

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 1 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Ejercicio 1.- Manejando aritmética decimal de cinco dígitos con truncamiento, calcular el intervalo en que la representación de la función $f(x) = 1 + x^3$ es uno.

Para realizar esta evaluación, el ordenador realiza los siguientes pasos

Representa x : $fl(x) = \tilde{x}$

Lo eleva al cuadrado y almacena el resultado: $fl(\tilde{x}^2)$

Multiplica la representación de x por el resultado anterior y lo almacena $fl(fl(\tilde{x}^2) \cdot \tilde{x})$

Suma el resultado a 1

Si este resultado del paso 3 es menor que la unidad de redondeo del ordenador δ , el paso 4 no modifica el valor 1. Puesto que trabajamos con cinco cifras decimales $\delta = 1 \times 10^{-5}$. Partiremos del que verifica que

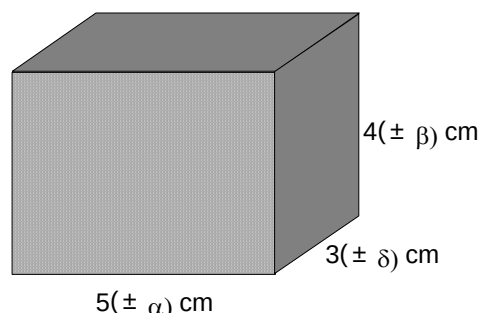
$$\tilde{x}^3 < 10^{-5} \rightarrow \tilde{x} < 0.0215443 \dots$$

Probemos las operaciones mencionadas con diferentes valores normalizados.

$\tilde{x}$	$(\tilde{x})^2$	$fl(\tilde{x}^2)$	$fl(\tilde{x}^2) \cdot \tilde{x}$	$fl(fl(\tilde{x}^2) \cdot \tilde{x})$
0.2×10^{-1}	0.4×10^{-3}	0.4×10^{-3}	0.8×10^{-5}	0.8×10^{-5}
0.21×10^{-1}	0.441×10^{-3}	0.441×10^{-3}	0.9261×10^{-5}	0.9261×10^{-5}
0.215×10^{-1}	0.46225×10^{-3}	0.46225×10^{-3}	0.9938375×10^{-5}	0.99383×10^{-5}
0.2154×10^{-1}	0.463972×10^{-3}	0.46397×10^{-3}	$0.99939138 \times 10^{-5}$	0.99939×10^{-5}
0.21544×10^{-1}	$0.464143936 \times 10^{-3}$	0.46414×10^{-3}	$0.999943216 \times 10^{-5}$	0.99994×10^{-5}
0.21545×10^{-1}	$0.464187025 \times 10^{-3}$	0.46419×10^{-3}	$1.000097355 \times 10^{-5}$	1.0001×10^{-5}

Se tiene que la mayor representación que verifica la condición es 0.21544×10^{-1} , por tanto todos los valores cuya representación sea esa, o menor, darán valores menores que δ y por tanto la representación de la función será 1. El conjunto de valores que cumplen la condición es el intervalo abierto $(-0.21545 \times 10^{-1}, 0.21545 \times 10^{-1})$

Ejercicio 2.- Un paralelepípedo rectangular tiene lados de 3 4 y 5 cm, medidos al centímetro más cercano. ¿Cuáles son las mejores cotas superior e inferior para su volumen?. ¿Y para su área superficial?



El volumen del paralelepípedo viene dado por el producto de sus tres dimensiones. Considerando que las medidas son exactas, esto es, no tienen error, se tiene:

$$V_e = x \cdot y \cdot z = 5 \times 4 \times 3 = 60 \text{ cm}^3$$

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 2 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Si se considera que las medidas son aproximadas, se obtiene

$$V_a = (x + \alpha) \cdot (y + \beta) \cdot (z + \gamma) \text{ donde } |\alpha|, |\beta|, |\gamma| < 0.5\text{cm}$$

El volumen máximo se obtiene cuando las medidas aproximan por defecto los valores reales, es decir, las medidas reales exceden las tomadas en medio centímetro. En este caso $\alpha = \beta = \gamma = 0.5\text{cm}$, y el valor del volumen máximo es $V_a^{\max} = 5.5 \times 4.5 \times 3.5 = 86.625\text{cm}$. Por el contrario, el volumen mínimo se obtiene al aproximar por exceso, en cuyo caso $\alpha = \beta = \gamma = -0.5\text{cm}$ y el volumen es $V_a^{\min} = 4.5 \times 3.5 \times 2.5 = 39.375\text{cm}$.

Calculamos los errores absolutos y relativos,

	<i>Error total</i>	<i>Error relativo</i>
<i>Volumen máximo</i>	$e = 86.625 - 60 = 26.625$	$\varepsilon = \frac{26.625}{86.375} = 30.74\%$
<i>Volumen mínimo</i>	$e = 39.375 - 60 = -20.625$	$\varepsilon = \frac{20.625}{39.375} = -52.38\%$

De forma similar se realiza el cálculo de la superficie. Considerando que las medidas son exactas se tiene $S_e = 2(x \cdot y + x \cdot z + y \cdot z) = 2(5 \times 4 + 5 \times 3 + 4 \times 3) = 94\text{cm}^2$

Si se considera que las medidas son aproximadas, se obtiene

$$S_a = 2((x + \alpha) \cdot (y + \beta) + (x + \alpha) \cdot (z + \gamma) + (y + \beta) \cdot (z + \gamma)) \text{ donde } |\alpha|, |\beta|, |\gamma| < 0.5\text{cm}$$

Utilizando las mismas hipótesis que en el caso del volumen, los valores máximos y mínimos, así como sus errores vienen dados por

	<i>Superficie</i>	<i>Error total</i>	<i>Error relativo</i>
<i>Máxima</i>	119.5	25.5	21.34%
<i>Mínima</i>	71.5	-22.5	-31.47

Ejercicio 3.- El siguiente algoritmo calcula $\sqrt{a}$, sin utilizar raíces cuadradas, con una exactitud arbitraria que depende del número de términos de la sucesión calculados. Evaluar los 5 primeros términos y compararlos con la solución exacta, indicando el error absoluto y relativo, para $a=2$

Paso 1 Leer a, ε

Paso 2 $x_0 \leftarrow a$

Paso 3 $n \leftarrow 1$

Paso 4 Repetir mientras $|x_{n+1} - x_n| \geq \varepsilon$ Pasos 5-6

Paso 5 $n \leftarrow n+1$

Paso 6 $x_n \leftarrow \frac{1}{2} \left(x_{n-1} + \frac{a}{x_{n-1}} \right)$

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 3 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Paso 7 Imprimir x_n

Recordamos que

$$E_n = |x_n - \sqrt{a}| \quad e_n = \left| \frac{x_n - \sqrt{a}}{\sqrt{a}} \right| \text{ y puesto que } \sqrt{2} = 1.4142135623730950488... \text{ se tiene}$$

n	x_n	E_n	e_n
0	2	0'585786	0'414214
1	3/2	$0'857864 \times 10^{-1}$	$0'606602 \times 10^{-1}$
2	17/12	$0'245310 \times 10^{-2}$	$0'173461 \times 10^{-2}$
3	577/408	$0'212390 \times 10^{-5}$	$0'150183 \times 10^{-5}$
4	665857/470832	$0'159472 \times 10^{-11}$	$0'112764 \times 10^{-11}$

Ejercicio 4.- Acotar el error que se comete al sumar n números con representación exacta

Sean: x_i los valores a sumar, tenemos que sus correspondientes errores relativos de representación ε_i son todos nulos $x_i = fl(x_i) = \tilde{x}_i$. Puesto que

$$x_1 + x_2 - fl(x_1 + x_2) = (x_1 + x_2) - (\tilde{x}_1 + \tilde{x}_2) = \underbrace{((x_1 + x_2) - (\tilde{x}_1 + \tilde{x}_2))}_{\text{Err.Propagado}} + \underbrace{((\tilde{x}_1 + \tilde{x}_2) - (\tilde{x}_1 + \tilde{x}_2))}_{\text{Err.Representación}}$$

no existirá error propagado en la suma, aunque sí puede aparecer un error de representación por

redondeo. Si representamos por $S_i = \sum_{j=1}^i x_j$ el valor exacto de la suma, y por $\tilde{S}_i = fl(\tilde{S}_{i-1} + x_i)$ la misma operación realizada por el ordenador, tendremos

$$S_2 = x_1 + x_2 \quad \tilde{S}_2 = fl(x_1 + x_2) = (x_1 + x_2)(1 - \delta_2)$$

$$S_2 - \tilde{S}_2 = (x_1 + x_2)\delta_2, \text{ donde } \delta_2 \text{ es el error de redondeo de la operación}$$

$$S_3 = x_1 + x_2 + x_3 \quad \tilde{S}_3 = fl(\tilde{S}_2 + x_3) = ((x_1 + x_2)(1 - \delta_2) + x_3)(1 - \delta_3) \\ = x_1(1 - \delta_2)(1 - \delta_3) + x_2(1 - \delta_2)(1 - \delta_3) + x_3(1 - \delta_3)$$

$$S_3 - \tilde{S}_3 \approx x_1(\delta_2 + \delta_3) + x_2(\delta_2 + \delta_3) + x_3(\delta_3)$$

Por inducción

$$S_n = \sum_{k=1}^n x_k \quad \tilde{S}_n = fl(\tilde{S}_{n-1} + x_n) = (((x_1 + x_2)(1 - \delta_2) + x_3)(1 - \delta_3) \cdots + x_n)(1 - \delta_n) \\ = x_1(1 - \delta_2) \cdots (1 - \delta_n) + x_2(1 - \delta_2) \cdots (1 - \delta_n) + \cdots + x_n(1 - \delta_n)$$

$$S_n - \tilde{S}_n \approx x_1(\delta_2 + \delta_3 + \cdots \delta_n) + x_2(\delta_2 + \delta_3 + \cdots \delta_n) + x_3(\delta_3 + \cdots \delta_n) + \cdots + x_{n-1}(\delta_{n-1} + \delta_n) + x_n(\delta_n)$$

Como $\delta_i < \varepsilon$ (error de redondeo de la representación), se tiene

$$|S_n - \tilde{S}_n| \leq \varepsilon(|x_1|(n-1) + |x_2|(n-1) + |x_3|(n-2) + |x_4|(n-3) + \cdots + |x_{n-1}|(2) + |x_n|)$$

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 4 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Se observa que la cota depende del orden en que se tomen los sumandos, por tanto la suma no es conmutativa. Asimismo, para disminuir el valor de esa cota, es conveniente ordenar los sumandos de menor a valor en valor absoluto antes de realizar la suma.

Ejercicio 5.- Acotar el error que se comete al sumar n veces un número con representación aproximada.

Sea: $\tilde{x}$ la representación aproximada de x , δ el error relativo de redondeo de la representación $\tilde{x} = x(1 - \delta)$ y ε la cota del error relativo de redondeo operacional.

Denotamos por $S_i = \sum_{j=1}^i x$ al resultado exacto de la operación y por $\tilde{S}_i = fl(\hat{S}_{i-1} + \tilde{x})$ al calculado por el ordenador.

Entonces

$$S_n - \tilde{S}_n = \left(\sum_{i=1}^N x - \sum_{i=1}^N \tilde{x} \right) + \left(\sum_{i=1}^N \tilde{x} - \tilde{S}_N \right)$$

Evaluando cada término

$$\left(\sum_{i=1}^N x - \sum_{i=1}^N \tilde{x} \right) = nx\delta$$

$$\begin{aligned} \left(\sum_{i=1}^N \tilde{x} - \tilde{S}_N \right) &= \tilde{x}(\varepsilon_2 + \varepsilon_3 + \cdots \varepsilon_n) + \tilde{x}(\varepsilon_2 + \varepsilon_3 + \cdots \varepsilon_n) + \tilde{x}(\varepsilon_3 + \cdots \varepsilon_n) + \cdots \tilde{x}(\varepsilon_{n-1} + \varepsilon_n) + \tilde{x}(\varepsilon_n) = \\ &= \tilde{x}(2\varepsilon_2 + 3\varepsilon_3 + \cdots n\varepsilon_n) \end{aligned}$$

Acotando

$$\text{E. Absoluto: } |S_n - \hat{S}_n| \leq |nx\delta| + |\tilde{x}| \{2\varepsilon + 3\varepsilon + \cdots n\varepsilon\} \leq |nx\delta| + |x(1-\delta)| \varepsilon \frac{(n-1)(n+2)}{2}$$

$$\text{E. Relativo: } \left| \frac{S_n - \hat{S}_n}{S_n} \right| \leq |\delta| + |(1-\delta)| \varepsilon \frac{(n-1)(n+2)}{2n}$$

Ejercicio 6.- Acotar el error que se comete al sumar n números con representación aproximada.

Sea: $\tilde{x}_i$ la representación aproximada de x_i , δ_i el error relativo de redondeo de la representación $\tilde{x}_i = x_i(1 - \delta_i)$ y ε la cota del error relativo de redondeo operacional.

Denotamos por $S_i = \sum_{j=1}^i x_j$ al resultado exacto de la operación y por $\tilde{S}_i = fl(\hat{S}_{i-1} + \tilde{x}_i)$ al calculado por el ordenador.

Entonces

$$S_n - \tilde{S}_n = \left(\sum_{i=1}^N x_i - \sum_{i=1}^N \tilde{x}_i \right) + \left(\sum_{i=1}^N \tilde{x}_i - \tilde{S}_N \right)$$

Evaluando cada término

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 5 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

$$\left(\sum_{i=1}^N x_i - \sum_{i=1}^N \tilde{x}_i \right) = \sum_{i=1}^N x_i \delta_i$$

$$\left(\sum_{i=1}^N \tilde{x}_i - \tilde{S}_N \right) = \tilde{x}_1 (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n) + \tilde{x}_2 (\varepsilon_3 + \dots + \varepsilon_n) + \tilde{x}_3 (\varepsilon_4 + \dots + \varepsilon_n) + \dots + \tilde{x}_{n-1} (\varepsilon_n) + \tilde{x}_n (\varepsilon_n)$$

Por tanto

$$\begin{aligned} S_n - \tilde{S}_n &= x_1 \delta_1 + x_1 (1 - \delta_1) (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n) + \sum_{i=1}^N x_i \left(\delta_i + (1 - \delta_i) \sum_{j=i}^N \varepsilon_j \right) \approx \\ &\approx x_1 (\delta_1 + \varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n) + \sum_{i=1}^N x_i \left(\delta_i + \sum_{j=i}^N \varepsilon_j \right) \end{aligned}$$

donde $|\delta_i|, |\varepsilon_i| < \delta \quad i = 1, \dots, n$. Acotando

$$\text{E. Absoluto: } |S_n - \hat{S}_n| \leq |x_1| n \delta + |x_2| n \delta + |x_3| (n-1) \delta + \dots + |x_n| 2 \delta$$

$$\text{E. Relativo: } \left| \frac{S_n - \hat{S}_n}{S_n} \right|$$

Ejercicio 7.- Acotar el error que se comete al multiplicar n números con representación exacta

Sean: x_i los valores a multiplicar, tenemos que sus correspondientes errores relativos de representación ε_i son todos nulos. Por tanto no existirá error propagado del producto. Sin embargo si aparecerá un error de redondeo. Si representamos por $P_n = \prod_{k=1}^n x_k$ el valor exacto del

producto, y por $\tilde{P}_n = fl(\tilde{P}_{n-1} \cdot x_n)$ la misma operación realizada por el ordenador, tendremos

$$P_2 = x_1 \cdot x_2 \quad \tilde{P}_2 = (x_1 \cdot x_2)(1 - \delta_2)$$

$$P_2 - \tilde{P}_2 = (x_1 \cdot x_2) \delta_2, \text{ donde } \delta \text{ es el error de redondeo de la operación}$$

$$P_3 = x_1 \cdot x_2 \cdot x_3 \quad \tilde{P}_3 = fl(\tilde{P}_2 \cdot x_3) = \tilde{P}_2 \cdot x_3 (1 - \delta_3) = (x_1 \cdot x_2 \cdot x_3)(1 - \delta_2)(1 - \delta_3)$$

$$P_3 - \tilde{P}_3 = (x_1 \cdot x_2 \cdot x_3)[1 - (1 - \delta_2)(1 - \delta_3)] \approx (x_1 \cdot x_2 \cdot x_3)(\delta_2 + \delta_3)$$

Por inducción

$$P_n = \prod_{k=1}^n x_k \quad \tilde{P}_n = fl(\tilde{P}_{n-1} \cdot x_n) = (x_1 \cdot x_2 \cdot x_3 \dots x_n)(1 - \delta_2)(1 - \delta_3) \dots (1 - \delta_n)$$

$$P_n - \tilde{P}_n = (x_1 \cdot x_2 \cdot x_3 \dots x_n)[1 - (1 - \delta_2)(1 - \delta_3) \dots (1 - \delta_n)] \approx (x_1 \cdot x_2 \cdot x_3 \dots x_n)(\delta_2 + \delta_3 \dots + \delta_n)$$

Como $\delta_i < \varepsilon$ (error de redondeo de la representación), se tiene

$$|P_n - \tilde{P}_n| \leq (|x_1| \cdot |x_2| \cdot |x_3| \dots |x_n|)(n-1)\varepsilon$$

Se observa que, al contrario que la adición, el producto sí es conmutativo.

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 6 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Ejercicio 8.- Acotar el error que se comete al multiplicar n veces un número con representación aproximada

Sea: $\tilde{x}$ la representación aproximada de x , δ el error relativo de redondeo de la representación $\tilde{x} = x(1 - \delta)$ y ε la cota del error relativo de redondeo operacional.

Denotamos por $P_n = \prod_{k=1}^n x$ al resultado exacto de la operación y por $\tilde{P}_n = fl(\tilde{P}_{n-1} \cdot x_n)$ al calculado por el ordenador. Entonces

$$P_n - \tilde{P}_n = \left(\prod_{i=1}^n x - \prod_{i=1}^n \tilde{x} \right) + \left(\prod_{i=1}^n \tilde{x} - \tilde{P}_n \right)$$

Evaluando cada término

$$\left(\prod_{i=1}^n x - \prod_{i=1}^n \tilde{x} \right) = x^n [1 - (1 - \delta)^n] \approx x^n n \delta$$

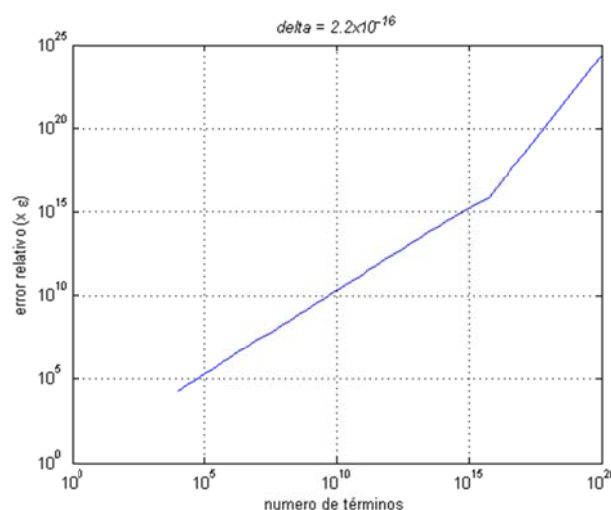
$$\left(\prod_{i=1}^n \tilde{x} - \tilde{P}_n \right) = (\tilde{x})^n (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n) = x^n (1 - \delta)^n (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n) \approx x^n (1 - n\delta) (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n)$$

Acotando

$$\text{E. Absoluto: } |P_n - \tilde{P}_n| \leq |nx^n \delta| + |x^n (1 - n\delta) (\varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_n)| \leq (n + |1 - n\delta|(n-1)) |\varepsilon| \cdot |x^n|$$

$$\text{E. Relativo: } \left| \frac{P_n - \tilde{P}_n}{P_n} \right| \leq (n + |1 - n\delta|(n-1)) |\varepsilon|$$

Se observa que el comportamiento del error es básicamente lineal hasta que $|1 - n\delta| \approx 0$, momento en que pasa a ser potencial



 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 7 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Ejercicio 9.- Dado el sistema $S = \begin{cases} 31.69x + 14.31y = 45.00 \\ 13.11x + 5.89y = 19.00 \end{cases}$ de solución exacta $x=7.2, y=-12.8$.

Resolverlo mediante el siguiente sistema equivalente

$$S = \begin{cases} 31.69x + 14.31y = 45.00 \\ [(13.11 \cdot 14.31) - (31.69 \cdot 5.89)]y = (13.11 \cdot 45.00) - (31.69 \cdot 19.00) \end{cases}$$

usando aritmética de 4 dígitos con corte. Analícese los resultados mediante un estudio de los errores.

$$\begin{cases} 31.69\tilde{x} + 14.31\tilde{y} = 45.00 \\ (187.6 - 186.6)\tilde{y} = 589.9 - 602.1 \end{cases} \rightarrow \begin{cases} \tilde{x} = \frac{45.00 + 12.2 \cdot 14.31}{31.69} = 6.926 \\ \tilde{y} = -12.2 \end{cases} \rightarrow \begin{cases} \varepsilon_x \approx 0.030 \\ \varepsilon_y \approx 0.047 \end{cases}$$

Realizamos el estudio de los errores, analizando la ecuación modificada (2ª ecuación)

$$[(13.11 \cdot 14.31) - (31.69 \cdot 5.89)]y = (13.11 \cdot 45.00) - (31.69 \cdot 19.00)$$

$$(187.6041 - 186.6541)y = 589.95 - 602.11 \rightarrow (1 + \varepsilon_1)y = -12.2(1 + \varepsilon_2)$$

$$\text{donde } \varepsilon_1 = -0.05 \text{ y } \varepsilon_2 = \frac{-0.04}{12.2} \approx 0.00328$$

$$y = -12.2 \frac{1 + \varepsilon_2}{1 + \varepsilon_1} = -12.2(1 + \varepsilon_2)(1 - \varepsilon_1 + (\varepsilon_1)^2 - (\varepsilon_1)^3 + \dots) \approx -12.2(1 + 0.04918)$$

$$y = \tilde{y}(1 + \varepsilon_y) \Rightarrow E_y = \tilde{y}\varepsilon_y \approx -12.2(0.04918) \approx -0.599996$$

Analizando los errores en x

$$31.69x + 14.31(-12.2)(1 + \varepsilon_y) = 45.00 \rightarrow 31.69x = 45.00 + 14.31(12.2)(1 + \varepsilon_y)$$

$$31.69x = 45.00 + 174.582(1 + \varepsilon_y) = 45.00 + 174.5(1 + \eta)(1 + \varepsilon_y) \rightarrow \eta = \frac{0.082}{174.5}$$

$$x = \frac{219.5}{31.69} + \frac{174.5}{31.69}(\varepsilon_y + \eta + \eta\varepsilon_y) = 6.926(1 + \zeta) + 6.926\zeta' \rightarrow \zeta' = \frac{174.5}{6.926 \cdot 31.69}(\varepsilon_y + \eta + \eta\varepsilon_y)$$

$$\text{donde } \zeta \approx 6.86 \times 10^{-5} \text{ y } \zeta' \approx 0.0395.$$

$$\text{La mayor parte del error se debe a } \frac{174.5}{6.926 \cdot 31.69} \varepsilon_y \approx 0.0391$$

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 8 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Ejercicio 10.- El desarrollo de Taylor de la función e^x viene dado por $\sum_{k=0}^{\infty} \frac{x^k}{k!}$. Calcular usando el polinomio de Taylor de orden 9, el valor de e^{-5} mediante las fórmulas

$$a) e^{-5} \approx \sum_{k=0}^9 \frac{(-5)^k}{k!} = \sum_{k=0}^9 \frac{(-1)^k 5^k}{k!} \quad b) e^{-5} = \frac{1}{e^5} \approx \frac{1}{\sum_{k=0}^9 \frac{5^k}{k!}}$$

El valor aproximado de e^{-5} es $6.737946999085 \times 10^{-3}$. Indicar, justificadamente, cuál de las fórmulas anteriores consigue la mayor precisión.

Comparando las fórmulas, se observa que en ambas los nuevos términos que se añaden son cada vez más pequeños. Con la primera fórmula se van alternando los signos, restando términos con signo alterno, y en la segunda en el denominador.

Tabulamos los valores pedidos (se han obtenido más términos de los pedidos para mostrar la evolución):

n	$(5)^n / n!$	$\sum_{k=0}^n \frac{(-5)^k}{k!}$	$\sum_{k=0}^n \frac{(5)^k}{k!}$	$\left(\sum_{k=0}^n \frac{(5)^k}{k!} \right)^{-1}$
0	1.0000	1.0000	1.0000	1.0000
1	5.0000	-4.0000	6.0000	0.1667
2	12.5000	8.5000	18.5000	0.05405
3	20.8333	-12.3333	39.3333	0.02542
4	26.0417	13.7083	65.3750	0.01530
5	26.0417	-12.3333	91.4167	0.01094
6	21.7014	9.3681	113.1181	0.008840
7	15.5010	-6.1329	128.6190	0.007775
8	9.6881	3.5552	138.3072	0.007230
9	5.3823	-1.8271	143.6895	0.006959
10	2.6911	0.8640	146.3806	0.006832
11	1.2232	-0.3592	147.6038	0.006775
12	0.5097	0.1505	148.1135	0.006752
13	0.1960	-0.04556	148.3096	0.006743
14	0.0700	0.02446	148.3796	0.006739
15	0.0233	-0.001119	148.4029	0.006738
16	0.0073	0.008412	148.4102	0.006738
17	0.0021	-0.006267	148.4124	0.006738
18	0.0006	0.006863	148.4130	0.006738
19	0.0002	0.006706	148.4131	0.006738
20	0.0000	0.006745	148.4131	0.006738

Aunque se ha visto que la fórmula más precisa es la (b), vamos a realizar un análisis de ambas. Por un lado, el error total se puede descomponer como el error del algoritmo más el error operacional, así se tiene que

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 9 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Error del algoritmo $e^x - \sum_{k=0}^{n-1} \frac{x^k}{k!} = \sum_{k=n}^{\infty} \frac{x^k}{k!} = \frac{\xi^n}{n!} \quad \xi \in (0, x)$

Error operacional $\sum_{k=0}^{n-1} \frac{x^k}{k!} - \tilde{S}_n = S_n - \hat{S}_n \approx x_0 (\delta_0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \dots + \varepsilon_{n-1}) + \sum_{i=1}^{n-1} x_i \left(\delta_i + \sum_{j=i}^{n-1} \varepsilon_j \right)$

representando por S_n y $\tilde{S}_n$ el cálculo exacto y el aproximado respectivamente.

Fórmula (a)

$$e^{-5} - \tilde{S}_9 = \left(e^{-5} - \sum_{k=0}^9 \frac{(-5)^k}{k!} \right) + \left(\sum_{k=0}^9 \frac{(-5)^k}{k!} - \tilde{S}_9 \right) \approx \frac{\xi^{10}}{10!} + x_0 (\delta_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_{10-1}) + \sum_{i=1}^9 x_i \left(\delta_i + \sum_{j=i}^9 \varepsilon_j \right) \quad \xi \in (0, -5)$$

$$|e^{-5} - \tilde{S}_9| \leq \frac{|\xi|^{10}}{10!} + |x_0| 10\varepsilon + \sum_{i=1}^9 |x_i| (10-i+1) \varepsilon = \frac{|\xi|^{10}}{10!} + 10\varepsilon + \sum_{i=1}^9 \left| \frac{(-5)^i}{i!} \right| (11-i) \varepsilon$$

Fórmula (b)

$$e^5 - \tilde{S}_9 = \left(e^5 - \sum_{k=0}^9 \frac{5^k}{k!} \right) + \left(\sum_{k=0}^9 \frac{5^k}{k!} - \tilde{S}_9 \right) \approx \frac{\xi^{10}}{10!} + x_0 (\delta_0 + \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_{10-1}) + \sum_{i=1}^9 x_i \left(\delta_i + \sum_{j=i}^9 \varepsilon_j \right) \quad \xi \in (0, 5)$$

$$|e^5 - \tilde{S}_9| \leq \frac{|\xi|^{10}}{10!} + |x_0| 10\varepsilon + \sum_{i=1}^9 |x_i| (10-i+1) \varepsilon = \frac{|\xi|^{10}}{10!} + 10\varepsilon + \sum_{i=1}^9 \left| \frac{5^i}{i!} \right| (11-i) \varepsilon$$

Ejercicio 11.- Calcular el error propagado por la función potencial $y = x^a$

Como ya se ha visto, para calcular el error debemos acotar las derivadas parciales de la función con respecto a sus parámetros

$$\frac{\partial y}{\partial x} = ax^{a-1} \quad \frac{\partial y}{\partial a} = x^a \ln x$$

El error absoluto viene dado por

$$E_y \leq \sum_k \left| \frac{\partial y}{\partial x_k} \right| E_{x_k} = |ax^{a-1}| E_x + |x^a \ln x| E_a$$

y dividiendo por el valor exacto, tenemos el correspondiente error relativo

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 10 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

$$\varepsilon_y \leq \sum_k \left| \frac{x_k}{y} \frac{\partial y}{\partial x_k} \right| \varepsilon_{x_k} = \left| \frac{x}{x^a} a x^{a-1} \right| \varepsilon_x + \left| \frac{a}{x^a} x^a \ln x \right| \varepsilon_a = |a| \varepsilon_x + |a \ln x| \varepsilon_a$$

Si a es exacto, entonces $\varepsilon_y \leq |a| \varepsilon_x$, luego el error aumenta en las potencias y disminuye en las raíces.

En el caso exponencial $y = e^a$, considerando que tanto e como a se calculan de forma exacta, se tiene que $x = e$ y su error por tanto es nulo ($E_x = \varepsilon_x = 0$).

Ejercicio 12.- Calcular el error propagado por las funciones trigonométricas

Como ya se ha visto, para calcular el error debemos acotar las derivadas parciales de la función con respecto a sus parámetros

$$y = \operatorname{sen} x \Rightarrow \begin{cases} E_y \leq |\cos x| E_x \\ \varepsilon_y \leq \left| \frac{x}{\operatorname{tg} x} \right| \varepsilon_x \end{cases} . \text{ Se observa que } \varepsilon_y \xrightarrow{x \rightarrow 0} \varepsilon_x$$

$$y = \operatorname{sen} x \Rightarrow \begin{cases} E_y \leq |\cos x| E_x \\ \varepsilon_y \leq \left| \frac{x}{\operatorname{tg} x} \right| \varepsilon_x \end{cases} . \text{ Se observa que } \varepsilon_y \xrightarrow{x \rightarrow \pi/2} \infty$$

$$y = \operatorname{tg} x \Rightarrow \begin{cases} E_y \leq \left| \frac{1}{\cos^2 x} \right| E_x \\ \varepsilon_y \leq \left| \frac{2x}{\operatorname{sen} 2x} \right| \varepsilon_x \end{cases}$$

$$y = \cot \operatorname{tg} x \Rightarrow \begin{cases} E_y \leq \left| \frac{-1}{\operatorname{sen}^2 x} \right| E_x \\ \varepsilon_y \leq \left| \frac{2x}{\operatorname{sen} 2x} \right| \varepsilon_x \end{cases}$$

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 11 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

Ejercicio 13.- La sucesión de términos $1, 3^{-1}, 3^{-2}, \dots$ puede generarse con cualesquiera de las siguientes relaciones recurrentes:

$$a) \begin{cases} p_0 = 1 \\ p_1 = \frac{1}{3} \\ p_n = \left(\frac{5}{6}\right)p_{n-1} - \left(\frac{1}{6}\right)p_{n-2} \end{cases} \quad b) \begin{cases} p_0 = 1 \\ p_1 = \frac{1}{3} \\ p_n = \left(\frac{5}{3}\right)p_{n-1} - \left(\frac{4}{9}\right)p_{n-2} \end{cases}$$

- (a) Calcule, usando aritmética de redondeo y de truncatura de 5 dígitos, los 20 primeros términos de las sucesiones y compare los resultados.
- (b) Estudie la estabilidad de los algoritmos numéricos.

n	Algoritmo a Redondeo	Algoritmo a Truncatura	Algoritmo b Redondeo	Algoritmo b Truncatura
0	1.000000	1.000000	1.000000	1.000000
1	.333330	.333330	.333330	.333330
2	.111100	.111100	.111120	.111090
3	.037027	.037030	.037050	.037000
4	.012339	.012343	.012365	.012292
5	.004111	.004114	.004142	.004041
6	.001369	.001371	.001408	.001272
7	.000456	.000457	.000506	.000323
8	.000152	.000152	.000217	-.000026
9	.000050	.000051	.000137	-.000187
10	.000017	.000017	.000132	-.000300
11	.000006	.000006	.000159	-.000417
12	.000002	.000002	.000207	-.000562
13	.000001	.000001	.000274	-.000751
14	.000000	.000000	.000365	-.001002
15	.000000	.000000	.000486	-.001336
16	.000000	.000000	.000648	-.001781
17	.000000	.000000	.000864	-.002375
18	.000000	.000000	.001152	-.003166
19	.000000	.000000	.001536	-.004221

Estabilidad del Algoritmo a:

La relación, matricialmente, viene dada por, $\begin{pmatrix} p_n \\ p_{n-1} \end{pmatrix} = \begin{pmatrix} \frac{5}{6} & -\frac{1}{6} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{n-1} \\ p_{n-2} \end{pmatrix}$ luego

$\begin{pmatrix} p_n \\ p_{n-1} \end{pmatrix} = \begin{pmatrix} \frac{5}{6} & -\frac{1}{6} \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} p_1 \\ p_0 \end{pmatrix}$. Calculando los valores propios de la matriz, se obtiene $\lambda_1 = \frac{1}{3}$ y

$\lambda_2 = \frac{1}{2}$. Por tanto $p_n = \alpha \left(\frac{1}{3}\right)^n + \beta \left(\frac{1}{2}\right)^n$.

Operando con aritmética real

 UNIVERSIDAD DE OVIEDO DEPARTAMENTO DE MATEMÁTICAS	Asignatura	Métodos Numéricos	Página 12 de 12
	Tema	Aritmética Finita (Representación y errores)	
	Autor	César Menéndez Fernández	

$$\begin{cases} p_0 = 1 = \alpha \left(\frac{1}{3}\right)^0 + \beta \left(\frac{1}{2}\right)^0 \\ p_1 = \frac{1}{3} = \alpha \left(\frac{1}{3}\right)^1 + \beta \left(\frac{1}{2}\right)^1 \end{cases} \rightarrow \begin{cases} \alpha = 1 \\ \beta = 0 \end{cases} \text{ y por tanto } p_n = \left(\frac{1}{3}\right)^n$$

Pero al trabajar con aritmética finita, se llega a

$$\begin{cases} p_0 = 1 = \alpha \left(\frac{1}{3}\right)^0 + \beta \left(\frac{1}{2}\right)^0 \\ p_1 = 0,33333 = \alpha \left(\frac{1}{3}\right)^1 + \beta \left(\frac{1}{2}\right)^1 \end{cases} \rightarrow \begin{cases} \alpha = 1.00002 \\ \beta = -0.00002 \end{cases}, \text{ con lo cual}$$

$E_n = \left(\frac{1}{3}\right)^n - \left[1.00002 \left(\frac{1}{3}\right)^n - 0.00002 \left(\frac{1}{2}\right)^n\right] = 0.00002 \left[\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n\right]$ y como $\lim_{n \rightarrow \infty} E_n = 0$, el algoritmo será estable.

Algoritmo b:

Repitiendo el proceso $\begin{pmatrix} p_n \\ p_{n-1} \end{pmatrix} = \begin{pmatrix} \frac{5}{3} & -\frac{4}{9} \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} p_1 \\ p_0 \end{pmatrix}$, siendo los valores propios obteniendo $\lambda_1 = \frac{1}{3}$

y $\lambda_2 = \frac{4}{3}$. Por tanto $p_n = \alpha \left(\frac{1}{3}\right)^n + \beta \left(\frac{4}{3}\right)^n$. Operando con aritmética real

$$\begin{cases} p_0 = 1 = \alpha \left(\frac{1}{3}\right)^0 + \beta \left(\frac{4}{3}\right)^0 \\ p_1 = \frac{1}{3} = \alpha \left(\frac{1}{3}\right)^1 + \beta \left(\frac{4}{3}\right)^1 \end{cases} \rightarrow \begin{cases} \alpha = 1 \\ \beta = 0 \end{cases} \text{ y por tanto } p_n = \left(\frac{1}{3}\right)^n$$

Pero al trabajar con aritmética finita, se llega a

$$\begin{cases} p_0 = 1 = \alpha \left(\frac{1}{3}\right)^0 + \beta \left(\frac{4}{3}\right)^0 \\ p_1 = 0,33333 = \alpha \left(\frac{1}{3}\right)^1 + \beta \left(\frac{4}{3}\right)^1 \end{cases} \rightarrow \begin{cases} \alpha = 1.00000\bar{3} \\ \beta = -0.00000\bar{3} \end{cases}, \text{ con lo cual}$$

$E_n = \left(\frac{1}{3}\right)^n - 1.00000\bar{3} \left(\frac{1}{3}\right)^n + 0.00000\bar{3} \left(\frac{4}{3}\right)^n = 0.00000\bar{3} \left\{ \left(\frac{4}{3}\right)^n - \left(\frac{1}{3}\right)^n \right\}$ y como $\lim_{n \rightarrow \infty} E_n \rightarrow \infty$, el algoritmo será inestable.