ARITH 21, Austin, TX, 2012

On the componentwise accuracy of complex floating point division with an FMA

Claude-Pierre Jeannerod
INRIA

Nicolas Louvet
UCB Lyon 1

Jean-Michel Muller
CNRS

Aric research team, LIP

Complex division

Given $x = a + ib$ and $y = c + id$, $z = x/y$ is can be expressed as

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

Several issues when performed in floating point arithmetic:

- Loss of componentwise accuracy, due to cancellations;
- Spurious overflow/underflow, e.g., $c^2 + d^2$ overflows, while the real and imaginary parts are representable.

In this talk:

- We focus on accuracy problems.
- We assume that a Fused Multiply-Add (FMA) instruction is available.

Assumptions and notation

- We assume a standard radix-2, precision-p ($p \geqslant 2$), floating point arithmetic.
- RN$(\cdot)$ denotes rounding 'to the nearest even'.
- An FMA instruction is available:
  given three floating point numbers x, y and z, it computes RN$(xy + z)$;
- $u = 2^{-p}$ denotes the unit roundoff;
- ulp to denote the unit in the last place function:
  $\mathrm{ulp}(0) = 0$ and for any nonzero real number t, $\mathrm{ulp}(t) = 2^{\lfloor \log_2(t) \rfloor + 1 - p}$.
- We assume that underflows and overflows never occur, so that

$$|\mathrm{RN}(t) - t| \leqslant \tfrac{1}{2}\mathrm{ulp}(t) \leqslant u|t| \quad \text{for any real number } t.$$

High normwise accuracy

Let $x = a + ib$ and $y = c + id$ with floating point coefficients.

$z = x/y$ can be evaluated without FMA as

$$\operatorname{Re}\widehat{z} = \mathrm{RN}\left(\frac{\mathrm{RN}\big(\mathrm{RN}(ac) + \mathrm{RN}(bd)\big)}{\mathrm{RN}\big(\mathrm{RN}(c^2) + \mathrm{RN}(d^2)\big)}\right),$$

and a similar expression for the imaginary part.

In the absence of underflow and overflow [Baudin, Smith, 2012],

$$\frac{|\widehat{z} - z|}{|z|} < \underbrace{(3+\sqrt{5})}_{=5.236\cdots} u + 13u^2.$$

$\hookrightarrow$The computed $\widehat{z}$ is always highly accurate in the normwise sense.

Componentwise relative error

$$E_c(\widehat{z}) := \max\left(\frac{|\mathrm{Re}\,\widehat{z} - \mathrm{Re}\,z|}{|\mathrm{Re}\,z|}, \frac{|\mathrm{Im}\,\widehat{z} - \mathrm{Im}\,z|}{|\mathrm{Im}\,z|}\right).$$

The classic division formula can be very inaccurate in the componentwise sense.

We provide an example ($p = 53$) such that:

- $E_c(\widehat{z}) \approx 9.0 \times 10^{15} \gg 1$ with the classic algorithm without FMA:

$$\mathrm{Re}\,\widehat{z} = \mathrm{RN}\left(\frac{\mathrm{RN}\big(\mathrm{RN}(ac) + \mathrm{RN}(bd)\big)}{\mathrm{RN}\big(\mathrm{RN}(c^2) + \mathrm{RN}(d^2)\big)}\right).$$

- $E_c(\widehat{z}) \approx 4.5 \times 10^{15} \gg 1$ with the classic algorithm with FMA:

$$\mathrm{Re}\,\widehat{z} = \mathrm{RN}\left(\frac{\mathrm{RN}\big(ac + \mathrm{RN}(bd)\big)}{\mathrm{RN}\big(c^2 + \mathrm{RN}(d^2)\big)}\right).$$

$\hookrightarrow$ How to obtain high componentwise accuray?

Kahan's algorithm for $g = ac + bd$

Kahan(a, b, c, d)
$\widehat{w} := \mathrm{RN}(bd)$;
$e := \mathrm{RN}(bd - \widehat{w})$; // $e = bd - \widehat{w}$ is computed exactly
$\widehat{f} := \mathrm{RN}(ac + \widehat{w})$; // $\widehat{f}$ is a naive evaluation of $ac + bd$ with FMA
$\widehat{g} := \mathrm{RN}(\widehat{f} + e)$; // compensation step
return $\widehat{g}$;

Kahan's algorithm always approximates $ac + bd$ to high relative accuracy.

Theorem ([Jeannerod, Louvet, Muller, 2012])

In the absence of underflow/overflow,

$$\frac{|\widehat{g} - g|}{|g|} \leqslant 2u.$$

This bound is proved to be asymptotically optimal, via the explicit construction of inputs for which $\frac{|\widehat{g}-g|}{|g|} = 2u - O(u^2)$.

Complex division with Kahan's algorithm

Kahan's algorithm to evaluate the numerators and the denominator in

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

then performing two floating point divisions.

$\hookrightarrow$ High componentwise accuracy can easily be obtained: $E_c(\widehat{z}) \leqslant 5u + O(u^2)$.

1. However, among the three terms in the division formula only one can cancel. In particular, a simpler scheme should be used for the denominator.
   $\rightarrow$ Accuracy of simpler algorithms for the sum of two nonnegative products.

2. We deduce two accurate FMA-based algorithms for complex division:
   - a straight-line algorithm,
   - an algorithm using tests to reduce error bound.

   $\rightarrow$ The sharpness of the error bounds was investigated.

Outline

Given a, b, c, d four floating point numbers, we focus in this part on the approximation of $g = ac + bd$ under the assumption

$$ac \geqslant 0 \qquad \text{and} \qquad bd \geqslant 0.$$

We consider the following ways of evaluating g:

- $\widehat{f_o} = \mathrm{RN}\big(\mathrm{RN}(ac) + \mathrm{RN}(bd)\big)$,
- $\widehat{f} = \mathrm{RN}\big(ac + \mathrm{RN}(bd)\big)$,
- $\widehat{g} = \mathrm{Kahan}(a, b, c, d)$.

Known results:

- From [Brent, Percival, Zimmermann, 2007]:
  $\hookrightarrow$ $2u$ is a relative error bound for $\widehat{f_o}$.

- From [Jeannerod, Louvet, Muller, 2012]:
  $\hookrightarrow$ $\mathrm{ulp}(g)$ and $2u$ are sharp absolute/relative error bounds for $\widehat{g}$.

Summary of the error bounds

$\widehat{r}$	bound on $\lvert\widehat{r}-g\rvert$		bound on $\frac{\lvert\widehat{r}-g\rvert}{g}$	
$\widehat{f}_o = \mathrm{RN}(\mathrm{RN}(ac)+\mathrm{RN}(bd))$	$\frac{5}{4}\mathrm{ulp}(g)$		$2u$	[BrePerZim07]
$\widehat{f} = \mathrm{RN}(ac+\mathrm{RN}(bd))$	$\mathrm{ulp}(g)$		$2u$	
$\widehat{g} = \mathrm{Kahan}(a,b,c,d)$	$\mathrm{ulp}(g)$	[JeLoMu12]	$2u$	[JeLoMu12]
$\mathrm{ulp}(bd) \leqslant \mathrm{ulp}(ac)$, $\widehat{f} = \widehat{g}$	$\frac{3}{4}\mathrm{ulp}(g)$		$\frac{3}{2}u$	

- The optimality or the asymptotic optimality (as $p \to \infty$) of these error bounds is established using examples of the form $ac + bd$ parametrized by p.
- We investigated the sharpness of these bounds for the evaluation of sums of squares $c^2 + d^2$: in most cases the bounds remain asymptotically optimal.
- When no example parametrized by p were found (p odd), the quality of the error bounds is illustrated with numerical examples in precisions $p = 53, 113$.

The case ulp(bd) $\leqslant$ ulp(ac)

The error bounds ulp(g) and $2u$ for the general case can be improved:

Theorem

If ulp(bd) $\leqslant$ ulp(ac) *then* $|\widehat{f} - g| \leqslant \frac{3}{4}\text{ulp}(g)$ *and* $\frac{|\widehat{f}-g|}{g} \leqslant \frac{3}{2}u$.

Example

For $p \geqslant 6$, let

$$a = 2^{p-1} + 2^{\lfloor \frac{p}{2} \rfloor}, \quad b = c = 2^{p-1} + 2^{\lceil \frac{p}{2} \rceil - 1}, \quad d = 2^{p-1} + 2^{\lfloor \frac{p}{2} \rfloor - 1}.$$

In this example, we have ulp(bd) = ulp(ac), and

- $|\widehat{f} - g| = \frac{3}{4}\text{ulp}(g)$ $\Rightarrow$ optimal absolute error bound
- $\frac{|\widehat{f}-g|}{g} = \frac{3}{2}u - O(u^{3/2})$ $\Rightarrow$ asymptotically optimal relative error bound

Outline

A straight-line algorithm

CompDivS$(a + ib,\ c + id)$

$\widehat{\delta} := \mathrm{RN}(c^2 + \mathrm{RN}(d^2))$ // $c^2 + d^2$ $\rightsquigarrow 2u$

$\widehat{g}_{\mathrm{re}} := \mathrm{Kahan}(a, b, c, d)$

$\widehat{g}_{\mathrm{im}} := \mathrm{Kahan}(b, -a, c, d)$ // $ac + bd$ and $bc - ad$ $\rightsquigarrow 2u$

$\widehat{z}_{\mathrm{re}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{re}}/\widehat{\delta}\right)$

$\widehat{z}_{\mathrm{im}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{im}}/\widehat{\delta}\right)$ // $\mathrm{Re}\,(\widehat{z})$ and $\mathrm{Im}\,(\widehat{z})$ $\rightsquigarrow u$

return $\widehat{z}_{\mathrm{re}} + i\,\widehat{z}_{\mathrm{im}}$

Property

For $p \geqslant 5$ and in the absence of underflow and overflow, $E_c(\widehat{z}) \leqslant 5u + 13u^2$.

For p odd, the sharpness of the bound illustrated through numerical examples:

p	$\lvert\widehat{z}_{\mathrm{re}} - \mathrm{Re}\,z\rvert/(u\lvert\mathrm{Re}\,z\rvert)$
53	4.9987 . . .
113	4.9987 . . .

For p even, the bound is proved to be asymptotically optimal.

Assuming p is even, let

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

Defining $R = 2^{p/2}$, it can be check that

$$\operatorname{Re} z = -\frac{2R^3 + 5R^2 - 4R}{2R^6 + 4R^5 + 4R^4 - 8R^2 + 8},$$

$$\frac{\widehat{g}_{\mathrm{re}}}{\widehat{\delta}} = \frac{-R^3 - \frac{5}{2}R^2}{R^6 + 2R^5} = \underbrace{-\frac{1}{R^3} - \frac{1}{2R^4}}_{=\operatorname{Re}\widehat{z}} + O\left(\frac{1}{R^5}\right).$$

Then we deduce

$$\frac{|\operatorname{Re}\widehat{z} - \operatorname{Re} z|}{|\operatorname{Re} z|} = \frac{5}{R^2} - O\left(\frac{1}{R^3}\right) = 5u - O(u^{3/2}),$$

which is equivalent to the componentwise bound $5u + 13u^2$ as $p \to \infty$.

If tests are allowed

CompDivT$(a+ib,\ c+id)$

if $|d| \leqslant |c|$ **then** $\widehat{\delta} := \mathrm{RN}(c^2 + \mathrm{RN}(d^2))$ // $|d| \leqslant |c| \implies \mathrm{ulp}(c^2) \leqslant \mathrm{ulp}(d^2)$

else $\widehat{\delta} := \mathrm{RN}(d^2 + \mathrm{RN}(c^2))$ // $c^2 + d^2 \rightsquigarrow 1.5u$

$\widehat{g}_{\mathrm{re}} := \mathrm{Kahan}(a, b, c, d)$

$\widehat{g}_{\mathrm{im}} := \mathrm{Kahan}(b, -a, c, d)$ // $ac + bd$ and $bc - ad \rightsquigarrow 2u$

$\widehat{z}_{\mathrm{re}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{re}}/\widehat{\delta}\right)$

$\widehat{z}_{\mathrm{im}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{im}}/\widehat{\delta}\right)$ // $\mathrm{Re}\,(\widehat{z})$ and $\mathrm{Im}\,(\widehat{z}) \rightsquigarrow u$

return $\widehat{z}_{\mathrm{re}} + i\,\widehat{z}_{\mathrm{im}}$

Property

For $p \geqslant 6$ and in the absence of underflow and overflow, $E_c(\widehat{z}) \leqslant 4.5u + 9u^2$.

p	$\lvert\widehat{z}_{\mathrm{re}} - \mathrm{Re}\,z\rvert/(u\lvert\mathrm{Re}\,z\rvert)$
24	4.4932 . . .
53	4.4421 . . .
113	4.4620 . . .

If tests are allowed

CompDivT$(a + ib,\ c + id)$

if $|d| \leqslant |c|$ **then** $\widehat{\delta} := \mathrm{RN}(c^2 + \mathrm{RN}(d^2))$ // $|d| \leqslant |c| \implies \mathrm{ulp}(c^2) \leqslant \mathrm{ulp}(d^2)$

else $\widehat{\delta} := \mathrm{RN}(d^2 + \mathrm{RN}(c^2))$ // $c^2 + d^2 \rightsquigarrow 1.5u$

$\widehat{g}_{\mathrm{re}} := \mathrm{Kahan}(a, b, c, d)$

$\widehat{g}_{\mathrm{im}} := \mathrm{Kahan}(b, -a, c, d)$ // $ac + bd$ and $bc - ad \rightsquigarrow 2u$

$\widehat{z}_{\mathrm{re}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{re}}/\widehat{\delta}\right)$

$\widehat{z}_{\mathrm{im}} := \mathrm{RN}\left(\widehat{g}_{\mathrm{im}}/\widehat{\delta}\right)$ // $\mathrm{Re}\,(\widehat{z})$ and $\mathrm{Im}\,(\widehat{z}) \rightsquigarrow u$

return $\widehat{z}_{\mathrm{re}} + i\,\widehat{z}_{\mathrm{im}}$

On architectures supporting the minNumMag and maxNumMag operations, the first two lines may be replaced by the following straight-line code:

$\underline{c} := \mathrm{maxNumMag}(c, d)$;

$\underline{d} := \mathrm{minNumMag}(c, d)$;

// Here, $\underline{c}^2 + \underline{d}^2 = c^2 + d^2$ and $|\underline{d}| \leqslant |\underline{c}|$.

$\widehat{\delta} := \mathrm{RN}(\underline{c}^2 + \mathrm{RN}(\underline{d}^2))$;

Conclusion

By combining Kahan's algorithm with cheaper schemes, we have proposed two complex division algorithms that take advantage of the FMA instruction :

1. A straight-line algorithm, with $E_c(\widehat{z}) \leqslant 5u + O(u^2)$.
   $\hookrightarrow$ bound proved to be asymptotically optimal when p is even.

2. A version in which $|c|$ and $|d|$ are compared to ensure $E_c(\widehat{z}) \leqslant 4.5u + O(u^2)$.

Future work:

- Overhead induced by the use of Kahan's algorithm?
- Extensions to non binary radices and other rounding modes.
- Using scaling techniques to avoid spurious under/overflows while preserving high componentwise accuracy [Smith, 1962], [Stewart, 1985], [Priest, 2004], [Baudin, Smith, 2012].

References


R. P. Brent, C. Percival, and P. Zimmermann, *Error bounds on complex floating-point multiplication*, Mathematics of Computation **76** (2007), 1469–1481.


Michael Baudin and Robert L. Smith, *A robust complex division in Scilab*, October 2012, Available at `http://arxiv.org/abs/1210.4539`.


C.-P. Jeannerod, N. Louvet, and J.-M. Muller, *Further analysis of Kahan's algorithm for the accurate computation of* 2×2 *determinants*, Mathematics of Computation (2012), to appear. Preliminary version available at `http://hal-ens-lyon.archives-ouvertes.fr/ensl-00649347/en/`.


D. M. Priest, *Efficient scaling for complex division*, ACM Trans. Math. Software **30** (2004), no. 4.


Robert L. Smith, *Algorithm 116: Complex division*, Comm. ACM **5** (1962), no. 8, 435.


G. W. Stewart, *A note on complex division*, ACM Trans. Math. Software **11** (1985), no. 3, 238–241.