CS 450: Numerical Anlaysis

Chapter 1 – Scientific Computing

Lecture 2

Floating Point

Edgar Solomonik

Department of Computer Science
University of Illinois at Urbana-Champaign

Review

- Last lecture introduced the notions of *roundoff* and *truncation* error.
  - roundoff error concerns floating point error due to finite precision
  - truncation error concerns error incurred due to algorithmic approximation, e.g. the representation of a function by a finite Taylor series

$$f(x+h) \approx g_x(h) = \sum_{i=0}^{k} \frac{f^{(i)}(x)}{i!} h^i$$

truncation error is $|f(x+h) - g_x(h)| \leq \left| \sum_{i=k+1}^{\infty} \frac{f^{(i)}(x)}{i!} h^i \right| \in O(h^{k+1})$ as $h \to 0$

- To study the propagation of roundoff error in arithmetic we can use the notion of conditioning.

$$\kappa_x(f) = \lim_{h \to 0} \left| \frac{(f(x+h) - f(x))/f(x)}{h/x} \right| = \left| \frac{f'(x)\,x}{f(x)} \right|$$

condition number of f at x

Floating Point Numbers

- **Scientific Notation**

$2.13798 \cdot 10^{17}$ (exponent)

significant digits (significand)

error limited to variations in the least significant digit

- **Significand (Mantissa) and Exponent** Given x with s leading bits $x_0, \ldots, x_{s-1}$

$x_0 . x_1 x_2 \cdots x_{s-1} \cdot 2^{e_0 \ldots e_k}$

significand; k exponent bits $\Rightarrow$ range $[-2^{2^k}, 2^{2^k}]$

UFL

representable normalized numbers

Rounding Error

- **Maximum Relative Representation Error (Machine Epsilon)**

$$\epsilon_{mach} = 2^{1-s}$$ s is # digits in the significand

$$\forall x \in [2^{-2^k}, 2^{2^k}]$$

$$\frac{|fl(x) - x|}{|x|} \leq \epsilon_{mach}$$

Rounding Error in Operations (I)

- **Addition and Subtraction** (subtraction is addition with sign bit flipped)

Catastrophic cancellation:

rel. err. $\leq 10^{-8}$

$1.3728410(4) \times 10^{-6}$ } 9 digits of accuracy

rel. err. $\leq 10^{-3}$

$-\ 1.3728300(2) \times 10^{-6}$

$> 1.1020000 \times 10^{-11}$ } 4 digits of accuracy

absolute error stayed the same,
but magnitude of value decreased significantly

$\Longrightarrow$ large relative error

Rounding Error in Operations (II)

- **Multiplication and Division**

$$2.1314 \times 10^{3}$$
$$\times\ 1.8912 \times 10^{-6}$$

$$x_1.x_2x_3x_4x_5 \times 10^{-3}$$

$$\frac{|fl(fl(x)\cdot fl(y)) - x\cdot y|}{|x\cdot y|} = \text{rel.err}$$

$$\text{rel.err} \le \frac{|(1+\varepsilon_{mach})((1+\varepsilon_{mach})x\cdot(1+\varepsilon_{mach})y) - xy|}{|xy|}$$

$$= \frac{|(1+\varepsilon_{mach})^3 xy - xy|}{|xy|} \simeq \frac{|(1+3\varepsilon_{mach})xy - xy|}{|xy|} \le 3\varepsilon_{mach}$$

Exceptional and Subnormal Numbers

- **Exceptional Numbers** = not normalized

$1/0 = INF$

$0/0 = \infty - \infty = NaN$

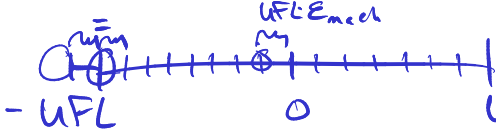
0

- **Subnormal (Denormal) Number Range**

norm: $1.x_1 \ldots x_{s-1}$ s = # digits in significand

$0.x_1x_2x_3 \ldots x_{s-1} \times 2^{-2^k}$

smallest subnormal is $0.0 \ldots 01 \cdot 2^{-2^k}$



← underflow limit for normalized numbers 2^{-2^k}

- **Gradual Underflow: Avoiding underflow in addition**

underflow means result is not a normalized machine number