Magnitud de los errores por truncamiento y por redondeo

Lamentablemente, la literatura especializada sobre el tratamiento de errores es escasa por lo que sin embargo resulta muy importante conocer la magnitud de los errores que se cometen, en este caso, en el desarrollo de métodos numéricos. Un estudio sobre errores muy difundido entre la comunidad dedicada al desarrollo del Análisis numérico es la desarrollada por Daniel McCracken. El referido estudio está enfocado al manejo de datos numéricos en computadora y pertenece a un momento histórico en el cual los recursos de cómputo eran aún muy limitados en comparación con los disponibles en los inicios del siglo XXI. En realidad, las conclusiones de McCracken siguen vigentes hoy en día.

Una aportación importante sobre el estudio de los errores consiste en la cuantificación de la magnitud de los que se comenten en el manejo de los datos en forma inherente al uso de la aritmética de punto flotante. Mc Craken concluye que las magnitudes de los errores cometidos por truncamiento son mayores a las cometidas por el uso del redondeo simétrico (McCracken y Dorn, 1984). Asimismo, se concluye también que la magnitud del error por redondeo simétrico es independiente de la cantidad en sí misma siendo producto del tamaño de la mantisa que se utilice para hacer los cálculos. El máximo error absoluto debido al redondeo simétrico se calcula a través de la expresión:

$$\frac{1}{2} \cdot 10^{-t+1} \qquad \text{donde t es el tamaño de la mantisa}$$

Ejemplo. Utilizando una mantisa de 3 cifras, determine el máximo error absoluto cometido en las siguientes cifras:

1. 10.334
2. 123293.967

En ambos casos, las cantidades están definidas con una mantisa de tamaño tres, $t = 3$, para lo cual sustituyendo en la ecuación correspondiente:

$$\frac{1}{2} \cdot 10^{-t+1} = \frac{1}{2} \cdot 10^{-3+1} = 0{,}0005$$

Se observa que las cantidades 1 y 2 son muy diferentes en cuanto a magnitud; no obstante, el máximo error absoluto presente en cada una de ellas es igual.

Es importante establecer que en la realización de cálculos no es trascendente conocer el signo algebraico de los errores, lo importante es conocer la diferencia entre los valores de trabajo, es decir, su distancia en valor absoluto. Esta debe ser siempre menor que una cantidad de error permitida para considerar válido el cálculo. En la práctica de la Ingeniería, a esta cantidad de error permitida se le conoce como *tolerancia*.

Las tolerancias suelen expresarse en forma de porcentajes (errores relativos) y casi siempre están enfocadas hacia el número de cifras significativas que deben utilizarse en la aproximación. Se puede demostrar que si el siguiente criterio se cumple, puede tenerse la seguridad de que el resultado es correcto en al menos n cifras significativas:

$$tol = (0{,}5x10^{2-n}) \qquad [\,\%]$$

Ejemplo.Calcule el valor de la función e^1 utilizando la serie:

$$e^x = \sum_{i=0}^{n} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

variando el número de términos de la serie utilizados y utilizando cinco cifras exactas. Para este ejemplo, la tolerancia es $tol = 0{,}5 \cdot 10^{2-5} = 0{,}00050$. Si se considera como valor real el obtenido directamente de una calculadora, el resultado se muestra en la siguiente tabla:

Cuadro 1: Errores en el cálculo de series infinitas

Término	**Valor**	**Error**
1	1	1.71828
2	2	0.71828
3	2.5	0.21828
4	2.66667	0.05161
5	2.70833	0.00995
6	2.71667	0.00161
7	2.71806	0.00022

Una segunda aportación del estudio de McCracken es el establecimiento de un proceso para medir la propagación de los errores ocasionados por el uso de la aritmética de punto flotante. A partir del establecimiento del máximo error absoluto cometido y de la operación aritmética utilizada se demuestra que en este tipo de procesos el orden en que se realiza las operaciones sí modifica el resultado.

Ejemplo. Sumar las cantidades siguientes, primero en orden ascendente y luego en orden descendente, considerando una mantisa normalizada de cuatro dígitos así como redondeo simétrico en cada operación intermedia; por otra parte, realice la suma exacta (con todos los dígitos posibles en un calculadora) y considere este valor como exacto. Calcule el error relativo que se comete en cada caso.

1. $0{,}2685x10^4$
2. $0{,}9567x10^3$
3. $0{,}0053x10^2$
4. $0{,}1111x10^1$

Para las alternativas solicitadas, en las tablas respectivas se mostrará la cantidad normalizada así como el subtotal, es decir, la suma redondeada en una mantisa normalizada de tamaño 4.

El valor *exacto*, obtenido a través de una calculadora es: 3643,341.

El procedimiento consiste en normalizar las cantidades (igualando el exponente de la base diez en cada cantidad) y sumarlas en forma ascendente o descendentes, según sea el caso; en la suma de cada par de cantidades, se redondea el resultado manteniendo la mantisa en el tamaño preestablecido.

Cuadro 2: Suma descendente

Cantidad	**Cantidad Normalizada**	**Subtotal**
$0{,}2685x10^4$	$0{,}2685x10^4$	
$0{,}9567x10^3$	$0{,}09567x10^4$	$0{,}3642x10^4$
$0{,}0053x10^2$	$0{,}0001x10^4$	$0{,}3643x10^4$
$0{,}1111x10^1$	$0{,}0001x10^4$	$0{,}3644x10^4$

Cuadro 3: Suma ascendente

Cantidad	**Cantidad Normalizada**	**Subtotal**
$0{,}1111x10^1$	$0{,}1111x10^1$	
$0{,}0053x10^2$	$0{,}0530x10^1$	$0{,}1614x10^1$
$0{,}9567x10^3$	$95{,}67.x10^1$	$95{,}8341x10^1$
$0{,}2685x10^4$	$268{,}5x10^1$	$363{,}3341x10^1$

Cuadro 4: Comparación de resultados

	Resultado	**Error absoluto**	**Error relativo**
Valor exacto	3643,341		
Suma descendente	$0{,}3664x10^4$	20,659	0,56703 %
Suma ascendente	$363{,}3341x10^1$	10	0,27447 %

En el cuadro dos se muestra la suma ascendente y en el cuadro tres se muestra la suma en forma descendente. Finalmente, los resultados se incluyen en el cuadro cuatro.

Finalmente, este estudio arroja tres importantes conclusiones que deben considerarse en el diseño de algoritmos para ejecutar métodos numéricos.

Las conclusiones de McCracken son las siguientes:

1. Cuando se van a sumar y/o restar números, se debe trabajar siempre con los más pequeños primero.
2. De ser posible, evitar la sustracción de dos números aproximadamente iguales. Una expresión que contenga dicha sustracción puede a menudo ser reescrita para evitarla.
3. Una expresión del tipo $a(b-c)$ puede reescribirse de la forma $ab-ac$ y $\frac{(a-b)}{c}$ como $\frac{a}{c}-\frac{b}{c}$. Si hay números aproximadamente iguales dentro del paréntesis, ejecutar la resta antes que la multiplicación. Esto evitará complicar el problema con errores de redondeo adicionales.
4. Cuando no se aplica ninguna de las reglas anteriores, debe minimizarse el número de operaciones aritméticas.

Queda como labor voluntaria analizar estas conclusiones y comprobar la forma en que fueron obtenidas.

5. Convergencia y estabilidad de un método numérico

Matemáticamente, la *convergencia* es la propiedad de algunas sucesiones y series de tender progresivamente a un límite, de tal forma, si este límite existe, se dice que la sucesión o la serie *converge*. En forma análoga, si un método numérico en su funcionamiento iterativo nos proporciona aproximaciones cada vez más cercanas al valor buscado, se dice que el método converge. La convergencia se mide a través de los errores; si el error entre dos aproximaciones sucesivas se reduce, el método converge; se debe cumplir que:

$$|x_n - x_{n-1}| \leq |x_{n-1} - x_{n-2}|$$

Es decir, la diferencia enésima $(x_n - x_{n-1})$ debe ser menor que la diferencia $(n-1)$ésima $x_{n-1} - x_{n-2}$.

Se dice que un sistema (o un proceso) es *estable* si a pequeñas variaciones en la entrada o en la excitación corresponden pequeñas variaciones en la salida o en la respuesta. La estabilidad de un método numérico tiene que ver con la manera en que los errores numéricos se propagan a lo largo del algoritmo. Cuando un método converge, lo más deseable es que en los resultados que se obtengan, los niveles de error se disminuyan en la forma más rápida posible. Sin embargo, ocurre que durante la operación del algoritmo, ya sea por el manejo de los datos numéricos o bien por la naturaleza propia del modelo matemático con el que se esté trabajando, los errores entre aproximaciones no disminuyan en forma progresiva, sino que incluso aumenten en alguna etapa del proceso para después reducirse mostrando un comportamiento aleatorio.

La robustez de un método numérico radica en su convergencia y su estabilidad. Pueden utilizarse métodos cuya prueba de convergencia indique la pertinencia de su uso, pero que durante su aplicación se obtengan resultados inestables que repercutan en el número de iteraciones y en consecuencia en el tiempo invertido en la solución. El ideal lo constituyen métodos que a la vez de ser convergentes resulten estables.

6. Aproximación de funciones por medio de polinomios

Particularmente en el manejo de funciones trascendentes, la solución analítica de problemas puede ser difícil y complicada; incluso esta situación podría ocurrir en el ámbito de la solución numérica. Cuando esto ocurre, una herramienta de solución posible es utilizar una representación aproximada de la función a través de funciones más sencillas. Algunas de estas aproximaciones son:

- Funciones periódicas (senos y cosenos) a través de las series de Fourier
- Segmentar la función a través de una secuencia de líneas rectas
- La series de Taylor

La expansión en series de Taylor busca obtener una aproximación a $f(x)$ a través de un polinomio de la forma:

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + a_{n-3} x^{n-3} + ... + a_1 x + a_0 \tag{1}$$

en la vecindad del punto $x = x_0$ para sus primeras n derivadas. Por lo anterior, se requiere que $f(x)$ tenga $n-1$ derivadas en el intervalo $a \leq x \leq b$, es decir que:

$$\begin{aligned} P(X_0) &= f(X_0) \\ P^{'}(X_0) &= f^{'}(X_0) \\ P^{''}(X_0) &= f^{''}(X_0) \\ &\vdots \\ P^{(n)}(X_0) &= f^{(n)}(X_0) \end{aligned} \tag{2}$$

Es necesario determinar los coeficientes a_i del polinomio (1) y las derivadas valuadas en $X_0 = 0$:

$$\begin{aligned} P(0) = a_0 &\Rightarrow a_0 = f(0) \\ P^{'}(0) = a_1 &\Rightarrow a_1 = f^{'}(0) \\ P^{''}(0) = 2a_2 &\Rightarrow a_2 = \tfrac{1}{2} f^{''}(0) \\ P^{'''}(0) = 3!a_3 &\Rightarrow a_3 = \tfrac{1}{3!} f^{'''}(0) \\ &\vdots \\ P^{(k)}(0) = k!a_k &\Rightarrow a_k = \tfrac{1}{k!} f^{(k)}(0) \end{aligned} \tag{3}$$

para $n = 0, 1, 2, 3, 4, \cdots, n$

Sustituyendo (2) y (3) en (1):

$$P(X) = f(0) + f^{'}(0)X + \frac{f^{''}(0)}{2!}X^2 + \frac{f^{'''}(0)}{3!}X^3 + \cdots + \frac{f^{(n)}(0)}{n!}X^n = \sum_{k=0}^{n} \frac{X^k}{n!} f^{(k)}(0) \tag{4}$$

La expresión (6) representa la Serie de McLaurin.

En un caso particular, es probable que se requiera que el polinomio $P(X)$ sea igual a la función $f(x)$ en un punto X diferente de cero, es decir, $X = a \neq 0$, se procede de la misma forma:

$$\begin{aligned} P(X_a) &= f(X_a) \\ P^{'}(X_a) &= f^{'}(X_a) \\ P^{''}(X_a) &= f^{''}(X_a) \\ &\vdots \\ P^{(n)}(X_a) &= f^{(n)}(X_a) \end{aligned} \tag{5}$$

Esta consideración genera un crecimiento de la abscisa, por lo cual la expresión general queda:

$$P(X) = \sum_{k=0}^{n} \frac{(X-a)^k}{n!} f^{(k)}(a) \tag{6}$$

A la ecuación (5) se le conoce polinomio de Taylo de grado n para la función $f(x)$ en el punto $x = a$.

Ejemplo. Calcule los polinomios de Taylor de grados 1 y 3 para $f(x) = \cos x$ en $x = \frac{\pi}{2}$.

Sean:

$$P_1(X) = \sum_{k=0}^{1} \frac{(X-\frac{\pi}{2})^k}{n!} f^(k)(\frac{\pi}{2}) = \frac{(X-\frac{\pi}{2})^0}{0!} f^{(0)}(\frac{\pi}{2}) + \frac{(x-\frac{\pi}{2})^1}{1!} f'(\frac{\pi}{2}) \tag{7}$$

y

$$P_3(X) = \sum_{k=0}^{3} \frac{(X-\frac{\pi}{2})^k}{n!} f^(k)(\frac{\pi}{2}) = \frac{(X-\frac{\pi}{2})^0}{0!} f^{(0)}(\frac{\pi}{2}) + \frac{(x-\frac{\pi}{2})^1}{1!} f'(\frac{\pi}{2}) + \frac{(X-\frac{\pi}{2})^2}{2!} f''(\frac{\pi}{2}) + \frac{(X-\frac{\pi}{2})^3}{3!} f'''(\frac{\pi}{2}) \tag{8}$$

Las derivadas valuadas en $\frac{\pi}{2}$:

$$\begin{gathered} f(X) = \cos(X) \Rightarrow f(\tfrac{\pi}{2}) = 0 \\ f^{'}(X) = -\sin(X) \Rightarrow f^{'}(\pi 2) = -1 \\ f^{''}(X) = -\cos(X) \Rightarrow f^{''}(\pi 2) = 0 \\ f^{'''}(X) = \sin(X) \Rightarrow f^{'''}(\pi 2) = 1 \end{gathered} \tag{9}$$

Sustituyendo (9) en (7) y en (8):

$$P_1(X) = 1 \cdot 0 + \frac{(X-\frac{\pi}{2})}{1} \cdot (-1) = -X + \frac{\pi}{2}$$

$$P_1(X) = 1 \cdot 0 + \frac{(X-\frac{\pi}{2})}{1} \cdot (-1) + \frac{(X-\frac{\pi}{2})^2}{2} \cdot (0) + \frac{(X-\frac{\pi}{2})^3}{6} \cdot (1)$$

$$P_1(X) = 0{,}167X^3 - 0{,}785X^2 + 0{,}234X + 0{,}925$$

La figura (1) muestra gráficamente cada una una de las aproximaciones a $f(x)$.

6.1. Residuo del polinomio de Taylor

No se debe perder de vista que el polinomio de Taylor es una aproximación a la función $f(x)$; conlleva un error que no suele ser considerado pero que en función de su orden pudiera llegar a ser significativo. De tal forma, a la expresión:

$$f(X) = \sum_{k=0}^{n} \frac{(X-a)^k}{n!} f^{(k)}(a) + E_n(X) \tag{10}$$

se le conoce como Fórmula de Taylor con residuo. Para calcular $E_n(X)$ se evalúa la ecuación (6.1) para diversos órdenes:

Primer orden:

$$f(X) = f(a) + (X-a) f^{'}(a) + E_1(X)$$

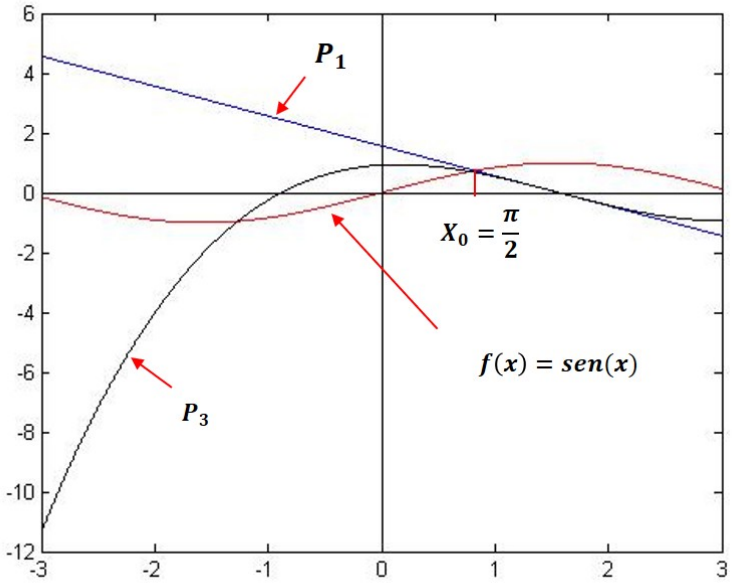

Figura 1: Aproximaciones a la función $seno(x)$

Despejando $E_1(X)$:

$$E_1(X) = f(X) - f(a) - (X - a)f'(a) \tag{11}$$

La ecuación (11) puede expresarse de forma integral:

$$E_1(X) = \int_a^X f'(t)dt - f'(a)\int_a^X dt = \int_a^X \left[f'(X) - f'(a)\right] dt \tag{12}$$

Integrando por partes:

$$\begin{array}{cc} u = f'(t) - f'(a) & du = f''(t)dt \\ dv = dt & v = t \end{array}$$

$$E_1(X) = \left|\left[f'(t) - f'(a)\right] \cdot t\right|_a^X - \int_a^X t \cdot f''(t)dt$$

$$E_1(X) = \left[f'(X) - f'(a)\right] \cdot X - \left[f'(a) - f'(a)\right] \cdot a - \int_a^X t \cdot f''(t)dt$$

$$E_1(X) = X\int_a^X f''(t)dt - \int_a^X t \cdot f''(t)dt$$

$$E_1(X) = \int_a^X (X-t)f^{''}(t)dt \tag{13}$$

Para un segundo orden el resultado es:

$$E_2(X) = \frac{1}{2!}\int_a^X (X-t)^2 f^{'''}(t)dt \tag{14}$$

Y para el n-ésimo orden:

$$E_1(X) = \frac{1}{n!}\int_a^X (X-t)^n f^{(n+1)}(t)dt \tag{15}$$

La ecuación (15) es el error cometido al aproximar la función $f(X)$ con un polinomio de Taylor de grado n.

6.2. Estimación del error de la aproximación de Taylor

Dato que la aproximación de Taylor representa una serie con un número infinito de términos, no es posible encontrar un valor exacto para $E_n(X)$, por lo que es necesario hacer algunas consideraciones: supóngase que m y M son los valores mínimo y máximo respectivamente, que adquiere la función $f^{(n+1)}(t)$ en el intervalo $[a, X]$. Sustituyendo estos supuestos en la ecuación (15):

$$[E_m(X)]_m = m\frac{(X-a)^{(n+1)}}{(n+1)!} \quad ; \quad [E_M(X)]_M = M\frac{(X-a)^{(n+1)}}{(n+1)!}$$

Ambas expresiones son las cotas de error, es decir:

$$m\frac{(X-a)^{(n+1)}}{(n+1)!} \leq E_n(X) \leq= M\frac{(X-a)^{(n+1)}}{(n+1)!}$$

Establecer los valores de m y M es un problema complicado. En aplicaciones reales se toma un criterio práctico que consiste en evaluar el término $n+1$ de la serie en algún punto de interés Xque esté en la vecindad de $X_0 = a$. Por ejemplo, se desea estimar el error cometido al aproximar $y = \sin(X)$, para $X_0 = 0$, a través de un polinomio de sexto orden. El polinomio de Taylor de $\sin(X)$ es ampliamente conocido:

$$\sin(X) = X - \frac{X^3}{3!} + \frac{X^5}{5!}$$

El siguiente término de la serie $(n+1)$ para estimar el error es: $E_6 \leq \frac{X^7}{7!}$. Se define como una desigualdad porque el valor real del error estará más cerca del punto pivote, en este caso de $X_0 = 0$, por lo que el error será menor. Esto se comprueba con los siguientes valores de X_0:

$$\begin{array}{ll} X = \pi & E_6(\pi) = \frac{pi^7}{7!} = 0{,}5993 \\ X = \frac{\pi}{2} & E_6(\frac{\pi}{2}) = \frac{\pi}{2^{7\cdot 7}} = 0{,}0047 \\ X = \frac{\pi}{4} & E_6(\frac{\pi}{4}) = \frac{\pi}{4^{7\cdot 7}} = 0{,}000037 \end{array}$$