On complex multiplication and division with an FMA

Jean-Michel Muller

Joint work with C.-P. Jeannerod, P. Kornerup and N. Louvet

INVA 2014

Thank you!

ありがとうございます

Complex multiplication and division

Given complex numbers $x = a + ib$ and $y = c + id$, their product $z = xy$ can be expressed as

$$z = ac - bd + i(ad + bc);$$

and their quotient x/y can be expressed as

$$q = \frac{ac + bd}{c^2 + d^2} + i\,\frac{bc - ad}{c^2 + d^2}.$$

In floating-point arithmetic, several issues:

- accuracy,
- spurious overflow/underflow (e.g., $c^2 + d^2$ overflows, whereas the real and imaginary parts of q are representable);

Here: focus on accuracy problems. Scaling techniques to avoid spurious overflow/underflow dealt with in further work.

We assume that an FMA instruction is available.

Floating-Point numbers, roundings

Precision-p binary FP number: either 0 or

$$x = X \cdot 2^{e_x - p + 1},$$

where X and $e_x \in \mathbb{Z}$, with $2^{p-1} \leqslant |X| \leqslant 2^p - 1$. We denote this set by $\mathbb{F}_p$.

- unlimited exponent range $\rightarrow$ results valid in usual FP arithmetic unless underflow/overflow occurs;
- X: integral significand of x;
- $2^{1-p} \cdot X$: significand of x;
- e_x: exponent of x.

This presentation: binary arithmetic only.

Floating-Point numbers, roundings

- In general, the sum, product, quotient, etc., of two FP numbers is not an FP number: it must be rounded;
- correct rounding: *Rounding function* $\circ$, and when $(a \top b)$ is performed, the returned value is $\circ(a \top b)$;
- default rounding function: round to nearest even: RN;
- for any real number t, $|\mathrm{RN}(t) - t| \leqslant u \cdot |t|$, where $u = 2^{-p}$;

$\longrightarrow$ all arithmetic operations performed with relative error $\leqslant u$;

- we assume that an FMA instruction is available: computes RN(ab+c).

Here: we try to analyze several simple algorithms.

Componentwise and normwise relative errors

When $\widehat{z}$ approximates z:

- componentwise error:

$$\max\left\{\left|\frac{\mathrm{Re}\,(z)-\mathrm{Re}\,(\widehat{z})}{\mathrm{Re}\,(z)}\right|;\left|\frac{\mathrm{Im}\,(z)-\mathrm{Im}\,(\widehat{z})}{\mathrm{Im}\,(z)}\right|\right\};$$

- normwise error:

$$\left|\frac{z-\widehat{z}}{z}\right|.$$

Choosing between both kinds of error depends on the application.

- componentwise error $\leqslant \epsilon \Rightarrow$ normwise error $\leqslant \epsilon$;
- the converse is not true.

Naive multiplication algorithm without an FMA

$$\boxed{\mathcal{A}_0 : (a+ib, c+id) \mapsto \mathrm{RN}\big(\mathrm{RN}(ac) - \mathrm{RN}(bd)\big) + i \cdot \mathrm{RN}\big(\mathrm{RN}(ad) + \mathrm{RN}(bc)\big)}$$

- componentwise error: can be huge (yet finite);
- Normwise accuracy: studied by Brent, Percival, and Zimmermann (2007). The computed value has the form

$$\widehat{z}_0 = z(1+\epsilon), \qquad |\epsilon| < \sqrt{5}\, u,$$

$\rightarrow$ the normwise relative error $|\widehat{z}_0/z - 1|$ is always $\leqslant \sqrt{5} \cdot u$.

For any $p \geqslant 2$ they provide FP numbers a, b, c, d for which $|\widehat{z}_0/z - 1| = \sqrt{5}\, u - O(u^2) \rightarrow$ the relative error bound $\sqrt{5}\, u$ is asymptotically optimal as $u \rightarrow 0$ (or, equivalently, as $p \rightarrow +\infty$).

Can we do better if an FMA instruction is available?

Naive multiplication algorithm with an FMA

With an FMA, the simple way of evaluating $ac - bd + i(ad + bc)$ becomes:

$$\boxed{\mathcal{A}_1 : (a + ib, c + id) \mapsto \mathrm{RN}\big(ac - \mathrm{RN}(bd)\big) + i \cdot \mathrm{RN}\big(ad + \mathrm{RN}(bc)\big)}$$

Algorithm $\mathcal{A}_1$ is just one of 4 variants that differ only in the choice of the products to which the FMA operations apply.

- componentwise error: can be huge (even infinite);
- normwise error:
  - for any of these 4 variants the computed complex product $\widehat{z}_1$ satisfies

$$|\widehat{z}_1 - z| \leqslant 2u|z| \qquad (1)$$

  - we build inputs a, b, c, d for which $|\widehat{z}_1/z - 1| = 2u - O(u^{1.5})$ as $u \to 0 \Rightarrow$ the relative error bound (1) is asymptotically optimal (given later on).

$\longrightarrow$ the FMA improves the situation from a normwise point of view.

The CHT algorithm

Given FP numbers a and b, the error $e = ab - \mathrm{RN}(ab)$ satisfies

$$e = \mathrm{RN}\big(ab - \mathrm{RN}(ab)\big)$$

- $\longrightarrow$ it is computed exactly with an FMA;
- $\longrightarrow$ compensated algorithms: we "re-inject" that error later on in the calculation.

(without an FMA and using only $+$, $-$, $\times$, the cheapest algorithm we are aware of for computing e uses 17 operations)

Cornea, Harrison, and Tang use this property in the following algorithm to evaluate

$$r = ab + cd$$

accurately in 7 floating-point operations.

The CHT algorithm

We approximate

$$r = ab + cd$$

by $\widehat{r}$ obtained as follows

$$
\boxed{
\begin{array}{l}
\textbf{algorithm } \mathrm{CHT}(a, b, c, d) \\
\quad \widehat{w}_1 := \mathrm{RN}(ab); \qquad \widehat{w}_2 := \mathrm{RN}(cd); \\
\quad e_1 := \mathrm{RN}(ab - \widehat{w}_1);\ e_2 := \mathrm{RN}(cd - \widehat{w}_2); \quad \text{// exact operations} \\
\quad \widehat{f} := \mathrm{RN}(\widehat{w}_1 + \widehat{w}_2); \\
\quad \widehat{e} := \mathrm{RN}(e_1 + e_2); \\
\quad \widehat{r} := \mathrm{RN}\left(\widehat{f} + \widehat{e}\right); \\
\quad \textbf{return } \widehat{r};
\end{array}
}
$$

Cornea, Harrison, and Tang show that the error is $\mathcal{O}(u)$.

Properties of the CHT algorithm

- we have shown that
$$|\widehat{r} - r| \leqslant 2u \cdot |r| \tag{2}$$
- we build a "generic example" parameterized by p, that shows that the bound (2) is asymptotically optimal (as $u \to 0$);
- for instance, in double precision arithmetic, with our generic example, error
$$u \times 1.9999999999999922284\cdots$$
is attained.

Application of CHT to the complex product

- Evaluate separately the real and imaginary parts of $z = ac - bd + i(ad + bc)$ using CHT;
- uses 14 floating-point operations.

$$\boxed{\mathcal{A}_2 : (a + ib, c + id) \mapsto \mathrm{CHT}(a, c, -b, d) + i \cdot \mathrm{CHT}(a, d, b, c)}$$

- componentwise error $\leqslant 2u$ (asymptotically optimal);
- consequence: normwise error $\leqslant 2u$.

The normwise bound is also asymptotically optimal.

Application of CHT to the complex product

Theorem 1

Let $a, b \in \mathbb{F}_p$ *be given by*

$$a = RD\Big((1-2^{-p})\sqrt{2^{p-2}}\Big), \qquad b = 2^{p-1} + \Big\lfloor \sqrt{2^{p-2}} \Big\rfloor + 1, \tag{3}$$

where, for $t \in \mathbb{R}$, $RD(t) = \max\{f \in \mathbb{F}_p \,:\, f \leqslant t\}$ *denotes rounding down in* $\mathbb{F}_p$*. Let also* $\widehat{z}_2$ *be the approximation to* $z = (a+ib)^2$ *computed by algorithm* $\mathcal{A}_2$*. If* $p \geqslant 5$ *then, barring underflow and overflow,*

$$|\widehat{z}_2/z - 1| > 2u - 8u^{1.5} - 6u^2.$$

To be compared with the upper bound $2u$.

Kahan's algorithm for $ab + cd$

algorithm Kahan(a, b, c, d)
  $\widehat{w} := \mathrm{RN}(cd)$;
  $e := \mathrm{RN}(cd - \widehat{w})$;   // this operation is exact: $e = cd - \widehat{w}$.
  $\widehat{f} := \mathrm{RN}(ab + \widehat{w})$;
  $\widehat{r} := \mathrm{RN}\left(\widehat{f} + e\right)$;
  return $\widehat{r}$;

- 4 operations (CHT needed 7);
- $e = cd - \mathrm{RN}(cd)$ computed exactly thanks to the FMA instruction;
- it is added to $\widehat{f}$ in order to yield the approximation $\widehat{r}$ to $r = ab + cd$.

We have shown that

$$|\widehat{r} - r| \leqslant 2u|r|, \tag{4}$$

and that this bound is asymptotically optimal (as $u \to 0$).

Application of Kahan's algorithm to the complex product

- Evaluate separately the real and imaginary parts of $z = ac - bd + i(ad + bc)$ using Kahan's algorithm;
- uses 8 floating-point operations (instead of 14 with CHT);

$$\boxed{\mathcal{A}_3 : (a + ib, c + id) \mapsto \mathrm{Kahan}(a, c, -b, d) + i \cdot \mathrm{Kahan}(a, d, b, c)}$$

- componentwise error $\leqslant 2u$ (asymptotically optimal);
- consequence: normwise error $\leqslant 2u$.

The normwise bound is asymptotically optimal.

Theorem 2

Let $a, b \in \mathbb{F}_p$ *be given by*

$$a = \mathrm{pred}\left(\sqrt{2^{p-2}}\right), \qquad b = 2^{p-1} + \left\lfloor \sqrt{2^{p-2}} \right\rfloor + 1,$$

where, for $t \in \mathbb{R}_{>0}$, $\mathrm{pred}(t) = \max\{f \in \mathbb{F}_p \,:\, f < t\}$ *denotes the predecessor of* t *in* $\mathbb{F}_p$. *Let also* $\widehat{z}_1$ *and* $\widehat{z}_3$ *be the approximations to* $z = (a + ib)^2$ *computed by algorithms* $\mathcal{A}_1$ *and* $\mathcal{A}_3$, *respectively. If* $p \geqslant 5$ *then, barring underflow and overflow,*

$$|\widehat{z}_h/z - 1| > 2u - 8u^{1.5} - 4u^2, \qquad h \in \{1, 3\}.$$

Conclusion on complex multiplication

- the availability of an FMA makes it possible to replace the classical normwise accuracy bound $\sqrt{5}u$ by $2u$ with simple algorithms,
- this new bound is sharp (asymptotically optimal with Algorithms $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$),
- if normwise error only is at stake, the simplest algorithm (naive multiplication with FMA: Algorithm $\mathcal{A}_1$) is juste fine,
- however if we also want to reduce the componentwise error the multiplication based on Kahan's algorithm (i.e., Algorithm $\mathcal{A}_3$) is to be preferred.

More on this: `http://perso.ens-lyon.fr/jean-michel.muller/JeKoLoMu13-submission.pdf`

A few words on complex division with an FMA

$$q = \frac{ac + bd}{c^2 + d^2} + i\,\frac{bc - ad}{c^2 + d^2}.$$

- here: componentwise error only;
- basic idea: separately compute $ac + bd$, $bc - ad$, and $c^2 + d^2$ using one of the previously seen methods (naive without FMA, naive with FMA, CHT, Kahan);
- $c^2 + d^2$ is a special case: cancellation cannot occur;
- notice that $ac + bd$ and $bc - ad$ cannot cancel simultaneously:
  - if $abcd > 0$ then ac and bd have the same sign;
  - otherwise bc and $-ad$ have the same sign (unless one of the inputs is zero: case easily dealt with).
- straight-line algorithms: no tests.

The special case $c^2 + d^2$

- more generally, computation of $ac + bd$ with ab and cd of the same sign;
- the naive method, the naive method with FMA, Kahan's algorithm have the same relative error bound $2u$;
- the bound is sharp, even if we restrict ourselves to $c^2 + d^2$.

Consequence: for $c^2 + d^2$, the naive algorithm (3 operations) or the naive-with-FMA algorithm (2 operations) suffice.

A $5u + \mathcal{O}(u^2)$ componentwise error bound

- we use Kahan's algorithm or CHT for $ac + bd$ and $bc - ad$
- we use naive-with-fma for $c^2 + d^2$;
- consider the real part $(ac + bd)/(c^2 + d^2)$. We have:

$$\mathrm{Re}\,(\widehat{q}) = \frac{\mathrm{Re}\,(q)(1+\epsilon)}{(c^2+d^2)(1+\epsilon')}(1+\epsilon''),$$

  where $|\epsilon|, |\epsilon'| \leqslant 2u$, and where the relative error $|\epsilon''|$ of FP division is bounded by u.

$\rightarrow$ the real part of $\widehat{q}$ has the form $(1+\theta) \cdot \mathrm{Re}\,(q)$, with

$$|\theta| \leqslant \frac{1+2u}{1-2u}(1+u) - 1,$$

- the same holds for the imaginary part.

Consequence: componentwise error $\leqslant 5u + 13u^2$.

A $5u + \mathcal{O}(u^2)$ componentwise error bound

The obtained algorithm is:

algorithm CompDivS($a + ib$, $c + id$)
$\widehat{\delta} := \mathrm{RN}(c^2 + \mathrm{RN}(d^2))$;
$\widehat{g}_{\mathrm{re}} := \mathrm{Kahan}(a, b, c, d)$; // evaluates $ac + bd$
$\widehat{g}_{\mathrm{im}} := \mathrm{Kahan}(b, -a, c, d)$; // evaluates $bc - ad$
$\widehat{q}_{\mathrm{re}} := \mathrm{RN}\big(\widehat{g}_{\mathrm{re}}/\widehat{\delta}\big)$; $\widehat{q}_{\mathrm{im}} := \mathrm{RN}\big(\widehat{g}_{\mathrm{im}}/\widehat{\delta}\big)$;
return $\widehat{q}_{\mathrm{re}} + i\,\widehat{z}_{\mathrm{im}}$;

When p is even the bound is asymptotically optimal

If we choose:

$$
\begin{aligned}
a &= 2^p - 5 \cdot 2^{\frac{p}{2}-1}, \\
b &= -2^{-\frac{p}{2}} \cdot \left(2^p - 5 \cdot 2^{\frac{p}{2}-1} + 3\right), \\
c &= 2^p - 2, \\
d &= 2^{\frac{p}{2}+1} \cdot \left(2^{p-1} + 2^{\frac{p}{2}-1}\right),
\end{aligned}
$$

then the quotient $\widehat{q}$ computed by CompDivS satisfies

$$
\frac{|\mathrm{Re}\,(\widehat{q}) - \mathrm{Re}\,(q)|}{|\mathrm{Re}\,(q)|} = 5u - O(u^{3/2}). \tag{5}
$$

When p is odd...

We have no proof of asymptotic optimality, however:

p	example
53	$a = 2^{52} + 1$
	$b = -142398041$
	$c = 2^{52}$
	$d = 94906267 \cdot 2^{52} = \left(1 + \lceil 2^{53/2} \rceil\right) \cdot 2^{52}$
	$\lvert \widehat{z}_{\mathrm{re}} - \operatorname{Re} z \rvert / \lvert \operatorname{Re} z \rvert = 4.9987\ldots \times u$
113	$a = 2^{112} + 1$
	$b = -152857240142482713$
	$c = 2^{112}$
	$d = 101904826760412363 \cdot 2^{112}$
	$\lvert \widehat{z}_{\mathrm{re}} - \operatorname{Re} z \rvert / \lvert \operatorname{Re} z \rvert = 4.9999\ldots \times u$

Table 1: Relative error in $\widehat{z}_{\mathrm{re}}$ computed using CompDivS close to the upper bound $5u + 13u^2$.

Conclusion on complex division

- accurate complex division is feasible with simple algorithms: componentwise error bound $5u + 13u^2$;
- that bound is asymptotically optimal (at least for even p);
- to be done: use scaling techniques to avoid spurious overflow/underflow.

Thank you for your attention.