Clase No. 2:

Errores por la representación de punto flotante y propagación de errores

MAT–251

Dr. Alonso Ramírez Manzanares
CIMAT, A.C.
e-mail: alram@cimat.mx
web: http://www.cimat.mx/~alram/met_num/

Dr. Joaquín Peña Acevedo
CIMAT A.C.
e-mail: joaquin@cimat.mx

Errores al realizar operaciones aritméticas (I)

- Hay que especificar el tipo de redondeo que se afectúa.

Al calcular 525000×0.0365 usando tres dígitos de precisión, tenemos que

$$a = 525000 = 0.525 \times 10^6, \qquad b = 0.365 \times 10^{-1}$$

$$fl(ab) = fl(0.191625 \times 10^5) = \begin{cases} 0.192 \times 10^5 & \text{Redondeo hacia arriba} \\ 0.191 \times 10^5 & \text{Redondeo hacia abajo} \end{cases}$$

Errores al realizar operaciones aritméticas (I)

- Hay que especificar el tipo de redondeo que se afectúa.

Al calcular 525000×0.0365 usando tres dígitos de precisión, tenemos que

$$a = 525000 = 0.525 \times 10^6, \qquad b = 0.365 \times 10^{-1}$$

$$fl(ab) = fl(0.191625 \times 10^5) = \begin{cases} 0.192 \times 10^5 & \text{Redondeo hacia arriba} \\ 0.191 \times 10^5 & \text{Redondeo hacia abajo} \end{cases}$$

- La asociatividad en la suma puede no ser válida.

  Ejemplo en base 10 con tres dígitos de precisión y redondeo hacia el más cercano:

$$a = 0.100 \times 10, \quad b = 0.480 \times 10^{-2}, \quad c = 0.450 \times 10^{-2}$$

Errores al realizar operaciones aritméticas (I)

- Hay que especificar el tipo de redondeo que se afectúa.

Al calcular 525000×0.0365 usando tres dígitos de precisión, tenemos que

$$a = 525000 = 0.525 \times 10^6, \qquad b = 0.365 \times 10^{-1}$$

$$fl(ab) = fl(0.191625 \times 10^5) = \begin{cases} 0.192 \times 10^5 & \text{Redondeo hacia arriba} \\ 0.191 \times 10^5 & \text{Redondeo hacia abajo} \end{cases}$$

- La asociatividad en la suma puede no ser válida.

  Ejemplo en base 10 con tres dígitos de precisión y redondeo hacia el más cercano:

$$a = 0.100 \times 10, \quad b = 0.480 \times 10^{-2}, \quad c = 0.450 \times 10^{-2}$$

$$\begin{aligned} fl(fl(a+b)+c) &= fl(0.100 \times 10 + 0.450 \times 10^{-2}) = 0.100 \times 10 = a \\ fl(a+fl(b+c)) &= fl(0.100 \times 10 + 0.930 \times 10^{-2}) = 0.101 \times 10 \end{aligned}$$

Errores al realizar operaciones aritméticas (II)

- Errores por sustracción o error por cancelación.

Ejemplo 1. Sea $f(x) = (1-\cos x)/x^2$. Para $x = 1.2 \times 10^{-5}$ y una precisión a 10 decimales, se tiene que

$$\cos x = 0.9999999999 \implies 1-\cos x = 0.0000000001$$
$$\implies \frac{1-\cos x}{x^2} = \frac{10^{-10}}{1.44 \times 10^{-10}} \approx 0.6944....$$

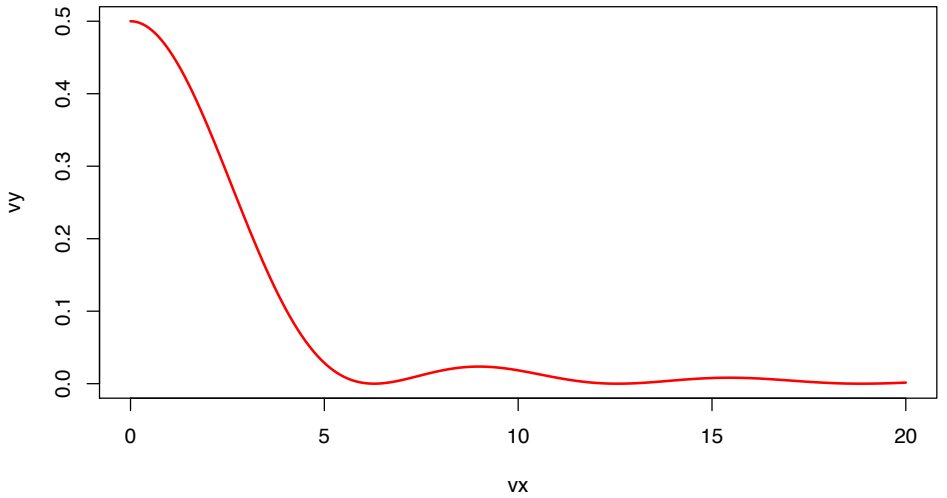
Errores al realizar operaciones aritméticas (II)

- Errores por sustracción o error por cancelación.

Ejemplo 1. Sea $f(x) = (1 - \cos x)/x^2$. Para $x = 1.2 \times 10^{-5}$ y una precisión a 10 decimales, se tiene que

$$\cos x = 0.9999999999 \implies 1 - \cos x = 0.0000000001$$
$$\implies \frac{1 - \cos x}{x^2} = \frac{10^{-10}}{1.44 \times 10^{-10}} \approx 0.6944....$$

El resultado es incorrecto. Resulta que $0 \leq f(x) < 0.5$ para todo $x \neq 0$.



Errores al realizar operaciones aritméticas (III)

Para evitarlo, podemos usar $\cos x = 1 - 2\sin^2(x/2)$.

$$f(x) = \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2.$$

$$x = 1.2 \times 10^{-5} \qquad \Longrightarrow \qquad f(x) = \frac{1}{2}\left(\frac{0.0000060000}{0.0000060000}\right)^2 = 0.5$$

Ejemplo 2. Consideremos la función

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right)$$

Se puede ver que la función es creciente para $x \geq 0$.
Si queremos evaluarla en la computadora cuando x va aumentado de valor, ¿qué resultados obtendremos?

Errores al realizar operaciones aritméticas (IV)

En lugar de evaluar $f(x)$ podemos utilizar

$$g(x) = \frac{x}{\sqrt{x+1} + \sqrt{x}}.$$

y comparar los resultados con $h(x) = \sqrt{x}/2$ usando variables tipo `double`:

x	f(x)	g(x)	h(x)
1.0e+05	1.581134877e+02	1.581134877e+02	1.581138830e+02
1.0e+06	4.999998750e+02	4.999998750e+02	5.000000000e+02
1.0e+07	1.581138790e+03	1.581138791e+03	1.581138830e+03
1.0e+08	5.000000056e+03	4.999999988e+03	5.000000000e+03
1.0e+09	1.581139077e+04	1.581138830e+04	1.581138830e+04
1.0e+10	4.999994417e+04	5.000000000e+04	5.000000000e+04
1.0e+11	1.581152901e+05	1.581138830e+05	1.581138830e+05
1.0e+12	5.000038072e+05	5.000000000e+05	5.000000000e+05
1.0e+13	1.578591764e+06	1.581138830e+06	1.581138830e+06
1.0e+14	5.029141903e+06	5.000000000e+06	5.000000000e+06
1.0e+15	1.862645149e+07	1.581138830e+07	1.581138830e+07
1.0e+16	0.000000000e+00	5.000000000e+07	5.000000000e+07
1.0e+17	0.000000000e+00	1.581138830e+08	1.581138830e+08
1.0e+18	0.000000000e+00	5.000000000e+08	5.000000000e+08
1.0e+19	0.000000000e+00	1.581138830e+09	1.581138830e+09

Errores al realizar operaciones aritméticas (V)

Y si en vez de `double` usamos variables tipo `float` se obtiene:

x	f(x)	g(x)	h(x)
1.0e+00	4.142135382e-01	4.142135680e-01	5.000000000e-01
1.0e+01	1.543471813e+00	1.543471217e+00	1.581138849e+00
1.0e+02	4.987525940e+00	4.987562180e+00	5.000000000e+00
1.0e+03	1.580810547e+01	1.580743790e+01	1.581138802e+01
1.0e+04	4.997253418e+01	4.999875259e+01	5.000000000e+01
1.0e+05	1.586914062e+02	1.581134949e+02	1.581138763e+02
1.0e+06	4.882812500e+02	4.999998779e+02	5.000000000e+02
1.0e+07	2.441406250e+03	1.581138794e+03	1.581138794e+03
1.0e+08	0.000000000e+00	5.000000000e+03	5.000000000e+03
1.0e+09	0.000000000e+00	1.581138770e+04	1.581138867e+04
1.0e+10	0.000000000e+00	5.000000000e+04	5.000000000e+04
1.0e+11	0.000000000e+00	1.581138906e+05	1.581138750e+05
1.0e+12	0.000000000e+00	5.000000000e+05	5.000000000e+05
1.0e+13	0.000000000e+00	1.581138750e+06	1.581138875e+06
1.0e+14	0.000000000e+00	5.000000000e+06	5.000000000e+06
1.0e+15	0.000000000e+00	1.581138800e+07	1.581138800e+07
1.0e+16	0.000000000e+00	5.000000000e+07	5.000000000e+07
1.0e+17	0.000000000e+00	1.581138720e+08	1.581138880e+08

Errores al realizar operaciones aritméticas (VI)

Ejemplo 3. Considere las siguientes expresiones:

$$
\begin{aligned}
s_1 &= 10^{22} + 17 - 10 + 130 - 10^{22} \\
s_2 &= 10^{22} - 10 + 130 - 10^{22} + 17 \\
s_3 &= 10^{22} + 17 - 10^{22} - 10 + 130 \\
s_4 &= 10^{22} - 10 - 10^{22} + 130 + 17 \\
s_5 &= 10^{22} - 10^{22} + 17 - 10 + 130 \\
s_6 &= 10^{22} + 17 + 130 - 10^{22} - 10
\end{aligned}
$$

En teoría deberían dar el mismo resultado.
¿Cuales son los valores que se obtienen en la computadora?

Propagación del error en la suma

Supongamos que tenemos dos números reales x, y con el mismo signo, y que

$$fl(x) = x(1+\delta_x), \qquad fl(y) = y(1+\delta_y)$$

El error relativo de la suma $x+y$ es

Propagación del error en la suma

Supongamos que tenemos dos números reales x, y con el mismo signo, y que

$$fl(x) = x(1+\delta_x), \qquad fl(y) = y(1+\delta_y)$$

El error relativo de la suma $x+y$ es

$$\delta_{x+y} = \frac{[fl(x)+fl(y)]-(x+y)}{x+y} = \frac{fl(x)-x}{x+y} + \frac{fl(y)-y}{x+y} = \delta_x \frac{x}{x+y} + \delta_y \frac{y}{x+y}$$

$$\left|\delta_{x+y}\right| \leq u \frac{|x|+|y|}{|x+y|} = u$$

Propagación del error en la suma

Supongamos que tenemos dos números reales x, y con el mismo signo, y que

$$fl(x) = x(1+\delta_x), \qquad fl(y) = y(1+\delta_y)$$

El error relativo de la suma $x+y$ es

$$\delta_{x+y} = \frac{[fl(x)+fl(y)]-(x+y)}{x+y} = \frac{fl(x)-x}{x+y} + \frac{fl(y)-y}{x+y} = \delta_x \frac{x}{x+y} + \delta_y \frac{y}{x+y}$$

$$\left|\delta_{x+y}\right| \leq u \frac{|x|+|y|}{|x+y|} = u$$

Para la resta se tiene algo similar:

$$\delta_{x-y} = \delta_x \frac{x}{x-y} - \delta_y \frac{y}{x-y}$$

Ejemplo

El polinomio de Rump se define como

$$R(x,y) = \frac{33375}{100}y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + \frac{55}{10}y^8 + \frac{x}{2y}$$

Si evaluamos este polinomio usando 'double' y 'long double' se tiene que

$$\begin{aligned} \text{(float)} \quad & R(77617,\ 33096) &&= && 6.33825 \times 10^{29} \\ \text{(double)} \quad & R(77617,\ 33096) &&= && 1.1726039400532 \\ \text{(long double)} \quad & R(77617,\ 33096) &&= && 1.1726039400531786\ldots \end{aligned}$$

Ejemplo

El polinomio de Rump se define como

$$R(x,y) = \frac{33375}{100}y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + \frac{55}{10}y^8 + \frac{x}{2y}$$

Si evaluamos este polinomio usando 'double' y 'long double' se tiene que

$$\begin{aligned} \text{(float)} \quad R(77617,\ 33096) &= 6.33825 \times 10^{29} \\ \text{(double)} \quad R(77617,\ 33096) &= 1.1726039400532 \\ \text{(long double)} \quad R(77617,\ 33096) &= 1.172603940053178 63\ldots \end{aligned}$$

Realizando las operaciones con fracciones, obtenemos

$$R(77617, 33096) = -\frac{54767}{66192} \approx -0.8273960599$$

Ejemplo (I)

Si

$$\begin{aligned} R_1(x,y) &= \frac{33375}{100}y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) \\ R_2(x,y) &= \frac{55}{10}y^8 \\ R_3(x,y) &= \frac{x}{2y} \end{aligned}$$

Entonces

$$\begin{aligned} R_1(77617,\ 33096) &= -79171113406689613611011347015249428 50, \\ R_2(77617,\ 33096) &= \ \ 79171113406689613611011347015249428 48, \\ R_3(77617,\ 33096) &= \frac{77617}{66192} \end{aligned}$$

Ejemplo (II)

Repetimos los cálculos usando 16 dígitos de precisión

$$\begin{aligned} R_1(77617,\ 33096) &= -7.917111340668963 \times 10^{36}, \\ R_2(77617,\ 33096) &= 7.917111340668962 \times 10^{36}, \\ R_3(77617,\ 33096) &= 1.172603940053179 \end{aligned}$$

$$R(77617,\ 33096) = -1 \times 10^{21} + 1.172603940053179 = -1 \times 10^{21}$$

Propagación del error en el producto

Si $fl(x) = x(1+\delta_x)$, $\quad fl(y) = y(1+\delta_y)$, entonces

$$\delta_{xy} = \frac{fl(x)fl(y) - xy}{xy} = \frac{xy(1+\delta_x)(1+\delta_y) - xy}{xy} = \delta_x + \delta_y + \delta_x\delta_y$$

$$|\delta_{xy}| \leq 2u + u^2$$

Para el caso de la división se tiene que:

$$\delta_{\frac{x}{y}} = \frac{\delta_x - \delta_y}{1+\delta_y}$$

Error combinando sumas y productos

Consideremos tres números reales x, y, z, y queremos calcular $x(y+z)$. Entonces, en lugar de operar x, y, z operamos con $x(1+\delta_x) = x+\epsilon_x$, $y(1+\delta_y) = y+\epsilon_y$ y $z(1+\delta_z) = z+\epsilon_z$

$$(x+\epsilon_x)(y+\epsilon_y+z+\epsilon_z) = x(y+z)+x(\epsilon_y+\epsilon_z)+(y+z)\epsilon_x+\epsilon_x(\epsilon_y+\epsilon_z)$$

Entonces el error es

$$E = x(\epsilon_y+\epsilon_z)+(y+z)\epsilon_x+\epsilon_x(\epsilon_y+\epsilon_z)$$

Error combinando sumas y productos

Consideremos tres números reales x, y, z, y queremos calcular $x(y+z)$. Entonces, en lugar de operar x, y, z operamos con $x(1+\delta_x) = x+\epsilon_x$, $y(1+\delta_y) = y+\epsilon_y$ y $z(1+\delta_z) = z+\epsilon_z$

$$(x+\epsilon_x)(y+\epsilon_y+z+\epsilon_z) = x(y+z)+x(\epsilon_y+\epsilon_z)+(y+z)\epsilon_x+\epsilon_x(\epsilon_y+\epsilon_z)$$

Entonces el error es

$$E = x(\epsilon_y+\epsilon_z)+(y+z)\epsilon_x+\epsilon_x(\epsilon_y+\epsilon_z)$$

Si suponemos que $|\epsilon_i| < \epsilon$, entonces

$$|E| \leq 2\epsilon|x|+|y+z|\epsilon+2\epsilon^2$$

Evaluación de polinomios (I)

Evaluamos el polinomio cúbico

$$p(x) = ax^3 + bx^2 + cx + d$$

donde

$$\begin{aligned} a &= 1.000, \\ b &= -89998.304, \\ c &= 2699898236.405, \\ d &= -26998473559412.543, \end{aligned} \tag{1}$$

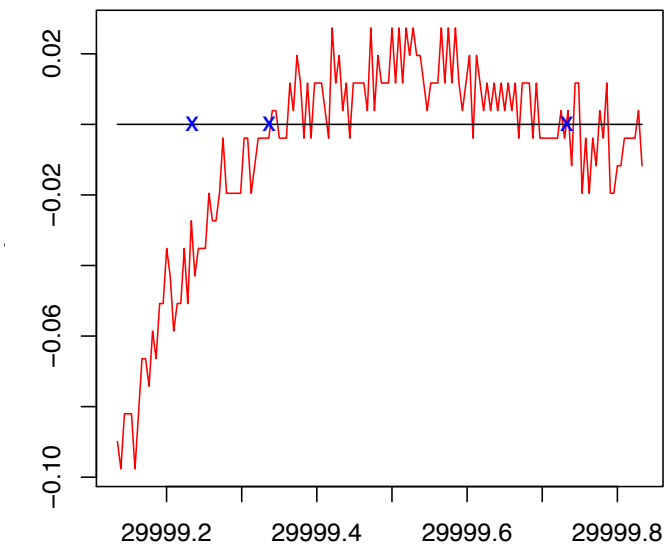
Evaluación de polinomios (I)

Evaluamos el polinomio cúbico

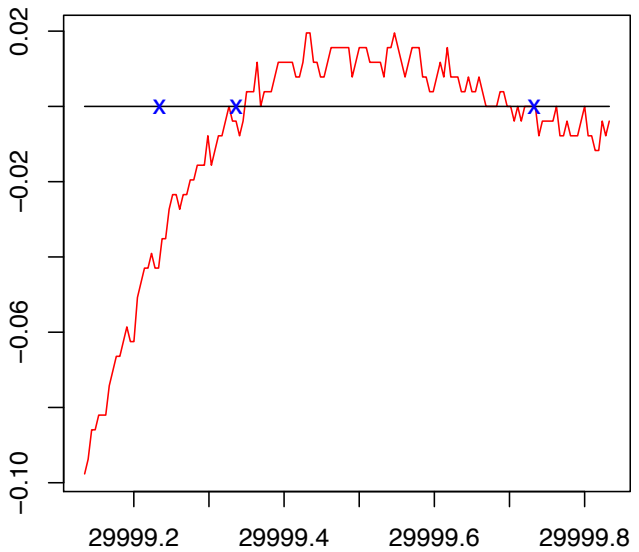
$$p(x) = ax^3 + bx^2 + cx + d$$

donde

$$\begin{aligned} a &= 1.000, \\ b &= -89998.304, \\ c &= 2699898236.405, \\ d &= -26998473559412.543, \end{aligned} \tag{1}$$



$((ax^3 + bx^2) + cx) + d$



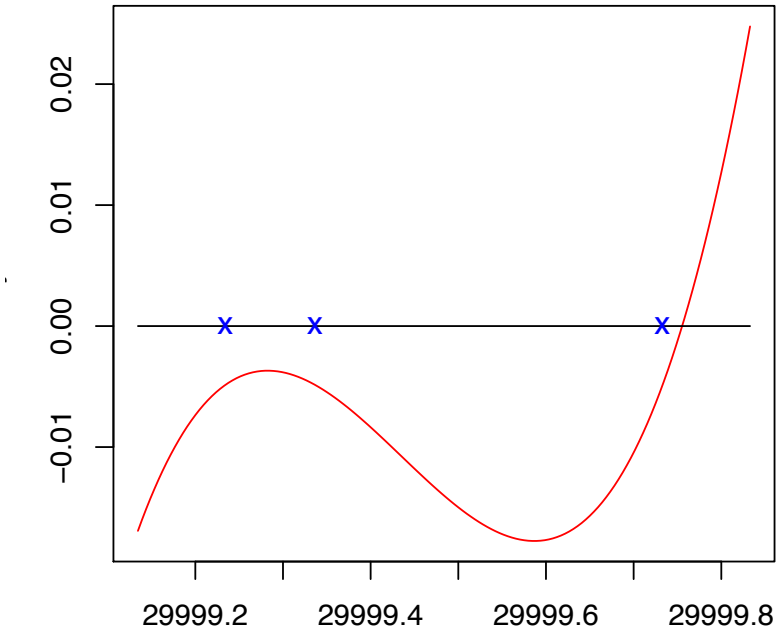
$(((ax + b)x)x + c)x + d$

Evaluación de polinomios (II)

Usando algunas estrategias, como las que se describen en

> *S. Graillat, P. Langlois and N. Louvet. Algorithms for accurate, validated and fast polynomial evaluation. Japan J. Indust. Appl. Math., vol. 26, pp. 191–214, 2009*

se obtiene lo siguiente:



Evaluación de polinomios (III)

El polinomio $p(x)$ se obtuvo al desarrollar la expresión

$$p(x) = (x - s_1)(x - s_2)(x - s_3)$$

donde

$$s_1 = 29999.234288122, \quad s_2 = 29999.336581462, \quad s_3 = 29999.733055670,$$

de modo que

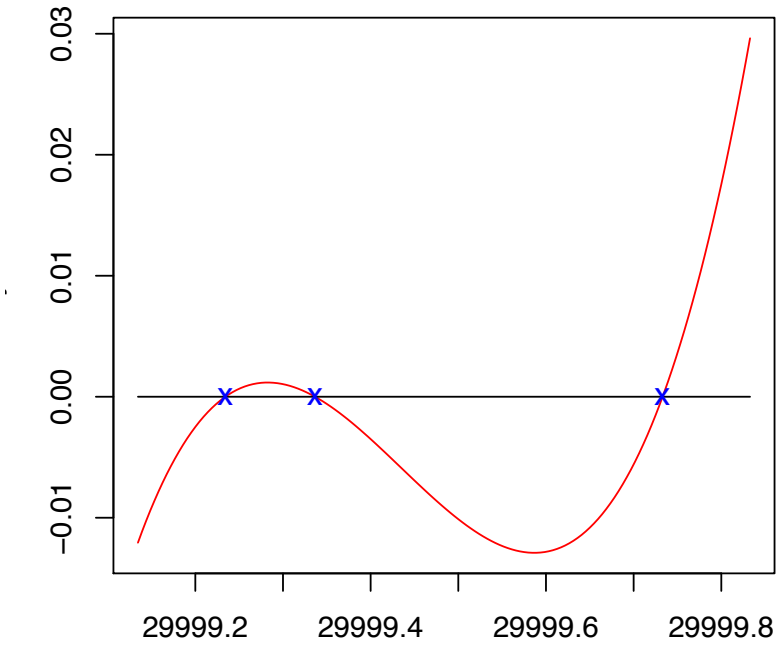
$$\begin{aligned} a &= 1 \\ b &= -s_1 - s_2 - s_3 \\ c &= s_2 s_3 + s_1 s_3 + s_1 s_2 \\ d &= -s_1 s_2 s_3 \end{aligned}$$

Los puntos indicados con 'x' en la gráfica anterior indican la posición de las raíces.
Es posible estimar los valores y_i que corrigen el cálculo de los coeficientes del polinomio

$$\begin{aligned} a &= fl(a) + y_1 \qquad b = fl(b) + y_2 \\ c &= fl(c) + y_3 \qquad d = fl(d) + y_4 \end{aligned}$$

Evaluación de polinomios (IV)

Entonces $p(x) = Horner(p, x) + y(x)$ y al evaluarlo se obtiene



En general, se puedan estrategias para calcular los coeficientes de un polinomio dadas sus raíces:

> *Calvetti, D. and Reichel, Lothar. On the evaluation of polynomial coefficients. Numerical Algorithms 33, pp. 153-161, 2003.*