ARITMÉTICA del PUNTO FLOTANTE

Clase 13
18/10
ALN 2022

¿Cómo representamos en la computadora un nº real?

La primera limitación es que solo podemos representar un subconjunto finito ~~subconjunto~~ de nº reales. Esto se debe a que podemos utilizar un conjunto finito de "bits" para representar un nº.

Esta limitación tiene como consecuencia que no podremos representar nº arbitrariamente grandes o cercanos a cero, y además existirán obviamente "gaps" (agujeros) entre nº representados.

Una forma eficiente de representar los nº es utilizar el modelo de "notación científica" de un número, i.e, su expansión decimal o binaria tipo 6.02×10^{23}. Veamos esto.

Si queremos multiplicar nº con nº pequeños la notación científica es de mucha ayuda. Por ejemplo.

$$\left.\begin{array}{l} 300.000.000 = 3 \times 10^{8} \\ \times \quad 0{,}0000002 = 2 \times 10^{-7} \end{array}\right] = \quad 6 \times 10^{1} = 60$$

Observar que esta representación de los números es muy eficiente ya que la cantidad de ceros no tiene mucha relevancia. En este sentido lo que "cuenta" es el 3 y el 2 (que son la mantisa) y luego el exponente indica hacia donde mover la coma.
Por ejemplo $1{,}234 \times 10^{e}$, al mover el exponente e lo que estamos moviendo es la coma. De aquí la terminología del coma flotante.
punto

Veamos algunos ejemplos.

El nº 123,45 se puede escribir en base 10 como.

...	10^2	10^1	10^0	10^{-1}	10^{-2}	...
	0	1	2	3	4	5	0	...

Análogamente en base 2 podemos escribir

$\frac{1}{2} = 0,5$ como

	2^3	2^2	2^1	2^0	2^{-1}	2^{-2}	
...	0	0	0	0	1	0	...

0,1 como

...	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}					
	0	0	0	0	1	1	0	0	1	1	0 0 1 1 ...

se repite.

Básicamente la representación de un número en la computadora se describe de la siguiente manera:

	$\frac{1}{2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	....									2^2	2^1	2^0	
	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	1

16-bits representación

mantisa (10) exponente (6)

signo

signo { 0 = +, 1 = −

$$\left(1\cdot\frac{1}{2} + 1\cdot\frac{1}{2^2} + \frac{1}{2^4}\right) \times 2^{[2^2+2^1+2^0]=7} = 2^6 + 2^5 + 2^3 = \ldots$$

Formalicemos lo anterior a cualquier base y mantisa.

Sea $\beta > 1$, $t \geq 1$, $e_{min} \in -\mathbb{N}$, $e_{max} \in \mathbb{N}$. Los nº flotantes son

$$y = \pm\, m \times \beta^{e-t}$$

donde β es la base (2, 10, 16)

t es la precisión, $e_{min} \leq e \leq e_{max}$ el exponente

m: mantisa con $0 \leq m \leq \beta^t - 1$

La mantisa $m = d_1 \beta^{t-1} + d_2 \beta^{t-2} + \cdots + d_{t-1}\beta + d_t$

con $0 \leq d_i \leq \beta - 1$, i.e.

$$y = \pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t} \right) \times \beta^e$$

o a veces se escribe $y = \pm \, . \, d_1 d_2 \cdots d_t \times \beta^e$.

(El punto flotante está normalizado si $d_1 \neq 0$, y se define por $\mathbb{F}$ a este conjunto)

Función Redondeo

Se define $fl : \mathbb{R} \to \mathbb{F}$ que asocia a un nº real x el punto flotante $fl(x) \in \mathbb{F}$ más cercano. (A menos de pts equidistantes, está bien definida. En esos puntos podemos decir que toma el valor de la derecha por ej.)

- Decimos que x es un "overflow" si $|x| > \max_{y \in \mathbb{F}} |y|$, y "underflow" si $0 < |x| < \min_{y \in \mathbb{F}} |y|$.

- Proposición: $\forall x \in [\min_{y \in \mathbb{F}} |y| \, , \max_{y \in \mathbb{F}} |y|] \;\; \exists \, \delta \in [0, \varepsilon_{machine}]$ tal que $fl(x) = x(1+\delta)$, es decir

$$\left| \frac{fl(x) - x}{x} \right| \leq \varepsilon_{machine}$$

siendo $\varepsilon_{machine} := \frac{\beta^{1-t}}{2}$.

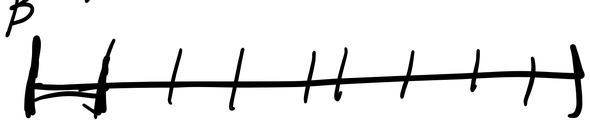
Dem: Supongamos que $x \in [\beta^e, \beta^{e+1}]$. $(e \leq e_{max} - 1)$

Si $x = \beta^e$, entonces $y = \left(\frac{1}{\beta}\right) \times \beta^{e+1} = \beta^e = x$ y para todo x está representado exactamente.

Si $x > \beta^e$, entonces los ptos flotantes (en F) que son mayores a β^e (y en el intervalo $[\beta^e, \beta^{e+1}]$) se pueden escribir como

$$y = \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}\right) \times \beta^{e+1}, \quad \text{con } d_1 \neq 0 \text{ y } 0 \leq d_i \leq \beta - 1.$$

Luego, la distancia entre dos nº consecutivos es igual a $\frac{\beta^{e+1}}{\beta^t} = \beta^e \cdot \beta^{1-t} = \beta^{e-2 \cdot e_{mach}}$



y por lo tanto $|fl(x) - x| \leq \frac{\beta^e \varepsilon_{mach}}{2} < |x| \cdot \varepsilon_m \leq \beta^{e\left(\frac{1}{\beta} + \frac{1}{\beta^t}\right)\beta^{e+1}}$

$$\Rightarrow \left|\frac{fl(x) - x}{x}\right| \leq \varepsilon_m. \qquad \square$$

Comentarios: 1) La norma IEEE utiliza base $\beta = 2$, $t = 53$ y por lo tanto el $\varepsilon_{machine} = 2^{-53} \simeq 1.11 \times 10^{-16}$.

2) A veces el ε machine se define como $\min_{y \in F,\, y > 1} |1 - y| = \beta^{1-t}$.

Para eso es fácil ver que y es de la forma $y = \left(\frac{1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t}\right) \times \beta$ y por lo tanto la solución es $y^* = \left(\frac{1}{\beta} + \frac{1}{\beta^t}\right) \cdot \beta = 1 + \frac{1}{\beta^{t-1}} \Rightarrow \min_{y \in F,\, y > 1} |1 - y| = \frac{1}{\beta^{t-1}}$.

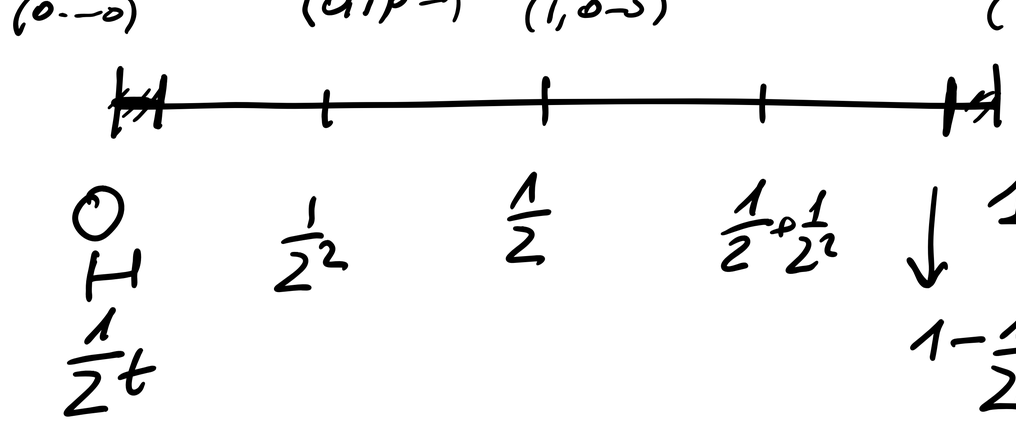
Para fijar ideas trabajamos con $\beta = 2$, $t = 53$ como en IEEE. (64-bits) como $t = 53$ (tamaño mantisa) quedan 11 para exponente, i.e $2^{11} = 2048$ elecciones para el exponente, que va de -1024 a 1024 (de hecho de -1022 a 1023 donde el 0 y el 1 tienen dígitos prefijados.)

En decimales $2^{1024} \simeq 2 \times 10^{308}$ por lo que la coma puede moverse 308 espacios hacia la derecha e izquierda. (Osea overflow y underflow no parecen ser un problema.)

Recordar de la expresión binaria tenemos

$$\frac{d_1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_t}{2^t}$$

nos da una elección de cada punto de la partición de $[0,1]$ en intervalitos de tamaño $\frac{1}{2^t}$.

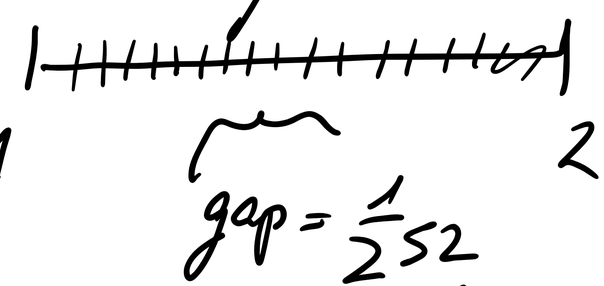


$$1 - \frac{1}{2^t} = \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^t} = (111\cdots1)_2$$

Luego en el intervalo $[1,2]$ tenemos que $\mathbb{F} \cap [1,2]$ son de la forma

$$\left(\frac{1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_t}{2^t}\right) \times 2 = 1 + \frac{d_2}{2} + \cdots + \frac{d_t}{2^{t-1}}.$$

y por lo tanto tenemos una partición equiespaciada de 2^{52} puntos en ese intervalo



Sin embargo el intervalo $[2,4]$ tiene la misma cantidad de puntos equiespaciados: $\left(\frac{1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_t}{2^t}\right) \times 2^2$

O sea el espacio (gap) ahora es $\frac{1}{2^t} \cdot 2^2 = \frac{1}{2^{t-2}} = \frac{1}{2^{51}}$.

Procediendo de la misma manera tenemos que en el intervalo $[2^j, 2^{j+1}]$ los puntos en F están espaciados $\frac{1}{2^{t-j-1}}$.

en $[1,2]$ los ptos son $1, 1+\frac{1}{2^{t-1}}, 1+2\cdot\frac{1}{2^{t-1}}, 1+3\cdot\frac{1}{2^{t-1}}, \ldots 1+2^{t-1}\cdot\frac{1}{2^{t-1}}$

en $[2,4]$, $2+\frac{1}{2^{51}}; 2+2\cdot\frac{1}{2^{51}}, \ldots$

Aunque la diferencia puede ser cada vez mayor, los errores relativos se mantienen como muestra la proposición probada.

Aritmética del Punto Flotante

Hasta ahora hemos representado digitalmente los números reales pero no hemos dicho nada sobre como hacer cálculos con ellos.

Todos los cálculos matemáticos se reducen a operaciones aritméticas elementales como $+, \times, \div, -$. Estas operaciones matemáticas en $\mathbb{R}$ tienen en análogo en F. Le agregamos un círculo para diferenciarlos $\oplus, \ominus, \otimes, \oslash$.

Los computadores se diseñan de manera tal que el siguiente principio es válido. Si x e $y \in F$, y $*$ es alguno de los op. elementales anteriores, y $\circledast$ es su análogo en punto flotante entonces $x \circledast y$ es exactamente $x \circledast y = fl(x * y)$.

Luego con este ppio se concluye el siguiente axioma:

Axioma Fundamental de la Aritmética del Punto Flotante

$\forall x, y \in F$, $\exists \varepsilon$ / $|\varepsilon| < \varepsilon_{machine}$ tq $x \circledast y = (x * y)(1+\varepsilon)$.

i.e toda operación en F es exacta a menos de un error relativo de tamaño $\varepsilon_{machine}$.