Further analysis of Kahan's algorithm for the accurate computation of 2 x 2 determinants

Claude-Pierre Jeannerod, Nicolas Louvet, Jean-Michel Muller

MATHEMATICS OF COMPUTATION
Volume 82, Number 284, October 2013, Pages 2245–2264
S 0025-5718(2013)02679-8
Article electronically published on March 4, 2013

FURTHER ANALYSIS OF KAHAN'S ALGORITHM FOR THE ACCURATE COMPUTATION OF 2×2 DETERMINANTS

CLAUDE-PIERRE JEANNEROD, NICOLAS LOUVET, AND JEAN-MICHEL MULLER

Abstract. We provide a detailed analysis of Kahan's algorithm for the accurate computation of the determinant of a 2×2 matrix. This algorithm requires the availability of a fused multiply-add instruction. Assuming radix-β, precision-p floating-point arithmetic with β even, $p \geq 2$, and barring overflow or underflow we show that the absolute error of Kahan's algorithm is bounded by $(\beta+1)/2$ ulps of the exact result and that the relative error is bounded by $2u$ with $u = \frac{1}{2}\beta^{1-p}$ the unit roundoff. Furthermore, we provide input values showing that i) when $\beta/2$ is odd—which holds for 2 and 10, the two radices that matter in practice—the absolute error bound is optimal; ii) the relative error bound is asymptotically optimal, that is, for such input the ratio (relative error)$/2u$ has the form $1 - O(\beta^{-p})$. We also give relative error bounds parametrized by the relative order of magnitude of the two products in the determinant, and we investigate whether the error bounds can be improved when adding constraints: When the products in the determinant have opposite signs, which covers the computation of a sum of squares, or when Kahan's algorithm is used for computing the discriminant of a quadratic equation.

1. Introduction

Expressions of the form $ad \pm bc$ with a, b, c, d some floating-point numbers arise naturally in many numerical computations. Examples include complex multiplication and division; discriminant of quadratic equations; cross-products and 2D determinants (e.g., for geometric predicates [14]). Unfortunately, the naive way of computing $ad \pm bc$ may lead to very inaccurate results, due to catastrophic cancellations. Hence it is of interest to compute $ad \pm bc$ accurately.

Concerning complex multiplication and division, the result may be inaccurate at least when we consider the real part and imaginary part of a complex product or quotient separately. For instance, for the complex product $z = z_1 z_2$, assuming $\hat{z}$ is the computed value, the naive method may lead to large values of the componentwise relative errors $|\Re(\hat{z}) - \Re(z)|/|\Re(z)|$ and $|\Im(\hat{z}) - \Im(z)|/|\Im(z)|$, although Brent, Percival, and Zimmermann [2] have shown that in precision-p binary floating-point arithmetic, the normwise relative error $|\hat{z} - z|/|\hat{z}|$ is always bounded by $\sqrt{5} \cdot 2^{-p}$ (which is a very sharp bound; for instance, in IEEE 754 binary32 arithmetic, they could build examples for which the normwise relative error is $2^{-p} \cdot \sqrt{4.9999899864}$).

An algorithm was proposed by Kahan [12] for the accurate computation of the discriminant $b^2 - ac$ of a quadratic equation $ax^2 - 2bx + c = 0$ with binary floating-point coefficients. Boldo [1] then gave a formal proof of the high relative accuracy of this algorithm, assuming IEEE 754 double-precision floating-point arithmetic [8], and allowing underflows in the intermediate computations.

Received by the editor December 7, 2011 and, in revised form, January 17, 2012.
2010 *Mathematics Subject Classification.* Primary 65G50.

Another algorithm for the computation of discriminants using specifically the fused multiply-add (FMA) instruction is stated in [11, p. 15]. The FMA instruction, which evaluates expressions of the form $ab+c$ with one rounding error instead of two, was first implemented on the IBM RS/6000 processor [7, 15]. It is currently available on several processors like the IBM PowerPC [9], the HP/Intel Itanium [3], the Fujitsu SPARC64 VI, and the STI Cell. More importantly, the FMA instruction is included in the new IEEE 754-2008 standard for floating-point arithmetic [8], so that within a few years, it will probably be available on most general-purpose processors. Experiments are provided in [11] that illustrate the high relative accuracy of the algorithm, but no error bound is provided.

For computing $ad \pm bc$, an algorithm attributed to Kahan by Higham [5, p. 65] can be used when an FMA instruction is available. Kahan's algorithm for computing $ad - bc$ is Algorithm 1 below. Here and hereafter, for any real number t we denote by $\mathrm{RN}(t)$ the floating-point number in radix β and precision p that is nearest to it and, in case of a tie, whose least significant digit is even (*roundTiesToEven* in [8]).

Algorithm 1 Kahan's way to compute $x = ad - bc$ with fused multiply-adds.

$\hat{w} \leftarrow \mathrm{RN}(bc)$
$e \leftarrow \mathrm{RN}(\hat{w} - bc)$ // this operation is exact: $e = \hat{w} - bc$.
$\hat{f} \leftarrow \mathrm{RN}(ad - \hat{w})$
$\hat{x} \leftarrow \mathrm{RN}(\hat{f} + e)$
return $\hat{x}$

Thus, Kahan's algorithm can be implemented in IEEE floating-point arithmetic using one multiplication, two independent FMA operations, and one addition. The fact that the error $\hat{w} - bc$ is computed exactly is a classic property of the FMA operation: it can be traced back at least to 1996 in Kahan's lecture notes [10] and is mentioned later on by several authors (see for example [13, Fig. 2] and [4, 16]), but was probably known at the time it was decided to include the FMA in the instruction set of the IBM RS/6000 processor.

Higham [5, solution to Problem 2.25] (or Problem 2.27 in [6]) shows that in the absence of underflow and overflow, Algorithm 1 approximates $x = ad - bc$ with high relative accuracy as long as $u|bc| \gg |x|$ does not hold, where $u = \frac{1}{2}\beta^{1-p}$ is the unit roundoff. The purpose of this paper is to establish—again, in the absence of underflow and overflow—that Kahan's algorithm *always* achieves high relative accuracy, and to give tight bounds on both the relative error $|\hat{x} - x|/|x|$ and the absolute error $|\hat{x} - x|$.

Absolute errors will be bounded by ulps of the exact result, using the function $t \mapsto \mathrm{ulp}(t)$ defined over the reals as follows [2]: $\mathrm{ulp}(0) = 0$ and for t nonzero, $\mathrm{ulp}(t)$ is the unique integer power of β such that $\beta^{p-1} \le |t|/\mathrm{ulp}(t) < \beta^p$. In particular, $u = \frac{1}{2}\mathrm{ulp}(1)$ and

$$\mathrm{ulp}(t) \le 2u|t| \quad \text{for any real number } t. \tag{1.1}$$

Main results. Barring underflow and overflow and under mild assumptions on β and p, we show that our absolute error bound is optimal and that our relative error bound is asymptotically optimal. Here, *optimal* means that the error bound is attainable, and *asymptotically optimal* means there are inputs a, b, c, d for which the ratio (error)/(error bound) has the form $1 - O(\beta^{-p})$.

Theorem 1.1. *If no underflow or overflow occurs then* $|\hat{x} - x| \leq \frac{\beta+1}{2}\mathrm{ulp}(x)$ *and, when* $\beta/2$ *is odd and* $p \geq 4$*, this absolute error bound is optimal.*

Combining this result with (1.1), we immediately deduce the relative error bound $(\beta + 1)u$. However, the next theorem shows that the factor $\beta + 1$ can be improved to 2, which is both smaller and independent of the radix, and that this constant is essentially the best possible.

Theorem 1.2. *If no underflow or overflow occurs then* $|\hat{x} - x| \leq 2u|x|$ *and, when* β *is even, this relative error bound is asymptotically optimal.*

Note that for both theorems the assumptions on β and p are satisfied by *all* the formats of IEEE floating-point arithmetic [8].

Remark 1.3. A floating-point number $\hat{t}$ is a *faithful rounding* of a real number t if it equals t or one of the two floating-point numbers surrounding t [17]. In particular, if $\hat{t}$ is a faithful rounding of t then $|\hat{t}-t| \leq \mathrm{ulp}(t)$. Theorem 1.1 implies that Kahan's algorithm sometimes generates an absolute error as large as $\frac{\beta+1}{2}$ ulps, which shows that a faithfully rounded result is not always returned.

Outline. This paper is organized as follows:

- Section 2 gives the main definitions and assumptions, recalls the classic error analysis of Algorithm 1, and introduces several useful properties.
- In Section 3 we show that Algorithm 1 is always accurate by bounding the relative error by $(\beta + 1)u + \beta u^2$, and the absolute error by β ulps of x. Although we provide sharper bounds later on in the paper, we have kept this section because the properties it contains will be needed in the next sections and also because these first bounds are relatively easy to derive.
- Section 4 presents our two main results: the absolute error is bounded by $(\beta + 1)/2$ ulps of x (Theorem 1.1) and the relative error is bounded by $2u$ (Theorem 1.2).
- In Section 5 we give relative error bounds parametrized by the difference $\sigma =$ (exponent of ad) $-$ (exponent of bc). These bounds are smaller than $2u$ as soon as $\sigma \leq -p - 3$ or $\sigma \geq 3$, and tend to u as $|\sigma| \to \infty$. Such results should be useful if, depending on the problem under consideration, we know further that the inputs satisfy $|ad| \gg |bc|$ or $|ad| \ll |bc|$.
- Section 6 concludes with some special cases. First, we consider the case of the computation of $ad - bc$ when ad and bc have opposite signs. This situation covers in particular the computation of a^2+b^2, which occurs when computing 2D Euclidean norms and performing complex divisions. Then we consider, in binary floating-point arithmetic, the special case of evaluation of $a^2 - bc$ or $ad - b^2$, which covers the computation of the discriminant of a quadratic equation.

2. Preliminaries

2.1. **Definitions and assumptions.** Throughout this paper $\mathbb{F}$ denotes the set $\{0\} \cup \{S \cdot \beta^{e-p+1} : S, e \in \mathbb{Z},\ \beta^{p-1} \leq |S| < \beta^p\}$ of radix-β, precision-p floating-point numbers, assuming that

$$\beta \text{ is even}, \quad p \geq 2, \quad \text{and the exponent range is unbounded;}$$

in addition, we assume that the inputs a, b, c, d to Algorithm 1 belong to $\mathbb{F}$. The variables $\hat{w}$, e, $\hat{f}$, $\hat{x}$ obtained by rounding to nearest are thus also in $\mathbb{F}$. *All the results in this paper are proved under such assumptions and remain true for IEEE floating-point arithmetic as long as underflow or overflow does not occur*, since $\beta \in \{2, 10\}$ and $p \geq 7$ for all the standard formats [8, p. 13].

Assuming an unbounded exponent range implies in particular that

$$|\mathrm{RN}(t) - t| \leq u|t| \quad \text{for any real number } t.$$

Hence the exact result t of a floating-point operation like multiply, add, or fused multiply-add is related to its correctly-rounded value $\hat{t} = \mathrm{RN}(t)$ by the identity below, referred to as the *standard model* of floating-point arithmetic [6, p. 40]:

$$\hat{t} = t\,(1 + \delta), \qquad |\delta| \leq u. \tag{2.1}$$

The standard model is not the only property of rounding to nearest, and we also have the following:

(i) $|\mathrm{RN}(t) - t| = \min_{s \in \mathbb{F}} |s - t| \leq \frac{1}{2}\mathrm{ulp}(t)$;
(ii) $\mathrm{RN}(\beta^k t) = \beta^k \mathrm{RN}(t)$ for any $k \in \mathbb{Z}$;
(iii) $|\mathrm{RN}(t)| = \mathrm{RN}(|t|)$.

Furthermore, by definition of $\mathbb{F}$,

(iv) the significand of any s in $\mathbb{F}\backslash\{0\}$ is an integer such that $\beta^{p-1} \leq |S| < \beta^p$.

While Section 2.2 uses just the standard model, all our results from Section 2.3 onwards exploit at least one of the lower level properties (i)–(iv).

Finally, besides the variables x, $\hat{w}$, e, $\hat{f}$, $\hat{x}$ introduced in Algorithm 1, we define

$$f = ad - \hat{w},$$

from which it follows that

$$\hat{f} = \mathrm{RN}(f) \quad \text{and} \quad x = f + e.$$

Our analyses will repeatedly use f as well as the error terms ϵ_1 and ϵ_2 given by

$$\hat{f} = f + \epsilon_1 \quad \text{and} \quad \hat{x} = \hat{f} + e + \epsilon_2. \tag{2.2}$$

From the last three identities we deduce that

$$\hat{x} - x = \epsilon_1 + \epsilon_2; \tag{2.3}$$

also, for $|\epsilon_1|$ and $|\epsilon_2|$ being the absolute errors due to rounding f and $\hat{f} + e$ to nearest, we have

$$|\epsilon_1| \leq \tfrac{1}{2}\mathrm{ulp}(f) \quad \text{and} \quad |\epsilon_2| \leq \tfrac{1}{2}\mathrm{ulp}\big(\hat{f} + e\big). \tag{2.4}$$

2.2. **Rounding error analysis in the standard model.** By using the standard model (2.1), Higham [6, solution to Problem 2.27] concludes that Kahan's algorithm offers high relative accuracy as long as $u|bc| \not\gg |x|$. More precisely, we have the bound

$$|\hat{x} - x| \leq J|x| \quad \text{with} \quad J = 2u + u^2 + (u + u^2)u|bc|/|x|, \tag{2.5}$$

which can be derived as follows: since $\hat{x} = \mathrm{RN}\big(\hat{f} + e\big)$ and $\hat{f} = \mathrm{RN}(f)$,

$$\hat{x} = \big(f(1 + \delta_1) + e\big)(1 + \delta_2), \qquad |\delta_1|, |\delta_2| \leq u;$$

using $f = x - e$, we deduce that $\hat{x} - x = x(\delta_1 + \delta_2 + \delta_1\delta_2) - e\delta_1(1 + \delta_2)$ and then

$$|\hat{x} - x| \leq (2u + u^2)|x| + (u + u^2)|e|; \tag{2.6}$$

finally, applying $|e| = |\mathrm{RN}(bc) - bc| \leq u|bc|$ to (2.6) leads to the bound in (2.5).

However, with such a bound, high relative accuracy is ensured *a priori* only when $u|bc|$ is not "large" compared to $|x|$, which is not always the case. To see this, consider for example

$$(a, b, c, d) = (N - 1, N, N, N + 1) \quad \text{with } N = \beta^p - 1. \tag{2.7}$$

One may check that $a, b, c, d \in \mathbb{F}$ and that $|bc|/|x| = N^2 \geq (u^{-1} - 1)^2$. Thus, the relative error bound J can be as large as $1 + u + u^3 > 1$ and one cannot even conclude that Algorithm 1 always computes the sign of the result correctly.

Of course, that this bound can be large does not mean that the maximum error must be large too. In the above example the computation is in fact exact, since both x and $\hat{x}$ are equal to -1. Although Algorithm 1 does not usually provide the exact answer, we shall see in Section 3.1 that it always yields an approximation having high relative accuracy. To arrive at this conclusion we will bound $|e|/|x|$ independently of u^{-1} and then combine this bound with the inequality in (2.6).

2.3. **Preliminary properties.** Our analysis of Kahan's algorithm will use several basic properties which we introduce now. First, in the special cases where bc or f is a floating-point number, Algorithm 1 behaves ideally: as the property below shows, $\hat{x}$ is the correctly-rounded result and thus $|\hat{x} - x| \leq \frac{1}{2}\mathrm{ulp}(x) \leq u|x|$.

Property 2.1. If $bc \in \mathbb{F}$ or $f \in \mathbb{F}$ then $\hat{x} = \mathrm{RN}(x)$.

Proof. If $bc \in \mathbb{F}$ then $e = 0$, which implies $\hat{x} = \hat{f} = \mathrm{RN}(x)$. If $f \in \mathbb{F}$ then $\hat{f} = ad - \hat{w}$, so that $\hat{f} + e = x$ and then $\hat{x} = \mathrm{RN}(x)$. □

Therefore, we shall focus most of our efforts on analysing Algorithm 1 under the following genericity condition:

$$bc \notin \mathbb{F} \qquad \text{and} \qquad f \notin \mathbb{F}. \tag{C}$$

The next property gives two useful consequences of that condition:

Property 2.2. If Condition (C) holds then $abcd \neq 0$ and $x \neq 0$.

Proof. If $ad = 0$ then $f = -\hat{w}$ belongs to $\mathbb{F}$, so that $f \notin \mathbb{F}$ implies $ad \neq 0$. Since $bc \notin \mathbb{F}$ implies $bc \neq 0$, we deduce that (C) implies $abcd \neq 0$. If $x = 0$ then $ad = bc$ and thus $f = -e$. Since e belongs to $\mathbb{F}$, we conclude that $f \notin \mathbb{F}$ implies $x \neq 0$. □

When the floating-point numbers a, b, c, d are nonzero, which is implied by Condition (C), they can be written $a = A\,\beta^{e_a - p + 1}$, $b = B\,\beta^{e_b - p + 1}$, $c = C\,\beta^{e_c - p + 1}$, $d = D\,\beta^{e_d - p + 1}$ for some integers $e_a, e_b, e_c, e_d, A, B, C, D$ such that

$$\beta^{p-1} \leq |A|, |B|, |C|, |D| < \beta^p. \tag{2.8a}$$

Thus, the ratio ad/bc has the form $AD/BC \cdot \beta^\sigma$ with $\sigma \in \mathbb{Z}$ given by

$$\sigma = e_a + e_d - e_b - e_c. \tag{2.8b}$$

Furthermore, we can now associate to $\hat{w}$, e, f, $\hat{f}$, $\hat{x}$, and x the following integers:

$$\hat{W} = \mathrm{RN}(BC) \qquad \text{and} \qquad E = \begin{cases} \hat{W} - BC & \text{if } \sigma \geq 0, \\ \big(\hat{W} - BC\big)\beta^{-\sigma} & \text{if } \sigma < 0; \end{cases} \tag{2.9a}$$

$$F = \begin{cases} AD\beta^\sigma - \hat{W} & \text{if } \sigma \geq 0, \\ AD - \hat{W}\beta^{-\sigma} & \text{if } \sigma < 0, \end{cases} \qquad \text{and} \qquad \hat{F} = \mathrm{RN}(F); \tag{2.9b}$$

$$X = F + E \qquad \text{and} \qquad \hat{X} = \mathrm{RN}(\hat{F} + E). \tag{2.9c}$$

Property 2.3. Assume $abcd \neq 0$ and let $E, F, \hat{F}, X, \hat{X}$ be defined as in (2.9). Then there exists $\mu \in \mathbb{Z}$ such that $e = E\beta^\mu$, $f = F\beta^\mu$, $\hat{f} = \hat{F}\beta^\mu$, $x = X\beta^\mu$, and $\hat{x} = \hat{X}\beta^\mu$.

Proof. Let $\mu = \min\{e_a + e_d, e_b + e_c\} - 2p + 2$. Assume first that $\sigma \geq 0$. Then $X = AD\beta^\sigma - BC$ and $\mu = e_b + e_c - 2p + 2$, from which it follows that $X\beta^\mu = ad - bc = x$, as wanted. Furthermore, since $\mathrm{RN}(BC)\beta^\mu = \mathrm{RN}(BC\beta^\mu) = \mathrm{RN}(bc)$, we deduce that $F\beta^\mu = ad - \hat{w} = f$. It then follows that $E\beta^\mu = (X - F)\beta^\mu = x - f = e$, and that $\hat{F}\beta^\mu = \mathrm{RN}(F\beta^\mu) = \mathrm{RN}(f) = \hat{f}$. Finally, $\hat{X}\beta^\mu = \mathrm{RN}(\hat{F}\beta^\mu + E\beta^\mu) = \mathrm{RN}(\hat{f} + e) = \hat{x}$. The case where $\sigma < 0$ can be handled similarly, using now the identities $X = AD - BC\beta^{-\sigma}$ and $\mu = e_a + e_d - 2p + 2$. □

We conclude these preliminaries with three facts that will be useful in the sequel:

$$\beta^{2p-2} \leq |AD|, |BC| \leq \beta^{2p} - 2\beta^p + 1, \tag{2.10a}$$

$$\mathrm{ulp}(BC) = \begin{cases} \beta^{p-1} & \text{if } |BC| < \beta^{2p-1}, \\ \beta^p & \text{otherwise,} \end{cases} \tag{2.10b}$$

$$\beta^{2p-2} \leq |\hat{W}| = \mathrm{RN}(|BC|) \leq \beta^{2p} - \beta^p. \tag{2.10c}$$

3. Kahan's algorithm is always highly accurate: First bounds

3.1. **Bounding the relative error by $(\beta + 1)u + \beta u^2$.** We show in this section that Algorithm 1 always approximates x to high relative accuracy. For this, we begin by proving the following result.

Lemma 3.1. *If Condition (C) holds then $|e|/|x| \leq \beta - 1$.*

Proof. By Property 2.2 the numbers a, b, c, d, and x are nonzero, and using Property 2.3 gives $|e|/|x| = |E|/|X|$. Furthermore, since $f \notin \mathbb{F}$ we have $|F| > \beta^p$. To show that $|E|/|X|$ is bounded by $\beta - 1$ we consider separately the following three cases, depending on the value of σ.

■ If $\sigma \geq 0$ then $E = \mathrm{RN}(BC) - BC$ and, since $\mathrm{ulp}(BC) \leq \beta^p$, we have $|E| \leq \frac{1}{2}\beta^p$. Hence $|X| \geq |F| - |E| > \beta^p - \frac{1}{2}\beta^p = \frac{1}{2}\beta^p$, from which we deduce that $|E| < |X|$. Since $\beta \geq 2$, this implies $|E| \leq (\beta - 1)|X|$.

■ Assume now that $\sigma = -1$. In this case we have $E = \big(\mathrm{RN}(BC) - BC\big)\beta$. If $|BC| \leq \beta^{2p-1}$ then $|E| \leq \frac{1}{2}\beta^{p-1} \cdot \beta = \frac{1}{2}\beta^p$, and we proceed as for the case "$\sigma \geq 0$". If $|BC| > \beta^{2p-1}$ then, using $X = AD - BC\beta$ and $|A|, |D| \leq \beta^p - 1$ we see that $|X| \geq \beta|BC| - |AD| > \beta \cdot \beta^{2p-1} - (\beta^{2p} - 2\beta^p + 1) = 2\beta^p - 1$; hence $|X| \geq 2\beta^p$ and since in this case $|E| \leq \frac{1}{2}\beta^p \cdot \beta = \frac{1}{2}\beta^{p+1}$, we obtain $|E|/|X| \leq \beta/4$, which for $\beta \geq 2$ implies $|E| \leq (\beta - 1)|X|$.

■ Assume that $\sigma \leq -2$. Setting $i = -\sigma$, we have $i \geq 2$ and $|E| \leq \frac{1}{2}\beta^{p+i}$. Since $bc \notin \mathbb{F}$ we have $|B|, |C| > \beta^{p-1}$; from $X = AD - BC\beta^i$ and $|A|, |D| < \beta^p$ it follows that $|X| \geq (\beta^{p-1} + 1)^2\beta^i - (\beta^p - 1)^2$. Noting that $\beta^i - 1 \geq 0$, we obtain

$$|X| \geq \beta^{p+i} y, \qquad y := \beta^{p-2} + 2\beta^{-1} - (\beta^p - 2)\beta^{-i}. \tag{3.1}$$

For $\beta, i \geq 2$ we have $\beta^{-i} \leq \beta^{-2}$ and thus $y \geq 2\beta^{-1}$. Hence $|X| \geq 2\beta^{p+i-1}$, which together with $|E| \leq \frac{1}{2}\beta^{p+i}$ gives again $|E|/|X| \leq \beta/4 \leq \beta - 1$. □

Using this lemma we can prove the following relative error bound for Algorithm 1.

Proposition 3.2. $|\hat{x} - x| \le K|x|$ *with* $K = (\beta+1)u + \beta u^2$.

Proof. If (C) holds then applying Lemma 3.1 to (2.6) gives the result. If (C) does not hold then $|\hat{x} - x| \le u|x|$ by Property 2.1, and $u \le K$ for $\beta, p \ge 2$. □

One can check that $(\beta+1)u \le 3/4$ and $\beta u^2 \le 1/8$ for $\beta, p \ge 2$. Hence the relative error bound K satisfies $K \le 7/8$, from which it follows that

$$\hat{x} \text{ and } x \text{ have the same sign.} \tag{3.2}$$

Recall from Section 2.2 that this conclusion could not have been obtained using the classic relative error bound J in (2.5).

Furthermore, for all the formats of IEEE arithmetic u^2 is much smaller than u, so that $K \approx (\beta+1)u$. In particular, $K \approx 3u$ for radix 2.

Remark 3.3. This relative error bound K is not the best possible, and we will prove in Section 4 a bound equal to $2u$. However, it has been derived from Lemma 3.1, whose upper bound $\beta - 1$ on $|e|/|x|$ is asymptotically optimal for radix 2: for example, if $p \ge 6$ is even and if $(a, b, c, d) = (N+3, N, N'+3, N')$ with $N = 3 \cdot 2^{p-2}$ and $N' = 7 \cdot 2^{p-3} + \frac{2^{p-3}+1}{3}$ then one can check that $a, b, c, d \in \mathbb{F}$ and that $bc = 11 \cdot 2^{2p-4} + 5 \cdot 2^{p-1} \notin \mathbb{F}$, $e = -2^{p-1} \in \mathbb{F}$, $x = 2^{p-1}+1 \in \mathbb{F}$, and $f = x - e = 2^p + 1 \notin \mathbb{F}$. Thus, in this case, (C) holds and the ratio $|e|/|x|$ is in $1 - O(2^{-p})$ as $p \to \infty$.

3.2. **Bounding the absolute error by β ulps of x.** We now turn to absolute error bounds expressed in ulps of the exact result. Note first that using Proposition 3.2 together with the fact that $|t| \le \frac{\beta}{2u}\mathrm{ulp}(t)$ for any t leads immediately to $|\hat{x} - x| \le \frac{\beta K}{2u}\mathrm{ulp}(x) \approx \frac{\beta(\beta+1)}{2}\mathrm{ulp}(x)$. However, this bound can be improved to β ulps of the exact result by using the error terms ϵ_1 and ϵ_2 introduced in (2.2). To show this, we essentially bound $|\epsilon_1|$ and $|\epsilon_2|$ separately and then use the fact that (2.3) implies $|\hat{x} - x| \le |\epsilon_1| + |\epsilon_2|$. Our bounds for $|\epsilon_1|$ and $|\epsilon_2|$ are given in the lemma below; they will also be key ingredients for establishing the optimal or asymptotically optimal error bounds of Section 4.

Lemma 3.4. $|\epsilon_i| \le \frac{\beta}{2}\mathrm{ulp}(x)$ *for* $i = 1, 2$.

Proof. Assume first that Condition (C) holds. From $x = f + e$ and Lemma 3.1 we deduce that $|f| \le |x| + |e| \le \beta|x|$, which by definition of the ulp function implies $\mathrm{ulp}(f) \le \beta\mathrm{ulp}(x)$. Hence, using (2.4), $|\epsilon_1| \le \frac{\beta}{2}\mathrm{ulp}(x)$. To show a similar bound on $|\epsilon_2|$, first we combine (2.2) with $x = f + e$ to obtain

$$\hat{f} + e = x + \epsilon_1. \tag{3.3}$$

Then, using the upper bound on $|\epsilon_1|$ and since x is nonzero by Property 2.2, we get

$$\mathrm{ulp}\big(\hat{f} + e\big) \le \beta^{1-p}|x| + \tfrac{\beta}{2}\beta^{1-p}\mathrm{ulp}(x) < L\mathrm{ulp}(x), \qquad L = \beta(1+u).$$

For $\beta, p \ge 2$ one can check that $\beta < L \le \beta^2$, so that $\mathrm{ulp}\big(\hat{f} + e\big) < \beta^2\mathrm{ulp}(x)$. Hence, ulps being integer powers of β, $\mathrm{ulp}\big(\hat{f} + e\big) \le \beta\mathrm{ulp}(x)$ and we conclude that $|\epsilon_2| \le \frac{\beta}{2}\mathrm{ulp}(x)$ using (2.4).

If (C) does not hold then one can check that (ϵ_1, ϵ_2) equals $(\mathrm{RN}(x) - x, 0)$ if $bc \in \mathbb{F}$, and $(0, \mathrm{RN}(x) - x)$ if $f \in \mathbb{F}$. This implies $|\epsilon_i| \le \frac{1}{2}\mathrm{ulp}(x)$ for $i = 1, 2$ and the conclusion follows. □

An immediate consequence of these bounds on $|\epsilon_1|$ and $|\epsilon_2|$ is the following result.

Proposition 3.5. $|\hat{x} - x| \leq \beta \mathrm{ulp}(x)$.

Another consequence of Lemma 3.4 is an alternative proof of Proposition 3.2: using (3.3) we have $|\epsilon_2| \leq \frac{1}{2}\mathrm{ulp}(x+\epsilon_1) \leq u|x+\epsilon_1|$ and then

$$|\hat{x} - x| \leq u|x| + (1+u)|\epsilon_1|.$$

Now, $|\epsilon_1| \leq \frac{\beta}{2}\mathrm{ulp}(x) \leq \beta u|x|$ by Lemma 3.4, so that $|\hat{x} - x|/|x|$ is bounded by $u + (1+u)\beta u$, which is precisely the constant K of Proposition 3.2.

Remark 3.6. We will see in Section 4 that the constant β in the bound of Proposition 3.5 can be improved to $(\beta+1)/2$. However, the bounds on $|\epsilon_1|$ and $|\epsilon_2|$ on which this proposition relies are essentially the best possible: assuming rounding "to nearest even", β even, and $p \geq 5$, Example 3.7 below provides inputs a, b, c, d for which both $|\epsilon_1|$ and $|\epsilon_2|$ are asymptotically equivalent to $\frac{\beta}{2}\mathrm{ulp}(x)$ as p tends to infinity. (In fact, as shown in Example 4.4 the bound on $|\epsilon_1|$ is attainable assuming rounding "to nearest even", $\beta \geq 2$, and $p \geq 4$.)

Example 3.7 (Example for which the ratios $|\epsilon_i|/(\frac{\beta}{2}\mathrm{ulp}(x))$, $i = 1, 2$ both tend to 1 as $p \to \infty$, assuming rounding "to nearest even," β even, and $p \geq 5$)**.** We consider two cases, depending on the parity of p.

• When p is odd, let

$$\begin{aligned} a &= \beta^{p-1} + \beta^{p-3} + \tfrac{\beta}{2}\beta^{p-4}, \\ b &= \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-4} + \beta^{\frac{p-5}{2}}, \\ c &= \beta^{p-1} + \beta^{\frac{p-1}{2}} + 1, \\ d &= c. \end{aligned}$$

One can check that

$$ad = \beta^{2p-2} + \beta^{2p-4} + \tfrac{\beta}{2}\beta^{2p-5} + \beta^{\frac{3p-3}{2}} + \beta^{\frac{3p-7}{2}} + \tfrac{\beta}{2}\beta^{\frac{3p-9}{2}} + \beta^{p-1} + \beta^{p-3} + \tfrac{\beta}{2}\beta^{p-4},$$

$$bc = \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-5} + \beta^{\frac{3p-3}{2}} + \beta^{\frac{3p-7}{2}} + \tfrac{\beta}{2}\beta^{\frac{3p-9}{2}} + \beta^{p-1} + \beta^{p-3} + \tfrac{\beta}{2}\beta^{p-4} + \beta^{\frac{p-5}{2}},$$

so that $x = \beta^{2p-4} - \beta^{\frac{p-5}{2}}$, and $\mathrm{ulp}(x) = \beta^{p-4}$. Since $\mathrm{ulp}(bc) = \beta^{p-1}$,

$$\hat{w} = \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-5} + \beta^{\frac{3p-3}{2}} + \beta^{\frac{3p-7}{2}} + \tfrac{\beta}{2}\beta^{\frac{3p-9}{2}} + \beta^{p-1},$$

and $e = -\beta^{p-3} - \frac{\beta}{2}\beta^{p-4} - \beta^{\frac{p-5}{2}}$. Moreover, $f = \beta^{2p-4} + \beta^{p-3} + \frac{\beta}{2}\beta^{p-4}$, and since $\mathrm{ulp}(f) = \beta^{p-3}$ rounding "to nearest even" gives $\hat{f} = \beta^{2p-4} + 2\beta^{p-3}$ and $\epsilon_1 = \frac{\beta}{2}\beta^{p-4}$. Then, $\hat{f} + e = \beta^{2p-4} + \beta^{p-3} - \frac{\beta}{2}\beta^{p-4} - \beta^{\frac{p-5}{2}}$, and since $\mathrm{ulp}(\hat{f}+e) = \beta^{p-3}$, one has $\hat{x} = \mathrm{RN}(\hat{f}+e) = \beta^{2p-4}$ and $\epsilon_2 = -\frac{\beta}{2}\beta^{p-4} + \beta^{\frac{p-5}{2}}$. As a consequence,

$$|\epsilon_1| = \tfrac{\beta}{2}\mathrm{ulp}(x) \quad \text{and} \quad |\epsilon_2| = \left(\tfrac{\beta}{2} - \beta^{\frac{-p+3}{2}}\right)\mathrm{ulp}(x).$$

• When p is even, consider

$$\begin{aligned} a &= \beta^{p-1} + \beta^{p-2} + \tfrac{\beta}{2}\beta^{\frac{p-6}{2}}, \\ b &= \beta^{p-1} + \beta^{\frac{p}{2}} + \beta, \\ c &= \beta^{p-1} + \beta^{\frac{p}{2}-1} + \tfrac{\beta}{2}\beta^{\frac{p-6}{2}}, \\ d &= b. \end{aligned}$$

Then

$$ad = \beta^{2p-2} + \beta^{2p-3} + \beta^{\frac{3p-2}{2}} + \beta^{\frac{3p-4}{2}} + \tfrac{\beta}{2}\beta^{\frac{3p-8}{2}} + \beta^{p} + \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-3} + \tfrac{\beta}{2}\beta^{\frac{p-4}{2}},$$

$$bc = \beta^{2p-2} + \beta^{\frac{3p-2}{2}} + \beta^{\frac{3p-4}{2}} + \tfrac{\beta}{2}\beta^{\frac{3p-8}{2}} + \beta^{p} + \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-3} + \beta^{\frac{p}{2}} + \tfrac{\beta}{2}\beta^{\frac{p-4}{2}},$$

from which it follows that $x = \beta^{2p-3} - \beta^{\frac{p}{2}}$ and $\mathrm{ulp}(x) = \beta^{p-3}$. Since $\mathrm{ulp}(bc) = \beta^{p-1}$,

$$\hat{w} = \beta^{2p-2} + \beta^{\frac{3p-2}{2}} + \beta^{\frac{3p-4}{2}} + \frac{\beta}{2}\beta^{\frac{3p-8}{2}} + \beta^{p} + \beta^{p-1},$$

so that $e = -\frac{\beta}{2}\beta^{p-3} - \beta^{\frac{p}{2}} - \frac{\beta}{2}\beta^{\frac{p-4}{2}}$. On the other hand, $f = \beta^{2p-3} + \frac{\beta}{2}\beta^{p-3} + \frac{\beta}{2}\beta^{\frac{p-4}{2}}$, and since $\mathrm{ulp}(f) = \beta^{p-2}$, one has $\hat{f} = \beta^{2p-3} + \beta^{p-2}$ and $\epsilon_1 = \frac{\beta}{2}\beta^{p-3} - \frac{\beta}{2}\beta^{\frac{p-4}{2}}$. Therefore, $\hat{f}+e = \beta^{2p-3} + \beta^{p-2} - \frac{\beta}{2}\beta^{p-3} - \beta^{\frac{p}{2}} - \frac{\beta}{2}\beta^{\frac{p-4}{2}}$, and since $\mathrm{ulp}(\hat{f}+e) = \beta^{p-2}$, we deduce that $\hat{x} = \beta^{2p-3}$ and $\epsilon_2 = -\frac{\beta}{2}\beta^{p-3} + \left(1 + \frac{1}{2\beta}\right)\beta^{\frac{p}{2}}$. To summarize, in this case one has

$$|\epsilon_1| = \left(\frac{\beta}{2} - \frac{1}{2}\beta^{\frac{-p+4}{2}}\right)\mathrm{ulp}(x) \quad \text{and} \quad |\epsilon_2| = \left(\frac{\beta}{2} - \left(1 + \frac{1}{2\beta}\right)\beta^{\frac{-p+6}{2}}\right)\mathrm{ulp}(x).$$

4. Optimal or asymptotically optimal error bounds

The results obtained in the previous section already show that Kahan's algorithm is always highly accurate: the absolute error is at most β ulps of the exact result and the relative error is at most $Ku \approx (\beta+1)u$. In this section we shall improve these bounds to $(\beta+1)/2$ ulps and $2u$, respectively, and show that these new bounds are optimal or asymptotically optimal. First, we proceed by treating the cases $|\epsilon_2| \le \frac{1}{2}\mathrm{ulp}(x)$ and $|\epsilon_2| > \frac{1}{2}\mathrm{ulp}(x)$ separately. In the first case (Section 4.1) we show that the absolute error is bounded by $|\epsilon_1| + \frac{1}{2}\mathrm{ulp}(x)$ and that the relative error is bounded by $2u$. In the second case (Section 4.2) it turns out that smaller bounds can be obtained: the absolute error is at most $|\epsilon_1|$, while the relative error has the form $u + O(u^2)$. Then, combining these results with specific input examples, we can prove Theorems 1.1 and 1.2, that is, conclude under mild assumptions on β and p that $(\beta+1)/2$ ulps is the best possible absolute error bound (Section 4.3) and that $2u$ is an asymptotically optimal relative error bound (Section 4.4).

4.1. Absolute and relative error bounds when $|\epsilon_2| \le \frac{1}{2}\mathbf{ulp}(x)$.

Lemma 4.1. *If* $|\epsilon_2| \le \frac{1}{2}\mathrm{ulp}(x)$ *then* $|\hat{x} - x| \le |\epsilon_1| + \frac{1}{2}\mathrm{ulp}(x)$ *and* $|\hat{x} - x| \le 2u|x|$.

Proof. The absolute error bound follows from $|\hat{x} - x| \le |\epsilon_1| + |\epsilon_2|$, and it remains to show that the relative error is bounded by $2u$. If (C) does not hold or if $|\epsilon_1| \le \frac{1}{2}\mathrm{ulp}(x)$ then the result is clear, so assume that (C) holds and that $\frac{1}{2}\mathrm{ulp}(x) < |\epsilon_1|$. First, by using Lemma 3.4 and the absolute error bound just shown, we obtain

$$|\hat{x} - x| \le \tfrac{\beta+1}{2}\mathrm{ulp}(x). \tag{4.1}$$

Since $|\epsilon_1| \le \frac{1}{2}\mathrm{ulp}(f)$ by (2.4), we have $\mathrm{ulp}(x) \le \beta^{-1}\mathrm{ulp}(f)$; on the other hand, using $x = f + e$ and Lemma 3.1 we obtain $|f| \le |x| + |e| \le \beta|x|$ and thus $\mathrm{ulp}(f) \le \beta\mathrm{ulp}(x)$. Therefore,

$$\mathrm{ulp}(x) = \beta^{-1}\mathrm{ulp}(f).$$

Recalling from Property 2.2 that (C) implies $x \neq 0$, we deduce that

$$|x| < \beta^{p}\mathrm{ulp}(x) \le |x| + |e|.$$

By Property 2.3 and since $x \neq 0$ implies $\mathrm{ulp}(x) > 0$, we have

$$0 < \beta^{p} - \frac{|x|}{\mathrm{ulp}(x)} \le \eta, \qquad \eta = \frac{|E|}{\mathrm{ulp}(X)}.$$

This gives an upper bound on $\mathrm{ulp}(x)/|x|$ which, combined with (4.1), leads to $|\hat{x}-x| \le K'|x|$ with $K' = \frac{\beta+1}{2(\beta^p-\eta)}$; one may check that $K' \le 2u$ as soon as

$$\eta \le \tfrac{1}{2}(\beta-1)\beta^{p-1}. \tag{4.2}$$

On the other hand, since $f \notin \mathbb{F}$ by (C), we have $|F| > \beta^p$ and thus $\mathrm{ulp}(F) \ge \beta$. Now, $\mathrm{ulp}(f) = \beta\mathrm{ulp}(x)$ is equivalent to $\mathrm{ulp}(F) = \beta\mathrm{ulp}(X)$, and we obtain

$$\mathrm{ulp}(X) \ge 1 \qquad \text{and} \qquad |X| < |F|.$$

■ Assume $\sigma \ge 0$. In this case $|E| \le \frac{1}{2}\beta^p$. If $\mathrm{ulp}(X) \ge \beta$ then $\eta \le \frac{1}{2}\beta^{p-1}$ and, since $\beta \ge 2$, (4.2) follows. If $\mathrm{ulp}(X) = 1$ then (4.1) leads to

$$\frac{|\hat{X}-X|}{|X|} = \frac{|\hat{x}-x|}{|x|} \le \tfrac{\beta+1}{2} \cdot \frac{\mathrm{ulp}(x)}{|x|} = \tfrac{\beta+1}{2} \cdot \frac{1}{|X|}.$$

Hence $|\hat{X}-X| \le \frac{\beta}{2}+\frac{1}{2}$ and since $X, \hat{X} \in \mathbb{Z}$ and β is even, $|\hat{X}-X| \le \frac{\beta}{2}$. Thus $|\hat{x}-x|/|x| \le \frac{\beta}{2}/|X|$ and, using $|X| \ge |F|-|E| > \frac{1}{2}\beta^p$, we get $|\hat{x}-x| < 2u|x|$.

■ Assume $\sigma = -1$. If $|BC| \le \beta^{2p-1}$ then $|E| \le \frac{1}{2}\beta^p$ and we can conclude exactly as for the case "$\sigma \ge 0$" detailed above. Let us now assume $|BC| > \beta^{2p-1}$. Then $|E| \le \frac{1}{2}\beta^{p+1}$. If $\mathrm{ulp}(X) \ge \beta^2$ then $\eta \le \frac{1}{2}\beta^{p-1}$ and (4.2) follows, and we are left with the case where $\mathrm{ulp}(X) \le \beta$. Since $\mathrm{RN}(|BC|)$ is either equal to β^{2p-1} or at least $\beta^{2p-1}+\beta^p$, we consider these two subcases separately:

- If $\mathrm{RN}(|BC|) \ge \beta^{2p-1}+\beta^p$ then, using $|F| \ge \beta\,\mathrm{RN}(|BC|) - |AD|$, we get
$$|F| \ge \beta(\beta^{2p-1}+\beta^p) - (\beta^p-1)^2 > \beta^{p+1} + \tfrac{1}{2}\beta^p.$$
From $|X| \ge |F|-|E|$, we deduce that $|X| > \frac{1}{2}(\beta+1)\beta^p$. Since $\mathrm{ulp}(x)/|x| = \mathrm{ulp}(X)/|X|$, it follows from (4.1) and $\mathrm{ulp}(X) \le \beta$ that $|\hat{x}-x| < 2u|x|$.
- Let us now show that $\mathrm{RN}(|BC|) = \beta^{2p-1}$ is not possible (when $|BC| > \beta^{2p-1}$). If $BC > 0$ then $E < 0$ and, since $X = F+E$ and $|X| < |F|$, we must have $F > 0$. By definition of F this implies $AD > \mathrm{RN}(BC)\beta = \beta^{2p}$; similarly, if $BC < 0$ then $E > 0$, $F < 0$, and $AD < \mathrm{RN}(BC)\beta = -\beta^{2p}$; in both cases we have $|AD| > \beta^{2p}$, which is impossible for $|A|, |D| < \beta^p$.

■ Assume $\sigma \le -2$. Let $i = -\sigma$, so that $i \ge 2$ and $|E| \le \frac{1}{2}\beta^{p+i}$. First taking $i = 2$, we deduce from $|X| \ge \beta^2|BC| - |AD|$ that $|X| > \beta^2(\beta^{p-1}+1)^2 - \beta^{2p} = 2\beta^{p+1}+\beta^2$. This implies $\mathrm{ulp}(X) \ge \beta^2$. This lower bound is enough to get (4.2) when $|BC| \le \beta^{2p-1}$, since then $|E| \le \frac{1}{2}\beta^{p+1}$; when $|BC| > \beta^{2p-1}$ we have $|E| \le \frac{1}{2}\beta^{p+2}$ but $|X|$ can now be lower bounded by β^{2p}, so that $\mathrm{ulp}(X) \ge \beta^{p+1}$ and $\eta \le \frac{1}{2}\beta$, from which (4.2) follows. Let us now take $i \ge 3$. Reusing (3.1) gives $|X| \ge \beta^{p+i}y$ and one can check that $y \ge 1$ for $\beta, p \ge 2$. Hence $\mathrm{ulp}(X) \ge \beta^{i+1}$ and, recalling that $|E| \le \frac{1}{2}\beta^{p+i}$, we obtain $\eta \le \frac{1}{2}\beta^{p-1}$ and then (4.2). □

4.2. Absolute and relative error bounds when $|\epsilon_2| > \frac{1}{2}\mathbf{ulp}(x)$.

Lemma 4.2. *If $|\epsilon_2| > \frac{1}{2}\mathrm{ulp}(x)$ then $|\hat{x}-x| \le |\epsilon_1|$ and $|\hat{x}-x| \le L|x|$ with $L = \frac{|\epsilon_1|}{\beta^p\mathrm{ulp}(x)-|\epsilon_1|} \le \frac{u}{1-u}$.*

Proof. Note first that the assumption on ϵ_2 implies x must be nonzero. To see this, recall that the ulp function is nonnegative, and that $x = 0$ implies $f = -e = \hat{f}$ and thus $|\epsilon_2| = 0 \le \frac{1}{2}\mathrm{ulp}(x)$.

Using (2.4) leads to $\mathrm{ulp}(x) \le \beta^{-1}\mathrm{ulp}\big(\hat{f}+e\big)$ and, as x is nonzero, we obtain

$$|x| < \beta^p\mathrm{ulp}(x) \le \beta^{p-1}\mathrm{ulp}\big(\hat{f}+e\big) \le |\hat{f}+e|. \tag{4.3}$$

Since $\hat{x} = \mathrm{RN}(\hat{f} + e)$, we deduce from (4.3) that $\beta^p \mathrm{ulp}(x) \le |\hat{x}|$. On the other hand, $|\hat{x}| \le |x| + |\hat{x} - x| < (\beta^p + \beta)\mathrm{ulp}(x)$ by Proposition 3.5. Noting further that $\mathrm{ulp}\big(\beta^p \mathrm{ulp}(x)\big) = \beta \mathrm{ulp}(x)$ and that $\hat{x} \in \mathbb{F}$ leads to

$$|\hat{x}| = \beta^p \mathrm{ulp}(x). \tag{4.4}$$

In other words, $|\epsilon_2|$ is larger than $\frac{1}{2}\mathrm{ulp}(x)$ only when $\hat{x}$ is a power of β. Since $\hat{f} + e = x + \epsilon_1$ it also follows from (4.3) that $|x| < \beta^p \mathrm{ulp}(x) \le |x| + |\epsilon_1|$ and, using (4.4), we deduce that

$$\beta^p \mathrm{ulp}(x) - |\epsilon_1| \le |x| < |\hat{x}| = \beta^p \mathrm{ulp}(x).$$

From (3.2) and $|x| < |\hat{x}|$ we obtain $|\hat{x} - x| = |\hat{x}| - |x|$ and, consequently,

$$|\hat{x} - x| = \beta^p \mathrm{ulp}(x) - |x| \le |\epsilon_1|. \tag{4.5}$$

This implies in particular $|x| \ge \beta^p \mathrm{ulp}(x) - |\epsilon_1|$. Using Lemma 3.4, one can check that this lower bound on $|x| \ne 0$ is positive for any $\beta, p \ge 2$. Hence $|\hat{x} - x|/|x| \le L$ with $L = |\epsilon_1|/(\beta^p \mathrm{ulp}(x) - |\epsilon_1|)$, and it follows from $|\epsilon_1| \le \frac{\beta}{2}\mathrm{ulp}(x)$ that L is upper bounded by $\frac{\beta/2}{\beta^p - \beta/2} = \frac{u}{1-u}$. □

Remark 4.3. In fact, we have the following implication:

$$|\epsilon_1| \le \tfrac{1}{2}\mathrm{ulp}(x) \qquad \Rightarrow \qquad |\epsilon_2| \le \tfrac{1}{2}\mathrm{ulp}(x). \tag{4.6}$$

To see this, recall from (4.3) in the proof of Lemma 4.2 that if $|\epsilon_2|$ is larger than $\frac{1}{2}\mathrm{ulp}(x)$ then $|x| < y \le |\hat{f} + e|$ for some y in $\mathbb{F}$. Since $\hat{x} = \mathrm{RN}(\hat{f} + e)$, we have $\mathrm{sign}(\hat{x}) = \mathrm{sign}(\hat{f} + e)$ and $|\hat{x}| = \mathrm{RN}(|\hat{f} + e|)$. Hence, recalling from (3.3) that $\hat{f} + e = x + \epsilon_1$,

$$|\epsilon_2| = \big||\hat{x}| - |\hat{f} + e|\big| \le |\hat{f} + e| - y < |\hat{f} + e| - |x| \le |\epsilon_1|,$$

from which it follows that $|\epsilon_1|$ must be larger than $\frac{1}{2}\mathrm{ulp}(x)$ too.

4.3. **Proof of Theorem 1.1.** Let us first check that the absolute error is indeed always bounded as

$$|\hat{x} - x| \le \tfrac{\beta+1}{2}\mathrm{ulp}(x). \tag{4.7}$$

By Lemma 4.1 and Lemma 4.2 we have $|\hat{x} - x| \le |\epsilon_1| + \frac{1}{2}\mathrm{ulp}(x)$, and Lemma 3.4 ensures further that $|\epsilon_1| \le \frac{\beta}{2}\mathrm{ulp}(x)$, thus leading to (4.7).

Assuming rounding "to nearest even", $\beta/2$ odd, and $p \ge 4$ we provide in Example 4.4 below an input (a, b, c, d) for which the absolute error $|\hat{x} - x|$ is *equal* to its bound in (4.7). This concludes the proof of Theorem 1.1. □

Example 4.4 (Example for which the absolute error bound in (4.7) is achieved, assuming rounding "to nearest even", $\beta/2$ odd, and $p \ge 4$)**.** Consider

$$\begin{aligned} a &= \beta^{p-1} + \beta^{p-3}, \\ b &= \beta^{p-1} + \tfrac{\beta}{2}, \\ c &= \beta^{p-1} + \beta^{p-3} + \beta^{p-4}, \\ d &= \beta^{p-1} + \beta^{p-2} + \tfrac{\beta}{2}. \end{aligned}$$

One can check that $a, b, c, d \in \mathbb{F}$ and that

$$bc = \beta^{2p-2} + \beta^{2p-4} + \beta^{2p-5} + \tfrac{\beta}{2}\beta^{p-1} + \tfrac{\beta}{2}\beta^{p-3} + \tfrac{\beta}{2}\beta^{p-4}.$$

This gives $\mathrm{ulp}(bc) = \beta^{p-1}$ and, recalling that $\hat{w} = \mathrm{RN}(bc)$ and $e = \hat{w} - bc$, we obtain

$$\hat{w} = \beta^{2p-2} + \beta^{2p-4} + \beta^{2p-5} + \tfrac{\beta}{2}\beta^{p-1}$$

and

$$e = -\tfrac{\beta}{2}\beta^{p-3} - \tfrac{\beta}{2}\beta^{p-4}.$$

Now, $f = ad - \hat{w}$ yields $f = \beta^{2p-3} + \frac{\beta}{2}\beta^{p-3}$, so that $\mathrm{ulp}(f) = \beta^{p-2}$ and $\hat{f} = \mathrm{RN}(f) = \beta^{2p-3}$. Hence

$$\hat{f} + e = \beta^{2p-3} - \tfrac{\beta}{2}\beta^{p-3} - \tfrac{\beta}{2}\beta^{p-4},$$

which gives

$$\hat{x} = \mathrm{RN}(\hat{f} + e) = \begin{cases} \beta^{2p-3} - \frac{\beta}{2}\beta^{p-3} & \text{if } \frac{\beta}{2} \text{ is even,} \\ \beta^{2p-3} - \left(\frac{\beta}{2} + 1\right)\beta^{p-3} & \text{if } \frac{\beta}{2} \text{ is odd.} \end{cases}$$

On the other hand, $x = ad - bc$ leads to $x = \beta^{2p-3} - \frac{\beta}{2}\beta^{p-4}$, from which we deduce $\mathrm{ulp}(x) = \beta^{p-3}$ and

$$|\hat{x} - x| = \begin{cases} \frac{\beta-1}{2}\mathrm{ulp}(x) & \text{if } \frac{\beta}{2} \text{ is even,} \\ \frac{\beta+1}{2}\mathrm{ulp}(x) & \text{if } \frac{\beta}{2} \text{ is odd.} \end{cases}$$

Thus, the absolute error bound in (4.7) is attained in this case when $\beta/2$ is odd. In addition, $\epsilon_1 = \hat{f} - f = -\frac{\beta}{2}\mathrm{ulp}(x)$, which means that for rounding "to nearest even", $\beta \geq 2$, and $p \geq 4$ the bound $\frac{\beta}{2}\mathrm{ulp}(x)$ given in Lemma 3.4 for $|\epsilon_1|$ is attained.

Remark 4.5. When $\beta/2$ is *even*—a case which seems unlikely in practice—we believe that the bound $\frac{\beta+1}{2}\mathrm{ulp}(x)$ is attainable as well; although we have so far not been able to define generic worst cases, our exhaustive simulations for small values of β and p have found inputs leading to this bound. The simplest example is $\beta = 4$, $p = 4$, $a = 81 = 1101_4$, $b = 70 = 1012_4$, $c = 69 = 1011_4$, and $d = 72 = 1020_4$.

4.4. **Proof of Theorem 1.2.** Since for $\beta, p \geq 2$ we have $u/(1-u) \leq 2u$, Lemma 4.1 and Lemma 4.2 imply that the relative error is always bounded as

$$|\hat{x} - x| \leq 2u|x|. \tag{4.8}$$

Furthermore, assuming rounding "to nearest even" and β even, we provide in Example 4.6 below an input (a, b, c, d) such that $|\hat{x} - x|/2u|x|$ has the form $1 - O(\beta^{-p})$, thus showing the asymptotic optimality of the relative error bound in (4.8). □

Example 4.6 (Example for which the relative error is asymptotically equivalent to the bound $2u$ in (4.8), assuming rounding "to nearest even" and β even)**.** Consider

$$a = b = \beta^{p-1} + 1, \qquad c = \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2}, \qquad d = 2\beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2}.$$

Since β is even, these four numbers are in $\mathbb{F}$. Furthermore, one can check that

$$bc = \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-3} + \beta^{p-1} + \tfrac{\beta}{2}\beta^{p-2},$$

from which it follows that the representation of bc in radix β is

$$\overbrace{1\tfrac{\beta}{2}000\ldots01}^{p \text{ digits}}\overbrace{\tfrac{\beta}{2}000\ldots00}^{p-1 \text{ digits}}.$$

Consequently, rounding bc "to nearest even" gives

$$\hat{w} = \beta^{2p-2} + \tfrac{\beta}{2}\beta^{2p-3} + 2\beta^{p-1},$$

and we obtain $e = \hat{w} - bc = \frac{\beta}{2}\beta^{p-2}$ and $f = ad - \hat{w} = \beta^{2p-2} + \frac{\beta}{2}\beta^{p-2}$. In particular, the representation of f in radix β is

$$\overbrace{10000\ldots00}^{p \text{ digits}}\overbrace{\tfrac{\beta}{2}000\ldots00}^{p-1 \text{ digits}},$$

so that rounding "to nearest even" simply truncates f into $\hat{f} = \beta^{2p-2}$. Thus, $\hat{f} + e = (\beta^p + \frac{\beta}{2})\beta^{p-2}$ and we deduce that

$$\hat{x} = \mathrm{RN}(\hat{f} + e) = \beta^{2p-2}.$$

On the other hand, since a and b are equal,

$$x = a(d - c) = (\beta^{p-1} + 1)\beta^{p-1}.$$

Therefore, the relative error is given by $|\hat{x} - x|/|x| = \frac{1}{\beta^{p-1}+1} = \frac{1}{1+\beta^{1-p}} \cdot 2u$, and we conclude, for β fixed and when $p \to \infty$, that the ratio $|\hat{x} - x|/2u|x|$ is in $1 - O(\beta^{-p})$.

5. Behavior of the worst case relative error with respect to σ

When ad and bc are of similar magnitude the relative error in the computed $ad - bc$ can be close to the bound $2u$ of Theorem 1.2, as already shown in Example 4.6. However, when one of the products $|ad|$ or $|bc|$ is sufficiently larger than the other, we can expect Algorithm 1 to be almost as accurate as if only one rounding were performed, that is, we can expect relative error bounds of the form $u + O(u^2)$.

In this section, we show that this is indeed the case: assuming that a, b, c, d, and x are nonzero, we investigate how the worst case relative error varies with the parameter $\sigma = e_a + e_d - e_b - e_c$ introduced in (2.8b). This integer is a convenient indicator of whether the ratio $|ad|/|bc|$ is huge or tiny, and it turns out that Kahan's algorithm behaves as predicted as soon as $|\sigma|$ is large enough. Before providing precise statements and proofs, let us first illustrate this behavior with a numerical experiment in radix 2.

For small values of the precision p and fixed values of the parameter σ, one can perform exhaustive tests and compute the worst case relative error generated by Kahan's algorithm. Figure 1 plots the worst case relative error $|\hat{x} - x|/|x|$ versus σ, for $\sigma \in \{-24, \ldots, 13\}$ and $(\beta, p) = (2, 11)$, which corresponds to the binary16 interchange format [8]. We used an exhaustive search program for maximizing $|\hat{X} - X|/|X|$ for each value of σ considered, distinguishing between two cases: either ad and bc have the same sign ($abcd > 0$), or ad and bc have opposite signs ($abcd < 0$). The corresponding worst cases are listed explicitly in Table 1.

This numerical experiment illustrates the typical behavior, with respect to σ, of the worst case relative error generated by Kahan's algorithm:

- When $\sigma \notin \{-p - 2, \ldots, 2\}$, the relative error bound $2u$ is not optimal. Sharper relative error bounds are derived when $\sigma \leq -p - 3$ in Subsection 5.1, and when $\sigma \geq 3$ in Subsection 5.2. These relative error bounds are plotted in Figure 1, and their ratio to the unit roundoff is reported in the last column of Table 1. In particular, they show that the relative error is bounded by $(1 + \epsilon)u$ for some positive ϵ such that $\epsilon \to 0$ as $|\sigma| \to \infty$.

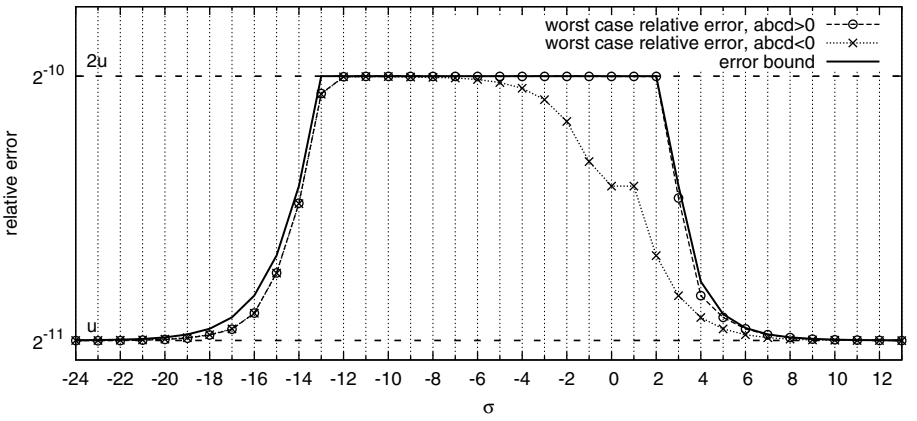

FIGURE 1. Worst case relative error for Algorithm 1, $\beta = 2$, $p = 11$.

- When $\sigma \in \{-p-2, \ldots, 2\}$, Figure 1 suggests that the relative error bound $2u$ is essentially the best possible. In fact, we show in Subsection 5.3 that at least in radix 2, there is a set of input values $\{A_\sigma, B_\sigma, C_\sigma, D_\sigma\}_{-p-2\le\sigma\le 2}$ whose associated relative errors tend to the plateau as $p \to \infty$.

The shape of the plots in Figure 1 reflects well the lack of symmetry between ad and bc in Kahan's algorithm. Thus, if σ is known to be negative then it may be useful to exchange the role of ad and bc, that is, to call Kahan's algorithm with (b, a, d, c) and return the opposite of $\hat{x}$.

5.1. Sharper relative error bound when $\sigma \le -p-3$.

Proposition 5.1. *If $\sigma \le -p-3$ then we have $|\hat{x} - x| \le (1+\epsilon)u|x|$, where $\epsilon := 2\beta^{p-1}/\left(\frac{\beta^{2p-2-\sigma}}{(\beta^p-1)^2} - 1\right)$ satisfies $0 < \epsilon < 2\beta^{-2}$. Furthermore, $\epsilon \to 0$ as $\sigma \to -\infty$.*

For $\beta, p \ge 2$, one can check that $1 + \epsilon \le 1/(1-u)$ iff $\sigma \le \zeta(\beta, p)$, where $\zeta(\beta, p) = -\left\lceil \log_\beta \left((4 - (2-\beta^{1-p})\beta^{1-p}) \cdot (\beta^p - 1)^2\right)\right\rceil$. In particular, $\zeta(\beta, p)$ equals $-2p-2$ if $\beta = 2 < p$, and $-2p-1$ if $\beta = 10$ or $\beta = p = 2$.

For the proof of Proposition 5.1, we first state the following two lemmas.

Lemma 5.2. *Let $y \in \mathbb{F}$ and $\delta, \epsilon \in \mathbb{R}$ be such that $\mathrm{RN}(y+\delta) = y + \delta + \epsilon$. If $\beta \ge 2$ and $|\delta| < \beta^{-1}\mathrm{ulp}(y)$ then $|\epsilon| \le |\delta|$.*

Proof. Note first that, by assumption, $y \ne 0$ and $|\delta| < \frac{1}{2}\mathrm{ulp}(y)$. Therefore, if $|y| \ne \beta^{p-1}\mathrm{ulp}(y)$ or if $y\delta \ge 0$ we have $\mathrm{RN}(y+\delta) = y$ and thus $\epsilon = -\delta$. The rest of the proof deals with the case where $|y| = \beta^{p-1}\mathrm{ulp}(y)$ and $y\delta < 0$. Let $\gamma = \beta^{-1}\mathrm{ulp}(y)$. When $|\delta| \le \frac{\gamma}{2}$ we have, as before, $\mathrm{RN}(y+\delta) = y$. When $\frac{\gamma}{2} < |\delta| < \gamma$, we can check that $\mathrm{RN}(y+\delta) = y + \mathrm{sign}(\delta)\gamma$, from which it follows that $\mathrm{sign}(\delta)\epsilon = \gamma - |\delta| \in (0, \frac{\gamma}{2}]$ and thus $|\epsilon| \le \frac{\gamma}{2} \le |\delta|$. □

Lemma 5.3. *For $\sigma \le -p-3$ and assuming (C), we have $|\epsilon_1| \le |ad| < \frac{1}{2}\mathrm{ulp}(x)$.*

Proof. We have $ad/bc = AD\beta^\sigma/BC$. Using $\beta^{p-1} \le |A|, |B|, |C|, |D| < \beta^p$ leads to $|ad| < \beta^{\sigma+2}|bc|$. Since $\hat{w} = \mathrm{RN}(bc)$ implies $\mathrm{ulp}(bc) \le \mathrm{ulp}(\hat{w})$, we arrive at $|ad| < \beta^{p+\sigma+2}\mathrm{ulp}(\hat{w})$. For $\beta \ge 2$ and $\sigma \le -p-3$, this gives $|ad| < \beta^{-1}\mathrm{ulp}(\hat{w})$ and

TABLE 1. Worst cases for Algorithm 1, $\beta = 2$, $p = 11$. The ratios $\frac{|\hat{x}-x|}{u|x|}$ and $\frac{\text{bound}}{u}$ have been rounded upward to four decimal places.

σ	A	B	C	D	$\frac{\lvert\hat{x}-x\rvert}{u\lvert x\rvert}$	A	B	C	D	$\frac{\lvert\hat{x}-x\rvert}{u\lvert x\rvert}$	$\frac{\text{bound}}{u}$
-24	1024	1536	1366	1024	0.9995	2047	−1536	1366	2047	0.9998	1.0005
-23	1024	1536	1366	1024	0.9994	2047	−1536	1366	2047	1.0000	1.0010
-22	2047	1792	1172	2047	0.9995	2047	−1536	1366	2047	1.0005	1.0020
-21	2047	1792	1172	2047	1.0005	2047	−1536	1366	2047	1.0015	1.0040
-20	2047	1792	1172	2047	1.0025	2047	−1536	1366	2047	1.0035	1.0079
-19	2047	1792	1172	2047	1.0064	2047	−1536	1366	2047	1.0074	1.0157
-18	2047	1792	1172	2047	1.0142	2047	−1536	1366	2047	1.0152	1.0313
-17	2047	1792	1172	2047	1.0298	2047	−1536	1366	2047	1.0308	1.0625
-16	2047	1051	1043	2047	1.0743	2047	−1056	1040	2047	1.0740	1.1249
-15	2047	1051	1043	2047	1.1938	2047	−1056	1040	2047	1.1932	1.2498
-14	2047	1051	1043	2047	1.4329	2047	−1056	1040	2047	1.4314	1.4997
-13	2047	1051	1043	2047	1.9113	2047	−1056	1040	2047	1.9078	2.0000
-12	2047	1792	1172	2047	1.9971	2047	−1536	1366	2047	1.9971	2.0000
-11	1550	1792	1172	1353	1.9981	1550	−1536	1366	1353	1.9981	2.0000
-10	1024	1792	1172	1024	1.9981	1024	−1536	1366	1024	1.9981	2.0000
-9	2031	1496	1408	1807	1.9981	1314	−2047	1025	1197	1.9952	2.0000
-8	1792	1496	1408	1024	1.9981	1280	−1536	1366	1024	1.9942	2.0000
-7	1915	1827	1163	1848	1.9989	1152	−1536	1366	1024	1.9903	2.0000
-6	1897	1831	1175	1831	1.9991	1088	−1536	1366	1024	1.9826	2.0000
-5	1551	1815	1191	1331	1.9990	1056	−1536	1366	1024	1.9674	2.0000
-4	1963	1575	1431	1277	1.9990	1040	−1536	1366	1024	1.9376	2.0000
-3	1405	1581	1445	1067	1.9990	1032	−1536	1366	1024	1.8807	2.0000
-2	1777	1681	1649	1519	1.9988	1028	−1536	1366	1024	1.7763	2.0000
-1	1113	1969	1361	1047	1.9986	1026	−1536	1366	1024	1.5988	2.0000
0	1027	1025	1025	1025	1.9981	1605	−1536	1514	1165	1.4992	2.0000
1	1088	1152	1076	1052	1.9981	1025	−1536	1366	1024	1.4990	2.0000
2	1040	1536	1450	1040	1.9981	1526	−1472	1456	1024	1.2493	2.0000
3	1024	2023	2007	1024	1.4531	1408	−1472	1456	1300	1.1244	1.5000
4	1153	1536	1370	1024	1.1244	1472	−1496	1408	1336	1.0620	1.1667
5	1072	1536	1370	1040	1.0620	1868	−1496	1408	1088	1.0308	1.0715
6	1057	1512	1408	1024	1.0308	1844	−1496	1408	1120	1.0152	1.0334
7	1043	1578	1536	1024	1.0152	1512	−1504	1504	1376	1.0074	1.0162
8	1033	1536	1450	1024	1.0074	1588	−1472	1456	1316	1.0035	1.0080
9	1029	1568	1504	1024	1.0035	1513	−1472	1456	1384	1.0015	1.0040
10	1027	1706	1536	1024	1.0015	1508	−1496	1408	1390	1.0005	1.0020
11	1026	2003	1570	1024	1.0005	1940	−1536	1382	1081	1.0000	1.0010
12	1025	1456	1440	1024	0.9998	1024	−1536	1366	1024	1.0000	1.0005
13	1483	1536	1434	1415	0.9996	1554	−1504	1504	1350	0.9996	1.0003

the inequality $|\epsilon_1| \le |ad|$ follows from Lemma 5.2. Moreover, $\sigma \le -p-3$ implies $|ad|/|x| \le 1/(\frac{\beta^{3p+1}}{(\beta^p-1)^2} - 1)$ and thus, as $\beta \ge 2$, $|ad| \le \frac{1}{2}\beta^{-p}|x| < \frac{1}{2}\text{ulp}(x)$. □

Proof of Proposition 5.1. Assume (C), for otherwise the result holds trivially. Using Lemma 5.3 and (4.6) gives $|\epsilon_1| \le |ad|$ and $|\epsilon_2| \le \frac{1}{2}\text{ulp}(x)$, so that $|\hat{x} - x| \le (u + \psi)|x|$ with $\psi := |ad|/|x|$. Now, $\psi^{-1} = |1 - bc/ad| \ge |bc/ad| - 1 = |BC|/|AD| \cdot \beta^{-\sigma} - 1 \ge \beta^{2p-2-\sigma}/(\beta^p - 1)^2 - 1$. Since $\sigma \le -2$, this lower bound is positive. Hence $\psi \le \epsilon u$ and thus $|\hat{x} - x| \le (1 + \epsilon)u|x|$. For fixed $\beta, p \ge 2$, we have $\epsilon = O(\beta^\sigma)$

as $\sigma \to -\infty$ and the limit follows; also, ϵ is a nondecreasing function of σ and one may check that $\epsilon < 2\beta^{-2}$ for $\sigma = -p-3$. □

5.2. **Sharper relative error bound when $\sigma \geq 3$.**

Proposition 5.4. *If $\sigma \geq 3$ then $|\hat{x} - x| < (1+\epsilon)u|x|$, where $\epsilon := \frac{\beta^{-1}}{\beta^{-2+\sigma}-1}$ satisfies $0 < \epsilon \leq \frac{1}{(\beta-1)\beta}$. Furthermore, $\epsilon \to 0$ as $\sigma \to +\infty$.*

For $\beta, p \geq 2$, $1+\epsilon \leq 1/(1-u)$ iff $\sigma \leq \zeta(\beta,p) := \lceil \log_\beta(2\beta^p + (\beta-1)\beta) \rceil$; In particular, we have $\zeta(2,p) = p+2$ and $\zeta(10,p) = p+1$.

For the proof of Proposition 5.4 we use the following two lemmas.

Lemma 5.5. *If $|e| \leq \frac{1}{2}\mathrm{ulp}(x)$ then $|\epsilon_1| \leq \frac{1}{2}\mathrm{ulp}(x)$.*

Proof. If $x = 0$ then by assumption $e = 0$, so that f, $\hat{f}$, ϵ_1 are zero and the result is true. Now assume $x \neq 0$. By (2.4) the only nontrivial case is when $\mathrm{ulp}(x) < \mathrm{ulp}(f)$. In this case, $x \neq 0$ leads to $|x| < \beta^p \mathrm{ulp}(x) \leq \beta^{p-1}\mathrm{ulp}(f) \leq |f|$, so that $|x| < y \leq |f|$ for some y in $\mathbb{F}$. Since $\hat{f} = \mathrm{RN}(f)$, we have $\mathrm{sign}(\hat{f}) = \mathrm{sign}(f)$ and $|\hat{f}| = \mathrm{RN}(|f|)$. Hence $|\epsilon_1| = \big|\mathrm{RN}(|f|) - |f|\big| \leq |f| - y < |f| - |x| \leq |e| \leq \frac{1}{2}\mathrm{ulp}(x)$. □

Lemma 5.6. *For $\sigma \geq 3$ and assuming (C), we have $|\epsilon_2| \leq |e| < \frac{1}{\beta}\mathrm{ulp}(x)$.*

Proof. Since $\sigma \geq 0$, we deduce from (2.9) and (2.10) that $|E| \leq \frac{1}{2}\mathrm{ulp}(BC) \leq \frac{1}{2}\beta^p$ and that both $|F|$ and $|X|$ are lower bounded by $\beta^{2p-2+\sigma} - \beta^{2p}$. Hence, for $\beta \geq 2$ and $\sigma \geq 3$, the ratios $|e|/|f| = |E|/|F|$ and $|e|/|x| = |E|/|X|$ are upper bounded by $\frac{\beta^{-p}}{2(\beta-1)} \leq \beta^{-p-1}$. This implies first that $|e| \leq \beta^{-p-1}|f| < \beta^{-1}\mathrm{ulp}(f) \leq \beta^{-1}\mathrm{ulp}(\hat{f})$ and, since $\mathrm{RN}(\hat{f}+e) = \hat{f} + e + \epsilon_2$, Lemma 5.2 gives $|\epsilon_2| \leq |e|$. On the other hand, $|e| \leq \beta^{-p-1}|x| < \beta^{-1}\mathrm{ulp}(x)$ and the conclusion follows. □

Proof of Proposition 5.4. Assume (C), for otherwise the result holds trivially. Since $\beta \geq 2$ and $\sigma \geq 3$, we deduce from Lemma 5.5 and Lemma 5.6 that $|\epsilon_1| \leq \frac{1}{2}\mathrm{ulp}(x)$ and $|\epsilon_2| \leq |e|$. Hence, recalling that (C) implies x is nonzero, $|\hat{x} - x| \leq (u+\psi)|x|$ with $\psi = |e|/|x| = |E|/|X|$. Now, since $\sigma \geq 0$ we have $|E| \leq \frac{1}{2}\beta^p$ and $|X| \geq |AD|\beta^\sigma - |BC| > \beta^{2p}(\beta^{-2+\sigma} - 1)$, so that $\psi < \epsilon u$ with ϵ as above. For any given $\beta \geq 2$, we have $\epsilon \leq \frac{1}{\beta^2-\beta}$ when $\sigma \geq 3$, and $\epsilon = O(\beta^{-\sigma})$ as $\sigma \to +\infty$, which concludes the proof. □

5.3. **Sharpness of the bound $2u$ for $-p-2 \leq \sigma \leq 2$ in radix 2.** To prove that in radix 2 the relative error bound $2u$ is essentially the best possible over the range $-p-2 \leq \sigma \leq 2$, Table 2 gives parametrized bad cases for which $|\hat{x} - x|/|x| \sim 2^{-p+1}$ as $p \to +\infty$. These examples have been first determined experimentally (by only focusing on input numbers of the form 2^{p-1}, or $2^{p-1} + 2^i$ with $p-2 \leq i \leq 0$, or $2^p - 1$, or $2^p - 1 - 2^i$ with $p-2 \leq i \leq 0$), to give a hint about the "bit patterns" that would lead to such cases. These guessed bit patterns have then been proven to actually correspond to cases for which the relative error tends to $2u$ as $p \to \infty$.

6. Concluding remarks: Special cases

Let us conclude with some remarks about the behavior of Kahan's algorithm in two special cases. We first consider the case where ad and bc have opposite signs, which covers in particular sums of squares $a^2 + b^2$. We then consider the evaluation of $y^2 - zt$, an expression that occurs for instance when computing the discriminant of a quadratic equation.

TABLE 2. Parametrized examples for $-p-2 \le \sigma \le 2$ for which the relative error produced by Algorithm 1 is asymptotically equivalent to the bound $2^{1-p} = 2u$ as p tends to infinity.

σ	A	B	C	D
$-p-2$	$2^p - 1$	$-\left(2^{p-1} + 2^{\lfloor \frac{p-1}{2} \rfloor}\right)$	$2^{p-1} + 2^{\lfloor \frac{p-2}{2} \rfloor}$	$2^p - 1$
$-p-1$	$2^p - 1$	$-(2^p - 1)$	$2^{p-1} + 1$	$2^p - 1$
$-p$	$2^p - 2$	$-(2^p - 1)$	$2^{p-1} + 1$	$2^{p-1} + 1$
$-p+1$	2^{p-1}	$-(2^p - 1)$	$2^{p-1} + 1$	2^{p-1}
$-p+2$	$2^p - 2^{p-2} - 1$	$-(2^p - 1)$	$2^{p-1} + 1$	2^{p-1}
$(-p+2, -p/2]$	$2^{p-1} + 2^{-\sigma}$	$-(2^p - 1)$	$2^{p-1} + 1$	2^{p-1}
$(-p/2, -2]$	$2^{p-1} + 2^{-\sigma}$	$2^{p-1} + 2^{p+\sigma-1}$	$2^{p-1} + 2^{-\sigma-1}$	$2^{p-1} + 2^{-\sigma-1}$
-1	$2^p - 2$	$2^{p-1} + 1$	$2^{p-1} + 1$	$2^{p-1} + 1$
0	$2^p - 5$	$2^p - 3$	$2^{p-1} + 1$	$2^{p-1} + 1$
1	$2^{p-1} + 2^{p-3}$	$2^{p-1} + 2^{p-2}$	$2^{p-1} + 1$	$2^{p-1} + 1$
2	$2^{p-1} + 2^{p-3}$	$2^p - 2$	$2^{p-1} + 2^{p-2} - 1$	$2^{p-1} + 1$

6.1. **Case where ad and bc have opposite signs.** In this special case a significantly smaller error bound in ulps can be derived, whereas, at least in radix 2, the error bound given by Theorem 1.2 remains asymptotically optimal.

Proposition 6.1. *If ad and bc have opposite signs then $|\hat{x} - x| \le \mathrm{ulp}(x)$.*

Proof. From Lemmas 4.1 and 4.2, $|\hat{x} - x| \le |\epsilon_1| + \frac{1}{2}\mathrm{ulp}(x)$ and it suffices to check that $|\epsilon_1| \le \frac{1}{2}\mathrm{ulp}(x)$. If ad and bc are of opposite signs then $|bc| \le |ad - bc| = |x|$ and thus $\mathrm{ulp}(bc) \le \mathrm{ulp}(x)$. Hence $|e| \le \frac{1}{2}\mathrm{ulp}(bc) \le \frac{1}{2}\mathrm{ulp}(x)$, which by Lemma 5.5 implies $|\epsilon_1| \le \frac{1}{2}\mathrm{ulp}(x)$, as wanted. □

The example below shows that in radix 2 and under the constraint $\mathrm{sign}(ad) \ne \mathrm{sign}(bc)$, both the improved absolute error bound of Proposition 6.1 and the relative error bound of Theorem 1.2 are asymptotically optimal.

Example 6.2. Consider

$$a = 2^p - 2, \quad b = -(2^p - 1)2^p, \quad c = d = 2^{p-1} + 1.$$

Then $bc < 0 < ad$ and one may check that $|\hat{x} - x| = (1 - 2^{-p} - 2^{1-2p})\mathrm{ulp}(x)$. Moreover, the relative error satisfies

$$\frac{|\hat{x} - x|}{|x|} = \frac{2^{3p-1} - 2^{2p-1} - 2^p}{2^{3p-1} + 2^{2p} - 2^p - 2} \cdot 2u.$$

Hence, both $|\hat{x} - x|/\mathrm{ulp}(x)$ and $|\hat{x} - x|/(2u|x|)$ are in $1 - O(2^{-p})$ as $p \to \infty$.

A particularly important occurrence of the case "$\mathrm{sign}(ad) \ne \mathrm{sign}(bc)$" is the evaluation of sums of squares, that is, expressions of the form $x = a^2 + b^2$ obtained by setting d to a and c to $-b$. For radix 2, the next three examples illustrate the sharpness of the bounds in Proposition 6.1 and Theorem 1.2 when evaluating sums of squares with Kahan's algorithm.

Example 6.3 shows that the absolute error bound in Proposition 6.1 is asymptotically optimal if p is even, and optimal if p is odd. Example 6.4 shows that, at least when p is even, the relative error bound $2u$ of Theorem 1.2 remains essentially the best possible. When p is odd, we did not manage to build such a generic example. However, Example 6.5 shows that a relative error close to $2u$ is attainable for the binary64 ($p = 53$) and binary128 ($p = 113$) arithmetics [8].

Example 6.3. If $p \geq 6$ is even, consider

$$a = d = 5 \cdot 2^{p-3} + 2 \quad \text{and} \quad b = -c = 3 \cdot 2^{p-2} + 1,$$

and, if $p \geq 7$ is odd, let

$$a = d = 2^{p-1} + 2^{\frac{p-1}{2}} \quad \text{and} \quad b = -c = 3 \cdot 2^{p-2} + 2^{\frac{p-1}{2}}.$$

Then the associated absolute errors are, respectively, $(1 - 5 \cdot 2^{-p})\,\mathrm{ulp}(x)$ and $\mathrm{ulp}(x)$.

Example 6.4. If p is even, let

$$a = d = 2^{p-1} \quad \text{and} \quad b = -c = (2^{p-1} + 2^{\frac{p}{2}-1} + 1)\, 2^{p/2}.$$

Then

$$\frac{|\hat{x} - x|}{|x|} = \frac{2^{p-1} - 2^{\frac{p}{2}} - 1}{2^{p-1} + 2^{\frac{p}{2}} + 3 + 2^{1-\frac{p}{2}} + 2^{1-p}} \cdot 2u$$

and we deduce that $|\hat{x} - x|/(2u|x|)$ is in $1 - O(2^{-p/2})$ as p tends to infinity.

Example 6.5. When $p = 53$, which corresponds to the binary64 format, taking

$$a = d = 8426657115275263 \quad \text{and} \quad b = -c = 302232031373205690122240,$$

gives $|\hat{x} - x|/(2u|x|) = 0.999000553067209\ldots$.

In the binary128 floating-point format, $p = 113$ and taking

$$\begin{array}{rcrcl} a & = & d & = & 97152742001491501330707333660016 63, \\ b & = & -c & = & 37414441915739171179399509762260948528898146041856 0, \end{array}$$

gives $|\hat{x} - x|/(2u|x|) = 0.999008178703665\ldots$.

6.2. **Computation of discriminants.** Now assume we want to evaluate $y^2 - zt$ from $y, z, t \in \mathbb{F}$, for instance, for getting the discriminant of a quadratic equation. Denoting the output of Algorithm 1 by $\mathbf{Kdet}(a, b, c, d)$, the value $y^2 - zt$ can clearly be approximated in two different ways, as $\mathbf{Kdet}(y, z, t, y)$ or $-\mathbf{Kdet}(z, y, y, t)$. However, unlike arbitrary determinants and sums of squares, the expression $y^2 - zt$ lacks symmetry (one square and one product instead of two products or two squares) and it is natural to ask whether the bounds in Theorems 1.1 and 1.2 can be improved if we restrict to one of those two evaluation choices.

The answer is "no" at least in radix 2: as we shall see in the two examples below, for both ways of computing the discriminant the relative error bound $2u$ remains asymptotically optimal and the absolute error bound $\frac{3}{2}\mathrm{ulp}(x)$ remains optimal.

Example 6.6. Let $\hat{x} = \mathbf{Kdet}(y, z, t, y)$ with

$$y = z = 2^{p-1} + 3 \quad \text{and} \quad t = 2^{p-1} + 1.$$

One can check that $|\hat{x} - x|/|x|$ is equal to $2u/(1 + 6 \cdot 2^{-p})$. Now, if we use instead, $\hat{x} = -\mathbf{Kdet}(z, y, y, t)$, then it can be checked that the inputs

$$y = z = 2^{p-1} + 1 \quad \text{and} \quad t = 2^{p-1} + 3$$

lead to the relative error $|\hat{x}-x|/|x| = 2u/(1+2^{1-p})$. Thus, for both ways of getting $y^2 - zt$ the relative error can have the form $1 - O(2^{-p})$ as p tends to infinity.

Example 6.7. Assume $p \geq 7$ and consider $\hat{x} = \mathbf{Kdet}(y, z, t, y)$ for y, z, t such that

$$y = 3 \cdot 2^{p-2} - 2, \qquad z = 2^p - 1, \qquad t = 9 \cdot 2^{p-4} - 6.$$

It can be checked that $x = y^2 - zt = 57 \cdot 2^{p-4} - 2$, $\mathrm{ulp}(x) = 4$, and $\hat{x} = 57 \cdot 2^{p-4} - 8$, from which we deduce $|\hat{x} - x| = \frac{3}{2}\mathrm{ulp}(x)$. When $\hat{x} = -\mathbf{Kdet}(z, y, y, t)$, considering

$$y = 5 \cdot 2^{p-3} - 1, \qquad z = 5 \cdot 2^{p-3} - 2, \qquad t = 5 \cdot 2^{p-3} - 3$$

with $p \geq 6$ leads to $x = y^2 - zt = 15 \cdot 2^{p-3} - 5$, $\mathrm{ulp}(x) = 2$, and $\hat{x} = 15 \cdot 2^{p-3} - 8$, which also gives an absolute error equal to $\frac{3}{2}\mathrm{ulp}(x)$.

References

1. S. Boldo, *Kahan's algorithm for a correct discriminant computation at last formally proven*, IEEE Transactions on Computers **58** (2009), no. 2, 220–225. MR2655570
2. R. P. Brent, C. Percival, and P. Zimmermann, *Error bounds on complex floating-point multiplication*, Mathematics of Computation **76** (2007), 1469–1481. MR2299783 (2008b:65062)
3. M. Cornea, J. Harrison, and P. T. P. Tang, *Scientific computing on Itanium®-based systems*, Intel Press, Hillsboro, OR, 2002.
4. Y. Hida, X. S. Li, and D. H. Bailey, *Algorithms for quad-double precision floating-point arithmetic*, Proceedings of the 15th IEEE Symposium on Computer Arithmetic (ARITH-16) (Vail, CO) (N. Burgess and L. Ciminiera, eds.), June 2001, pp. 155–162.
5. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, first ed., SIAM, Philadelphia, PA, USA, 1996. MR1368629 (97a:65047)
6. ———, *Accuracy and Stability of Numerical Algorithms*, second ed., SIAM, Philadelphia, PA, USA, 2002. MR1927606 (2003g:65064)
7. E. Hokenek, R. K. Montoye, and P. W. Cook, *Second-generation RISC floating point with multiply-add fused*, IEEE Journal of Solid-State Circuits **25** (1990), no. 5, 1207–1213.
8. 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`.
9. R. M. Jessani and C. H. Olson, *The floating-point unit of the PowerPC 603e microprocessor*, IBM Journal of Research and Development **40** (1996), no. 5, 559–566.
10. W. Kahan, *Lecture notes on the status of IEEE-754*, PDF file accessible at `http://www.cs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF`, 1996.
11. ———, *Matlab's loss is nobody's gain*, Available at `http://www.cs.berkeley.edu/~wkahan/MxMulEps.pdf`, 1998.
12. ———, *On the cost of floating-point computation without extra-precise arithmetic*, Available at `http://http.cs.berkeley.edu/~wkahan/Qdrtcs.pdf`, 2004.
13. A. H. Karp and P. Markstein, *High-precision division and square root*, ACM Transactions on Mathematical Software **23** (1997), no. 4, 561–589. MR1671702
14. G. Melquiond and S. Pion, *Formally certified floating-point filters for homogeneous geometric predicates*, Theoretical Informatics and Applications **41** (2007), no. 1, 57–69. MR2330043 (2008e:68147)
15. R. K. Montoye, E. Hokonek, and S. L. Runyan, *Design of the IBM RISC System/6000 floating-point execution unit*, IBM Journal of Research and Development **34** (1990), no. 1, 59–70.
16. Y. Nievergelt, *Scalar fused multiply-add instructions produce floating-point matrix arithmetic provably accurate to the penultimate digit*, ACM Transactions on Mathematical Software **29** (2003), no. 1, 27–48. MR2001452
17. S. M. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation part I: Faithful rounding*, SIAM Journal on Scientific Computing **31** (2008), no. 1, 189–224. MR2460776 (2009k:65081)

INRIA, Laboratoire LIP (CNRS, ENS de Lyon, INRIA, UCBL), Université de Lyon — 46, allée d'Italie, 69364 Lyon cedex 07, France
E-mail address: claude-pierre.jeannerod@ens-lyon.fr

UCBL, Laboratoire LIP (CNRS, ENS de Lyon, INRIA, UCBL), Université de Lyon — 46, allée d'Italie, 69364 Lyon cedex 07, France
E-mail address: nicolas.louvet@ens-lyon.fr

CNRS, Laboratoire LIP (CNRS, ENS de Lyon, INRIA, UCBL), Université de Lyon — 46, allée d'Italie, 69364 Lyon cedex 07, France
E-mail address: jean-michel.muller@ens-lyon.fr