Numer Algor (2016) 73:735–760
DOI 10.1007/s11075-016-0115-x

ORIGINAL PAPER

Sharp error bounds for complex floating-point inversion

Claude-Pierre Jeannerod[1] · Nicolas Louvet[2] · Jean-Michel Muller[3] · Antoine Plet[4]

Received: 26 August 2015 / Accepted: 18 February 2016 / Published online: 11 March 2016

Abstract We study the accuracy of the classic algorithm for inverting a complex number given by its real and imaginary parts as floating-point numbers. Our analyses are done in binary floating-point arithmetic, with an unbounded exponent range and in precision p; we also assume that the basic arithmetic operations $(+, -, \times, /)$ are rounded to nearest, so that the roundoff unit is $u = 2^{-p}$. We bound the largest relative error in the computed inverse either in the componentwise or in the normwise sense. We prove the componentwise relative error bound $3u$ for the complex inversion algorithm (assuming $p \geqslant 4$), and we show that this bound is asymptotically optimal (as $p \to \infty$) when p is even, and sharp when using one of the basic IEEE 754 binary formats with an odd precision ($p = 53, 113$). This componentwise bound obviously

✉ Jean-Michel Muller
jean-michel.muller@ens-lyon.fr

Claude-Pierre Jeannerod
claude-pierre.jeannerod@inria.fr

Nicolas Louvet
nicolas.louvet@ens-lyon.fr

Antoine Plet
antoine.plet@ens-lyon.fr

1 Inria, Laboratoire LIP (CNRS, ENSL, Inria, UCBL), Université de Lyon, Lyon, France

2 UCBL, Laboratoire LIP (CNRS, ENSL, Inria, UCBL), Université de Lyon, Lyon, France

3 CNRS, Laboratoire LIP (CNRS, ENSL, Inria, UCBL), Université de Lyon, Lyon, France

4 ENS de Lyon, Laboratoire LIP (CNRS, ENSL, Inria, UCBL), Université de Lyon, Lyon, France

leads to the same bound $3u$ for the normwise relative error. However, we prove that the smaller bound $2.707131u$ holds (assuming $p \geqslant 24$) for the normwise relative error, and we illustrate the sharpness of this bound for the basic IEEE 754 binary formats ($p = 24, 53, 113$) using numerical examples.

Keywords Floating-point arithmetic · Rounding error analysis · Complex inversion

1 Introduction

This paper deals with the accuracy of the inversion of a nonzero complex number given by its real and imaginary parts as floating-point numbers. We assume that the underlying floating-point arithmetic has radix 2 and precision $p \geqslant 2$, and we also assume an unbounded exponent range, which means that our results apply to practical floating-point calculations according to the IEEE 754 standard [6] as long as underflow and overflow do not occur.

Given a nonzero complex number $a + ib$, its inverse satisfies

$$z = R + i\,I, \qquad R = \frac{a}{a^2 + b^2}, \qquad I = -\frac{b}{a^2 + b^2}. \tag{1}$$

Assuming a and b are floating-point numbers and denoting by RN a round-to-nearest function, we focus in this paper on the approximation $\widehat{z} = \widehat{R} + i\,\widehat{I}$ that can be computed classically in floating-point arithmetic according to

$$\widehat{R} = \mathrm{RN}\left(\frac{a}{\mathrm{RN}\big(\mathrm{RN}(a^2) + \mathrm{RN}(b^2)\big)}\right) \tag{2}$$

for the real part, and with a similar expression for the imaginary part $\widehat{I}$. This scheme corresponds to Algorithm 1 below.

Algorithm 1 Inversion of a nonzero complex floating-point number $a + ib$.

$s_a \leftarrow \mathrm{RN}(a^2)$
$s_b \leftarrow \mathrm{RN}(b^2)$
$s \leftarrow \mathrm{RN}(s_a + s_b)$
$\widehat{R} \leftarrow \mathrm{RN}(a/s)$
$\widehat{I} \leftarrow \mathrm{RN}(-b/s)$
return $\widehat{R} + i\,\widehat{I}$

We provide an accuracy analysis of Algorithm 1, for both the componentwise relative error $\mathrm{E_C} = \max\big(|R - \widehat{R}|/|R|, |I - \widehat{I}|/|I|\big)$ and the normwise relative error $\mathrm{E_N} = |z - \widehat{z}|/|z|$. In each case, we bound the *largest* error value by a function $B(p)$

depending only on the precision p, and study the tightness of that bound. In this context, we typically distinguish between three levels of quality:

- If we can show that there exist some inputs $a + ib$ parametrized by p and for which the bound is attained for every $p \geqslant p_0$ (for a given $p_0 \geqslant 2$), then we say that the bound is *optimal*.
- If we can show that there exist some inputs parametrized by p and for which the relative error $E(p)$ satisfies $E(p)/B(p) \to 1$ as $p \to \infty$, then we say that the bound is *asymptotically optimal*.
- In some cases, we did not manage to establish (asymptotic) optimality, but have found input examples—either parametrized by p, or just for some values of p of practical interest (like those corresponding to the basic IEEE 754 formats)—for which $E(p)$ is very close to $B(p)$. In this case, we say that the bound is *sharp*. (See [13] for a similar use of the word "sharp".)

The componentwise relative error generated by Algorithm 1 can easily be bounded as $\mathrm{E_C} \leqslant 3u + O(u^2)$, where $u = 2^{-p}$ is the unit roundoff. Our first contribution is to show that the term $\mathcal{O}(u^2)$ can in fact be removed, which leads to the simpler bound $\mathrm{E_C} \leqslant 3u$ (assuming $p \geqslant 4$). Furthermore, when p is even, we show that this bound is asymptotically optimal by providing floating-point numbers a and b parametrized by p and for which $\mathrm{E_C} \geqslant 3u - \frac{31}{2}u^{\frac{3}{2}} + \mathcal{O}(u^2)$; when p is odd, we show that the bound $3u$ is sharp, especially for the corresponding basic IEEE 754 binary formats ($p = 53, 113$).

The normwise relative error bound $\mathrm{E_N} \leqslant 3u + \mathcal{O}(u^2)$ can be found in [4, p. 30], and a direct application of our componentwise error analysis leads further to $\mathrm{E_N} \leqslant 3u$. The second main contribution of our paper is to show that if $p \geqslant 10$ then the following smaller bound holds: $\mathrm{E_N} < \gamma u + 9u^2$, where γ is an explicit constant in $(2.70712, 2.70713)$. When using for example the IEEE 754 binary32 format ($p = 24$), this implies $\mathrm{E_N} < 2.707131u$. The techniques and the case distinction we use to prove this bound are inspired from [13], but we also extensively use real analysis and differentiation for the treatment of each case. We provide numerical examples to show that the bound we obtain is sharp for the basic IEEE 754 formats ($p = 24, 53, 113$).

Several authors [2, 8, 10, 11] have suggested ways of avoiding spurious overflows and underflows in complex division, and some of them may be used in Algorithm 1. Of course, if the computation introduces further rounding errors, which is the case for example in Smith's method [10], then our error bounds may not hold anymore. However, following the technique developed by Priest in [8], it is possible to scale a and b by a power of two in order to avoid spurious overflows and underflows without introducing new rounding errors: in that case, our analyses are valid if neither overflow nor underflow occurs during the computation. Nonetheless, we do not deal with scaling techniques here, and focus only on the largest error assuming an unbounded exponent range.

Outline Section 2 is devoted to the componentwise relative error analysis of Algorithm 1, and Section 3 to its normwise relative error analysis. We conclude in Section 4 with some remarks on the implications of these error analyses for complex floating-point division. The technical parts of the proofs that can be skipped at first reading are gathered in Appendix A.

Assumptions and notation For any real number t, we denote by $\mathrm{RN}(t)$ the binary floating-point number that is nearest to t, with a tie-breaking strategy preserving the following properties:

- $\mathrm{RN}(2^k t) = 2^k \mathrm{RN}(t)$ for any integer k;
- $\mathrm{RN}(-t) = -\mathrm{RN}(t)$.

In particular, either the *roundTiesToEven* or the *roundTiesToAway* rounding direction attributes defined in the IEEE 754 standard [6] can be used.

Throughout the paper, we also rely on the following relative error bound [7, p. 232]: for any real number t,

$$\mathrm{RN}(t) = t(1+\epsilon) \quad \text{with} \quad |\epsilon| \leqslant \frac{u}{1+u}. \tag{3}$$

Note that (3) implies the well-known inequality $|\mathrm{RN}(t) - t| \leqslant u|t|$; see [5, p. 38].

Finally, we use the notation $\mathrm{ufp}(t)$ (*unit in the first place*, introduced in [9]) to denote the weight of the most significant bit of t: if $t \neq 0$ then $\mathrm{ufp}(t)$ is the unique integer power of two such that $1 \leqslant |t|/\mathrm{ufp}(t) < 2$, and $\mathrm{ufp}(0) = 0$. The usual ulp function (*unit in the last place*) is related to the ufp function through the relation $\mathrm{ulp}(t) = 2u \cdot \mathrm{ufp}(t)$, so that

$$|t - \mathrm{RN}(t)| \leqslant \frac{1}{2}\mathrm{ulp}(t) = \mathrm{ufp}(t)u. \tag{4}$$

2 Componentwise error bound

In this section, we focus on the componentwise relative error of Algorithm 1. We note first that since $a+ib$ is nonzero, $R = a/(a^2+b^2)$ and $I = -b/(a^2+b^2)$ cannot both be zero, and that if one of them is zero then the returned result is very accurate. Assume for example that $R = 0$ (the case $I = 0$ is similar). In that case, $a = 0$ and $I = -1/b$. Using the bound in (3), it is then easily checked that the values $\widehat{R}$ and $\widehat{I}$ returned by the algorithm are as follows:

- $\widehat{R} = 0$, which means that the real part is computed exactly;
- $\widehat{I} = -\mathrm{RN}(b/\mathrm{RN}(b^2))$ and the relative error on the imaginary part is bounded by $2u$ (and thus smaller than the bound we are going to give in the general case).

Therefore, the rest of this section is devoted to analyzing $\mathrm{E_C} = \max\big(|R - \widehat{R}|/|R|, |I - \widehat{I}|/|I|\big)$ for R and I nonzero. Repeated applications of the bound in (3) give immediately $\mathrm{E_C} \leqslant 3u + \mathcal{O}\big(u^2\big)$. We show below that if $p \geqslant 4$ then the $\mathcal{O}\big(u^2\big)$ term can in fact be removed, leading to the simpler bound $3u$.

To do this, we prove that if $p \neq 3$ then the relative error bound $u/(1+u)$ in (3) can be replaced by $u/(1+3u)$ when evaluating a square $\mathrm{RN}\big(a^2\big)$ instead of a general product. (When $p = 3$, it is easily checked that the bound $u/(1+u)$ is attained when squaring the floating-point numbers $3/2 \cdot 2^e$, $e \in \mathbb{Z}$.) This slight refinement will turn out to be enough to show that Algorithm 1 satisfies $\mathrm{E_C} \leqslant 3u$.

Lemma 1 *Let a be a floating-point number. If* $p \neq 3$ *then* $|a^2 - (2+2u)| \geqslant 4u^2$.

Proof If $|a| < 1$ then $|a^2 - (2+2u)| > 1+2u$, and the result follows from the fact that $1 + 2u > 4u^2$ when $p > 0$. Assume now that $|a| \geqslant 1$. To handle this case, we show first that

$$a^2 = 2 + 2u \quad \Rightarrow \quad p = 3. \tag{5}$$

Since $|a|$ is a floating-point number not smaller than 1, there exists a positive integer A such that $|a| = A \cdot 2^{1-p} = A \cdot 2u$. The equality $a^2 = 2 + 2u$ is thus equivalent to $A^2 = (2^p + 1) \cdot 2^{p-1}$ and, using the (unique) decomposition $A = (2B+1) \cdot 2^C$ with $B, C \in \mathbb{N}$, it can also be rewritten $(2B + 1)^2 \cdot 2^{2C} = (2^p + 1) \cdot 2^{p-1}$. Now, $p > 0$ implies that $2^p + 1$ is odd and at least 3, so $B \neq 0$ and $(2B + 1)^2 = 2^p + 1$. The latter equality can be rewritten as $4B(B + 1) = 2^p$ and its unique solution over $\mathbb{N}^2_{>0}$ is $(B, p) = (1, 3)$, so (5) follows.

If $p \neq 3$ then, by (5) we have $a^2 \neq 2+2u$, that is, $A^2 \neq (2^p+1)\cdot 2^{p-1}$. Since the latter inequality involves only integers, it is equivalent to $|A^2 - (2^p + 1) \cdot 2^{p-1}| \geqslant 1$ and thus to $|a^2 - (2 + 2u)| \geqslant 4u^2$. □

Lemma 2 *Let a be a floating-point number. If* $p \neq 3$ *then* $\mathrm{RN}\big(a^2\big) = a^2(1+\epsilon)$ *with* $|\epsilon| \leqslant u/(1 + 3u)$.

Proof We can assume that $1 \leqslant a < 2$. If $a = 1$ then $\mathrm{RN}\big(a^2\big) = a^2$ and the result is clear. If $1 < a < \sqrt{2}$ then it follows from a being a floating-point number that $p \geqslant 4$ and that a belongs to the non-empty interval $[1 + 2u, \sqrt{2})$. Consequently, $1 + 4u < a^2 < 2$ and thus $|\epsilon| \leqslant u\,\mathrm{ufp}(a^2)/a^2 = u/a^2 < u/(1 + 4u)$. Finally, if $\sqrt{2} < a < 2$ then $2 < a^2 < 4$ and, by Lemma 1, it suffices to consider the following four subcases:

- If $2 < a^2 \leqslant 2 + 2u - 4u^2$ then $\mathrm{RN}\big(a^2\big) = 2$ and, therefore,

$$|\epsilon| = 1 - \frac{2}{a^2} \leqslant 1 - \frac{2}{2 + 2u - 4u^2} \leqslant \frac{u}{1 + 3u}.$$

- If $2 + 2u + 4u^2 \leqslant a^2 < 2 + 4u$ then $\mathrm{RN}\big(a^2\big) = 2 + 4u$ and, therefore,

$$|\epsilon| = \frac{2 + 4u}{a^2} - 1 \leqslant \frac{2 + 4u}{2 + 2u + 4u^2} - 1 \leqslant \frac{u}{1 + 3u}.$$

- If $2 + 4u \leqslant a^2 < 2 + 6u$ then $\mathrm{RN}\big(a^2\big) = 2 + 4u$ and, therefore,

$$|\epsilon| = 1 - \frac{2 + 4u}{a^2} \leqslant 1 - \frac{2 + 4u}{2 + 6u} = \frac{u}{1 + 3u}.$$

- If $2 + 6u \leqslant a^2 < 4$ then $\mathrm{ufp}(a^2) = 2$ and $|\epsilon| \leqslant 2u/a^2 \leqslant 2u/(2+6u) = u/(1+3u)$. $\square$

Theorem 1 *If $p \geqslant 4$ then the componentwise relative error for Algorithm 1 satisfies* $\mathrm{E_C} \leqslant 3u$.

Proof Due to the symmetry of Algorithm 1, it suffices to show that $|R - \widehat{R}| \leqslant 3u|R|$. From (3) and Lemma 2 we have

$$s_a = a^2(1+\epsilon_a), \qquad s_b = b^2(1+\epsilon_b), \qquad s = (s_a+s_b)(1+\epsilon_s), \qquad \widehat{R} = \frac{a}{s}(1+\epsilon_R)$$

with $|\epsilon_a|, |\epsilon_b| \leqslant u/(1+3u)$ and $|\epsilon_s|, |\epsilon_R| \leqslant u/(1+u)$. Hence

$$\widehat{R} = \frac{a}{a^2(1+\epsilon_a) + b^2(1+\epsilon_b)} \cdot \frac{1+\epsilon_R}{1+\epsilon_s}$$

and, using $R = a/(a^2+b^2)$, we deduce that $\varphi R \leqslant \widehat{R} \leqslant \varphi' R$ with

$$\varphi := \frac{1 - \frac{u}{1+u}}{(1 + \frac{u}{1+3u})(1 + \frac{u}{1+u})} \qquad \text{and} \qquad \varphi' := \frac{1 + \frac{u}{1+u}}{(1 - \frac{u}{1+3u})(1 - \frac{u}{1+u})}.$$

It is easily checked that $\varphi > 1 - 3u$ and $\varphi' = 1 + 3u$, which completes the proof. $\square$

We conclude this section by showing that the componentwise bound $\mathrm{E_C} \leqslant 3u$ is sharp. More precisely, when the precision p is even, the following example shows that the componentwise error bound $3u$ is asymptotically optimal as $p \to \infty$. Assuming an even $p \geqslant 12$, let us consider the following binary floating-point numbers in precision p:

$$\begin{aligned} a &= 2^{\frac{p}{2}-1} + 5 \cdot 2^{-2} + 2^{-\frac{p}{2}+2}, \\ b &= 2^{p-1} + 2^{\frac{p}{2}-1} + 1. \end{aligned}$$

With these values as inputs of Algorithm 1, we have (the details are provided in Appendix A.1):

$$\begin{aligned} s_a &= 2^{p-2} + 5 \cdot 2^{\frac{p}{2}-2} + 11 \cdot 2^{-1}, \\ s_b &= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 3 \cdot 2^{p-1}, \\ s &= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 2^{p+1}. \end{aligned}$$

From this we deduce

$$\frac{a}{s} = 2^{-\frac{3p}{2}+1} + 2^{-2p} - 2^{-\frac{5p}{2}+1} - 2^{-3p+2} + \mathcal{O}\big(2^{-\frac{7p}{2}}\big),$$

and $\mathrm{ulp}\big(\frac{a}{s}\big) = 2^{-\frac{5p}{2}+2}$. Then, defining the floating-point number τ by

$$\tau = 2^{-\frac{3p}{2}+1} + 2^{-2p} - 2^{-\frac{5p}{2}+2},$$

Table 1 Examples with p odd and a componentwise relative error close to $3u$

p	Inputs a and b	E_C/u
15	$a = 16732$	2.93047...
	$b = 23252{\cdot}2^3$	
17	$a = 66078$	2.96359...
	$b = 93014{\cdot}2^8$	
19	$a = 131435$	2.98509...
	$b = 370969{\cdot}2^8$	
53	$a = 4508053433127332$	2.97894...
	$b = 6369149602646415{\cdot}2^{16}$	
113	$a = 5192393427440123027423416459819356$	2.97647...
	$b = 7343016638055329519853569740503421{\cdot}2^{16}$	

it can be checked that

$$\left|\frac{a}{s} - \tau\right| = \frac{2^{-\frac{5p}{2}+1} + 2^{-\frac{7p}{2}+5}}{1 + 2^{-\frac{p}{2}+1} + 2^{-p+3}} < \frac{1}{2}\mathrm{ulp}\Big(\frac{a}{s}\Big).$$

Hence $\widehat{R} = \mathrm{RN}\big(\frac{a}{s}\big) = \tau$, which together with $R = a/(a^2 + b^2)$ leads to

$$\frac{R - \widehat{R}}{R} = 3u - \frac{31}{2}u^{\frac{3}{2}} + \mathcal{O}\Big(u^2\Big).$$

As a consequence, in this example the componentwise relative error in the computed $\widehat{z}$ is at least $3u - \frac{31}{2}u^{\frac{3}{2}} + \mathcal{O}\big(u^2\big)$, which shows the asymptotic optimality (as $p \to \infty$) of the bound when p is even.

When p is odd, we have not found an input set parametrized by the precision to prove the asymptotic optimality of the error bound $3u$. However, we illustrate the sharpness of the bound by numerical examples in Table 1.

3 Normwise error bound

In this section, we are interested in the normwise relative error of Algorithm 1, that is,

$$\mathrm{E_N} = \sqrt{a^2 + b^2}\sqrt{(R - \widehat{R})^2 + (I - \widehat{I})^2}.$$

The analysis is done in radix 2 and precision p, and we assume that overflows and underflows never occur. If we apply directly the componentwise bound obtained in Section 2, we end up with the normwise error bound $\mathrm{E_N} \leqslant 3u$. In this section, we establish the following result, which achieves a smaller bound by keeping track of the correlations between the various rounding errors committed by the algorithm.

Table 2 Examples with a normwise relative error close to γu

p	Inputs a and b	E_N/u
24	$a = 11863283$	2.69090...
	$b = 11865457 \cdot 2^{12}$	
53	$a = 4503599709991314$	2.70679...
	$b = 6369051770002436 \cdot 2^{26}$	
113	$a = 2^{112}$	2.70559...
	$b = 7343016637207171132572330391109909 \cdot 2^{56}$	

Theorem 2 *If $p \geqslant 10$ then the normwise relative error for Algorithm 1 satisfies $\mathrm{E_N} \leqslant \gamma u + 9u^2$, where γ is defined by*

$$\gamma = \frac{\sqrt{8778980525057 + 16793600(8\sqrt{2} - \sqrt{127}) - 550842155008\sqrt{254}}}{8192\,(16 - \sqrt{254})}, \quad (6)$$

and is such that $\gamma \in (2.70712, 2.70713)$.

If $p \geqslant 10$, $\mathrm{E_N} < 2.70713u + 9u^2$ is therefore a rigorous bound for the normwise error of Algorithm 1. It should also be noticed that the second order term in the error bound can be absorbed by the first order term, at the cost of a slight overestimation: for example, for $p \geqslant 24$, we have $9u = 9 \cdot 2^{-24} < 10^{-6}$ so that $\mathrm{E_N} < 2.707131u$. The numerical examples listed in Table 2 show that the error bound of Theorem 2 is sharp for the basic IEEE 754 formats ($p = 24, 53, 113$).

3.1 Preliminaries

The first step in the error analysis of Algorithm 1 is to reduce the input domain. Since the function RN is symmetric with respect to zero, the signs of a and b are not relevant and we can assume that both a and b are nonnegative. Swapping the inputs a and b does not affect the relative error; moreover, if $a = 0$, then a simple analysis, based on (3), leads to the upper bound $2u$ for $\mathrm{E_N}$, so we can assume that $0 < a \leqslant b$. Finally, multiplying or dividing by two both a and b does not affect either the relative error, and we can restrict the analysis to the case $1 \leqslant b < 2$.

From the definition of the ufp function and this input range reduction, we know that $\mathrm{ufp}(b^2) \in \{1, 2\}$. Moreover, b is a floating-point number, so that $1 \leqslant b \leqslant 2 - 2u$ and thus $1 \leqslant b^2 \leqslant 4 - 4u$. Since $4 - 4u$ is a floating-point number, and using the monotonicity of the rounding function RN, we deduce that $1 \leqslant s_b < 4$. Using again the monotonicity of RN, we also deduce that $0 < s_a < 4$. Hence $1 < s_a + s_b < 8$, which implies $\mathrm{ufp}(s_a + s_b) \in \{1, 2, 4\}$.

We now define δ_a, δ_b, δ_s, δ_R, and δ_I as follows:

$$\begin{aligned}
s_a &= a^2 + \delta_a u, && |\delta_a| \leqslant \mathrm{ufp}\big(a^2\big),\\
s_b &= b^2 + \delta_b u, && |\delta_b| \leqslant \mathrm{ufp}\big(b^2\big),\\
s &= s_a + s_b + \delta_s u, && |\delta_s| \leqslant \mathrm{ufp}\big(s_a + s_b\big),\\
\widehat{R} &= \frac{a}{s} + \delta_R u, && |\delta_R| \leqslant \mathrm{ufp}\left(\frac{a}{s}\right),\\
\widehat{I} &= -\left(\frac{b}{s} + \delta_I u\right), && |\delta_I| \leqslant \mathrm{ufp}\left(\frac{b}{s}\right).
\end{aligned}$$

Let us also define $\delta = \delta_a + \delta_b + \delta_s$ and $\epsilon = \frac{|\delta|}{a^2+b^2}$, so that $|\delta|u$ and ϵu are the absolute and relative errors, respectively, in the evaluation of $a^2 + b^2$. Since $0 < a \leqslant b$, $\mathrm{ufp}\big(b^2\big) \leqslant 2$ and $\mathrm{ufp}(s_a + s_b) \leqslant 4$, we deduce that $|\delta| \leqslant 8$. Moreover, it can be deduced from (3) that $\epsilon \leqslant 2$. (This bound on ϵ already appeared in [3, p. 1471].)

With these notations, we have

$$R - \widehat{R} = \frac{a}{s(a^2+b^2)}\delta u - \delta_R u,$$

and since a similar expression holds for $I - \widehat{I}$, we arrive at

$$\frac{\mathrm{E}_\mathrm{N}^2}{u^2} = \left(a^2 + b^2\right)\left(\delta_R^2 + \delta_I^2\right) - 2\frac{\delta(a\delta_R + b\delta_I)}{a^2 + b^2 + \delta u} + \left(\frac{\delta}{a^2 + b^2 + \delta u}\right)^2.$$

Then, using the triangular inequality, we obtain

$$\begin{aligned}
\frac{\mathrm{E}_\mathrm{N}^2}{u^2} \leqslant{}& \left(a^2 + b^2\right)\left(\mathrm{ufp}\left(\frac{a}{s}\right)^2 + \mathrm{ufp}\left(\frac{b}{s}\right)^2\right)\\
&+ 2\frac{|\delta|\Big(\mathrm{ufp}\left(\frac{a}{s}\right)a + \mathrm{ufp}\left(\frac{b}{s}\right)b\Big)}{a^2 + b^2 - |\delta|u} + \left(\frac{\delta}{a^2 + b^2 - |\delta|u}\right)^2.
\end{aligned}$$

For $p \geqslant 2$, $\epsilon u < 1$ and we use the equality $\frac{1}{a^2+b^2-|\delta|u} = \frac{1}{a^2+b^2}\left(1 + \frac{\epsilon}{1-\epsilon u}u\right)$ and the inequality $\left(1 + \frac{\epsilon}{1-\epsilon u}u\right)^2 \leqslant 1 + \frac{2\epsilon}{(1-\epsilon u)^2}u$ to get

$$\mathrm{E}_\mathrm{N}^2 \leqslant f_2(a,b)u^2 + f_3(a,b)u^3, \tag{7}$$

with

$$\begin{aligned}
f_2(a,b) ={}& \left(a^2 + b^2\right)\left(\mathrm{ufp}\Big(\frac{a}{s}\Big)^2 + \mathrm{ufp}\Big(\frac{b}{s}\Big)^2\right)\\
&+ 2\frac{|\delta|\Big(\mathrm{ufp}\big(\frac{a}{s}\big)a + \mathrm{ufp}\big(\frac{b}{s}\big)b\Big)}{a^2 + b^2} + \left(\frac{\delta}{a^2 + b^2}\right)^2
\end{aligned} \tag{8}$$

and

$$f_3(a,b) = 2\Big(\mathrm{ufp}\Big(\frac{a}{s}\Big)a + \mathrm{ufp}\Big(\frac{b}{s}\Big)b\Big)\frac{\epsilon^2}{1 - \epsilon u} + \frac{2\epsilon^3}{(1 - \epsilon u)^2}.$$

From (4), we have

$$\mathrm{ufp}\left(\frac{a}{s}\right)a + \mathrm{ufp}\left(\frac{b}{s}\right)b \leqslant \frac{a^2+b^2}{s} \leqslant \frac{a^2+b^2}{a^2+b^2-|\delta|u} = \frac{1}{1-\epsilon u},$$

and since $0 \leqslant \epsilon \leqslant 2$, it follows that $f_3(a,b) \leqslant \frac{2\epsilon^2(1+\epsilon)}{(1-\epsilon u)^2} < 25$ for $p \geqslant 10$. Moreover, if f_2 is upper bounded by κ, we can conclude from (7) that

$$\mathrm{E_N} \leqslant \sqrt{\kappa}u + \frac{25}{2\sqrt{\kappa}}u^2. \tag{9}$$

3.2 Taking care of some corner cases

We can first roughly bound f_2 using the inequality $\mathrm{ufp}(t) \leqslant |t|$, valid for any real t, which will allow us to conclude in some particular cases and to further reduce the input domain. From (8) we have

$$\begin{aligned} f_2(a,b) &\leqslant \left(\frac{a^2+b^2}{a^2+b^2-|\delta|u}\right)^2 + 2\frac{|\delta|\left(a^2+b^2\right)}{\left(a^2+b^2\right)\left(a^2+b^2-|\delta|u\right)} + \left(\frac{\delta}{a^2+b^2}\right)^2 \\ &= \left(1+\epsilon+\frac{\epsilon}{1-\epsilon u}u\right)^2. \end{aligned}$$

This last bound is increasing with respect to ϵ and u (*i.e.*, decreasing with respect to the precision p). Therefore, if $\epsilon \leqslant 1 + \frac{\sqrt{2}}{2} + u$, and as soon as $p \geqslant 5$, we have $f_2(a,b) \leqslant \left(2 + \frac{\sqrt{2}}{2} + 3u\right)^2$ and, from (9),

$$\mathrm{E_N} \leqslant \left(2 + \frac{\sqrt{2}}{2}\right)u + 8u^2. \tag{10}$$

Below are five cases that lead to the inequality $\epsilon \leqslant 1 + \frac{\sqrt{2}}{2} + u$, so they can be ignored in the following analysis.

- If $a = b$, then $s_a = s_b$ and $s = s_a + s_b$ so that $\delta_s = 0$ and one can check that $\epsilon \leqslant 1$. In this case, the previous bound (10) holds and we can continue the analysis assuming that
$$a < b. \tag{11}$$
- If $b = 1$, then $s_b = b^2 = 1$ and $\delta_b = 0$. Moreover, from (11) we have $a < 1$, so that $s_a < 1$, which implies $\mathrm{ufp}(1 + s_a) = 1$ and $\epsilon \leqslant 1$. Again, the bound (10) holds and we can continue the analysis assuming that $1 < b$. In fact, since b is a floating-point number, we can assume that
$$1 + 2u \leqslant b. \tag{12}$$
- If $a = 1$, then $\delta_a = 0$ and we can distinguish three cases. If $\mathrm{ufp}\left(b^2\right) = 1$ then $\mathrm{ufp}(1 + s_b) = 2$ and $\epsilon \leqslant \frac{3}{2}$. If $\mathrm{ufp}\left(b^2\right) = 2$ then either $\mathrm{ufp}(1 + s_b) = 2$ which implies $\epsilon \leqslant \frac{4}{3}$, or $\mathrm{ufp}(1 + s_b) = 4$ and then $\epsilon \leqslant \frac{3}{2} + u$. In all these cases, (10) holds, hence we can assume now that
$$a \neq 1. \tag{13}$$

- If $a^2+b^2 < \mathrm{ufp}\,(s_a+s_b)$, then we have $(s_a+s_b)-\mathrm{ufp}(s_a+s_b) < (\delta_a+\delta_b)u \leqslant (a^2+b^2)u < \mathrm{ufp}(s_a+s_b)u = \frac{1}{2}\mathrm{ulp}(s_a+s_b)$. Since $\mathrm{ufp}(s_a+s_b)$ is a floating-point number, we can deduce that $s = \mathrm{RN}(s_a+s_b) = \mathrm{ufp}(s_a+s_b)$ hence $\epsilon \leqslant 1$ and (10) holds. In the following, we can then assume that

$$\mathrm{ufp}(s_a+s_b) \leqslant a^2+b^2. \tag{14}$$

- One last case is when $s_a+s_b \geqslant \sqrt{2}\mathrm{ufp}(s_a+s_b)$. In this case, $\epsilon \leqslant 1+\frac{\sqrt{2}}{2}+u$ and the previous bound (10) holds. Therefore, we now assume that

$$s_a+s_b < \sqrt{2}\mathrm{ufp}(s_a+s_b). \tag{15}$$

3.3 Overview of the case analysis

The analysis goes through the possible values of $\mathrm{ufp}(s_a+s_b)$, which are 1, 2, and 4. In each case, we first deduce upper bounds for $\mathrm{ufp}\big(b^2\big)$, $\mathrm{ufp}\big(\frac{a}{s}\big)$, and $\mathrm{ufp}\big(\frac{b}{s}\big)$. This leads to a new function g, which is greater than or equal to f_2, and which depends on a and b as well as on a third parameter, e, defined as the unique integer such that

$$\mathrm{ufp}\Big(a^2\Big) = 2^{-e}.$$

The function g does not involve floating-point operations anymore and can be seen as a continuous and differentiable function over real inputs. We then look for an upper bound on this function over a restricted domain D containing all the floating-point numbers we are interested in. For this latter step, we mainly use real analysis, especially partial derivatives. In some cases, we can maximize with respect to a and b at the same time. The last step is always to maximize with respect to e, using the change of variable $x = 2^{-e}$ and considering x as a continuous variable.

The analysis is split into seven cases depending on the values of some ufp functions involved in the definition (8) of f_2. Note that, since $a^2 < 4$, we have $e \geqslant -1$. In each case but the last one, we end up with a bound smaller than or equal to $\big(2+\frac{\sqrt{2}}{2}\big)^2$ for f_2, from which we conclude using (9) that $\mathrm{E_N} \leqslant \big(2+\frac{\sqrt{2}}{2}\big)u+5u^2$. The last case is similar although we have a slightly larger bound γ^2+20u for f_2 (we have $2+\frac{\sqrt{2}}{2} = 2.70710\ldots$, while $\gamma = 2.70712\ldots$), which leads to $\mathrm{E_N} \leqslant \gamma u + 9u^2$. The table below summarizes the bounds in each case, under the assumptions (11) to (15).

$\mathrm{ufp}(s_a+s_b)$	$\mathrm{ufp}\big(b^2\big)$	e	$\mathrm{ufp}\big(\frac{a}{s}\big)$	f_2	$\mathrm{E_N}$
1	1	$\geqslant 2$	$\leqslant 2^{-\frac{e}{2}}$	6.565	$2.6u$
4	2	$=-1$	$\leqslant \frac{1}{4}$	$\big(2+\frac{\sqrt{2}}{2}\big)^2$	$\big(2+\frac{\sqrt{2}}{2}\big)u+5u^2$
		$\geqslant 0$	$\leqslant 2^{-2-\frac{e}{2}}$	$\big(\frac{7}{4}+\frac{\sqrt{2}}{2}\big)^2$	$2.5u$
2	1	$\geqslant 1$	$\leqslant 2^{-1-\frac{e}{2}}$	$\big(\frac{7}{4}+\frac{\sqrt{3}}{2}\big)^2$	$2.65u$
		$=0$	$\leqslant \frac{1}{4}$	$\big(\frac{5}{2}\big)^2$	$\frac{5}{2}u+5u^2$
	2	$\geqslant 1$	$\leqslant 2^{-\frac{3+e}{2}}$	$\big(2+\frac{\sqrt{2}}{2}\big)^2$	$\big(2+\frac{\sqrt{2}}{2}\big)u+5u^2$
		$\geqslant 2$, even	$= 2^{-1-\frac{e}{2}}$	γ^2+20u	$\gamma u+9u^2$

We give all the details of the analysis of the first case. For the other cases, we only give a sketch of the analysis, while deferring the details to Appendices A.2 to A.7.

3.4 Case $\mathrm{ufp}(s_a + s_b) = 1$

In this case, we can deduce from (15) that $1 \leqslant s_a + s_b < \sqrt{2}$. As a consequence, we must have $b < \sqrt{2}$ (otherwise we would have $s_a + s_b > 2$), hence

$$\mathrm{ufp}\left(b^2\right) = 1.$$

Since $s_a < \sqrt{2} - 1 < \frac{1}{2}$ and $s_a = \mathrm{RN}\left(a^2\right)$, we have $a^2 < \frac{1}{2}$, and

$$e \geqslant 2.$$

Moreover, we know from (12) that $b \geqslant 1+2u$ so we have $b^2 \geqslant b(1+2u) \geqslant b+2u$, which is a floating-point number because $\mathrm{ufp}(b) = 1$. Consequently $s_b \geqslant b + 2u$ and $s \geqslant s_a + s_b - u \geqslant s_a + b + u > b$, hence $\frac{b}{s} < 1$, which implies

$$\mathrm{ufp}\left(\frac{b}{s}\right) \leqslant \frac{1}{2}.$$

Finally, $s = \mathrm{RN}(s_a + s_b) \geqslant 1$ so $\frac{a}{s} \leqslant a < 2^{\frac{1-e}{2}}$ and

$$\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-\frac{e}{2}}.$$

Therefore, using (8) we deduce in this case that $f_2(a, b) \leqslant g_1(a, b, e)$, with

$$g_1(a, b, e) := \left(a^2 + b^2\right)\left(2^{-e} + \frac{1}{4}\right) + 2\frac{\left(2 + 2^{-e}\right)\left(2^{-\frac{e}{2}}a + \frac{b}{2}\right)}{a^2 + b^2} + \left(\frac{2 + 2^{-e}}{a^2 + b^2}\right)^2.$$

Let us now characterize explicitly the domain over which we will bound $g_1(a, b, e)$. First, we know that $2^{-\frac{e}{2}} \leqslant a < 2^{\frac{1-e}{2}}$. Next, since $s_a + s_b < \sqrt{2}$ and $s_a > 0$, we have $s_b < \sqrt{2}$, so that $b^2 < \sqrt{2} + u$ and $1 < b < \sqrt{\sqrt{2} + u}$. Finally, we have $a^2 + b^2 \leqslant s_a + \mathrm{ufp}\left(a^2\right)u + s_b + \mathrm{ufp}\left(b^2\right)u < \sqrt{2} + \frac{5}{4}u$, which concludes the domain analysis: it suffices to look for an upper bound for g_1 over the domain

$$D_1 := \left\{(a, b, e) \mid 2^{-\frac{e}{2}} \leqslant a < 2^{\frac{1-e}{2}}, 1 \leqslant b < \sqrt{\sqrt{2} + u}, a^2+b^2 < \sqrt{2}+\frac{5}{4}u, e \geqslant 2\right\}.$$

We now compute the partial derivatives of g_1 with respect to a and b,

$$\frac{\partial g_1}{\partial a} = 2a\left(2^{-e} + \frac{1}{4}\right) + \frac{2 + 2^{-e}}{a^2 + b^2}2^{1-\frac{e}{2}} - 4a\frac{\left(2 + 2^{-e}\right)\left(2^{-\frac{e}{2}}a + \frac{b}{2}\right)}{\left(a^2 + b^2\right)^2} - 4a\frac{\left(2 + 2^{-e}\right)^2}{\left(a^2 + b^2\right)^3},$$

$$\frac{\partial g_1}{\partial b} = 2b\left(2^{-e} + \frac{1}{4}\right) + \frac{2 + 2^{-e}}{a^2 + b^2} - 4b\frac{\left(2 + 2^{-e}\right)\left(2^{-\frac{e}{2}}a + \frac{b}{2}\right)}{\left(a^2 + b^2\right)^2} - 4b\frac{\left(2 + 2^{-e}\right)^2}{\left(a^2 + b^2\right)^3},$$

and the next step is to prove that they are both negative over the domain D_1. Since $\frac{1}{b}\frac{\partial}{\partial b}g_1(a, b, e) - \frac{1}{a}\frac{\partial}{\partial a}g_1(a, b, e) = \frac{2+2^{-e}}{a^2+b^2}\left(\frac{1}{b} - \frac{1}{a}2^{1-\frac{e}{2}}\right) < 0$ over D_1, it is sufficient to prove that $\frac{\partial}{\partial a} g_1(a,b,e) < 0$. Since $2a\frac{2+2^{-e}}{a^2+b^2} > 0$, we can rewrite this inequality as

$$\frac{\left(2^{-e}+\frac{1}{4}\right)\left(a^2+b^2\right)}{2+2^{-e}} + \frac{2^{-\frac{e}{2}}}{a} < 2\frac{2^{-\frac{e}{2}}a+\frac{b}{2}}{a^2+b^2} + 2\frac{2+2^{-e}}{\left(a^2+b^2\right)^2}.$$

This inequality follows from the following three relations:

$$\frac{\left(2^{-e}+\frac{1}{4}\right)\left(a^2+b^2\right)}{2+2^{-e}} + \frac{2^{-\frac{e}{2}}}{a} < \frac{\sqrt{2}+\frac{5}{4}u}{4} + 1 \quad \text{for } (a,b,e) \in D_1,$$

$$\frac{\sqrt{2}+\frac{5}{4}u}{4} + 1 < \frac{1}{\sqrt{2}+\frac{5}{4}u} + \frac{4}{\left(\sqrt{2}+\frac{5}{4}u\right)^2} \quad \text{for } p \geqslant 3,$$

$$\frac{1}{\sqrt{2}+\frac{5}{4}u} + \frac{4}{\left(\sqrt{2}+\frac{5}{4}u\right)^2} < 2\frac{2^{-\frac{e}{2}}a+\frac{b}{2}}{a^2+b^2} + 2\frac{2+2^{-e}}{\left(a^2+b^2\right)^2} \quad \text{for } (a,b,e) \in D_1.$$

Since both $\frac{\partial g_1}{\partial a}$ and $\frac{\partial g_1}{\partial b}$ are negative over D_1, since $(a,b,e) \in D_1$ implies $a \geqslant 2^{-\frac{e}{2}}$ and $b \geqslant 1$, and since $(2^{-\frac{e}{2}},1,e) \in D_1$, we deduce that $g_1(a,b,e) \leqslant g_1(2^{-\frac{e}{2}},1,e) =: h_1(x)$, with $x = 2^{-e}$ and

$$h_1(x) = (x+1)\left(x+\frac{1}{4}\right) + \frac{(x+2)(2x+1)}{x+1} + \left(\frac{x+2}{x+1}\right)^2.$$

Since $e \geqslant 2$, we have $0 < x \leqslant \frac{1}{4}$ and

$$h_1'(x) = \frac{8x^4+37x^3+63x^2+43x+1}{4(x+1)^3} > 0.$$

Overall, we thus have $f_2(a,b) \leqslant g_1(a,b,e) \leqslant h_1(x) \leqslant h_1\left(\frac{1}{4}\right) = 6.565$.

3.5 Case $\mathrm{ufp}(s_a + s_b) = 4$

From (15) and (11), we know that $4 \leqslant s_a + s_b < 4\sqrt{2}$ and $s_a < s_b$. As a consequence, we have $2 < s_b$ which implies $2 < b^2$, so that

$$\mathrm{ufp}\left(b^2\right) = 2 \quad \text{and} \quad \sqrt{2} < b \leqslant 2 - 2u.$$

Since 4 is a floating-point number, we have $s = \mathrm{RN}(s_a + s_b) \geqslant 4$ and $\frac{b}{s} \leqslant \frac{b}{4} < \frac{1}{2}$ hence

$$\mathrm{ufp}\left(\frac{b}{s}\right) \leqslant \frac{1}{4}.$$

In the same way, $\frac{a}{s} \leqslant \frac{a}{4} < 2^{-\frac{3+e}{2}}$ so that

$$\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-2-\frac{e}{2}}.$$

We now distinguish two subcases, namely $e = -1$ and $e \geqslant 0$.

3.5.1 Subcase $e = -1$

We have $\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-\frac{3}{2}}$, hence $\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant \frac{1}{4}$, thus we deduce from (8) that $f_2(a,b) \leqslant g_2(a,b)$ with

$$g_2(a,b) := \frac{a^2+b^2}{8} + \frac{4(a+b)}{a^2+b^2} + \left(\frac{8}{a^2+b^2}\right)^2.$$

From (15), we know that $s_a + s_b < 4\sqrt{2}$, which implies $a^2+b^2 < 4\sqrt{2}+4u$. The domain of interest is then given by

$$D_2 := \{(a,b) \mid \sqrt{2} \leqslant a \leqslant b < 2,\ a^2+b^2 < 4\sqrt{2}+4u\}.$$

Computing the partial derivatives of g_2 with respect to a and b, and proving that they are both negative over the domain D_2 (detailed computations are in Appendix A.2), we end up with $f_2(a,b) \leqslant g_2(\sqrt{2},\sqrt{2}) = \left(2+\frac{\sqrt{2}}{2}\right)^2$.

3.5.2 Subcase $e \geqslant 0$

Since $|\delta| \leqslant \mathrm{ufp}\left(a^2\right) + \mathrm{ufp}\left(b^2\right) + \mathrm{ufp}(s_a+s_b) = 6+2^{-e}$ and $\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-2-\frac{e}{2}}$, from (8) we get $f_2(a,b) \leqslant g_3(a,b,e)$ with

$$g_3(a,b,e) := \frac{(a^2+b^2)\left(2^{-e}+1\right)}{16} + \frac{(6+2^{-e})(2^{-\frac{e}{2}}a+b)}{2\left(a^2+b^2\right)} + \left(\frac{6+2^{-e}}{a^2+b^2}\right)^2.$$

From (14), a^2+b^2 is lower bounded by 4, and we restrict the analysis of g_3 to the domain

$$D_3 := \{(a,b,e) \mid 2^{-\frac{e}{2}} \leqslant a \leqslant 2^{\frac{1-e}{2}},\ \sqrt{2} \leqslant b < 2, 4 \leqslant a^2+b^2 < 4\sqrt{2}+4u, e \geqslant 0\}.$$

First, it can be checked that the partial derivative of g_3 with respect to b is negative over D_3 (details are in Appendix A.3). Since $b \geqslant \sqrt{4-a^2}$, and $(a,b,e) \in D_3$ implies $(a,\sqrt{4-a^2},e) \in D_3$, we deduce that $g_3(a,b,e) \leqslant g_3(a,\sqrt{4-a^2},e)$, where

$$g_3(a,\sqrt{4-a^2},e) = \frac{2^{-e}+1}{4} + \frac{(6+2^{-e})(2^{-\frac{e}{2}}a+\sqrt{4-a^2})}{8} + \frac{\left(6+2^{-e}\right)^2}{16}.$$

We then compute $\frac{\partial}{\partial a} g_3(a,\sqrt{4-a^2},e) = \frac{6+2^{-e}}{8}\left(2^{-\frac{e}{2}} - \frac{a}{\sqrt{4-a^2}}\right)$, which is nonnegative because $a^2 \leqslant \frac{2a^2}{1+2^{-e}} \leqslant \frac{4\cdot 2^{-e}}{1+2^{-e}}$. Since $(2^{\frac{1-e}{2}}, \sqrt{4-2^{1-e}}, e) \in D_3$, we have $g_3(a,b,e) \leqslant g_3(2^{\frac{1-e}{2}}, \sqrt{4-2^{1-e}}, e) =: h_3(x)$, with $x = 2^{-e}$ and

$$h_3(x) = \frac{x+1}{4} + \frac{(6+x)(\sqrt{2}x+\sqrt{4-2x})}{8} + \frac{(6+x)^2}{16}.$$

Since

$$h_3'(x) = 1 + \frac{x}{8}\left(1+\sqrt{2}\right) + \frac{\sqrt{4-2x}}{8} + \frac{x+6}{8}\left(\sqrt{2} - \frac{1}{\sqrt{4-2x}}\right)$$

is positive for $0 < x \leqslant 1$, we deduce $f_2(a,b) \leqslant h_3(1) = \left(\frac{7}{4} + \frac{\sqrt{2}}{2}\right)^2 = 6.037\ldots$.

3.6 Case $\mathrm{ufp}(s_a + s_b) = 2$

From (14) we have $2 \leqslant a^2 + b^2$, and from (15) we have $2 \leqslant s_a + s_b < 2\sqrt{2}$ hence

$$e \geqslant 0.$$

Since 2 is a floating-point number, we know that $s \geqslant 2$. Therefore $\frac{a}{s} < 2^{-\frac{1+e}{2}}$, hence

$$\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-1-\frac{e}{2}}, \tag{16}$$

and $\frac{b}{s} < 1$ so that

$$\mathrm{ufp}\left(\frac{b}{s}\right) \leqslant \frac{1}{2}.$$

We handle separately the two possible values, 1 and 2, for $\mathrm{ufp}\big(b^2\big)$.

3.6.1 Subcase $\mathrm{ufp}\big(b^2\big) = 1$

We distinguish the cases $e \geqslant 1$ and $e = 0$.

- *Subsubcase* $e \geqslant 1$: From (8) we have $f_2(a,b) \leqslant g_4(a,b,e)$ with

$$g_4(a,b,e) := \frac{\left(a^2+b^2\right)\left(2^{-e}+1\right)}{4} + \frac{\left(3+2^{-e}\right)\left(2^{-\frac{e}{2}}a+b\right)}{a^2+b^2} + \left(\frac{3+2^{-e}}{a^2+b^2}\right)^2.$$

  From (14), we know that a^2+b^2 is lower bounded by 2. On the other hand, we have $a^2+b^2 \leqslant s_a + s_b + \left(\mathrm{ufp}\big(a^2\big) + \mathrm{ufp}\big(b^2\big)\right)u < 2\sqrt{2} + 2u$ and $1 < b < \sqrt{2}$, hence we can restrict the analysis to the domain

$$D_4 := \{(a,b,e) \mid 2^{-\frac{e}{2}} \leqslant a < 2^{\frac{1-e}{2}}, 1 < b < \sqrt{2}, 2 \leqslant a^2+b^2 < 2\sqrt{2}+2u, e \geqslant 1\}.$$

  We can first compute the partial derivative of g_4 with respect to b and prove it is negative over D_4 for $p \geqslant 4$ (see the details in Appendix A.4). Since $(a, \sqrt{2-a^2}, e)$ is in D_4, we deduce that $g_4(a,b,e) \leqslant g_4(a, \sqrt{2-a^2}, e)$, and we have

$$g_4(a, \sqrt{2-a^2}, e) = \frac{2^{-e}+1}{2} + \frac{\left(3+2^{-e}\right)\left(2^{-\frac{e}{2}}a + \sqrt{2-a^2}\right)}{2} + \frac{\left(3+2^{-e}\right)^2}{4}.$$

Next, we can compute the derivative of $g_4(a, \sqrt{2-a^2}, e)$ with respect to a (see Appendix A.4) and check that the maximum is attained at $a_0 = 2^{-\frac{e}{2}}\sqrt{\frac{2}{1+2^{-e}}}$, so that $g_4(a,b,e) \leqslant g_4(a_0, \sqrt{2-a_0^2}, e) =: h_4(x)$ with

$$h_4(x) = \frac{x+1}{2} + \frac{3+x}{2}\left(x\sqrt{\frac{2}{1+x}} + \sqrt{2 - \frac{2x}{1+x}}\right) + \frac{(3+x)^2}{4}.$$

Since $h_4'(x) > 0$ for $0 < x \leqslant \frac{1}{2}$, we conclude that $f_2(a,b) \leqslant g_4(a,b,e) \leqslant h_4\left(\frac{1}{2}\right) = \left(\frac{7}{4} + \frac{\sqrt{3}}{2}\right)^2$.

- *Subsubcase* $e = 0$: According to (13), we assume that $1 < a$, so that $\mathrm{ufp}\left(b^2\right) = \mathrm{ufp}\left(a^2\right) = 1$. It follows that $s \geqslant s_a + s_b - 2u \geqslant a^2 + b^2 - 4u$, hence $\frac{a}{s} \leqslant \frac{a}{a^2+b^2-4u}$. Since a and b are both floating-point numbers, and from (11), we know that $b \geqslant a + 2u$ so that $b^2 - 4u > a^2$. By computing its partial derivative, it can then be checked that $\frac{a}{a^2+b^2-4u}$ is increasing with respect to a, which implies $\frac{a}{s} \leqslant \frac{b-2u}{(b-2u)^2+b^2-4u}$. This last expression is decreasing with respect to b, and since $b \geqslant 1 + 2u$ we deduce $\frac{a}{s} \leqslant \frac{1}{2(1+2u^2)} < \frac{1}{2}$. Thus,

$$\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant \frac{1}{4}.$$

In the same way, it can be derived from $\frac{b}{s} \leqslant \frac{b}{a^2+b^2-4u}$ that

$$\mathrm{ufp}\left(\frac{b}{s}\right) \leqslant \frac{1}{4}.$$

Combining these bounds on $\mathrm{ufp}\left(\frac{a}{s}\right)$ and $\mathrm{ufp}\left(\frac{b}{s}\right)$ with (8) gives $f_2(a,b) \leqslant g_5(a,b)$, where

$$g_5(a,b) := \frac{a^2+b^2}{8} + \frac{2\,(a+b)}{a^2+b^2} + \frac{16}{\left(a^2+b^2\right)^2}.$$

Hence it remains to bound $g_5(a,b)$ over the domain D_5 defined by

$$D_5 := \{(a,b) \mid 1 \leqslant a \leqslant b < \sqrt{2} \text{ and } a^2 + b^2 < 2\sqrt{2} + 2u\}.$$

In this domain, we have $\frac{\partial}{\partial b} g_5(a,b) < 0$ (details are in Appendix A.5), so that $g_5(a,b) \leqslant g_5(a,a) = \frac{a^2}{4} + \frac{4}{a^4} + \frac{2}{a}$, which is maximal for $a = 1$. Therefore, we deduce that $f_2(a,b) \leqslant g_5(a,b) \leqslant g_5(1,1) = \left(\frac{5}{2}\right)^2$.

3.6.2 *Subcase* $\mathrm{ufp}\left(b^2\right) = 2$

In this paragraph, $a^2 < 1$ (otherwise we would have $s_a + s_b \geqslant 2 + 1$ while from (15) we have $s_a + s_b < 2\sqrt{2}$), hence $e \geqslant 1$. We split the inequality (16) into two possible cases. Either $\mathrm{ufp}\left(\frac{a}{s}\right) < 2^{-1-\frac{e}{2}}$ which implies $\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-\frac{3+e}{2}}$, or $\mathrm{ufp}\left(\frac{a}{s}\right) = 2^{-1-\frac{e}{2}}$ in which case e is even.

- *Subsubcase* $\mathrm{ufp}\left(\frac{a}{s}\right) < 2^{-1-\frac{e}{2}}$: We deduce from (8) and $|\delta| \leqslant 4 + 2^{-e}$ that $f_2(a,b) \leqslant g_6(a,b,e)$ with

$$g_6(a,b,e) := \frac{\left(a^2+b^2\right)\left(2^{-1-e}+1\right)}{4} + \frac{\left(4+2^{-e}\right)\left(2^{-\frac{1+e}{2}}a+b\right)}{a^2+b^2} + \left(\frac{4+2^{-e}}{a^2+b^2}\right)^2.$$

We can compute the derivatives of g_6 (details are provided in Appendix A.6) with respect to a and b and prove that they are negative over the domain

$$D_6 := \{(a,b,e) \mid 2^{-\frac{e}{2}} \leqslant a < 2^{\frac{1-e}{2}}, \sqrt{2} \leqslant b < 2, \\ 2 \leqslant a^2+b^2 < 2\sqrt{2} + \left(2+2^{-e}\right)u, e \geqslant 1\}.$$

For $(a,b,e) \in D_6$, we deduce that $g_6(a,b,e) \leqslant g_6(2^{-\frac{e}{2}}, \sqrt{2}, e) =: h_6(x)$ with

$$h_6(x) = \frac{(x+2)\left(\frac{x}{2}+1\right)}{4} + \frac{\sqrt{2}\,(4+x)\left(\frac{x}{2}+1\right)}{x+2} + \left(\frac{4+x}{x+2}\right)^2, \quad x = 2^{-e}.$$

We can maximize $h_6(x)$ for $0 < x \leqslant \frac{1}{2}$, which leads to $f_2(a,b) \leqslant h_6(0) = \left(2+\frac{\sqrt{2}}{2}\right)^2$.

- *Subsubcase* $\mathrm{ufp}\left(\frac{a}{s}\right) = 2^{-1-\frac{e}{2}}$: In this case, e is even, hence $e \geqslant 2$. We have $f_2(a,b) \leqslant g_7(a,b,e)$ with

$$g_7(a,b,e) := \frac{\left(a^2+b^2\right)\left(2^{-e}+1\right)}{4} + \frac{\left(4+2^{-e}\right)\left(2^{-\frac{e}{2}}a+b\right)}{a^2+b^2} + \left(\frac{4+2^{-e}}{a^2+b^2}\right)^2.$$

The lower bound $2^{-\frac{e}{2}}$ for a does not lead to a sufficiently tight bound for f_2 in this case: to get a better bound, we exploit further the hypothesis $\mathrm{ufp}\left(\frac{a}{s}\right) = 2^{-1-\frac{e}{2}}$. This gives $s2^{-1-\frac{e}{2}} \leqslant a$, which implies $a^2 - 2^{1+\frac{e}{2}}a + b^2 + \delta u \leqslant 0$, hence

$$a \geqslant 2^{\frac{e}{2}} - \sqrt{2^e - 2 + (4+2^{-e})u} = a_0 + \eta(u)$$

with

$$a_0 = 2^{\frac{e}{2}} - \sqrt{2^e-2}, \quad \eta(u) < 0, \quad |\eta(u)| \in \mathcal{O}(u).$$

Therefore, we analyze g_7 over the domain

$$D_7 := \{(a,b,e) \mid a_0 + \eta(u) \leqslant a < 2^{\frac{1-e}{2}}, \sqrt{2} \leqslant b < 2, \\ 2 \leqslant a^2+b^2 < 2\sqrt{2} + \left(2+2^{-e}\right)u, e \geqslant 2, e \text{ even}\}.$$

First, we can compute the partial derivative of g_7 with respect to b and prove (see Appendix A.7) that it is negative over the domain D_7, hence we know that $g_7(a,b,e) \leqslant g_7(a,\sqrt{2},e)$.

It can be checked that $a_0 + \eta(u)$ belongs to $\left[2^{-1-\frac{e}{2}}, 2^{\frac{1-e}{2}}\right]$ and that $g_7(a,\sqrt{2},e)$ is decreasing with respect to a over $\left[2^{-1-\frac{e}{2}}, 2^{\frac{1-e}{2}}\right]$;

see Appendix A.7. We then deduce that $g_7(a, \sqrt{2}, e) \leqslant g_7(a_0 + \eta(u), \sqrt{2}, e)$, for any $(a, b, e) \in D_7$.

Next, it can be proved that $g_7(a_0+\eta(u), \sqrt{2}, e) \leqslant g_7(a_0, \sqrt{2}, e)+20u$ (again, the details are provided in Appendix A.7). As a consequence, for any $(a, b, e) \in D_7$ we have $g_7(a, b, e) \leqslant g_7(a_0, \sqrt{2}, e) + 20u$.

The last step is to bound $g_7(a_0, \sqrt{2}, e)$ for e an even positive integer. With $y = \sqrt{1 - 2^{1-e}}$, we have $g_7(a_0, \sqrt{2}, e) =: h_7(y)$ with $h_7(y)$ a rational function in y. The variable y belongs to $\left[\sqrt{2}/2, 1\right]$, and $h_7'(y) = \frac{P(y)}{32\,(y+1)^2}$ with

$$\begin{aligned} P(y) = 3y^7 + 11y^6 - 5y^5 \; &- \; (12\sqrt{2} + 85)y^4 - (32\sqrt{2} + 143)y^3 \\ &- \; (23 - 8\sqrt{2})y^2 + (64\sqrt{2} + 113)y + 36\sqrt{2} + 33. \end{aligned}$$

Using Descartes' rule of signs, one can check that P has exactly one root in the interval $\left[\sqrt{2}/2, 1\right]$, and since the evaluation of P is positive at $\sqrt{1 - 2^{-5}}$ and negative at $\sqrt{1 - 2^{-7}}$, we deduce that h_7 is increasing over $\left[\sqrt{2}/2, \sqrt{1 - 2^{-5}}\right]$ and decreasing over $\left[\sqrt{1 - 2^{-7}}, 1\right]$. Comparing the values of h_7 at the points $\sqrt{1 - 2^{-5}}$ and $\sqrt{1 - 2^{-7}}$, we conclude that $h_7(\sqrt{1 - 2^{-7}})$ is an upper bound for h_7.

Finally, it can be checked that $h_7(\sqrt{1 - 2^{-7}}) = \gamma^2$ hence we get $f_2(a, b) \leqslant \gamma^2 + 20u$. From (9), we derive the final upper bound $\gamma u + 9u^2$ for $\mathrm{E_N}$, which concludes the proof of Theorem 2.

4 Implications for complex floating-point division

Let us conclude with some remarks about complex division. The conventional complex division algorithm for computing an approximation $\widehat{z} = \widehat{R} + i\,\widehat{I}$ of $(a+ib)/(c+id)$ in floating-point arithmetic consists in evaluating the real part as

$$\widehat{R} = \mathrm{RN}\left(\frac{\mathrm{RN}(\mathrm{RN}(ac) + \mathrm{RN}(bd))}{\mathrm{RN}\big(\mathrm{RN}\big(c^2\big) + \mathrm{RN}\big(d^2\big)\big)}\right) \tag{17}$$

and using a similar scheme for the imaginary part. An approximate quotient $\widehat{z}$ can also be obtained by first computing an inverse of $c + id$ using Algorithm 1, and then multiplying it by $a + ib$ by means of the classic complex multiplication algorithm. Note that both algorithms require 3 additions/subtractions, 6 multiplications, and 2 divisions.

Normwise relative accuracy analyses of the method based on (17) can be found in [4, 5] and [12]. To our knowledge, the best known upper bound for the normwise relative error generated by this method is $(3 + \sqrt{5})u + \mathcal{O}\big(u^2\big) \approx 5.2u$: as noted in [1], this bound can be derived from the bound $\sqrt{5}u$ from [3] on the normwise relative error for the classic complex multiplication algorithm. On the other hand, it can be checked using Theorem 2 and, again, the bound $\sqrt{5}u$ from [3] that the algorithm combining inversion and multiplication admits the smaller normwise error bound $(\gamma + \sqrt{5})u + \mathcal{O}(u^2) \approx 4.9u$. The following examples of complex quotients in precision $p = 11$ show that in both cases the largest normwise relative error cannot be bounded by $\gamma u + \mathcal{O}(u^2) \approx 2.7u$ as for inversion:

- with $a + ib = 1575 + i\,1419$ and $c + id = 1457 + i\,1480$, using (17) gives $|\widehat{z} - z|/(u|z|) = 4.67973\ldots$;
- dividing $1506 + i\,1512$ by $1491 + i\,1504$ using the inversion-multiplication approach leads to $|\widehat{z} - z|/(u|z|) = 4.34446\ldots$

However, these examples are not sufficient to conclude about the sharpness of the bounds $(3 + \sqrt{5})u + \mathcal{O}(u^2)$ and $(\gamma + \sqrt{5})u + \mathcal{O}(u^2)$, and further investigation is needed to understand the accuracy of complex floating-point division.

Acknowledgments We thank the associate editor and the anonymous reviewers for their helpful comments and suggestions.

Appendix

A Details omitted in the proofs

A.1 Asymptotic optimality of the componentwise error bound

We briefly detail the computations of s_a, s_b and s in the example parametrized by p given in Section 2. We assume that $p \geqslant 12$ is even, and we recall that

$$\begin{aligned} a &= 2^{\frac{p}{2}-1} + 5 \cdot 2^{-2} + 2^{-\frac{p}{2}+2}, \\ b &= 2^{p-1} + 2^{\frac{p}{2}-1} + 1. \end{aligned}$$

- Computation of $s_a = \mathrm{RN}(a^2)$:

$$\begin{aligned} a^2 &= 2^{p-2} + 5 \cdot 2^{\frac{p}{2}-2} + 11 \cdot 2^{-1} + 2^{-4} + 10 \cdot 2^{-\frac{p}{2}} + 2^{-p+4} \\ \mathrm{ulp}\left(a^2\right) &= 2^{-1} \\ \widetilde{s_a} &:= 2^{p-2} + 5 \cdot 2^{\frac{p}{2}-2} + 11 \cdot 2^{-1} \\ |a^2 - \widetilde{s_a}| &= 2^{-4} + 10 \cdot 2^{-\frac{p}{2}} + 2^{4-p} \\ &\leqslant 2^{-4} + 10 \cdot 2^{-6} + 2^{-8} \\ &< 2^{-2} = \frac{1}{2}\mathrm{ulp}\left(a^2\right) \end{aligned}$$

Hence $s_a = \widetilde{s_a}$.

- Computation of $s_b = \mathrm{RN}\left(b^2\right)$:

$$
\begin{aligned}
b^2 &= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 2^p + 2^{p-2} + 2^{\frac{p}{2}} + 1 \\
\widetilde{s_b} &:= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 3 \cdot 2^{p-1} \\
\mathrm{ulp}\left(b^2\right) &= 2^{p-1} \\
|b^2 - \widetilde{s_b}| &= 2^{p-2} - 2^{\frac{p}{2}} - 1 \\
&< 2^{p-2} = \frac{1}{2}\mathrm{ulp}\left(b^2\right)
\end{aligned}
$$

Hence $s_b = \widetilde{s_b}$.

- Computation of $s = \mathrm{RN}(s_a + s_b)$:

$$
\begin{aligned}
s_a + s_b &= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 3 \cdot 2^{p-1} + 2^{p-2} + 5 \cdot 2^{\frac{p}{2}-2} + 11 \cdot 2^{-1} \\
\widetilde{s} &= 2^{2p-2} + 2^{\frac{3p}{2}-1} + 2^{p+1} \\
\mathrm{ulp}(s_a + s_b) &= 2^{p-1} \\
|s_a + s_b - \widetilde{s}| &= 2^{p-2} - 5 \cdot 2^{\frac{p}{2}-2} - 11 \cdot 2^{-1} \\
&< 2^{p-2} = \frac{1}{2}\mathrm{ulp}(s_a + s_b)
\end{aligned}
$$

Hence $s = \widetilde{s}$.

A.2 Partial derivatives of g_2

Computing the partial derivatives of g_2 with respect to a and b gives

$$
\begin{aligned}
\frac{\partial g_2}{\partial a} &= \frac{a}{4} + \frac{4}{a^2+b^2} - \frac{8a(a+b)}{\left(a^2+b^2\right)^2} - \frac{256a}{\left(a^2+b^2\right)^3}, \\
\frac{\partial g_2}{\partial b} &= \frac{b}{4} + \frac{4}{a^2+b^2} - \frac{8b(a+b)}{\left(a^2+b^2\right)^2} - \frac{256b}{\left(a^2+b^2\right)^3}.
\end{aligned}
$$

First, we know that $b > a$ so $\frac{1}{b}\frac{\partial}{\partial b} g_2(a,b) < \frac{1}{a}\frac{\partial}{\partial a} g_2(a,b)$. We just have to prove that $\frac{\partial}{\partial a} g_2(a,b) < 0$, that is,

$$
\frac{\left(a^2+b^2\right)^2}{4} + \frac{4\left(a^2+b^2\right)}{a} < 8(a+b) + \frac{256}{a^2+b^2}.
$$

Since for $(a,b) \in D_2$ we have $\sqrt{2} < a, b$, and $a^2 + b^2 < 4\sqrt{2} + 4u$, it is enough to check that

$$
\frac{\left(4\sqrt{2}+4u\right)^2}{4} + \frac{4\left(4\sqrt{2}+4u\right)}{\sqrt{2}} < 16\sqrt{2} + \frac{256}{4\sqrt{2}+4u},
$$

which holds for $p \geqslant 2$.

A.3 Partial derivative of g_3

We compute the partial derivative of g_3 with respect to b, and check that this derivative is negative over the domain D_3. We have

$$\frac{\partial g_3}{\partial b} = \frac{b\left(2^{-e}+1\right)}{8} + \frac{6+2^{-e}}{2\left(a^2+b^2\right)} - b\frac{\left(6+2^{-e}\right)\left(2^{-\frac{e}{2}}a+b\right)}{\left(a^2+b^2\right)^2} - 4b\frac{\left(6+2^{-e}\right)^2}{\left(a^2+b^2\right)^3},$$

and we check that

$$\frac{b\left(2^{-e}+1\right)}{8} + \frac{6+2^{-e}}{2\left(a^2+b^2\right)} < b\frac{\left(6+2^{-e}\right)\left(2^{-\frac{e}{2}}a+b\right)}{\left(a^2+b^2\right)^2} + 4b\frac{\left(6+2^{-e}\right)^2}{\left(a^2+b^2\right)^3}.$$

Multiplying both sides by $\frac{(a^2+b^2)^2}{b(6+2^{-e})}$ and since $1 \leqslant b$, it is enough to prove that

$$\frac{\left(2^{-e}+1\right)\left(a^2+b^2\right)^2}{8\left(6+2^{-e}\right)} + \frac{a^2+b^2}{2} < 2^{-\frac{e}{2}}a + b + 4\frac{6+2^{-e}}{a^2+b^2}.$$

This follows from the following sequence of three inequalities

$$\begin{aligned}
\frac{\left(2^{-e}+1\right)\left(a^2+b^2\right)^2}{8\left(6+2^{-e}\right)} + \frac{a^2+b^2}{2} &< \frac{2\left(4\sqrt{2}+4u\right)^2}{48} + \frac{4\sqrt{2}+4u}{2},\\
\frac{2\left(4\sqrt{2}+4u\right)^2}{48} + \frac{4\sqrt{2}+4u}{2} &< 4\frac{6}{4\sqrt{2}+4u} + 1 \quad \text{for } p \geqslant 3,\\
4\frac{6}{4\sqrt{2}+4u} + 1 &< 4\frac{6+2^{-e}}{a^2+b^2} + 2^{-\frac{e}{2}}a + b.
\end{aligned}$$

A.4 Partial derivatives of g_4

The partial derivative of g_4 with respect to b is given by

$$\frac{\partial g_4}{\partial b} = \frac{b\left(2^{-e}+1\right)}{2} + \frac{3+2^{-e}}{a^2+b^2} - 2b\frac{\left(2^{-\frac{e}{2}}a+b\right)\left(3+2^{-e}\right)}{\left(a^2+b^2\right)^2} - 4b\frac{\left(3+2^{-e}\right)^2}{\left(a^2+b^2\right)^3}.$$

We want to prove that $\frac{\partial}{\partial b}g_4(a, b, e) < 0$ or, equivalently, that

$$\frac{\left(a^2+b^2\right)^2\left(2^{-e}+1\right)}{2\left(3+2^{-e}\right)} + \frac{a^2+b^2}{b} < 2\left(2^{-\frac{e}{2}}a+b\right) + 4\frac{3+2^{-e}}{a^2+b^2}.$$

This inequality can be derived from the following ones:

$$\frac{\left(a^2+b^2\right)^2\left(2^{-e}+1\right)}{2\left(3+2^{-e}\right)}+\frac{a^2+b^2}{b} < \frac{2\left(2\sqrt{2}+2u\right)^2}{6}+2\sqrt{2}+2u,$$

$$\frac{\left(2\sqrt{2}+2u\right)^2}{3}+2\sqrt{2}+2u < 2+\frac{12}{2\sqrt{2}+2u} \quad \text{for } p \geqslant 4,$$

$$2+\frac{12}{2\sqrt{2}+2u} < 2\left(2^{-\frac{e}{2}}a+b\right)+4\frac{3+2^{-e}}{a^2+b^2}.$$

The partial derivative of $g_4(a, \sqrt{2-a^2}, e)$ with respect to a is

$$\frac{\partial}{\partial a}g_4(a,\sqrt{2-a^2},e)=\frac{3+2^{-e}}{2}\left(2^{-\frac{e}{2}}-\frac{a}{\sqrt{2-a^2}}\right),$$

which is zero if $a = a_0$ with $a_0 = 2^{-\frac{e}{2}}\sqrt{\frac{2}{1+2^{-e}}}$, positive if $a < a_0$, and negative if $a > a_0$.

A.5 Partial derivative of g_5

We have

$$\frac{\partial g_5}{\partial b}=\frac{b}{4}+\frac{2}{a^2+b^2}-\frac{4\left(a+b\right)}{\left(a^2+b^2\right)^2}b-\frac{64}{\left(a^2+b^2\right)^3}b,$$

and it can be checked that this partial derivative is negative using the following inequalities:

$$\frac{\left(a^2+b^2\right)^2}{4}+\frac{2}{b}\left(a^2+b^2\right) < \frac{\left(2\sqrt{2}+2u\right)^2}{4}+2\left(2\sqrt{2}+2u\right),$$

$$\frac{\left(2\sqrt{2}+2u\right)^2}{4}+2\left(2\sqrt{2}+2u\right) < 8+\frac{64}{2\sqrt{2}+2u} \quad \text{for } p \geqslant 2,$$

$$8+\frac{64}{2\sqrt{2}+2u} < 4\left(a+b\right)+\frac{64}{a^2+b^2}.$$

A.6 Partial derivatives of g_6

The partial derivatives of g_6 with respect to a and b are

$$\frac{\partial g_6}{\partial a}=\frac{a}{4}\left(2^{-e}+2\right)+\frac{4+2^{-e}}{a^2+b^2}2^{-\frac{1+e}{2}}-2a\frac{\left(2^{-\frac{1+e}{2}}a+b\right)\left(4+2^{-e}\right)}{\left(a^2+b^2\right)^2}-4a\frac{\left(4+2^{-e}\right)^2}{\left(a^2+b^2\right)^3}$$

and

$$\frac{\partial g_6}{\partial b}=\frac{b}{4}\left(2^{-e}+2\right)+\frac{4+2^{-e}}{a^2+b^2}-2b\frac{\left(2^{-\frac{1+e}{2}}a+b\right)\left(4+2^{-e}\right)}{\left(a^2+b^2\right)^2}-4b\frac{\left(4+2^{-e}\right)^2}{\left(a^2+b^2\right)^3}.$$

For $(a,b,e) \in D_6$, it can be checked that $\frac{\partial g_6}{\partial a}(a,b,e) < 0$ and $\frac{\partial g_6}{\partial b}(a,b,e) < 0$ as follows. Note first that $2^{-\frac{e}{2}} \leqslant a$ implies

$$\frac{4+2^{-e}}{a^2+b^2}2^{-\frac{1+e}{2}} \leqslant \frac{4+2^{-e}}{\sqrt{2}\left(a^2+b^2\right)}a,$$

and that $\sqrt{2} \leqslant b$ implies

$$\frac{4+2^{-e}}{a^2+b^2} \leqslant \frac{4+2^{-e}}{\sqrt{2}\left(a^2+b^2\right)}b.$$

Thus, the same expression can be used as an upper bound for both $\frac{1}{a}\frac{\partial g_6}{\partial a}$ and $\frac{1}{b}\frac{\partial g_6}{\partial b}$. Then, multiplying it by $\frac{(a^2+b^2)^2}{4+2^{-e}}$, it is enough to prove that

$$\frac{\left(a^2+b^2\right)^2\left(2^{-1-e}+1\right)}{2\left(4+2^{-e}\right)} + \frac{a^2+b^2}{\sqrt{2}} < 2\left(2^{-\frac{1+e}{2}}a+b\right)+4\frac{4+2^{-e}}{a^2+b^2}.$$

This last inequality follows from the following three ones:

$$\frac{\left(a^2+b^2\right)^2\left(2^{-1-e}+1\right)}{2\left(4+2^{-e}\right)} + \frac{a^2+b^2}{\sqrt{2}} < \frac{\left(2\sqrt{2}+\left(2+\frac{1}{2}\right)u\right)^2\left(\frac{1}{4}+1\right)}{8}+2+\frac{2+\frac{1}{2}}{\sqrt{2}}u,$$

$$\frac{\left(2\sqrt{2}+\left(2+\frac{1}{2}\right)u\right)^2\left(\frac{1}{4}+1\right)}{8}+2+\frac{2+\frac{1}{2}}{\sqrt{2}}u < 2\sqrt{2}+\frac{16}{2\sqrt{2}+\left(2+\frac{1}{2}\right)u} \quad \text{for } p \geqslant 2,$$

and

$$2\sqrt{2}+\frac{16}{2\sqrt{2}+\left(2+\frac{1}{2}\right)u} < 2\left(2^{-\frac{1+e}{2}}a+b\right)+4\frac{4+2^{-e}}{a^2+b^2}.$$

A.7 Analysis of g_7

In this section, we provide some details about the analysis of g_7 that were omitted in Section 3.6.2.

- Let us first maximize g_7 with respect to b. We have

$$\frac{\partial g_7}{\partial b} = \frac{b}{2}\left(2^{-e}+1\right)+\frac{4+2^{-e}}{a^2+b^2}-2b\frac{\left(2^{-\frac{e}{2}}a+b\right)\left(4+2^{-e}\right)}{\left(a^2+b^2\right)^2}-4b\frac{\left(4+2^{-e}\right)^2}{\left(a^2+b^2\right)^3}.$$

We want to prove that $\frac{\partial}{\partial b}g_7(a,b,e) < 0$ over D_7. Multiplying by $\frac{(a^2+b^2)^2}{(4+2^{-e})b}$ and using the inequality $\frac{1}{b} < 1$, we only need to prove that

$$\frac{\left(a^2+b^2\right)^2\left(2^{-e}+1\right)}{2\left(4+2^{-e}\right)}+a^2+b^2 < 2\left(2^{-\frac{e}{2}}a+b\right)+4\frac{4+2^{-e}}{a^2+b^2}.$$

Since $e \geqslant 2$, we can derive this inequality for $p \geqslant 2$ from the three following ones using the definition of D_7:

$$\frac{\left(a^2+b^2\right)^2\left(2^{-e}+1\right)}{2\left(4+2^{-e}\right)}+a^2+b^2<\frac{\left(2\sqrt{2}+\left(2+\frac{1}{4}\right)u\right)^2\left(\frac{1}{4}+1\right)}{8}+2\sqrt{2}+\left(2+\frac{1}{4}\right)u,$$

$$\frac{\left(2\sqrt{2}+\left(2+\frac{1}{4}\right)u\right)^2\left(\frac{1}{4}+1\right)}{8}+2\sqrt{2}+\left(2+\frac{1}{4}\right)u<2\sqrt{2}+\frac{16}{2\sqrt{2}+\left(2+\frac{1}{4}\right)u},$$

and

$$2\sqrt{2}+\frac{16}{2\sqrt{2}+\left(2+\frac{1}{4}\right)u}<2\left(2^{-\frac{e}{2}}a+b\right)+4\frac{4+2^{-e}}{a^2+b^2}.$$

Therefore, g_7 is decreasing with respect to b, and for all (a,b,e) in D_7, $g_7(a,b,e)\leqslant g_7(a,\sqrt{2},e)$.

- We now maximize $g_7(a,\sqrt{2},e)$ with respect to a. Let us recall that in D_7,

$$a\geqslant a_0+\eta(u)=2^{\frac{e}{2}}-\sqrt{2^e-2+(4+2^{-e})u};$$

and prove that

$$a_0+\eta(u)\geqslant 2^{-1-\frac{e}{2}}.$$

Using the notation $x=2^{-e}$, the inequality $a_0+\eta(u)\geqslant 2^{-1-\frac{e}{2}}$ is equivalent to $(\frac{1}{4}-u)x\geqslant -1+4u$ which holds for $p\geqslant 2$ since $\frac{1}{4}-u\geqslant 0\geqslant -1+4u$.

Moreover, we have

$$\begin{aligned}\frac{(a^2+2)^2}{a(4+2^{-e})}\frac{\partial}{\partial a}g_7(a,\sqrt{2},e) &= \frac{(2^{-e}+1)(a^2+2)^2}{2(4+2^{-e})}+\frac{2^{-\frac{e}{2}}}{a}(a^2+2)\\ &\quad -2\left(2^{-\frac{e}{2}}a+\sqrt{2}\right)-4\frac{4+2^{-e}}{a^2+2},\end{aligned}$$

with $\frac{(a^2+2)^2}{a(4+2^{-e})}>0$ for $a\in I:=\left[2^{-1-\frac{e}{2}},2^{\frac{1-e}{2}}\right]$. For $e\geqslant 2$ and $a\in I$, we have

$$\frac{(a^2+2)^2}{a(4+2^{-e})}\frac{\partial}{\partial a}g_7(a,\sqrt{2},e)<\frac{125}{128}+5-2\sqrt{2}-\frac{32}{5}<0.$$

As a consequence, $g_7(a,\sqrt{2},e)$ is decreasing with respect to a over I, and since $a_0+\eta(u)\in I$, the maximum of $g_7(a,\sqrt{2},e)$ for $a\in\left[a_0+\eta(u),2^{\frac{1-e}{2}}\right]$ is $g_7(a_0+\eta(u),\sqrt{2},e)$.

Thus, for (a,b,e) in D_7, we have $g_7(a,b,e)\leqslant g_7(a,\sqrt{2},e)\leqslant g_7(a_0+\eta(u),\sqrt{2},e)$.

- Let us prove that $g_7(a_0+\eta(u),\sqrt{2},e)\leqslant g_7(a_0,\sqrt{2},e)+20u$. For this purpose, we first show that $|\eta(u)|<2u$. We have

$$|\eta(u)|=\sqrt{2^e-2+(4+2^{-e})u}-\sqrt{2^e-2}$$

and a short calculation shows that $|\eta(u)| < 2u$. It can also be checked that $a_0 < 2^{\frac{1-e}{2}}$ using again $x = 2^{-e}$ and a short calculation. Since $e \geqslant 2$, this implies $a_0 \leqslant \frac{\sqrt{2}}{2} < 1$. Let us now consider

$$\lambda_0(u) = \frac{1}{(a_0 + \eta(u))^2 + 2}.$$

We have

$$\lambda_0(u) = \frac{1}{a_0^2 + 2} - \frac{2a_0 + \eta(u)}{(a_0^2 + 2)((a_0 + \eta(u))^2 + 2)}\eta(u),$$

and using $-2u < \eta(u) \leqslant 0$, we deduce

$$\lambda_0(u) < \frac{1}{a_0^2 + 2} + a_0 u.$$

Moreover, we have

$$\begin{aligned}\lambda_0(u)^2 &= \left(\frac{1}{a_0^2 + 2}\right)^2 - \frac{4a_0}{(a_0^2 + 2)^2((a_0 + \eta(u))^2 + 2)}\eta(u) \\ &\quad + \frac{(2a_0 + \eta(u))^2 - 2((a_0 + \eta(u))^2 + 2)}{(a_0^2 + 2)^2((a_0 + \eta(u))^2 + 2)^2}\eta(u)^2,\end{aligned}$$

and using both $-2u < \eta(u) \leqslant 0$ and $a_0 < 1$, we also deduce

$$\lambda_0(u)^2 < \left(\frac{1}{a_0^2 + 2}\right)^2 + a_0 u.$$

From the definition of g_7, using $0 < a_0 + \eta(u) < a_0$ and the previous upper bounds on $\lambda_0(u)$ and $\lambda_0(u)^2$, we obtain

$$\begin{aligned}g_7(a_0 + \eta(u), \sqrt{2}, e) &< \frac{\left(a_0^2 + 2\right)\left(2^{-e} + 1\right)}{4} + \left(4 + 2^{-e}\right)\left(2^{-\frac{e}{2}} a_0 + \sqrt{2}\right)\left(\frac{1}{a_0^2 + 2} + a_0 u\right) \\ &\quad + (4 + 2^{-e})^2\left(\frac{1}{(a_0^2 + 2)^2} + a_0 u\right),\end{aligned}$$

and $g_7(a_0+\eta(u), \sqrt{2}, e) < g_7(a_0, \sqrt{2}, e)+(4+2^{-e})\left(2^{-\frac{e}{2}} a_0 + \sqrt{2} + 4 + 2^{-e}\right) a_0 u$. The inequality $g_7(a_0 + \eta(u), \sqrt{2}, e) < g_7(a_0, \sqrt{2}, e) + 20u$ then follows from $e \geqslant 2$ and $a_0 \leqslant \frac{\sqrt{2}}{2}$.

- Finally, we check that the function h_7 is increasing over $\left[\sqrt{2}/2, \sqrt{1 - 2^{-5}}\right]$ and decreasing over $\left[\sqrt{1 - 2^{-7}}, 1\right]$. We have $h_7(y) = \frac{H(y)}{64(y+1)}$ with

$$\begin{aligned}H(y) &= y^7 + 3y^6 - 7y^5 - (8\sqrt{2} + 45)y^4 - (16\sqrt{2} + 53)y^3 \\ &\quad + (64\sqrt{2} + 113)y^2 + (144\sqrt{2} + 315)y + 72\sqrt{2} + 249.\end{aligned}$$

Hence $h_7'(y) = \frac{P(y)}{32\,(y+1)^2}$ where P is the polynomial

$$\begin{aligned} P(y) &= 3y^7 + 11y^6 - 5y^5 - (12\sqrt{2} + 85)y^4 - (32\sqrt{2} + 143)y^3 \\ &\quad - (23 - 8\sqrt{2})y^2 + (64\sqrt{2} + 113)y + 36\sqrt{2} + 33. \end{aligned}$$

This polynomial has 0 or 2 positive roots according to Descartes' rule of signs (there are two sign changes in the sequence of coefficients). Moreover,

$$\begin{aligned} P(y+1) &= 3y^7 + 32y^6 + 124y^5 + (160 - 12\sqrt{2})y^4 - (208 + 80\sqrt{2})y^3 \\ &\quad - (784 + 160\sqrt{2})y^2 - (640 + 64\sqrt{2})y - 96 + 64\sqrt{2}, \end{aligned}$$

with only one sign change, so there is exactly one root of P greater than 1 and at most one root of P in $\left[\sqrt{2}/2, 1\right]$. Since $P(\sqrt{1 - 2^{-5}}) > 0$ and $P(\sqrt{1 - 2^{-7}}) < 0$, we deduce that $P(y)$ is positive for $y \in \left[\sqrt{2}/2, \sqrt{1 - 2^{-5}}\right]$, and negative for $y \in \left[\sqrt{1 - 2^{-7}}, 1\right]$, which implies that h_7 is increasing over the former interval, and decreasing over the latter.

References

1. Baudin, M.: Error bounds of complex arithmetic. Available at http://forge.scilab.org/upload/compdiv/files/complexerrorbounds_v0.2.pdf (2011)
2. Baudin, M., Smith, R.L.: A robust complex division in Scilab. Available at http://arxiv.org/abs/1210.4539 (2012)
3. Brent, R., Percival, C., Zimmermann, P.: Error bounds on complex floating-point multiplication. Math. Comput. **76**(259), 1469–1481 (2007)
4. Champagne, W.P.: On finding roots of polynomials by hook or by crook. Master's thesis, University of Texas (1964)
5. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, second edn. SIAM, Philadelphia, PA, USA (2002)
6. IEEE Computer Society: IEEE Standard for Floating-Point Arithmetic. IEEE Standard 754–2008 (2008). http://ieeexplore.ieee.org/servlet/opac?punumber=4610933
7. Knuth, D.E.: The Art of Computer Programming, Volume 2, Seminumerical Algorithms, third edn. Addison-Wesley, Reading, MA, USA (1998)
8. Priest, D.M.: Efficient scaling for complex division. ACM Trans. Math. Softw. **30**(4), 389–401 (2004)
9. Rump, S.M., Ogita, T., Oishi, S.: Accurate floating-point summation, Part I: Faithful rounding. SIAM J. Sci. Comput. **31**(1), 189–224 (2008)
10. Smith, R.L.: Algorithm 116: Complex division. Commun. ACM **5**(8), 435 (1962)
11. Stewart, G.W.: A note on complex division. ACM Trans. Math. Softw. **11**(3), 238–241 (1985)
12. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford University Press (1965)
13. Ziv, A.: Sharp ULP rounding error bound for the hypotenuse function. Math. Comput. **68**(227), 1143–1148 (1999)