A rounding error analysis of the Cornea-Harrison-Tang method in radix β

Claude-Pierre Jeannerod

▶ **To cite this version:**

Claude-Pierre Jeannerod. A rounding error analysis of the Cornea-Harrison-Tang method in radix β. 2014. hal-01050021v1

HAL Id: hal-01050021

https://inria.hal.science/hal-01050021v1

Preprint submitted on 25 Jul 2014 (v1), last revised 23 Sep 2015 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A rounding error analysis of the Cornea-Harrison-Tang method in radix β

Claude-Pierre Jeannerod *

July 15, 2014

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 that, unlike previous studies, is not restricted to binary floating-point arithmetic but holds for any radix β. We show first that an asymptotically optimal bound on the relative error of this method is $2u + 2u^2 + O(u^3)$, where $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff in radix β and precision p. Then we show that the possibility of removing the $O(u^2)$ term from this bound depends on the radix parity and the tie-breaking strategy used for rounding: if β is odd or rounding is *to nearest even*, then the simpler bound $2u$ is obtained, while if β 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 $2u + \frac{1}{\beta}u^2$. All these results hold provided underflows and overflows do not occur and under some mild assumptions on β 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)
$p_1 := \text{RN}(ab)$; $p_2 := \text{RN}(cd)$;
$e_1 := \text{RN}(ab - p_1)$; $e_2 := \text{RN}(cd - p_2)$; // these two operations are exact.
$r := \text{RN}(p_1 + p_2)$; $e := \text{RN}(e_1 + e_2)$;
$\widehat{x} := \text{RN}(r + e)$;
return $\widehat{x}$

The main feature of this algorithm is its use of the fused multiply-add operation to compute the rounding errors of the two multiplications *exactly*, so that $e_1 = ab - p_1$ and $e_2 = cd - p_2$. The rounded sum e of these error terms is then

*Inria & laboratoire LIP (CNRS, ENS de Lyon, Inria, UCBL), Université de Lyon, France.

added to the (possibly highly inaccurate) rounded sum r of the two products in order to obtain an approximation $\widehat{x}$ having a tiny relative error.

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 showed in [1, pp. 273–275] 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 [7], 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:

- Does the bound $2u + O(u^2)$ hold beyond radix $\beta = 2$, that is, for $\beta > 2$ and u equal to $\frac{1}{2}\beta^{1-p}$?
- 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 [7] 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 $\mathrm{RN}(r + e_1)$. Thus, despite a lack of symmetry, this approach saves three floating-point operations compared with algorithm CHT; furthermore, barring underflow and overflow, it was shown in [5] that its relative error is bounded by $2u$, and that this bound is asymptotically optimal when β is even. Thus, it is important to understand whether this simple $O(u^2)$-free bound $2u$ can still be achieved in the more difficult case of 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 $p \geqslant 6$. Our second contribution is to show that, perhaps surprisingly, the answer to the second question depends on the parity of β and the way RN breaks ties: in some cases (say, when β 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, 7]), 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 β 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 $t \in \mathbb{R}$,

- $|\mathrm{RN}(t) - t| = \min_{s \in \mathbb{F}} |s - t|$,
- $\mathrm{RN}(\beta^i t) = \beta^i \mathrm{RN}(t)$ for all $i \in \mathbb{Z}$,
- $\mathrm{RN}(-t) = -\mathrm{RN}(t)$.

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 verified by roundTiesToEven and roundTiesToAway, the two specifications of rounding to nearest given in the IEEE 754-2008 standard [3, p. 16]: 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*

$$\widehat{x} = x(1+\epsilon), \qquad |\epsilon| < \begin{cases} 2u & \text{if } \beta \text{ is odd or } \mathrm{RN}(1+u) = 1, \\ 2u + 2u^2 + O(u^3) & \text{otherwise.} \end{cases}$$

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

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 β 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 β is even, $\mathrm{RN}(1+u) = 1+2u$, and $p \geqslant 3$. 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*

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

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* formats specified by the IEEE 754-2008 standard; see [6]. Similarly, it is easily checked that the assumptions $\beta^{p-1} \geqslant 24$ and $p \geqslant 3$ are 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 β and unit roundoff u. Then, in the absence of underflow and overflow, algorithm CHT has a relative error less than $2u$ when* RN *is roundTiesToEven, and less than $2u + O(u^2)$ 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 $2u + \frac{1}{\beta}u^2$.

Outline and additional notation. The rest of this paper is devoted to the proof of Theorems 1 and 2. Section 2 first identifies the inputs for which the upper bounds in Theorem 1 are immediate, and then introduces the tools needed to handle the remaining cases. Those cases are then analyzed in detail in Sections 3 and 4, depending on whether ab and cd have the same sign or not, and via a number of subcases treated separately. Overall, this case analysis leads to ten different error bounds, which are then summarized in Section 5.1, where Theorem 1 is proved. The lower bound given in Theorem 2 is established independently in Section 5.2.

A useful tool for our analyses will be the *unit in the first place* function [8], denoted by ufp and defined for $t \in \mathbb{R}$ by

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

Whenever this is more convenient, we shall also use the *unit in the last place* function, which is denoted by ulp and such that

$$\mathrm{ulp}(t) = \mathrm{ufp}(t) \cdot \beta^{1-p} \qquad \text{for all } t \in \mathbb{R}.$$

In particular, by definition of $\mathbb{F}$, RN, u, we have the classical relations

$$|\mathrm{RN}(t) - t| \leqslant \frac{1}{2}\mathrm{ulp}(t) = u\,\mathrm{ufp}(t) \leqslant u|t| \qquad \text{for all } t \in \mathbb{R}.$$

2 Preliminaries

We begin this section by handling trivial cases, for which the accuracy of the CHT algorithm is obviously smaller than $2u$, and by restricting the input set using symmetry arguments.

Assume first that either ab or cd or x is zero. In this case, by propagating the equality $\mathrm{RN}(0) = 0$ within the algorithm, we see that $\widehat{x} = \mathrm{RN}(x)$, so that the relative error is bounded by the unit roundoff:

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

Assume now that

$$ab \neq 0 \quad \text{and} \quad cd \neq 0 \quad \text{and} \quad x \neq 0.$$

Since $\mathrm{CHT}(a,b,c,d) = \mathrm{CHT}(c,d,a,b)$, we can exchange ab and cd to ensure

$$|ab| \geqslant |cd|.$$

Furthermore, 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 to

$$ab \geqslant 0.$$

Overall, it thus remains to analyze the case where a, b, c, d are such that

$$ab > 0 \quad \text{and} \quad -ab < cd \leqslant ab. \tag{2}$$

Deriving the bounds of Theorem 1 for inputs as in (2) will be the purpose of Sections 3 and 4. For this, we shall combine three techniques, which we review in Sections 2.1–2.3 below.

2.1 Inequalities resulting from the refined standard model

As noted in [6], the standard model (1) can be refined slightly by replacing u by the attainable bound

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

Applying this refined model 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,

$$\widehat{x} = x(1+\epsilon_4)(1+\epsilon_5) + (e_1+e_2)(\epsilon_3-\epsilon_4)(1+\epsilon_5). \tag{3}$$

On the other hand, the definition of ϵ_1 and ϵ_2 implies that

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

Hence, writing

$$K = \frac{|ab| + |cd|}{|ab + cd|}, \tag{4}$$

we arrive at the following inequalities:

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + \max\left\{|\epsilon_1|, |\epsilon_2|\right\} \cdot |\epsilon_3 - \epsilon_4|(1 + \epsilon_5) \cdot K \\ &\leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + 2u_1^2(1 + u_1) \cdot K. \end{aligned} \tag{5}$$

Remark. When using u_1 instead of u the bound $2u + 7u^2 + O(u^3)$ obtained in [7] immediately becomes $2u + 5u^2 + O(u^3)$. This sharper bound, however, is not enough for our purposes, since it assumes $\beta = 2$ and has a $O(u^2)$ term no matter what the tie-breaking strategy.

2.2 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 β of those used in [7] for radix two) will be useful when dealing with the case in Section 4.1, where p_1 and $-p_2$ are so close to each other that $p_1 + p_2$ is computed exactly.

Property 1. *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 β^{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 2. *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 β^{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 1, if $i = e_a + e_b$ then $ab - \mathrm{RN}(ab)$ is an integer multiple of β^{i-2p+2}. Else $i = e_a + e_b + 1$, but then $|ab| = |AB|\beta^{1-2p+i} \leqslant (\beta^p - 1)^2 \beta^{1-2p+i} = (1 - \frac{2u}{\beta})^2 \beta^{i+1}$. □

2.3 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.

Then, having defined u_k, we can state the following two properties, which indicate the range constraints implied by large enough relative errors. Property 3 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 4, 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 3. *Let $k \in \mathbb{R}_{\geqslant 0}$. 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 strict lower bound on $|t|/\mathrm{ufp}(t)$ follows from the fact that $t \notin \mathbb{F}$. To establish the upper bound, it suffices to note that if it does not hold, then $|t| \geqslant (1+ku)\mathrm{ufp}(t)$, which implies $\frac{|\mathrm{RN}(t)-t|}{|t|} \leqslant \frac{u\,\mathrm{ufp}(t)}{|t|} \leqslant \frac{u\,\mathrm{ufp}(t)}{(1+ku)\,\mathrm{ufp}(t)} = u_k$. □

Property 4. *Given $k \in \mathbb{R}_{\geqslant 0}$, 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:

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

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

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

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

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

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

Hence the expression for interval I_j follows from (6). Recalling from Property 3 that $t < 1 + ku$, we deduce from (6b) 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

$$\frac{1 + 2ju}{1 - u_k} < t. \tag{7a}$$

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

$$t \leqslant 1 + (2j+1)u. \tag{7b}$$

From (7) we deduce the definition of interval $\widetilde{I}_j$. Finally, using again Property 3 we have $t < 1 + ku$, which together with (7a) 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 in Sections 4.2.3 and 4.2.4 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 simpler intervals $\mathcal{I}_j$ and $\widetilde{\mathcal{I}}_j$, defined by degree-2 polynomials in u as follows: 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 ℓ is defined in Property 4 as $\ell = \lceil (k-1)/2 \rceil$, it is easily checked that

$$I_j \subset \mathcal{I}_j, \qquad \widetilde{I}_{j-1} \subset \widetilde{\mathcal{I}}_{j-1}, \qquad j = 1, \ldots, \ell. \tag{8}$$

3 Analysis when ab and cd have the same sign

When ab and cd have the same sign, the assumption in (2) implies that both ab and cd are positive and that K in (4) is equal to one. Using (5), 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 bound $|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant 2u_1 + u_1^2$ 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.

3.1 Case where $\epsilon_4 \leqslant u_2$ or $\epsilon_5 \leqslant u_2$

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

$$-2u_1 + u_1^2 \leqslant \epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5 \leqslant u_1 + u_2 + u_1u_2.$$

Since $u_1 + u_2 + u_1u_2 \geqslant 2u_1 - u_1^2$, we have

$$|\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| \leqslant u_1 + u_2 + u_1u_2,$$

which ensures $|\widehat{x} - x|/|x| \leqslant u_1 + u_2 + u_1u_2 + 2u_1^2 + 2u_1^3$ or, after simplification,

$$\frac{|\widehat{x} - x|}{|x|} \leqslant 2u\frac{1 + 3u + 3u^2}{(1+u)^3} = 2u - 2u^4 + O(u^5). \tag{9}$$

3.2 Case where $\epsilon_4 > u_2$ and $\epsilon_5 > u_2$

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 3 with $k = 2$ gives

$$\beta^i < p_1 + p_2 < (1+2u)\beta^i$$

for some integer i. Since ϵ_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

$$(1-u_1)^2 x \leqslant r \leqslant (1+u_1)^2 x. \tag{10}$$

Using the lower bound in (10) thus leads to

$$\begin{aligned} |e| &\leqslant \frac{u_1 + u_1^2}{(1-u_1)^2} r = \frac{(u_1 + u_1^2)(1+2u)}{(1-u_1)^2}\beta^i \\ &< 2u\beta^i \qquad \text{for } \beta^{p-1} \geqslant 4. \end{aligned}$$

Consequently,

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

Now, the assumption $\epsilon_5 > u_2$ implies that when rounding $r + e$ to nearest rounding up occurs and, by Property 3, that $r + e$ must be less than $(1+2u)\beta^i$. 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, since (10) implies $|r - x|/|x| \leqslant 2u_1 + u_1^2 = 2u\frac{1+3u/2}{(1+u)^2}$, we conclude that

$$\frac{|\widehat{x} - x|}{|x|} \leqslant 2u\frac{1 + \frac{3}{2}u}{(1+u)^2} = 2u - u^2 + O(u^4) \qquad \text{for } \beta^{p-1} \geqslant 4. \tag{11}$$

4 Analysis when ab and cd have opposite signs

When ab and cd have opposite signs, the assumption in (2) can be rewritten

$$-ab < cd < 0 < ab. \tag{12}$$

This implies

$$K = \frac{ab + |cd|}{ab - |cd|} \tag{13}$$

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 $p_1 + p_2 \geqslant 0$.

4.1 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 β the sum $p_1 + p_2$ is computed exactly by Sterbenz's theorem [9, p. 138] (see also [2, p. 45]). Hence

$$\epsilon_4 = 0$$

and, using (3),

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant |\epsilon_5| + \frac{|(e_1 + e_2)\epsilon_3|}{|x|}(1 + \epsilon_5) \\ &\leqslant u_1 + u_1^2(1 + u_1)K. \end{aligned} \tag{14}$$

If $K \leqslant 1/u_1$ then we deduce immediately from the latter bound that the relative error on x is bounded by $2u_1 + u_1^2$ as in (11). Hence the rest of this section is devoted to handling the case

$$K > \frac{1}{u_1} = \frac{1}{u} + 1. \tag{15}$$

(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 (12), (13), and (15) 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.

4.1.1 Case $\beta^i \leqslant |cd| < ab < \beta^{i+1}$

In this case, $\mathrm{ulp}(ab) = \mathrm{ulp}(cd) = \beta^{i-p+1}$ and, since $|e_1| \leqslant \frac{1}{2}\mathrm{ulp}(ab)$ and $|e_2| \leqslant \frac{1}{2}\mathrm{ulp}(cd)$, we obtain

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

On the other hand, Property 1 implies the existence of integers E_1, E_2 such that

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

Consequently, $e_1+e_2 = (E_1+E_2)\cdot\beta^{i-2p+1}$ and the integer E_1+E_2 must satisfy $|E_1 + E_2| \leqslant \beta^p$. This means $e_1 + e_2 \in \mathbb{F}$ or, equivalently,

$$\epsilon_3 = 0.$$

It then follows immediately from (14) that $|\widehat{x} - x|/|x| \leqslant |\epsilon_5| \leqslant u_1$, that is,

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{u}{1+u} = u - u^2 + O(u^3). \tag{16}$$

4.1.2 Case $\beta^{i-1} \leqslant |cd| < \beta^i \leqslant ab < \beta^{i+1}$

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

$$|e_1 + e_2| \leqslant \frac{1}{2}\beta^{i-p+1} + \frac{1}{2}\beta^{i-p}. \tag{17}$$

Property 1 still gives

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

for some integer E_1, and by applying Property 2 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 (17) gives $|E_1 + E_2| \leqslant \frac{1}{2}\beta^p + \frac{1}{2}\beta^{p-1} < \beta^p$, from which we deduce $e_1 + e_2 \in \mathbb{F}$, that is,

$$\epsilon_3 = 0.$$

It follows that we have here the same bound as in Section 4.1.1, namely

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{u}{1+u} = u - u^2 + O(u^3).$$

■ Assume now that $|cd| \leqslant \left(1 - \frac{2u}{\beta}\right)^2 \beta^i$. To handle this case, we shall derive a lower bound on $|x| = ab - |cd|$ together with an upper bound on $|(e_1+e_2)\epsilon_3| = |e - (e_1 + e_2)|$, and then conclude using (14).

Since $ab \geqslant \beta^i$, we have

$$x \geqslant \Big(\frac{4u}{\beta} - \frac{4u^2}{\beta^2}\Big)\beta^i.$$

On the other hand, (17) implies $|e_1 + e_2| < \beta^{i-p+1}$ and thus $\mathrm{ufp}(e_1 + e_2) \leqslant \beta^{i-p}$. Therefore, it follows from $e = \mathrm{RN}(e_1+e_2)$ that $|e-(e_1+e_2)| \leqslant u\,\mathrm{ufp}(e_1+e_2) \leqslant u\beta^{i-p}$, that is,

$$\big|e - (e_1 + e_2)\big| \leqslant \frac{\beta}{2}\beta^{i-2p}.$$

Applying (14) then gives

$$\frac{|\widehat{x} - x|}{|x|} \leqslant u_1 + (1 + u_1)\varphi, \qquad \varphi := \frac{\frac{\beta}{2}\beta^{-2p}}{\frac{4u}{\beta} - \frac{4u^2}{\beta^2}}.$$

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

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \frac{3}{2}u. \tag{18}$$

4.2 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, noticing that $K = 1 + \frac{2}{\psi-1}$, we deduce that

$$K < \frac{3 - u_1}{1 - 3u_1} = 3 + 8u + O(u^2). \tag{19}$$

Combining (19) and (5), we obtain

$$\frac{|\widehat{x} - x|}{|x|} \leqslant |\epsilon_4 + \epsilon_5 + \epsilon_4\epsilon_5| + 2u_1^2(1 + u_1)\frac{3 - u_1}{1 - 3u_1}. \tag{20}$$

For the same reason as in Section 3, applying to (20) 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 always have the form $2u+5u^2+O(u^3)$. In order to achieve the bounds claimed in Theorem 1, we shall refine further this analysis by examining separately four cases defined by the pair (ϵ_4, ϵ_5).

4.2.1 Case where ϵ_4 and ϵ_5 have opposite signs

In this case

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

so (20) immediately gives a bound of the form $u + O(u^2)$:

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant u \cdot \frac{1 + 6u + 13u^2 + 6u^3}{(1-2u)(1+u)^3} \\ &< 2u \quad \text{for } \beta^{p-1} \geqslant 4. \end{aligned} \tag{21}$$

4.2.2 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 (20) then leads to

$$\begin{aligned} \frac{|\widehat{x} - x|}{|x|} &\leqslant 2u \cdot \frac{1 + \frac{15}{2}u + 26u^2 + \frac{89}{2}u^3 + 19u^4}{(1-2u)(1+7u)(1+u)^3} = 2u - u^2 + O(u^3) \\ &< 2u \quad \text{for } \beta^{p-1} \geqslant 24. \end{aligned} \tag{22}$$

4.2.3 Case where $\epsilon_4 > u_7$ and $\epsilon_5 > u_7$

In this case, we will derive a relative error bound that depends on radix parity and the tie-breaking strategy of rounding to nearest. More precisely, defining the condition

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

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

$$\frac{|\widehat{x} - x|}{|x|} \leqslant \begin{cases} 2u - \eta \ \text{ for some } \eta > 0 & \text{if (C) holds,} \\ \frac{2u+2u^2}{1-2u^2} = 2u + 2u^2 + O(u^3) & \text{otherwise.} \end{cases} \tag{23}$$

To establish (23), we shall apply Property 4 in order to obtain suitables 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 ϵ_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 β 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 4 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

$$r \in \{1+2u, 1+4u, 1+6u\}. \tag{24}$$

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

$$\begin{aligned} |e| &\leqslant (1+u_1)\big(|e_1| + |e_2|\big) \\ &\leqslant u(1+u_1)\big(p_1 + |p_2|\big) \\ &< 3u(1+u_1)(p_1+p_2) \\ &< 4u \qquad\qquad \text{when } \beta^{p-1} \geqslant 10. \end{aligned} \tag{25}$$

From (24) and (25) we deduce that $1 - 2u < r + e < 1 + 10u$. Now, $1 + 10u$ is easily seen to be at most β and, 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 4 (i) with $k = 7$, we deduce from $\epsilon_5 > u_7$ that

$$r + e \in \mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3. \tag{26}$$

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 (26), 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 β and the tie-breaking strategy of RN. If in addition β is odd or RN rounds $1+u$ down to 1, as is the case when condition (C) holds, then we have further

$$-u < e \quad \text{and} \quad -u < e_1 + e_2 \tag{27}$$

(see Appendix A 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. To obtain the range of $e_1 + e_2$, we use the fact that $e_1 + e_2 = e(1+\delta)$ with $|\delta| \leqslant u$; for the lower bound on $e_1 + e_2$ in the case $e \in \mathcal{I}_1^{(1,C)}$, we use further the second inequality in (27). Finally, since $x = p_1 + p_2 + e_1 + e_2$, the range of x is obtained 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.

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
$\mathcal{I}_1^{(1)}$	$1 + 2u$	$[-u - u^2, -u + 7u^2 - 6u^3)$	$[1 - u^2, 1 + 13u^2 - 6u^3)$
$\mathcal{I}_1^{(1,C)}$	$1 + 2u$	$(-u, -u + 7u^2 - 6u^3)$	$(1, 1 + 13u^2 - 6u^3)$
$\mathcal{I}_1^{(2)}$	$1 + 4u$	$[u - u^2, u + 5u^2 + 4u^3)$	$[1 + 2u - u^2, 1 + 2u + 11u^2 + 4u^3)$
$\mathcal{I}_1^{(3)}$	$1 + 6u$	$[3u - 3u^2, 3u + 5u^2 + 2u^3)$	$[1 + 4u - 3u^2, 1 + 4u + 11u^2 + 2u^3)$

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

$$r = 1 + 4u$$

and, using (26), we deduce that

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

Then we proceed in the same way as in the previous case. First, if condition (C) holds, we can check (see Appendix A) that

$$-3u < e \quad \text{and} \quad -3u < e_1 + e_2, \tag{28}$$

which leads to replacing $\mathcal{I}_2^{(1)}$ by

$$\begin{aligned}\mathcal{I}_2^{(1,C)} &:= \mathcal{I}_2^{(1)} \backslash \{-3u\} \\ &= (-3u, -3u + 6u^2).\end{aligned}$$

Second, we deduce for each of the intervals $\mathcal{I}_2^{(1)}$, $\mathcal{I}_2^{(1,C)}$, $\mathcal{I}_2^{(2)}$, $\mathcal{I}_2^{(3)}$ the data collected in Table 2. In particular, in the case $e \in \mathcal{I}_2^{(1,C)}$ the lower bound on $e_1 + e_2$ is the one given in (28), while in the case $e \in \mathcal{I}_2^{(1)}$ we use

$$-3u - 2u^2 \leqslant e_1 + e_2. \tag{29}$$

(The latter bound, shown in Appendix A, slightly improves upon the straightforward bound $-3u - 3u^2$ obtained using $e_1 + e_2 = e(1 + \delta)$ with $|\delta| \leqslant u$.)

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
$\mathcal{I}_2^{(1)}$	$1 + 2u$	$[-3u - 2u^2, -3u + 9u^2 - 6u^3)$	$[1 - 2u^2, 1 + 13u^2 - 6u^3)$
$\mathcal{I}_2^{(1,C)}$	$1 + 2u$	$(-3u, -3u + 9u^2 - 6u^3)$	$(1, 1 + 13u^2 - 6u^3)$
$\mathcal{I}_2^{(2)}$	$1 + 4u$	$[-u - u^2, -u + 5u^2 - 4u^3)$	$[1 + 2u - u^2, 1 + 2u + 9u^2 - 4u^3)$
$\mathcal{I}_2^{(3)}$	$1 + 6u$	$[u - u^2, u + 3u^2 + 2u^3)$	$[1 + 4u - u^2, 1 + 4u + 7u^2 + 2u^3)$

■ If $p_1 + p_2 \in \mathcal{I}_3$ then

$$r = 1 + 6u$$

and, using (26) and recalling from (25) that e cannot be smaller than $-4u$, we deduce that

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

Proceeding as in the two previous cases, we then obtain the value 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 $\widehat{x} \geqslant x$ and to the following relative error bounds:

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

From these ten cases and for $\beta^{p-1} \geqslant 10$, it is easily deduced that when discarding the two intervals $\mathcal{I}_1^{(1)}$ and $\mathcal{I}_2^{(1)}$, the error is always less than $2u$, while it is at most $\frac{2u+2u^2}{1-2u^2}$ when discarding $\mathcal{I}_1^{(1,C)}$ and $\mathcal{I}_2^{(1,C)}$. This shows (23) and, therefore, concludes the analysis of the case where $\epsilon_4 > u_7$ and $\epsilon_5 > u_7$.

4.2.4 Case where $\epsilon_4 < -u_7$ and $\epsilon_5 < -u_7$

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

$$\frac{|\widehat{x}-x|}{|x|} \leqslant \frac{2u+3u^2}{1+2u+3u^2} = 2u - u^2 - O(u^3). \tag{30}$$

This bound shall be obtained in the same way as in Section 4.2.3, 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 4 (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 4.2.3, the bound $|e| < 4u$ established in (25) still holds. It then follows from $\epsilon_5 < -u_7$ that $1 < r + e < \beta$ and

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

and, using again Property 4 (ii) with $k = 7$, that

$$r + e \in \widetilde{\mathcal{I}}_0 \cup \widetilde{\mathcal{I}}_1 \cup \widetilde{\mathcal{I}}_2. \tag{31}$$

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 (31) 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 4.2.3 we can set up the table below.

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

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

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

$$r = 1 + 2u.$$

Consequently, (31) implies

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

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.

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

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

$$r = 1+4u.$$

Using (31), 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, all less than $2u$:

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

Finally, it is easily checked that when $\beta^{p-1} \geqslant 10$ the largest of these bounds is the sixth one, which proves (30) and finishes the case where $\epsilon_4 < -u_7$ and $\epsilon_5 < -u_7$.

5 Proofs of the theorems

5.1 Proof of Theorem 1

According to the analysis performed in Sections 2–4, the value $\widehat{x}$ computed by algorithm CHT satisfies $\widehat{x} = x(1+\epsilon)$ with $|\epsilon|$ bounded as shown in the table below. Recall from Section 4.2.3 that (C) denotes the condition "β is odd or $\mathrm{RN}(1+u) = 1$."

Equation	Upper bound on $\lvert\epsilon\rvert$	Extra condition
(1)	u	
(9)	$2u - 2u^4 + O(u^5)$	
(11)	$2u - u^2 + O(u^4)$	$\beta^{p-1} \geqslant 4$
(16)	$u - u^2 + O(u^3)$	
(18)	$\frac{3}{2}u$	
(21)	$u + O(u^2)$	$\beta^{p-1} \geqslant 4$
(22)	$2u - u^2 + O(u^3)$	$\beta^{p-1} \geqslant 24$
(23)	$\begin{cases} 2u - \eta \text{ for some } \eta > 0 & \text{if (C) holds,} \\ 2u + 2u^2 + O(u^3) & \text{otherwise;} \end{cases}$	$\beta^{p-1} \geqslant 10$
(30)	$2u - u^2 - O(u^3)$	$\beta^{p-1} \geqslant 10$

It is easily checked that if (C) holds, then these bounds on $|\epsilon|$ are all less than $2u$, and that otherwise the largest bound is the second one in Equation (23). Furthermore, we know from [4, Corollary 4.1] that if $\beta^{p-1} \geqslant 12$, 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 upper bounds in Theorem 1.

5.2 Proof of Theorem 2

For β even, we know from [6, Theorem 3.2] that there exist $a, b \in \mathbb{F}$ such that

$$ab = 1 + u. \tag{32}$$

Since $1+u$ is exactly halfway between 1 and $1+2u$, RN will round it away from zero, so that

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

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,

$$cd = -\Big(u + \frac{2}{\beta}u^2 - 4(1-\frac{1}{\beta})u^3\Big), \tag{33}$$

which implies

$$u < |cd| < u + \frac{2}{\beta}u^2 = u + \frac{1}{2}\mathrm{ulp}(u)$$

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 have $e = \mathrm{RN}(e_1+e_2) = \mathrm{RN}(cd) = p_2$, that is

$$e = -u.$$

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

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

Finally, we deduce from (32) and (33) 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 $2u + \frac{1}{\beta}u^2$ when $p \geqslant 3$. This concludes the proof of Theorem 2.

A Proofs of (27), (28), and (29)

Lemma 1.

- *If β is even then u and $3u$ are in $\mathbb{F}$.*

- *If β is odd then u and $3u$ are midpoints for $\mathbb{F}$, and their expansions in radix β have the form*

$$u = (\delta.\delta\delta\cdots)_\beta \cdot \beta^{-p} \qquad \textit{and} \qquad 3u = (1.\delta\delta\delta\cdots)_\beta \cdot \beta^{1-p},$$

where $\delta = (\beta-1)/2$.

Proof. By definition, $u = \frac{\beta}{2}\cdot\beta^{-p}$ and thus $3u = \frac{3\beta}{2}\cdot\beta^{-p}$. If β is even then $\beta/2$ and $3\beta/2$ are integers less than β^p, which implies $u \in \mathbb{F}$ and $3u \in \mathbb{F}$. Assume now that β is odd. In this case $\beta/2 = \delta\cdot\beta/(\beta-1) = \delta\cdot\sum_{i=0}^{\infty}\beta^{-i}$, which gives the announced expansions in radix β. It remains to check that u is a midpoint for $\mathbb{F}$. (For $3u$, this can be verified similarly.) From the radix-β 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 (27). 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, β is odd or $\mathrm{RN}(1+u) = 1$.

If β 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 β 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-β digits, we must have $-u \neq e_1 + e_2$. If β 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 (28). This amounts to checking that $-3u < e$ and $-3u < e_1 + e_2$ when

(i) $r = 1 + 4u$;

(ii) $-3u \leqslant e \in \mathbb{F}$;

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

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

For this we can proceed as in the proof of (27), with $3u$ replacing u. □

Proof of (29). The goal here is to show that if $e=\mathrm{RN}(e_1+e_2)$ satisfies $-3u \leqslant e$, then $-3u-2u^2 \leqslant e_1+e_2$. By definition of RN and ulp, we have

$$|e-(e_1+e_2)| \leqslant \frac{1}{2}\mathrm{ulp}(e_1+e_2) \leqslant \frac{1}{2}\mathrm{ulp}(e) \leqslant \frac{1}{2}\mathrm{ulp}(3u).$$

Now, $3u=\frac{3}{2}\beta^{1-p}$ and, since $\frac{3}{2}\in[1,\beta)$ for $\beta \geqslant 2$, we deduce that

$$\mathrm{ulp}(3u)=\beta^{2-2p}=4u^2.$$

Consequently, $|e-(e_1+e_2)| \leqslant 2u^2$ and thus $e_1+e_2 \geqslant e-2u^2 \geqslant -3u-2u^2$. □

References

[1] M. CORNEA, J. HARRISON, AND P. T. P. TANG, *Scientific Computing on Itanium®-based Systems*, Intel Press, Hillsboro, OR, 2002.

[2] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, second ed., 2002.

[3] IEEE COMPUTER SOCIETY, *IEEE Standard for Floating-Point Arithmetic*, IEEE Standard 754-2008, August 2008. Available at `http://ieeexplore.ieee.org/servlet/opac?punumber=4610933`.

[4] C.-P. JEANNEROD, P. KORNERUP, N. LOUVET, AND J.-M. MULLER, *Error bounds on complex floating-point multiplication with an FMA*, September 2013. Available at `http://hal.inria.fr/hal-00867040`.

[5] C.-P. JEANNEROD, N. LOUVET, AND J.-M. MULLER, *Further analysis of Kahan's algorithm for the accurate computation of* 2×2 *determinants*, Math. Comp. **82** (2013), 2245–2264.

[6] C.-P. JEANNEROD AND S. M. RUMP, *On relative errors of floating-point operations: optimal bounds and applications*, January 2014. Available at `http://hal.inria.fr/hal-00934443`.

[7] J.-M. MULLER, *On the error of computing* $ab+cd$ *using Cornea, Harrison and Tang's method*, September 2013; to appear in ACM Trans. Math. Software. Available at `http://hal.inria.fr/ensl-00862910`.

[8] S. M. RUMP, T. OGITA, AND S. OISHI, *Accurate floating-point summation, Part I: Faithful rounding*, SIAM J. Sci. Comput., 31 (2008), pp. 189–224.

[9] P. H. STERBENZ, *Floating-Point Computation*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.