CS 450: Numerical Anlaysis

Chapter 1 – Scientific Computing

Lecture 2

Floating Point

Vector Norms

Edgar Solomonik

Department of Computer Science
University of Illinois at Urbana-Champaign

January 20, 2018

Floating Point Numbers

- **Scientific Notation**

$\underbrace{3.124} \times 10^8 \qquad [3.123 \;,\; 3.125] \times 10^8$

$2.10\underline{3} \times 10^{-3} \qquad [2.102 \;,\; 2.104] \times 10^8$

rel error is 10^{-3}

- **Significand (Mantissa) and Exponent**

$\underbrace{1.\overbrace{01011}}_{\text{significand}} \times 2^4$ $\qquad$ <u>normalized</u>

Rounding Error

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

$1.\underbrace{0110110}_{\rightarrow}1_m$

$\uparrow 2^{-7}$

digits in significand = $\boxed{12}$

exponent range ~ $[-1023, 1024]$

UFL: $1 \cdot 2^{-1023}$

OFL = $1 \cdot 2^{1024}$

Machine epsilon = $2^{-12} = \varepsilon_{mac}$

- **Rounding Error Analysis**

$1024 = 1 \cdot 2^{10}$

$+ \frac{1}{4}$

2^{-2}

2^9

$1.\underbrace{0\,0 \ldots 0}_{12} \cdot 2^{10}$

2^8

smallest number we can add to 1, so that $fl_m(1+\varepsilon_{mac}) \neq 1$

$1.000 \ldots 0$

$+ 0.000 \ldots 1$ — ε_{mac}

$1.000 \ldots 0 \cdot 2^{-12}$

Rounding Error in Operations

- **Addition and Subtraction**

$$1.0110 \rightarrow \text{approximate to relative precision } 2^{-4}$$

$$-\,1.0100$$

$$= 0.0010$$

$$= 1.0\overbrace{000} \cdot 2^{-3}$$

- **Multiplication and Division**

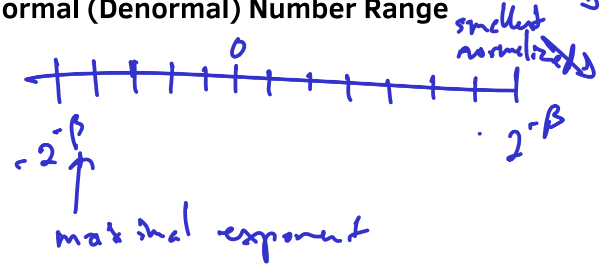
$$z = \underset{1.0111}{\uparrow x} \cdot \underset{1.1100}{\uparrow y} = 1.101\underset{\sim}{0}$$

Overflow leads to a result of ±∞

$$0/0 = \infty + (-\infty) = \text{NaN}$$

Subnormal Numbers

- **Subnormal (Denormal) Number Range**



Smallest subnormal

$0.00001 \cdot 2^{-\beta}$

$1.0000 \cdot 2^{-\beta} = \varepsilon_{mac} \cdot 2^{-\beta}$

$0.00100 \cdot 2^{-\beta}$

$0.01001 \cdot 2^{-\beta}$

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

$x - y = 0$ if and only if $x = y$