Introducción al Análisis Numérico. Errores

María González Taboada

Departamento de Matemáticas

27 de febrero de 2007

Esquema:

Redondeo y redondeo a cero en el sistema decimal

- Sea x un número decimal en notación científica normalizada:

$$x = \pm 0.d_1 d_2 \cdots \times 10^n = \pm \left(\sum_{k=1}^{+\infty} d_k \times 10^{-k} \right) \times 10^n$$

con $1 \leq d_1 \leq 9$ y $0 \leq d_k \leq 9$, para $k \geq 2$.

- Para aproximar x por un número x^* con una precisión de p dígitos:
  - Truncamiento o redondeo a cero.
  - Redondeo.

Redondeo a cero en el sistema decimal

$$x = \pm 0.d_1 d_2 \cdots \times 10^n = \pm \left(\sum_{k=1}^{+\infty} d_k \times 10^{-k} \right) \times 10^n$$

- **Truncamiento o redondeo a cero con p cifras**:

$$x \approx x^* = \pm\, 0.d_1 d_2 \ldots d_p \times 10^n$$

Ejemplo:

(a) *Si $x = 0{,}99995$ y $p = 4$, su aproximación por redondeo a cero es $x^* = 0{,}9999 \times 10^0$.*

(b) *Si $x = 0{,}4332609$ y $p = 3$, su aproximación por redondeo a cero es $x^* = 0{,}433 \times 10^0$.*

Redondeo en el sistema decimal

$$x = \pm 0.d_1 d_2 \cdots \times 10^n = \pm \left(\sum_{k=1}^{+\infty} d_k \times 10^{-k} \right) \times 10^n$$

- **Redondeo con p cifras**:

$$x \approx x^* = \begin{cases} \pm 0.d_1 d_2 \ldots d_p \times 10^n & \text{si } 0 \leq d_{p+1} \leq 4 \\ \pm (0.d_1 d_2 \ldots d_p + 10^{-p}) \times 10^n & \text{si } 5 \leq d_{p+1} \leq 9 \end{cases}$$

Ejemplo:

(a) *Si* $x = 0{,}99995$ *y* $p = 4$*, su aproximación por redondeo es* $x^* = 0{,}1 \times 10^1$*.*

(b) *Si* $x = 0{,}4332609$ *y* $p = 3$*, su aproximación por redondeo es* $x^* = 0{,}433 \times 10^0$*.*

Redondeo y redondeo a cero en el sistema binario

- Supongamos que usamos un formato de coma flotante de precisión p.
- Si la mantisa de un número x contiene más de p dígitos binarios, x no se puede almacenar de forma exacta.
  En su lugar se almacena x^*, obtenido por uno de los métodos siguientes:
  - **Truncamiento o redondeo a cero**:
    Se almacenan los p primeros dígitos binarios de la mantisa, prescindiendo del resto.
  - **Redondeo**:
    La mantisa se trunca a p dígitos y, si el dígito $p+1$ es 1, se le suma $(0{,}0\ldots01)_2 = 2^{-p+1}$.

Redondeo vs. redondeo a cero en el sistema binario

- Puede probarse que:

$$x^* = x(1+\gamma)$$

con:

- $\gamma \in [-2^{-p+1}, 0]$, si se usa redondeo a cero,
- $\gamma \in [-2^{-p}, 2^{-p}]$, si se usa redondeo.

($p = 24$ en precisión simple y $p = 53$ en precisión doble).

- Por tanto, el peor error posible es el doble cuando se usa redondeo a cero que cuando se usa redondeo.

Redondeo vs. redondeo a cero en el sistema binario

- Cuando se usa **redondeo a cero**, el error es del mismo signo que el número almacenado:

$$x - x^* = -\gamma x \qquad -2^{-p+1} \leq \gamma \leq 0$$

- Cuando se usa **redondeo**, el error puede ser negativo o positivo:

$$x - x^* = -\gamma x \qquad -2^{-p} \leq \gamma \leq 2^{-p}$$

- Por tanto, cuando se realizan muchas operaciones aritméticas, la propagación del error es mejor cuando se usa redondeo.

Redondeo vs. redondeo a cero en el sistema binario

Ejemplo:

Sea $x = (1{,}10011)_2 = (1{,}59375)_{10}$.

- *Su aproximación por redondeo a cero con* 5 *dígitos es* $x_0^* = (1{,}1001)_2 = (1{,}5625)_{10}$.
  En este caso, $\gamma = -0{,}019607... \in [-2^{-4}, 0]$.

- *Su aproximación por redondeo con* 5 *dígitos es* $x^* = (1{,}1010)_2 = (1{,}625)_{10}$.
  En este caso, $\gamma = +0{,}019607... \in [-2^{-5}, 2^{-5}]$.

Para tener en cuenta...

- Algunos números con una expresión decimal finita tienen una expresión binaria infinita. Por ejemplo,

$$(0{,}1)_{10} = (0{,}0001100110011\ldots)_2$$

- Estos números no se pueden representar de forma exacta en el ordenador.
- Como consecuencia, pueden obtenerse resultados erróneos al trabajar con ellos. No conviene, por ejemplo, utilizarlos como valores finales de la variable de control en un ciclo.

Para tener en cuenta...

- En los lenguajes de programación que disponen de precisión simple y doble (como Fortran y C) hay que especificar el tipo de las constantes correctamente.
  En caso contrario, podemos obtener resultados erróneos debido a los errores de redondeo.

- En Matlab, todos los cálculos se hacen en precisión doble.

Error absoluto y error relativo

- Sea x un número real y sea $\widehat{x}$ una aproximación de x.
- Se llama error absoluto entre x y $\widehat{x}$ al valor:

$$e_a = |x - \widehat{x}|$$

- Se llama error relativo entre x ($x \neq 0$) y $\widehat{x}$ al valor:

$$e_r = \frac{|x - \widehat{x}|}{|x|} = \frac{e_a}{|x|}$$

Error absoluto y error relativo

Ejemplo:

x	$\widehat{x}$	e_a	e_r
$0{,}3000 \times 10^{1}$	$0{,}3100 \times 10^{1}$	$0{,}1$	$0{,}333\overset{\frown}{3} \times 10^{-1}$
$0{,}3000 \times 10^{-3}$	$0{,}3100 \times 10^{-3}$	$0{,}1 \times 10^{-4}$	$0{,}333\overset{\frown}{3} \times 10^{-1}$
$0{,}3000 \times 10^{4}$	$0{,}3100 \times 10^{4}$	$0{,}1 \times 10^{3}$	$0{,}333\overset{\frown}{3} \times 10^{-1}$

Cifras significativas

- Se dice que $\widehat{x}$ aproxima a x con t cifras o dígitos significativos cuando t es el mayor entero no negativo tal que:

$$\frac{|\widehat{x} - x|}{|x|} \leq 5 \times 10^{-t}.$$

Ejemplo:

(a) $\widehat{x} = 124{,}45$ *aproxima a* $x = 123{,}45$ *con* 2 *cifras significativas:*

$$\frac{|\widehat{x} - x|}{|x|} = \frac{1}{123{,}45} = 0{,}0081 \leq 5 \times 10^{-2}$$

Cifras significativas

Ejemplo:

(b) $\widehat{x} = 0{,}0012445$ *aproxima a* $x = 0{,}0012345$ *con* 2 *cifras significativas:*

$$\frac{|\widehat{x} - x|}{|x|} = \frac{0{,}00001}{0{,}0012345} = 0{,}0081 \leq 5 \times 10^{-2}$$

(c) $\widehat{x} = 999{,}8$ *aproxima a* $x = 1000$ *con* 4 *cifras significativas:*

$$\frac{|\widehat{x} - x|}{|x|} = \frac{0{,}2}{1000} = 0{,}0002 \leq 5 \times 10^{-4}\,.$$