INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Error Bounds on Complex Floating-Point Multiplication

Richard Brent — Colin Percival — Paul Zimmermann

N° 6068

December 2006

Thème SYM

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--6068--FR+ENG

Error Bounds on Complex Floating-Point Multiplication

Richard Brent, Colin Percival, Paul Zimmermann

Thème SYM — Systèmes symboliques
Projet Cacao

Rapport de recherche n° 6068 — December 2006 — 22 pages

Abstract: Given floating-point arithmetic with t-digit base-β significands in which all arithmetic operations are performed as if calculated to infinite precision and rounded to a nearest representable value, we prove that the product of complex values z_0 and z_1 can be computed with maximum absolute error $|z_0|\,|z_1|\,\frac{1}{2}\beta^{1-t}\sqrt{5}$. In particular, this provides relative error bounds of $2^{-24}\sqrt{5}$ and $2^{-53}\sqrt{5}$ for IEEE 754 single and double precision arithmetic respectively, provided that overflow, underflow, and denormals do not occur.

We also provide the numerical worst cases for IEEE 754 single and double precision arithmetic. Finally, we consider generic worst cases and briefly discuss Karatsuba multiplication.

Key-words: IEEE 754, floating-point number, complex multiplication, roundoff error, error analysis

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Bornes d'erreur pour la multiplication de nombres flottants complexes

Résumé : On considère une arithmétique flottante de t chiffres de précision en base β, où tous les calculs sont effectués avec arrondi au plus proche. Nous montrons que le produit de deux nombres complexes z_0 et z_1 peut être calculé avec une erreur absolue d'au plus $|z_0|\,|z_1|\,\frac{1}{2}\beta^{1-t}\sqrt{5}$. Ceci fournit des bornes de $2^{-24}\sqrt{5}$ et $2^{-53}\sqrt{5}$ respectivement pour l'erreur relative dans les formats simple et double précision du standard IEEE 754, en supposant qu'aucun dépassement de capacité ni nombre dénormalisé n'intervient.

Nous donnons également les pires cas pour les formats simple et double précision du standard IEEE 754. Enfin, nous considérons des pires cas génériques, et nous évoquons brièvement la multiplication de Karatsuba.

Mots-clés : IEEE 754, nombre flottant, multiplication complexe, erreur d'arrondi, analyse d'erreur

In memory of Erin Brent (1947–2005)

1. Introduction

In an earlier paper [2], the second author made the claim that the maximum relative error which can occur when computing the product $z_0 z_1$ of two complex values using floating-point arithmetic is $\epsilon\sqrt{5}$, where ϵ is the maximum relative error which can result from rounded floating-point addition, subtraction, or multiplication. While reviewing that paper a few years later, the other two authors noted that the proof given was incorrect, although the result claimed was true.

Since the bound of $\epsilon\sqrt{8}$ which is commonly used [1] is suboptimal, we present here a corrected proof of the tighter bound. Interestingly, by explicitly finding worst-case inputs, we can demonstrate that our error bound is effectively optimal.

Throughout this paper, we concern ourselves with floating-point arithmetic with t-digit base-β significands, denote by $\mathrm{ulp}(x)$ for $x \neq 0$ the (unique) power of β such that $\beta^{t-1} \leq |x|/\mathrm{ulp}(x) < \beta^t$, and write $\epsilon = \frac{1}{2}\mathrm{ulp}(1) = \frac{1}{2}\beta^{1-t}$; we also define $\mathrm{ulp}(0) = 0$. We use the notations $x \oplus y$, $x \ominus y$, and $x \otimes y$ to represent rounded floating-point addition, subtraction, and multiplication of the values x and y.

2. An Error Bound

Theorem 1. *Let $z_0 = a_0 + b_0 i$ and $z_1 = a_1 + b_1 i$, with a_0, b_0, a_1, b_1 floating-point values with t-digit base-β significands, and let $z_2 = ((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) + ((a_0 \otimes b_1) \oplus (b_0 \otimes a_1))i$ be computed. Providing that no overflow or underflow occur, no denormal values are produced, arithmetic results are correctly rounded to a nearest representable value, $z_0 z_1 \neq 0$, and $\epsilon \leq 2^{-5}$, the relative error*

$$\left|z_2 (z_0 z_1)^{-1} - 1\right|$$

is less than $\epsilon\sqrt{5} = \frac{1}{2}\beta^{1-t}\sqrt{5}$.

Proof. Let a_0, b_0, a_1, and b_1 be chosen such that the relative error is maximized. By multiplying z_0 and z_1 by powers of i and/or taking complex conjugates, we can assume without loss of generality that

$$0 \leq a_0, b_0, a_1, b_1 \tag{1}$$

$$b_0 b_1 \leq a_0 a_1 \tag{2}$$

and given our assumptions that overflow, underflow, and denormals do not occur, and that rounding is performed to a nearest representable value, we can conclude that for any x occurring in the computation, the error introduced when rounding x is at most $\frac{1}{2}\mathrm{ulp}(x)$ and is strictly less than $\epsilon \cdot x$.

We note that the error $|\Im(z_2 - z_0 z_1)|$ in the imaginary part of z_2 is bounded as follows:

$$\begin{aligned} |\Im(z_2 - z_0 z_1)| &= |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 b_1 + b_0 a_1)| \\ &\leq |a_0 \otimes b_1 - a_0 b_1| + |b_0 \otimes a_1 - b_0 a_1| \\ &\quad + |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| \end{aligned}$$

and consider two cases:

Case I1: $\mathrm{ulp}(a_0 b_1 + b_0 a_1) < \mathrm{ulp}(a_0 \otimes b_1 + b_0 \otimes a_1)$

Using first the definition of ulp and second the assumption above, we must have

$$a_0 b_1 + b_0 a_1 < \beta^t \mathrm{ulp}(a_0 b_1 + b_0 a_1) \leq a_0 \otimes b_1 + b_0 \otimes a_1$$

and therefore

$$\begin{aligned} \left|(a_0 \otimes b_1 + b_0 \otimes a_1) - \beta^t \mathrm{ulp}(a_0 b_1 + b_0 a_1)\right| &< (a_0 \otimes b_1 + b_0 \otimes a_1) - (a_0 b_1 + b_0 a_1) \\ &\leq |a_0 \otimes b_1 - a_0 b_1| + |b_0 \otimes a_1 - b_0 a_1| \\ &\leq \epsilon \cdot (a_0 b_1 + b_0 a_1). \end{aligned}$$

However, $\beta^t \mathrm{ulp}(a_0 b_1 + b_0 a_1)$ is a representable floating-point value; so given our assumption that rounding is performed to a nearest representable value, we must now have

$$|((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| < \epsilon \cdot (a_0 b_1 + b_0 a_1).$$

Case I2: $\mathrm{ulp}(a_0 \otimes b_1 + b_0 \otimes a_1) \leq \mathrm{ulp}(a_0 b_1 + b_0 a_1)$

From our assumption that the results of arithmetic operations are correctly rounded, we obtain

$$\begin{aligned} |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| &\leq \frac{1}{2} \mathrm{ulp}(a_0 \otimes b_1 + b_0 \otimes a_1) \\ &\leq \frac{1}{2} \mathrm{ulp}(a_0 b_1 + b_0 a_1) \\ &\leq \epsilon \cdot (a_0 b_1 + b_0 a_1). \end{aligned}$$

Combining these two cases with the earlier-stated bound, we obtain

$$\begin{aligned} |\Im(z_2 - z_0 z_1)| &\leq |a_0 \otimes b_1 - a_0 b_1| + |b_0 \otimes a_1 - b_0 a_1| \\ &\quad + |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| \\ &< \epsilon \cdot (a_0 b_1) + \epsilon \cdot (b_0 a_1) + \epsilon \cdot (a_0 b_1 + b_0 a_1) \\ &= \epsilon \cdot (2 a_0 b_1 + 2 b_0 a_1). \end{aligned}$$

Now that we have a bound on the imaginary part of the error, we turn our attention to the real part, and consider the following four cases (where the examples given apply to $\beta = 2$):

$$\begin{array}{ll} \mathrm{ulp}(b_0 b_1) \leq \mathrm{ulp}(a_0 a_1) \leq \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) & \text{e.g., } z_0 = z_1 = 0.8 + 0.1i \\ \mathrm{ulp}(b_0 b_1) < \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0 a_1) & \text{e.g., } z_0 = z_1 = 0.8 + 0.4i \\ \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \leq \mathrm{ulp}(b_0 b_1) < \mathrm{ulp}(a_0 a_1) & \text{e.g., } z_0 = z_1 = 0.8 + 0.7i \\ \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(b_0 b_1) = \mathrm{ulp}(a_0 a_1) & \text{e.g., } z_0 = z_1 = 0.8 + 0.8i. \end{array}$$

Since we have assumed that $b_0 b_1 \leq a_0 a_1$, we know that $\mathrm{ulp}(b_0 b_1) \leq \mathrm{ulp}(a_0 a_1)$, and thus these four cases cover all possible inputs. Consequently, it suffices to prove the required bound for each of these four cases.

Case R1: $\mathrm{ulp}(b_0 b_1) \le \mathrm{ulp}(a_0 a_1) \le \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)$

Note that the right inequality can only be strict if $a_0 \otimes a_1$ rounds up to a power of β and $b_0 b_1 = 0$.

We observe that

$$a_0 \otimes a_1 - b_0 \otimes b_1 < a_0 a_1 - b_0 b_1 + \epsilon \cdot (a_0 a_1 + b_0 b_1)$$

and bound the real part of the complex error as follows:

$$\begin{aligned}
|\Re(z_2 - z_0 z_1)| &\le |a_0 \otimes a_1 - a_0 a_1| + |b_0 \otimes b_1 - b_0 b_1| \\
&\quad + |((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) - (a_0 \otimes a_1 - b_0 \otimes b_1)| \\
&\le \frac{1}{2}\mathrm{ulp}(a_0 a_1) + \frac{1}{2}\mathrm{ulp}(b_0 b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \\
&\le \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) + \frac{1}{2}\mathrm{ulp}(b_0 b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \\
&< 2\epsilon \cdot (a_0 \otimes a_1 - b_0 \otimes b_1) + \epsilon \cdot (b_0 b_1) \\
&< \epsilon \cdot (2a_0 a_1 - b_0 b_1) + \epsilon^2 \cdot (2a_0 a_1 + 2b_0 b_1).
\end{aligned}$$

Applying the triangle inequality, we now observe that

$$\begin{aligned}
|z_2 - z_0 z_1| &= \sqrt{\Re(z_2 - z_0 z_1)^2 + \Im(z_2 - z_0 z_1)^2} \\
&< \epsilon \sqrt{(2a_0 a_1 - b_0 b_1)^2 + (2a_0 b_1 + 2b_0 a_1)^2} + \epsilon^2 \cdot (2a_0 a_1 + 2b_0 b_1) \\
&\le \epsilon \sqrt{\frac{32}{7}|z_0 z_1|^2 - \frac{4}{7}(a_0 b_1 - b_0 a_1)^2 - \frac{1}{7}(2a_0 a_1 - 5b_0 b_1)^2} + 2\epsilon^2 |z_0 z_1| \\
&\le \epsilon \left(\sqrt{32/7} + 2\epsilon \right) |z_0 z_1| < \epsilon \sqrt{5}\, |z_0 z_1|
\end{aligned}$$

as required.

Case R2: $\mathrm{ulp}(b_0 b_1) < \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0 a_1)$

Noting that $\mathrm{ulp}(x) < \mathrm{ulp}(y)$ implies $\mathrm{ulp}(x) \le \beta^{-1}\mathrm{ulp}(y) \le \frac{1}{2}\mathrm{ulp}(y)$, we obtain

$$\begin{aligned}
|\Re(z_2 - z_0 z_1)| &\le \frac{1}{2}\mathrm{ulp}(a_0 a_1) + \frac{1}{2}\mathrm{ulp}(b_0 b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \\
&\le \frac{7}{8}\mathrm{ulp}(a_0 a_1) \\
&\le \epsilon \cdot \left(\frac{7}{4} a_0 a_1 \right)
\end{aligned}$$

and therefore

$$\begin{aligned}
|z_2 - z_0 z_1| &= \sqrt{\Re(z_2 - z_0 z_1)^2 + \Im(z_2 - z_0 z_1)^2} \\
&< \epsilon \sqrt{\left(\frac{7}{4} a_0 a_1\right)^2 + (2a_0 b_1 + 2b_0 a_1)^2} \\
&= \epsilon \sqrt{\frac{1024}{207} |z_0 z_1|^2 - \frac{196}{207}(a_0 b_1 - b_0 a_1)^2 - \frac{1}{3312}(79 a_0 a_1 - 128 b_0 b_1)^2} \\
&\leq \epsilon \sqrt{1024/207}\, |z_0 z_1| < \epsilon \sqrt{5}\, |z_0 z_1|
\end{aligned}$$

as required.

Case R3: $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \leq \mathrm{ulp}(b_0 b_1) < \mathrm{ulp}(a_0 a_1)$

In this case, there is no rounding error introduced in computing the difference between $a_0 \otimes a_1$ and $b_0 \otimes b_1$ since $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \leq \mathrm{ulp}(b_0 b_1) \leq \mathrm{ulp}(b_0 \otimes b_1)$ and $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0 a_1) \leq \mathrm{ulp}(a_0 \otimes a_1)$. Also,

$$\mathrm{ulp}(b_0 b_1) \leq \frac{1}{\beta} \mathrm{ulp}(a_0 a_1) \leq \frac{1}{2} \mathrm{ulp}(a_0 a_1)$$

so we have

$$\begin{aligned}
|\Re(z_2 - z_0 z_1)| &\leq \frac{1}{2} \mathrm{ulp}(a_0 a_1) + \frac{1}{2} \mathrm{ulp}(b_0 b_1) \\
&\leq \frac{3}{4} \mathrm{ulp}(a_0 a_1) \\
&\leq \epsilon \cdot \left(\frac{3}{2} a_0 a_1\right)
\end{aligned}$$

and consequently

$$\begin{aligned}
|z_2 - z_0 z_1| &= \sqrt{\Re(z_2 - z_0 z_1)^2 + \Im(z_2 - z_0 z_1)^2} \\
&< \epsilon \sqrt{\left(\frac{3}{2} a_0 a_1\right)^2 + (2a_0 b_1 + 2b_0 a_1)^2} \\
&= \epsilon \sqrt{\frac{256}{55} |z_0 z_1|^2 - \frac{36}{55}(a_0 b_1 - b_0 a_1)^2 - \frac{1}{220}(23 a_0 a_1 - 32 b_0 b_1)^2} \\
&\leq \epsilon \sqrt{256/55}\, |z_0 z_1| < \epsilon \sqrt{5}\, |z_0 z_1|
\end{aligned}$$

as required.

Case R4: $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(b_0 b_1) = \mathrm{ulp}(a_0 a_1)$

In this case, there is again no rounding error introduced in computing the difference between $a_0 \otimes a_1$ and $b_0 \otimes b_1$, so we obtain

$$\begin{aligned} |\Re(z_2 - z_0 z_1)| &\leq |a_0 \otimes a_1 - a_0 a_1| + |b_0 \otimes b_1 - b_0 b_1| \\ &< \epsilon \cdot (a_0 a_1 + b_0 b_1) \end{aligned}$$

and consequently

$$\begin{aligned} |z_2 - z_0 z_1| &= \sqrt{\Re(z_2 - z_0 z_1)^2 + \Im(z_2 - z_0 z_1)^2} \\ &< \epsilon \sqrt{(a_0 a_1 + b_0 b_1)^2 + (2a_0 b_1 + 2b_0 a_1)^2} \\ &= \epsilon \sqrt{5\,|z_0 z_1|^2 - (a_0 b_1 - b_0 a_1)^2 - 4(a_0 a_1 - b_0 b_1)^2} \\ &\leq \epsilon \sqrt{5}\,|z_0 z_1| \end{aligned}$$

as required. □

3. Worst-case multiplicands for $\beta = 2$

Having proved an upper bound on the relative error which can result from floating-point rounding when computing the product of complex values, we now turn to a more number-theoretic problem: finding precise worst-case inputs for $\beta = 2$. Starting with the assumption that some inputs produce errors very close to the proven upper bound, we will repeatedly reduce the set of possible inputs until an exhaustive search becomes feasible.

Theorem 2. *Let $\beta = 2$ and assume that $z_0 = a_0 + b_0 i \neq 0$ and $z_1 = a_1 + b_1 i \neq 0$, where a_0, b_0, a_1, b_1 are floating-point values with t-digit base-β significands, and $z_2 = ((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) + ((a_0 \otimes b_1) \oplus (b_0 \otimes a_1))i$ are such that*

$$0 \leq a_0, b_0, a_1, b_1 \tag{1}$$

$$b_0 b_1 \leq a_0 a_1 \tag{2}$$

$$b_0 a_1 \leq a_0 b_1 \tag{3}$$

$$1/2 \leq a_0 a_1 < 1 \tag{4}$$

and no overflow, underflow, or denormal values occur during the computation of z_2. Assume further that the results of arithmetic operations are correctly rounded to a nearest representable value and that

$$\frac{|z_2 - z_0 z_1|}{|z_0 z_1|} > \epsilon\sqrt{5 - n\epsilon} > \epsilon \cdot \max\left(\sqrt{1024/207}, \sqrt{32/7} + 2\epsilon\right) \tag{5}$$

for some positive integer n. Then

$$\begin{aligned}
a_0a_1 &= 1/2 + (j_{aa} + 1/2)\epsilon + k_{aa}\epsilon^2 \\
a_0b_1 &= 1/2 + (j_{ab} + 1/2)\epsilon + k_{ab}\epsilon^2 \\
b_0a_1 &= 1/2 + (j_{ba} + 1/2)\epsilon + k_{ba}\epsilon^2 \\
b_0b_1 &= 1/2 + (j_{bb} + 1/2)\epsilon + k_{bb}\epsilon^2
\end{aligned}$$

for some integers j_{xy}, k_{xy} satisfying

$$0 \le j_{aa}, j_{ab}, j_{ba}, j_{bb} < \frac{n}{4}$$

$$|k_{aa}|, |k_{bb}| < n$$

$$|k_{ab}|, |k_{ba}| < \frac{n}{2}$$

and $a_0 \neq b_0$, $a_1 \neq b_1$.

Proof. From equation (5), we note that $\epsilon \le n\epsilon < 11/207 < 2^{-4}$; we will use this trivial bound later without explicit comment.

From the proof of Theorem 1, we know that Case R4 must hold, i.e., there is no error introduced in the computation of the difference between $a_0 \otimes a_1$ and $b_0 \otimes b_1$, and $\mathrm{ulp}(b_0b_1) = \mathrm{ulp}(a_0a_1)$. From inequalities (2) and (4) above, this implies that

$$1/2 \le b_0b_1 \le a_0a_1 < 1$$

$$|\Re(z_2 - z_0z_1)| \le |a_0 \otimes a_1 - a_0a_1| + |b_0 \otimes b_1 - b_0b_1| \le \epsilon.$$

We can now obtain lower bounds on $|z_0z_1|$ and $|z_2 - z_0z_1|$, using the fact that $(a_0a_1)(b_0b_1) = (a_0b_1)(b_0a_1)$:

$$\begin{aligned}
|z_0z_1|^2 &= \left(a_0^2 + b_0^2\right)\left(a_1^2 + b_1^2\right) \\
&= (a_0a_1)^2 + (a_0b_1)^2 + (b_0a_1)^2 + (b_0b_1)^2 \\
&\ge (1/2)^2 + (a_0b_1)^2 + \frac{(1/2)^4}{(a_0b_1)^2} + (1/2)^2 \ge 1 \\
|z_2 - z_0z_1|^2 &> |z_0z_1|^2 \epsilon^2(5 - n\epsilon) \ge \epsilon^2(5 - n\epsilon)
\end{aligned}$$

as well as an upper bound on $|z_0z_1|$:

$$\begin{aligned}
|z_0z_1|^2 \cdot \frac{1024\epsilon^2}{207} &< |z_2 - z_0z_1|^2 \\
&= |\Re(z_2 - z_0z_1)|^2 + |\Im(z_2 - z_0z_1)|^2 \\
&< \epsilon^2 + (\epsilon \cdot (2a_0b_1 + 2b_0a_1))^2 \\
&\le \epsilon^2 + 4\epsilon^2 |z_0z_1|^2 \\
|z_0z_1|^2 &< \frac{207}{196}
\end{aligned}$$

We now note that

$$\begin{aligned}(a_0b_1)^2 &\le |z_0z_1|^2 - (a_0a_1)^2 - (b_0b_1)^2 \\ &\le \frac{207}{196} - \frac{1}{4} - \frac{1}{4} = \frac{109}{196}\end{aligned}$$

so $b_0a_1 \le a_0b_1 \le \sqrt{109/196} < 1$ and $a_0 \otimes b_1 + b_0 \otimes a_1 \le \sqrt{109/49} \cdot (1+\epsilon) < 2$; this implies that $\mathrm{ulp}(b_0a_1) \le \mathrm{ulp}(a_0b_1) \le \mathrm{ulp}(1/2)$ and $\mathrm{ulp}(a_0 \otimes b_1 + b_0 \otimes a_1) \le \mathrm{ulp}(1)$, and therefore

$$\begin{aligned}|a_0 \otimes b_1 - a_0b_1| &\le \epsilon/2 \\ |b_0 \otimes a_1 - b_0a_1| &\le \epsilon/2 \\ |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| &\le \epsilon \\ |\Im(z_2 - z_0z_1)| &\le \epsilon/2 + \epsilon/2 + \epsilon = 2\epsilon\end{aligned}$$

which allows us to place upper bounds on $|z_2 - z_0z_1|$ and $|z_0z_1|$:

$$|z_2 - z_0z_1|^2 = |\Re(z_2 - z_0z_1)|^2 + |\Im(z_2 - z_0z_1)|^2 \le (\epsilon)^2 + (2\epsilon)^2 = 5\epsilon^2$$

$$|z_0z_1|^2 < \frac{|z_2 - z_0z_1|^2}{\epsilon^2(5 - n\epsilon)} \le \frac{5}{5 - n\epsilon}.$$

Combining the known lower bound $\epsilon^2(5 - n\epsilon)$ for $|z_2 - z_0z_1|^2$ with the upper bounds on the error contributed by each individual rounding step, we find that

$$\begin{aligned}\epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon &< |a_0 \otimes a_1 - a_0a_1| \le \epsilon/2 \\ \epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon &< |b_0 \otimes b_1 - b_0b_1| \le \epsilon/2 \\ \epsilon/2 - (2 - \sqrt{4 - n\epsilon})\epsilon &< |a_0 \otimes b_1 - a_0b_1| \le \epsilon/2 \\ \epsilon/2 - (2 - \sqrt{4 - n\epsilon})\epsilon &< |b_0 \otimes a_1 - b_0a_1| \le \epsilon/2\end{aligned}$$

and similarly, by combining the upper bound on $|z_0z_1|^2$ with lower bound of $1/2$ for each pairwise product, we obtain

$$\begin{aligned}1/2 \le b_0b_1 \le a_0a_1 &\le \sqrt{\frac{5}{5 - n\epsilon} - \frac{3}{4}} = \sqrt{\frac{5 + 3n\epsilon}{20 - 4n\epsilon}} \\ 1/2 \le b_0a_1 \le a_0b_1 &\le \sqrt{\frac{5}{5 - n\epsilon} - \frac{3}{4}} = \sqrt{\frac{5 + 3n\epsilon}{20 - 4n\epsilon}}.\end{aligned}$$

Now consider the possible values for a_0a_1 which satisfy these restrictions. Since it is the product of two values which are expressible using t digits of significand, a_0a_1 can be exactly represented using $2t$ digits of significand; but since $1/2 \le a_0a_1 < 1$, this implies that a_0a_1 is an integer multiple of ϵ^2. There is therefore at least one pair of integers j_{aa}, k_{aa} with $0 \le j_{aa} < \epsilon^{-1}/2$, $|k_{aa}| \le \epsilon^{-1}/2$ for which

$$a_0a_1 = 1/2 + (j_{aa} + 1/2)\epsilon + k_{aa}\epsilon^2.$$

Since $a_0 \otimes a_1$ is the closest multiple of ϵ to a_0a_1, this implies that

$$\begin{aligned}\epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon < |a_0 \otimes a_1 - a_0a_1| &= \epsilon/2 - |k_{aa}|\,\epsilon^2 \\ |k_{aa}|\,\epsilon < 1 - \sqrt{1 - n\epsilon} &< 1 - (1 - n\epsilon) = n\epsilon\end{aligned}$$

i.e., $|k_{aa}| < n$, and similarly

$$1/2 + j_{aa}\epsilon \le a_0 a_1 \le \sqrt{\frac{5 + 3n\epsilon}{20 - 4n\epsilon}} < \sqrt{1/4 + n\epsilon/4} < 1/2 + \frac{n\epsilon}{4}$$

i.e., $0 \le j_{aa} < n/4$.

Applying the same argument to $a_0 b_1$, $b_0 a_1$, and $b_0 b_1$ allows us to infer that they possess the same structure, as required. To complete the proof, we note that the rounding errors from the products $a_0 a_1$ and $b_0 b_1$ must be in opposite directions (in order that they accumulate when subtracted), while the rounding errors from the products $a_0 b_1$ and $b_0 a_1$ must be in the same direction (in order that they accumulate when added); consequently, we must have $a_0 \ne b_0$ and $a_1 \ne b_1$. □

Corollary 1. *Assume that the preconditions of Theorem 2 are satisfied, and assume further that*

$$\frac{1}{2} \le a_0 < 1 \tag{6}$$

and $n \le 2\epsilon^{-1/2}$. Then

$$\frac{1}{2} < a_0, b_0, a_1, b_1 < 1.$$

Proof. Assume that $a_1 \ge 1$. Then we can write

$$a_0 = 1/2 + A\epsilon$$

$$a_1 = 1 + 2B\epsilon$$

for some $0 \le A, B < (2\epsilon)^{-1}$. From Theorem 2, we have

$$1/2 + (A + B)\epsilon + 2AB\epsilon^2 = a_0 a_1 = 1/2 + (j_{aa} + 1/2)\epsilon + k_{aa}\epsilon^2$$

for some $0 \le j_{aa} < n/4$, $|k_{aa}| < n$.

As a result, we must have $A + B \le n/4 \le 1/2 \cdot \epsilon^{-1/2}$, and since $0 \le A, B$ this implies $0 \le 2AB\epsilon^2 \le \epsilon/8$. However, by reducing the equation above modulo ϵ, we find that $2AB\epsilon^2 \equiv \epsilon/2 + k_{aa}\epsilon^2$, which contradicts our bounds on $2AB\epsilon^2$. Consequently, we can conclude that $a_1 < 1$. Now we note that $a_0 a_1 > 1/2$ and $a_0 < 1$, so $a_1 > 1/2$, and we have both of the bounds required for a_1.

Applying the same argument to the other products provides the same bounds for a_0, b_0, and b_1. □

Corollary 2. *Assume that the preconditions of Corollary 1 are satisfied, and assume further that $n \le \epsilon^{-1/2}$ and $\epsilon \le 2^{-6}$. Then*

$$j_{aa} - j_{ab} - j_{ba} + j_{bb} = 0,$$

$$|a_0 - b_0| \cdot |a_1 - b_1| < 3n\epsilon^2.$$

Proof. From Theorem 2, we obtain that

$$a_0(a_1 - b_1) = a_0a_1 - a_0b_1 = (j_{aa} - j_{ab})\epsilon + (k_{aa} - k_{ab})\epsilon^2$$

where $|j_{aa} - j_{ab}| < \frac{n}{4}$, $|k_{aa} - k_{ab}| < \frac{3n}{2}$, and since $a_0 > \frac{1}{2}$ (from Corollary 1), we can conclude that $|a_1 - b_1| < \frac{n}{2}\epsilon + 3n\epsilon^2$. Since a_1 and b_1 are integer multiples of ϵ and $3n\epsilon^2 < \epsilon/2$, we conclude that $|a_1 - b_1| \le \frac{n}{2}\epsilon$. Applying the same argument to the product $a_1(a_0 - b_0)$ provides the same bound for $|a_0 - b_0|$.

We now note that

$$\begin{aligned}\left|(j_{aa} - j_{ab} - j_{ba} + j_{bb})\epsilon + (k_{aa} - k_{ab} - k_{ba} + k_{bb})\epsilon^2\right| &= |a_0 - b_0| \cdot |a_1 - b_1| \\ &\le \left(\frac{n}{2}\epsilon\right)^2 \\ &< \frac{\epsilon}{4}\end{aligned}$$

from our assumed upper bound on n, and consequently we can conclude that $j_{aa} - j_{ab} - j_{ba} + j_{bb} = 0$. Finally, this allows us to write

$$\begin{aligned}|a_0 - b_0| \cdot |a_1 - b_1| &= |k_{aa} - k_{ab} - k_{ba} + k_{bb}|\, \epsilon^2 \\ &< 3n\epsilon^2\end{aligned}$$

as required. □

Corollary 3. *Assume that the preconditions of Corollary 1 are satisfied, and assume further that $n \le \frac{1}{4}\epsilon^{-1/2}$. Then*

$$(a_0 - b_0)(a_1 - b_1) = 2(j_{aa} - j_{ab})(j_{aa} - j_{ba})\epsilon^2$$
$$(a_0 - b_0)(a_1 - b_1)k_{aa} = (k_{aa} - k_{ab})(k_{aa} - k_{ba})\epsilon^2$$

Proof. For brevity and clarity, we will write $(a_0 - b_0)(a_1 - b_1) = x\epsilon^2$ and note that x is an integer between $-3n$ and $3n$, from Corollary 2. Then

$$\begin{aligned}xa_0a_1 &= \frac{x}{2} + x\left(j_{aa} + \frac{1}{2}\right)\epsilon + xk_{aa}\epsilon^2, \\ xa_0a_1 &= \frac{a_0(a_1 - b_1)}{\epsilon} \cdot \frac{a_1(a_0 - b_0)}{\epsilon} \\ &= \left((j_{aa} - j_{ab}) + (k_{aa} - k_{ab})\epsilon\right)\left((j_{aa} - j_{ba}) + (k_{aa} - k_{ba})\epsilon\right) \\ &= (j_{aa} - j_{ab})(j_{aa} - j_{ba}) + \left((j_{aa} - j_{ab})(k_{aa} - k_{ba}) + (j_{aa} - j_{ba})(k_{aa} - k_{ab})\right)\epsilon \\ &\quad + (k_{aa} - k_{ab})(k_{aa} - k_{ba})\epsilon^2.\end{aligned}$$

Consequently,

$$\begin{aligned}&x - 2(j_{aa} - j_{ab})(j_{aa} - j_{ba}) \\ &\qquad = \left(2(j_{aa} - j_{ab})(k_{aa} - k_{ba}) + 2(j_{aa} - j_{ba})(k_{aa} - k_{ab}) - (2j_{aa} + 1)x\right)\epsilon \\ &\qquad\quad + \left(2(k_{aa} - k_{ab})(k_{aa} - k_{ba}) - 2k_{aa}x\right)\epsilon^2,\end{aligned}$$

$$\begin{aligned}|x - 2(j_{aa} - j_{ab})(j_{aa} - j_{ba})| &\leq \left(2\frac{n}{4}\frac{3n}{2} + 2\frac{n}{4}\frac{3n}{2} + 3\left(\frac{n}{2}+1\right)n\right)\epsilon \\ &\quad + \left(2\frac{3n}{2}\frac{3n}{2} + 6n^2\right)\epsilon^2 \\ &= \left(3n^2 + 3n\right)\epsilon + \frac{21}{2}n^2\epsilon^2 \\ &\leq \frac{3}{16} + \frac{3}{4}\sqrt{\epsilon} + \frac{21}{32}\epsilon < 1,\end{aligned}$$

and since the only integer with absolute value less than one is zero, we can conclude that $x = 2(j_{aa} - j_{ab})(j_{aa} - j_{ba})$ as required.

We now consider $xa_0a_1\epsilon^{-2}$ modulo $\frac{1}{2}\epsilon^{-1}$, and note that

$$\begin{aligned}xk_{aa} &\equiv xa_0a_1\epsilon^{-2} \\ &\equiv (k_{aa} - k_{ab})(k_{aa} - k_{ba})\end{aligned}$$

and further that

$$\begin{aligned}|xk_{aa} - (k_{aa} - k_{ab})(k_{aa} - k_{ba})| &\leq 3n \cdot n + \frac{3n}{2} \cdot \frac{3n}{2} \\ &= \frac{21n^2}{4} \leq \frac{21\epsilon^{-1}}{64} < \frac{1}{2}\epsilon^{-1}\end{aligned}$$

and therefore $xk_{aa} = (k_{aa} - k_{ab})(k_{aa} - k_{ba})$. □

Theorem 3. *Let $\beta = 2$ and assume that $z_0 = a_0 + b_0 i$, $z_1 = a_1 + b_1 i$, and $z_2 = ((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) + ((a_0 \otimes b_1) \oplus (b_0 \otimes a_1))i$ are such that*

(1) $$0 \leq a_0, b_0, a_1, b_1$$

(2) $$b_0b_1 \leq a_0a_1$$

(3) $$b_0a_1 \leq a_0b_1$$

(4) $$1/2 \leq a_0a_1 < 1$$

(6) $$1/2 \leq a_0 < 1$$

and no overflow, underflow, or denormal values occur during the computation of z_2. Assume further that the results of arithmetic operations are correctly rounded to the nearest representable value, and that

(5) $$\frac{|z_2 - z_0z_1|}{|z_0z_1|} > \epsilon\sqrt{5 - n\epsilon} > \epsilon \cdot \max\left(\sqrt{1024/207}, \sqrt{32/7} + 2\epsilon\right)$$

for some $n < \frac{1}{4}\epsilon^{-1/2}$ *and* $\epsilon \leq 2^{-6}$. *Then there exist integers* c_0, d_0, α_0, β_0, c_1, d_1, α_1, β_1 *satisfying*

$$\begin{aligned}
a_0 &= \frac{c_0}{d_0}(1+\alpha_0\epsilon) & b_0 &= \frac{c_0}{d_0}(1+\beta_0\epsilon)\\
a_1 &= \frac{c_1}{d_1}(1+\alpha_1\epsilon) & b_1 &= \frac{c_1}{d_1}(1+\beta_1\epsilon)\\
&\gcd(c_0,d_0) = 1 & &\frac{d_0}{2} \leq c_0 \leq d_0\\
&\gcd(c_1,d_1) = 1 & &\frac{d_1}{2} \leq c_1 \leq d_1\\
&2c_0c_1 = d_0d_1 < 3n & &\frac{1}{2} < a_0, b_0, a_1, b_1 < 1\\
&\alpha_0 \equiv \beta_0 \equiv -\epsilon^{-1} \pmod{d_0} & &\alpha_0 \neq \beta_0\\
&\alpha_1 \equiv \beta_1 \equiv -\epsilon^{-1} \pmod{d_1} & &\alpha_1 \neq \beta_1\\
&\min(\alpha_0,\beta_0) + \min(\alpha_1,\beta_1) \geq 0 & &\max(|\alpha_0|,|\beta_0|)\cdot\max(|\alpha_1|,|\beta_1|) < n
\end{aligned}$$

Proof. Let the values j_{aa}, j_{ab}, j_{ba}, j_{bb}, k_{aa}, k_{ab}, k_{ba}, and k_{bb} be as constructed in Theorem 2, and further let $g_0 = \gcd(j_{aa} - j_{ab}, (a_1 - b_1)/\epsilon)$. From Corollary 1 we know that $1/2 < a_1, b_1 < 1$, so a_1 and b_1 are multiples of ϵ; consequently g_0 must be an integer. By the same argument, $g_1 = \gcd(j_{aa} - j_{ba}, (a_0 - b_0)/\epsilon)$ is an integer.

Now note that

$$g_0 | (a_1 - b_1)\epsilon^{-1} | (a_1 - b_1)a_0\epsilon^{-2} = (j_{aa} - j_{ab})\epsilon^{-1} + (k_{aa} - k_{ab})$$

and since $g_0 | (j_{aa} - j_{ab})$, we can conclude that $g_0 | (k_{aa} - k_{ab})$. By the same argument, $g_1 | (k_{aa} - k_{ba})$.

We now write

$$\begin{aligned}
c_0 &= \frac{j_{aa} - j_{ab}}{g_0} & d_0 &= \frac{a_1 - b_1}{g_0\epsilon} & e_0 &= \frac{k_{aa} - k_{ab}}{g_0}\\
c_1 &= \frac{j_{aa} - j_{ba}}{g_1} & d_1 &= \frac{a_0 - b_0}{g_1\epsilon} & e_1 &= \frac{k_{aa} - k_{ba}}{g_1}
\end{aligned}$$

and note that these values are all integers; further, from Corollary 3 we have $d_0d_1k_{aa} = e_0e_1$ and $d_0d_1 = 2c_0c_1$, and since $\gcd(c_0,d_0) = \gcd(c_1,d_1) = 1$ by construction, this implies $\gcd(c_0,c_1) = 1$.

We now observe that

$$a_0 = \frac{a_0(a_1 - b_1)}{a_1 - b_1} = \frac{c_0g_0\epsilon + e_0g_0\epsilon^2}{d_0g_0\epsilon} = \frac{c_0 + e_0\epsilon}{d_0}$$

$$a_1 = \frac{a_1(a_0 - b_0)}{a_0 - b_0} = \frac{c_1g_1\epsilon + e_1g_1\epsilon^2}{d_1g_1\epsilon} = \frac{c_1 + e_1\epsilon}{d_1}$$

and therefore

$$\begin{aligned}\frac{1}{2} + \left(j_{aa} + \frac{1}{2}\right)\epsilon + k_{aa}\epsilon^2 &= a_0 a_1 \\ &= \frac{c_0 c_1}{d_0 d_1} + \frac{c_0 e_1 + e_0 c_1}{d_0 d_1}\epsilon + \frac{e_0 e_1}{d_0 d_1}\epsilon^2 \\ &= \frac{1}{2} + \frac{c_0 e_1 + e_0 c_1}{d_0 d_1}\epsilon + k_{aa}\epsilon^2\end{aligned}$$

and thus (using $d_0 d_1 = 2c_0 c_1$)

$$c_0 c_1 \left(2j_{aa} + 1\right) = c_0 e_1 + e_0 c_1.$$

Consequently $c_0 | e_0 c_1$ and $c_1 | c_0 e_1$, and since $\gcd(c_0, c_1) = 1$ it follows that $c_0 | e_0$ and $c_1 | e_1$. Writing $e_0 = c_0 \alpha_0$, $e_1 = c_1 \alpha_1$ for integers α_0, α_1, we now have

$$a_0 = \frac{c_0}{d_0}(1 + \alpha_0 \epsilon) \qquad a_1 = \frac{c_1}{d_1}(1 + \alpha_1 \epsilon)$$

and taking $\beta_0 = \alpha_0 + 2c_1 g_1$, $\beta_1 = \alpha_1 + 2c_0 g_0$, we have

$$b_0 = \frac{c_0}{d_0}(1 + \beta_0 \epsilon) \qquad b_1 = \frac{c_1}{d_1}(1 + \beta_1 \epsilon)$$

as required.

The remaining conditions can be obtained by remembering that a_0, b_0, a_1, and b_1 are integer multiples of ϵ, and by using the bounds on j_{xy} and k_{xy} given in Theorem 2. □

Corollary 4. *In IEEE 754 single-precision arithmetic ($\beta = 2$, $t = 24$, $\epsilon = 2^{-24}$), using "nearest even" rounding mode, the values*[1]

$$a_0 = \frac{3}{4} \qquad b_0 = \frac{3}{4}(1 - 4\epsilon) \qquad a_1 = \frac{2}{3}(1 + 11\epsilon) \qquad b_1 = \frac{2}{3}(1 + 5\epsilon)$$

result in a relative error $\delta \approx \epsilon\sqrt{5 - 168\epsilon} \approx \epsilon\sqrt{4.9999899864}$ in z_2, and δ is the worst possible provided that overflow, underflow, and denormals do not occur.

Proof. Straightforward computation for the values given establishes that

$$\begin{aligned}a_0 a_1 &= \frac{1}{2}(1 + 11\epsilon) & a_0 \otimes a_1 &= \frac{1}{2}(1 + 12\epsilon) \\ b_0 b_1 &= \frac{1}{2}(1 + \epsilon - 20\epsilon^2) & b_0 \otimes b_1 &= \frac{1}{2} \\ \Re(z_0 z_1) &= 5\epsilon + 10\epsilon^2 & \Re(z_2) &= 6\epsilon \\ a_0 b_1 &= \frac{1}{2}(1 + 5\epsilon) & a_0 \otimes b_1 &= \frac{1}{2}(1 + 4\epsilon) \\ b_0 a_1 &= \frac{1}{2}(1 + 7\epsilon - 44\epsilon^2) & b_0 \otimes a_1 &= \frac{1}{2}(1 + 6\epsilon) \\ \Im(z_0 z_1) &= 1 + 6\epsilon - 22\epsilon^2 & \Im(z_2) &= 1 + 4\epsilon\end{aligned}$$

[1] Note that while $\frac{2}{3}$ is not an IEEE 754 single-precision value, $\frac{2}{3}(1 + 5\epsilon)$ and $\frac{2}{3}(1 + 11\epsilon)$ are, since $\epsilon^{-1} + 5 \equiv \epsilon^{-1} + 11 \equiv 0 \pmod 3$.

$$|z_2 - z_0 z_1|^2 = \epsilon^2(5 - 108\epsilon + \mathrm{O}(\epsilon^2))$$
$$|z_0 z_1|^2 = 1 + 12\epsilon + \mathrm{O}(\epsilon^2)$$

and the ratio of these provides the error as stated.

To prove that this is the worst possible relative error, we note that the mappings $z_0 \to z_0 i$, $z_1 \to z_1 i$, $(z_0, z_1) \to (\bar{z_0}, \bar{z_1})$, $(z_0, z_1) \to (z_1, z_0)$, $z_0 \to z_0 \cdot 2^j$, and $z_1 \to z_1 \cdot 2^k$ do not affect the relative error in z_2; consequently, this allows us to assume without loss of generality that conditions (1-4) and (6) are satisfied by the worst-case inputs. Using the results of Theorem 3, an exhaustive computer search (taking about five minutes in MAPLE on the second author's 1.4 GHz laptop) completes the proof. □

Corollary 5. *In IEEE 754 double-precision arithmetic ($\beta = 2$, $t = 53$, $\epsilon = 2^{-53}$), using "nearest even" rounding mode, the values*

$$a_0 = \frac{3}{4}(1 + 4\epsilon) \qquad b_0 = \frac{3}{4} \qquad a_1 = \frac{2}{3}(1 + 7\epsilon) \qquad b_1 = \frac{2}{3}(1 + 1\epsilon)$$

result in a relative error in z_2 of approximately $\epsilon\sqrt{5 - 96\epsilon} \approx \epsilon\sqrt{4.9999999999999893}$, and this is the worst possible provided that overflow, underflow, and denormals do not occur.

Proof. Straightforward computation for the values given establishes that

$$|z_2 - z_0 z_1|^2 = \epsilon^2(5 - 36\epsilon + \mathrm{O}(\epsilon^2))$$
$$|z_0 z_1|^2 = 1 + 12\epsilon + \mathrm{O}(\epsilon^2)$$

and the ratio of these provides the error as stated.

As in Corollary 4, an exhaustive search using the results of Theorem 3 (again, taking just a few minutes) completes the proof. □

For $\beta = 2$ and $t > 6$, the constructions given in Corollaries 4 and 5 for a_0, b_0, a_1, b_1 provide for even and odd t respectively relative errors of $\epsilon\sqrt{5 - 168\epsilon + \mathrm{O}(\epsilon^2)}$ and $\epsilon\sqrt{5 - 96\epsilon + \mathrm{O}(\epsilon^2)}$. We believe that these are the worst-case inputs for all sufficiently large t when $\beta = 2$.

4. A note on methods

The existence of this paper serves a strong demonstration of the power of experimental mathematics. The initial result — the upper bound of $\sqrt{5}\epsilon$ — was discovered experimentally seven years ago, on the basis of testing a few million random single-precision products.

Experimental methods became even more important when it came to the results concerning worst-case inputs. Here the approach taken was to perform an exhaustive search, taking several hours on the second author's laptop, of IEEE single-precision inputs, using only a few arguments from Theorem 1 to prune the search. Once the worst few sets of inputs had been enumerated, it became clear that they possessed the structure described in Theorem 3, and it was natural to conjecture that this structure would be satisfied by the worst-case inputs in any precision. As is common with such problems, once the required result was known, constructing a proof was fairly straightforward.

References

1. N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, 2002.
2. C. Percival, *Rapid multiplication modulo the sum and difference of highly composite numbers*, Math. Comp. **72** (2002), 387–395.

Appendix

The content of this report up to here will appear in *Mathematics of Computation*. This appendix gives some additional results that do not appear in the *Mathematics of Computation* article.

4.1. **A bit of history.** In an article entitled *Error Bounds for Polynomial Evaluation and Complex Arithmetic* published in the IMA Journal of Numerical Analysis (1986), Frank W. J. Olver proves the bound of $\sqrt{16/3}$ for complex multiplication, and follows with this remark:

> *"Indeed, in unpublished work R. P. Brent has demonstrated that in base* 2*, for example, [the error term] can be reduced to* $\sqrt{5}$ *. . . "*

Thus this "unpublished work" is now published, even if it took twenty years!

4.2. **Karatsuba Multiplication.** The $\sqrt{5}$ bound we obtained here holds for classical complex multiplication, with 4 real products. Karatsuba's algorithm enables one to perform a complex multiplication with only 3 real products, for example:

$$p_1 = (a+b)(c-d), \quad p_2 = ad, \quad p_3 = bc,$$

then $(a+bi)(c+di) = (p_1+p_2-p_3)+(p_2+p_3)i$.

It is natural to ask what the bound becomes in that case. The worst case could come when all the errors accumulate in the real part and there is no error in the imaginary part. Take for example $a \approx b \approx c \approx -d \approx 1$ and assume that all rounding errors are maximal and contribute to the same direction:

$$\begin{aligned}
a \oplus b &\approx a+b+2\epsilon \approx 2 \\
c \ominus d &\approx c-d+2\epsilon \approx 2 \\
p_1 &\approx (a+b)(c-d)+12\epsilon \approx 4 \\
p_2 &\approx ad+\epsilon \approx -1 \\
p_3 &\approx bc-\epsilon \approx 1 \\
p_2 \ominus p_3 &\approx ad-bc+4\epsilon \approx -2 \\
p_2 \oplus p_3 &\approx ad+bc \approx 0 \\
p_1 \oplus (p_2 \ominus p_3) &\approx ac-bd+16\epsilon \approx 2 \\
|z_2 - z_0 z_1| &\approx 16\epsilon \\
|z_0 z_1| &\approx 2
\end{aligned}$$

This may give a relative error of up to 8ϵ. Note that there are different ways to compute $p_1+p_2-p_3$, which may give different rounding errors: $p_1\oplus(p_2\ominus p_3)$ as above, or $(p_1\oplus p_2)\ominus p_3$, or even $(p_1 \ominus p_3) \oplus p_2$.

The worst case we found with the above way of computing $p_1+p_2-p_3$ is $z_1 = 260+278i$, $z_2 = 268-278i$, with a precision of 8 bits, which gives $a\oplus b = 536$, $c\ominus d = 544$, $p_1 = 290816$, $p_2 = -72192$, $p_3 = 74752$, $p_2 \oplus p_3 = 2560$, $p_2 \ominus p_3 = -147456$, and $p_1 \oplus (p_2 \ominus p_3) = 143360$. The rounded product is thus $z_2 = 143360+2560i$, whereas the exact one is $z_0z_1 = 146964+2224i$, with a relative error of about 6.30ϵ.

4.3. **Getting rid of the 2nd-order term ϵ in Case R1.** The bound we get in Case R1 is the following:

$$|z_2 - z_0z_1| < \epsilon\left(\sqrt{32/7}+2\epsilon\right)|z_0z_1|\,.$$

This is the only case among R1, …, R4 where we get a 2nd-order term ϵ^2. We show how to get rid of that term in the case $\beta = 2$.

Lemma 1. *Let $x > 0$ be rounded to a value y, with rounding to nearest. Then $|y-x| \le \frac{\epsilon}{1+\epsilon}x$.*

Proof. Remember $\epsilon = \frac{1}{2}\mathrm{ulp}(1) = \frac{1}{2}\beta^{1-t}$. Without loss of generality, one can assume $1 \le x < \beta$. Then $1 \le y \le \beta$, with an absolute error $|y - x| \le \epsilon$.

If $x < 1+\epsilon$, x is rounded to $y = 1$, thus the error is $x - 1 \le \frac{\epsilon}{1+\epsilon}x$. If $1+\epsilon \le x$, the error is at most $\epsilon \le \frac{\epsilon}{1+\epsilon}x$. ☐

We will now prove the following result, from which it follows $|z_2 - z_0z_1| \le \epsilon\sqrt{32/7}|z_0z_1|$.

Lemma 2. *Assume radix $\beta = 2$, precision $t \ge 2$, and we are in Case R1, i.e., $\mathrm{ulp}(b_0b_1) \le \mathrm{ulp}(a_0a_1) \le \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)$. Then:*

$$|\Re(z_2 - z_0z_1)| \le \epsilon \cdot (2a_0a_1 - b_0b_1).$$

Proof. Without loss of generality, we assume $1 \le a_0a_1 < 2$. Let $k \ge 0$ be the exponent difference between a_0a_1 and b_0b_1, i.e., $\mathrm{ulp}(a_0a_1) = 2^k\mathrm{ulp}(b_0b_1)$.

First assume $k = 0$; then $1 \le b_0 \otimes b_1 \le 2$. The only possibility that Case R1 holds is that $a_0 \otimes a_1 - b_0 \otimes b_1 = 1$, which implies $a_0 \otimes a_1 = 2$ and $b_0 \otimes b_1 = 1$. We thus have $2 - \epsilon \le a_0a_1 \le 2$ and $1 \le b_0b_1 \le 1 + \epsilon$. The real error is bounded by 2ϵ, and $2a_0a_1 - b_0b_1 \ge 2(2-\epsilon) - (1+\epsilon) \ge 3(1-\epsilon) \ge 2$. Thus $|\Re(z_2 - z_0z_1)| \le \epsilon(2a_0a_1 - b_0b_1)$.

Now assume $k \ge 1$, i.e., $2^{-k} \le b_0b_1 < 2^{1-k}$. Since $b_0 \otimes b_1 \ge 2^{-k}$, $a_0 \otimes a_1 - b_0 \otimes b_1$ is an integer multiple of $2^{1-k}\epsilon$, and is larger or equal to 1 by hypothesis. We distinguish three cases, depending on the position of $a_0 \otimes a_1 - b_0 \otimes b_1$ with respect to $1+\epsilon$ and $1+3\epsilon$.

First assume $1 + 3\epsilon \le a_0 \otimes a_1 - b_0 \otimes b_1$. The error when rounding a_0a_1 is bounded by:

$$\epsilon \le \frac{\epsilon}{1+3\epsilon}(a_0 \otimes a_1 - b_0 \otimes b_1).$$

The same bounds holds for the error when rounding $a_0 \otimes a_1 - b_0 \otimes b_1$, and that on b_0b_1 is bounded by $\frac{\epsilon}{1+\epsilon}b_0b_1$ from Lemma 1. The difference between $\epsilon(2a_0a_1 - b_0b_1)$ and the sum of those three bounds has the same sign as $4a_0a_1 - 7b_0b_1 + \epsilon(4a_0a_1 - 5b_0b_1)$, which is non-negative since $b_0b_1 \le \frac{1}{2}a_0a_1$. (Otherwise $b_0 \otimes b_1 \ge \frac{1}{2}a_0 \otimes a_1$, and $a_0 \otimes a_1 - b_0 \otimes b_1 \le \frac{1}{2}a_0 \otimes a_1 \le 1$.)

Now assume $1 + \epsilon < a_0 \otimes a_1 - b_0 \otimes b_1 < 1 + 3\epsilon$. Since $a_0 \otimes a_1 - b_0 \otimes b_1$ is an integer multiple of $2^{1-k}\epsilon$, we have

$$1 + \epsilon + 2^{1-k}\epsilon \le a_0 \otimes a_1 - b_0 \otimes b_1 \le 1 + 3\epsilon - 2^{1-k}\epsilon,$$

and $a_0 \otimes a_1 - b_0 \otimes b_1$ is rounded to $1 + 2\epsilon$, with an error at most $\epsilon(1 - 2^{1-k})$. The real error is thus bounded by

$$\frac{\epsilon + \epsilon(1 - 2^{1-k})}{1 + \epsilon}(a_0 \otimes a_1 - b_0 \otimes b_1) + \frac{\epsilon}{1 + \epsilon} b_0 b_1,$$

using again Lemma 1, and $a_0 \otimes a_1 - b_0 \otimes b_1 > 1 + \epsilon$. In the case $k = 1$, the difference between $\epsilon(2a_0a_1 - b_0b_1)$ and the error bound has the same sign as $(1+\epsilon)a_0a_1 - (1+2\epsilon)b_0b_1$, which is nonnegative since as above we have $b_0b_1 \le \frac{1}{2}a_0a_1$. For $k \ge 2$, using $b_0b_1 \le 2^{1-k}a_0a_1$, the difference between $\epsilon(2a_0a_1 - b_0b_1)$ and the error bound has the same sign as $1 - 2^{1-k} + \epsilon(2^{1-k} - 2) \ge 0$.

Now assume $1 \le a_0 \otimes a_1 - b_0 \otimes b_1 \le 1 + \epsilon$. In that case $a_0 \otimes a_1 - b_0 \otimes b_1$ is rounded down to 1. If a_0a_1 is rounded down and b_0b_1 is rounded up, then $a_0 \otimes a_1 - b_0 \otimes b_1 \le a_0a_1 - b_0b_1$, thus the total error is bounded by

$$2\epsilon(a_0a_1 - b_0b_1) + \epsilon b_0b_1 \le \epsilon(2a_0a_1 - b_0b_1).$$

If either a_0a_1 is rounded up or b_0b_1 is rounded down, the three errors have different signs, so the total error is bounded by the absolute sum of the two largest bounds, those for a_0a_1 and $a_0 \otimes a_1 - b_0 \otimes b_1$. Let $a_0 \otimes a_1 - b_0 \otimes b_1 = 1 + \eta$, with $\eta \le \epsilon$. The total error is thus bounded by $\epsilon + \eta$. On the other side, we have $a_0 \otimes a_1 = (a_0 \otimes a_1 - b_0 \otimes b_1) + b_0 \otimes b_1 \ge 1 + \eta + 2^{-k}$, and thus $a_0a_1 \ge 1 + \eta + 2^{-k} - \epsilon$. It follows $2a_0a_1 - b_0b_1 \ge 2(1 + \eta - \epsilon)$. Hence

$$\epsilon(2a_0a_1 - b_0b_1) - (\epsilon + \eta) = (1 - 2\epsilon)(\epsilon - \eta) \ge 0.$$

□

As a consequence, we can simplify Eq. (5) from Theorems 2 and 3 into:

$$\frac{|z_2 - z_0z_1|}{|z_0z_1|} > \epsilon\sqrt{5 - n\epsilon} > \epsilon \cdot \sqrt{1024/207}.$$

4.4. **Generic Worst Cases for Binary Precision t.** We prove here the statement at the end of Section 3.

Theorem 4. *For $\beta = 2$ and $t \ge 7$, the worst cases are:*

$$a_0 = \frac{3}{4}, b_0 = \frac{3}{4}(1 - 4\epsilon), a_1 = \frac{2}{3}(1 + 11\epsilon), b_1 = \frac{2}{3}(1 + 5\epsilon) \quad \textit{for } t \textit{ even},$$

and

$$a_0 = \frac{3}{4}(1 + 4\epsilon), b_0 = \frac{3}{4}, a_1 = \frac{2}{3}(1 + 7\epsilon), b_1 = \frac{2}{3}(1 + \epsilon) \quad \textit{for } t \textit{ odd}.$$

Proof. We first compute the relative error for those cases, and then prove this is the largest possible error for a given value of t.

First assume t even. We have $a_0a_1 = \frac{1}{2}(1+11\epsilon)$, which is rounded $\frac{1}{2}(1+12\epsilon)$; $b_0b_1 = \frac{1}{2}(1+\epsilon-20\epsilon^2)$ is rounded to $1/2$; $a_0b_1 = \frac{1}{2}(1+5\epsilon)$ is rounded to $\frac{1}{2}(1+4\epsilon)$; and $b_0a_1 = \frac{1}{2}(1+7\epsilon-44\epsilon^2)$ is rounded to $\frac{1}{2}(1+6\epsilon)$. Thus $a_0\otimes a_1 - b_0\otimes b_1 = 6\epsilon$ (exact since we are in Case R4), and $a_0\otimes b_1 + b_0\otimes a_1 = 1+5\epsilon$ is rounded to $1+4\epsilon$. We thus have $z_2 = 6\epsilon + i(1+4\epsilon)$, whereas $z_0z_1 = (5\epsilon+10\epsilon^2)+i(1+6\epsilon-22\epsilon^2)$ thus $z_2 - z_0z_1 = (\epsilon-10\epsilon^2)+i(-2\epsilon+22\epsilon^2) = \epsilon[(1-10\epsilon)+i(-2+22\epsilon)]$ which gives

$$\begin{aligned} |z_2 - z_0z_1|^2 &= \epsilon^2(5-108\epsilon+584\epsilon^2), \\ |z_0z_1|^2 &= 1+12\epsilon+17\epsilon^2-164\epsilon^3+584\epsilon^4 \end{aligned}$$

and taking the ratio gives $\epsilon^2(5-168\epsilon+2515\epsilon^2-\ldots)$.

Now assume t odd. We have $a_0a_1 = \frac{1}{2}(1+11\epsilon+28\epsilon^2)$, which is rounded $\frac{1}{2}(1+12\epsilon)$; $b_0b_1 = \frac{1}{2}(1+\epsilon)$ is rounded to $1/2$; $a_0b_1 = \frac{1}{2}(1+5\epsilon+4\epsilon^2)$ is rounded to $\frac{1}{2}(1+6\epsilon)$; and $b_0a_1 = \frac{1}{2}(1+7\epsilon)$ is rounded to $\frac{1}{2}(1+8\epsilon)$. Thus $a_0\otimes a_1 - b_0\otimes b_1 = 6\epsilon$ (exact since we are in Case R4), and $a_0\otimes b_1 + b_0\otimes a_1 = 1+7\epsilon$ is rounded to $1+8\epsilon$. We thus have $z_2 = 6\epsilon+i(1+8\epsilon)$, whereas $z_0z_1 = (5\epsilon+14\epsilon^2)+i(1+6\epsilon+2\epsilon^2)$ thus $z_2-z_0z_1 = (\epsilon-14\epsilon^2)+i(2\epsilon-2\epsilon^2) = \epsilon[(1-14\epsilon)+i(2-2\epsilon)]$ which gives

$$\begin{aligned} |z_2 - z_0z_1|^2 &= \epsilon^2(5-36\epsilon+200\epsilon^2), \\ |z_0z_1|^2 &= 1+12\epsilon+65\epsilon^2+164\epsilon^3+200\epsilon^4 \end{aligned}$$

and taking the ratio gives $\epsilon^2(5-96\epsilon+1027\epsilon^2-\ldots)$.

Now assume t even, $t \geq 20$. Since $\frac{1}{4}\epsilon^{-1/2} \geq 256$, we can apply Theorem 3 with $n = 168$.

Assume we apply Theorem 3 with $n = 97$. We first assume that none of α_0, β_0, α_1, β_1 is zero. We have $a_0a_1 = c_0c_1/(d_0d_1)(1+\alpha_0\epsilon)(1+\alpha_1\epsilon) = 1/2(1+\alpha_0\epsilon)(1+\alpha_1\epsilon) = 1/2[1+(\alpha_0+\alpha_1)\epsilon+(\alpha_0\alpha_1)\epsilon^2]$.

The 2nd order term $(\alpha_0\alpha_1)\epsilon^2 < n\epsilon^2 \leq 97\epsilon^2 < \epsilon/2$. Thus a_0a_1 rounds:

- to $1/2[1+(\alpha_0+\alpha_1)\epsilon]$ if $\alpha_0+\alpha_1$ is even or ≤ 0,
- to $1/2[1+(\alpha_0+\alpha_1+\text{sign}(\alpha_0\alpha_1))\epsilon]$ otherwise, where $\text{sign}(\alpha_0\alpha_1) = \pm 1$.

Similarly, b_0b_1 rounds to:

- $1/2[1+(\beta_0+\beta_1)\epsilon]$ if $\beta_0+\beta_1$ is even or ≤ 0,
- $1/2[1+(\beta_0+\beta_1+\text{sign}(\beta_0\beta_1))\epsilon]$ otherwise.

Thus $a_0a_1 - b_0b_1 = 1/2(\alpha_0+\alpha_1-\beta_0-\beta_1)\epsilon$ if $\alpha_0+\alpha_1$ and $\beta_0+\beta_1$ are both even or ≤ 0, or three other cases, with a general form of $1/2(\alpha_0+\alpha_1-\beta_0-\beta_1+c)\epsilon$ where $-2 \leq c \leq 2$; and $a_0a_1 - b_0b_1$ rounds to itself since we are in Case R4.

$a_0b_1 = 1/2(1+\alpha_0\epsilon)(1+\beta_1\epsilon)$ rounds:

- to $1/2[1+(\alpha_0+\beta_1)\epsilon]$ if $\alpha_0+\beta_1$ is even or ≤ 0,
- to $1/2[1+(\alpha_0+\beta_1+\text{sign}(\alpha_0\beta_1))\epsilon]$ otherwise.

$b_0a_1 = 1/2(1+\beta_0\epsilon)(1+\alpha_1\epsilon)$ rounds

- to $1/2[1+(\beta_0+\alpha_1)\epsilon]$ if $\beta_0+\alpha_1$ is even or ≤ 0,

- to $1/2[1 + (\beta_0 + \alpha_1 + \mathrm{sign}(\beta_0\alpha_1))\epsilon]$ otherwise.

Thus $a_0 \otimes b_1 + b_0 \otimes a_1 = 1/2[2 + (\alpha_0 + \beta_1 + \beta_0 + \alpha_1 + d')\epsilon]$ where $-2 \le d' \le 2$.

Since $\min(\alpha_0, \beta_0) + \min(\alpha_1, \beta_1) \ge 0$, we have

$$\alpha_0 + \beta_1 + \beta_0 + \alpha_1 \ge 2[\min(\alpha_0, \beta_0) + \min(\alpha_1, \beta_1)] \ge 0.$$

Moreover with the condition $\alpha_0 \ne \beta_0$ and $\alpha_1 \ne \beta_1$, we have

$$\alpha_0 + \beta_1 + \beta_0 + \alpha_1 \ge 2[\min(\alpha_0, \beta_0) + \min(\alpha_1, \beta_1)] + 2.$$

Thus $\alpha_0 + \beta_1 + \beta_0 + \alpha_1 + d' \ge 0$.

Since $a_0 \otimes b_1 + b_0 \otimes a_1$ is a multiple of 2^{-p-1} and it rounds to a value ≥ 1, i.e., a multiple of 2^{-p+1}, the rounding error is of the form $k2^{-p-1}$ with $-2 \le k \le 2$. Thus the total rounding error on the imaginary part is of the form $1/2d\epsilon$ where $-4 \le d \le 4$.

Thus we have:

$$\begin{aligned} \Re(z_2) &= 1/2(\alpha_0 + \alpha_1 - \beta_0 - \beta_1 + c)\epsilon \\ \Im(z_2) &= 1/2[2 + (\alpha_0 + \beta_1 + \beta_0 + \alpha_1 + d)\epsilon] \end{aligned}$$

where $-2 \le c \le 2$ and $-4 \le d \le 4$.

The exact product $z_0 z_1$ has:

$$\begin{aligned} \Re(z_0 z_1) &= 1/2(1 + \alpha_0\epsilon)(1 + \alpha_1\epsilon) - 1/2(1 + \beta_0\epsilon)(1 + \beta_1\epsilon) \\ &= 1/2[(\alpha_0 + \alpha_1 - \beta_0 - \beta_1)\epsilon + (\alpha_0\alpha_1 - \beta_0\beta_1)\epsilon^2] \\ \Im(z_0 z_1) &= 1/2[2 + (\alpha_0 + \alpha_1 + \beta_0 + \beta_1)\epsilon + (\alpha_0\beta_1 + \beta_0\alpha_1)\epsilon^2] \end{aligned}$$

The difference is then:

$$\begin{aligned} \Re(z_2 - z_0 z_1) &= 1/2[c\epsilon - (\alpha_0\alpha_1 - \beta_0\beta_1)\epsilon^2] \\ \Im(z_2 - z_0 z_1) &= 1/2[d\epsilon - (\alpha_0\beta_1 + \beta_0\alpha_1)\epsilon^2] \end{aligned}$$

with $|c| \le 2$ and $|d| \le 4$.

Thus neglecting 2nd order terms we have $|z_2 - z_0 z_1|^2 \approx 1/4[c^2 + d^2]\epsilon^2$, and $|z_0 z_1|^2 \approx 1$, so the ratio is $\approx 1/4[c^2 + d^2]\epsilon^2$. To get an error near $5\epsilon^2$, we need $|c| = 2$ and $|d| = 4$, which implies that:

- $\alpha_0 + \alpha_1$ is odd and > 0,
- $\beta_0 + \beta_1$ is odd and > 0,
- $\mathrm{sign}(\alpha_0\alpha_1) = -\mathrm{sign}(\beta_0\beta_1)$ (if say α_0 is zero, then α_1 is odd, and the even-rule implies that we replace $\mathrm{sign}(\alpha_0\alpha_1)$ by 1 if $\alpha_1 + 1$ is a multiple of 4, and -1 otherwise),
- $\alpha_0 + \beta_1$ is odd and > 0,
- $\beta_0 + \alpha_1$ is odd and > 0,
- $\mathrm{sign}(\alpha_0\beta_1) = \mathrm{sign}(\beta_0\alpha_1)$, with the same convention as above in case one is zero.

If none is zero, the conditions $\mathrm{sign}(\alpha_0\alpha_1) = -\mathrm{sign}(\beta_0\beta_1)$ and $\mathrm{sign}(\alpha_0\beta_1) = \mathrm{sign}(\beta_0\alpha_1)$ are incompatible.

Thus we can assume without loss of generality that $\alpha_0 = 0$, which gives:

- α_1 is odd and > 0, say $4k + l_1$ with $l_1 = 1$ or 3,
- $\beta_0 + \beta_1$ is odd and > 0,

- $\text{sign}(l_1 - 2) = -\text{sign}(\beta_0\beta_1)$,
- β_1 is odd and > 0, say $4j + m_1$ with $m_1 = 1$ or 3,
- $\beta_0 + \alpha_1$ is odd and > 0,
- $\text{sign}(m_1 - 2) = \text{sign}(\beta_0\alpha_1)$.

This implies:

- $m_1 \neq l_1$, thus $\alpha_1 - \beta_1 = 2 \bmod 4$,
- β_0 even, > 0 for $l_1 = 1$, < 0 for $l_1 = 3$,
- $\alpha_1 > 0$, odd,
- $\beta_1 > 0$, odd.

We then get:

$$\begin{aligned}\Re(z_2 - z_0z_1) &= 1/2[c\epsilon + \beta_0\beta_1\epsilon^2]\\ \Im(z_2 - z_0z_1) &= 1/2[d\epsilon - \beta_0\alpha_1\epsilon^2]\end{aligned}$$

with $|c| = 2$ and $|d| = 4$.

Since $c = \text{sign}(\alpha\alpha_1) - \text{sign}(\beta_0\beta_1)$, and $\text{sign}(d) = \text{sign}(\beta_0\alpha_1)$, we can write:

$$\begin{aligned}\Re(z_2 - z_0z_1) &= 1/2c[\epsilon - 1/2|\beta_0\beta_1|\epsilon^2]\\ \Im(z_2 - z_0z_1) &= 1/2d[\epsilon - 1/4|\beta_0\alpha_1|\epsilon^2]\end{aligned}$$

with $|c| = 2$ and $|d| = 4$.

Thus $|z_2 - z_0z_1|^2 = [\epsilon - 1/2|\beta_0\beta_1|\epsilon^2]^2 + [2\epsilon - 1/2|\beta_0\alpha_1|\epsilon^2]^2 = \epsilon^2[5 - (|\beta_0\beta_1| + 2|\beta_0\alpha_1|)\epsilon + 1/4(|\beta_0\beta_1|^2 + |\beta_0\alpha_1|^2)\epsilon^2]$.

Now $|z_0|^2 = (c_0/d_0)^2[(1 + \alpha_0\epsilon)^2 + (1 + \beta_0\epsilon)^2] = (c_0/d_0)^2[2 + 2(\alpha_0 + \beta_0)\epsilon + \alpha_0^2\beta_0^2\epsilon^2]$; $|z_1|^2 = (c_1/d_1)^2[(1 + \alpha_1\epsilon)^2 + (1 + \beta_1\epsilon)^2] = (c_1/d_1)^2[2 + 2(\alpha_1 + \beta_1)\epsilon + \alpha_1^2\beta_1^2\epsilon^2]$.

Thus neglecting 2nd order terms and using $\alpha_0 = 0$:

$$|z_0z_1|^2 \approx 1 + (\beta_0 + \alpha_1 + \beta_1)\epsilon.$$

Finally $|z_2 - z_0z_1|^2/|z_0z_1|^2$ is (neglecting 2nd order terms):

$$\epsilon^2\frac{5 - (|\beta_0\beta_1| + 2|\beta_0\alpha_1|)\epsilon}{1 + (\beta_0 + \alpha_1 + \beta_1)\epsilon} \approx \epsilon^2[5 - (|\beta_0\beta_1| + 2|\beta_0\alpha_1| + 5(\beta_0 + \alpha_1 + \beta_1))\epsilon].$$

Since $\alpha_1, \beta_1 > 0$, this simplifies to:

$$\epsilon^2[5 - (|\beta_0|(\beta_1 + 2\alpha_1) + 5(\beta_0 + \alpha_1 + \beta_1))\epsilon],$$

thus we are looking for the minimum of:

$$E = |\beta_0|(\beta_1 + 2\alpha_1) + 5(\beta_0 + \alpha_1 + \beta_1)$$

under the contraints [remember $\alpha_0 = 0$]:

- (a_1) either $\alpha_1 = 4k + 1$ for $k \geq 0$, $\beta_1 = 4j + 3$ for $j \geq 0$, β_0 even > 0,
- (a_2) or $\alpha_1 = 4k + 3$ for $k \geq 0$, $\beta_1 = 4j + 1$ for $j \geq 0$, β_0 even < 0,
- (b) and $\beta_0 + \beta_1 > 0$,
- (c) and $\beta_0 + \alpha_1 > 0$.

Since $\alpha_0 = 0$, and $\alpha_0 \equiv 2^t \bmod d_0$, this implies that d_0 is a power of two.

If $d_0 = 2$, then $c_0 = 1$ (because $d_0/2 \le c_0 \le d_0$ and $\gcd(c_0, d_0) = 1$), thus $c_1 = d_1$ (because $2c_0c_1 = d_0d_1$), which contradicts $\gcd(c_1, d_1) = 1$.

Thus d_0 is at least 4. Now the constraint $\beta_0 \equiv 0 \bmod d_0$, together with $\beta_0 \neq \alpha_0$, implies that $|\beta_0| \ge 4$.

d_1 cannot be 2, since otherwise c_0 and c_1 would be both powers of 2. (More generally d_1 cannot be divisible by 2, since 4 divides d_0, then 2 divides c_1, and $\gcd(c_1, d_1) = 1$.) Thus d_1 is at least 3. Now the constraint $\alpha_1 = \beta_1 \bmod d_1$ implies that α_1 and β_1 differ by at least $d_1 \ge 3$.

For $\beta_0 > 0$, the expression E above simplifies to:

$$\begin{aligned} E &= \beta_0(\beta_1 + 2\alpha_1) + 5(\beta_0 + \alpha_1 + \beta_1) \\ &= \beta_0(4j + 3 + 2(4k + 1)) + 5(\beta_0 + 4k + 1 + 4j + 3) \\ &= (4j + 8k + 10)\beta_0 + (20j + 20k + 20). \end{aligned}$$

Since $\beta_0 \ge 4$, this gives $E \ge 36j + 52k + 60$. Since $\alpha_1 = 4k + 1$ and $\beta_1 = 4j + 3$ must differ by at least 3, we cannot have both $k = j = 0$, thus the smallest value is obtained for $j = 1$ and $k = 0$, with $E \ge 96$. For that case to apply, since $\alpha_1 \equiv -2^t \bmod d_1$, we need for $d_1 = 3$ that is t odd. (Otherwise if $d_1 > 3$, necessarily $d_1 \ge 5$, thus α_1 and β_1 differ by at least 5. Then $j = k = 0$ does not work; same for $j = 1, k = 0$ which would give $\alpha_1 = 1, \beta_1 = 7$ which would imply $d_1 = 3$; same for $j = 0, k = 1$ which would give $\alpha_1 = 5, \beta_1 = 3$; same for $j = k = 1$ which would give $\alpha_1 = 5, \beta_1 = 7$; thus the smallest solution would be $j = 2, k = 1$ which gives $E \ge 184$.)

For $\beta_0 < 0$, say $\beta_0 = -i$, E simplifies to:

$$\begin{aligned} E &= i(\beta_1 + 2\alpha_1) + 5(-i + \alpha_1 + \beta_1) \\ &= i(4j + 1 + 2(4k + 3)) + 5(-i + 4k + 3 + 4j + 1) \\ &= (4j + 8k + 2)i + (20j + 20k + 20). \end{aligned}$$

Since $|\beta_0| \ge 4$, i.e., $i \ge 4$, this gives $E \ge 36j + 52k + 28$. Since $\alpha_1 = 4k + 3$ and $\beta_1 = 4j + 1$ must differ by at least 3, and we must have $\min(\alpha_0, \beta_0) + \min(\alpha_1, \beta_1) \ge 0$, i.e. $\min(4k + 3, 4j + 1) \ge 4$, we need $j, k \ge 1$. We cannot have $j = k = 1$ since this would give $(\alpha_1, \beta_1) = (7, 5)$ which do not differ by ≥ 3, nor $j = 2, k = 1$ which would give $(\alpha_1, \beta_1) = (7, 9)$, thus the smallest possible value is $j = 1, k = 2$ which gives $E \ge 168$. $\square$

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399