Clase No. 1 (Segunda parte):

Errores por la representación de los números

MAT–251

Fuentes de error

Todo proceso computacional que trata con la solución de problema matemático involucra errores:

- Errores de modelación del problema.
- Errores al usar aproximaciones matemáticas.
- Errores debidos a la representación de los números en la computadora y la aritmética.
- Errores de programación.

Fuentes de error

Todo proceso computacional que trata con la solución de problema matemático involucra errores:

- Errores de modelación del problema.
- Errores al usar aproximaciones matemáticas.
- Errores debidos a la representación de los números en la computadora y la aritmética.
- Errores de programación.

En este momento, analizamos el penúltimo punto de la lista.

Error en la representación numérica (I)

El número de punto flotante $fl(x)$ que representa a un número real x es de la forma

$$fl(x) = \pm(0.d_1 d_2 \cdots d_p)_\beta \times \beta^e = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_p}{\beta^p}\right) \times \beta^e$$

donde $d_i \in \{0, 1, \ldots, \beta - 1\}$. A $d_1 d_2 \ldots d_p$ se llama la *fracción* y p determina la precisión.

Decimos que la representación es *normalizada* si $d_1 \neq 0$. En caso contrario, se dice que es *subnormal*.

Tenemos que si el exponente $e \in [e_{\min}, e_{\max}]$, entonces el rango de valores para números con representación normalizada es

$$0.100\ldots0 \times \beta^{e_{\min}} \leq |x| \leq 0.\{\beta-1\}\{\beta-1\}\ldots\{\beta-1\} \times \beta^{e_{\max}}.$$

Esto es, $\quad \beta^{e_{\min}-1} \leq |x| \leq \left(1 - \dfrac{1}{\beta^p}\right)\beta^{e_{\max}}.$

Error en la representación numérica (II)

Ejemplo: Si ϵ es el épsilon de la máquina, entonces

$$\begin{aligned} fl(1.0) &= (0.10\ldots00)_\beta \times \beta^1 \\ fl(1.0+\epsilon) &= (0.10\ldots01)_\beta \times \beta^1 \end{aligned}$$

La distancia entre el número máquina $fl(1.0)$ y su consecutivo es

$$\epsilon = fl(1.0+\epsilon) - fl(1.0) = (0.00\ldots01)_\beta \times \beta = \frac{1}{\beta^p}\,\beta = \beta^{1-p}.$$

Por otra parte, el número máquina anterior a $fl(1.0)$ es

$$(0.(\beta-1)(\beta-1)\ldots(\beta-1))_\beta \times \beta^0.$$

La distancia entre ellos es

$$1 - \left(\frac{\beta-1}{\beta} + \frac{\beta-1}{\beta^2} + \cdots + \frac{\beta-1}{\beta^p}\right) = 1 - (\beta-1)\frac{\beta^p-1}{\beta^p(\beta-1)} = \beta^{-p} = \beta^{1-p}\beta^{-1} = \frac{\epsilon}{\beta}.$$

Error en la representación numérica (III)

Para el caso general, si tenemos que

$$x = (0.d_1 d_2 \ldots d_p d_{p+1} \ldots)_\beta \times \beta^e,$$

los números de máquina más cercanos a x son

$$\begin{aligned} x_- &= (0.d_1 d_2 \ldots d_p)_\beta \times \beta^e, \\ x_+ &= [(0.d_1 d_2 \ldots d_p)_\beta + \beta^{-p}] \times \beta^e. \end{aligned}$$

Si x_- es el más cercano, el error absoluto es

$$|x - x_-| = (0.0 \ldots 0 d_{p+1} \ldots)_\beta \times \beta^e = (0.d_{p+1} \ldots)_\beta \times \beta^{e-p} \leq \beta^{e-p}.$$

Entonces el error relativo es

$$\left| \frac{x - x_-}{x} \right| \leq \frac{\beta^{e-p}}{(0.d_1 \ldots d_{p+1} \ldots)_\beta \times \beta^e} = \frac{\beta^{-p}}{(0.d_1 \ldots d_{p+1} \ldots)_\beta} \leq \frac{\beta^{-p}}{\frac{1}{\beta}} = \beta^{1-p} = \epsilon.$$

Error en la representación numérica (IV)

puesto que d_1 debe ser 1 y $(0.d_1 \ldots d_{p+1} \ldots)_\beta \geq (0.1)_\beta = \frac{1}{\beta}$.

Por otra parte, si x_+ es el más cercano, entonces

$$|x - x_+| \leq \frac{1}{2}|x_+ - x_-| = \frac{1}{2}\beta^{-p}\beta^{e},$$

por lo que

$$\frac{|x - x_+|}{|x|} \leq \frac{1}{2}\frac{\beta^{-p}\beta^{e}}{\frac{1}{\beta}\beta^{e}} = \frac{\beta^{1-p}}{2} = \frac{\epsilon}{2}$$

Definimos la *unidad de error de redondeo* u como

$$u = \begin{cases} \epsilon & \text{para redondeo hacia abajo} \\ \frac{\epsilon}{2} & \text{para redondeo hacia arriba} \end{cases}$$

Error de redondeo

La relación entre un número real y el número de máquina que lo representa está dada por $fl(x) = x(1+\delta)$, donde $|\delta| < u$.

Operaciones con números de punto flotante (I)

Dados dos números máquina a y b, en el modelo estándar de aritmética de punto flotante se tiene que

$$fl(a \circ b) = (a \circ b)(1 + \delta)$$

donde $\circ$ es uno de los operadores $\{+, -, \times, /\}$, y $|\delta| < u$.

Con este modelo podemos ver que

$$fl(a + b) = fl(b + a),$$

pero si queremos calcular la suma $a + b + c$, entonces

$$fl(fl(a + b) + c) \neq fl(a + fl(b + c))$$