Error bounds on complex floating-point multiplication

Paul Zimmermann, INRIA/LORIA, Nancy, France

(joint work with Richard Brent and Colin Percival)

December 14th, 2005

Notations

- t-digit base β f-p arithmetic
- no underflow/overflow
- all roundings to nearest (even)

$\circ(x)$ is the rounding to nearest of x

$$a \oplus b = \circ(a+b), \qquad a \otimes b = \circ(a \cdot b)$$

$\mathrm{ulp}(x)$ is the “unit in last place” of x:

$$\beta^{t-1}\mathrm{ulp}(x) \leq |x| < \beta^{t}\mathrm{ulp}(x)$$

Complex Multiplication

$$z_0 = a_0 + b_0 i, \qquad z_1 = a_1 + b_1 i$$

$$z_0 z_1 = (a_0 a_1 - b_0 b_1) + (a_0 b_1 + b_0 a_1)i$$

$$z_2 = ((a_0 \otimes a_1) \ominus (b_0 \otimes b_1)) + ((a_0 \otimes b_1) \oplus (b_0 \otimes a_1))i$$

What is the largest relative error?

$$\frac{|z_2 - z_0 z_1|}{|z_0 z_1|}$$

Plan

- previous work
- proof of the $\sqrt{5}$ bound
- worst-cases for base $\beta = 2$
- future work

References:

Rapid multiplication modulo the sum and difference of highly composite numbers, C. Percival, Math. of Comp., 2003.

Error bounds on complex floating-point multiplication, R. Brent, C. Percival, P. Z., submitted to Math. of Comp., 2005, 12 pages.

Higham's Bound

N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, 2002.

$$|z_2 - z_0 z_1| \leq \epsilon\sqrt{8}|z_0 z_1|$$

where $\epsilon = \frac{1}{2}\mathrm{ulp}(1) = \frac{1}{2}\beta^{1-t}$.

Higham's Bound (sketch)

$$|\mathcal{I}(z_2 - z_0 z_1)| \leq 2\epsilon \cdot (a_0 b_1 + b_0 a_1)$$

$$|\mathcal{R}(z_2 - z_0 z_1)| \leq 2\epsilon \cdot (a_0 a_1) + O(\epsilon^2)$$

$$\begin{aligned}\sqrt{\mathcal{R}^2 + \mathcal{I}^2} &\leq \epsilon\sqrt{4(a_0 a_1)^2 + 4(a_0 b_1 + b_0 a_1)^2} + O(\epsilon^2)\\ &\leq \epsilon\sqrt{8(a_0 a_1 - b_0 b_1)^2 + 8(a_0 b_1 + b_0 a_1)^2} + O(\epsilon^2)\end{aligned}$$

A Maple Proof

```
> e := 8*(a0*b1+a1*b0)^2 + 8*(a0*a1-b0*b1)^2
       - 4*(a0*b1+a1*b0)^2 - 4*(a0*a1)^2:
> expand(e);
    2    2                        2    2        2    2
4 a0   b1   - 8 a0 b1 a1 b0 + 4 a1   b0   + 4 a0   a1

             2    2
     + 8 b0   b1
```

This is:

$$4(a_0b_1 - a_1b_0)^2 + 4a_0^2a_1^2 + 8b_0^2b_1^2$$

A 10-line but Wrong Proof

[. . .] we observe that if $2b_0b_1 \geq a_0a_1$ *there is no error introduced by the subtraction [6]; further, if* $2b_0b_1 < a_0a_1$ *then the total error introduced in computing* b_0b_1 *and performing the subtraction is bounded by* $\epsilon(a_0a_1 - b_0b_1)$*.*

$\beta = 2$, $t = 5$, $z_0 = 28 + 17i$, $z_1 = 31 + 18i$

Total error on b_0b_1 and subtraction: $16 - (-2) = 18$

$\epsilon(a_0a_1 - b_0b_1) = 17.5625$

Our Main Result

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*

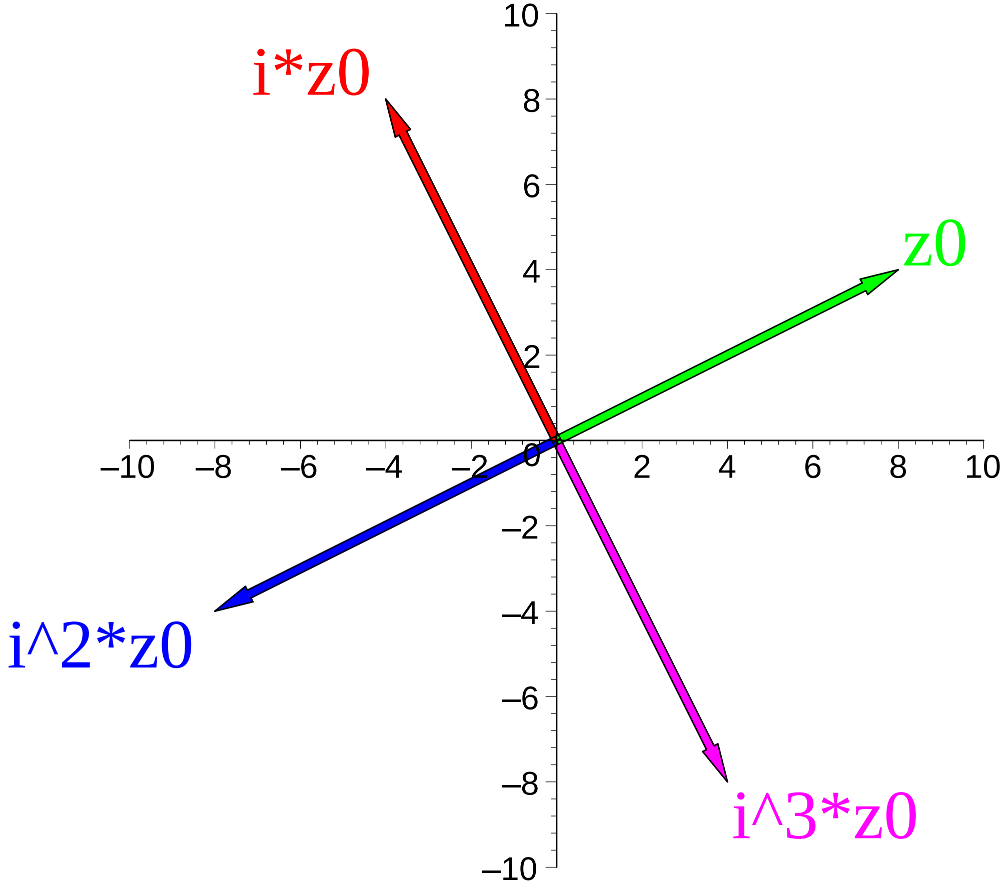
$$|z_2(z_0 z_1)^{-1} - 1|$$

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

Symmetries

Let $\mathcal{R}(a_0, b_0, a_1, b_1) := (a_0 \otimes a_1) \ominus (b_0 \otimes b_1)$ and $\mathcal{I}(a_0, b_0, a_1, b_1) := (a_0 \otimes b_1) \oplus (b_0 \otimes a_1)$.

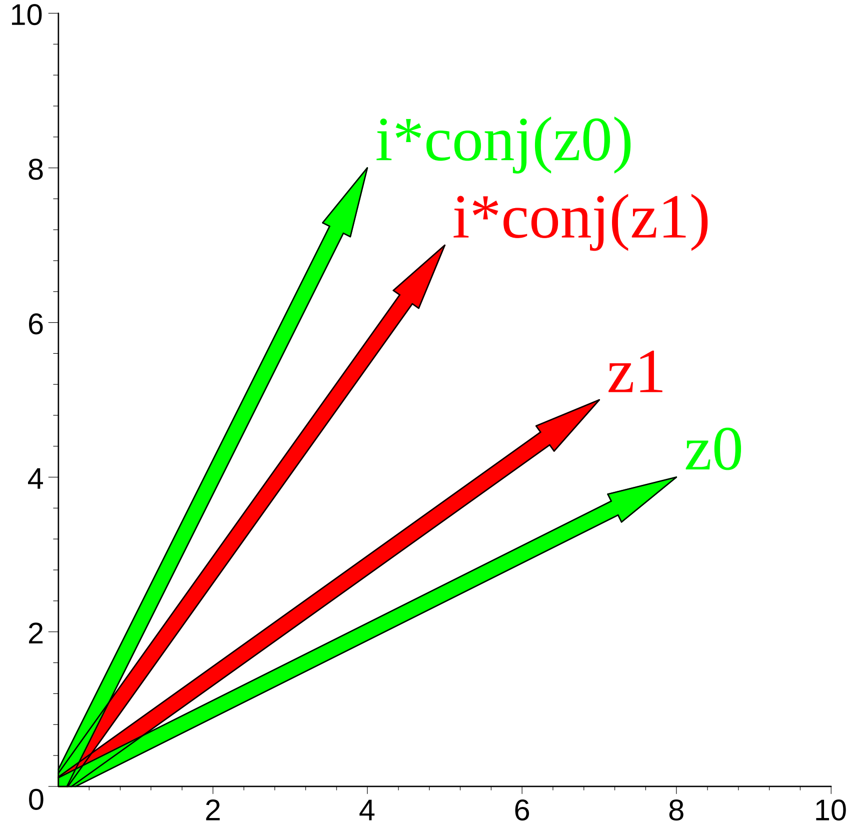
The change $z_0 \to z_0 i$ gives $(a_0, b_0) \to (-b_0, a_0)$, and $\mathcal{R} \to -\mathcal{I}, \mathcal{I} \to \mathcal{R}$, thus the relative error on z_2 is unchanged.



The same holds for $z_1 \to z_1 i$. We can thus assume z_0 and z_1 in the 1st quadrant:

$$a_0, b_0, a_1, b_1 \geq 0.$$

Similarly, $(z_0, z_1) \to (i\bar{z_0}, i\bar{z_1})$ gives $\mathcal{R} \to -\mathcal{R}$, $\mathcal{I} \to \mathcal{I}$.



We can thus assume $z_0 z_1$ is in the 1st quadrant:

$$b_0 b_1 \leq a_0 a_1$$

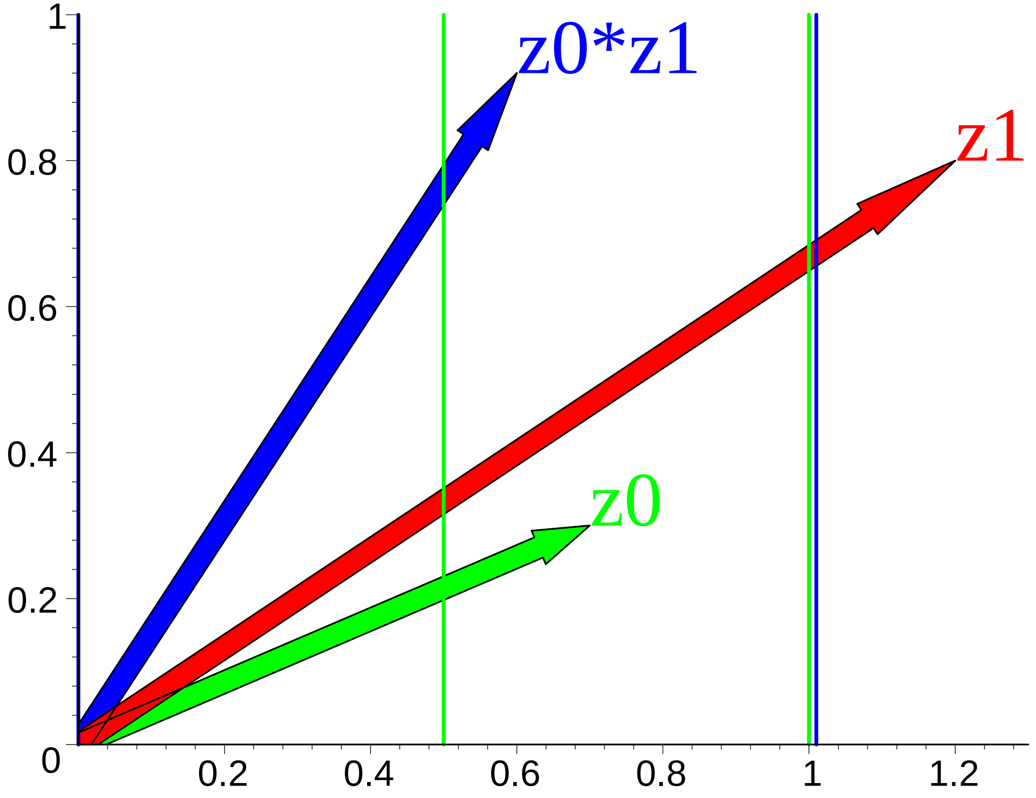
By exchanging z_0 and z_1, we can assume

$$b_0 a_1 \leq a_0 b_1$$

Then by $z_0 \to z_0 \cdot 2^j$ and $z_1 \to z_1 \cdot 2^k$, we can assume

$$\frac{1}{2} \leq a_0 < 1, \qquad \frac{1}{2} \leq a_0 a_1 < 1.$$

In the sequel, we assume all those inequalities hold.


z0*z1
z1
z0
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
1.2

Proof of Theorem 1 (sketch)

(1) bound on the imaginary part: two cases (I1, I2)

$$|\mathcal{I}(z_2 - z_0 z_1)| \leq \epsilon \cdot (2a_0 b_1 + 2b_0 a_1)$$

(2) bound on the real part: four cases (R1, R2, R3, R4)

$$|\mathcal{R}(z_2 - z_0 z_1)| \leq \epsilon \cdot (\lambda a_0 a_1 + \mu b_0 b_1) + \gamma \epsilon^2 \cdot (a_0 a_1 + b_0 b_1)$$

with different λ, μ, γ;

(3) from (1) and (2) we deduce:

$$|z_2 - z_0 z_1| \leq \nu\epsilon \cdot |z_0 z_1|$$

Preliminary Lemma

Lemma. *For any real x, let $y = \circ(x)$, we have:*

$$|y - x| \le \frac{1}{2}\mathrm{ulp}(x),$$

$$|y - x| < \epsilon \cdot |x|.$$

First bound trivial for $\mathrm{ulp}(x) = \mathrm{ulp}(y)$. Otherwise $y = \beta^j$ and $|y - x| \le \frac{1}{2\beta}\mathrm{ulp}(y) = \frac{1}{2}\mathrm{ulp}(x)$.

The 2nd follows from the 1st, with $\beta^{t-1}\mathrm{ulp}(x) \le |x|$ (equality if $|x| = \beta^j$ only) and $\epsilon = \frac{1}{2}\beta^{1-t}$.

The Imaginary Part

$$|\mathcal{I}(z_2 - z_0 z_1)| \leq |a_0 \otimes b_1 - a_0 b_1| + |b_0 \otimes a_1 - b_0 a_1|$$
$$+ \quad |((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) - (a_0 \otimes b_1 + b_0 \otimes a_1)|$$

Two cases:

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

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

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

Exceptional case.

Example: $z_0 = 0.1011 + 0.1000i$, $z_1 = 0.1100 + 0.1110i$.

$a_0 b_1 + b_0 a_1 = 0.11111010$

$a_0 \otimes b_1 = 0.1010$, $b_0 \otimes a_1 = 0.0110$,

$a_0 \otimes b_1 + b_0 \otimes a_1 = 1.000$

Remark: $a_0 \otimes b_1 + b_0 \otimes a_1$ is not necessarily a power of 2.

Consider $t = 5$, $z_0 = 30 + 19i$, $z_1 = 19 + 22i$, then

$a_0 b_1 + b_0 a_1 = 1021$, $a_0 \otimes b_1 + b_0 \otimes a_1 = 672 + 368 = 1040$.

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

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

Thus:

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

Since $\beta^t \mathrm{ulp}(a_0 b_1 + b_0 a_1)$ is representable:

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

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

Usual case.

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

In both cases (I1 and I2), we have

$$
\begin{aligned}
|((a_0 \otimes b_1) \oplus (b_0 \otimes a_1)) &- (a_0 \otimes b_1 + b_0 \otimes a_1)| \\
&\leq \epsilon \cdot (a_0 b_1 + b_0 a_1)
\end{aligned}
$$

thus:

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

A $\sqrt{6}$ Bound

```
> e := 6*(a0*b1+a1*b0)^2 + 6*(a0*a1-b0*b1)^2
       - 4*(a0*b1+a1*b0)^2 - 4*(a0*a1)^2:
> expand(e);
    2   2                          2   2        2   2
2 a0  b1  - 8 a0 b1 a1 b0 + 2 a1  b0  + 2 a0  a1


           2   2
     + 6 b0  b1
```

This is:

$$2(a_0 b_1 - b_0 a_1)^2 + 2(a_0 a_1 - b_0 b_1)^2 + 4(b_0 b_1)^2$$

A $\sqrt{4}$ Bound?

We have:

$$|\mathcal{I}(z_2 - z_0 z_1)| \leq 2\epsilon \cdot (a_0 b_1 + b_0 a_1)$$

If we had:

$$|\mathcal{R}(z_2 - z_0 z_1)| \leq 2\epsilon \cdot (a_0 a_1 - b_0 b_1)$$

we would get:

$$|z_2 - z_0 z_1|^2 \leq 4\epsilon^2 |z_0 z_1|^2$$

and thus:

$$|z_2 - z_0 z_1| \leq 2\epsilon |z_0 z_1|$$

Instead of $2 = \sqrt{4}$ we get $\sqrt{5}$ only ...

The Real Part

Let $A = \mathrm{ulp}(a_0 a_1)$, $B = \mathrm{ulp}(b_0 b_1)$,
$C = \mathrm{ulp}(a_0 \otimes a_1 - b_0 \otimes b_1)$. By hypothesis: $B \leq A$.

R1: $B \leq A \leq C$

R2: $B < C < A$

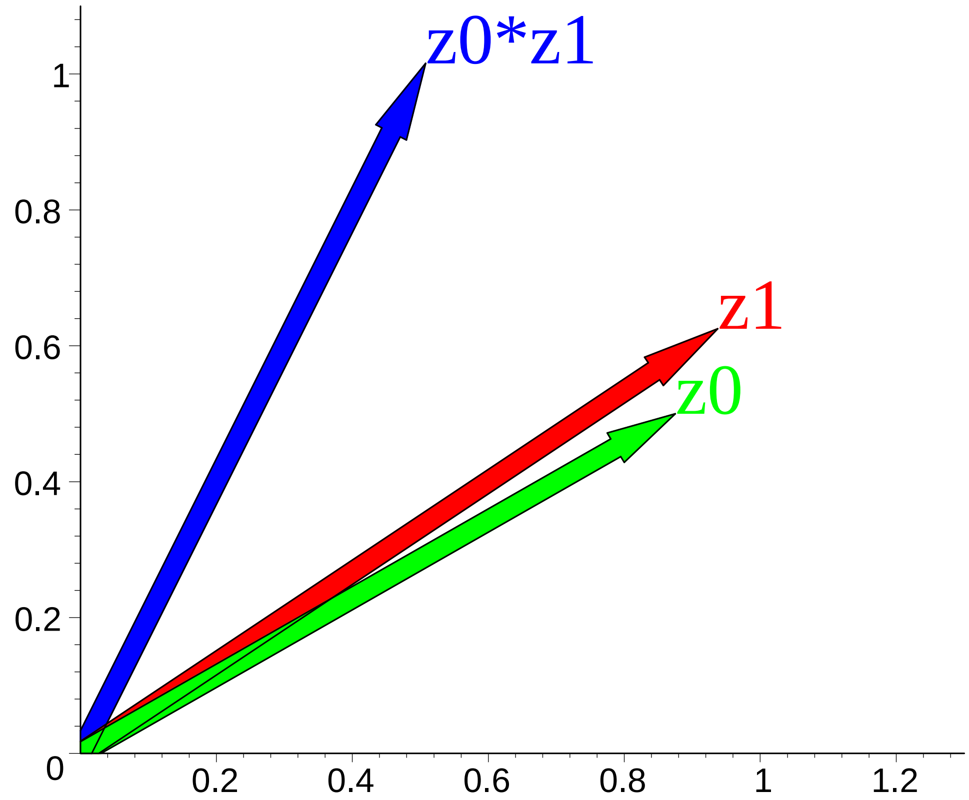
R3: $C \leq B < A$

R4: $C < B = A$

Case R1: $B \leq A \leq C$

Example: $\beta = 2,\ t = 4,\ z_0 = 14 + 8i,\ z_1 = 15 + 10i$

$a_0 \otimes a_1 - b_0 \otimes b_1 = 208 - 80 = 128,\ a_0 a_1 = 210$



$$|\mathcal{R}(z_2 - z_0 z_1)| < \epsilon \cdot (2a_0 a_1 - b_0 b_1) + \epsilon^2 \cdot (2a_0 a_1 + 2b_0 b_1)$$

which gives:

$$|z_2 - z_0 z_1| \le \epsilon(\sqrt{32/7} + 2\epsilon)|z_0 z_1|$$

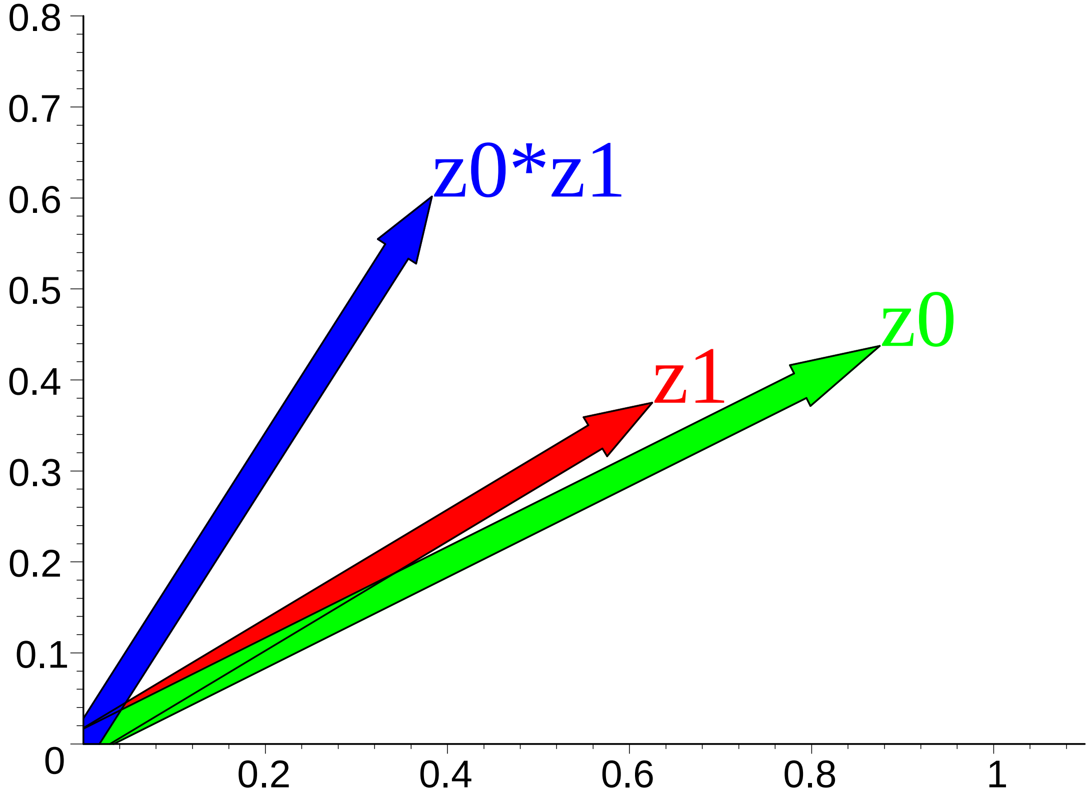
For $\epsilon \le 2^{-5}$:

$$\sqrt{32/7} + 2\epsilon \approx 2.138 + 2\epsilon \le 2.201 \le \sqrt{5} \approx 2.236$$

Case R2: $B < C < A$

Example: $\beta = 2,\ t = 3,\ z_0 = 14 + 7i,\ z_1 = 10 + 6i$

$b_0 b_1 = 42,\ a_0 \otimes a_1 - b_0 \otimes b_1 = 128 - 40 = 88,\ a_0 a_1 = 140,$



$$|\mathcal{R}(z_2 - z_0 z_1)| < \epsilon \cdot (7/4 \cdot a_0 a_1)$$

which gives:

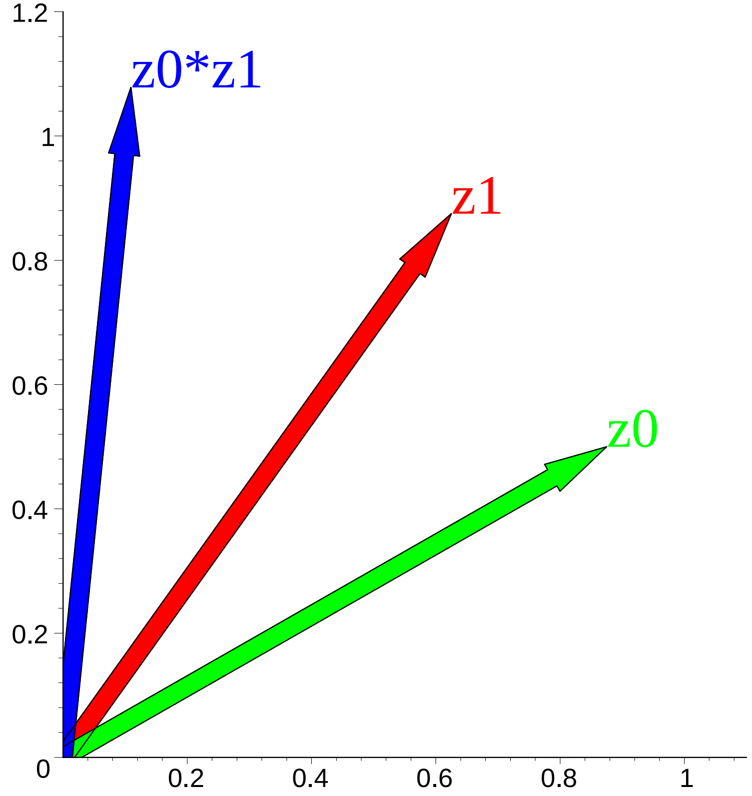
$$|z_2 - z_0 z_1| \le \epsilon \sqrt{1024/207} |z_0 z_1|$$

And $\sqrt{1024/207} \approx 2.224 \le \sqrt{5} \approx 2.236$

Case R3: $C \leq B < A$

Example: $\beta = 2,\ t = 3,\ z_0 = 7 + 4i,\ z_1 = 5 + 7i$

$$a_0 \otimes a_1 - b_0 \otimes b_1 = 32 - 28 = 4,\ b_0 b_1 = 28,\ a_0 a_1 = 35$$



$$|\mathcal{R}(z_2 - z_0 z_1)| < \epsilon \cdot (3/2 \cdot a_0 a_1)$$

Since $\frac{3}{2} \leq \frac{7}{4}$, we get a better bound than R2:

$$|z_2 - z_0 z_1| \leq \epsilon \sqrt{256/55} |z_0 z_1|$$

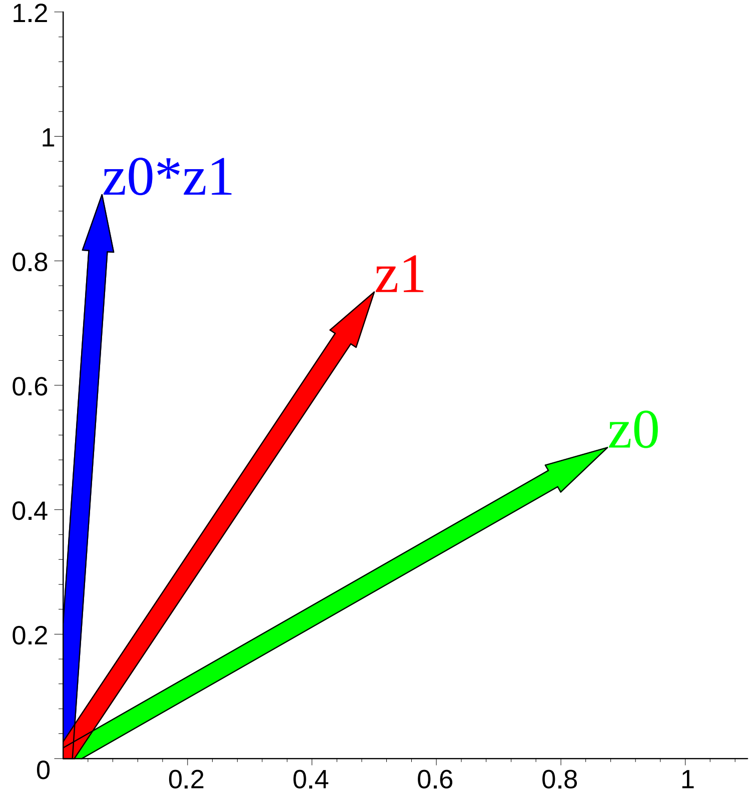
And $\sqrt{256/55} \approx 2.157 \leq \sqrt{5} \approx 2.236$

Case R4:

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

Example: $\beta = 2,\ t = 3,\ z_0 = 7 + 4i,\ z_1 = 4 + 6i$

$a_0 \otimes a_1 - b_0 \otimes b_1 = 28 - 24 = 4,\ b_0 b_1 = 24,\ a_0 a_1 = 28$



Sterbenz: $a_0 \otimes a_1 - b_0 \otimes b_1$ is exact.

$$|\mathcal{R}(z_2 - z_0 z_1)| \leq |a_0 \otimes a_1 - a_0 a_1| + |b_0 \otimes b_1 - b_0 b_1| < \epsilon \cdot (a_0 a_1 + b_0 b_1)$$

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

Worst-Case Multiplicands for $\beta = 2$

Theorem 2. *Assume*

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

for some positive integer n, *then* $a_0 \neq b_0$, $a_1 \neq b_1$, *and:*

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

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

$$0 \leq j_{aa}, j_{ab}, j_{ba}, j_{bb} < \frac{n}{4}, \quad |k_{aa}|, |k_{bb}| < n, \quad |k_{ab}|, |k_{ba}| < \frac{n}{2}$$

Proof of Theorem 2 (sketch)

$\sqrt{5 - n\epsilon} > \sqrt{1024/207}$ gives $n\epsilon < \frac{11}{207} \approx 0.053$

Thus $1/2 \le a_0a_1, a_0b_1, b_0a_1, b_0b_1 \lessapprox 1/2 + \frac{11}{828} \approx 0.513$

Case R4 must hold: $a_0 \otimes a_1 - b_0 \otimes b_1$ is exact, and $\mathrm{ulp}(b_0b_1) = \mathrm{ulp}(a_0a_1)$.

We get a lower bound on $|z_2 - z_0z_1|$, an upper bound on $|z_0z_1|$, from which we deduce tight bounds:

$$\epsilon/2 - (1 - \sqrt{1 - n\epsilon})\epsilon < |a_0 \otimes a_1 - a_0a_1| \le \epsilon/2$$

and similarly for $|b_0 \otimes b_1 - b_0b_1|$, ...

Conclude by noticing that a_0a_1 is an integer multiple of ϵ^2

Worst-Case in Single Precision

Corollary 4. *In IEEE 754 single-precision arithmetic ($\epsilon = 2^{-24}$), the worst-case values are:*

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

with a relative error $\epsilon\sqrt{5 - 168\epsilon} \approx \epsilon\sqrt{4.9999899864}$.

Worst-Case in Double Precision

Corollary 5. *In IEEE 754 double-precision arithmetic ($\epsilon = 2^{-53}$), the worst-case values are:*

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

with a relative error $\epsilon\sqrt{5-96\epsilon} \approx \epsilon\sqrt{4.9999999999999893}$.

Conjecture

For precision t large enough, the worst-cases are as in Corollary 4 (single precision) for even precision, and as in Corollary 5 (double precision) for odd precision.

In particular, the worst-case for quadruple precision $t = 113$ would be as for double precision.

Applications

- correctly rounded complex multiply (separate relative error on real and imaginary parts)

- complex floating-point FFT (Percival's paper):

Theorem. *The FFT allows computation of the cyclic convolution $z = x * y$ of two vectors of length $N = 2^n$ of complex values such that*

$$|z' - z|_\infty < |x| \cdot |y| \cdot [(1+\epsilon)^{3n}(1+\epsilon\sqrt{5})^{3n+1}(1+\alpha)^{3n} - 1],$$

where $|\cdot|$ denotes the Euclidean norm, and
$\alpha > |(\omega^k)' - (\omega^k)|$, $\omega = e^{\frac{2\pi i}{N}}$.

Applications (2)

If $\omega^k = x + yi$ is correctly rounded, $\alpha = \epsilon/\sqrt{2}$:

$\mathrm{err}(x), \mathrm{err}(y) \leq \frac{1}{2}\epsilon$,

$$|z' - z|_\infty < |x| \cdot |y| \cdot [(1+\epsilon)^{3n}(1+\epsilon\sqrt{5})^{3n+1}(1+\epsilon/\sqrt{2})^{3n} - 1]$$

Improvement: from $1 + 1/\sqrt{2} + \sqrt{8}$ to $1 + 1/\sqrt{2} + \sqrt{5}$, about 13%.

Example: multiply two degree 524288 polynomials with digits in $[-5000, 5000]$, or 2 million digit numbers.

Open Problems

- simplify the 3-page proof of Theorem 1
- get rid of the ϵ^2 term in Case R1
- prove the conjecture
- find the worst-cases for any β
- get ω^k correctly rounded ...

Percival: linear-time algorithm for max error of 1.5ϵ

Lemma. *For any real x, let $y = \circ(x)$, we have:*

$$|y - x| < \frac{\epsilon}{1+\epsilon}|x|.$$

Proof. We can assume $1 \leq x < 2$.

If $1 + \epsilon \leq x$:

$$|y - x| \leq \epsilon \leq \epsilon \frac{x}{1+\epsilon}$$

If $x = 1 + \lambda$ with $0 \leq \lambda < \epsilon$:

$$|y - x| = \lambda \leq \frac{\epsilon}{1+\epsilon}(1 + \lambda)$$

Since:

$$\lambda(1 + \epsilon) \leq \epsilon(1 + \lambda)$$