CAPíTULO 1

Aritmética de Punto Flotante

1. Números de Punto Flotante

Un conjunto F de números de *punto flotante* está caracterizado por los siguientes parámetros:

(1) La base del sistema β.
(2) El número de dígitos n en la mantisa.
(3) Un exponente $m \leq e \leq M$.

donde β, n, m, e son enteros.

Cada número en el sistema de punto flotante F tiene la forma

$$\pm(.d_1 d_2 \cdots d_n)_\beta \beta^e,$$

donde $d_i = 0, 1, \ldots, \beta - 1$, $i = 1, 2, \ldots, n$, $(.d_1 d_2 \cdots d_n)_\beta$ es una β–fracción llamada *mantisa*, e es un entero llamado el *exponente*.

El sistema F de punto flotante está *normalizado* si $d_1 \neq 0$. en general, todos los números flotantes se normalizan con la excepción del cero, en el cual $d_1 = d_2 = \cdots = d_n = 0$.

El conjunto F es discreto y finito. Su cardinalidad, el número de elementos que lo constituyen, está dada por

$$2(\beta - 1)\beta^{n-1}(M - m + 1) + 1.$$

Como una consecuencia de la finitud de F para representar al conjunto de números reales $\mathbb{R}$, existirá una infinidad de números en $\mathbb{R}$ que no pueden representarse en forma exacta en F.

Sea x un número real denotemos por $\text{fl}(x)$ el número en F que es más *cercano* a x. La diferencia entre x y $\text{fl}(x)$ se llama *error de rendondeo*, éste depende de la magnitud de x y es por lo tanto medido relativo a x

$$\delta(x) = \frac{\text{fl}(x) - x}{x}, \quad x \neq 0, \tag{1.1}$$

luego $|\delta(x)|$ es el error relativo introducido en la representación de x en el sistema de punto flotante F. De (1.1) obtenemos

$$\text{fl}(x) = x(1 + \delta(x)).$$

Hay dos formas generalmente usadas para convertir un número real x a un n–β número flotante $\text{fl}(x)$: *redondeado* o *truncado*. Cuando se redondea, $\text{fl}(x)$ se elige como el número de punto flotante normalizado más cercano a x, si hay empate se usa alguna regla especial, por ejemplo, se toma el de la derecha. Si se trunca, $\text{fl}(x)$ se escoge como el número flotante normalizado más cercano entre 0 y x, esto es, se toman algunas d's y se desprecian otras.

A continuación encontraremos una cota para $\delta(x)$ que sea independiente de x. En el sistema de números de punto flotante F, el número cuyo valor absoluto es el más pequeño está dado por

$$+(.100\cdots 0)_\beta \beta^m = \beta^{m-1},$$

el sucesor inmediato de β^{m-1} se encuentra sumándole a éste el número

$$+(.000\cdots 1)_\beta \beta^m = \beta^{m-n}.$$

Se concluye entonces, que la distancia entre dos números consecutivos en el intervalo $[\beta^{m-1}, \beta^m]$ es β^{m-n}.

En forma similar se demuestra que la distancia entre dos números consecutivos que pertenezcan a cualquier intervalo de la forma $[\beta^j, \beta^{j+1}]$ $j = m, \ldots M-1$ está dada por

$$\beta^{j+1-n}. \tag{1.2}$$

Hemos demostrado un aspecto singular de la distribución de los números del sistema de punto flotante F:

- *estos no están igualmente espaciados a través de todo su rango, sino únicamente cuando se encuentran entre potencias sucesivas de la base β.*

Sunpongamos que $x \in [\beta^j, \beta^{j+1}]$ para alguna $j = m, \ldots M-1$, si el número flotante $\mathrm{fl}(x)$ que representa a x es seleccionado por redondeo, entonces de acuerdo a (1.2) el error introducido es a lo más $(1/2)\beta^{j+1-n}$, si $\mathrm{fl}(x)$ se selecciona por truncamiento el error es a lo más β^{j+1-n}. Lo anterior nos da la *medida del error de redondeo absoluto*

$$|\mathrm{fl}(x) - x| \leq \begin{cases} \frac{1}{2}\beta^{j+1-n} & \text{redondeo,} \\ \beta^{j+1-n} & \text{truncamiento.} \end{cases} \tag{1.3}$$

El *error de redondeo relativo* $|\delta(x)|$ se obtiene dividiendo al error de redondeo absoluto por $|x|$. Ya que $0 < \beta^j \leq |x|$ obtenemos

$$|\delta(x)| \leq \begin{cases} \frac{1}{2}\beta^{1-n} & \text{redondeo,} \\ \beta^{1-n} & \text{truncamiento.} \end{cases}$$

El *épsilon de la máquina* se define como

$$\varepsilon = \begin{cases} \frac{1}{2}\beta^{1-n} & \text{redondeo,} \\ \beta^{1-n} & \text{truncamiento.} \end{cases}$$

Consecuente con lo anterior se cumple que $|\delta(x)| \leq \varepsilon$ para toda x. La cota para $\delta(x)$ independiente de la x es el épsilon de la máquina.

Dado $x \in \mathbb{R}$ su flotante $\mathrm{fl}(x)$ se define como

$$\mathrm{fl}(x) = x(1+\delta), \quad |\delta| \leq \epsilon.$$

La exactitud de la aritmética de punto flotante está entonces caracterizada por el épsilon de la máquina.

Supongamos que estamos en el intervalo $[1, \beta]$. Queremos calcular $1 \oplus \varepsilon$, donde $\oplus$ denota la suma entre números que pertenecen a F

$$1 \oplus \varepsilon = \mathrm{fl}(1+\varepsilon).$$

Si el error introducido en la representación de $1+\varepsilon$ es por redondeo, entonces éste queda localizado en la mitad del intervalo $[1, 1+\beta^{1-n}]$. Ya que la distancia entre dos números consecutivos de F que pertenecen al intervalo $[1,\beta]$ es β^{1-n} el número de punto flotante $1 \oplus \varepsilon$ es tomado como $1+\beta^{1-n} > 1$. Sin embargo, si $0 < \varepsilon_1 < \varepsilon$ obtenemos por un procedimiento análogo al anterior que $1 \oplus \varepsilon_1 = 1$. Un resultado similar es encontrado cuando el error introducido en la representacion de $1 \oplus \varepsilon$ es por truncamiento.

El épsilon de la máquina se caracteriza por ser el menor número positivo de F que satisface

$$1 \oplus \varepsilon > 1.$$

---

2. Modelo de Aritmética en F

La aritmética en el sistema numérico de punto flotante F permite aproximar a la del sistema de números reales $\mathbb{R}$. Como notación emplearemos $\oplus, \ominus, \otimes, \oslash$ para indicar las aproximaciones a las operaciones aritméticas $+, -, \times, /$ de $\mathbb{R}$:

$$\begin{aligned} x \oplus y &= \mathrm{fl}(x+y), \\ x \ominus y &= \mathrm{fl}(x-y), \\ x \otimes y &= \mathrm{fl}(x \times y), \\ x \oslash y &= \mathrm{fl}(x/y), \end{aligned}$$

El modelo que asumiremos para la artimética en F es el siguiente

$$\mathrm{fl}(x \,\mathrm{op}\, y) = (x \,\mathrm{op}\, y)(1+\delta), \quad |\delta| \leq \varepsilon, \quad \mathrm{op} = +, -, *, /.$$

Para efectuar operaciones en forma manual en este modelo aritmético, por cada operación $+, -, \times, /$ encontrada hágala en aritmética exacta, normalice el resultado, trunque o redondeé de acuerdo al número de dígitos permitidos.