On Errors and Error Bounds

The value of any arithmetic scheme is reflected in the quality of error bounds an analyst can construct for computations in that arithmetic. Examples that show that one system of arithmetic produces a smaller error in some specific computation are often irrelevant. In real life, we don't know the true answer, so we don't know the error. We may be able to get a bound on the error, either analytically or by interval arithmetic.

Since the bound depends on the style of arithmetic, we may prefer arithmetic that provides lower error bounds for equal analytic effort. Comparing error bounds is tricky, of course, since they reflect the skill of the analyst and the intended audience as well as properties of the arithmetic. That's why it's so gratifying when a single analysis like the one below so graphically demonstrates the superiority of one system of arithmetic over another.

The Last Example on Gradual Underflow?

W. Kahan suggested that Smith's algorithm for complex divide be analyzed under gradual underflow and flush to zero. My analysis surprised me in the strength of its argument for gradual underflow.

Hopefully we all agree that complex divide is a relevant useful operation. I hope that everyone can follow the error analysis below. We have two complex numbers p+iq and r+is. p, q, r, and s are all single precision KCS numbers, normalized or zero, but r and s are not both zero. The usual formula for the quotient is

$$\left\{\frac{pr+qs}{r^2+s^2}\right\} + \left\{\frac{qr-ps}{r^2+s^2}\right\} i$$

but this familiar formula is prone to intermediate underflow and overflow and requires 3+, 6*, and 2/ operations. Surprisingly, the computation can be rearranged with fewer operations, 3+, 3*, and 3/, with less chance of intermediate over/underflow:

if $|s| \le |r|$ then compute

$$z := \left\{\frac{p+q\left(\frac{s}{r}\right)}{r+s\left(\frac{s}{r}\right)}\right\} + \left\{\frac{q-p\left(\frac{s}{r}\right)}{r+s\left(\frac{s}{r}\right)}\right\} i$$

else compute

$$z := \left\{\frac{p\left(\frac{r}{s}\right)+q}{r\left(\frac{r}{s}\right)+s}\right\} + \left\{\frac{q\left(\frac{r}{s}\right)-p}{r\left(\frac{r}{s}\right)+s}\right\} i$$

This formula dates back at least to R. Smith in 1963; you can find it in Knuth volume II, p. 195.

CLAIM: If no exception other than underflow or inexact is raised, the indicated formulas produce a computed complex result $z + \delta z$ that differs from the correct result z by no more than a few units in the last place of $|z|$.

This claim is possible in KCS with gradual underflow but no comparably simple statement can be made in a system with flush to zero or UN symbols in place of gradual underflow. Note that the claim does not imply that both components of complex z are individually accurate to a few units in the last place.

The computation should be executed in normalizing mode. In Warning mode, Invalid Result may be raised on one of the final divides if p and q are tiny.

I won't analyze the entire computation; instead, let's just look at the denominator

$$D \equiv r + s \cdot \left(\frac{s}{r}\right)$$

in the case $|s| \le |r|$. r must be normalized but s can be zero, denormalized or normalized. The model of arithmetic is

computed(x op y) := (x op y)(1 + ε) + δ

where $|\varepsilon| \le \rho = 2^{-24}$ and $|\delta| \le \mu = 2^{-127}$ for flush to zero (FZ) and 2^{-150} for gradual underflow (GU); further we may choose ε and δ so that $\varepsilon\delta = 0$.

Denoting computed values by ~ we find

$$\widetilde{\frac{s}{r}} = \frac{s}{r}(1+\varepsilon_1)+\delta_1 .$$

This can't overflow; in fact $s/r \le 1$.

$$\widetilde{s\cdot\left(\frac{s}{r}\right)} = s\cdot\widetilde{\left(\frac{s}{r}\right)}(1+\varepsilon_2)+\delta_2$$

This can't overflow either.

$$\tilde{D} \equiv \widetilde{r+s\cdot\left(\frac{s}{r}\right)} = \left\{r + \widetilde{s\cdot\left(\frac{s}{r}\right)}\right\}(1+\varepsilon_3) .$$

A true add always occurs so no underflow is possible; overflow can happen if r and s are huge.

$$\tilde{D} = r(1+\varepsilon_3) + s\cdot\left(\frac{s}{r}\right)(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) + s\cdot\delta_1(1+\varepsilon_2)(1+\varepsilon_3) + \delta_2(1+\varepsilon_3)$$

Thus the absolute error

$$|\tilde{D}-D| = \left| r\varepsilon_3 + s\cdot\left(\frac{s}{r}\right)\left\{(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)-1\right\} + s\,\delta_1(1+\varepsilon_2)(1+\varepsilon_3) + \delta_2(1+\varepsilon_3) \right|$$

$$\le |r|\rho + \left|\frac{s^2}{r}\right|\cdot 3\rho + |s|\cdot\mu + \mu$$

$$\le \left(|r| + 3\left|\frac{s^2}{r}\right|\right)\rho + \mu$$

and the relative error

$$\left|\frac{\tilde{D}-D}{D}\right| \le \left(\frac{|r|+3\left|\frac{s^2}{r}\right|}{|r|+\left|\frac{s^2}{r}\right|}\right)\rho + \frac{\mu}{|r+s(\frac{s}{r})|} \le 2\rho + \frac{\mu}{|r|}$$

So the relative uncertainty in the denominator is bounded by:

	GU	FZ
ρ	2^{-24}	2^{-24}
minimum $\lvert r\rvert$	2^{-126}	2^{-126}
μ	2^{-150}	2^{-127}
Relative error bound, independent of r	$3 \cdot 2^{-24}$	$\frac{1}{2}$

In this analysis, with gradual underflow the uncertainty due to underflow is actually less than the uncertainty due to roundoff. With flush to zero, the uncertainty due to underflow overwhelms the uncertainty due to roundoff.

The error bound $2\rho + \frac{\mu}{|r|}$ is realistic. For instance, take

	GU	FZ
r	$3 \cdot 2^{-126}$	$3 \cdot 2^{-126}$
s	2^{-126}	2^{-126}
$\widetilde{\left(\frac{s}{r}\right)}$	$.AAAAAB_{16}\ 2^{-1}$	$.AAAAAB_{16}\ 2^{-1}$
$\widetilde{s\cdot\frac{s}{r}}$	$.555556_{16}\ 2^{-126}$	0
$\widetilde{r+s\left(\frac{s}{r}\right)} \equiv \tilde{D}$	$3.555558_{16}\ 2^{-126}$	$3 \cdot 2^{-126}$
D correct	$(10/3)\ 2^{-126}$	$(10/3)\ 2^{-126}$
Relative error	$.8 \cdot 2^{-24}$	.1
Computed bound on error	$2.4 \cdot 2^{-24}$	.17