Numerical Precision Analysis

Dr. Gordon Cichon

- Interests: Chip-Design, Compiler Construction, Computer Graphics, Signal Processing
- System architect and algorithmic researcher at NVIDIA, Philips, and Infineon
- Compiler experience:
  e.g., independent port of „Self“ compiler to x86
- Studies: CS in Munich (MS in 1996),
  EE in Dresden (PhD in 2004)
- email: gordon@cichon.com

Outline

- Number Representation
- Error Propagation Rules
- Effects of Error Propagation
- Forward Error Analysis
- Example

Definition: Error

- Correct Value: x
- Represented Value: x‘

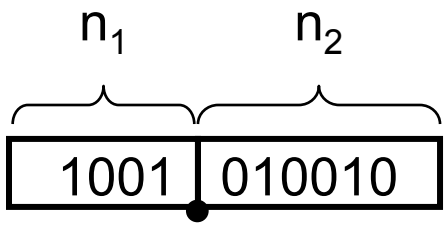
- Absolute Error

$$\Delta x = \left| x' - x \right|$$

- Relative Error

$$\varepsilon_x = \frac{\Delta x}{\left| x \right|} = \frac{\left| x' - x \right|}{\left| x \right|}$$

Fixed Point Representation



$$n = n_1 + n_2$$

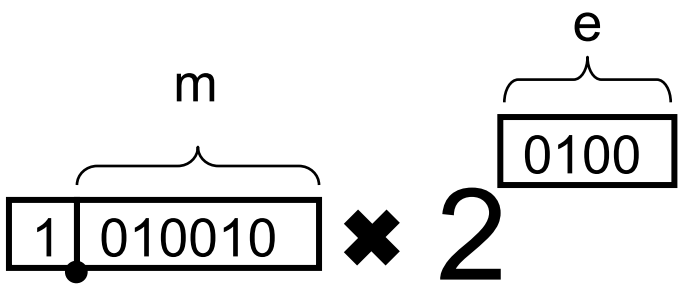
$$\Delta x = 2^{-n_2 - 1}$$

$$\varepsilon_x = \frac{2^{-n_2 - 1}}{|x|}$$

$$x = \pm\left(\alpha_{n_1-1}2^{n_1-1} + \alpha_{n_1-2}2^{n_1-2} + \ldots + \alpha_0 2^0 + \alpha_{-1}2^{-1} + \alpha_{-2}2^{-2} + \ldots + \alpha_{-n_2}2^{-n_2}\right)$$

$$\alpha_i = 0 \qquad \text{or} \qquad \alpha_i = 1$$

Floating Point Representation



$$\Delta x = 2^{-m-1} \cdot |x|$$

$$\varepsilon_x = 2^{-m-1}$$

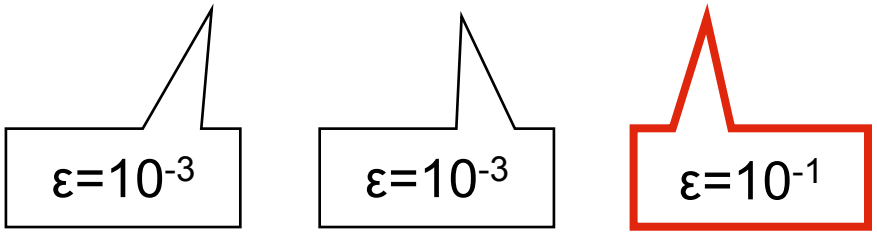
$$x = \pm\left(1 + \alpha_1 2^{-1} + \alpha_2 2^{-2} + \ldots + \alpha_m 2^{-m}\right) \times 2^{\beta}$$

Error Propagation of Primitive Operations

	Addition	Multiplication
	$x = a + b$	$x = a \times b$
Fixed Point	$\Delta x = 0$	$\Delta x = 2^{-n_2 - 1}$
Floating Point	$\varepsilon_x = \varepsilon_a \frac{a}{a+b} + \varepsilon_b \frac{b}{a+b}$	$\varepsilon_x = \varepsilon_a + \varepsilon_b$

Effects

Cancelation: 1.01 - 1.00 = 0.01



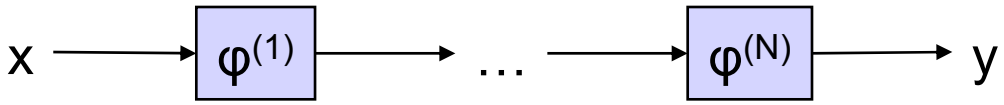
Expected error after N operations:

$$\varepsilon^{(N)} = \sqrt{N} \cdot \varepsilon$$

Forward Error Analysis

- Idea: What happens to the result if an intermediate value changes a little?

$$y = \varphi(x) \qquad y' = y + \Delta y \doteq y + D\varphi(x) \cdot \Delta x$$



$$\varphi = \varphi^{(N)} \circ \ldots \circ \varphi^{(1)}$$

$$\varepsilon^{(i)} \doteq \underbrace{\sum_j \frac{\eta_j}{\varphi(\eta_j)} \cdot \frac{\partial \varphi(\eta_j)}{\delta \eta_j}}_{=: \mathbf{c}} \cdot \varepsilon_{\eta_j}$$

Condition

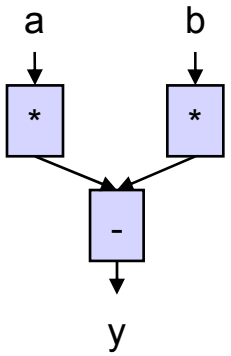
Example

Calculation of a^2-b^2

$\eta_1 := a * a$

$\eta_2 := b * b$

$y := \eta_1 - \eta_2$

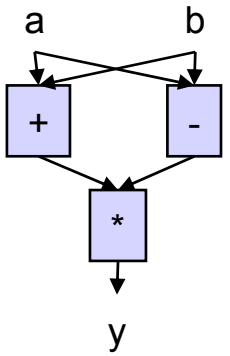


Calculation of (a+b)(a-b)

$\eta_1 := a + b$

$\eta_2 := a - b$

$y := \eta_1 * \eta_2$



a = 1.51, b = 1.50, precision: 3 digits

exact: $3{,}01 \cdot 10^{-2}$

$\eta_1 := 1.51 * 1.51 = 2{,}28$

$\eta_2 := 1.50 * 1.50 = 2{,}25$

$y := 2{,}28\text{-}2{,}25 = \mathbf{3{,}00} \cdot 10^{-2}$

$\eta_1 := 1{,}51 + 1{,}50 = 3{,}01$

$\eta_2 := 1{,}51 - 1{,}50 = 0{,}01$

$y := 3{,}01 * 0{,}01 = \mathbf{3{,}01} \cdot 10^{-2}$

Error Analysis

Calculation of a^2-b^2

$$\varphi^{(1)}(a,b) = \begin{bmatrix} a^2 \\ b^2 \end{bmatrix}$$

$$\varphi^{(2)}(\eta_1,\eta_2) = \eta_1 - \eta_2$$

$$\varepsilon = 2a\varepsilon_a - 2b\varepsilon_b + a^2\varepsilon_{\eta_1} - b^2\varepsilon_{\eta_2} + \left(a^2 - b^2\right)\varepsilon_y$$

$$\varepsilon_{\min} = 2a\varepsilon_a - 2b\varepsilon_b + \left(a^2 - b^2\right)\varepsilon_y$$

$$\varepsilon = a^2\varepsilon_{\eta_1} - b^2\varepsilon_{\eta_2}$$

Calculation of (a+b)(a-b)

$$\varphi^{(1)}(a,b) = \begin{bmatrix} a+b \\ a-b \end{bmatrix}$$

$$\varphi^{(2)}(\eta_1,\eta_2) = \eta_1 \cdot \eta_2$$

$$\varepsilon = 2a\varepsilon_a - 2b\varepsilon_b + \left(a^2 - b^2\right)\left(\varepsilon_{\eta_1} + \varepsilon_{\eta_2} + \varepsilon_y\right)$$

$$\varepsilon = \left(a^2 - b^2\right)\left(\varepsilon_{\eta_1} + \varepsilon_{\eta_2}\right)$$

Summary & Further Reading

- Number Representation in Fixed Point and Floating Point
- Error Propagation Rules
- Forward Error Analysis
- Application Example

- Josef Stoer: „Numerische Mathematik 1“, Springer, Berlin, 1971 (In German)