CS 370 Numerical Computation

Keven Qiu
Instructor: Christopher Batty
Winter 2024

Part I

Floating Point Numbers

Chapter 1

Floating Point Number Systems

The real numbers $\mathbb{R}$ are infinite in extent and in density. The standard (partial) solution is to use floating point numbers to approximate the reals.

Definition: Floating Point Number System

An approximate representation of $\mathbb{R}$ using a finite number of bits.

An analytic solution is an exact solution, whereas a numerical solution is the approximate solution.

We can express a real number as an infinite expansion relative to some base β. For example, in base 10

$$\frac{73}{3} = 24.3333\ldots = 2 \times 10^1 + 4 \times 10^0 + 3 \times 10^{-1} + \cdots$$

After expressing the real number in the desired base β, we multiply by a power of β to shift it into a normalized form:

$$0.d_1 d_2 d_3 d_4 \cdots \times \beta^p$$

where d_i are digits in base β, i.e., $0 \leq d_i < \beta$, normalized implies we shift to ensure $d_1 \neq 0$, and the exponent p is an integer.

Density (or precision) is bounded by limiting the number of digits, t. Extent (or range) is bounded by limiting the range of values for exponent p.

Definition: Floating Point Representation

The finite form

$$\pm 0.d_1 d_2 \cdots d_t \times \beta^p$$

for $L \leq p \leq U$ and $d_1 \neq 0$.

The four integer parameters (β, t, L, U) characterize a specific floating point system F.

Overflow/underflow errors:

- If the exponent p is too big or too small, our system cannot represent the number.
- When arithmetic operations generate such a number, this is called overflow or underflow.
- For underflow, we simply round to 0.
- For overflow, we typically produce a $\pm\infty$ or NaN.

IEEE single precision (32 bits) has $(\beta = 2, t = 24, L = -126, U = 127)$ and IEEE double precision (64 bits) has $(\beta = 2, t = 53, L = -1022, U = 1023)$.

Unlike fixed point, floating point numbers are not evenly spaced.

There are two ways to convert reals to floats:

1. Round-to-nearest: rounds to closest available number in F.
2. Truncation: rounds to next number in F towards 0.

1.1 Measuring Error

Our algorithms will compute approximate solutions to problems.

Let x_{exact} be the true analytical solution and x_{approx} be the approximate numerical solution.

Definition: Absolute Error

$$E_{abs} = |x_{exact} - x_{approx}|$$

Definition: Relative Error

$$E_{rel} = \frac{|x_{exact} - x_{approx}|}{x_{exact}}$$

Relative error is often more useful because it is independent of the magnitudes of the numbers involved and related the number of significant digits in the result.

A result is correct to roughly s digits if $E_{rel} \approx 10^{-s}$ or

$$0.5 \times 10^{-s} \leq E_{rel} < 5 \times 10^{-s}$$

For floating point system F, the relative error between $x \in \mathbb{R}$ and its floating point approximation, $fl(x)$, has a bound, E, such that

$$(1 - E)\,|x| \leq |fl(x)| \leq (1 + E)\,|x|$$

Definition: Machine Epsilon/Unit Round-Off Error

The maximum relative error E for converting a real into a floating point system.

It is defined as the smallest value such that $fl(1+E) > 1$ under the given floating point system.

These definitions give a rule $fl(x) = x(1+\delta)$ for some $|\delta| \leq E$. δ may be positive or negative. E is defined as positive.

For a FP system $F = (\beta, t, L, U)$:

- Rounding to nearest: $E = \frac{1}{2}\beta^{1-t}$.
- Truncation: $E = \beta^{1-t}$.

Example: Find E for $F = (\beta = 10, t = 3, L = -5, U = 5)$.

Under round to nearest:

$$E = \frac{1}{2}(10)^{1-3} = 5 \times 10^{-3}$$

Consider the smallest representable number above 1. We have $1 = 0.100 \times 10^1$ in F. The next largest is 0.101×10^1. For $fl(1+E)$ to exceed 1, we must add $0.0005 \times 10^1 = 5 \times 10^{-3}$ to get halfway to the next number, where rounding occurs.

Under truncation:

$$E = 10^{1-3} = 10^{-2}$$

1.2 Arithmetic with Floating Point

IEEE standard requires that for $w, z \in F$,

$$w \oplus z = fl(w+z) = (w+z)(1+\delta)$$

where $\oplus$ is the floating point addition.

This rule only applies to *individual* FP operations. So it is not generally true that

$$(a \oplus b) \oplus c = a \oplus (b \oplus c) = fl(a+b+c)$$

The result is order-dependent and associativity is broken.

Consider the relative error of $(a \oplus b) \oplus c$ for $a, b, c \in F$.

$$
\begin{aligned}
E_{rel} &= \frac{|(a \oplus b) \oplus c - (a + b + c)|}{|a + b + c|} \\
&= \frac{|(a + b)(1 + \delta_1) \oplus c - (a + b + c)|}{|a + b + c|} \\
&= \frac{|((a + b)(1 + \delta_1) + c)(1 + \delta_2) - a - b - c|}{|a + b + c|} \\
&= \frac{|a + b + c + (a + b)\delta_1 + (a + b + c + (a + b)\delta_1)\delta_2 - a - b - c|}{|a + b + c|} \\
&= \frac{|(a + b)\delta_1 + (a + b)\delta_1\delta_2 + (a + b + c)\delta_2|}{|a + b + c|} \\
&\leq \frac{|(a + b)\delta_1| + |(a + b)\delta_1\delta_2| + |(a + b + c)\delta_2|}{|a + b + c|} && \text{(Triangle inequality)} \\
&\leq \frac{|a + b|\,|\delta_1| + |a + b|\,|\delta_1\delta_2|}{|a + b + c|} + |\delta_2| \\
&\leq \frac{|a + b|\,E + |a + b|\,E^2}{|a + b + c|} + E && (|\delta| \leq E) \\
&\leq \frac{|a + b|}{|a + b + c|}(E + E^2) + E
\end{aligned}
$$

This is the bound on the relative error for $(a \oplus b) \oplus c$. We can weaken the bound slightly to make it symmetric.

$$
\begin{aligned}
E_{rel} &\leq \frac{|a + b|}{|a + b + c|}(E + E^2) + \frac{|a + b + c|}{|a + b + c|}E \\
&\leq \left(\frac{|a| + |b| + |c|}{|a + b + c|}\right)(E + E^2) + \left(\frac{|a| + |b| + |c|}{|a + b + c|}\right)E \\
&\leq \left(\frac{|a| + |b| + |c|}{|a + b + c|}\right)(2E + E^2)
\end{aligned}
$$

This bound will apply to both $(a \oplus b) \oplus c$ and $a \oplus (b \oplus c)$, i.e. order does not matter.

$\frac{|a|+|b|+|c|}{|a+b+c|}$ is small when the denominator is small. This occurs when quantities have differing signs and similar magnitudes, leading to cancellation.

Catastrophic cancellation occurs when subtracting of about the same magnitude, when the input numbers contain error. All signficant digits cancel out, so the result will have no correct digits.

Part II

Interpolation and Splines

Part III

Ordinary Differential Equations

Part IV

Fourier Transforms

Part V

Numerical Linear Algebra