Introducción al Análisis numérico y tratamiento de errores

Ing. Jesús Javier Cortés Rosas
M. en A. Miguel Eduardo González Cárdenas
M. en A. Víctor D. Pinilla Morán *

2011

Resumen

Definición de Análisis Numérico. Necesidad del uso de los métodos numéricos. Definición, clasificación y cuantificación de errores. Aproximación numérica.

1. Definición de Análisis Numérico

El *Análisis Numérico* es una rama de las matemáticas[**?**] que, mediante el uso de algoritmos iterativos, obtiene soluciones numéricas a problemas en los cuales la matemática simbólica (o analítica) resulta poco eficiente y en consecuencia no puede ofrecer una solución. En particular, a estos algortimos se les denomina *métodos numéricos*.

Por lo general los métodos numéricos se componen de un número de pasos finitos que se ejecutan de manera lógica, mejorando aproximaciones iniciales a cierta cantidad, tal como la raíz de una ecuación, hasta que se cumple con cierta cota de error. A esta operación cíclica de mejora del valor se le conoce como *iteración*.

Ejemplo. Uno de los ejercicios más comunes en los cursos básicos de Álgebra universitaria consiste en encontrar las raíces de un polinomio. El estudiante conoce principios tales como que el polinomio posee n raíces, donde n es el grado del polinomio. Conoce también que es posible que existan exclusivamente raíces reales o bien, una combinación entre raíces reales y raíces complejas, existiendo estas últimas en parejas conjugadas. El método de solución comúnmente utilizado es la división sintética (que es un método numérico). El estudiante aplica el método tantas veces como sea necesario para lograr que el residuo de la división sea cero, o muy cercano a cero.

No obstante, este procedimiento podría dejar insatisfecho a un estudiante acusioso pues aún cuando existen mecanismos para elegir un valor incial de una raíz, se invierte mucho tiempo mejorando este valor incial; adicionalmente es complicado obtener las raíces complejas, cosa que usualmente debe lograrse a través de un cambio de variable y del uso de la fórmula general para ecuaciones de segundo grado. Finalmente, este proceso sólo es aplicable en polinomios; no es posible su aplicación en ecuaciones trascendentes.

*Facultad de Ingeniería, UNAM. Profesores de tiempo completo del Departamento de Matemáticas Aplicadas de la División de Ciencias Básicas

El análisis numérico es una alternativa muy eficiente para la resolución de ecuaciones, tanto algebraicas (polinomios) como trascendentes teniendo una ventaja muy importante respecto a otro tipo de métodos: La repetición de instrucciones lógicas (iteraciones), proceso que permite mejorar los valores inicialmente considerados como solución. Dado que se trata siempre de la misma operación lógica, resulta muy pertinente el uso de recursos de cómputo para realizar esta tarea.

2. Necesidad del uso del análisis numérico

El desarrollo y el auge del uso del análisis numérico corre en forma paralela al desarrollo tecnológico de la computación[**?**]. Las computadoras (y en consecuencia también las calculadoras) están facultadas para realizar una multitud practicamente infita de operaciones algebraicas en intervalos de tiempo muy pequeños; esto las convierte en la herramienta ideal para la aplicación de los métodos numéricos. De hecho, el análisis numérico resulta ser la manera natural de resolver modelos matemáticos (de naturaleza algebraica o trascendente tanto para la matemática continua como para la discreta) a través de la computadora.

Por otra parte, como consecuencia directa de la aplicación de soluciones numéricas y del crecimiento de recursos computacionales, se ha logrado también la incorporación de la simulación matemática como una forma de estudio de diversos sistemas.

Sin embargo debe haber claridad en el sentido de que el análisis numérico no es la panacea en la solución de problemas matemáticos.

Consecuencia de lo anteriormente dicho consiste en que, por lo general, los métodos numéricos arrojan soluciones numéricas. Si en determinado caso se desea obtener soluciones analíticas deberá recurrir a los procedimientos algebraicos acostumbrados. Por otra parte, las soluciones numéricas resultan ser *aproximaciones*, es decir, en pocas ocasiones son soluciones exactas.

Como se analizará en su oportunidad, las soluciones numéricas conllevan una cota de error. Este error, que si bien puede ser tan pequeño como los recursos de cálculo lo permitan, siempre está presente y debe considerarse su manejo en el desarrollo de las soluciones requeridas.

Es muy posible que se conozca de diversos sistemas de cómputo que proporcionen soluciones analíticas. Estos sistemas no sustituyen a los métodos numéricos, de hecho son un complemento en el proceso integral del modelado de sistemas físicos que son el elemento fundamental de la práctica de la Ingeniería.

3. Definición de errores

Una actividad frecuente del profesional de la Ingeniería consiste en trabajar con modelos matemáticos representativos de un fenómeno físico. Estos modelos son abstracciones matemáticas que distan mucho de representar exactamente al fenómeno bajo estudio debido principalmente a las carencias y dificultades que aún posee el humano de la comprensión total de la naturaleza.

Como consecuencia de esto existen diferencias entre los resultados obtenidos experimentalmente y los emanados propiamente del modelo matemático.

A las diferencias cuantitativas entre los dos modelos se les denomina *Errores*.

Ejemplo. Sea h la altura a la que se encuentra un cuerpo, g la constante de la aceleración de la gravedad y t el tiempo que dura la caída, se define al modelo matemático como:

$$t = \sqrt{\frac{2h}{g}}$$

Resulta lógico pensar que al realizar los cálculos utilizando el anterior modelo se obtendrán resultados que diferirán de las mediciones que pudieran obtenerse en el desarrollo del experimento.

4. Clasificación de los errores

Las diferencias (errores) son múltiples y de diversa naturaleza, aunque pueden separse en dos grupos genéricos:

- Los errores[**?**] que provienen del modelado teórico (o abstracción matemática) del fenómeno real; estos errores se denominan *Errores del modelo o inherentes*. Los errores inherentes son producto de factores intrínsecos a la naturaleza, al ambiente y las personas mismas. Los errores inherentes son imposibles de remediar aunque pueden minimizarse; en consecuencia, no pueden cuantificarse.

  Se distinguen dos tipos de errores inherentes: Las *incertidumbres* hacen referencia a las dimensiones físicas que nunca podrán ser medidas en forma exacta debido a la naturaleza de la materia y a las imperfecciones de los instrumentos de medición. Las *verdaderas equivocaciones* son las situaciones que se producen en la lectura de instrumentos de medición o en el traslado de información y que son inadvertidas a las personas; un claro ejemplo de estas situaciones es la denominada *ceguera de taller*.

- Los *errores del método* son producto de la limitante en la representación y manipulación de cantidades numéricas utilizadas en los cálculos necesarios en el desarrollo del modelo matemático. Es de destacar que los dispositivos de cálculo (tales como calculadoras y computadoras) utilizan y manipulan cantidades en forma imprecisa.

  Existen dos grandes tipos de errores del método: El *truncamiento* se provoca ante la imposibilidad de manipular, por parte de un instrumento de cómputo, una cantidad infinita de términos o cifras. Los términos o cifras omitidas (que son infinitas en número) introducen un error en los resultados calculados. El *redondeo* se produce por el mismo motivo que el truncamiento pero, a diferencia de éste, las cifras omitidas sí son consideradas en la cifra resultante. Esta consideración se hace aplicando el siguiente esquema al dígito menos significativo (dms) de la cifra a redondear de acuerdo al siguiente esquema:

  1. Si el dms es mayor a 5, se incrementa en una unidad la cifra anterior.
  2. Si el dms es menor a 5, la cifra anterior no se modifica.
  3. Si el dms es igual a 5, deberá observarse a la cifra anterior; si ésta es par no sufre modificación, pero por el contrario, si es impar, deberá incrementarse en una unidad.

  Quizás se conozca una versión práctica y popular del redondeo simétrico en el cual la consideración tres se incluye en la primera de este esquema. Finalmente, existen también esquemas que permiten minimizar la ocurrencia de estos errores, de igual forma es importante destacar que los errores del método sí pueden ser cuantificados.

5. Cuantificación de errores

Los errores se cuantifican de dos formas diferentes:

1. Error Absoluto. El error absoluto es la diferencia absoluta que existe entre un valor real y un aproximado. Está dado por la siguiente fórmula:

$$E = \mid V_{Real} - V_{Aprox} \mid$$

   El error absoluto recibe este nombre ya que posee las mismas dimensiones que la variable bajo estudio.

2. Error relativo. El error relativo corresponde a la expresión en porcentaje de un error absoluto; en consecuencia, este error es adimesional.

$$e = \frac{\mid V_{Real} - V_{Aprox} \mid}{V_{Real}} x100\,\%$$

La diferencia entre la preferencia en el uso de los dos tipos de error consiste precisamente en la presencia de las dimensiones físicas. Debido a las unidades de medición utilizadas, el manejo y la percepción del error absoluto suele ser engañoso o difícil de comprender rapidamente. Sin embargo, el manejo de porcentajes (o valores relativos) resulta más natural y sencillo de comprender. Sin embargo, el uso de estos dos tipos de errores está sujeto siempre al objetivo de las actividades desarrolladas.

Consideraciones sobre el Valor Real (V_{Real})

La expresiones que definen a los errores absoluto y relativo requieren del conocimiento de la variable V_{Real} que representa un valor ideal que no posee error alguno. Como podrá suponerse, en la práctica resulta imposible determinar este valor.

Una práctica común en los análisis elementales sobre errores es considerar como un valor real a los resultados arrojados por la medición experimental de los fenómenos y a los valores aproximados como los proporcionados por los modelos matemáticos (o viceversa). El lector ha percibido que en ambos valores existe un error, por lo cual ninguno de ellos puede ser considerado como valor real. En realidad, ambos valores son valores aproximados.

Para lograr un resultado coherente, en la práctica debe sustituirse al valor real por un valor que se considere posee un error menor. Por ejemplo, en un proceso de mediciones suele utilizarse como valor real a los valores nominales citados en las especificaciones de los objetos a medir.

En el caso del análisis numérico, dado que los resultados se obtienen a partir de procesos iterativos que mejoran resultados inicialmente seleccionados, debe partirse del supuesto que el último valor obtenido posee un nivel menor de error que el valor previo. Dado lo anterior, los errores absoluto y relativo se calcularán de la siguiente forma:

Error absoluto:

$$E = \mid V_i - V_{i-1} \mid$$

Error relativo:

$$e = \frac{\mid V_i - V_{i-1} \mid}{V_i} x100\,\%$$

En ambos casos, V_i es el valor de la última iteración i y V_{i-1} es el valor de la iteración anterior $i-1$.

Magnitud de los errores por truncamiento y por redondeo

Lamentablemente, la literatura especializada sobre el tratamiento de errores es escasa y sin embargo resulta muy importante el poder 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 errores 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. 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 distancia absoluta, o error abosluto, 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. 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.

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

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 números 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.

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 %

3. Una expresión del tipo $a(b-c)$ puede reescribirse de la forma $ab-ac$ y $\frac{(a-b)}{c}$ puede reescribirse 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.

Referencias

[1] http://es.wikipedia.org. *Análisis numérico*. 2006.

[2] Lloyd Trefethen. The definition of numerical analysis. Bulletin of the Institute for Mathematics and Application, 1992.

[3] Mc Cracken. *Métodos numéricos y programación en Fortran con aplicaciones en ingeniería y ciencias*. México, 1967.