Errores numéricos

- Antes de continuar se pide que escriba un programita muy sencillo:
  1) guardar en una variable 1 dividido por 3
  2) restar de 1 esa variable, y repetir otras dos veces
  3) si el resultado es cero imprimir “verdadero”, y sino “falso”
- Sorprendentemente el programa produce un resultado erróneo. Esto nos pone frente a los *errores numéricos*. Esto puede acarrear resultados catastróficos (al final de la clase se muestran algunos casos)
- En cálculos numéricos debemos estar preparados para manejar estos errores numéricos.

Errores numéricos

- Los errores numéricos provienen de la diferencia entre los valores calculados y los valores reales
- La palabra *error* no significa aquí *equivocación*
- En todos los cálculos con computadoras y en todos los métodos numéricos es necesario manejar el tema de *errores*

Fuentes de errores

En aproximaciones en ingeniería se introducen errores (entendidos aqui como diferencia entre el valor calculado y el valor real) de diferentes fuentes:

- Errores en la formulación del problema matemático
- Errores en la aproximación de la geometría
- Errores en los datos físicos
- Errores del método numérico de aproximación
- Errores en la resolución del sistema de ecuaciones
- Errores de redondeo

Errores numéricos

- En los métodos numéricos manejaremos dos clases de errores:
  - Errores de truncamiento
    También llamados *errores algorítmicos*.
    Son los provenientes del método numérico de aproximación. Por ejemplo: el área calculada con los rectángulos, en el ejemplo dado, no es igual al área encerrada por la curva en la integración definida. Se los suele llamar *errores de truncamiento*, ya que en muchos casos provienen de truncar una serie infinita.
  - Errores de redondeo
    Son los introducidos por la computadora digital que tiene un espacio finito para almacenar sus variables.
    En este caso entran tanto los denominados errores por truncamiento (hacia abajo) como por redondeo (hacia arriba). En ambos casos se engloban en el término *errores de redondeo*
- A continuación veremos errores de redondeo.
- Los errores de truncamiento serán tratados en los capítulos siguientes.

Aritmética de las computadoras

- Normalmente usamos una base **decimal** para escribir los números.
  Ej:
  $$1563 = 1 \times 10^3 + 5 \times 10^2 + 6 \times 10^1 + 3 \times 10^0$$
- Un número *natural* se puede representar:

  $$N = a_k \times 10^k + a_{k-1} \times 10^{k-1} + \ldots a_1 \times 10^1 + a_0 \times 10^0$$

  con $a_i \in \{0, 1, 2, \ldots 9\}$ y se escribe:

  $$N = a_k a_{k-1} \ldots a_1 a_0 = (a_k a_{k-1} \ldots a_1 a_0)_{(10)}$$

Aritmética de las computadoras

- Un número *real* se puede representar:

$$X = s\ a_k a_{k-1} \ldots a_1 a_0 . b_1 b_2 \ldots_{(10)}$$

$$X = s \left(\sum_{i=0}^{k} a_i 10^i + \sum_{i=1}^{\infty} b_i 10^{-i} \right)_{(10)}$$

donde $s = \begin{cases} + \\ - \end{cases}$

- Def:
  Dígitos significativos: cantidad de dígitos a_i y b_i.
  (Empezando por el primero no nulo)

Aritmética de las computadoras

- Def:
  Representación normalizada:

$$X = s\,(0.a_1a_2\ldots)_{(10)} \times 10^m$$

donde

$$X = \underbrace{s}_{signo}\ \underbrace{(0.a_1a_2\ldots)_{(10)}}_{mantisa} \times \underbrace{10}_{base}\ \underbrace{m}_{exp}$$

Ej:

$$1563 = 0.1563 \times 10^4$$

$$2100000 = 0.21 \times 10^7$$

$$0.00017 = 0.17 \times 10{-}3$$

La mantisa es ≥ 0.1 y < 1

Aritmética de las computadoras

- Base binaria

$$1563 = 1 \times 2^{10} + 1 \times 2^9 + 0 \times 2^8 + 0 \times 2^7 + 0 \times 2^6 + 0 \times 2^5$$

$$+1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$

Los a_i son ahora 0 o 1. La base es 2.

$$1563 = 1024 + 512 + 16 + 8 + 2 + 1$$

Como se hizo en base decimal, en base binaria se puede representar:

$$1563_{(10)} = 11000011011_{(2)}$$

Aritmética de las computadoras

- Representación binaria normalizada
  Ventajas:
  - un número $\rightarrow$ string de bits (no preciso el punto)
  - el dígito a_1 no preciso guardarlo (es siempre 1)
- Se denomina *floating point*
- Se destina una cantidad de bits para la mantisa (que determina la precisión); y otra cantidad para el exponente (que determina el rango que puede representar)

Aritmética de las computadoras

Representación estándar IEEE

- Precisión simple (32 bits)
  - 1 bit para el signo; 8 bits p/exponente; y 23 p/ mantisa
  - El rango que puede representar va de 2^{-126} a 2^{127} o sea 1.17×10^{-38} a 1.7×10^{38}. Por debajo o encima da *underflow* o *overflow*.
  - El error es $2^{-24} \simeq 6 \times 10^{-8}$
- Precisión doble (64 bits)
  - 1 bit para el signo; 11 bits p/exponente; y 52 p/ mantisa
  - El rango que puede representar va de 2^{-1022} a 2^{1023} o sea 10^{-308} a 10^{308} aprox.
  - El error es $2^{-53} \simeq 10^{-16}$

Errores de redondeo

- Representación binaria normalizada

$$x = s\,(0.a_1a_2\ldots a_ka_{k+1}\ldots)_{(2)} \times 2^m \;=\; \pm q \times 2^m$$

  donde la mantisa q está entre $\frac{1}{2} \leq q < 1$
- El *número de máquina* más cercano, por defecto:

$$x' = s\,(0.a_1a_2\ldots a_k)_{(2)} \times 2^m$$

  se retienen $k-1$ bits ($a_1 = 1$) y descarta el resto
  Se obtiene un número por defecto. $\rightarrow$ *truncamiento* o *poda* o *cancelación*
- El número por exceso:

$$x'' = s\,\left((0.a_1a_2\ldots a_k)_{(2)} + 2^{-k}\right) \times 2^m$$

Errores de redondeo

- El número de máquina más próximo a x lo llamamos $fl(x)$
- En el *redondeo*
  - si el bit $a_{k+1} = 0 \rightarrow fl(x) = x'$
  - si el bit $a_{k+1} = 1 \rightarrow fl(x) = x''$
- El error al usar $fl(x)$ en vez de x se llama *error de redondeo* (ya sea que esté podado o redondeado).
- Error absoluto

$$|x - fl(x)|$$

- Error relativo

$$\frac{|x - fl(x)|}{|x|}$$

siempre que $|x| \neq 0$

Errores de redondeo

Epsilon de máquina ϵ

- La diferencia entre las representaciones por exceso y por defecto:

$$x'' - x' = 2^{m-k}$$

- La representación en punto flotante es de ellas la más cercana a x, de donde, el error absoluto:

$$|x - fl(x)| \leq \frac{1}{2}(x'' - x') = \frac{1}{2}2^{m-k} = 2^{m-k-1}$$

- Error relativo

$$\frac{|x - fl(x)|}{|x|} \leq \frac{2^{m-k-1}}{q\ 2^m} = \frac{2^{-k-1}}{q} \leq \frac{2^{-k-1}}{\frac{1}{2}} = 2^{-k}$$

donde q es la mantisa cuyo valor esta entre $\frac{1}{2}$ y 1.
(recordar que $(0.1)_2 = (0.5)_{10}$)

Errores de redondeo

- Puede ponerse:

$$fl(x) = x(1+\delta)$$

  con $|\delta| \leq \epsilon$
- Donde $\epsilon = 2^{-k}$
- El ϵ de máquina varía según la computadora.
- Para computadoras con palabras de 32 bits
  - simple precisión: $\epsilon \simeq 10^{-7}$
  - doble precisión: $\epsilon \simeq 10^{-15}$
- Para computadoras con palabras de 64 bits
  - simple precisión: $\epsilon \simeq 10^{-14}$
  - doble precisión: $\epsilon \simeq 10^{-28}$
- Luego

$$\frac{|x - fl(x)|}{|x|} \leq \epsilon$$

Errores de redondeo

- Cifras significativas
  Se dice que x^* aproxima a x con t cifras (o dígitos) significativas, si t es el entero, no negativo, más grande para el cual

$$\frac{|x - fl(x)|}{|x|} \leq 0.5 \times 10^{-t}$$

Ejemplos:

- Sea la aproximación para $\pi = 3.141592$. El número 3.1416 aproxima al anterior con 5 cifras significativas, ya que $\frac{|x-fl(x)|}{|x|} \simeq 2.5 \times 10^{-6} \leq 0.5 \times 10^{-5}$
- Una medición con error relativo $1‰$ contiene 2 cifras significativas, ya que $0.001 \leq 0.5 \times 10^{-2}$

Propagación de errores

- Los errores de redondeo, por la aritmética de la máquina, se pueden propagar con las operaciones, acumulándose.

Propagación de errores

En operación suma o resta

- Sea la suma de dos número x e y, cuyas representaciones en punto flotante son $fl(x)$ y $fl(y)$, y los errores de redondeo η_x y η_y:

$$x = fl(x) + \eta_x$$

$$y = fl(y) + \eta_y$$

- El error absoluto de la suma:

$$|(x+y) - (fl(x) + fl(y))| = |\eta_x| + |\eta_y|$$

- El error absoluto de la suma (o resta), es la suma de los errores absolutos de cada sumando.

Propagación de errores

En operación producto

- Sea el producto de dos número x e y, cuyas representaciones en punto flotante son $fl(x)$ y $fl(y)$, y los errores de redondeo η_x y η_y:
- El error absoluto del producto:

$$|(xy) - (fl(x)fl(y))| = |(xy) - (x - \eta_x)(y - \eta_y)| = |x\eta_y + y\eta_x - \eta_x\eta_y|$$

- El error relativo (despreciando el término de orden superior):

$$\frac{|(xy) - (fl(x)fl(y))|}{|xy|} = \frac{\eta_x}{x} + \frac{\eta_y}{y}$$

- El error relativo del producto, es la suma de los errores relativos de cada factor.

Sustracción de números próximos

- Sean x e y números próximos:

$$fl(x) = (0.a_1a_2 \ldots a_pa_{p+1} \ldots a_k)10^n$$
$$fl(y) = (0.a_1a_2 \ldots a_pb_{p+1} \ldots b_k)10^n$$

- Restando:

$$fl(x) - fl(y) = (0.00 \ldots 0c_{p+1} \ldots c_k)10^n$$

$$fl(x) - fl(y) = (0.c_{p+1} \ldots c_k)10^{n-p}$$

donde se designa $c_{p+1} = a_{p+1} - b_{p+1}$, etc.

- El resultado tiene a lo sumo $k - p$ cifras significativas.

Sustracción de números próximos

Ejemplo:

$$x = 0.3721478693$$

$$y = 0.3720230572$$

$$x - y = 0.0001248121$$

En una computadora de 5 dígitos:

$$fl(x) = 0.37215$$

$$fl(y) = 0.37202$$

$$fl(x - y) = 0.00013$$

El error relativo es grande ($\simeq 4\,\%$)
El resultado se expresa 0.13000×10^{-3}, pero los tres últimos dígitos (ceros) carecen de valor.

Evaluación de funciones

- Sea evaluar

$$f(x) = x^3 - 6.1\,x^2 + 3.2\,x + 1.5$$

en $x = 4.71$ usando aritmética de 3 dígitos

- El valor exacto:

$$f(4.71) = 104.487111 - 135.32301 + 15.072 + 1.5 = -14.263899$$

- El valor con tres dígitos (truncado):

$$f(4.71) = 104. - 135. + 15.0 + 1.5 = -13.5$$

- El error relativo:

$$\left|\frac{-14.263899 + 13.5}{-14.263899}\right| \simeq 0.05$$

Evaluación de funciones

- La misma función puede reescribirse en forma anidada:

$$f(x) = x^3 - 6.1\,x^2 + 3.2\,x + 1.5 = ((x - 6.1)\,x + 3.2)\,x + 1.5$$

- Evaluando esta expresión con tres dígitos (truncado):

$$f(4.71) = ((4.71 - 6.1)\,4.71 + 3.2)\,4.71 + 1.5 = -14.2$$

- El error relativo:

$$\left|\frac{-14.263899 + 14.2}{-14.263899}\right| \simeq 0.0045$$

diez veces menor que en el caso anterior.

Evaluación de funciones

- las funciones deberían escribirse siempre en forma anidada antes de evaluarlas numéricamente
- en este caso hemos visto como se ha disminuido el error de truncamiento
- tambien disminuye la cantidad de operaciones: en el primer caso se precisan 4 multiplicaciones y 3 adiciones; en el segundo 2 multiplicaciones y tres adiciones.

Problemas matemáticos y solución numérica