Clase 3
Análisis del Error

3.1 Números en Formato Decimal

El uso del **formato binario** cuando realizamos operaciones es incómodo y muy tedioso ya que se necesitan muchas cifras para representar valores numéricos.

Por dicho motivo, usualmente el formato usado es el decimal o la **base 10**.

La representación de un número x en formato decimal o en base 10 en el formato llamado punto flotante es la siguiente

$$x = \pm 0.d_1 d_2 \ldots d_k \times 10^n, \qquad d_i \in \{0, \ldots, 9\}, d_1 \neq 0$$

El valor k es el número de cifras significativas del número x y n es el exponente del número.

Vimos anteriormente que si usamos la representación binaria de 64 bits, el valor máximo de cifras significativas es $k_{máx} = 16$

Ejemplo 1:

La representación punto-flotante de los números siguientes es:

- $x = 31.41593$ es $fl(x) = 0.3141596 \times 10^2$
- $x = -0.00004356$ es $fl(x) = -0.4356 \times 10^{-4}$
- $x = 19875.24$ es $fl(x) = 0.1987524 \times 10^5$

3.1.1 Representación en punto flotante

Si el número de cifras significativas del número x supera el valor k (número máximo de cifras significativas con las que trabajamos), la representación punto-flotante de x, $fl\left(x\right)$ será una aproximación de x.

Para hallar dicha aproximación, podemos hacer dos cosas:

- **Cortar (Truncamiento).** En este caso si

$$x = \pm 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots \times 10^n,$$

  la representación de punto flotante es

$$fl\left(x\right) = \pm 0.d_1 d_2 \ldots d_k \times 10^n$$

- **Redondear.** En este caso si

$$x = \pm 0.d_1 d_2 \ldots d_k d_{k+1} d_{k+2} \ldots \times 10^n,$$

  la representación de punto flotante es

$$fl\left(x\right) = \begin{cases} \pm 0.d_1 d_2 \ldots \left(d_k + 1\right) \times 10^n & \textbf{si } d_{k+1} \geq 5 \\ \pm 0.d_1 d_2 \ldots d_k \times 10^n & \textbf{si } d_{k+1} < 5 \end{cases}$$

3.1.2 Corte y Redondeo

En el caso de **redondear**, si $d_{k+1} \geq 5$ y $d_k + 1 = 10$, la representación **punto-flotante** de x será $fl\left(x\right) = \pm 0.d_1 \ldots \left(d_{k-1} + 1\right) 0$. Si $d_{k-1} + 1 = 10$, la representación **punto-flotante** de x será $fl\left(x\right) = \pm 0.d_1 \ldots \left(d_{k-2} + 1\right) 00$ y así sucesivamente.

Ejemplo 2:

Supongamos que trabajamos con $k = 4$ **cifras significativas.** La representación de los siguientes números vale:

- $x = 1234.5678$, $fl\left(x\right) = 0.1234 \times 10^4$ si cortamos y $fl\left(x\right) = 0.1235$ si redondeamos.
- $x = -0.00004599881234$, $fl\left(x\right) = -0.4599 \times 10^{-4}$ si cortamos y $fl\left(x\right) = -0.4600 \times 10^{-4}$ si redondeamos

Ejercicio 1:

Construir un *programa* en **Python** que permita representar un número real en formato de coma flotante a través del método del redondeo con k cifras significativas.

Programa de Google Colab

Este programa se ha desarrollado usando Google Colab. Para ver el programa haz clic en el logo.

3.2 Errores Absoluto y Relativo

Definición 1: Definición de Error Absoluto y Relativo

Sea x un valor real y $\widehat{x}$ una aproximación del mismo. Definiremos error absoluto de x a la cantidad $e_a(x) = |x - \widehat{x}|$ y error relativo a la cantidad $e_r(x) = \frac{|x-\widehat{x}|}{|x|}$

Ejemplo 3:

Calculemos los errores absolutos y relativos para los valores numéricos del ejemplo anterior cuando truncamos o redondeamos

- $x = 1234.5678$
  - **corte (truncamiento):** $fl_c(x) = 0.1234 \times 10^4$ **entonces** $e_a(x) = |1234.5678 - 1234| = 0.5678$ **y** $e_r(x) = \frac{e_a(x)}{|x|} = \dfrac{0.5678}{1234.5678} = 0.000459918038 \approx 4.599 \times 10^{-4} \approx 0.459 \times 10^{-3}$
  - **redondeo:**
    $fl_r(x) = 0.1235 \times 10^4$ **entonces** $e_a(x) = |1234.5678 - 1235| = 0.4322$ **y** $e_r(x) = \frac{e_a(x)}{|x|} = \dfrac{0.4322}{1234.5678} = 0.000350082029 \approx 3.500 \times 10^{-4} \approx 0.350 \times 10^{-3}$

- $x = -0.00004599881234$
  - **corte (truncamiento):** $fl_c(x) = -0.4599 \times 10^{-4}$ **entonces** $e_a(x) = |-0.00004599881234 - (-0.00004599)| = 0.000000008812 \approx 0.88 \times 10^{-8}$ **y** $e_r(x) = \frac{e_a(x)}{|x|} = \dfrac{0.000000008812}{|-0.00004599881234|} = 0.000191570163 \approx 1.916 \times 10^{-4} \approx 0.192 \times 10^{-3}$
  - **redondeo:** $fl_r(x) = -0.4600 \times 10^{-4}$ **entonces** $e_a(x) = |-0.00004599881234 - (-0.000046)| = 0.000000001188 \approx 0.19 \times 10^{-8}$ **y** $e_r(x) = \frac{e_a(x)}{|x|} = \dfrac{0.000000001188}{0.00004599881234} = 0.000025819362 \approx 2.582 \times 10^{-5} \approx 0.258 \times 10^{-4}$

Ejercicio 2:

Crear un *programa* en **Python** que calcule el error absoluto y el error relativo dado un número y su redondeo a k cifras significativas.

Programa de Google Colab

Este programa se ha desarrollado usando Google Colab. Para ver el programa haz clic en el logo.

Observaciones

- Mirando los ejemplos anteriores, vemos que los **errores absolutos** dependen de las magnitudes de los valores x: en el primer ejemplo los errores absolutos son de orden de 10^{-3} y en el segundo del orden de 10^{-8}
- En cambio, los **errores relativos** no se ven afectados por dichas magnitudes. Por dicho motivo, si queremos estudiar los errores sin tener en cuenta el orden de los valores x, hay que usar los errores relativos. Vemos que en los dos ejemplos los errores relativos son

del orden de 10^{-3} y 10^{-4} ya que recordemos que trabajamos con 4 **cifras decimales significativas**

Taller

Escribir para cada uno de los siguientes ejercicios un *programa* en **Python** que determine el error absoluto y el error relativo al resolver las siguientes ecuaciones lineales. El programa debe tener un parámetro que le permita configurar las cifras significativas del cálculo para los números reales, tales como raíces y fracciones.

1. $\sqrt{2}x + \frac{\sqrt{3}}{2} = \sqrt{3}x + \frac{\sqrt{2}}{3}$
2. $\left(1+\sqrt{2}\right)x - \sqrt{3} = \left(1+\sqrt{5}\right)x + \sqrt{7}$
3. $\frac{x+\frac{\sqrt{2}}{3}}{x-\frac{\sqrt{3}}{5}} = \frac{\frac{1}{2}+\frac{\sqrt{5}}{7}}{\frac{1}{3}+\frac{\sqrt{7}}{11}}$
4. $\displaystyle\sum_{n=1}^{7} \frac{1}{n}x = 89$
5. $\displaystyle\sum_{n=2}^{7} \frac{1}{n^2}x - n = 51$
6. $2\sin\left(\frac{\pi}{4}\right)x + 2\cos\left(\frac{\pi}{6}\right)x + 2\tan\left(\frac{\pi}{3}\right)x = 3\sqrt{3} + \sqrt{2}$
7. $\frac{\sqrt{34x}}{68} = \sqrt{131}\sin\left(\frac{\pi}{4}\right)$
8. $\displaystyle\frac{1}{135}\sum_{i=1}^{9} i^2 x = 4199$
9. $\left(2^3-1\right)x = 5\left(2^5 - \frac{x}{5}\right) + \frac{63+\sum_{s=1}^{2^5} s^2}{5^6}$
10. $\frac{5^6\left(2^3-1\right)x+11^2x}{3^2\sin\left(\frac{\pi}{4}\right)} = \frac{7^{\left(2^{\sum_{i=1}^{7}(2i-3)^5-11^4+20}\right)-5^6x+\sqrt{14641}x}}{\sqrt[3]{729}\cos\left(\frac{\pi}{4}\right)}$

3.3 Aritmética de Dígitos Finitos

3.3.1 Cifras Significativas

Vamos a formalizar la definición de aproximación de k cifras significativas:

Definición 2:
Diremos que la aproximación $\widehat{x}$ del valor x tiene k cifras significativas si el error relativo de la aproximación está acotado por:

$$e_r\left(x\right) = \frac{|x - \widehat{x}|}{|x|} \leq 5 \times 10^{-k}$$

Observamos que los dos ejemplos anteriores, las aproximaciones por corte y redondeo tenían 4 cifras significativas ya que en todos los casos los errores relativos estaban acotados por 5×10^{-4}

3.3.2 Operaciones Básicas

Aparte de los **errores de redondeo** que tenemos en la representación de los números, las operaciones básicas como la **suma, resta, multiplicación o división** realizadas en los computadores no son exactas ya que se cometen errores.

Las operaciones anteriores se realizan en **formato binario** en los computadores y son básicamente operaciones lógicas o de desplazamiento de bits.

A dicho tipo de **aritmética** se le denomina **aritmética de dígitos finitos**.

Sean $\oplus, \ominus, \odot$ y $\oslash$ las operaciones de la suma, resta, multiplicación y división, respectivamente, que realiza el computador.

Dados dos valores x e y, cuando nos planteamos el resultado de una operación entre ellos, $x + y$, $x - y$, $x \cdot y$ o x/y en realidad obtenemos lo siguiente: $fl\left(x\right)\oplus fl\left(y\right)$, $fl\left(x\right)\ominus fl\left(y\right)$, $fl\left(x\right)\odot fl\left(y\right)$ y $fl\left(x\right)\oslash fl\left(y\right)$ y los resultados se indican de la forma siguiente para poner de manifiesto los errores cometidos en las operaciones:

$$\begin{aligned}
fl\left(x\right)\oplus fl\left(y\right) &= fl\left(fl\left(x\right)+fl\left(y\right)\right)\\
fl\left(x\right)\ominus fl\left(y\right) &= fl\left(fl\left(x\right)-fl\left(y\right)\right)\\
fl\left(x\right)\odot fl\left(y\right) &= fl\left(fl\left(x\right)\cdot fl\left(y\right)\right)\\
fl\left(x\right)\oslash fl\left(y\right) &= fl\left(fl\left(x\right)/fl\left(y\right)\right)
\end{aligned}$$

Ejemplo 4:

Supongamos que trabajamos con 4 cifras significativas. Nos planteamos la operación siguiente

$$fl\left(x\right)\oplus fl\left(y\right)$$

con $x=\frac{1}{3}$ e $y=\frac{345}{7}$. Encontrar el valor real y el valor aproximado.

Como trabajamos con 4 cifras significativas, el valor de $fl\left(x\right)$ será $fl\left(x\right)=0.3333$. Por otra parte el valor de $fl\left(y\right)$ será $fl\left(y\right)=49.29$.

En las aproximaciones anteriores, hemos redondeado ya que vimos que redondear es mejor que cortar.

El valor de $fl\left(x\right)\oplus fl\left(y\right)$ será:

$$\begin{aligned}
fl\left(x\right)\oplus fl\left(y\right) &= fl\left(fl\left(x\right)+fl\left(y\right)\right)\\
&= fl\left(0.3333+49.29\right)\\
&= fl\left(49.6233\right)\\
&= 49.62
\end{aligned}$$

Fijémonos que tenemos los errores siguientes:

- Las aproximaciones de x e y a 4 cifras significativas
- El error que hemos cometido en la operación de la suma.

El valor real de la suma es

$$\frac{1}{3}+\frac{345}{7}=\frac{1042}{21}=49.619047619$$

con esto podemos calcular el error absoluto y el error relativo

$$\begin{aligned}
e_a\left(x\right) &= \left|49.619047619-49.62\right|\\
&= 0.000952380952\\
&\approx 9.524\times 10^{-4}
\end{aligned}$$

$$
\begin{aligned}
e_r(x) &= \frac{|49.619047619 - 49.62|}{|49.619047619|} \\
&= 0.000019193858 \\
&\approx 1.919 \times 10^{-5} \leq 5 \times 10^{-4}
\end{aligned}
$$

Ahora nos planteamos el problema de la multiplicación.

$$
\begin{aligned}
fl(x) \odot fl(y) &= fl(fl(x) \cdot fl(y)) \\
&= fl(0.3333 \cdot 49.29) \\
&= fl(16.428357) \\
&= 16.43
\end{aligned}
$$

Calculemos el error absoluto y el error relativo. Para ello tenemos que

$$
\frac{1}{3} \cdot \frac{345}{7} = \frac{115}{7} = 16.428571428571
$$

así que

$$
\begin{aligned}
e_a(x) &= |16.428571428571 - 16.43| \\
&= 0.001428571429 \\
&\approx 1.429 \times 10^{-3}
\end{aligned}
$$

y

$$
\begin{aligned}
e_r(x) &= \frac{e_a(x)}{|x|} = \frac{0.001428571429}{16.428571428571} \\
&= 0.000086956522 \\
&\approx 8.696 \times 10^{-5} \leq 5 \times 10^{-4}
\end{aligned}
$$

Ejercicio 3:

Crear un *programa* en **Python** para calcular la suma y el producto de dos números usando los redondeos a k cifras significativas.

Programa de Google Colab

Este programa se ha desarrollado usando Google Colab. Para ver el programa haz clic en el logo.