1 Errors (Section 2.2)

Some errors are unavoidable, but others may be reduced. For example,

- Representing a real number on a standardized binary computer.
- The subtraction of similar numbers.

(Example.) Suppose we have a hypothetical decimal machine[a] with 5 digits of precision to the right of the period. Define[b] $x = \mathbf{0.12417}515$ and $y = \mathbf{0.12411}405$. Then, the exact difference is given by

$$x - y = \mathbf{0.00006}110.$$

We know that $\text{fl}(x) = 0.12418$ and $\text{fl}(y) = 0.12411$. Then, the floating-point difference is given by

$$\text{fl}(x) - \text{fl}(y) = 0.00007.$$

The relative error is then

$$\frac{|\,\text{fl}(x-y) - (\text{fl}(x) - \text{fl}(y))|}{|x-y|} = \frac{|0.00006110 - 0.00007|}{|0.00006110|} \approx 0.145662847791.$$

[a] i.e., a machine that represents number in decimal, base 10, form instead of the usual binary, base 2, form.
[b] The bolded number represents the digits that aren't cut off.

1.1 Subtraction of Nearly Equal Quantities

Our goal is to, obviously, avoid some errors. Subtraction (of very similar numbers) is one such case where these errors might occur. An easy way to do this is to try to get rid of subtraction in our operations.

(Example.) Suppose x becomes small (towards 0), and define

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

We can rewrite this as follows:

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

This minimizes the error since there's no subtraction.

Theorem 1.1: Loss of Precision

For positive normalized binary machine numbers x and y where $x > y$, if $2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$ for some p, q, then $x - y$ loses at most p significant digits and at least q significant digits.

(Exercise.) Note that $x - \sin(x)$ involves some error if $x \approx \sin(x)$ when x is small. How do we compute this expression with little cancellation?

We can make use of the Taylor series,

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \dots.$$

We end up with

$$x - \sin(x) \approx \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \frac{x^9}{9!} + \dots.$$

By factoring, we get an expression for which we can use Horner's method on,

$$x - \sin(x) \approx \frac{x^3}{3!}\left(1 - \frac{x^2}{4\cdot 5}\left(1 + \frac{x^2}{6\cdot 7}\left(1 - \frac{x^2}{8\cdot 9}\left(\ldots\right)\right)\right)\right).$$

We use the inequality with $p = 1$ to determine the value of x so only 1 digit of accuruacy is lost in $x - \sin(x)$;

$$2^{-1} = \frac{1}{2} \leq 1 - \frac{y}{x} = 1 - \frac{\sin(x)}{x}, |x| \geq 1.9.$$

In this case, we use this series when $|x| < 1.9$ and the actual expression otherwise. In other words,

$$x - \sin(x) = \begin{cases} \frac{x^3}{3!}\left(1 - \frac{x^2}{4\cdot 5}\left(1 + \frac{x^2}{6\cdot 7}\left(1 - \frac{x^2}{8\cdot 9}\right)\right)\right) & |x| < 1.9 \\ x - \sin(x) & \text{otherwise} \end{cases}.$$

Note that, if x is indeed near 0, we can truncate the series.

1.2 Floating-Point Error Analysis

Let $\circledcirc = \{+, -, \times, /\}$. If x and y are machine numbers, the result is represented by another machine number,

$$x \circledcirc y = \text{fl}(x \circledcirc y).$$

(Example.) Consider the decimal machine with 5 digits (uses 10 digits for intermediate computations). Let x and y be machine numbers so that

$$x = 0.31426 \times 10^3, \quad y = 0.32577 \times 10^5.$$

Then,

$$\begin{aligned} x + y &= 0.31426 \times 10^3 + 0.32577 \times 10^5 \\ &= 0.0031426 \times 10^3 10^2 + 0.32577 \times 10^5 \\ &= 0.0031426000 \times 10^5 + 0.3257700000 \times 10^5 \\ &= 0.3289126000 \times 10^5 \end{aligned}$$

Then, rounding to the fifth digit gives us

$$\text{fl}(x + y) = 0.32891 \times 10^5.$$