

Elementos de Cálculo Numérico/Cálculo Numérico

Clase 1

Primer Cuatrimestre 2021

Notación O y o

Dadas $f(x), g(x)$ funciones, $A = x_0, x_0^\pm, \pm\infty$

- $f(x) = O(g(x))$ cuando $x \rightarrow A$ sii
existe $C > 0$ tal que

$$\left| \frac{f(x)}{g(x)} \right| \leq C, \quad x \rightarrow A$$

- $f(x) = o(g(x))$ cuando $x \rightarrow A$ sii

$$\left| \frac{f(x)}{g(x)} \right| \rightarrow 0, \quad x \rightarrow A$$

- $f(x) \sim g(x)$ sii $f(x) = O(g(x))$ y $g(x) = O(f(x))$

Notación O y o: ejemplos

1 $\sin(x) = O(x)$ si $x \rightarrow 0$

2 $1 - \cos(x) = O(x^2)$ si $x \rightarrow 0$

3 $\ln(x) = o(x)$ si $x \rightarrow +\infty$

4 $\ln(x) = o(x^{-1})$ si $x \rightarrow 0^+$

5 $1 + 2 + \dots + n = O(n^2)$ si $n \rightarrow +\infty$

6 $5n^3 + n^2 = O(n^3)$ si $n \rightarrow +\infty$

7 $1 + 4 + \dots + n^2 = O(n^3)$ si $n \rightarrow +\infty$

8 $\sin(x) \sim x$ si $x \rightarrow 0$

Notación O y o: propiedades

- 1 Si $f(x) = o(g(x))$, entonces $f(x) = O(g(x))$
- 2 Si $f(x) = O(g(x))$ y $g(x) = O(h(x))$, entonces

$$f(x) = O(h(x))$$

- 3 Si $f(x) = O(g(x))$ y $g(x) = o(h(x))$, entonces

$$f(x) = o(h(x))$$

- 4 Si $f_1(x) = O(g_1(x))$ y $f_2(x) = O(g_2(x))$, entonces

$$f_1(x) + f_2(x) = O(|g_1(x)| + |g_2(x)|)$$

- 5 \sim es una relación de equivalencia

Errores

Valor exacto: y

Valor aproximado: \hat{y}

Error absoluto: $E_{\text{abs}} = |y - \hat{y}|$

Error relativo:

$$E_{\text{rel}} = \frac{|y - \hat{y}|}{|y|}$$

Ejemplo: $y = 0.000032173$, $\hat{y} = 0.000032174$

$$E_{\text{abs}} = |0.000032173 - 0.000032179| = 6 \times 10^{-9}$$

$$E_{\text{rel}} = \frac{6 \times 10^{-9}}{3.2173 \times 10^{-5}} \cong 2 \times 10^{-4}$$

Dígitos significativos

$y = 32.17$: cuatro dígitos significativos

$y = 0.00003217$: cuatro dígitos significativos

$y = 32.1700$: seis dígitos significativos

Ejemplo: $y = 0.000032173$, $\hat{y} = 0.000032179$

cuatro dígitos significativos correctos $\cong E_{\text{rel}}$

Ejemplo: $y = 0.1302$, $\hat{y} = 0.1299$

un dígitos significativo correcto $\not\cong E_{\text{rel}} = 2.3 \times 10^{-4}$

Dígitos significativos

$\text{round}_k(y)$: redondeo a k dígitos

k dígitos correctos: $\text{round}_k(y) = \text{round}_k(\hat{y})$

No está bien definido

$$y = 0.19949 \text{ y } \hat{y} = 0.19951$$

$$\text{round}_4(y) = \text{round}_4(\hat{y}) = 0.1995,$$

$$\text{round}_3(y) = 0.199 \neq \text{round}_3(\hat{y}) = 0.200,$$

$$\text{round}_2(y) = \text{round}_2(\hat{y}) = 0.20,$$

k dígitos correctos: máx k tal que $\text{round}_k(y) = \text{round}_k(\hat{y})$

Depende de la base elegida!

Números representables

t : número de dígitos

m : mantisa, número entero entre 2^t y $2^{t+1} - 1$

e : exponente, número entero entre e_{\min} y e_{\max}

s : signo, $s = \pm 1$

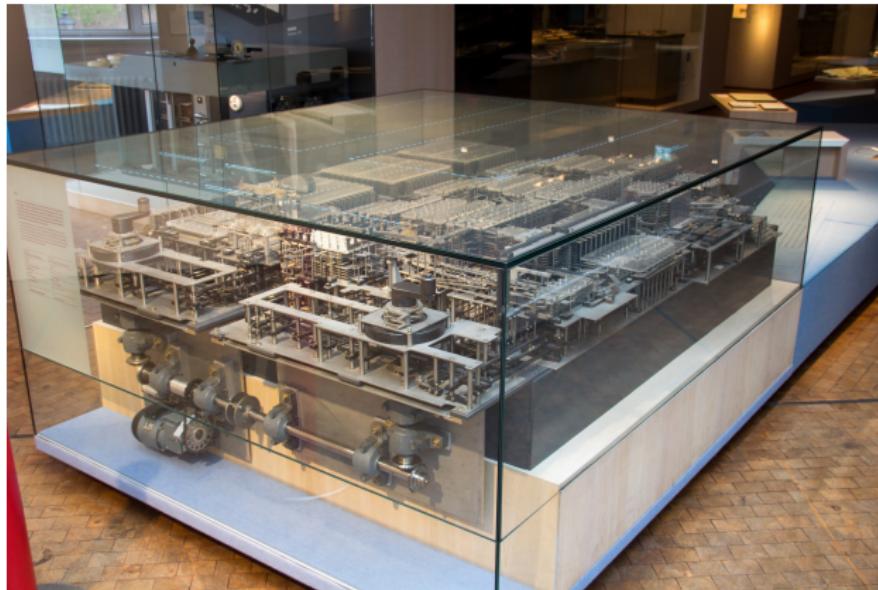
$$y = s \times m \times 2^{e-t}$$

Conjunto de números representables:

$$\mathbf{F} = \{s \times m \times 2^{e-t}\} \cup \{0\}$$

Primera computadora de punto flotante

Konrad Zuse de 1936 a 1937 Z1



By Photograph by Mike Peel (www.mikepeel.net)., CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=58211423>

Aritmética de punto flotante

$t = 2$, $e_{\min} = -2$ y $e_{\max} = 2$, entonces $m = 4, \dots, 7$

| | | mantisa | | | |
|-----------|----|---------|--------|-------|--------|
| | | 4 | 5 | 6 | 7 |
| exponente | -2 | 0.25 | 0.3125 | 0.375 | 0.4375 |
| | -1 | 0.5 | 0.625 | 0.75 | 0.875 |
| | 0 | 1.0 | 1.25 | 1.5 | 1.75 |
| | 1 | 2.0 | 2.5 | 3.0 | 3.5 |
| | 2 | 4.0 | 5.0 | 6.0 | 7.0 |

Tabla: Números de punto flotante con $t = 2$, $e_{\min} = -2$ y $e_{\max} = 2$.



Fig.: Conjunto F_+ .

Norma IEEE 754 de doble precisión

registro: 64 bits

signo: 1 bit

$$s = \pm 1$$

exponente: 11 bits

$$-1022 \leq e \leq 1023$$

mantisa: 52 bits

$$2^{52} \leq m \leq 2^{53} - 1$$

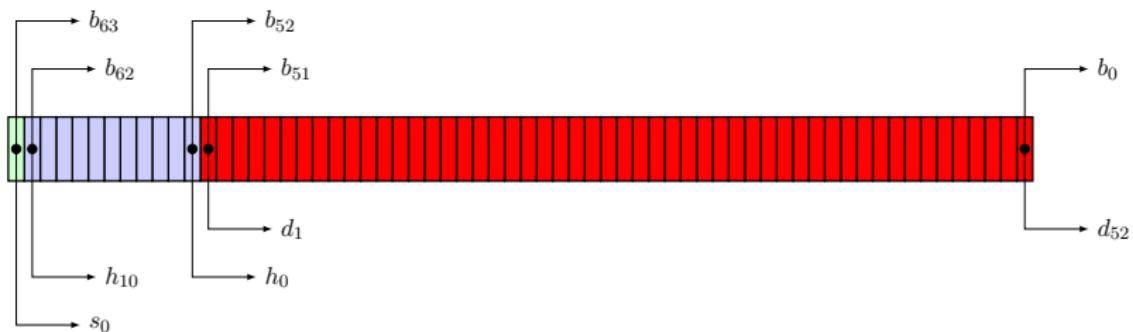


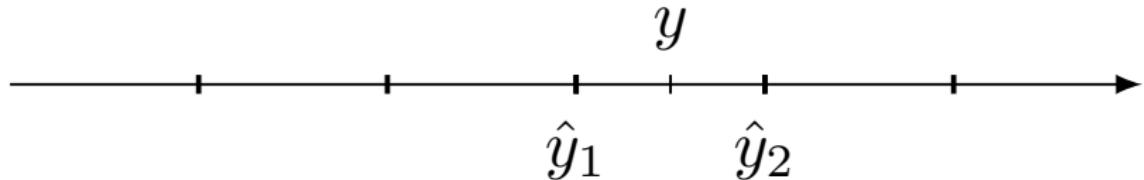
Fig.: Representación de números de doble precisión

Función fl

$fl : \mathbb{R} \rightarrow \mathbf{F}$ se define como

$$fl(y) = \left\{ \hat{y} \in \mathbf{F} : |y - \hat{y}| = \min_{z \in \mathbf{F}} |y - z| \right\}$$

Si $|y - \hat{y}_1| = |y - \hat{y}_2|$, se toma un criterio:



- \hat{y} máximo/mínimo valor absoluto
- \hat{y} mantisa par

Función fl

$$\lambda = \min \mathbf{F}_+$$

$$\Lambda = \max \mathbf{F}_+$$

$|y| < \lambda$: underflow pérdida de precisión

$|y| > \Lambda$: overflow catastrófico

$$\epsilon = \sup\{x > 0 : fl(1 + x) = 1\} = 2^{-t-1}$$



Fig.: Conjunto \mathbf{F}_+ .

Error absoluto

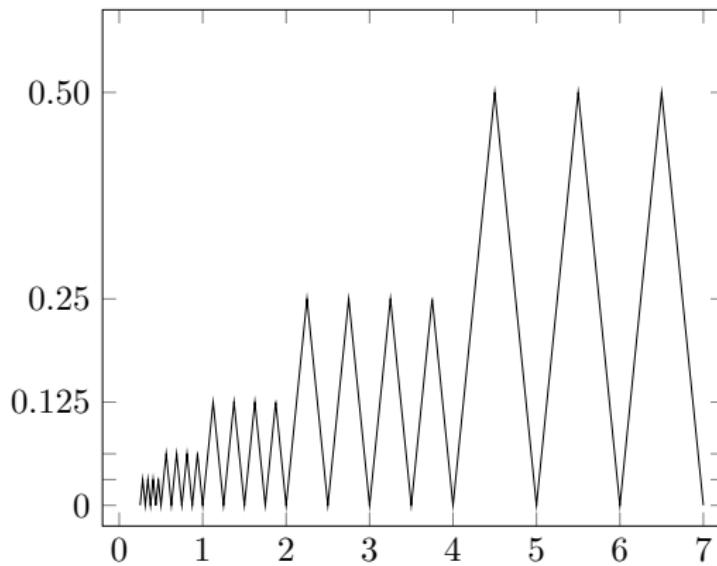


Fig.: Error absoluto: $|x - fl(x)|$

Error relativo

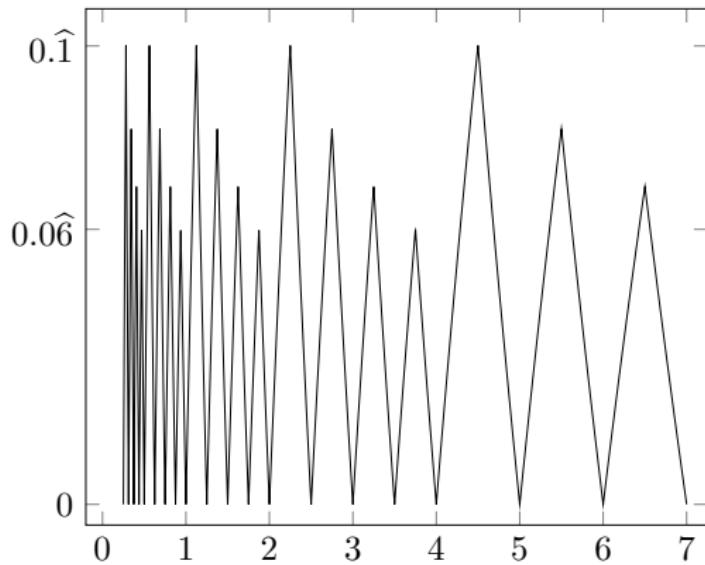


Fig.: Error relativo: $|x - fl(x)|/|x|$

Máximo error relativo

Proposición: Si $x \in \mathbb{R}$ verifica $\lambda < |x| < \Lambda$, entonces existe δ con $|\delta| \leq \epsilon/(1 + \epsilon) (\cong \epsilon)$ tal que $fl(x) - x = \delta x$.

El error relativo de representación está acotado si no hay underflow ni overflow

Operaciones aritméticas: $* = +, -, \times, /$

Operaciones de máquina: $\circledast = \oplus, \ominus, \otimes, \oslash$

Suposición: Si $x, y \in \mathbf{F}$, existe δ con $|\delta| \leq \epsilon$ tal que

$$(x \circledast y) - fl(x * y) = \delta(x * y)$$

$x \circledast y$ es la mejor representación de $x * y$

Normas IEEE 754-1985, IEEE 754-2008

Operaciones en punto flotante

Con varias operaciones el error relativo puede crecer mucho

Ejemplo: sistema decimal con cuatro dígitos

$$a = 2.357 \times 10^1, b = 1.723 \times 10^2, c = -1.368 \times 10^{-1}$$

$$a + b \times c = -6.4 \times 10^{-4}$$

$$b \otimes c = fl(b \times c) = fl(-23.5706) = -2.357 \times 10^1$$

$$a \oplus (b \otimes c) = fl(a + fl(b \times c)) = 0$$

Error relativo:

$$\frac{|(a + b \times c) - (a \oplus (b \otimes c))|}{|a + b \times c|} = 1$$

Operaciones en punto flotante

A veces se implementa la operación ternaria $a \oplus (b \otimes c)$

Si p es el producto interno de $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$p = \mathbf{x} \cdot \mathbf{y} = x_1 \times y_1 + \cdots + x_n \times y_n$$

$$p = x_1 \times y_1$$

$$j = 2$$

for $j \leq n$ **do**

$$p = p + x_j \times y_j$$

$$j = j + 1$$

end for

Operaciones en punto flotante

Si $a, b, c \in \mathbf{F}$ y $a, b, c > 0$

Existe $\delta_1 \in (-\epsilon, \epsilon)$ tal que $b \otimes c = (1 + \delta_1)(b \times c)$

Existe $\delta_2 \in (-\epsilon, \epsilon)$

$$\begin{aligned} a \oplus (b \otimes c) &= fl(a + (1 + \delta_1)(b \times c)) \\ &= (1 + \delta_2)(a + (1 + \delta_1)(b \times c)) \end{aligned}$$

$$(1 - \epsilon)^2(a + b \times c) < a \oplus (b \otimes c) < (1 + \epsilon)^2(a + b \times c)$$

$$(1 - \epsilon)^2 \cong 1 - 2\epsilon, (1 + \epsilon)^2 \cong 1 + 2\epsilon$$

Existe $\delta \in (-2\epsilon, 2\epsilon)$ tal que

$$(a \oplus (b \otimes c)) - (a + b \times c) = \delta(a + b \times c)$$

Errores catastróficos

Misiles Patriot americanos en Dharan

Cohete Ariane 5

Plataforma petrolífera Sleipner A

Millenium Bridge

https://www.academia.edu/3524260/Teoria_de_errores

Complejidad

Complejidad (punto flotante) de un algoritmo

complejidad = número de flops

Ejemplo: producto interno $p = p + x_j y_j$, $j = 1, \dots, n$

n productos, $n - 1$ sumas

Antes los productos eran muchos más costosos que las sumas

Asignación y comparación: costo despreciable

Leer de disco: ∞

Complejidad

La complejidad es del algoritmo, no del problema

Algoritmo 1: $1.2 + 3.2 \times 1.7 + 0.4 \times 1.7^2 + 1.5 \times 1.7^3$

complejidad = 9 (3 sumas y 6 productos)

Algoritmo 2:

$$p = 1.2, s = 1.7$$

$$p = p + 3.2 \times s, s = 1.7 \times s$$

$$p = p + 0.4 \times s, s = 1.7 \times s$$

$$p = p + 1.5 \times s$$

complejidad = 8 (3 sumas y 5 productos)

Algoritmo 3: $1.2 + 1.7 \times (3.2 + 1.7 \times (0.4 + 1.7 \times 1.5))$

complejidad = 6 (3 sumas y 3 productos)

Paralelismo

Máquinas vectoriales: Cray I



De Rama - Trabajo propio, CC BY-SA 2.0 fr,
<https://commons.wikimedia.org/w/index.php?curid=345865>

Paralelismo

Clúster



De NASA Ames Research Center/Tom Trower.