u - 2u^2, 1 + 5u]}_{=: \widetilde{\mathcal{I}}_2}.$$

Since $\epsilon_4 < 0$, we have the strict inequality

$$r < p_1 + p_2$$

and then, no matter what the tie-breaking strategy of rounding to nearest,

$$r \in \{1, 1 + 2u, 1 + 4u\}.$$

Since $\beta^{p-1} \geqslant 10$ and since the right endpoint of $\widetilde{\mathcal{I}}_2$ is not larger than the one of $\mathcal{I}_3$ from Section B.2.3, the bound $|e| < 4u$ established in (31) still holds. From $\beta^{p-1} \geqslant 10$ and $\epsilon_5 < -u_7$ it then follows that $1 < r + e < \beta$ and

$$\widehat{x} < r + e$$

and, using again Property 2 (ii) with $k = 7$, that

$$r + e \in \widetilde{\mathcal{I}}_0 \cup \widetilde{\mathcal{I}}_1 \cup \widetilde{\mathcal{I}}_2. \tag{35}$$

**Analysis depending on whether $p_1 + p_2$ belongs to $\widetilde{\mathcal{I}}_0$, $\widetilde{\mathcal{I}}_1$, or $\widetilde{\mathcal{I}}_2$.** As in the previous section, we will now consider each of these three cases in turn in order to deduce values for $\widehat{x}$ and intervals for $x$.

■ If $p_1 + p_2 \in \widetilde{\mathcal{I}}_0$ then

$$r = 1$$

and, using (35) together with the fact that $e$ is less than $4u$, we deduce that

$$e \in \underbrace{(u - 6u^2, u]}_{=: \widetilde{\mathcal{I}}_0^{(1)}} \cup \underbrace{(3u - 4u^2, 3u]}_{=: \widetilde{\mathcal{I}}_0^{(2)}}.$$

Proceeding as in Section B.2.3 we can set up the table below.

■ If $p_1 + p_2 \in \widetilde{\mathcal{I}}_1$ then

$$r = 1 + 2u.$$

Consequently, (35) implies

$$e \in \underbrace{(-u - 6u^2, -u]}_{=: \widetilde{\mathcal{I}}_1^{(1)}} \cup \underbrace{(u - 4u^2, u]}_{=: \widetilde{\mathcal{I}}_1^{(2)}} \cup \underbrace{(3u - 2u^2, 3u]}_{=: \widetilde{\mathcal{I}}_1^{(3)}},$$

Table 4: Ranges or values of $e$, $\widehat{x}$, $e_1+e_2$, $x$ in the case $p_1+p_2 \in \widetilde{\mathcal{I}}_0$.

| $e$ | $\widehat{x}$ | $e_1+e_2$ | $x$ |
|---|---|---|---|
| $\widetilde{\mathcal{I}}_0^{(1)}$ | $1$ | $(u-7u^2+6u^3, u+u^2]$ | $(1+2u-13u^2+6u^3, 1+2u+u^2]$ |
| $\widetilde{\mathcal{I}}_0^{(2)}$ | $1+2u$ | $(3u-7u^2+4u^3, 3u+3u^2]$ | $(1+4u-13u^2+4u^3, 1+4u+3u^2]$ |

Table 5: Ranges or values of $e$, $\widehat{x}$, $e_1+e_2$, $x$ in the case $p_1+p_2 \in \widetilde{\mathcal{I}}_1$.

| $e$ | $\widehat{x}$ | $e_1+e_2$ | $x$ |
|---|---|---|---|
| $\widetilde{\mathcal{I}}_1^{(1)}$ | $1$ | $(-u-7u^2-6u^3, -u+u^2]$ | $(1+2u-11u^2-6u^3, 1+2u+u^2]$ |
| $\widetilde{\mathcal{I}}_1^{(2)}$ | $1+2u$ | $(u-5u^2+4u^3, u+u^2]$ | $(1+4u-9u^2+4u^3, 1+4u+u^2]$ |
| $\widetilde{\mathcal{I}}_1^{(3)}$ | $1+4u$ | $(3u-5u^2+2u^3, 3u+3u^2]$ | $(1+6u-9u^2+2u^3, 1+6u+3u^2]$ |

and in each case the value of $\widehat{x}$ and the ranges of $e_1+e_2$ and $x$ are as shown in Table 5.

■ If $p_1+p_2 \in \widetilde{\mathcal{I}}_2$ then

$$r = 1+4u.$$

Using (35), we deduce

$$e \in \underbrace{(-3u-6u^2, -3u]}_{=:\,\widetilde{\mathcal{I}}_2^{(1)}} \cup \underbrace{(-u-4u^2, -u]}_{=:\,\widetilde{\mathcal{I}}_2^{(2)}} \cup \underbrace{(u-2u^2, u]}_{=:\,\widetilde{\mathcal{I}}_2^{(3)}}$$

and for each of these three intervals, the corresponding information about $\widehat{x}$, $e_1+e_2$, and $x$ appears in Table 6.

Table 6: Ranges or values of $e$, $\widehat{x}$, $e_1+e_2$, $x$ in the case $p_1+p_2 \in \widetilde{\mathcal{I}}_2$.

| $e$ | $\widehat{x}$ | $e_1+e_2$ | $x$ |
|---|---|---|---|
| $\widetilde{\mathcal{I}}_2^{(1)}$ | $1$ | $(-3u-9u^2-6u^3, -3u+3u^2]$ | $(1+2u-11u^2-6u^3, 1+2u+3u^2]$ |
| $\widetilde{\mathcal{I}}_2^{(2)}$ | $1+2u$ | $(-u-5u^2-4u^3, -u+u^2]$ | $(1+4u-7u^2-4u^3, 1+4u+u^2]$ |
| $\widetilde{\mathcal{I}}_2^{(3)}$ | $1+4u$ | $(u-3u^2+2u^3, u+u^2]$ | $(1+6u-5u^2+2u^3, 1+6u+u^2]$ |

**Conclusion.** For $\beta^{p-1} \geqslant 10$, the second and fourth columns of Tables 4–6 imply $\widehat{x} \leqslant x$ and thus the following relative error bounds:

$$\frac{|\widehat{x}-x|}{|x|} = 1 - \frac{\widehat{x}}{x} \leqslant \begin{cases} \frac{2u+u^2}{1+2u+u^2} = 2u - 3u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_0^{(1)}, \\ \frac{2u+3u^2}{1+4u+3u^2} = 2u - 5u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_0^{(2)}, \\ \frac{2u+u^2}{1+2u+u^2} = 2u - 3u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_1^{(1)}, \\ \frac{2u+u^2}{1+4u+u^2} = 2u - 7u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_1^{(2)}, \\ \frac{2u+3u^2}{1+6u+3u^2} = 2u - 9u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_1^{(3)}, \\ \frac{2u+3u^2}{1+2u+3u^2} = 2u - u^2 - O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_2^{(1)}, \\ \frac{2u+u^2}{1+4u+u^2} = 2u - 7u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_2^{(2)}, \\ \frac{2u+u^2}{1+6u+u^2} = 2u - 11u^2 + O(u^3) & \text{if } e \in \widetilde{\mathcal{I}}_2^{(3)}. \end{cases}$$

Finally, it is easily checked that all these bounds are less than $2u - u^2$, which proves (17) and finishes the case where $\epsilon_4 < -u_7$ and $\epsilon_5 < -u_7$.

# C Proofs of (33) and (34)

We begin with the following lemma, which will be used to prove (33) .

**Lemma 1.** *If $\beta$ is even then $u \in \mathbb{F}$. If $\beta$ is odd then $u$ is a midpoint for $\mathbb{F}$ and its expansion in radix $\beta$ has the form*

$$u = (\delta.\delta\delta\cdots)_\beta \cdot \beta^{-p}, \qquad \delta := \frac{\beta-1}{2}.$$

*Proof.* By definition, $u = \frac{\beta}{2} \cdot \beta^{-p}$. If $\beta$ is even then $\beta/2$ is an integer less than $\beta^p$, which implies $u \in \mathbb{F}$. Assume now that $\beta$ is odd. In this case $\beta/2 = \delta \cdot \beta/(\beta-1) = \delta \cdot \sum_{i=0}^{\infty} \beta^{-i}$, which gives the announced expansion in radix $\beta$. It remains to check that $u$ is a midpoint for $\mathbb{F}$. From the radix-$\beta$ expansion of $u$, we deduce that the two consecutive elements $f_1, f_2$ of $\mathbb{F}$ such that $f_1 < u < f_2$ are

$$f_1 = (\delta.\underbrace{\delta\delta\cdots\delta}_{p\,-\,1})_\beta \cdot \beta^{-p} \quad \text{and} \quad f_2 = f_1 + 2u \cdot \beta^{-p}.$$

The associated midpoint for $\mathbb{F}$ is thus $\frac{f_1+f_2}{2} = f_1 + u \cdot \beta^{-p} = \big(\sum_{i=0}^{p-1} \delta\beta^{-i} + \sum_{i=0}^{\infty} \delta\beta^{-i-p}\big) \cdot \beta^{-p} = \sum_{i=0}^{\infty} \delta\beta^{-i} \cdot \beta^{-p}$, which is precisely $u$. □

*Proof of (33).* This amounts to checking that $-u < e$ and $-u < e_1 + e_2$ when

(i) $r = 1 + 2u$;

(ii) $-u \leqslant e \in \mathbb{F}$;

(iii) $\widehat{x} := \mathrm{RN}(r+e) > r+e$;

(iv) condition (C) holds, that is, $\beta$ is odd or $\mathrm{RN}(1+u) = 1$.

If $\beta$ is odd, then Lemma 1 implies $-u \notin \mathbb{F}$ and it follows from (ii) that $-u < e$. If $\mathrm{RN}(1+u) = 1$, then $e \neq -u$, for otherwise (i) and (iii) would yield $\mathrm{RN}(1+u) > 1+u$, a contradiction. Hence, we have in both cases

$$-u < e.$$

Let us now check that $-u < e_1 + e_2$. If $\beta$ is odd, then $-u < e$ with $e = \mathrm{RN}(e_1+e_2)$ and, by Lemma 1, $-u$ is a midpoint for $\mathbb{F}$. The definition of RN thus implies that $-u \leqslant e_1 + e_2$. Now, since $e_1, e_2$ are in $\mathbb{F}$ and since, on the other hand, $-u$ has infinitely many radix-$\beta$ digits, we must have $-u \neq e_1 + e_2$. If $\beta$ is even then Lemma 1 implies $-u \in \mathbb{F}$, so that by the monotonicity of RN the strict inequality $-u < \mathrm{RN}(e_1+e_2)$ shown above leads to $-u < e_1 + e_2$. □

*Proof of (34).* This amounts to checking that $-\frac{5}{2}u \leqslant e_1 + e_2$ when

(i) $p_1 + p_2 \in [1+3u, 1+3u+4u^2)$;

(ii) $0 < -p_2 < \frac{1}{2}p_1$;

(iii) $p_1, p_2 \in \mathbb{F}$.

Applying (i) gives $1+3u-p_2 \leqslant p_1 < 1+3u+4u^2-p_2$ and then, using (ii) and $u \leqslant 1/4$, we obtain

$$1+3u < p_1 < 2+6u+8u^2 \leqslant 2+8u.$$

Since $p_1 \in \mathbb{F}$ by (iii), we deduce that

$$1+4u \leqslant p_1 \leqslant \begin{cases} 2+4u & \text{if } \beta = 2, \\ 2+6u & \text{if } \beta > 2. \end{cases} \tag{36}$$

Let us now show that $|p_2| < 1$. By (ii) and (36) we have $|p_2| = -p_2 < \frac{1}{2}p_1 \leqslant 1+3u$ and thus $|p_2| \leqslant 1+2u$ because $p_2$ is in $\mathbb{F}$ by (iii). For contradiction, assume that $|p_2| \geqslant 1$, that is, $p_2 \in \{-1, -1-2u\}$. It follows that $p_1 \geqslant 1+3u-p_2 \geqslant 2+3u$ and then $p_1 \geqslant 2+4u$, since $p_1$ is in $\mathbb{F}$. Combining the latter inequality with (36), we see that $p_1 = 2+4u$ if $\beta = 2$, and $p_1 \in \{2+4u, 2+6u\}$ if $\beta > 2$. Consequently,

$$p_1 + p_2 \in \begin{cases} \{1+2u, 1+4u\} & \text{if } \beta = 2, \\ \{1+2u, 1+4u, 1+6u\} & \text{if } \beta > 2. \end{cases}$$

Thus, in both cases, $p_1+p_2$ is in $\mathbb{F}$, which contradicts the fact that $\epsilon_4$ is nonzero. Therefore,

$$|p_2| < 1 \qquad \text{for any } \beta \geqslant 2. \tag{37}$$

From (36) and (37) we deduce that $\mathrm{ufp}(p_1) \leqslant 2$ and $\mathrm{ufp}(p_2) \leqslant \beta^{-1}$ for any $\beta \geqslant 2$. Since $|e_1| \leqslant u\,\mathrm{ufp}(p_1)$ and $|e_2| \leqslant u\,\mathrm{ufp}(p_2)$, we conclude that $|e_1+e_2| \leqslant (2+\beta^{-1})u \leqslant \frac{5}{2}u$. □

## Acknowledgments

This work was supported in part by the French National Research Agency, under grant ANR-13-INSE-0007 (MetaLibm project). I am also grateful to the referees for carefully reading the manuscript and offering helpful suggestions.

## References

[1] Marius Cornea, John Harrison, and Ping Tak Peter Tang. *Scientific Computing on Itanium®-based Systems*. Intel Press, Hillsboro, OR, USA, 2002.

[2] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, second edition, 2002.

[3] IEEE Computer Society. *IEEE Standard for Floating-Point Arithmetic, IEEE Standard 754-2008*. IEEE Computer Society, New York, August 2008. `http://ieeexplore.ieee.org/servlet/opac?punumber=4610933`.

[4] Claude-Pierre Jeannerod, Peter Kornerup, Nicolas Louvet, and Jean-Michel Muller. Error bounds on complex floating-point multiplication with an FMA, 2015. Preprint available at `https://hal.inria.fr/hal-00867040v4`.

[5] Claude-Pierre Jeannerod, Nicolas Louvet, and Jean-Michel Muller. Further analysis of Kahan's algorithm for the accurate computation of $2 \times 2$ determinants. *Mathematics of Computation*, 82(284):2245–2264, 2013.

[6] Claude-Pierre Jeannerod and Siegfried M. Rump. On relative errors of floating-point operations: optimal bounds and applications, 2014. Preprint available at `https://hal.inria.fr/hal-00934443`.

[7] Donald E. Knuth. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms*. Addison-Wesley, Reading, MA, USA, third edition, 1998.

[8] Jean-Michel Muller. On the error of computing $ab + cd$ using Cornea, Harrison and Tang's method. *ACM Trans. Math. Software*, 41(2):7:1–7:8, 2015.

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

[10] Pat H. Sterbenz. *Floating-Point Computation*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.