Unidad 1: Punto Flotante

Carlos Alliera

Tipos de error

Supongamos que queremos estimar una magnitud θ por medio de una aproximación $\hat{\theta}$, los errores que podemos cometer se pueden clasificar en:

- Error Absoluto:

$$E_A := |\hat{\theta} - \theta|$$

- Error Relativo:

$$E_R := \frac{E_A}{|\theta|} = \frac{|\hat{\theta} - \theta|}{|\theta|}$$

Representación de números de máquina

Épsilon de la máquina

El menor número tal que, en la máquina

$$1 + \varepsilon \neq 1$$

es el épsilon de la máquina.

Este número también es conocido como error de redondeo unitario.

No se confunda!

Las máquinas no reconocen a ε como 0, ni siquiera a valores $\delta < \varepsilon$, sino que el ε es el menor número *no despreciable* cuando es sumado a 1.

Representación de números de máquina

Dado un $x \in \mathbb{R}$ cualquiera, el *redondeado* de x, $fl(x)$ es el número de máquina que verifica:

$$fl(x) = x^* = x(1+\delta) \text{ con } |\delta| \leq \varepsilon = 5 \cdot 10^{-m}$$

Por lo anterior, el error relativo de aproximar x con x^* es δ.

Representación de números de máquina

La siguiente representación de números de máquina:

$$x^* = 0, a_1 a_2 ... a_m \cdot 10^l, \ a_i \in \{n \in \mathbb{N}_0, \ n \leq 9\} \ \forall i, \ |a_1| > 0$$

se la denomina de punto flotante.
En este caso decimos que conocemos a x con $m \in \mathbb{N}$ dígitos significativos.
El exponente l verifica $-M_1 \leq l \leq M_2$ donde $M_1, \ M_2$ dependen de la máquina.

Problema

Sumar los números $x = 12,\ y = 0,004$ y $z = 0,003$ y empleamos una aritmética de 4 dígitos y base 10 empleando el método de redondeo. En este caso tenemos 3 números de máquina, pues:

$$x = fl(x) = 0,12 \cdot 10^2,\ y = fl(y) = 0,4 \cdot 10^{-2},\ z = fl(z) = 0,3 \cdot 10^{-2}$$

Si sumamos $(x + y) + z$ en ese orden:

$$fl(x + y) = fl(12,004) = fl(0,12004 \cdot 10^2) = 0,12 \cdot 10^2 = fl(x)$$

pues al observar el quinto dígito y ver que es menor a 5, la aritmética de cuatro dígitos *redondea para abajo*.

Entonces al sumar el número que nos falta:

$$fl(fl(x+y)+z) = fl(x+z) = fl(12,003) =$$

$$= fl(0,12003 \cdot 10^2) = 0,12 \cdot 10^2 = fl(x)$$

Nos queda que para nuestra máquina:

$$fl(fl(x+y)+z) = fl(x)$$

Sin embargo, si sumamos de menor a mayor: $(z + y) + x$:

$$fl(y + z) = fl(0,007) = 0,7 \cdot 10^{-2}$$

y luego:

$$fl(fl(z + y) + x) = fl(0,12007 \cdot 10^{2}) = 0,1201 \cdot 10^{2}$$

pues *redondea para arriba* ya que el quinto dígito es mayor a 5.
Este resultado se aproxima más al verdadero valor de $x + y + z$.

Sean $a,\ b \in \mathbb{R}$ dos números cualesquiera, y $\star$ es una operación algebraica, entonces los flotantes se calculan en cada paso del desarrollo de la operación:

Lo que hace la máquina	Lo que debe dar
$fl(fl(a) \star fl(b))$	$fl(a \star b)$

En cada instancia de cálculo de redondeos se pueden perder dígitos significativos, lo que ocasiona errores de cálculo.

Un problema de parcial

Este ejercicio integró el primer parcial del segundo cuatrimestre de 2010.
Sea $a \in \mathbb{R} - \{0\}$ un número de máquina. Se quiere calcular a^4 con aritmética de punto flotante.
Probar que el error relativo de este cálculo se puede acotar por

$$3\varepsilon + 3\varepsilon^2 + \varepsilon^3$$

siendo ε el correspondiente épsilon de máquina.

Un problema de parcial

Se sabe que

$$a^4 = a \cdot (a \cdot (a \cdot a)) \Rightarrow fl(a^4) = (a^* \cdot (a^* \cdot (a^* \cdot a^*)^*)^*)^* =$$

pero como $a^* = a$ (pues es de máquina)

$$= (a \cdot (a \cdot (a \cdot a)^*)^*)^* = (a \cdot (a \cdot (a^2(1+\delta_1))^*)^* =$$

$$= (a \cdot a^3(1+\delta_1)(1+\delta_2))^* = a^4(1+\delta_1)(1+\delta_2)(1+\delta_3) \leq$$

$$\leq a^4(1+\varepsilon)(1+\varepsilon)(1+\varepsilon) = a^4(1+\varepsilon)^3 = a^4(1+3\varepsilon+3\varepsilon^2+\varepsilon^3)$$

Un problema de parcial

Planteamos la fórmula del error relativo:

$$\frac{|fl(a^4) - a^4|}{|a^4|} = \frac{|\cancel{a^4}^{1}(1+\delta_1)(1+\delta_2)(1+\delta_3) - \cancel{a^4}^{1}|}{|\cancel{a^4}^{1}|} =$$

$$= |(1+\delta_1)(1+\delta_2)(1+\delta_3) - 1| = |((1+\delta_1) + \delta_2(1+\delta_1))(1+\delta_3) - 1| =$$

$$= |\cancel{1} + \delta_1 + \delta_3(1+\delta_1) + \delta_2(1+\delta_1)(1+\delta_3) - \cancel{1}| =$$

$$\leq |\delta_1| + |\delta_3|(1+|\delta_1|) + |\delta_2|(1+|\delta_1|)(1+|\delta_3|) \leq \varepsilon + \varepsilon(1+\varepsilon) + \varepsilon(1+\varepsilon)^2 =$$

$$= 2\varepsilon + \varepsilon^2 + \varepsilon(1 + 2\varepsilon + \varepsilon^2) = \boxed{3\varepsilon + 3\varepsilon^2 + \varepsilon^3 = O(\varepsilon)}$$

Para pensar

Se dispone de una máquina que utiliza el método de redondeo en base 10 y mantisa de 2 dígitos.
Respetando el orden de la suma, calcular:

$$1 + 1/2 + 1/4 + 1/8 + 1/16$$

Si se sabe que $1 + 1/2 + 1/4 + 1/8 + 1/16 = 1,9375$,

1. Estimar el error relativo.
2. Sumar de otra forma para que el error relativo sea menor y estimarlo en este caso.

Trucos para evitar errores

- Binomio de Newton

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

- Polinomio de Taylor.
- Igualdades Trigonométricas
  1. $\operatorname{sen}(x+y) = \operatorname{sen}(x)\cos(y) + \operatorname{sen}(y)\cos(x)$
  2. $\cos(x+y) = \cos(x)\cos(y) - \operatorname{sen}(y)\operatorname{sen}(x)$
- Multiplicar y dividir por el *conjugado*.

Práctica 1: Ejercicio 4

Enunciado

a) Sean a y b dos número de máquina. Demostrar que el error relativo que se comete al calcular a^2b con aritmética de punto flotante se puede acotar por $2\varepsilon + O(\varepsilon^2)$, donde ε es el épsilon de la máquina asociado a una aritmética de punto flotante.

b) Demostrar que si en cambio $a,\ b \in \mathbb{R}$ son dos números reales arbitrarios, entonces dicho error se puede acotar por $5\varepsilon + O(\varepsilon^2)$

Resolución de la parte a)

Como a y b son número de máquina, entonces verifican:

$$a = fl(a) = a^* \text{ y } b = fl(b) = b^*$$

es decir que $\delta_a = \delta_b = 0$.

Planteamos la fórmula del error relativo:

$$\frac{|fl(fl(fl(a)fl(a))fl(b)) - a^2b|}{|a^2b|} = \frac{|fl(fl(a^2)b) - a^2b|}{|a^2b|} =$$

$$= \frac{|fl(a^2(1+\delta_{a^2})b) - a^2b|}{|a^2b|} = \frac{|a^2b(1+\delta_{a^2})(1+\delta_{\times b}) - a^2b|}{|a^2b|} =$$

porque

$$fl(a^2) = a^2(1+\delta_{a^2})$$

y $\delta_{\times b}$ es el δ que se obtiene tras multiplicar por b

ahora reordenamos un poco...

$$= \frac{|a^2 b(1 + \delta_{a^2} + \delta_{\times b} + \delta_{a^2}\delta_{\times b}) - a^2 b|}{|a^2 b|} =$$

$$= \frac{|\cancel{a^2 b}(\cancel{1} + \delta_{a^2} + \delta_{\times b} + \delta_{a^2}\delta_{\times b} - \cancel{1})|}{|\cancel{a^2 b}|} =$$

$$|\delta_{a^2} + \delta_{\times b} + \delta_{a^2}\delta_{\times b}| \leq |\delta_{a^2}| + |\delta_{\times b}| + |\delta_{a^2}||\delta_{\times b}| \leq \varepsilon + \varepsilon + \varepsilon^2 = 2\varepsilon + O(\varepsilon^2)$$

que era lo que se quería probar.

Resolución de la parte b)

En este caso

$$fl(a) = a(1+\delta_a) \text{ y } fl(b) = b(1+\delta_b)$$

al plantear nuevamente el error relativo:

$$\frac{|fl(fl(fl(a)fl(a))fl(b)) - a^2b|}{|a^2b|} = \frac{|fl(fl(a^2(1+\delta_a)^2)b(1+\delta_b)) - a^2b|}{|a^2b|} =$$

$$= \frac{|fl(a^2(1+\delta_a)^2(1+\delta_{a^2})b(1+\delta_b)) - a^2b|}{|a^2b|}$$

Resolución de la parte b)

observemos que cada operación $(*)$ con fl "aporta" un nuevo $1+\delta_*$:

$$= \frac{|a^2b(1+\delta_a)^2(1+\delta_{a^2})(1+\delta_b)(1+\delta_{\times b}) - a^2b|}{|a^2b|} = \overset{1}{\cancel{1}} \qquad \overset{1}{\cancel{|a^2b|}}$$

$$= |(1+\delta_a)^2(1+\delta_{a^2})(1+\delta_b)(1+\delta_{\times b}) - 1| \underbrace{\leq 5\varepsilon + O(\varepsilon^2)}_{\text{Probar esto}}$$

A partir de acá, distribuya, calcule con cuidado, acote correctamente y concluye el ejercicio.

Estamos acostumbrados al desarrollo en base decimal de números, pero no es la única, y no siempre es la ideal.
Sean un número $x \in \mathbb{R}$ y una *base* $\beta \in \mathbb{N}$, decimos que el desarrollo en base β de x es:

$$(x)_\beta = a_1 a_2 ... a_k, \quad k \in \mathbb{N}$$

con $a_i \in \mathbb{N}_0,\ a_i < \beta$ si se verifica:

$$x = a_0\beta^k + a_1\beta^{k-1} + ... + a_{k-2}\beta^2 + a_{k-1}\beta^1 + a_k\beta^0 = \sum_{j=0}^{k} a_j \beta^{k-j}$$

En general, si usamos una máquina que trabaja en una base β, con $m \in \mathbb{N}$ dígitos significativos el error de redondeo de aproximar $x = 0, a_1 .... * 10^l$ con el $x^* = fl(x) = 0, a_1 ....a_m * 10^l$,

$$E_R(x) = \frac{|x - x^*|}{|x|} \leq \frac{1}{2}\beta^{l-m} := \varepsilon$$

y en el caso del error relativo para el truncamiento:

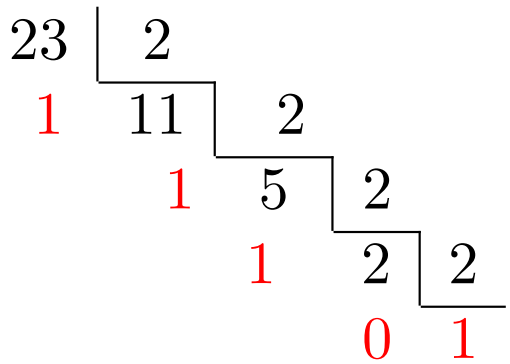
$$E_R(x) = \frac{|x - x^*|}{|x|} \leq \beta^{l-m}$$

Base Decimal vs Base Binaria

En base decimal es muy fácil desarrollar cualquier número:

$$(23)_{10} = 2 * 10^1 + 3 * 10^0$$

pero si queremos la base binaria del número 23 hay que obtenerla así:



$\longleftarrow$ Mire así

$10111 = (23)_2$ pues:

$$23 = 1 * 2^4 + 0 * 2^3 + 1 * 2^2 + 1 * 2^1 + 1 * 2^0$$

Representación binaria: Ejemplo

Queremos hallar la expresión binaria de $0,6$:

$$0,6 = c_1 2^{-1} + c_2 2^{-2} + c_3 2^{-3} + ...$$

$$\times 2 \Rightarrow 1,2 = c_1 + c_2 2^{-1} + c_3 2^{-2} + ... \Rightarrow c_1 = 1 \Rightarrow \quad 0,2 = c_2 2^{-1} + c_3 2^{-2} + c_4 2^{-3} ...$$

$$\times 2 \Rightarrow 0,4 = c_2 + c_3 2^{-1} + c_4 2^{-2} + ... \Rightarrow c_2 = 0 \Rightarrow \quad 0,4 = c_3 2^{-1} + c_4 2^{-2} + c_5 2^{-3} ...$$

$$\times 2 \Rightarrow 0,8 = c_3 + c_4 2^{-1} + c_5 2^{-2} + ... \Rightarrow c_3 = 0 \Rightarrow \quad 0,8 = c_4 2^{-1} + c_5 2^{-2} + c_6 2^{-3} ...$$

$$\times 2 \Rightarrow 1,6 = c_4 + c_5 2^{-1} + c_6 2^{-2} + ... \Rightarrow c_4 = 1 \Rightarrow \quad 0,6 = c_5 2^{-1} + c_6 2^{-2} + c_7 2^{-3} + ...$$

Se repite, entonces:

$$\boxed{(0,6)_2 = 0,\widehat{1001}}$$

Ejemplo 1

Cuando restamos dos números muy parecidos, el error de redondeo se puede propagar de manera preocupante.
Por ejemplo, sean $p = 0,54617$ y $q = 0,54601$ y usamos para calcular $p - q$ una máquina que usa 4 cifras significativas.
Al redondear los números se obtiene:

$$p^* = 0,5462,\ q^* = 0,5460 \Rightarrow p^* - q^* = 0,0002 = 0,2 * 10^{-3}$$

el error relativo es:

$$\frac{|(p^* - q^*) - (p - q)|}{|p - q|} = \frac{|0,0002 - 0,00016|}{0,00016} = \frac{1}{4} = 0,25$$

El resultado tiene una cifra significativa. (Un desastre!)

Ejemplo 1

El truncamiento... no mejora las cosas

$$p^* = 0,5461,\ q^* = 0,5460 \Rightarrow p^* - q^* = 0,0001 = 0,1 * 10^{-3}$$

el error relativo es:

$$\frac{|(p^* - q^*) - (p - q)|}{|p - q|} = \frac{|0,0001 - 0,00016|}{0,00016} = \frac{3}{8} = 0,375$$

Un error mayor.
La pérdida de precisión se puede resolver reformulando el problema, como veremos ahora.

Ejemplo 2

Con aritmética de 4 dígitos, hallar las soluciones de:

$$x^2 + 62,1x + 1 = 0$$

mediante la clásica fórmula se obtiene que las raíces aproximadamente son:

$$x_1 = -0,01610723 \qquad x_2 = -62,08390$$

Pero al usar la fórmula para el cálculo de x_1 con la máquina se llega a lo siguiente,

$$fl(x_1) = \frac{-62,1 + 62,07}{2} = -0,015$$

Ejemplo 2

que produce un error relativo:

$$\frac{|x_1^* - x_1|}{|x_1|} = \frac{|-0,015 + 0,01610723|}{0,01610723} \approx 0,06874$$

Sin embargo, si se calcula la raíz con una fórmula alternativa:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \cdot \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

nos queda

$$fl(x_1) = \left(\frac{-2}{62,1 + 62,07}\right)^* = -0,0161$$

cuyo error relativo es bastante menor:

$$\frac{|x_1^* - x_1|}{|x_1|} = \frac{|-0,0161 + 0,01610723|}{0,01610723} \approx 4,488 * 10^{-4}$$

¿Qué error se comete al aproximar $f'(x_0)$ con $d_h = \frac{f(x_0+h)-f(x_0)}{h}$?
Sea $d_h^* = fl(d_h)$ y se supone que $f \in C^2$

$$|f'(x_0) - d_h^*| = |f'(x_0) - d_h + d_h - d_h^*| \leq |f'(x_0) - d_h| + |d_h - d_h^*|$$

por Taylor

$$f(x_0 + h) = f(x_0) + hf'(x_0) + h^2 \frac{f''(\xi)}{2}, \quad \xi \text{ entre } x_0 \text{ y } x_0 + h$$

entonces:

$$d_h = f'(x_0) + h\frac{f''(\xi)}{2} \Rightarrow |f'(x_0) - d_h| = \left| h\frac{f''(\xi)}{2} \right| \to 0$$

Se puede ver tras algunas cuentas que:

$$|d_h - d_h^*| \leq |d_h|(2\varepsilon + \varepsilon^2) + \varepsilon^2(1+\varepsilon)^2 \underbrace{\frac{|f(x_0+h)| + |f(x_0)|}{h}}_{\to\infty} \text{ cuando } h \to 0$$

Aclaración

Quizá le resulte raro que el último término de la expresión anterior tienda a ∞ cuando $h \to 0$, en realidad para aproximar una derivada mediante d_h previamente se fija un valor de h preferentemente pequeño cuando trabajamos con una máquina.