Units:

1. Floating points
2. Interpolation
3. Ordinary Differential Equations
4. Fourier Analysis
5. Numerical Linear Algebra

Grades: Assignment: 32%, 4%, Midterm: 28%, Exam: 40%

Midterm at June 26, 7pm

Printed course notes at Media doc in DC.

1 Floating Point Arithmetic

Real numbers are:

- infinite in extent
- infinite in density

Floating point system is an approximate representation of real numbers using a finite number of bits.

Consider the sum

$$12 + \sum_{i=1}^{100} 0.01 = 13$$

If we perform the sum one addition at a time, retaining two digts of accuracy at each step:

$$12 + 0.01 + 0.01 + \ldots$$

The numerical answer is 12. Performing the addition from the opposite direction will yield the correct answer.

We can express a real number as an infinite expansion relative to some base β. After expressing the real number in the desired base β, we multiply it by a power of β to shift it into a normalized form

$$0.d_1 d_2 d_3 d_4 \cdots \times \beta^{\rho}$$

where

- d_i are digits in base B, i.e., $0 \leq d_i \leq \beta$
- normalized means $d_1 \neq 0$
- exponent ρ is an integer

1.1 Floating Point System

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

Our floating point representation then has the form:

$$\begin{cases} \pm 0.d_1 d_2 \ldots d_t \times \beta^\rho & : L \leq \rho \leq U, d_1 \neq 0 \\ 0 \end{cases}$$

The four integer parameters $\{\beta, t, L, U\}$ characterize a specific floating point system, F.

If $\rho > U$ or $\rho < L$, our system cannot represent the number. When an arithmetic operation generates such a number, it's called overflow or underflow

The floating point standards are almost always directly implemented in the hardware. The two most common standardized floating point systems are:

- IEEE single precision (32 bits): $\beta = 2, t = 24, L = -126, U = 127$
- IEEE double precision (64 bits): $\beta = 2, t = 53, L = -1022, U = 1033$
- Fixed number of digits are the decimal; integer representation scaled by a fixed scale factor; eg. 10234×10^{-3}
- Floating point number systems let the radix point float to represent a wider range.
- Floating point numbers are not evenly spaced

1.2 Measuring Errors

Absolute error:

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

Relative error:

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

Relative error is more useful:

- independent of magnitudes of numbers involved
- related to the number of significant digits in the result

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

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