MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

ERROR BOUNDS ON COMPLEX FLOATING-POINT MULTIPLICATION

RICHARD BRENT, COLIN PERCIVAL, AND PAUL ZIMMERMANN

In memory of Erin Brent (1947–2005)

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.

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 $\mathrm{ulp}(0) = 0$. The notations $x \oplus y$, $x \ominus y$, and $x \otimes y$ represent rounded-to-nearest 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 $\epsilon\sqrt{5} = \frac{1}{2}\beta^{1-t}\sqrt{5}$.

©1997 American Mathematical Society

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

$$0 \le a_0, b_0, a_1, b_1 \tag{1}$$
$$b_0 b_1 \le a_0 a_1 \tag{2}$$

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 $\frac{1}{2}\text{ulp}(x)$ 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:

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

and consider two cases:

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

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

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

and therefore

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

However, $\beta^t \text{ulp}(a_0 b_1 + b_0 a_1)$ 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: $\text{ulp}(a_0 \otimes b_1 + b_0 \otimes a_1) \le \text{ulp}(a_0 b_1 + b_0 a_1)$

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

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

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

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

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 $\beta = 2$):

$$\begin{aligned}
&\mathrm{ulp}(b_0b_1) \le \mathrm{ulp}(a_0a_1) \le \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) && \text{e.g., } z_0 = z_1 = 0.8 + 0.1i\\
&\mathrm{ulp}(b_0b_1) < \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0a_1) && \text{e.g., } z_0 = z_1 = 0.8 + 0.4i\\
&\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \le \mathrm{ulp}(b_0b_1) < \mathrm{ulp}(a_0a_1) && \text{e.g., } z_0 = z_1 = 0.8 + 0.7i\\
&\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(b_0b_1) = \mathrm{ulp}(a_0a_1) && \text{e.g., } z_0 = z_1 = 0.8 + 0.8i.
\end{aligned}$$

Since we have assumed that $b_0b_1 \le a_0a_1$, we know that $\mathrm{ulp}(b_0b_1) \le \mathrm{ulp}(a_0a_1)$, 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: $\mathrm{ulp}(b_0b_1) \le \mathrm{ulp}(a_0a_1) \le \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)$

Note that the right inequality can only be strict if $a_0 \otimes a_1$ rounds up to a power of β and $b_0b_1 = 0$.

We observe that

$$a_0 \otimes a_1 - b_0 \otimes b_1 < a_0a_1 - b_0b_1 + \epsilon \cdot (a_0a_1 + b_0b_1)$$

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

$$\begin{aligned}
|\Re(z_2 - z_0z_1)| &\le |a_0 \otimes a_1 - a_0a_1| + |b_0 \otimes b_1 - b_0b_1|\\
&\qquad + |((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) - (a_0 \otimes a_1 - b_0 \otimes b_1)|\\
&\le \frac{1}{2}\mathrm{ulp}(a_0a_1) + \frac{1}{2}\mathrm{ulp}(b_0b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)\\
&\le \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) + \frac{1}{2}\mathrm{ulp}(b_0b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)\\
&< 2\epsilon \cdot (a_0 \otimes a_1 - b_0 \otimes b_1) + \epsilon \cdot (b_0b_1)\\
&< \epsilon \cdot (2a_0a_1 - b_0b_1) + \epsilon^2 \cdot (2a_0a_1 + 2b_0b_1).
\end{aligned}$$

Applying the triangle inequality, we now observe that

$$\begin{aligned}
|z_2 - z_0z_1| &= \sqrt{\Re(z_2 - z_0z_1)^2 + \Im(z_2 - z_0z_1)^2}\\
&< \epsilon\sqrt{(2a_0a_1 - b_0b_1)^2 + (2a_0b_1 + 2b_0a_1)^2} + \epsilon^2 \cdot (2a_0a_1 + 2b_0b_1)\\
&\le \epsilon\sqrt{\frac{32}{7}|z_0z_1|^2 - \frac{4}{7}(a_0b_1 - b_0a_1)^2 - \frac{1}{7}(2a_0a_1 - 5b_0b_1)^2} + 2\epsilon^2|z_0z_1|\\
&\le \epsilon\left(\sqrt{32/7} + 2\epsilon\right)|z_0z_1| < \epsilon\sqrt{5}\,|z_0z_1|
\end{aligned}$$

as required.

Case R2: $\mathrm{ulp}(b_0b_1) < \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0a_1)$

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

$$\begin{aligned}
|\Re(z_2 - z_0z_1)| &\le \frac{1}{2}\mathrm{ulp}(a_0a_1) + \frac{1}{2}\mathrm{ulp}(b_0b_1) + \frac{1}{2}\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)\\
&\le \frac{7}{8}\mathrm{ulp}(a_0a_1)\\
&\le \epsilon \cdot \left(\frac{7}{4}a_0a_1\right)
\end{aligned}$$

and therefore

$$\begin{aligned}|z_2 - z_0z_1| &= \sqrt{\Re(z_2 - z_0z_1)^2 + \Im(z_2 - z_0z_1)^2} \\ &< \epsilon\sqrt{\left(\frac{7}{4}a_0a_1\right)^2 + (2a_0b_1 + 2b_0a_1)^2} \\ &= \epsilon\sqrt{\frac{1024}{207}\,|z_0z_1|^2 - \frac{196}{207}(a_0b_1 - b_0a_1)^2 - \frac{1}{3312}(79a_0a_1 - 128b_0b_1)^2} \\ &\le \epsilon\sqrt{1024/207}\,|z_0z_1| < \epsilon\sqrt{5}\,|z_0z_1|\end{aligned}$$

as required.

Case R3: $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \le \mathrm{ulp}(b_0b_1) < \mathrm{ulp}(a_0a_1)$

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 $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) \le \mathrm{ulp}(b_0b_1) \le \mathrm{ulp}(b_0 \otimes b_1)$ and $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(a_0a_1) \le \mathrm{ulp}(a_0 \otimes a_1)$. Also,

$$\mathrm{ulp}(b_0b_1) \le \frac{1}{\beta}\mathrm{ulp}(a_0a_1) \le \frac{1}{2}\mathrm{ulp}(a_0a_1)$$

so we have

$$\begin{aligned}|\Re(z_2 - z_0z_1)| &\le \frac{1}{2}\mathrm{ulp}(a_0a_1) + \frac{1}{2}\mathrm{ulp}(b_0b_1) \\ &\le \frac{3}{4}\mathrm{ulp}(a_0a_1) \\ &\le \epsilon \cdot \left(\frac{3}{2}a_0a_1\right)\end{aligned}$$

and consequently

$$\begin{aligned}|z_2 - z_0z_1| &= \sqrt{\Re(z_2 - z_0z_1)^2 + \Im(z_2 - z_0z_1)^2} \\ &< \epsilon\sqrt{\left(\frac{3}{2}a_0a_1\right)^2 + (2a_0b_1 + 2b_0a_1)^2} \\ &= \epsilon\sqrt{\frac{256}{55}\,|z_0z_1|^2 - \frac{36}{55}(a_0b_1 - b_0a_1)^2 - \frac{1}{220}(23a_0a_1 - 32b_0b_1)^2} \\ &\le \epsilon\sqrt{256/55}\,|z_0z_1| < \epsilon\sqrt{5}\,|z_0z_1|\end{aligned}$$

as required.

Case R4: $\mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1) < \mathrm{ulp}(b_0b_1) = \mathrm{ulp}(a_0a_1)$

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

$$\begin{aligned}|\Re(z_2 - z_0z_1)| &\le |a_0 \otimes a_1 - a_0a_1| + |b_0 \otimes b_1 - b_0b_1| \\ &< \epsilon \cdot (a_0a_1 + b_0b_1)\end{aligned}$$

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

$$\begin{aligned}
a_0 a_1 &= 1/2 + (j_{aa} + 1/2)\epsilon + k_{aa}\epsilon^2 \\
a_0 b_1 &= 1/2 + (j_{ab} + 1/2)\epsilon + k_{ab}\epsilon^2 \\
b_0 a_1 &= 1/2 + (j_{ba} + 1/2)\epsilon + k_{ba}\epsilon^2 \\
b_0 b_1 &= 1/2 + (j_{bb} + 1/2)\epsilon + k_{bb}\epsilon^2
\end{aligned}$$

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

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

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

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

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 $\mathrm{ulp}(b_0 b_1) = \mathrm{ulp}(a_0 a_1)$. From inequalities (2) and (4) above, this implies that

$$1/2 \le b_0 b_1 \le a_0 a_1 < 1$$
$$|\Re(z_2 - z_0 z_1)| \le |a_0 \otimes a_1 - a_0 a_1| + |b_0 \otimes b_1 - b_0 b_1| \le \epsilon.$$

We can now obtain lower bounds on $|z_0 z_1|$ and $|z_2 - z_0 z_1|$, using the fact that $(a_0 a_1)(b_0 b_1) = (a_0 b_1)(b_0 a_1)$:

$$\begin{aligned} |z_0 z_1|^2 &= \left(a_0^2 + b_0^2\right)\left(a_1^2 + b_1^2\right) \\ &= (a_0 a_1)^2 + (a_0 b_1)^2 + (b_0 a_1)^2 + (b_0 b_1)^2 \\ &\ge (1/2)^2 + (a_0 b_1)^2 + \frac{(1/2)^4}{(a_0 b_1)^2} + (1/2)^2 \ge 1 \\ |z_2 - z_0 z_1|^2 &> |z_0 z_1|^2 \epsilon^2 (5 - n\epsilon) \ge \epsilon^2 (5 - n\epsilon) \end{aligned}$$

as well as an upper bound on $|z_0 z_1|$:

$$\begin{aligned} |z_0 z_1|^2 \cdot \frac{1024\epsilon^2}{207} &< |z_2 - z_0 z_1|^2 \\ &= |\Re(z_2 - z_0 z_1)|^2 + |\Im(z_2 - z_0 z_1)|^2 \\ &< \epsilon^2 + (\epsilon \cdot (2a_0 b_1 + 2b_0 a_1))^2 \\ &\le \epsilon^2 + 4\epsilon^2 |z_0 z_1|^2 \\ |z_0 z_1|^2 &< \frac{207}{196} \end{aligned}$$

We now note that

$$\begin{aligned} (a_0 b_1)^2 &\le |z_0 z_1|^2 - (a_0 a_1)^2 - (b_0 b_1)^2 \\ &\le \frac{207}{196} - \frac{1}{4} - \frac{1}{4} = \frac{109}{196} \end{aligned}$$

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

$$|a_0 \otimes b_1 - a_0 b_1| \le \epsilon/2$$
$$|b_0 \otimes a_1 - b_0 a_1| \le \epsilon/2$$
$$|((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)| \le \epsilon$$
$$|\Im(z_2 - z_0 z_1)| \le \epsilon/2 + \epsilon/2 + \epsilon = 2\epsilon$$

which allows us to place upper bounds on $|z_2 - z_0 z_1|$ and $|z_0 z_1|$:

$$|z_2 - z_0 z_1|^2 = |\Re(z_2 - z_0 z_1)|^2 + |\Im(z_2 - z_0 z_1)|^2 \le (\epsilon)^2 + (2\epsilon)^2 = 5\epsilon^2$$
$$|z_0 z_1|^2 < \frac{|z_2 - z_0 z_1|^2}{\epsilon^2 (5 - n\epsilon)} \le \frac{5}{5 - n\epsilon}.$$

Combining the known lower bound $\epsilon^2(5 - n\epsilon)$ for $|z_2 - z_0 z_1|^2$ with the upper bounds on the error contributed by each individual rounding step, we find that

$$\epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon < |a_0 \otimes a_1 - a_0 a_1| \le \epsilon/2$$

$$\epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon < |b_0 \otimes b_1 - b_0 b_1| \leq \epsilon/2$$
$$\epsilon/2 - (2 - \sqrt{4 - n\epsilon})\epsilon < |a_0 \otimes b_1 - a_0 b_1| \leq \epsilon/2$$
$$\epsilon/2 - (2 - \sqrt{4 - n\epsilon})\epsilon < |b_0 \otimes a_1 - b_0 a_1| \leq \epsilon/2$$

and similarly, by combining the upper bound on $|z_0 z_1|^2$ with lower bound of $1/2$ for each pairwise product, we obtain

$$1/2 \leq b_0 b_1 \leq a_0 a_1 \leq \sqrt{\frac{5}{5 - n\epsilon} - \frac{3}{4}} = \sqrt{\frac{5 + 3n\epsilon}{20 - 4n\epsilon}}$$
$$1/2 \leq b_0 a_1 \leq a_0 b_1 \leq \sqrt{\frac{5}{5 - n\epsilon} - \frac{3}{4}} = \sqrt{\frac{5 + 3n\epsilon}{20 - 4n\epsilon}}.$$

Now consider the possible values for $a_0 a_1$ which satisfy these restrictions. Since it is the product of two values which are expressible using t digits of significand, $a_0 a_1$ can be exactly represented using $2t$ digits of significand; but since $1/2 \leq a_0 a_1 < 1$, this implies that $a_0 a_1$ is an integer multiple of ϵ^2. There is therefore at least one pair of integers j_{aa}, k_{aa} with $0 \leq j_{aa} < \epsilon^{-1}/2$, $|k_{aa}| \leq \epsilon^{-1}/2$ for which

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

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

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

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

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

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

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

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

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

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

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

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

$$a_0 = 1/2 + A\epsilon$$
$$a_1 = 1 + 2B\epsilon$$

for some $0 \leq A, B < (2\epsilon)^{-1}$. 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 $0 \leq j_{aa} < n/4$, $|k_{aa}| < n$.

As a result, we must have $A + B \leq n/4 \leq 1/2 \cdot \epsilon^{-1/2}$, and since $0 \leq A, B$ this implies $0 \leq 2AB\epsilon^2 \leq \epsilon/8$. 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 $a_0a_1 > 1/2$ 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 $n \leq \epsilon^{-1/2}$ and $\epsilon \leq 2^{-6}$. 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 $|a_1 - b_1| \leq \frac{n}{2}\epsilon$. 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

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

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 $n \leq \frac{1}{4}\epsilon^{-1/2}$. 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,

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

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

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*

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

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

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

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

$$1/2 \leq a_0 < 1 \tag{6}$$

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*

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

for some $n < \frac{1}{4}\epsilon^{-1/2}$ *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

$$
\begin{aligned}
a_0 &= \frac{a_0(a_1-b_1)}{a_1-b_1} = \frac{c_0g_0\epsilon + e_0g_0\epsilon^2}{d_0g_0\epsilon} = \frac{c_0+e_0\epsilon}{d_0}\\
a_1 &= \frac{a_1(a_0-b_0)}{a_0-b_0} = \frac{c_1g_1\epsilon + e_1g_1\epsilon^2}{d_1g_1\epsilon} = \frac{c_1+e_1\epsilon}{d_1}
\end{aligned}
$$

and therefore

$$
\begin{aligned}
\frac{1}{2} + \left(j_{aa}+\frac{1}{2}\right)\epsilon + k_{aa}\epsilon^2 &= a_0a_1\\
&= \frac{c_0c_1}{d_0d_1} + \frac{c_0e_1+e_0c_1}{d_0d_1}\epsilon + \frac{e_0e_1}{d_0d_1}\epsilon^2\\
&= \frac{1}{2} + \frac{c_0e_1+e_0c_1}{d_0d_1}\epsilon + k_{aa}\epsilon^2
\end{aligned}
$$

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

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

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 largest possible relative error for for IEEE 754 single-precision inputs provided that overflow, underflow, and denormals do not occur.

Proof. Straightforward computation for the values given establishes that

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

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

and the ratio of these provides the error as stated.

To prove that this is the largest possible relative error for IEEE single-precision inputs, 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. □

[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 $\epsilon^{-1} + 5 \equiv \epsilon^{-1} + 11 \equiv 0 \pmod 3$.

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. $\square$

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.

Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
E-mail address: complex@rpbrent.com

IRMACS Centre, Simon Fraser University, Burnaby, BC, Canada
E-mail address: cperciva@irmacs.sfu.ca

INRIA Lorraine/LORIA, 615 rue du Jardin Botanique, F-54602 Villers-lès-Nancy Cedex, France
E-mail address: zimmerma@loria.fr