Números de máquina

Castillo Flores Junior 1, Cordero Gavilán Anthony 2 Aguirre Janampa Cristian 3, Nalvarte Yantas Kevin 4, Yoshimar 5,Alvitres Palomino Jean 6, García Sifuentes Tomás 7

Facultad de Ciencias 1*, Universidad Nacional de Ingeniería* 1*, e-mail:juniorcastillon6@gmail.com**
Facultad de Ciencias 2*, Universidad Nacional de Ingeniería* 2*, e-mail:anthony.cordero.g@uni.pe**

Palabras Claves: *mantiza, épsilon de máquina, truncamiento, aproximación, expansión de Taylor.*

Keywords: *mantissa, machine epsilon, clipping, aproach, expansion of Taylor*

1. INTRODUCCIÓN

En las últimas décadas el avance de la tecnología ha crecido considerablemente, tanto en ámbitos laborales como científicos, este último ayudado grandemente por el calculo por computador.
Aunque dispongamos de los calculos hechos por estas maquinas, ¿se debe confiar ciegamente en los cálculos realizados por esta?, ¿que peligro hay para la precisión de los cálculos científicos realizados por la máquina?. En casi todas las ramas de la ciencia, se utiliza alguna expresión numérica que involucra cálculos. Aun la más pequeña o rebuscada ecuación científica tiene una solución numérica, obtenida por operaciones aritméticas fundametales enseñadas desde muy temprana edad en los centros de estudios. En el presente informe, expondremos una simulación de operaciones de una máquina, mostrando la metodología que realiza la máquina para dar una respuesta a operaciones básicas como la suma, resta, multiplicación y la división entre 2 números reales.¿Cómo opera la máquina para realizar estas operaciones? Aquí trataremos de mostrar la forma de como lo hace. Además de cálculos básicos, la máquina puede expresar funciones matemáticas, esto debido a que dichas funciones se puede aproximar a expresiones conocidas por la máquina como la serie de Taylor.

2. CONCEPTOS PREVIOS

Números en la computadora: La aparición de los computadores ha hecho posible la resolución de problemas, que por su tamaño antes eran excluidos. Desafortunadamente los resultados son afectados por el uso de la aritmética de precisión finita, en la cual para cada número se puede almacenar tantos dígitos como lo permita el diseño del computador. Así de nuestra experiencia esperamos obtener siempre expresiones verdaderas como 2 + 2 = 4, 3 x 3 = 9, sin embargo, en la aritmética de precisión finita $\sqrt{5}$ no tiene un solo número fijo y finito, que lo representa. Como $\sqrt{5}$ no tiene una representación de dígitos finitos, en el interior del computador se le da un valor aproximado cuyo cuadrado no es exactamente 5, aunque con toda probabilidad estará lo bastante cerca a él para que sea aceptable.

Números en punto flotante: Los números en punto flotante son números reales de la forma:

$$\alpha.\beta^e$$

Donde α tiene un número de dígitos limitados, β es la base y e es el exponente que hace cambiar la posición al punto decimal. Un número real x tiene

la representacion punto flotante normalizada si:

$$x = \pm\alpha.\beta^e, \frac{1}{\beta} < \|a\| < 1 \quad (1)$$

En caso que x tenga representación en punto flotante normalizada entonces x = 0, d_1 $d_2...d_k$ donde:

$$d_1 \neq 0, 0 \leq d_i < \beta, i = 1, 2, 3, ... \quad (2)$$

y L ≤ e ≤ U. El conjunto de los números en punto flotante se le llama, conjunto de números de máquina. El conjunto de número de máquina es finito ya que si

$$x = \pm 0, d_1 d_2 d_3 d_4 ... d_t.b^e, \quad (3)$$

Conn d_1 hay β-1 posibles valores y para d_i, i = 2, 3, 4..., t hay β posibles asignaciones, luego existirán $(\beta-1).\beta.\beta.\beta = (\beta-1)\beta^{t-1}$, fracciones positivas. Pero como el número de exponentes es U−L+1 en total habrán $(\beta-1)\beta^{t-1}$.(U−L+1) números de máquina positivos y tomando los números de máquina negativos, el total de números de máquina es $2.(\beta-1)\beta^{t-1}$.(U−L+1). Si incluimos al cero en nuestros números, significa que cualquier número real debe ser representado por uno de los $2.(\beta\text{-}1)\beta^{t-1}$.(U−L+1)+1 números de máquina.

Épsilon de máquina: En una aritmética de punto flotante, se llama épsilon de la máquina (ϵ) al menor valor de una determinada mquina que cumple lo siguiente:

$$1.0+\epsilon > 1.0$$

El épsilon es el número decimal ms pequeño que, sumado a 1, la computadora nos arroja un valor diferente de 1, es decir, que no es redondeado. Representa la exactitud relativa de la aritmtica del computador. La existencia del épsilon de la mquina es una consecuencia de la precisión finita de la aritmética en punto flotante.

Aproximaciones : Como hemos dicho, los números pueden sufrir aproximaciones cuando se dan como datos de entrada a la máquina o como resultado de operaciones. Estas aproximaciones pueden ser de 2 maneras :

- **Truncamiento :** En este proceso el número se representa por medio del mayor número de la máquina menor que el número dado.
- **Redondeo :** En este proceso el número se representa por el número de máquina más cercano al número dado.
- **Observación :** Hay cálculos que pueden perjudicar la presición computacional, como por ejemplo se hace la resta de 2 números casi iguales, ya que se cancelan los dígitos principales.

Expansión de Taylor: Sea f la función cuyas derivadas existen en un intervalo I, y estas no tienen un tamaño desmesurado,es decir, están acotadas:

$$|f(\mathrm{x})| \leq \mathrm{k},\ \mathrm{x} \in I$$

Entonces, se verifíca que:

$$f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f'''(a)}{3!}(x-a)^3+... \quad (4)$$

$$\text{Donde } a \in I$$

Es decir una función infinitamente derivable en un intervalo puede representarse como un polinomio a partir de sus derivadas evaluadas en el punto de dicho intervalo. De manera ms compacta:

$$\sum_{i=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n \quad (5)$$

Dado que esta serie se expande al infinito, nosotros aproximaremos a un orden de $O(x^4)$ para poder hacer cálculos aproximados para ciertas funciones como el $sen(x)$ o el $cos(x)$.

Aritmética de punto flotante:

A continuación vamos a recordar las cuatro operaciones básicas que se realizan sobre este tipo de representaciones.

Suma.

Cuando sumamos o restamos dos números en coma flotante se deben comparar los exponentes y hacerlos iguales, para lo cual hay que desplazar o alinear uno de ellos respecto al otro. Por ejemplo, consideremos 10,375 + 6,34375 = 16,71875 o en binario :

$$\begin{array}{rr} & 1,0100110 * 2^3 \\ + & 1,1001011 * 2^2 \\ \hline \end{array}$$

Estos 2 números no tienen el mismo exponente, así que se desplaza la mantisa para hacer iguales los exponentes y entonces sumar :

$$\begin{array}{rr} & 1,0100110 * 2^3 \\ + & 1,1001011 * 2^3 \\ \hline & 10,0001100 * 2^3 \end{array}$$

Observe que el desplazamiento de $1,1001011 * 2^2$ pierde el uno delantero y luego de redondear el resultado se convierte en $0,1100110 * 2^3$. El resultado de la suma, $10,0001100 * 2^3$ (o $1,00001100 * 2^4$) es igual a 10000,1102 o 16.75. Esto no es igual a la respuesta exacta (16,71875). Es solo una aproximación debido al error del redondeo del proceso de la suma.

Es importante tener en cuenta que la aritmética de punto flotante en un computador (o calculadora) es siempre una aproximación. Las leyes de las matemáticas no siempre funcionan con números de punto flotante en un computador. Las matemáticas asumen una precisión infinita que un computador no puede alcanzar. Por ejemplo, las matemáticas enseñan que $(a+b)b = a$; sin embargo, esto puede ser exactamente cierto en un computador.

Resta

La resta trabaja muy similar y tiene los mismos problemas que la suma.

Considere un ejemplo : $16,75 - 15,9375 = 0,8125$

$$\begin{array}{rr} & 1,0000110 * 2^4 \\ - & 1,1111111 * 2^3 \\ \hline \end{array}$$

Desplazando $1,1111111 * 2^3$ da (redondeado arriba) $1,0000000 * 2^4$

$$\begin{array}{rr} & 1,0000110 * 2^4 \\ - & 1,0000000 * 2^4 \\ \hline & 0,0000110 * 2^4 \end{array}$$

Multiplicación y división

Para la multiplicación las mantisas son multiplicadas y los exponentes son sumandos. Considere $10,375 * 2,5 = 25,9375$:

$$\begin{array}{rr} & 1,0100110 * 2^3 \\ * & 1,0100000 * 2^1 \\ \hline & 10100110 \\ + & 10100110 \\ \hline & 1,10011111000000 * 2^4 \end{array}$$

Claro esta, el resultado real podría ser redondeado a $8bits$ para dar :

$1,1010000 * 2^4 = 11010,0002 = 26$

La división es más complicada, pero tiene problemas similares con errores de redondeo.

3. ANÁLISIS

4. OBSERVACIONES

5. CONCLUSIONES

---

1. I.K. Argyros, *Newton-like methods under mild differentiability conditions with error analysis,* Bull. Austral. Math. Soc. **37** (1988), 131-147.
2.