TEXTOS UNIVERSITARIOS
CIENCIAS

Métodos numéricos

Una introducción a las matemáticas del cálculo científico

René Escalante

UAH

Métodos numéricos

Una introducción a las matemáticas del cálculo científico

TEXTOS UNIVERSITARIOS
CIENCIAS

UAH

Plaza de San Diego, s/n
28801 Alcalá de Henares
www.uah.es

I.S.B.N.: 978-84-18979-91-0

Composición: Solana e Hijos, A. G., S.A.U.
Impresión y encuadernación: Solana e Hijos, A.G., S.A.U.
Impreso en España

Métodos numéricos

Una introducción a las matemáticas del cálculo científico

René Escalante

Universidad de Alcalá
EDITORIAL
UNIVERSIDAD DE ALCALÁ

Prefacio

Pretende ésta ser una obra de carácter introductorio, escrita en lengua castellana, sobre el relevante tema de las matemáticas numéricas y el cálculo científico, el cual abarca áreas tales como el cálculo y el análisis numérico, la optimización matemática y muchas otras relacionadas, como son el modelaje matemático y las diferentes estrategias utilizadas en la resolución práctica de problemas que involucran, todas ellas, la fusión de las matemáticas formales que conocemos con el uso exhaustivo de las máquinas de cálculo científico modernas.

Debido a la gran utilidad y amplia aplicabilidad de los métodos numéricos del cálculo científico en prácticamente todas las áreas de las matemáticas aplicadas, ingeniería y ciencias básicas en general, nos hemos propuesto en este libro describir y analizar los métodos más conocidos, pretendiendo preparar al lector para enfrentar resultados más especializados, muchos de los cuales se derivan de los expuestos aquí. El conocimiento básico requerido por el lector para usar este libro es el de estar familiarizado con algunos conceptos elementales del álgebra lineal y del cálculo infinitesimal.

En nuestro tratamiento de los diferentes temas buscamos mostrar un enfoque moderno y actualizado, no sólo respecto a la manera en que realizamos la exposición de estos, sino también respecto a la inclusión de algunos de los resultados del cálculo científico que han sido considerados recientemente por los especialistas de cada sub-área en sus trabajos académicos y de investigación, y que no necesariamente han sido abordados por obras publicadas en años anteriores.

Este libro se fue formando a partir de las numerosas notas de clase que fui acumulando en el curso de unos 30 años de docencia, tanto de pregrado como de postgrado, que mantuve en diferentes universidades como la Universidad Central de Venezuela (Caracas, Venezuela), la Universidad de Costa Rica (San José, Costa Rica), la Universidad Simón Bolívar (Caracas, Venezuela) y, más recientemente, la Universidad de Alcalá (Madrid, España). De tal forma que el libro puede ser utilizado tanto como un libro de texto por estudiantes de pregrado de nivel medio o avanzado, como por estudiantes de un curso de primer año de estudios de postgrado. Asimismo, persiguiendo claridad, concisión y completitud, consideramos que el texto contiene, en relativamente pocas páginas, todo el material necesario para su lectura individual, pudiendo ser utilizado también como un libro de texto en un curso tutorizado o como un libro de referencia por parte de investigadores que necesiten aplicar en su trabajo los elementos básicos del *cálculo científico.*

Mucha de la inspiración que me llevó a escribir este libro proviene de la lectura y estudio de las conocidas obras clásicas de Lanczos [44], Collatz [12], Isaacson y Keller [41], Luenberger [47], Dahlquist y Björck [15], Forsythe, Malcolm y Moler [29], Linz [45], Conte y De Boor [13], Atkinson [4], Kincaid y Cheney [43], [11], Stoer y Burlirsch [53], Heath [35], Quarteroni, Sacco y Saleri [52], entre las principales, y que mayor influencia tuvieron en mí durante mi formación académica y profesional.

Debido a que para la realización de este libro me basé principalmente en mis notas de clase, no recuerdo exactamente la fuente de algunos de los ejemplos y ejercicios mostrados en el manuscrito. Sin embargo, en compensación de ello, el lector encontrará citadas algunas de las referencias al comienzo de capítulos, secciones, ejemplos y ejercicios. A través de los ejercicios de cálculo computacional propuestos, hemos sugerido usar el ambiente de trabajo MATLAB®[1], el cual es fácil de aprender y usar; es además potente, exacto, robusto y rápido. Integra cálculo, visualización gráfica y programación en un ambiente abierto y flexible [26].

Si bien, a través del libro hemos usado una notación estándar, similar a la encontrada en los libros arriba citados, tan sólo mencionamos la inclusión en el manuscrito de algunos elementos simbólicos de uso frecuente. Siempre finalizaremos las demostraciones con el símbolo $\blacksquare$ y terminaremos los ejemplos y los ejercicios propuestos con los símbolos $\square$ y $\Diamond$, respectivamente. En el texto las dos palabras "ecuaciones diferenciales" se abreviarán como ED (EDs para el plural). Asimismo, usaremos EDOs para "ecuaciones diferenciales ordinarias", PVIs para "problemas de valores iniciales" y PVFs para "problemas con valores en la frontera". Además, con frecuencia usaremos SVD en lugar de "descomposición en valores singulares". Si a, b, c son escalares reales, usaremos la notación $a \ll b$ (o $b \gg a$) para "a es mucho menor que b" (o para "b es mucho mayor que a"), y $a \approx b$ para "a es aproximadamente igual a b" y significa que $|a - b| \ll c$, donde c se escoge de la manera más conveniente dependiendo del contexto (aunque, en general, no podemos decir, por ejemplo, que $0 \approx 10^{-7}$).

Aunque se ha querido mantener el libro tan corto como fuera posible, pero sin sacrificar claridad en la exposición. se ha incluido también algunos

[1]MATLAB es una marca registrada de The MathWorks, Inc., 24 Prime Park Way, Natick, MA 01760, Telf.: 001+508-647-7000, Fax: 001+508-647-7001.

ejercicios, tanto en el curso del desarrollo de la teoría, como al final de cada capítulo, donde ofrecemos una variedad de problemas para ser resueltos. Algunos de ellos están íntimamente relacionados con el desarrollo teórico del campo del análisis numérico, y otros están relacionados con los aspectos prácticos de los algoritmos descritos.

Reconocimientos.

Varios colegas han colaborado con el libro, bien sea a través de algunas discusiones o bien como resultado de relevantes comentarios sobre partes del texto de versiones anteriores. Más específicamente, vaya mi mayor agradecimiento a los Profesores Marcos Raydan, Débora Cores, Minaya Villasana y Saúl Buitrago, todos colegas del Departamento de Cómputo Científico y Estadística de la Universidad Simón Bolívar (Venezuela). Asimismo, quisiera expresar también un especial agradecimiento a los Profesores Juan Gerardo Alcázar Arribas, Marcos Marvá Ruiz y José Javier Martínez Fernández de las Heras del Departamento de Física y Matemáticas de la Universidad de Alcalá (España).

R. E.
Alcalá de Henares, 2022.

Contenido

Capítulo 1

Conceptos básicos

1.1 Métodos numéricos

Debemos primero hablar de dos conceptos fundamentales, como son el de **problema numérico** y el de **algoritmo**. El primero se refiere a una clara y no ambigüa descripción de la conexión funcional entre un conjunto de datos de entrada (las variables independientes) y un conjunto de datos de salida (las variables dependientes). Un algoritmo, para un determinado problema numérico, es una descripción completa de operaciones bien definidas, a través de las cuales cada conjunto de datos de entrada admisible es transformado en datos de salida. De manera que, un determinado problema numérico puede involucrar muchos algoritmos diferentes, cada uno de los cuales puede dar lugar a una respuesta aproximada con una precisión diferente.

Un **método numérico** es un procedimiento mediante el cual buscamos resolver, de manera aproximada, un problema matemático dado a través de la consideración de un problema numérico (por ejemplo, el conocido **método de Newton-Raphson** es un método numérico usado en el cálculo aproximado de un cero de una ecuación no lineal, ver Capítulo 3). Este concepto incluye el caso en que el problema matemático sea transformado en un problema más simple que, posiblemente implique, a posteriori, la formulación de otro problema numérico asociado (por ejemplo, la transformación de un problema que involucra una ecuación diferencial en un sistema de ecuaciones).

1.2 Tipos de errores y cotas

La capacidad de representar al conjunto de los números reales por una computadora es realmente limitada, ya que tan sólo contamos con una cantidad finita de números para representar a un conjunto de cardinalidad infinita (no numerable) de elementos. De allí que exista la necesidad de contar con conceptos, medios y estrategias que nos permitan la estimación numérica de los errores involucrados[1]. Con esta idea "en mente", comenzamos por introducir la siguiente definición.

Definición 1.1. *Si x es el valor exacto* (en la generalidad de los casos no conocido) y $\tilde{x}$ es su valor aproximado, definimos el **error absoluto** por $|x-\tilde{x}|$ y el **error relativo** por $\left|\frac{x-\tilde{x}}{x}\right|$, siempre que $x \neq 0$.

Para los problemas en los cuales la magnitud del valor verdadero puede ser muy grande, o muy pequeña, el error relativo puede ser más útil que el error absoluto.

Ejemplo 1:

i) Si $x = 0.2000 \times 10^1$ y $\tilde{x} = 0.2100 \times 10^1$, el error absoluto es 0.1000 y el error relativo es 0.5000×10^{-1}.

ii) Si $x = 0.2000 \times 10^{-3}$ y $\tilde{x} = 0.2100 \times 10^{-3}$, el error absoluto es 0.1×10^{-4} y el error relativo es 0.5000×10^{-1}.

iii) Si $x = 0.2000 \times 10^4$ y $\tilde{x} = 0.2100 \times 10^4$, el error absoluto es 0.1×10^3 y el error relativo es otra vez 0.5000×10^{-1}. $\square$

Observamos que el error relativo en i), ii) y iii) es el mismo aún cuando los errores absolutos son diferentes. En este sentido, el error relativo es más significativo, y puede pensarse en términos porcentuales; de manera que, en este ejemplo, podemos decir que el error es de 5%. Por ejemplo, si tenemos dos doblones y perdemos uno de ellos, parece que estamos perdiendo mucho (de hecho, estamos perdiendo el 50% de nuestro capital), pero si tenemos un millón de doblones y perdemos uno, la pérdida no parece ser muy importante (relativa a la cantidad de doblones inicial). En este último caso el error

[1] Existe un número importante de referencias sobre el tema de los errores. En general, la mayoría de los textos introductorios de cálculo o análisis numérico resultan ser adecuados para quien comienza a adentrarse en esta área (ver, por ejemplo, [7] y otras referencias al final del texto).

relativo es de 10^{-6} cuando en el primero es de 0.5, pero en ambos casos el error absoluto es de 1.

Los computadores representan los números reales en una forma denominada de **punto flotante**, que es semejante a la notación científica. Por ejemplo, un número N se almacena como $N = \pm .d_1 d_2 d_3 \ldots d_p \beta^e$, donde β es la **base** y las $d_i s$ son los **dígitos**. Para un computador, la base es usualmente 2, 8 o 16. Cada dígito es un número entero entre 0 y $\beta - 1$. Hay un número fijo de dígitos, la **precisión** p, y un **exponente** entero e, el cual está restringido a un **rango** de valores ($e \in [e_{min}, e_{max}]$). Si $d_1 \neq 0$, el sistema se denomina **sistema en punto flotante normalizado**.

En 1985 el *Institute for Electrical and Electronic Engineers* (IEEE), publicó el informe *Binary Floating Point Arithmetic Standar 754-1985*, en donde se establecieron los formatos para las precisiones simples, dobles y extendidas. Estos son los estándares utilizados por los fabricantes de PC's para el *hardware* de punto flotante.

El procesador numérico de las PC's suele utilizar una representación de 64 bits (dígitos binarios) para un número real. El primer bit representa el signo, s. Los siguientes 11 bits representan al **exponente** e y los 52 bits que siguen es una fracción en binario, f, llamada **mantisa**. La base para el exponente es 2. La **precisión** con la cual los números pueden almacenarse y los cálculos llevarse a cabo, dependen del número de dígitos de la mantisa y del rango del exponente usados para representar un número real. La condición finita de f es una limitación sobre la precisión p. Asimismo, la condición finita de e es una limitación sobre el rango.

Como 52 dígitos binarios corresponden a unos 16 o 17 dígitos decimales, supondremos que un número representado en este sistema tiene al menos 16 cifras decimales de precisión. El exponente de 11 dígitos representa un intervalo del 0 al $2^{11} - 1 = 2047$. A fin de que estos números se puedan representar, se resta 1023 del exponente, de manera que el intervalo del exponente sería en realidad $[-1023, 1024]$. Con el objeto de proporcionar una representación única de cada número en punto flotante, supondremos que la representación está normalizada. El uso de este sistema proporciona un número en punto flotante de la forma[2]

$$(-1)^{(s)_2} \times 2^{(e)_2 - 1023} \times (1.f)_2.$$

[2]En esta representación normalizada de un número en punto flotante distinto de cero, el primer bit en la mantisa es siempre 1, por lo que el mismo no se tiene que almacenar.

Por ejemplo, consideremos el número de máquina

$$(0110000000011011100100001000000000000000000000000000000000000000)_2.$$

Observemos que el primer bit de la izquierda es 0, lo cual indica que el número es positivo. Los siguientes 11 bits corresponden al exponente, y son equivalentes al número decimal

$$(e)_2 = 1.2^{10} + 1.2^9 + 0.2^8 + \ldots + 0.2^1 + 1.2^0 = 1537$$

De manera que, la parte exponencial del número es $2^{1027-1023} = 2^{514}$. También, los últimos 52 bits muestran que la mantisa es

$$(f)_2 = 1.\Big(\frac{1}{2}\Big)^1 + 1.\Big(\frac{1}{2}\Big)^3 + 1.\Big(\frac{1}{2}\Big)^4 + 1.\Big(\frac{1}{2}\Big)^5 + 1.\Big(\frac{1}{2}\Big)^8 + 1.\Big(\frac{1}{2}\Big)^{13}.$$

Por lo tanto, este número de máquina representa al número decimal

$$\begin{aligned}(-1)^{(s)_2} \times 2^{(e)_2-1023}(1.f)_2 &= (-1)^0.2^{514}\Big(1 + \Big(\frac{1}{2} + \frac{1}{8} + \\ &\qquad \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{8192}\Big)\Big) \\ &= (9.239472329847650)10^{154}.\end{aligned}$$

Es interesante observar que el mayor número de máquina menor que el anterior es

$$(0110000000011011100100000111111111111111111111111111111111111111)_2,$$

y el siguiente número de máquina mayor es

$$(0110000000011011100100001000000000000000000000000000000000000001)_2.$$

Esto significa que nuestro número de máquina original representa también a la mitad de los números reales que están entre él y los dos números de máquina mas próximos.

El menor número positivo normalizado que puede representarse tiene ceros en todas partes excepto en el bit del extremo derecho (que es un 1), y es igual a

$$2^{-1023}.(1 + 2^{-52}) \approx 10^{-308},$$

y el mayor tiene un cero al principio seguido de unos, y es equivalente a

$$2^{1024}.(1 + 1 - 2^{-52}) \approx 10^{308}.$$

Los números que aparecen en los cálculos y tienen una magnitud menor que 10^{-308} producen un *underflow* (por lo que se les asigna un valor de 0). Los números mayores que 10^{308} producen un *overflow* (lo cual interrumpe los cálculos).

Asimismo, el número total de valores que se pueden representar con $d_1 \neq 0$ está dado por

$$2(\beta - 1)(\beta^{\,p-1})(\text{nro. total de exponentes}) + 1.$$

(¿Por qué?)

Ejercicio 1:

a. ¿Cuántos números podemos representar de manera exacta en un sistema en base 10, con dos dígitos ($p = 2$) y exponentes 0 o 1? ¿Cuáles son los valores máximo y mínimo que en este sistema podemos representar? ¿Cuáles son los números positivos que pueden representarse con el exponente 1 y cuáles con el exponente 0? Entonces, ¿cuántos valores existen en el intervalo $[0, 0.99]$ y cuántos en el intervalo $[1.0, 1.9]$? $\Diamond$

b. En un sistema de punto flotante binario, los decimales corresponden a sumas de potencias negativas de 2. Por ejemplo,

$$(0.10)_2 = (1)\frac{1}{2} + (0)\frac{1}{4} = \frac{1}{2} \quad \text{y} \quad 0.11_2 = (1)\frac{1}{2} + (1)\frac{1}{4} = \frac{3}{4}.$$

¿Cuáles son los números positivos que pueden representarse en un sistema en punto flotante normalizado binario con dos dígitos y un exponente igual a 0? (la normalización exige que el primer dígito en el desarrollo sea 1). Escriba los números positivos con dos dígitos y exponentes -1 y 1. $\Diamond$

Nota: Por razones de claridad en la exposición, supondremos de ahora en adelante que los números de máquina se representan en la forma de punto flotante **decimal** normalizado (es decir, $\pm 0.d_1 d_2 \ldots d_p \times 10^n$, $1 \leq d_1 \leq 9$ y $0 \leq d_i \leq 9$, para cada $i = 2, 3, \ldots, p$).

1.2.1 Errores por redondeo

Existen dos formas de expresar un número que tenga más dígitos de los que pueden representarse por el sistema en punto flotante. La más simple consiste en truncar el número y desechar los dígitos que el sistema no puede manejar. El segundo método es el de redondear el número; el resultado depende del valor del primer dígito del grupo de dígitos que van a descartarse. Si el sistema permite trabajar con tan sólo n dígitos, y si el dígito $n+1$ está entre 0 y 4 (ambos inclusive), el redondeo produce el mismo resultado que el truncamiento. Si el dígito $n+1$ está entre 5 y 9, el dígito n-ésimo se incrementa en 1. Las imprecisiones que resultan de redondear o de truncar se conocen como **errores por redondeo**. En general, el redondeo acumula menos errores (durante los sucesivos cálculos) que los producidos por truncamiento, ya que el valor verdadero es mayor (o menor) que el valor redondeado alrededor de la mitad de las veces. Además, para la operación de truncamiento, el mayor error absoluto que pudiera producirse es del doble del de redondeo. Claro que el truncamiento no requiere de decisión alguna acerca de cambiar o no el último dígito retenido.

Ejemplo 2:
Consideremos la siguiente suma: $0.99 + 0.0044 + 0.0042$. Con aritmética exacta, el resultado es 0.9986. Sin embargo, si usamos una aritmética de tres dígitos, y las operaciones se realizan siguiendo el orden de izquierda a derecha, encontramos que $(0.99+0.0044)+0.0042 = 0.994+0.0042 = 0.998$. Por otra parte, si operamos primero los dos últimos números, tenemos que: $0.99+(0.0044+0.0042) = 0.99+0.0086 = 0.999$, lo cual demuestra el efecto del error por redondeo en un caso tan simple como éste. (¿Por qué sucede esto?) $\square$

Nota: Desde el punto de vista numérico es importante el orden en que sumamos.

Ejercicio 2:
Considere la suma de términos $S = 9.87+0.78+0.05+0.01$. Realice la suma de izquierda a derecha y luego de derecha a izquierda, en una máquina que tan sólo maneja una aritmética de tres cifras con redondeo. ¿Son iguales los resultados de realizar estas dos operaciones? Explique. $\diamondsuit$

Además de la representación imprecisa de los números en la computadora, la aritmética involucrada no es exacta. Supongamos que $fl(x)$ y $fl(y)$ son las representaciones en punto flotante de los números reales x y y, y que $\oplus$, $\ominus$, $\otimes$ y $\oslash$ representan las operaciones de suma, resta, multiplicación y división en la máquina, respectivamente. De manera que, la aritmética con un número finito de cifras estará dada por:

$$x \oplus y = fl(fl(x) + fl(y)), \quad x \ominus y = fl(fl(x) - fl(y)),$$

$$x \otimes y = fl(fl(x) \times fl(y)), \quad x \oslash y = fl(fl(x) \div fl(y))$$

Lo cual corresponde a realizar la aritmética exacta con las representaciones de x y y en punto flotante, para después transformar el resultado a su representación en punto flotante con un número finito de cifras.

Ejemplo 3:
Supongamos que $x = \frac{5}{7}$ y $y = \frac{1}{3}$ y que usamos truncamiento de cinco cifras para los cálculos. Así, $fl(x) = 0.71428 \times 10^0$ y $fl(y) = 0.33333 \times 10^0$. Algunos resultados numéricos son: $x \oplus y = 0.10476 \times 10^1$, $x \ominus y = 0.38095 \times 10^0$, $x \otimes y = 0.23809 \times 10^0$ y $x \oslash y = 0.21428 \times 10^1$. $\square$

Ejercicio 3:

a) Calcule los correspondientes errores absolutos y relativos en el Ejemplo 3 anterior. ¿Produce la aritmética seleccionada resultados satisfactorios? $\Diamond$

b) A los valores del Ejemplo 3 añadimos los valores $u = 0.714251$, $v = 98765.9$ y $w = 0.111111 \times 10^{-4}$, de manera que $fl(u) = 0.71425 \times 10^0$, $fl(v) = 0.98765 \times 10^5$ y $fl(w) = 0.11111 \times 10^{-4}$. Calcular $x \ominus u$, $(x \ominus u) \oslash w$, $(x \ominus u) \otimes v$ y $u \oplus v$, así como sus correspondientes errores absolutos y relativos. ¿Produce en este caso la aritmética seleccionada resultados satisfactorios? $\Diamond$

c) Repetir el Ejercicio (partes a) y b)) usando una aritmética de cinco cifras con redondeo. $\Diamond$

Otro concepto importante es el que se refiere a los dígitos de un núnero dado que empiezan con el dígito distinto de cero del extremo izquierdo y terminan con el dígito correcto del extremo derecho, incluyendo los ceros finales que son exactos.

Definición 1.2. *Decimos que el número $\tilde{x}$ aproxima a x con t* **cifras significativas** (*o* **dígitos significativos**) *si t es el mayor entero no negativo para el que*

$$\left| \frac{\tilde{x} - x}{x} \right| < 0.5 \times 10^{-t}.$$

Ejemplo 4:
Usando esta definición, vemos que 0.998 aproxima a la solución verdadera, $x = 0.9986$, con dos dígitos significativos, pues

$$\left| \frac{0.998 - 0.9986}{0.9986} \right| = 6.0084 \times 10^{-4} < 0.5 \times 10^{-2}.$$

Por otra parte, 0.999 aproxima la solución verdadera con tres dígitos significativos, pues

$$\left| \frac{0.999 - 0.9986}{0.9986} \right| = 4.0056 \times 10^{-4} < 0.5 \times 10^{-3}. \quad \square$$

Ejercicio 4:
Si $x = 0.097821$ y $\tilde{x} = 0.097437$, ¿con cuántas cifras significativas $\tilde{x}$ aproxima a x? $\Diamond$

1.2.2 Errores por cancelación

Otro ejemplo del efecto de cálculos inexactos se produce cuando el cálculo implica la substracción de dos números muy próximos entre sí. Con el objeto de evitar esta dificultad se recomienda buscar una reformulación del algoritmo inicialmente considerado.

Ejemplo 5:
Consideremos el problema de usar la fórmula cuadrática para resolver la ecuación de segundo grado

$$x^2 + bx + 1 = 0.$$

El efecto de redondear el discriminante $\Delta = \sqrt{b^2 - 4}$ se ilustra en este ejemplo. Notemos que para b grande ($b \gg 4$), Δ es bastante próxima a b. La fórmula cuadrática da $x_1 = \frac{-b+\Delta}{2}$ y $x_2 = \frac{-b-\Delta}{2}$. Si b es positivo el cálculo de x_1 involucra la diferencia de dos números muy próximos entre sí, lo cual representa una situación "catastrófica".

Por ejemplo, consideremos $x^2 + 70x + 1 = 0$, cuyas raíces tienen los valores

$$x_1 = -0.01428863\ldots \quad \text{y} \quad x_2 = -69.98571\ldots$$