MACM 316 Lecture 2 - More Computer Arithmetic

Alexander Ng

Wednesday, January 8, 2025

1 Floating Point Decimal Normalization

Can we write all real numbers in normalized scientific notation?

$$
\begin{aligned}
732.5051 &\to +0.7325051 \times 10^{+3} \\
-0.005612 &\to -0.5612 \times 10^{-2}
\end{aligned}
$$

For $x \in \mathbb{R}$, we can express it as:

$$x = \pm r \times 10^{\pm n}, \quad \text{where } \frac{1}{10} \leq r \leq 1.$$

In binary, we write:

$$x = \pm q \times 2^{\pm m}, \quad \text{where } \frac{1}{2} \leq q < 1.$$

Here, q is the mantissa and m is the integer exponent.

We limit r and q so that when $k < 1/\text{BASE}$, we can shift the decimal place and normalize the number further. When we have $1.x$, we rewrite it as $0.x \times \text{base}^1$.

2 Rounding or Chopping (Sources of Error)

Given $x = 0.a_1a_2 \ldots a_na_{n+1} \ldots a_m$ using m digits, rounding to n places follows:

- If $0 \leq a_{n+1} < 5$, then $x = 0.a_1a_2 \ldots a_n$.
- If $5 \leq a_{n+1} \leq 9$, then $x = 0.a_1a_2 \ldots (a_n + 1)$.

Example:

$$\text{round}(0.125) = 0.13,$$
$$\text{round}(-0.125) = -0.13.$$

Instead of rounding, truncation follows:

$$x = 0.a_1a_2 \ldots a_n.$$

Truncation introduces larger errors but is computationally cheaper than rounding.

3 Error

We define:

- Absolute error: $|p - p^*|$.
- Relative error: $\frac{|p-p^*|}{|p|}$.

Absolute error is used when magnitude matters, particularly for small values. Relative error is preferred when values differ in scale.

Example:

$$\text{Exact: } 0.1, \quad \text{Approximate: } 0.099,$$
$$\text{Relative Error: } \frac{|0.1 - 0.099|}{0.1} = 0.01.$$

t	5×10^{-t}	Is error within bound?
0	5	✓
1	0.5	✓
2	0.05	✓
3	0.005	×

Since $0.01 < 5 \times 10^{-2}$ but not 5×10^{-3}, we have two significant digits.

4 Computations and Machine Representation

Let $\mathrm{fl}(x)$ denote the machine representation of x. Computations on a machine follow:

$$\mathrm{fl}(\mathrm{fl}(x) + \mathrm{fl}(y)).$$

Each step introduces an error.

Example:

$$\begin{aligned} p &= 0.54617, \quad q = 0.54601, \\ r &= p - q = 0.00016. \end{aligned}$$

With 4-digit rounding,

$$\begin{aligned} p^* &= 0.5462, \quad q^* = 0.5460, \\ r^* &= p^* - q^* = -0.0002. \end{aligned}$$

Relative error:

$$\frac{|r - r^*|}{|r|} = 0.25.$$

A high relative error results when subtracting close numbers.

5 Minimizing Error

Consider computing $f(x) = \frac{1-\cos x}{x^2}$ for $\bar{x} = 1.2 \times 10^{-5}$.

With 10-digit rounding:

$$\begin{aligned} c &= \mathrm{fl}(\cos \bar{x}) = 0.9999999999, \\ 1 - c &= 0.0000000001. \end{aligned}$$

This results in a large error.

Using $\cos x = 1 - 2\sin^2(x/2)$:

$$f(x) = \frac{1}{2} \left(\frac{\sin(x/2)}{x/2} \right)^2.$$

This provides a more accurate computation.

Conclusion: Avoid subtracting close numbers. Use alternative representations like Taylor series or trigonometric identities.