Floating Point Arithmetic (II)

In n-digit base 10 floating point we have discovered that

$$\left|\frac{x - fl(x)}{x}\right| < \frac{1}{2} 10^{-n}$$

for any x which is in-range. This is good! But we can still get poor results if we're not careful.

Example.

$$x = \underbrace{1.0\ldots0}_{n}49\ldots90\ldots0$$

$$y = 1.0$$

Both are in-range, and y is a representable number with $y = fl(y)$.

However,

$$fl(x) = 1.0 = fl(y)$$

so the relative error in the result of computing $x-y$ is

$$\left| \frac{fl(fl(x)-fl(y)) - (x-y)}{(x-y)} \right| = 1.$$

That is terrible!

We are now going to see that the problem occurs (only) when $|x-y|$ is small compared to x, or $|1-y/x|$ is small.

Proposition. If $x>y>0$ are in-range for n-digit floating point base 10, then

$$\left| \frac{fl(fl(x)-fl(y)) - (x-y)}{x-y} \right| < \left(\frac{3}{2}10^{-n}\right)10^{p} \text{ if } 1-y/x > 10^{-p}.$$

Suppose that

$$x = a \times 10^r \qquad y = b \times 10^s$$

where $1 \leq a, b < 10$. We compute

$$fl(x) = fl(a \times 10^r) = fl(a) \times 10^r$$
$$fl(y) = fl(b) \times 10^s$$

so

$$fl(x) - fl(y) = (fl(a) - fl(b) \times 10^{s-r})10^r$$

and

$$fl(fl(x) - fl(y)) = fl(fl(a) - fl(b) \times 10^{s-r})10^r$$

Similarly,

$$x - y = (a - b10^{s-r})10^r$$

Now we may assume $x > y > 0$.
Further

$$\frac{fl(fl(x)-fl(y))-(x-y)}{x-y} = \frac{fl(fl(a)-fl(b)10^{s-r}) - (a-b10^{s-r})}{(a-b10^{s-r})}$$

Now $1 \leq a < 10$, so $1 \leq fl(a) \leq 10$.
Further, $a10^r = x > y = b10^s$, so

$$a > b10^{s-r} \quad \text{and} \quad fl(a) > fl(b)10^{s-r}.$$

Thus $fl(a) - fl(b)10^{s-r} < 10$, and
the absolute roundoff error ~~between~~

$$|fl(fl(a)-fl(b)10^{s-r}) - (fl(a)-fl(b)10^{s-r})|$$

is less than $\frac{1}{2}10^{-n}$. We write

$$fl(fl(a)-fl(b)10^{s-r}) = fl(a) - fl(b)10^{s-r} + \epsilon$$

where $|\epsilon| < \frac{1}{2}10^{-n}$.

Now we have

$$\left|fl(fl(a)-fl(b)10^{s-r})-(a-b10^{s-r})\right|$$

$$=\left|fl(a)-fl(b)10^{s-r}-a+b10^{s-r}+\epsilon\right|$$

$$=\left|fl(a)-a+(b-fl(b))10^{s-r}+\epsilon\right|$$

$$\leq |fl(a)-a|+|b-fl(b)|10^{s-r}+|\epsilon|$$

Now $1 \leq a,b < 10$, so $|fl(a)-a| < \frac{1}{2}10^{-n}$, and $|fl(b)-b| < \frac{1}{2}10^{-n}$. Further,

$$10 > a > b10^{s-r} > 10^{s-r},$$

so $s-r < 1$ and (since s,r are integers) we have $s-r \leq 0$. Thus

$$|fl(b)-b|10^{s-r} \leq |fl(b)-b| < \frac{1}{2}10^{-n}.$$

We have now shown that

$$\left| fl\left(fl(a) - fl(b)10^{s-r}\right) - \left(a - b10^{s-r}\right) \right| < \frac{3}{2}10^{-n}$$

Let's consider the denominator!

$$\left(a - b10^{s-r}\right) = a\left(1 - \frac{b10^{s}}{a10^{r}}\right)$$

$$= a\left(1 - \frac{y}{x}\right)$$

Thus if $\left(1 - \frac{y}{x}\right) > 10^{-p}$ we have

$a\left(1 - \frac{y}{x}\right) > 10^{-p}$ and

$$\left| \frac{fl(fl(x) - fl(y)) - (x-y)}{(x-y)} \right| < \left(\frac{3}{2}10^{-n}\right)10^{p}. \quad \square$$

Phew! That was a lot of work. Note that this can't be improved much (extra credit: prove it!).

Principle. Avoid subtracting nearly equal numbers in floating point arithmetic.

We can have similar problems when computing something like

$$7 \times \frac{500!}{499!} = 7 \times 500 = 3500$$

This ought to be exact, but

$7 \times 500!$ will overflow
$7/499!$ will underflow

So you have to do the algebra first.

Principle. Avoid multiplying and dividing by very large or very small numbers.

We now consider addition a little more closely. Recall that for $x, y > 0$

$$fl(fl(x) + fl(y)) = fl((1+\delta_x)x + (1+\delta_y)y)$$

$$= fl(x + y + \delta_x x + \delta_y y)$$

$$= (1 + \delta_{x+y})(x + y + \delta_x x + \delta_y y)$$

$$= x + y + \delta_x x + \delta_y y + \delta_{x+y} x + \delta_{x+y} y + \delta_{x+y}\delta_x x + \delta_{x+y}\delta_y y$$

where $|\delta_x|, |\delta_y|, |\delta_{x+y}| < \frac{1}{2} 10^{-n}$. We get

$$fl(fl(x) + fl(y)) = (1+\epsilon)(x+y)$$

with $|\epsilon| < \sim 2 \times 10^{-n}$.

So suppose we compute

$$\sum_{i=0}^{n} x_i = (\cdots(\underbrace{\underbrace{(x_0+x_1)}_{S_1}+x_2)}_{S_2}\cdots)+x_n)$$

in floating point, letting

$\hat{S}_i$ be the (floating point) result

at the i-th step. We saw

$$\hat{S}_1 = (1+\epsilon_1)(x_0+x_1)$$

$$\begin{aligned}\hat{S}_2 &= (1+\epsilon_2)(x_2+(1+\epsilon_1)(x_0+x_1)) \\ &= (1+\epsilon_2)(x_0+x_1+x_2+\epsilon_1(x_0+x_1)) \\ &= x_0+x_1+x_2+\epsilon_1(x_0+x_1) \\ &\quad + \epsilon_2(x_0+x_1+x_2) + \epsilon_1\epsilon_2(x_0+x_1)\end{aligned}$$

$$\begin{aligned}\hat{S}_n = x_0+\dots+x_n &+ \cancel{\epsilon_1 S_1 + \epsilon_2 S_2 + \dots + \epsilon_{n-1}S_{n-1}} \\ &+ \epsilon_1 S_1 + \dots + \epsilon_{n-1}S_{n-1} + (\text{terms with more products of } \epsilon_i\text{'s})\end{aligned}$$

Since all ~~terms~~ x_i are positive and all ϵ_i are positive in the worst-case scenario, we have

$$\hat{S}_n - (x_0 + \dots + x_n) \geq \epsilon_1 S_1 + \dots + \epsilon_{n-1} S_{n-1}$$

which may be as large as

$$\frac{1}{2} 10^{-n} (S_1 + \dots + S_{n-1})$$

Now given positive x_i the worst case scenario is that x_0 is the largest summand (and essentially equal to S_n) so we get

$$\hat{S}_n - S_n \approx \frac{1}{2} 10^{-n} (n-1) S_n$$

and

$$\frac{|\hat{S}_n - S_n|}{|S_n|} \approx \left(\frac{1}{2} 10^{-n}\right)(n-1).$$

Principle. Avoid summing large numbers of terms – the results may be less accurate than you would like!

<n-digit-floating-examples.nb>