UNIVERSIDAD TECNICA FEDERICO SANTA MARIA
DEPARTAMENTO DE INFORMÁTICA

INF 285 - Computación Científica
Ingeniería Civil Informática

02: Estándar de Punto Flotante y Pérdida de importancia

Números binarios con decimales

$$(B)_2 = ...b_2 b_1 b_0 \, . \, b_{-1} b_{-2} b_{-3} ...$$

$b_i \in \{0, 1\}$

$$((B)_2)_{10} = \sum_{-\infty}^{\infty} b_i \, 2^i$$

Números Binarios

Ejemplo 1

¿Qué número representa $\left(0.\overline{10}\right)_2$ *en base* 10*?*

Estándar de punto flotante

IEEE standard: conjunto de representación binaria.

Número de punto flotante: signo ($+$ o $-$), mantisa y un exponente.

precisión	signo	exponente	mantisa
single	1	8	23
double	1	11	52
long double	1	15	64

Número de punto flotante **normalizado**

$$\pm 1\,\textbf{.}\,bbb...b \times 2^{p}$$

$$(9.5)_{10} \rightarrow (1001\,\textbf{.}\,1)_2 \rightarrow +1\,\textbf{.}\,0011 \times 2^{3}$$

Estándar de punto flotante

$$1 \quad = +1.\underbrace{\boxed{0000000000\cdots0000000000}}_{52\text{ bits}} \times 2^0$$

$$2 \quad = +1.\boxed{0000000000\cdots0000000000} \times 2^1$$

¿Cuál es el siguiente número mayor a 1 representable en "double precisión"?

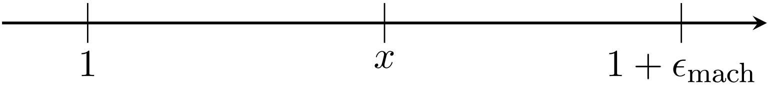
$$1 = +1.\underbrace{\boxed{0000000000\cdots0000000000}}_{52\text{ bits}}\times 2^0$$

$$1 + 2^{-52} = +1.\boxed{0000000000\cdots0000000001}\times 2^0$$

$$\epsilon_{\text{mach}} = 2^{-52} \approx 2 \times 10^{-16}$$

Estándar de punto flotante

¿Qué ocurre cuando queremos representar un número entre 1 y $1 + \epsilon_{\text{mach}}$?



$$9.4 = \left(1001\,\mathbf{.}\,\overline{0110}\right)_2 \qquad = +1\mathbf{.}\ \boxed{0010110\cdots011001100}\ 110\ldots \times 2^3$$

chopping: quitar los bits que sobran.
rounding: redondear hacia arriba si el bit es 1

- Sumar 1 al bit 52 si el bit 53 es 1.
- Mantener tal cual el bit 52 si el bit 53 es 0.
- Excepción: todos los bits después del bit 53 igual a 1 son 0's, se suma solo si el bit 52 es 1.

$$\mathrm{fl}(9.4) \qquad = +1\mathbf{.}\ \boxed{00101100\cdots11001101}\ \times 2^3 \ \ = 9.4 + 0.2 \times 2^{-49}$$

Relative rounding error:

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq \frac{1}{2}\epsilon_{\text{mach}}$$

Ejemplo:

$$\frac{|\mathrm{fl}(9.4) - 9.4|}{|9.4|} = \frac{0.2 \times 2^{-49}}{9.4} = \frac{8}{47} \times 2^{-52} \leq \frac{1}{2}\epsilon_{\text{mach}}$$

s	$e_1\, e_2\ \ldots\ e_{10}\, e_{11}$	$b_1\, b_2\ \ldots\ b_{51}\, b_{52}$
1 bit	11 bits	52 bits

$$\pm 1\,.\,b_1 b_2 ... b_{52} \cdot 2^p = \pm \left(1 + \sum_{i=1}^{52} b_i \cdot 2^{-i}\right) \cdot 2^p$$

$$p = e_1 \cdot 2^{10} + e_2 \cdot 2^9 + \cdots + e_{11} \cdot 2^0 - \underbrace{1023}_{2^{11-1}-1}$$

Exponente $e_1\, e_2 \cdots e_{11} = 11111111111$

- $+\infty$: signo 0 y mantisa 0.
- $-\infty$: signo 1 y mantisa 0.
- NaN $(not - a - number)$: algún bit de la mantisa $\neq 0$.

s	e_1	e_2	e_3	$\cdots$	e_{11}	b_1	b_2	$\cdots$	b_{52}	número	ejemplo
0	1	1	1	$\cdots$	1	0	0	$\cdots$	0	$+\infty$	$1/0$
1	1	1	1	$\cdots$	1	0	0	$\cdots$	0	$-\infty$	$-1/0$
1	1	1	1	$\cdots$	1	x	x	$\cdots$	x	NaN	$0/0$

Exponente $e_1\, e_2 \cdots e_{11} = 00000000000$:

$$\pm 0\,\blacksquare\, b_1 b_2 \ldots b_{52} \cdot 2^{-1022}$$

- El exponente es fijo.
- El número al costado del signo es 0 (no-normalizado).
- Los únicos bits modificables son los de la mantisa.

¿Qué sucede cuando los bits de la mantisa y el exponente son 0?

Pérdida de importancia

Ejemplo 2

Calcular $m_1 + m_2 + m_3$ *mediante los algoritmos:*

- *Algoritmo 1:* $(m_1 + m_2) + m_3$
- *Algoritmo 2:* $(m_1 + m_3) + m_2$

donde $m_1 = 1$, $m_2 = 3 \times 2^{-55}$ *y* $m_3 = -1$.

Pérdida de importancia

Ejemplo 3

$\sqrt{9.01} - 3$ *con un computador de 3 dígitos:*

- *Respuesta correcta:* 1.6662×10^{-3}.
- $\sqrt{9.01} \approx 3.0016662$, *no se obtienen dígitos significativos.*

Pérdida de importancia

Ejemplo 4

Raíces de $x^2 + 9^{12}x = 3$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \rightarrow x = \frac{-9^{12} \pm \sqrt{9^{24} + 4(3)}}{2}$$

Tomando 4 dígitos: $x_1 = -2.824 \times 10^{11}$ *(−)*, $x_2 = 0$ *(+)*.

Ejercicio 1

Evaluar computacionalmente las siguientes funciones a medida que x tienda a 0^+

$$E_1 = \frac{1 - \cos x}{\sin^2 x} \quad E_2 = \frac{1}{1 + \cos x}$$