Today’s Overview

- We will discuss
  - Floating-point number system
  - Roundoff errors
  - Loss of significance

Single precision

- Recall $(-1)^s \times 2^{c-127} \times (1.f)_2$
- $0 < c < (11\,111\,111)_2 = 255$

  $\Rightarrow -127 < c - 127 < 128$
- $1 \leq (1.f)_2 = 2 - 2^{-23}$
- Largest machine number: 3.4×10^{38}
- Smallest machine number: 1.2×10^{-38}
- Machine epsilon: smallest number $1 + \epsilon \neq 1$
  - $\epsilon = 2^{-24} \approx 6 \times 10^{-8} \Rightarrow$ **7 significant decimal digits**

Double precision

- Double precision $(-1)^s \times 2^{c-1023} \times (1.f)_2$
- 11 bits for exponent and 52 for mantissa
- $-1022 \leq c \leq 1023$
- Largest machine number: 1.8×10^{308}
- Smallest machine number: 2.2×10^{-308}
- Machine epsilon: $2^{-53} \approx 1.11 \times 10^{-16}$
  - **15 significant decimal digits**

Computer errors

- The process of replacing a number by its nearest machine number is called correct rounding; the error involved is called roundoff error.
- In general, we want to know how large roundoff error can be!
- If a number is overflow, roundoff error could be huge.

Rounding

- Suppose

$$x = (0.1b_2b_3b_4\dots b_{24}b_{25}b_{26}\dots)_2 \times 2^m$$

- Round down

$$x_- = (0.1b_2b_3b_4\dots b_{24})_2 \times 2^m$$

- Round up

$$x_+ = \left[(0.1b_2b_3b_4\dots b_{24})_2 + 2^{-24}\right] \times 2^m$$

Unit roundoff error

$$|x - x_-| \le \frac{1}{2}|x_+ - x_-| = 2^{-25+m}$$

$$\left|\frac{x - x_-}{x}\right| \le \frac{2^{-25+m}}{(0.1b_2b_3b_4\ldots)_2 \times 2^m} \le \frac{2^{-25}}{2^{-1}} = 2^{-24} = u$$

The unit roundoff error for a 32 bit binary computer is $u = 2^{-24}$, which is equivalent to machine epsilon.

Errors in arithmetic operations

Example

- Suppose we have a five-place decimal machine and have two numbers to add

$$x = 0.37218 \times 10^4, \qquad y = 0.71422 \times 10^{-1}$$

- Perform operations in double-length

$$\begin{aligned} x &= 0.37218\,00000 \times 10^4 \\ y &= 0.00000\,71422 \times 10^4 \\ \hline x + y &= 0.37218\,71422 \times 10^4 \end{aligned}$$

- Nearest machine number

$$z = 0.37219 \times 10^4$$

- Error involved

$$\frac{|x+y-z|}{|x+y|} = \frac{0.00000\,28578 \times 10^4}{0.37218\,71422 \times 10^4} \approx 0.77 \times 10^{-5}$$

FP machine number

- Define fl(x) be the FL machine number that corresponds to x.
- The function fl depends on the computer.
- For a 32-bit word-length computer, we have

$$\frac{|x - \mathrm{fl}(x)|}{|x|} \le u \qquad \left(u = 2^{-24}\right)$$

- The inequality can be expressed by

$$\mathrm{fl}(x) = x(1+\delta) \qquad \left(|\delta| \le 2^{-24}\right)$$

Arithmetic operations

$$\mathrm{fl}(x \odot y) = (x \odot y)(1 + \delta) \qquad (|\delta| \leq 2^{-24})$$

Example:

If x, y are real numbers in a 32-bit computer, estimate the relative roundoff error in computing (x+y).

Loss of Significance

Significant digits

- Significance of the digits diminishes from left to right.
- Every measured quantity involves an error whose magnitude depends on the nature of the measuring device.
- If a meter stick is used, it is not reasonable to get precision better than 1 millimeter, e.g., 2.3453 meters.
- The least significant digit should be in error by at most 5 units, i.e., measured result is rounded correctly!

Infinite precision

- If the side of a square is reported to be $s = 0.736$ meter, then error does not exceed 5 units in the third decimal place.
- The diagonal of the square

$$s\sqrt{2} \approx 0.104\,086\,1182 \times 10^1$$

should be reported as 0.1041×10^1.

- The infinite precision in $\sqrt{2}$=1.41421… does not convey any more precision to $s\sqrt{2}$ than was already present in s.

Loss of significance

- Consider to execute the statement at x=1/15

$$y \leftarrow x - \sin(x)$$

- Then

$$\begin{aligned} x &\leftarrow 0.66666\,66667 \times 10^{-1} \\ \sin(x) &\leftarrow 0.66617\,29492 \times 10^{-1} \\ x - \sin(x) &\leftarrow 0.00049\,37175 \times 10^{-1} \\ x - \sin(x) &\leftarrow 0.49371\,75000 \times 10^{-4} \end{aligned}$$

- Correct value

$$\frac{1}{15} - \sin\left(\frac{1}{15}\right) \approx 0.49371\,74327 \times 10^{-4}$$

Theorem on loss of precision

Exact how much significant binary digits are lost in subtraction x-y when x is close to y?

Let x and y be normalized floating-point machine numbers, where $x > y > 0$. If $2^{-p} \leq 1 - (y/x) \leq 2^{-q}$ for some positive integers p and q, then at most p and at least q significant binary bits are lost in the subtraction $x - y$.

- The closeness of x and y is measured by $|1 - \frac{y}{x}|$.
- Double precision may help.
- Taylor series may help

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

Avoid loss of significance

- Double precision
- Taylor series
- Rationalization
- Trigonometric identities
- Logarithmic properties
- Range reduction