Sharp error bounds for complex floating-point inversion

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

Sharp error bounds for complex floating-point inversion

Claude-Pierre Jeannerod Nicolas Louvet
Jean-Michel Muller Antoine Plet

September 8, 2015

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 in precision p, and we assume that the elementary arithmetic operations ($+$, $-$, $\times$, $/$) are rounded to nearest, so that the roundoff unit is $u = 2^{-p}$. 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 reasonably sharp when using one of the basic IEEE 754 binary formats with an odd precision ($p = 53, 113$). This componentwise bound obviously leads to the same bound $3u$ for the normwise relative error. However we prove that the significantly smaller bound $2.707131u$ holds (assuming $p \geqslant 24$) for the normwise relative error, and we illustrate the sharpness of this bound using numerical examples for the basic IEEE 754 binary formats ($p = 24, 53, 113$).

keywords: floating-point arithmetic, numerical error, complex inversion, error analysis, roundings.

1 Introduction

This paper deals with the accuracy of the inversion of a 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 can be expressed as

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

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

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

for the real part, and with a similar expression for the imaginary part $\widehat{I}$, which straightforwardly leads to Algorithm 1 below.

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

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

We provide an accuracy analysis of this algorithm, for both the componentwise relative error (assuming $a \neq 0$ and $b \neq 0$) defined by $\mathrm{E_C} = \max\big(|R - \widehat{R}|/|R|, |I - \widehat{I}|/|I|\big)$, and the normwise relative error $\mathrm{E_N} = |z - \widehat{z}|/|z|$ (assuming $z \neq 0$).

Of course, complex inversion is a particular case of complex division. However, the quotient computed by the classic division algorithm can be highly inaccurate in the componentwise sense (see for example [7, §1]), while the componentwise relative error generated by Algorithm 1 can easily be bounded by $3u + O(u^2)$, where $u = 2^{-p}$ is the unit roundoff. In fact, we prove in Section 2 that the $\mathcal{O}\big(u^2\big)$ term can be removed, leading to the simpler bound $3u$ (assuming $p \geqslant 4$). We also show that this bound is asymptotically optimal (as $p \to \infty$) when p is even, by providing floating-point numbers a and b parametrized by p, and for which $\mathrm{E_C}$ is at least $3u - \frac{31}{2}u^{\frac{3}{2}} + \mathcal{O}\big(u^2\big)$. When p is odd, we give numerical examples to show that the bound $3u$ is reasonably sharp, especially for the corresponding basic IEEE 754 binary formats ($p = 53, 113$).

Normwise relative accuracy analyses of the classic complex division algorithm can be found for example in [4],[13]. To our knowledge, the smallest bound on the normwise relative error for complex division is $(3+\sqrt{5})u+\mathcal{O}\big(u^2\big)$: as noted in [1, §3.5], this bound can be derived from the bound $\sqrt{5}u$ from [3] on the normwise relative error for the classic complex multiplication algorithm (note that $3 + \sqrt{5} \approx 5.24$). In the case of complex inversion with Algorithm 1, the bound $3u + \mathcal{O}\big(u^2\big)$ can be found in [4, p. 30], and a direct application of our componentwise bound $3u$ obviously leads to $\mathrm{E_N} \leqslant 3u$. However, we prove in Section 3 the significantly smaller bound $\mathrm{E_N} < \gamma u + 9u^2$ for the normwise error of Algorithm 1 (assuming $p \geqslant 10$), with γ a 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 [14], 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 quite sharp for the basic IEEE 754 formats ($p = 24, 53, 113$).

Several authors [11, 12, 9, 2] have suggested ways of avoiding spurious overflows and underflows in complex division. As already noticed, we do not deal with this problem here, and we only focus 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. Section 4 concludes the paper. 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 attribute defined in the IEEE 754 standard [6] can be used.

Throughout the paper, the relative error due to rounding is bounded as follows [8, p. 232]: for any real 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]).

We use the notation $\mathrm{ufp}(t)$ (*unit in the first place*, introduced in [10]) to denote the weight of the most significant digit of t. More precisely, $\mathrm{ufp}(0) = 0$, and if $t \neq 0$ then $\mathrm{ufp}(t)$ is the unique integer power of 2 such that $1 \leqslant \frac{|t|}{\mathrm{ufp}(t)} < 2$. 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

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

2 Componentwise error bound

In this section, we analyze $\mathrm{E_C} = \max\left(|R-\widehat{R}|/|R|, |I-\widehat{I}|/|I|\right)$ for Algorithm 1. Repeated applications of the bound $u/(1+u)$ in (3) give immediately $\mathrm{E_C} \leqslant 3u+\mathcal{O}(u^2)$. We show below that the $\mathcal{O}(u^2)$ term can in fact be removed, leading to the simpler bound $3u$.

To do this, we show that if $p \neq 3$ then the relative bound $u/(1+u)$ in (3) can be replaced by $u/(1+3u)$ when evaluating a square $\mathrm{RN}(a^2)$ 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 $\mathrm{RN}(a^2) = a^2(1+\epsilon)$ with $|\epsilon| \leqslant u/(1+3u)$.*

Proof. We can assume that $1 \leqslant a < 2$. If $a = 1$ then $\mathrm{RN}(a^2) = a^2$ and the result is clear. If $1 < a < \sqrt{2}$ then it follows from a being a floating-point number and $p \geqslant 4$ 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 $\mathrm{RN}(a^2) = 2$ 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 $\mathrm{RN}(a^2) = 2 + 4u$ 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 $\mathrm{RN}(a^2) = 2 + 4u$ 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 $|\epsilon| \leqslant 2u/a^2 \leqslant 2u/(2+6u) = u/(1+3u)$.

□

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

$$s_a = a^2(1+\epsilon_1), \qquad s_b = b^2(1+\epsilon_2), \qquad s = (s_a+s_b)(1+\epsilon_3), \qquad \widehat{R} = \frac{a}{s}(1+\epsilon_4)$$

with $|\epsilon_1|, |\epsilon_2| \leqslant u/(1+3u)$ and $|\epsilon_3|, |\epsilon_4| \leqslant u/(1+u)$. Hence

$$\widehat{R} = \frac{a}{a^2(1+\epsilon_1) + b^2(1+\epsilon_2)} \cdot \frac{1+\epsilon_4}{1+\epsilon_3}$$

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. □

We conclude this section by showing that the componentwise bound $\mathrm{E_C} \leqslant 3u$ is essentially 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 $p \geqslant 10$, 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

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

From this we deduce

$$\frac{a}{s} = 2^{-\frac{3p}{2}+1} + 2^{-2p} - 2^{-\frac{5p}{2}+1} - 15 \cdot 2^{-3p+2} + \mathcal{O}\big(2^{-\frac{7p}{2}}\big),$$

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

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

it can be checked that

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

hence $\widehat{R} = \mathrm{RN}\big(\frac{a}{s}\big) = \tau$, which leads to

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

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 did not find an input set parametrized by the precision p to prove the asymptotic optimality of the error bound $3u$. However, the sharpness of the bound is illustrated in this case by the numerical examples provided in Table 1.

p	example
15	$a = 16732$ $b = 23252{\cdot}2^3$ $\lvert\widehat{R} - R\rvert/(u\lvert R\rvert) = 2.93047\ldots$
17	$a = 66078$ $b = 93014{\cdot}2^8$ $\lvert\widehat{R} - R\rvert/(u\lvert R\rvert) = 2.96359\ldots$
19	$a = 131435$ $b = 370969{\cdot}2^8$ $\lvert\widehat{R} - R\rvert/(u\lvert R\rvert) = 2.98509\ldots$
53	$a = 4508053433127332$ $b = 6369149602646415{\cdot}2^{16}$ $\lvert\widehat{R} - R\rvert/(u\lvert R\rvert) = 2.97894\ldots$
113	$a = 5192393427440123027423416459819356$ $b = 7343016638055329519853569740503421{\cdot}2^{16}$ $\lvert\widehat{R} - R\rvert/(u\lvert R\rvert) = 2.97647\ldots$

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

3 Normwise error bound

In this section, we are interested in the relative normwise 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 $\mathrm{E_N} \leqslant 3u$. But we can do better and improve significantly this bound keeping track of the correlation between the errors. In this section, we prove the following result.

Theorem 2. *For $p \geqslant 10$, the normwise error in the approximate inverse $\widehat{z}$ computed by 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 enlargement: 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 quite sharp for the basic IEEE 754 formats ($p = 24, 53, 113$).

p	example
24	$a = 11863283$ $b = 11865457 \cdot 2^{12}$ $\lvert\widehat{z} - z\rvert/(u\lvert z\rvert) = 2.69090\ldots$
53	$a = 4503599709991314$ $b = 6369051770002436 \cdot 2^{26}$ $\lvert\widehat{z} - z\rvert/(u\lvert z\rvert) = 2.70679\ldots$
113	$a = 2^{112}$ $b = 7343016637207171132572330391109909 \cdot 2^{56}$ $\lvert\widehat{z} - z\rvert/(u\lvert z\rvert) = 2.70559\ldots$

Table 2: Examples with a normwise relative error close to γ.

3.1 Preliminaries

The first step in the error analysis of Algorithm 1 is to reduce the input domain. Since the RN function 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 non negative. Moreover, if $a = 0$, then a simple analysis, based on (3), leads to the upper bound $2u$ for $\mathrm{E_N}$. Then, swapping the inputs a and b does not affect the relative error, so we can assume that $0 < a \leqslant b$. At last, multiplying or dividing by 2 both a

and b does not affect either the relative error, and we can reduce the analysis to the case $1 \leqslant b < 2$. From the definition of the ufp function and this input reduction, we know that $\mathrm{ufp}\big(b^2\big) \in \{1, 2\}$ and $\mathrm{ufp}(s_a + s_b) \in \{1, 2, 4\}$.

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

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

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 respectively the absolute and relative errors in the evaluation of $a^2 + b^2$. In the rest of the section, e denotes the integer such that $\mathrm{ufp}\big(a^2\big) = 2^{-e}$.

With these notations, we have:

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

and a similar expression holds for the imaginary part, hence

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

Given a, b and δ, this function is maximal when $\delta\delta_R \leqslant 0$ and $\delta\delta_I \leqslant 0$ with $|\delta_R|$ and $|\delta_I|$ maximal, that is $|\delta_R| = \mathrm{ufp}\big(\frac{a}{s}\big)$ and $|\delta_I| = \mathrm{ufp}\big(\frac{b}{s}\big)$. As a consequence, we have

$$\begin{aligned}
\frac{\mathrm{E_N^2}}{u^2} \leqslant \left(a^2 + b^2\right)&\left(\mathrm{ufp}\big(\tfrac{a}{s}\big)^2 + \mathrm{ufp}\big(\tfrac{b}{s}\big)^2\right)\\
&+ 2\frac{|\delta|\left(\mathrm{ufp}\big(\frac{a}{s}\big)a + \mathrm{ufp}\big(\frac{b}{s}\big)b\right)}{a^2 + b^2 - |\delta|u} + \left(\frac{\delta}{a^2 + b^2 - |\delta|u}\right)^2.
\end{aligned}$$

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

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

with

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

and

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

From (4), we have

$$\operatorname{ufp}\big(\tfrac{a}{s}\big)a + \operatorname{ufp}\big(\tfrac{b}{s}\big)b \leqslant \frac{a^2+b^2}{s} \leqslant \frac{a^2+b^2}{a^2+b^2-|\delta|u} = \frac{1}{1-\epsilon u},$$

and we know from [3] that $\epsilon \leqslant 2$, so $f_3(a,b) \leqslant \frac{2\epsilon^2(1+\epsilon)}{(1-\epsilon u)^2} < 25$ 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 $\operatorname{ufp}(t) \leqslant |t|$, 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),

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

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 $\operatorname{ufp}(1+s_a) = 1$ 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 $\mathrm{ufp}\big(b^2\big) = 1$ then $\mathrm{ufp}(1 + s_b) = 2$ and $\epsilon \leqslant \frac{3}{2}$. If $\mathrm{ufp}\big(b^2\big) = 2$ 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, greater than or equal to f_2, with three parameters: a, b and e. This function g does not involve floating-point operations anymore and can be seen as a continuous and derivable function over real inputs. We then look for an upper bound for 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 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 upper bound (8) for f_2. 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 $\gamma^2 + 20u$ 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	$=2^{-2-\frac{e}{2}}$	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 Appendix A.

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

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

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

$$\mathrm{ufp}\big(\tfrac{b}{s}\big) \leqslant \tfrac{1}{2}.$$

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

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

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

$$g_1(a,b,e) := \big(a^2+b^2\big)\big(2^{-e}+\tfrac{1}{4}\big) + 2\frac{(2+2^{-e})\big(a2^{-\frac{e}{2}}+\frac{b}{2}\big)}{a^2+b^2} + \left(\frac{2+2^{-e}}{a^2+b^2}\right)^2.$$

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 $a^2+b^2 \leqslant s_a+\mathrm{ufp}\big(a^2\big)u+s_b+\mathrm{ufp}\big(b^2\big)u < \sqrt{2}+\frac{5}{4}u$ which concludes the domain analysis: we are looking for an upper bound for g_1 over the domain

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

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

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

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

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

and a small computation using the definition of D_1 shows that it is true for all $(a,b,e) \in D_1$.

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 $g_1(a,b,e) \leqslant g_1(2^{-\frac{e}{2}},1,e) =: h_1(x)$, with $x = 2^{-e}$ and

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

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

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

is clearly positive, hence we deduce $f_2(a,b) \leqslant h_1\left(\frac{1}{4}\right) = 6.565$.

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

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

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

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

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

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

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

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

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

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

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 §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 $e \geqslant 0$

In this case, using the inequality $\mathrm{ufp}\left(\frac{a}{s}\right) \leqslant 2^{-2-\frac{e}{2}}$ in (8), we have $f_2(a,b) \leqslant g_3(a,b,e)$ with

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

The domain over which we bound g_3 is

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

First, it can be checked that the partial derivative of g_3 with respect to b is negative over D_3 (details are in §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

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

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 non-negative 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

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

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 $f_2(a,b) \leqslant h_3(1) = \left(\frac{7}{4} + \frac{\sqrt{2}}{2}\right)^2$.

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

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

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

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

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

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

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

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

We have $a^2 + b^2 \leqslant s_a + s_b + \left(\mathrm{ufp}\left(a^2\right) + \mathrm{ufp}\left(b^2\right)\right)u < 2\sqrt{2} + 2u$ and $1 < b < \sqrt{2}$, hence we can restrict the analysis to the domain

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

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

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

We can next compute the derivative of $g_4(a,\sqrt{2-a^2},e)$ with respect to a (see §A.4) and check that the maximum is obtained 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

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

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

• **Subsubcase** $e = 0$**:** According to (13), we assume that $1 < a$, so that $\mathrm{ufp}(b^2) = \mathrm{ufp}(a^2) = 1$. 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 derivatives, 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,

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

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

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

Using these bounds on $\mathrm{ufp}\left(\frac{a}{s}\right)$ and $\mathrm{ufp}\left(\frac{b}{s}\right)$ in (8) we get $f_2(a,b) \leqslant g_5(a,b)$ with

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

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 §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 $g_5(a,b) \leqslant g_5(1,1) = \left(\frac{5}{2}\right)^2$.

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

In this paragraph, $a^2 < 1$ (otherwise we would have $s_a + s_b \geqslant 2 + 1$ while from (15) we have $2\sqrt{2} < s_a + s_b$), 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 that $f_2(a,b) \leqslant g_6(a,b,e)$ with

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

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

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

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

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

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

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

We can compute the partial derivative of g_7 with respect to b and prove that it is negative over the domain

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

Therefore, we know that $g_7(a,b,e) \leqslant g_7(a,\sqrt{2},e)$. Moreover, for $a \geqslant 2^{-1-\frac{e}{2}}$, we have $\frac{\partial}{\partial a} g_7(a,\sqrt{2},e) < 0$, so that $g_7(a,\sqrt{2},e)$ is decreasing with respect to a and is maximal at the minimal value of a which we will now determine. The lower bound $2^{-\frac{e}{2}}$ for a does not lead to a sufficiently tight bound for f_2: to get a better bound, we exploit further the hypothesis $\mathrm{ufp}\left(\frac{a}{s}\right) = 2^{-1-\frac{e}{2}}$. From this assumption, we deduce $s2^{-1-\frac{e}{2}} \leqslant a$, that is $a^2 - 2^{1+\frac{e}{2}}a + b^2 + \delta u \leqslant 0$, which implies

$$a \geqslant 2^{-\frac{e}{2}} \frac{2 - (4 + 2^{-e})u}{1 + \sqrt{1 - 2^{-e}(2 - (4+2^{-e})u)}} = a_0 + \eta(u)$$

with $a_0 := 2^{-\frac{e}{2}} \frac{2}{1+\sqrt{1-2^{1-e}}}$, $\eta(u) < 0$ and $\eta(u) \in \mathcal{O}(u)$. Then, it can be proved that $g_7(a,\sqrt{2},e) \leqslant g_7(a_0,\sqrt{2},e) + 20u$ (the details are provided in §A.7).

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 fraction over y. The variable y belongs to $\left[\sqrt{2}/2, 1\right]$, and $h_7'(y) = \frac{P(y)}{32\,(y+1)^2}$ where

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

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 $\mathrm{E_N}$ (details of the proof can be found in §A.7), which concludes the proof of Theorem 2.

4 Conclusion

We showed the componentwise relative error bound $3u$ for the complex inversion algorithm, and we proved that this bound is asymptotically optimal (as $p \to \infty$) when the precision p is even, and reasonably sharp when p is odd. We also proved the bound $\gamma u + 9u^2$, with $\gamma \in (2.70712, 2.70713)$ for the normwise relative error, and we have illustrated the sharpness of this bound using numerical examples for the basic IEEE 754 binary formats.

Let us conclude with a remark concerning floating-point division. The classic complex division algorithm for computing an approximate $\widehat{z} = \widehat{R} + i\,\widehat{I}$ of $(a+ib)/(c+id)$ in floating-point arithmetic is given by

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

with a similar formula for the imaginary part $\widehat{I}$. As mentioned in [1, §3.6], the smallest known upper bound for the normwise relative error generated using (17) is $(3+\sqrt{5})u + \mathcal{O}(u^2)$ (let us recall $3+\sqrt{5} \approx 5.24$). With $a+ib = 1$, formula (17) reduces to the classic algorithm for the complex inversion of $c+id$. However, in precision $p = 11$, dividing for instance $a + ib = 1575 + i\,1419$ by $c + id = 1457 + i\,1480$ leads to $|\widehat{z} - z|/(u|z|) = 4.67973\ldots$ This example suffices to show that the normwise relative error bound for complex division cannot be reduced to a bound of the form $\gamma u + \mathcal{O}(u^2)$ as in the particular case of complex inversion, but it is not very informative concerning the sharpness of the bound $(3+\sqrt{5})u + \mathcal{O}(u^2)$. Thus, in a future work we plan to investigate further the normwise accuracy of complex division.

References

[1] Baudin, M.: Error bounds of complex arithmetic (2011). Available at `http://forge.scilab.org/upload/compdiv/files/complexerrorbounds_v0.2.pdf`

[2] Baudin, M., Smith, R.L.: A robust complex division in Scilab (2012). Available at `http://arxiv.org/abs/1210.4539`

[3] Brent, R., Percival, C., Zimmermann, P.: Error bounds on complex floating-point multiplication. Mathematics of Computation **76**, 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). Available at `http://ieeexplore.ieee.org/servlet/opac?punumber=4610933`

[7] Jeannerod, C.P., Louvet, N., Muller, J.M.: On the componentwise accuracy of complex floating-point division with an FMA. In: 21st IEEE Symposium on Computer Arithmetic, ARITH 2013, Austin, TX, USA, April 7-10, 2013, pp. 83–90 (2013)

[8] Knuth, D.E.: The Art of Computer Programming, Volume 2, Seminumerical Algorithms, third edn. Addison-Wesley, Reading, MA, USA (1998)

[9] Priest, D.M.: Efficient scaling for complex division. ACM Transactions on Mathematical Software **30**(4) (2004)

[10] Rump, S.M., Ogita, T., Oishi, S.: Accurate floating-point summation, Part I: Faithful rounding. SIAM Journal on Scientific Computing **31**(1), 189–224 (2008)

[11] Smith, R.L.: Algorithm 116: Complex division. Communications of the ACM **5**(8), 435 (1962)

[12] Stewart, G.W.: A note on complex division. ACM Transactions on Mathematical Software **11**(3), 238–241 (1985)

[13] Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Oxford University Press (1965)

[14] Ziv, A.: Sharp ULP rounding error bound for the hypotenuse function. Mathematics of Computation **68**(227), 1143–1148 (1999)

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 $p \geqslant 10$ 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 $s_a = \mathrm{RN}\big(a^2\big)$:

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

Hence $s_a = \widetilde{s_a}$.

• Computation of $s_b = \mathrm{RN}\big(b^2\big)$:

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

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

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

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

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

Since $a > \sqrt{2}$, $b > \sqrt{2}$ and $a^2+b^2 < 4\sqrt{2}+4u$, it is enough to check that

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

which holds as soon as $p \geqslant 2$.

A.3 Partial derivatives 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

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

and we check that

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

Since $1 \leqslant b$, it is enough to prove:

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

This follows from the inequalities

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

A.4 Partial derivatives of g_4

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

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

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

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

This inequality can be derived from the following ones:

$$\begin{aligned}
\frac{\left(a^2+b^2\right)^2\left(2^{-e}+1\right)}{2\left(3+2^{-e}\right)} + \frac{a^2+b^2}{b} &< \frac{2\left(2\sqrt{2}+2u\right)^2}{6} + 2\sqrt{2} + 2u \\
\frac{\left(2\sqrt{2}+2u\right)^2}{3} + 2\sqrt{2} + 2u &< \frac{6}{\sqrt{2}+u} + 2 \text{ which holds when } p \geqslant 4 \\
\frac{12}{2\sqrt{2}+2u} + 2 &< 4\frac{3+2^{-e}}{a^2+b^2} + 2\left(2^{-\frac{e}{2}}a+b\right).
\end{aligned}$$

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

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

Note that $e \geqslant 1$ implies $a < 1$, and $\sqrt{2-a^2} > 1 > a$.

A.5 Partial derivatives of g_5

We have

$$\frac{\partial g_5}{\partial b} = \frac{1}{4}b - \frac{64}{\left(a^2+b^2\right)^3}b + \frac{2}{a^2+b^2} - \frac{4\left(a+b\right)}{\left(a^2+b^2\right)^2}b,$$

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

$$\begin{aligned}
\frac{\left(a^2+b^2\right)^2}{4} + \frac{2}{b}\left(a^2+b^2\right) &< \frac{\left(2\sqrt{2}+2u\right)^2}{4} + 2\left(2\sqrt{2}+2u\right), \\
\frac{\left(2\sqrt{2}+2u\right)^2}{4} + 2\left(2\sqrt{2}+2u\right) &< \frac{64}{2\sqrt{2}+2u} + 8 \text{ when } p \geqslant 2, \\
\frac{64}{2\sqrt{2}+2u} + 8 &< \frac{64}{a^2+b^2} + 4\left(a+b\right).
\end{aligned}$$

A.6 Partial derivatives of g_6

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

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

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

It can be checked that $\frac{\partial}{\partial a}g_6(a,b,e) < 0$ and $\frac{\partial}{\partial b}g_6(a,b,e) < 0$ using

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

and

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

Thus, we only need to prove that

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

This last inequality can be derived from the three following ones:

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

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

and

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

A.7 Analysis of g_7

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

• We first maximize g_7 with respect to b. We have

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

We want to prove that $\frac{\partial}{\partial b} g_7(a,b,e) < 0$ over D_7. Since $\frac{1}{b} < 1$, we only need to prove that

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

We can derive this inequality for $p \geqslant 2$ from the three following ones:

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

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

and

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

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

• We now maximize $g_7(a,\sqrt{2},e)$ with respect to a. We compute

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

with $\frac{(a^2+2)^2}{a(4+2^{-e})} > 0$. On the domain $e \geqslant 2$, $0 < 2^{-1-\frac{e}{2}} \leqslant a \leqslant 2^{\frac{1-e}{2}}$, we have

$$\frac{(a^2+2)^2}{a(4+2^{-e})}\frac{\partial}{\partial a} g_7(a,\sqrt{2},e) < \frac{125}{128} - \frac{32}{5} + \frac{5}{2}\frac{2^{-\frac{e}{2}}}{a} - 2\sqrt{2} < 0 \text{ since } a \geqslant 2^{-1-\frac{e}{2}}.$$

We deduce that $g_7(a,\sqrt{2},e)$ is decreasing with respect to a over $\left[2^{-1-\frac{e}{2}}, 2^{\frac{1-e}{2}}\right]$. Let us recall that $a \geqslant a_0 + \eta(u) = 2^{-\frac{e}{2}}\frac{2-(4+2^{-e})u}{1+\sqrt{1-2^{-e}(2-(4+2^{-e})u)}}$. It can be checked that $2^{-1-\frac{e}{2}} \leqslant a_0 + \eta(u)$ using the following equivalent inequalities ($x = 2^{-e}$ and we know that $0 < x \leqslant \frac{1}{4}$) :

$$\begin{aligned}
\frac{\sqrt{x}}{2} &\leqslant \frac{1}{\sqrt{x}}\left(1 - \sqrt{1-x\left(2-(4+x)u\right)}\right)\\
\sqrt{1-x\left(2-(4+x)u\right)} &\leqslant 1 - \frac{x}{2}\\
1-x\left(2-(4+x)u\right) &\leqslant 1 - x + \frac{x^2}{4}\\
0 &\leqslant 1 + \frac{x}{4} - (4+x)u\\
0 &\leqslant (1-4u)\left(1+\frac{x}{4}\right) \text{ which holds for } p \geqslant 2\,.
\end{aligned}$$

We deduce that $g_7(a, \sqrt{2}, e) \leqslant g_7(a_0 + \eta(u), \sqrt{2}, e)$.

• 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 $|\eta(u)| < 2u$:

$$\begin{aligned}
|\eta(u)| &= 2^{\frac{e}{2}} \left(\sqrt{1 - 2^{-e}(2 - (4 + 2^{-e})u)} - \sqrt{1 - 2^{1-e}} \right) \\
&= \frac{2^{\frac{e}{2}}}{\sqrt{1 - 2^{-e}(2 - (4 + 2^{-e})u)} + \sqrt{1 - 2^{1-e}}} \left(2^{1-e} - 2^{-e}(2 - (4 + 2^{-e})u \right) \\
&= \frac{2^{-\frac{e}{2}}}{\sqrt{1 - 2^{-e}(2 - (4 + 2^{-e})u)} + \sqrt{1 - 2^{1-e}}} (4 + 2^{-e})u \\
&\leqslant \frac{2^{-1}(4 + 2^{-2})}{\sqrt{2}} u \text{ since } e \geqslant 2 \\
&< 2u.
\end{aligned}$$

It can also be checked that $a_0 < 2^{\frac{1-e}{2}}$ using the following equivalent inequalities:

$$\begin{aligned}
\frac{1}{\sqrt{x}} \left(1 - \sqrt{1 - 2x}\right) &< \sqrt{2}\sqrt{x} \\
1 - \sqrt{2}x &< \sqrt{1 - 2x} \\
1 - 2\sqrt{2}x + 2x^2 &< 1 - 2x \\
0 &< \sqrt{2} - 1 - x, \text{ which holds since } x \leqslant \tfrac{1}{4}.
\end{aligned}$$

Since $e \geqslant 2$, this implies $a_0 \leqslant \frac{\sqrt{2}}{2} < 1$. Let us now consider

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

We have

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

and using $|\eta(u)| < 2u$, we deduce

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

Moreover, we have

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

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

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

As a consequence, from the definition of g_7, and the previous upper bounds on $\lambda_0(u)$ and $\lambda_0(u)^2$, we obtain

$$g_7(a_0+\eta(u),\sqrt{2},e) < \frac{\left(a_0^2+2\right)\left(2^{-e}+1\right)}{4} + (4+2^{-e})^2\left(\frac{1}{(a_0^2+2)^2}+a_0u\right) + \left(4+2^{-e}\right)\left(2^{-\frac{e}{2}}a_0+\sqrt{2}\right)\left(\frac{1}{a_0^2+2}+a_0u\right),$$

and $g_7(a_0+\eta(u),\sqrt{2},e) < g_7(a_0,\sqrt{2},e)+(4+2^{-e})\left(4+2^{-e}+2^{-\frac{e}{2}}a_0+\sqrt{2}\right)a_0u$. 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 $a_0 \leqslant \frac{\sqrt{2}}{2}$.

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

$$H(y) = y^7+3y^6-7y^5-(8\sqrt{2}+45)y^4-(16\sqrt{2}+53)y^3 + (62\sqrt{2}+113)y^2+(144\sqrt{2}+315)y+72\sqrt{2}+249.$$

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

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

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,

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

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.