

Métodos Numéricos

Propagación de errores y
Punto Flotante - S 02

Hermes Pantoja Carhuavilca

(hpantoja@utec.edu.pe)

Rósulo Pérez Cupe

(rperezc@utec.edu.pe)

Jimmy Mendoza Montalvo

(jmendozam@utec.edu.pe)

Máximo Obregón Ramos

(mobregon@utec.edu.pe)

Daniel Camarena Perez

(vcamarena@utec.edu.pe)



► Reinventa el mundo ◀



Profesores: Utec-Ciencias

Índice

- 1 Propagación de Errores
- 2 Punto Flotante



1 PROPAGACIÓN DE ERRORES



Logros de Aprendizaje

- Identifica cómo los errores en los valores numéricos pueden propagarse a través de operaciones y funciones matemáticas.

Propagación de errores

Al resolver un problema utilizando métodos numéricos, el error que se genera será consecuencia de un cúmulo de errores ocurridos en pasos sucesivos, se debe estudiar la mecánica de "propagación" de los mismos a lo largo del cálculo.

Propagación de errores

Propagación de errores en sumas y diferencias

Si tenemos como datos iniciales:

- $x \pm \xi_x$.
- $y \pm \xi_y$.

Sea su suma $q = x + y$ y su diferencia $q = x - y$

¿Cuál es la incertidumbre ξ_q ?

Propiedad

El error absoluto de la suma y de la diferencia de dos o mas magnitudes es aproximadamente igual a la suma de los errores absolutos de dichas magnitudes

$$q = x \pm y \Rightarrow \xi_q \approx \xi_x + \xi_y$$

Ejemplo 1

En un experimento se introducen dos líquidos en dos matraces y se quiere hallar la masa total del líquido. Se conocen:

- $M_1 = \text{Masa del matraz 1} + \text{contenido} = 540 \pm 10\text{g}$.
- $m_1 = \text{Masa del matraz 1} = 72 \pm 1\text{g}$.
- $M_2 = \text{Masa del matraz 2} + \text{contenido} = 940 \pm 20\text{g}$.
- $m_2 = \text{Masa del matraz 2} = 97 \pm 1\text{g}$.

La masa de líquido será:

$$M = \underbrace{(M_1 - m_1)}_{\text{líquido 1}} + \underbrace{(M_2 - m_2)}_{\text{líquido 2}} = 1311\text{g}$$

Su error:

$$\xi_M = \xi_{M_1} + \xi_{m_1} + \xi_{M_2} + \xi_{m_2} = 32\text{g}$$

El resultado se expresará:

$$M = 1311 \pm 32\text{g}$$

Propagación de errores

Funciones de una sola variable

Sea la función $f(x)$ dependiente de una sola variable independiente x . Considere que \tilde{x} es una aproximación de x ; es decir,

$$x = \tilde{x} \pm \Delta\tilde{x}.$$

Se desea estimar

$$\Delta f(\tilde{x}) = |f(x) - f(\tilde{x})|$$

$f(x)$ se desconoce, dado que x se desconoce. De acuerdo a la serie de Taylor:

$$f(x) = f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \frac{f''(\tilde{x})}{2}(x - \tilde{x})^2 + \dots$$

Aproximando:

$$f(x) - f(\tilde{x}) \approx f'(\tilde{x})(x - \tilde{x})$$

$$\Delta f(\tilde{x}) \approx |f'(\tilde{x})|\Delta\tilde{x}$$

Continuación...

Aquí $\Delta f(\tilde{x}) = |f(x) - f(\tilde{x})|$ puede representar una estimación del error de la función y $\Delta \tilde{x} = |x - \tilde{x}|$ puede representar la incertidumbre al medir x .

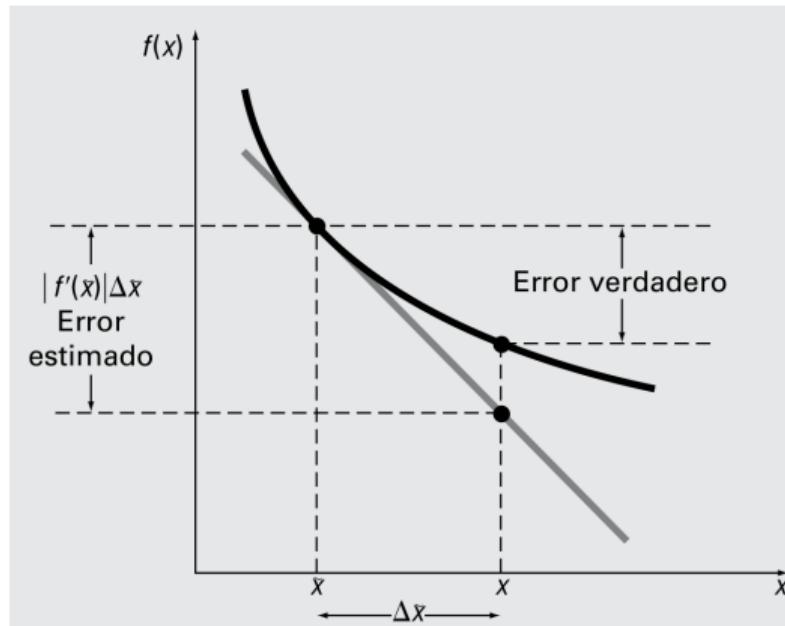


Figure: Representación gráfica de la propagación del error de primer orden.

Ejemplo 2

Ejemplo

Dado un valor de $\tilde{x} = 2.5$ con un error $\Delta\tilde{x} = 0.01$, estime el error resultante en la función $f(x) = x^3$.

Solución:

$$\Delta f(\tilde{x}) \approx 3 \times (2.5)^2 \times (0.01) = 0.1875$$

Dado que $f(2.5) = 15.625$, se pronostica que

$$f(2.5) = 15.625 \pm 0.1875$$

o que el valor verdadero se encuentra entre 15.4375 y 15.8125

Ejercicio 1 (EP 2023 - 1)

Verdadero o Falso: Se desea calcular el área de un círculo y su incertidumbre absoluta, considerando que el radio del círculo es $r = (10 \pm 1)$ cm. Sabemos que el área de un círculo se calcula como $A = \pi r^2$. Bajo estas condiciones, el área del círculo es 100π cm² y la incertidumbre absoluta asociada es 20π cm². ¿Es esto correcto?

Propagación de errores

Funciones de más de una variable

El enfoque anterior puede generalizarse a funciones que sean dependientes de más de una variable independiente.

Sean las medidas \tilde{x}_1 y \tilde{x}_2 con errores $\Delta\tilde{x}_1$ y $\Delta\tilde{x}_2$ usadas para calcular:

$$q = f(x_1, x_2)$$

Mediante una versión de la serie de Taylor para el caso de varias variables:

$$f(x_1, x_2) \approx f(\tilde{x}_1, \tilde{x}_2) + \frac{\partial f}{\partial x_1} \cdot (x_1 - \tilde{x}_1) + \frac{\partial f}{\partial x_2} \cdot (x_2 - \tilde{x}_2)$$

con lo que:

$$\xi_q = \Delta f(\tilde{x}_1, \tilde{x}_2) \approx \left| \frac{\partial f}{\partial x_1} \right| \cdot \Delta \tilde{x}_1 + \left| \frac{\partial f}{\partial x_2} \right| \cdot \Delta \tilde{x}_2$$

Propagación del error en producto

Si tenemos dos magnitudes:

- $x = \tilde{x} \pm \xi_x$.
- $y = \tilde{y} \pm \xi_y$.

Sea $q = xy$, podremos estimar ξ_q ?

Definimos la función $q = g(x, y) = xy$ y usando la serie de Taylor:

$$\Delta g(\tilde{x}, \tilde{y}) \approx \left| \frac{\partial g}{\partial x} \right| \Delta \tilde{x} + \left| \frac{\partial g}{\partial y} \right| \Delta \tilde{y}$$

Entonces: $\frac{\partial g}{\partial x} = y$ $\frac{\partial g}{\partial y} = x$.

$$\Delta g(\tilde{x}, \tilde{y}) \approx |\tilde{y}| \Delta \tilde{x} + |\tilde{x}| \Delta \tilde{y} \rightarrow \frac{\Delta g(\tilde{x}, \tilde{y})}{|\tilde{x}\tilde{y}|} \approx \frac{\Delta \tilde{x}}{|\tilde{x}|} + \frac{\Delta \tilde{y}}{|\tilde{y}|}$$

El error relativo del producto (cociente) es aproximadamente igual a la suma de errores relativos.

Ejemplo 3

Ejemplo

La deflexión y de la punta de un mástil en un bote de vela es

$$y = \frac{FL^4}{8EI}$$

donde F = una carga lateral uniforme (N/m), L = altura (m), E = el módulo de elasticidad (N/m^2) e I = el momento de inercia (m^4). Estime el error en y con los siguientes datos:

$$\tilde{F} = 750 \text{ N/m}$$

$$\Delta \tilde{F} = 30 \text{ N/m}$$

$$\tilde{L} = 9 \text{ m}$$

$$\Delta \tilde{L} = 0.03 \text{ m}$$

$$\tilde{E} = 7.5 \times 10^9 \text{ N/m}^2$$

$$\Delta \tilde{E} = 5 \times 10^7 \text{ N/m}^2$$

$$\tilde{I} = 0.0005 \text{ m}^4$$

$$\Delta \tilde{I} = 0.000005 \text{ m}^4$$

Solución

$$\Delta y(\tilde{F}, \tilde{L}, \tilde{E}, \tilde{I}) \approx \left| \frac{\partial y}{\partial F} \right| \Delta \tilde{F} + \left| \frac{\partial y}{\partial L} \right| \Delta \tilde{L} + \left| \frac{\partial y}{\partial E} \right| \Delta \tilde{E} + \left| \frac{\partial y}{\partial I} \right| \Delta \tilde{I}$$

$$\Delta y(\tilde{F}, \tilde{L}, \tilde{E}, \tilde{I}) \approx \frac{\tilde{L}^4}{8\tilde{E}\tilde{I}} \Delta \tilde{F} + \frac{\tilde{F}\tilde{L}^3}{2\tilde{E}\tilde{I}} \Delta \tilde{L} + \frac{\tilde{F}\tilde{L}^4}{8\tilde{E}^2\tilde{I}} \Delta \tilde{E} + \frac{\tilde{F}\tilde{L}^4}{8\tilde{E}\tilde{I}^2} \Delta \tilde{I}$$

Sustituyendo los valores adecuados se obtiene

$$\Delta y \approx 0.006561 + 0.002187 + 0.001094 + 0.00164 = 0.011482$$

Por consiguiente, $y = 0.164025 \pm 0.011482$ indica que y está entre 0.152543 m y 0.175507 m.

Ejercicio 2 (EP 2023-2)

La ubicación del centro de presiones (y_p) en una placa plana rectangular de base b y altura h sobre la cual actúa la fuerza hidrostática resultante se puede determinar mediante la expresión siguiente:

$$y_p = y_c + \frac{I_{xx,c}}{y_c A}$$

Donde, para la placa rectangular $I_{xx,c} = \frac{1}{2}bh^3$ y $A = bh$. Si se conoce de manera exacta que la base $b = 1\text{ m}$, pero los valores de $y_c = 5\text{ m}$ (coordenada y del centroide) y la altura $h = 2\text{ m}$ presentan una incertidumbre de $\pm 0.1\text{ m}$ y $\pm 0.05\text{ m}$, respectivamente, calcule el valor del error relativo porcentual en la determinación de y_p .

Ejercicio 3 (EP 2024-1)

Sean las magnitudes

$$\begin{cases} x = \frac{1}{3} \pm 0.01 \\ y = 1 - |2x - 1| \\ z = \operatorname{sen}^2\left(\frac{\pi}{2}y\right) \end{cases}$$

- a) Estime la incertidumbre de y
- b) Estime la incertidumbre de z



2 PUNTO FLOTANTE



Logros de Aprendizaje

- Representa un número en el Estandar IEEE 754 con precisión simple y doble.

Problema

Para resolver problemas numéricos con el computador, debemos trabajar con una cantidad de espacio finita y una precisión limitada.

- Los sistemas de precisión arbitraria todavía son costosos desde el punto de vista del tiempo de cómputo y desde el punto de vista del espacio requerido para su funcionamiento.
- Muchas veces hay una incertidumbre que no se puede eliminar en los valores de interés.
- Los modelos que se utilizan tienen una incertidumbre intrínseca por lo que no vale la pena mejorar la precisión más allá de cierto punto.

Es un hecho que debemos acostumbrarnos a vivir con los errores de representación, por lo que se hace necesario entender como ocurren.

Punto flotante

La representación de punto flotante es simplemente un tipo de notación científica que reconoce las limitaciones de espacio inherentes al computador.

El objetivo de la representación de punto flotante es ofrecer un sistema que sirva para las necesidades de cómputo modernas sin requerir que las operaciones se efectuen con precisión infinita.

IEEE-754

Los científicos que desarrollaron el cohete Ariane 5 vuelo 501 reutilizaron parte del código de su predecesor, el Ariane 4, pero los motores del cohete nuevo incorporaron también, sin que nadie se diera cuenta, un "bug" en una rutina aritmética en la computadora de vuelo. Esto provocó, el 4 de junio de 1996, que la computadora fallara segundos después del despegue del cohete; 0.5 segundos más tarde falló el ordenador principal de la misión. El

Ariane 5 se desintegró 40 segundos después del lanzamiento.



IEEE-754

Los errores en un sistema de punto flotante (*floating point* en inglés) pueden ser catastróficos:

- Vuelo inaugural del Ariane 5, 1996. ([Video 4'26"](#), [Reporte](#))
- Batería anti-misiles *Patriot*, 1991. ([Nota](#), [Artículo](#))
- Error de division del Pentium, 1994. ([Wikipedia](#))
- Elecciones parlamentarias en Alemania, 1992. ([Nota](#))

El [estándar IEEE-754-2008](#) reglamenta una cantidad de aspectos relevantes.
Nosotros vamos a concentrarnos solo en algunos aspectos.

Aritmética de Punto Flotante

Todo número real $x \neq 0$ puede escribirse

$$x = \pm (1.d_1 d_2 \dots d_k \dots)_2 \times 2^{E_x}$$

Un sistema de punto flotante se especifica por la base 2, el largo de la mantisa p , y límites para los exponentes de L y U .

El punto flotante de $x \neq 0$, en base 2 tiene la forma:

$$fl(x) = \pm (1.d_1 d_2 \dots d_p)_2 \times 2^E$$

donde $d_1 d_2 \dots d_p$ es la mantisa, el 1er bit de la mantisa es implícito, y E es el exponente entero de la punto flotante.

Se define el sistema de punto flotante F con cuatro elementos

$$F(2, p, L, U) = F(base, precision, expomin, expomax)$$

$$L \leq E \leq U$$



Ejemplo 1

Dado el sistema de punto flotante $F(2, 4, -6, 7)$,

- Hallar la forma del sistema de punto flotante.
- Hallar el número de representaciones.
- El más pequeño en positivos.
- El más grande.
- ¿ $0.5 \in F$? , ¿ $\frac{3}{4} \in F$? , ¿ $0.5 + \frac{3}{4} \in F$?

Estándar IEEE-754

Precisión Simple: 32 bits

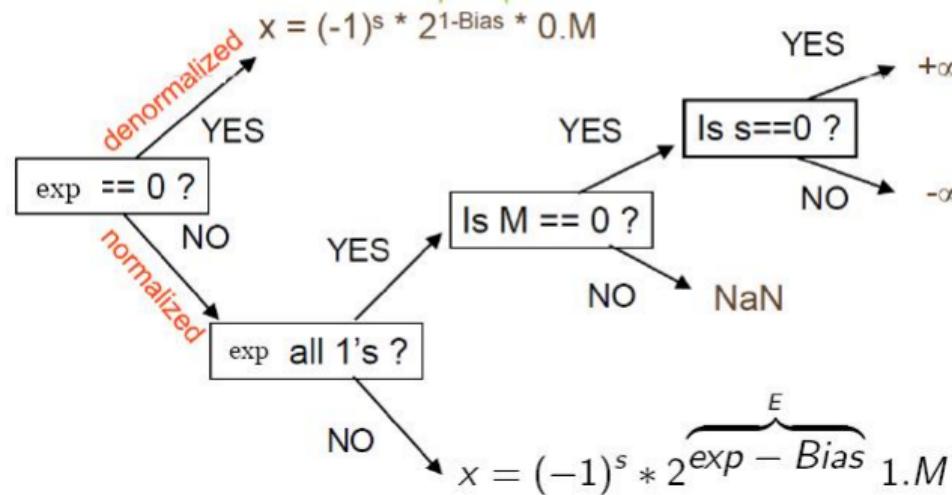
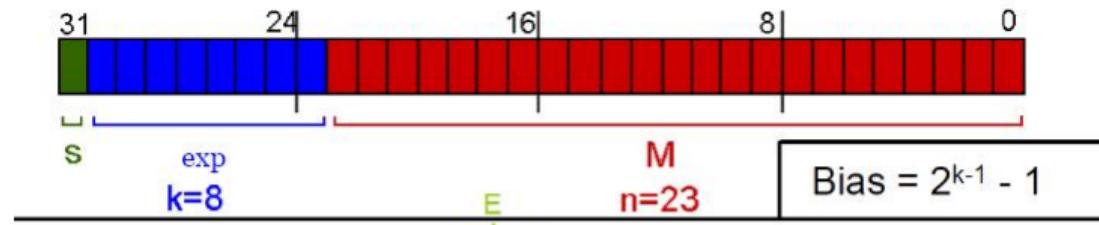


Precisión Doble: 64 bits



IEEE-754

La representación de punto flotante de Precisión Simple



Formato IEEE 754 en Precisión Simple (32 bits)

- Un número en punto flotante se representa con **32 bits**:
 - **Bit de signo (b_s)**: 1 bit — 0 para positivo, 1 para negativo.
 - **Exponente interno (E_{int})**: 8 bits con **sesgo (bias)**.
 - **Fracción o mantisa (f) en algunos casos denotada como M** : 23 bits de la parte fraccionaria de la mantisa.
- Cálculo del sesgo:

$$\text{bias} = 2^{8-1} - 1 = 127$$

- Exponente externo (real):

$$E_{\text{ext}} = E_{\text{int}} - 127$$

Relación entre exponentes y forma general

- Relación entre exponente interno y externo (números normalizados):

$$1 \leq E_{\text{int}} \leq 254 \quad \Rightarrow \quad -126 \leq E_{\text{ext}} \leq 127$$

- Forma general del número normalizado:

$$x = (-1)^{b_s} \times 2^{E_{\text{ext}}} \times (1 + f)$$

- En números normalizados, se asume un **1 implícito** antes del punto binario en la mantisa.

Casos especiales en IEEE 754

- **Subnormal (denormalizado):** $E_{\text{int}} = 0$ y $f \neq 0$

$$x = (-1)^{b_s} \times 2^{-126} \times f$$

(No se asume el "1" implícito)

- **Cero:** $E_{\text{int}} = 0$, $f = 0$ representa $+0$ o -0
- **Infinito:** $E_{\text{int}} = 255$, $f = 0$ representa $+\infty$ o $-\infty$
- **NaN (Not a Number):** $E_{\text{int}} = 255$, $f \neq 0$

Rango de Precisión Simple

- Los exponentes 00000000 y 11111111 son reservados.
- El valor más pequeño

- Exponente exp: $00000001_2 = 1$
- Exponente E: $E = \text{exp} - \text{Bias} = 1 - 127 = -126$
- Parte fraccionaria: 000...00

$$x_{min} = 1.0 \times 2^{-126}$$

- El valor más grande
- Exponente exp: $11111110_2 = 254$
- Exponente E: $E = \text{exp} - \text{Bias} = 254 - 127 = +127$
- Parte fraccionaria: 111...11

$$x_{max} = 1.111\dots11 \times 2^{127}$$

Ejemplo 2

Ejemplo

Representa -3.75 en representación de precisión simple.

Solución:

Normalizamos de acuerdo a las normas de la IEEE 754.

- $-3.75 = -11.11_2 = -1.111 \times 2^1$
- $Bias = 127$, por lo que $127 + 1 = 128$ (es el exponente actual).
- El primer 1 en la mantisa es implícito, por lo que tenemos:

1 10000000 11100000000000000000000000
- Desde que tenemos implícito 1 en el significando, esto es igual a :
 $-1.111_2 \times 2^{(128-127)} = -1.111_2 \times 2^1 = -11.11_2 = -3.75$

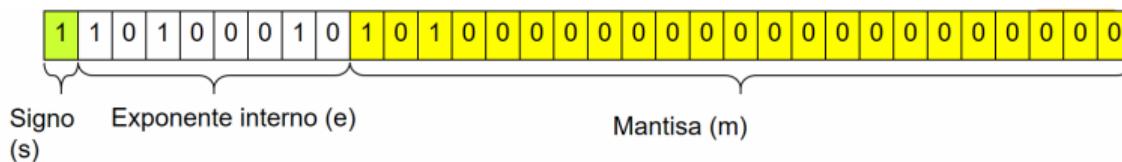
Ejemplo 3

Ejemplo

Identifique el números de puntos flotante correspondiente a la siguiente cadenas de bits

1 10100010 10100000000000000000000000000000

Solución:



$$\begin{aligned} \text{Valor} &= (-1)^s \times (1.m)_2 \times 2^{e-127} = (-1)^1 \times (1 + 0.10100000_2) \times 2^{10100010_2 - 127} \\ &= (-1) \times (1 + 0.625) \times 2^{162 - 127} \\ &= (-1) \times 1.625 \times 2^{35} = -5.5834 \times 10^{10} \end{aligned}$$

El epsilón de máquina

El Epsilón de máquina, ϵ , es definido como el número más pequeño tal que

$$1 + \epsilon > 1.$$

Es el número que mide la precisión de la máquina y por lo tanto el responsable del error de redondeo. El número de máquina ϵ se define como la diferencia entre 1 y el sucesor de 1.

Ejemplo 4

Dada la siguiente representación en punto flotante de un computador hipotético:

$$x = \pm(1.d_1d_2d_3)_2 \times 2^E$$

Consideré que este sistema tiene un rango de exponentes limitado de $E_{\min} = 0$ a $E_{\max} = 1$. ¿Determine cuantos bits tiene este sistema hipotético?. Explique cómo se determina este valor.

Solución:

- Sabemos que se cumple la desigualdad:

$$-\text{Bias} + 1 \leq E \leq \text{Bias} \rightarrow -(2^{k-1} - 1) + 1 \leq E \leq 2^{k-1} - 1$$

donde k es el número de bits del exponente interno.

- Se tiene:

$$\underbrace{-(2^{k-1} - 1) + 1}_{E_{\min}=0} \leq E \leq \underbrace{2^{k-1} - 1}_{E_{\max}=1}$$

- Así $2^{k-1} - 1 = 1 \rightarrow 2^{k-1} = 2 \rightarrow k - 1 = 1 \rightarrow k = 2$

Ejercicio 1 (EP 2024-1)

Dada la siguiente representación en punto flotante de un computador hipotético:

$$x = \pm(1.d_1d_2d_3d_4)_2 \times 2^E$$

Considere que este sistema tiene un rango de exponentes limitado de $E_{\min} = -2$ a $E_{\max} = 3$. ¿Cuál sería el menor número positivo diferente de cero que puede representarse en este sistema? .¿Determine cuantos bits tiene este sistema hipotético?. Explique cómo se determina este valor.

Ejercicio 2 (EP 2024-2)

En una prueba respecto a la precisión de resultados, se necesita representar al número $\frac{-4328}{7}$ en la notación coma flotante IEEE-754 con precisión simple, detalle paso a paso el proceso y luego halle la secuencia de bits.

Conclusiones

- La serie de Taylor provee una estimación de la propagación de incertidumbres de ciertas magnitudes al realizar un proceso que las involucre.
- Un computador realiza cálculos con precisión finita (inexacta) lo que requiere un sistema de representación numérica (IEEE 754).

Bibliografía

-  **Steven C. Chapra and Raymond P. Canale**
Métodos numéricos para ingenieros, 7a ed.
-  **Richard L. Burden and J. Douglas Faires**
Análisis numérico, 7a ed.

**Gracias por su
atención**

