Números de Máquina

Álgebra Lineal Computacional

Ciencias de Datos-Ciencias de la Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

1er Cuatrimestre 2025

Cálculo simbólico y numérico

Cálculo simbólico

$$x = \sqrt{2}, \qquad x^2 = 2$$

Cálculo numérico

$$x = 1.4142135623730951, \qquad x^2 = 2.0000000000000004$$

Cálculo simbólico y numérico

Ejemplo

- El número 517.23 en base 10 representa al número:

$$5 \cdot 10^2 + 1 \cdot 10^1 + 7 \cdot 10^0 + 2 \cdot 10^{-1} + 3 \cdot 10^{-2}$$

- El número 101.11 en base 2 representa al número:

$$1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$$

Bases de numeración

En general todo número $x \in \mathbb{R}, x \neq 0$, puede representarse en una base $b \in \mathbb{N}, b \geq 2$, de la forma

$$(x)_b = sg(x)\, \tilde{a}_k\, \tilde{a}_{k-1} \ldots \tilde{a}_0 \,.\, \tilde{a}_{-1}\, \tilde{a}_{-2} \ldots$$

donde

- los dígitos verifican $0 \leq \tilde{a}_i \leq b-1$
- $sg(x)$ es el signo de x
- numeramos los índices para que el punto se ubique entre $\tilde{a}_0$ y $\tilde{a}_{-1}$

La notación dada de x representa al número

$$x = \tilde{a}_k b^k + \tilde{a}_{k-1} b^{k-1} + \cdots + \tilde{a}_0 b^0 + \tilde{a}_{-1} b^{-1} + \tilde{a}_{-2} b^{-2} + \cdots$$

Punto flotante

Notación normalizada

- La representación anterior no da idea del orden de magnitud
- Versión *normalizada*: se ubica el punto justo a la izquierda de $\tilde{a}_k$
- Todo $x \neq 0$ se escribe

$$(x)_b = sg(x)\, 0\,.\, a_1\, a_2\, a_3 \ldots \times b^e$$

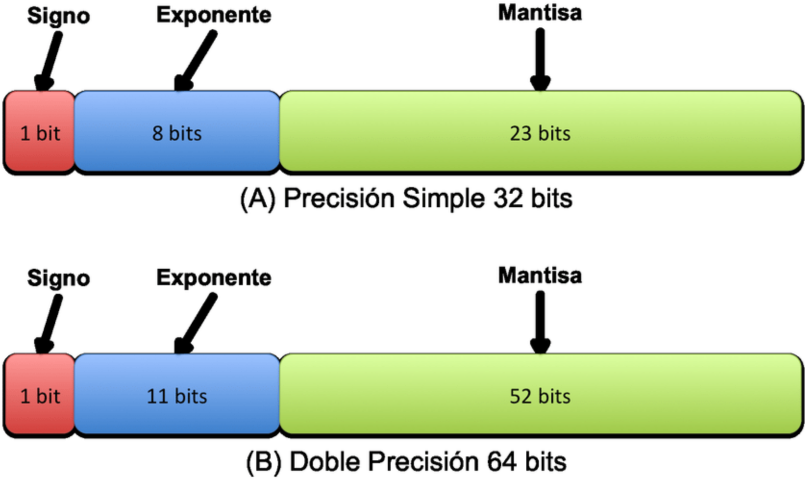
  donde $0 \leq a_i \leq b-1, a_1 \neq 0$ y $e \in \mathbb{Z}$
- La expresión $m = 0\,.\, a_1\, a_2\, a_3 \ldots$ se llama **mantisa** y e al **exponente**.

Punto flotante

Representación en máquina

- Se dedica porción limitada y fija en memoria a cada número.
- La cantidad de bits dedicados depende de la precisión a utilizar.
- Las computadoras representan la información en formato binario (bits)
- Estándar IEEE-754 de 2019 como convención para representar números:
  - Precisión *half*: 16 bits (2 bytes)
  - Precisión *single*: 32 bits (4 bytes)
  - Precisión *double*: 64 bits (8 bytes)
- Surgen *errores de redondeo* al representar números reales.
- Existe otra representación, usada para operaciones bancarias entre otras, llamada *punto fijo*

Estándar IEEE-754



Representación de los números de la forma $x = \pm m \times 2^e$

Estándar IEEE-754

Rangos

En doble precisión, la mayor mantisa que podemos guardar es

$$2^{52} - 1 \approx 0.45 \times 10^{16}$$

y los exponentes que podemos guardar son

$$e \leq 2^{11} - 2 = 2046$$

El exponente se toma con un *offset* de 1023 y entonces el mayor exponente es $2^{1023} \approx 10^{308}$

Doble precisión en representación decimal

Para tener una visión aproximada, podemos pensar que tenemos una mantisa $m \in [0, 1)$ con 16 digitos de precisión, y un exponente e en $[-308, 308]$. Los números que podemos representar tienen la forma

$$0, a_1 \ldots a_{16} \times 10^e$$

Ejemplos en punto flotante

Mantisa y exponente

- $\sqrt{2} \to 1.4142135623730950 = 0.14142135623730950 \times 10$
- $\sqrt{2} \times 1000 \to 1414.213562373095 = 0.1414213562373095 \times 10^4$
- $\sqrt{2}/1000 \to 0.001414213562373095 = 0.1414213562373095 \times 10^{-2}$

Operaciones

- $\sqrt{2} \times 2 \to 2.8284271247461903$
- $1 + \sqrt{2}/1000000 \to 1.0000014142135625$
- $5.0^{100} \to 7.888609052210118 \times 10^{69}$
- $10.0^{400} \to$ (OVERFLOW)
- $10.0^{-400} \to 0$ (UNDERFLOW)

También es útil este link donde pueden ver como se relaciona la representación binaria y el número decimal.

Distribución de los números de máquina

- Los números de máquina no se distribuyen de manera uniforme.
- Sean x_1 y x_2 dos números de máquina consecutivos $(0 < x_1 < x_2)$:

$$(x_1)_b = 0.a_1 a_2 \dots a_m \times b^e$$

y el siguiente número de máquina es

$$(x_2)_b = 0.a_1 a_2 \dots (a_m + 1) \times b^e$$

(salvo que $a_m = b - 1$ en cuyo caso habrá acarreo).
Luego,

$$x_2 - x_1 = b^{-m} b^e$$

- $x_2 - x_1$ crece con el exponente e.
- $x_2 - x_1$ es constante para cada exponente fijo dado.

Distribución de los números de máquina

- Algunos números positivos de máquina con $b = 2$ y $m = 4$
- Los números son equiespaciados únicamente entre potencias sucesivas de 2
- La cantidad de números de máquina en esos rangos se mantiene constante

Representando números

Truncamiento y redondeo

- Si $x \in \mathbb{R}$ no es de un número de máquina:
  - se puede elegir el más cercano para representarlo (*redondeo*)
  - se puede elegir el inmediato inferior (*truncado*)
- Llamamos $fl(x)$ a esta a la función que toma x y le asigna su número de máquina
- Noten que el error relativo cometido al redondear o truncar el número es uniforme a lo largo de todas las escalas.

En k dígitos

$$x = \pm 0, a_1 a_2 \dots a_k a_{k+1} a_{k+2} \dots \times 10^e$$

- Truncamiento: $fl(x) = t(x) = \pm 0, a_1 a_2 \dots a_k \times 10^e$
- Redondeo: $fl(x) = t(x + 5 \times 10^{e-(k+1)})$

Ejemplo de operación

Operaciones en la máquina

Modelamos la operatoria de la siguiente forma:

1. Obtenemos la representación en máquina de los operandos
2. Realizamos la operación en forma exacta
3. Obtenemos la representación en máquina del resultado

Ejemplo suma en máquina de x e y: $fl(fl(x) + fl(y))$.

Ejemplo suma con mantisa de 5 dígitos y truncado

Sean $x = 0.88888888 \times 10^7$ e $y = 0.1 \times 10^2$

$$\begin{aligned} x + y &= fl(fl(x) + fl(y)) \\ &= fl(fl(0.88888888 \times 10^7) + fl(0.1 \times 10^2)) \\ &= fl(0.88888 \times 10^7 + 0.1 \times 10^2) \\ &= 0.88888 \times 10^7 \end{aligned}$$

Quick facts

- ¿Cuál es el número más grande que puedo representar en la computadora?
  $0.99999999999999999 \times 10^{308} \approx 10^{308}$
- ¿Cuál es el número positivo más chico que puedo representar en la computadora?
  $0.000000000000001 \times 10^{-308} = 10^{-323}$
- ¿Cuál es el número siguiente al 1 en la computadora?
  1.000000000000001 (14 ceros)
- ¿Cuál es el número positivo más chico que le puedo sumar a 1 y obtener algo distinto de 1?
  $0.0000000000000005 = \frac{1}{2}10^{-15}$

Epsilon de máquina

Epsilon de máquina (ϵ)

Número más chico que le puedo sumar a 1 y se obtiene algo distinto de 1.

En doble precisión

$$\epsilon \approx 0.0000000000000005 = \frac{1}{2}10^{-15}$$

En general, para base b con mantisa de m dígitos y redondeo

$$\epsilon = \frac{1}{2}b^{1-m}$$

Precisión de la máquina

El epsilon de máquina coincide con el máximo error relativo que puedo obtener al convertir un número a número de máquina

$$\left|\frac{x - fl(x)}{x}\right| \leq \frac{1}{2} b^{1-m}$$

A este número se lo suele llamar precisión de la máquina.

Cancelación Catastrófica

$$\frac{1-\cos(x)}{x^2} \quad \text{vs.} \quad \frac{2\cdot\sin^2(\frac{x}{2})}{x^2}, \qquad -4\times 10^{-8} \le x \le 4\times 10^{-8}$$