Loss of significance (2.2)

It may be surprising, but after the last example, we can start to see that even adding or subtracting floating point numbers can lead to catastrophic results.

Example. $y = x - \sin(x)$, evaluated to 10 digits at $x = 1/15$.

$$x = 0.66666\ 66667 \times 10^{-1}$$

$$\sin x = 0.66617\ 29492 \times 10^{-1}$$

$$x - \sin x = 0.000\underset{4}{\blacksquare}9\ 37175 \times 10^{-1}$$

$$x - \sin x = 0.4937175\underset{\substack{\uparrow \\ \text{what goes here?}}}{\underline{\qquad}} \times 10^{-4}$$

When floating point math requires the computer to "invent" digits, they are always 0s. These are not significant digits since they are probably wrong.

Note: This shows you the difference between $fl(x-y)$, which is close to $x-y$, since $fl(x-y) = (x-y)(1+\delta)$, $|\delta| \leq 2^{-24}$, And $fl(fl(x)-fl(y))$ which isn't close.

Theorem. Let x and y be normalized floating point machine numbers, with $x > y > 0$. If

$$2^{-p} \leq 1-(y/x) \leq 2^{-q}, \quad p,q \in \mathbb{N}$$

then at most p and at least q significant bits are lost in computing $x-y$.

Proof. (We prove the second part)

Let $x = r \times 2^n$, $y = s \times 2^m$, $\frac{1}{2} \le r, s < 1$.

To compute $x-y$, since $x>y$, we may need to shift y. So we really compute

$$x-y = (r \times 2^n) - (s\, 2^{m-n} \times 2^n)$$

$$= (r - s\, 2^{m-n}) \times 2^n$$

Now

$$r - s\, 2^{m-n} = r\left(1 - \frac{s2^m}{r2^n}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-q},$$

so to normalize $x-y$, we need to shift at least q bits left (introducing q bits of spurious zeros). □

Example. How many bits are lost in
$$37.593621 - 37.584216 ?$$

We compute

$$1 - (y/x) = 0.000\,250\,175\,422$$

Now we recall that

$$\log_2 x = \frac{\log_{10} x}{\log_{20} 2}$$

so we compute

$$\log_2 0.000\,250\,175\,422 = -11.9647723$$

so

$$2^{-11} > 0.000\,250\,175\,422 > 2^{-12}$$

and between 11 and 12 bits are lost.

How can we avoid this?

Basically, the short answer is to avoid trying to compute such subtractions! Often, this can be done with some algebra.

Example.

$$f(x) = \sqrt{x^2+1} - 1.$$

This formula for $f(x)$ is bad for x near 0. But

$$f(x) = \frac{\left(\sqrt{x^2+1} - 1\right)\left(\cancel{1}\sqrt{x^2+1} + 1\right)}{\sqrt{x^2+1} + 1}$$

$$= \frac{x^2+1 - 1}{1+\sqrt{x^2+1}} = \frac{x^2}{1+\sqrt{x^2+1}}.$$

Example. Write code to evaluate

$$f(x) = x - \sin x.$$

The idea is to observe that

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots$$

and so

$$x - \sin x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots$$

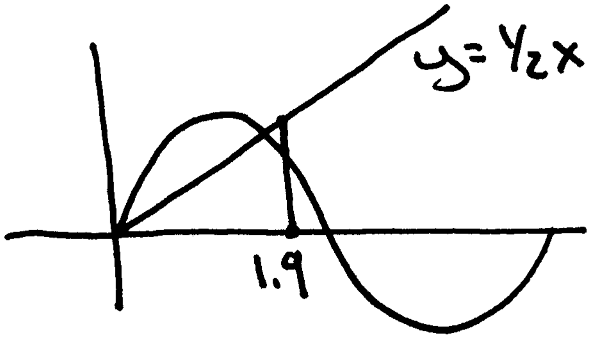
Now it's more efficient to evaluate $x - \sin x$ directly when it's safe. From the theorem, we see that if

$$1 > 1 - \frac{\sin x}{x} \geq 1/2,$$

we will have at most one bit of error. Now this requires

$$x - \sin x \geq 1/2\, x, \quad 1/2\, x \geq \sin x.$$

It turns out



$\sin 1.9 = 0.946300088$

$\frac{1}{2}(1.9) = .95$

so for $x \geq 1.9$, we can subtract safely.

This leaves us using the Taylor expansion on $[0, 1.9]$. It turns out that ten terms give x within 10^{-16} on this interval.

Example. $f(x) = e^x - e^{-2x}$.

(Similar trick involving Taylor series)

Example. Criticize the line of code

$$y := \cos^2\theta - \sin^2\theta;$$

When $\cos\theta = \sin\theta$, this will be inaccurate. Use

$$y := \cos 2\theta;$$

instead.

Example. Criticize

$$y := \ln x - 1;$$

For x near e, this will be inaccurate. But we observe

$$\ln x - 1 = \ln x - \ln e$$

$$= \ln(x/e)$$

and this form is ok to use.