APROXIMACION NUMERICA Y ERRORES

Comenzando por una definición etimológica tenemos

- *Error:* Equivocación (antónimo: certidumbre), incertidumbre
- *Incertidumbre* Falta de certidumbre (antónimo: certidumbre)

Las computadoras utilizan el sistema binario, es decir, solo usan los dígitos 0 y 1. Aunque la "comunicación" entre la computadora y el usuario es por medio del sistema decimal, las entradas o salidas de datos requieren una conversión al sistema binario internamente.

En modo general, sea un numero entero $\alpha > 1$, puede utilizarse para representar un sistema numérico, por tanto

$$(N)_\alpha = (n)_{10} = (N)_2$$

representa a un número en la base α . La existencia de procesos de conversión implica errores de redondeo que deben tenerse en cuenta. Cualquier número distinto de cero, puede ser representado en notación científica, por ejemplo, considérese al número **x**, si

- n es un número entero (positivo, negativo o cero)
- α es la base
- q es un número, tal que $\frac{1}{\alpha} \leq q < 1$

entonces podemos escribir al número **x** como $x = \pm q * \alpha^{\pm n}$

donde q se llama mantisa y n es el exponente. Siendo esta la representación de punto flotante normalizada para el número X. Esto significa que el primer dígito se encuentra a la derecha del punto y se suministran las potencias de la base α adecuadas

ARITMÉTICA DE PUNTO FLOTANTE

La unidad más pequeña de información que la computadora reconoce se denomina bit, es representado en la computadora por la presencia o abstinencia de un pulso electrónico, simbolizado por el hombre como cero o uno.

La longitud de palabra de la computadora es una restricción en la precisión con que se representan los números reales. Por ejemplo, si tenemos que la longitud de palabra es de 32 bits, cuya distribución es

- **s**= signo del número real X 1 bit
- s= signo del exponente n 1 bit
- E= exponente (entero | n|) 7 bits
- F= mantisa (número real | q|) 23 bits

en forma grafica se tiene

s	s	E	E	E	E	E	E	E	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

LONGITUD DE PALABRA 32 bits

Recordando que para el sistema binario se tiene α =2, siendo esta la especificación el exponente cumple con,

$$|n| \leq (1111111)_2 = 127$$

es decir que nuestra máquina puede calcular números reales en el intervalo

$$2^{-127} \approx 10^{-38} \text{ y } 2^{127} \approx 10^{38}$$

mientras que para los números enteros, se pueden usar todos los bits a excepción de uno, el reservado para el signo, es decir, en el intervalo

$$-(2^{31}-1) \text{ y } (2^{31}-1) = 2147483646$$

aún así, existen cálculos que requieren más precisión. Esto se logra asignando dos longitudes de palabras al número, esto se conoce como, calculo de doble precisión. Cuando se restringe a 24 bits para representar a la mantisa, implica que el bit menos significativo es 2^{-24}.

ERROR DE TRUNCAMIENTO Y ERROR DE REDONDEO.

Cuando se aproxima un número X, es necesario encontrar el número de máquina más adecuado. Por ejemplo, sí

$$X = (a_1 a_2 a_3 \quad a_{24} a_{25} a_{26} \cdots)_2 \times 2^n$$

Y nuestra computadora solo tiene capacidad para una mantisa de 24 bits. Se tienen dos opciones,

i. Se descartan los bits excedentes, esto es, $a_{25}a_{26}\cdots$, a este proceso se le llama truncamiento, además el valor del número máquina X^t se encuentra a la izquierda de X.

$$X^t = (a_1 a_2 \ldots a_{24})_2 \times 2^n$$

ii. Se descartan los bits excedentes, esto es, $a_{25}a_{26}\cdots$, y se aumenta una unidad al último bit. Siendo este el proceso de redondeo por exceso de X.

$$X = ((a_1 a_2 \ldots a_{24})_2 + 2^{-24}) \times 2^n$$

y se encuentra a la derecha del número X.

Se escoge aquella en la cual la distancia al número X sea menor. Generalmente *fl*(*x*) representa al número de máquina de punto flotante más cerca de X.

ERROR ABSOLUTO Y ERROR RELATIVO

Considerando la distancia entre el número de máquina adecuado y el número X, ya sea X'ó X''denominamos a dicha distancia error absoluto. Para el caso de haber elegido X'se escribe

$$\left|X - X'\right| \leq \frac{1}{2}\left|X' - X''\right| = 2^{n-25}$$

en caso de ser X''el número escogido se escribe

$$\left|X - X''\right| \leq \frac{1}{2}\left|X' - X''\right| = 2^{n-25}$$

y error relativo se define como el error absoluto entre el número X, esto es

i. Error relativo para el truncamiento

$$\left|\frac{X - X'}{X}\right| \leq \frac{2^{n-25}}{q \times 2^{n}} = \frac{2^{-25}}{q} \leq \frac{2^{-25}}{1/2} = 2^{-24}$$

cuando el error absoluto de truncamiento es menor que el error absoluto de redondeo.

ii. Error relativo para el redondeo por exceso

$$\left|\frac{X - X''}{X}\right| \leq \frac{2^{n-25}}{q \times 2^{n}} \leq 2^{-24}$$

cuando el error absoluto de redondeo es menor que el error absoluto del truncamiento.

Al número 2^{-24} se le conoce como error de redondeo unitario. Cuando en nuestros cálculos se generan números demasiado grandes (n excede el valor permitido) se tiene lo que se denomina error de desbordamiento por exceso. En el caso contrario, es decir, si se han generado números demasiado pequeños (n es menos del mínimo) se conoce como desbordamiento por defecto, en nuestro caso n>127 y n<-127 respectivamente. En el primer caso la computadora detiene el programa debido a la indeterminación generada, en el segundo caso se asigna el valor 0 y continua el proceso numérico.

Sea $\hat{X}$el número de máquina más cerca de X,

$$\left|\frac{X - \hat{X}}{X}\right| \leq 2^{-24}$$

Sí

$$\delta = \frac{X - \hat{X}}{X} \Rightarrow \hat{X} = X(1+\delta);\ |\delta| \le 2^{-24}$$

ó

$$\hat{X} = fl(x) = x(1+\delta)\ |\delta| \le E$$

con $E = 2^{-24}$ error de redondeo unitario para nuestra máquina.

En general, si una máquina funciona en base α y utiliza N posiciones para la mantisa de sus números de punto flotante, escribimos

$$fl(x) = x(1+\varepsilon)\ |\varepsilon| \le E$$

donde $E = \frac{1}{2}\alpha^{1-N}$ para el proceso de redondeo por exceso y $E = \beta^{1-N}$ en el caso de truncamiento.

Ejemplo:

Sea $X = (0.111..\ 11100..)_2 \times 2^{17}$, en donde la parte fraccionaria tiene 26 unos seguidos de ceros. Para una máquina de 32 bits para la longitud de palabra determine

$$X^t, X^r, X - X^t, X^r - X^t, X^r - X, |X - fl(x)| / X$$

Solución:

Dado que

$$X = (0.11\underbrace{..1}_{26}00)_2 \times 2^{17}$$

el valor truncado es simplemente

$$X^t = (0.1\overbrace{..1}^{24})_2 \times 2^{17}$$

en cuanto al valor de redondeo tenemos

$$X^r = ((0\ 11..1)_2 + 2^{-24}) \times 2^{17}$$

Ahora procedemos a determinar cual de los dos será el número máquina, para esto se requiere de calcular las siguientes diferencias

a. $$X - X^{+} = (0.110...)_2 \times 2^{-24} \times 2^{17} = (0.1100...)_2 \times 2^{-7} = \frac{3}{4} \times 2^{-7} = 3 \times 2^{-9}$$

b. $$X^{++} - X = (X^{++} - X^{+}) - (X - X^{+}) = 2^{-7} - (0.110..)_2 \times 2^{-7} = (1 - \frac{3}{4}) \times 2^{-7} = \frac{1}{4} \times 2^{-7} = 2^{-9}$$

eligiendo la menor diferencia tenemos que $fl(x) = X^{++}$, siendo el error de redondeo absoluto

$$|fl(x) - X| = \frac{1}{4} \times 2^{-7} = 2^{-9}$$

por tanto, el error de redondeo relativo es

$$\left|\frac{fl(x) - X}{X}\right| = \frac{2^{-9}}{(0.111...11100...)_2 \times 2^{17}} = \frac{2^{-9} \times 2^{-16}}{(0.111...11100...)_2} = \frac{2^{-25}}{(0.111...11100...)_2} < 2^{-25} < 2^{-9}$$

Ahora considere que se tienen dos números y se realiza una operación con ellos, primero debemos realizar la combinación correspondiente (suma, resta, multiplicación o división) posteriormente se normaliza continuando con un redondeo, y finalmente se almacena en memoria, entonces $fl(x \Theta y)$ es el número almacenado, que denota que se ha realizado una operación básica con *x* e *y*.

TEOREMA DEL ANÁLISIS DEL ERROR RELATIVO

"Sean X_0 , X_1 , ... , X_n números máquina positivos en una computadora, cuyo error de redondeo unitario es *E*. Entonces el error de redondeo relativo al calcular $\sum_{i=0}^{n} X_i$ de la manera usual es a lo sumo $(1 + E)^n - 1$. Esta cantiddad es aproximadamente *nE* ."

TEOREMA SOBRE LA PÉRDIDA DE PRECISIÓN

"Si *X* e *Y* son dos números de máquina binarios normalizados de punto flotante, positivos, tales que $X > Y$ con

$$2^{-q} \leq 1 - \frac{Y}{X} \leq 2^{-p}$$

entonces en la resta se pierden a lo más *q*, y al menos *p* bits significativos."

Cuando en un proceso informal los errores pequeños producidos en alguna etapa del cálculo se propagan en forma creciente en los cálculos posteriores y afectan el resultado del cálculo, es decir, son del orden del cálculo; entonces se dice que es un proceso numérico inestable.

Cuando los datos iniciales son variados con algún número pequeño y nuestros resultados difieren notoriamente, es decir, pequeños cambios en los datos iniciales

producen grandes variaciones en los resultados; entonces se dice que nuestro planteamiento del problema está mal condicionado.

$$f(x+h) - f(x) = f'(\varepsilon)h \approx hf'(x)$$

donde se ha aplicado el teorema del valor medio. Si *f'*(*x*) es pequeña, la variación en x, h, ha provocado un efecto de variación pequeño en la función *f(x)*.

Se observa que $\frac{x}{h}$ es el valor relativo de la perturbación en la variable independiente. El valor relativo de la perturbación en la variable dependiente es

$$\frac{f(x+h) - f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = \left[\frac{xf'(x)}{f(x)}\right]\left(\frac{h}{x}\right)$$

y se llama número de condición al cociente

$$\frac{xf'(x)}{f(x)}$$

SUMA

Si un número es el resultado de la adición de otros dos números, en donde

$$x = x_0 \pm \delta x$$
$$y = y_0 \pm \delta y$$

entonces

$$z_0 = x_0 + y_0$$
$$\delta z = \delta x + \delta y$$

o sea

$$z_0 \pm \delta z = (x_0 + y_0) \pm (\delta x + \delta y)$$

RESTA

Considerando los números definidos anteriormente

$$x = x_0 \pm \delta x$$
$$y = y_0 \pm \delta y$$

sea *z* la diferencia, entonces

$$z_0 = x_0 - y_0$$

$$\delta z = \delta x + \delta y$$

entonces

$$z_0 \pm \delta z = (x_0 - y_0) \pm (\delta x + \delta y)$$

Se concluye que tanto en la suma como en la resta de dos magnitudes el error de la operación se define como la suma de los errores en los números.

MULTIPLICACIÓN Y/O DIVISIÓN

Se consideran simultáneamente un producto y un cociente, las conclusiones son las mismas y el procedimiento se simplifica.

Sea

$$x = \frac{b \cdot c}{a}$$

donde b, c y a son los números con los errores δ b, δ c y δ a respectivamente. La magnitud x debe obtenerse a través de operaciones, un producto y un cociente. Obviamente, x tiene un error δ x, entonces

$$x \pm \delta x = \frac{(b \pm \delta b)(c + \delta c)}{a \pm \delta a}$$

Determinando el valor máximo

1. $$x + \delta x = \frac{(b + \delta b)(c + \delta c)}{a - \delta a}$$

Determinando el valor mínimo

2. $$x - \delta x = \frac{(b - \delta b)(c - \delta c)}{a + \delta a}$$

Se efectúa la diferencia de las expresiones 1 y 2 obteniendo

$$\delta x = \frac{1}{2}\left[\frac{(bc + b\,\delta c + c\,\delta b + \delta b\,\delta c)(a + \delta a) - (bc - b\,\delta c - c\,\delta b + \delta b\,\delta c)(a - \delta a)}{a^2 - (\delta a)^2}\right]$$

Como saca δ b<<1 y δ c<<1, cualquier combinación de productos de errores resulta insignificante, por lo que se obtiene

$$\frac{\delta x}{x} = \frac{\delta c}{c} + \frac{\delta b}{b} + \frac{\delta a}{a} \quad \delta x = \frac{ab\,\delta c + ac\,\delta b + bc\,\delta a}{a^2}$$

además

$$\frac{\delta x}{x} = \frac{\delta c}{c} + \frac{\delta b}{b} + \frac{\delta a}{a}$$

o sea que, el error relativo, de productos y/o cocientes es la suma de los errores relativos.