Máster en Materiales y Sistemas Sensores
para Tecnologías Medioambientales
(Erasmus Mundus)

NOTAS DE CÁLCULO NUMÉRICO

Damián Ginestar Peiró

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
DEL DISEÑO

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Capítulo 1

Errores y aritmética finita

1.1. Introducción

A la hora de realizar un determinado cálculo con un ordenador se pueden tener distintas fuentes de error:

- Errores de truncamiento.
- Error en los datos.
- Errores asociados a las aproximaciones numéricas utilizadas.

En este tema estudiaremos el origen de los errores de truncamiento derivados de utilizar una representación finita para los números y cómo se propagan los errores en los datos, que pueden deberse a los errores de truncamiento o a posibles errores experimentales, al realizar distintas operaciones aritméticas. Los errores derivados de utilizar una cierta aproximación numérica para realizar un cierto cálculo, los iremos estudiando a lo largo del curso.

1.2. Representación de los números en un ordenador

Por cuestiones de carácter técnico, los ordenadores utilizan una representación binaria de los números, donde los únicos dígitos posibles son el 0 o el 1, a diferencia de la representación decimal, o en base 10, que es la que utilizamos normalmente, donde se utilizan los dígitos del 0 al 9.

Dado un número p, este número en base b se escribe

$$(p)_b = B_n b^n + B_{n-1} b^{n-1} + \cdots + B_0 + B_{-1} b^{-1} + B_{-2} b^{-2} + \cdots \ .$$

Así, el número 32.5 en base 10 representa

$$(32{,}5)_{10} = 3\,10 + 2 + 5\,10^{-1}$$

y el número 10.01 en base 2 representa

$$(10{,}01)_2 = 1\,2 + 0 + 0\,2^{-1} + 1\,2^{-2} \ ,$$

o sea, es el número 2.25 en base 10.

Los ordenadores utilizan una representación de los números denominada *en coma flotante*. Esta representación utiliza un conjunto de dígitos comenzando por 0., denominado *mantisa* y un *exponente*. Por ejemplo, el número 32.5 en base 10, se puede expresar como

$$0{,}325\,10^2 \ .$$

En este caso la mantisa sería 325 y el exponente 2. De igual forma para el número 10.01 en base 2, tendríamos la representación

$$0{,}1001\,2^2 \ ,$$

con lo que la mantisa sería 1001 y el exponente 2.

Los ordenadores generalmente para representar un número real utilizan una posición de memoria (bit) para representar el signo, luego un cierto número de posiciones de memoria para representar el exponente y otro número de posiciones de memoria para representar la mantisa. La representación concreta cambia de unos ordenadores a otros. Por ejemplo, para un ordenador IBM cada posición de memoria está formada por 32 bits. Un número real p, en simple precisión se almacena de la forma siguiente: Una vez escrito p como

$$p = \pm p' \, 16^{(p''-64)} \ ,$$

con $p'' \geq 0$ y $\frac{1}{16} \leq p' < 1$. Se almacena primero el signo (0 si es + y 1 si es -), en los 7 bits siguientes se almacena el exponente p'' en base 2. En las 24 posiciones restantes se almacena la mantisa en base 2. Si el número está en doble precisión, se usan 64 bits para almacenarlo y para la mantisa se usan 56 bits.

Veamos un ejemplo, si queremos representar el número 1.1 en base 10. Lo primero que haremos es representar el número $(1{,}1)_{10}$ en base 2. La parte

entera del número es 1 con lo que queda de la misma forma. La parte decimal es 0.1 que podemos reescribir como

$$\begin{aligned} 0{,}1 &= 2^{-1}\,(0{,}2) = 2^{-4}\,(1{,}6) \\ &= 2^{-4}\,(1+0{,}6) = 2^{-4}+2^{-5}\,(1{,}2) \\ &= 2^{-4}+2^{-5}+2^{-8}1{,}6 = \cdots\ , \end{aligned}$$

así,

$$(0{,}1)_{10} = 0{,}0001100110011\cdots\ .$$

Por otra parte,

$$\begin{aligned} 1{,}1 &= \frac{1{,}1}{16}16 = 1{,}0\overline{0011}\,2^{-4}\,16^{(65-64)} \\ &= 0{,}00010\overline{0011}\,16^{(65-64)}\ . \end{aligned}$$

El exponente es $65 = (64+1) = 2^6+1$. Por lo tanto, el exponente en base 2 es

$$(65)_2 = 1000001\ ,$$

y, por tanto, la representación del número 1.1 en simple precisión es de la forma

0	1000001	00010001100110011001

Con este tipo de representación el número más alto que es posible obtener es

0	1111111	11111111111111111111111

que corresponde a

$$\left(2^{-1}+2^{-2}+\cdots+2^{-24}\right)16^{(127-64)} \approx 16^{23} \approx 7\,10^{75}\ .$$

Por otra parte, el número más pequeño que podemos representar es

$$\frac{1}{16}\,16^{-64} = 16^{-65} \approx 6\,10^{-79}\ ,$$

cuya representación es

0	0000000	00010000000000000000000

El hecho que se disponga de un número finito para representar los números reales hace que se produzcan errores al hacer operaciones aritméticas con los mismos. Estos errores se denominan *errores de truncamiento* o *errores de redondeo* dependiendo de la técnica que se utilice para cortar la representación de los números.

Veamos algunos ejemplos que ilustran los errores que se pueden producir al utilizar aritmética finita. Supongamos que se tiene un ordenador ideal que puede representar los números en base 10, pero utiliza una mantisa de dos números. En este ordenador la representación del número $x = 15\pi/34 \approx 1{,}38$ es $x^* = 0{,}14\,10^1$, con lo que se comete un error de un 1 %. Si ahora consideramos $y = \sqrt{2} \approx 1{,}428\ldots$, la representación de este número es $y^* = 0{,}14\ ,10^1$ y la resta de estos números será

$$x^* - y^* = 0\ ,$$

con lo que cometemos un error

$$|x - y| = 0{,}0282168034\ ,$$

que corresponde a un error del 100 %.

En general, al restar dos números reales próximos entre sí, x e y, se obtiene una diferencia $x^* - y^*$, que aproxima mal la verdadera diferencia $x - y$.

Otra operación en la que se pueden cometer errores al utilizar aritmética finita es la división de un número x por un número pequeño δ. Supongamos que x se aproxima por $x^* = x+\varepsilon$. Al dividir por δ, $x^*/\delta = x/\delta+\varepsilon/\delta$. El nuevo error es ε/δ. Si δ es un número pequeño, el error cometido será una cantidad grande. Un razonamiento similar se puede hacer cuando se multiplica un número x por otro número muy grande.

En ocasiones el error de redondeo que se comete al realizar un cálculo se puede minimizar reformulando el problema. Por ejemplo, supongamos que queremos calcular las raíces de la ecuación

$$ax^2 + bx + c = 0\ , \quad a \neq 0\ .$$

Estas raíces vienen dadas por las expresiones

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}\ , \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}\ .$$

Consideremos la ecuación

$$x^2 + 62{,}10x + 1 = 0\ ,$$

cuyas soluciones aproximadas son

$$x_1 = -0{,}01610723 \ , \quad x_2 = -62{,}08390 \ .$$

Supongamos que se calcula x_1 con una aritmética de 4 dígitos

$$\begin{aligned} \sqrt{b^2 - 4ac} &= \sqrt{(62{,}10)^2 - 4{,}000} = \sqrt{3856 - 4{,}000} = \\ &= \sqrt{3852} = 62{,}06 \ , \end{aligned}$$

así,

$$\begin{aligned} x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-62{,}10 + 62{,}06}{2{,}000} = \\ &= \frac{-0{,}04000}{2{,}000} = -0{,}02000 \ , \end{aligned}$$

que es una aproximación pobre de $x_1 = -0{,}01611$. Por otro lado, el cálculo de x_2 involucra la adición de números casi iguales y no presenta problemas,

$$\begin{aligned} x_2 &= \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{-62{,}10 - 62{,}06}{2{,}000} \\ &= \frac{-124{,}2}{2{,}000} = -62{,}10 \ . \end{aligned}$$

Para obtener una mejor aproximación de x_1 tenemos en cuenta que

$$\begin{aligned} x_1 &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{b^2 - (b^2 - 4ac)}{2a\left(-b - \sqrt{b^2 - 4ac}\right)} \\ &= \frac{-2c}{b + \sqrt{b^2 - 4ac}} = \frac{-2{,}000}{62{,}10 + 62{,}06} = -0{,}0161 \ . \end{aligned}$$

1.3. Propagación de errores

Supongamos que se quiere dar una cota al error que se comete al realizar un cálculo $y = f(x)$, si la variable x viene afectada de un error dado. Así, si $x = x^* + \Delta x$, y $f(x)$ es una función derivable, podemos escribir

$$y = f(x) = f\left(x^* + \Delta x\right) \approx f\left(x^*\right) + f'\left(x^*\right) \Delta x \ ,$$

con lo que el error

$$\varepsilon f(x^*) = \left|y - f\left(x^*\right)\right| \approx \left|f'\left(x^*\right)\right| \left|\Delta x\right| \ .$$

Si se tiene un problema multidimensional $y = f(x_1, x_2, \ldots, x_n)$, podemos escribir

$$\begin{aligned} |y - f(x_1^*, x_2^*, \ldots, x_n^*)| &\approx \left| \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \ldots, x_n^*)\, \Delta x_i \right| \\ &\leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \ldots, x_n^*) \right| |\Delta x_i| \ . \end{aligned}$$

Así una expresión para una cota del error cometido es de la forma

$$\varepsilon f(x_1^*, x_2^*, \ldots, x_n^*) = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \ldots, x_n^*) \right| |\Delta x_i| \ ,$$

que suele funcionar bien cuando n no es muy grande. En el caso que se tengan muchas variables esta expresión proporciona una cota muy conservadora del error cometido.

Otra posibilidad para definir un error es utilizar el *error cudrático medio*

$$\varepsilon f(x_1^*, \ldots, x_n^*) = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}(x_1^*, \ldots, x_n^*)\, \Delta x_i \right)^2} \ .$$

1.4. Ejercicios

1. Escribir en base 2 los números:

   $$\text{a)}\ \ 1.5\ , \quad \text{b)}\ 1.4\ , \quad \text{c)}\ 0.3\ \ .$$

2. Escribir la representación en un ordenador IBM en simple precisión de $x = -232$.

3. Escribir en base 2 el número cuya representación en base 10 es 100.3. Utilizar este resultado para obtener la representación en simple precisión de este número en un ordenador IBM.

4. Dado el sistema de ecuaciones

   $$\begin{aligned} 3x + ay &= 10\ , \\ 5x + by &= 20\ , \end{aligned}$$

   donde $a = 2{,}100 \pm 5\,10^{-4}$ y $b = 3{,}300 \pm 5\,10^{-4}$. Obtener una cota del error asociado a x e y al resolver el sistema.

5. Con qué exactitud se ha de medir el radio de una esfera y con cuántos decimales se ha de dar el número π, para que su volumen se conozca con un error realivo menor que el $0.01\,\%$.

6. Supongamos que se quiere hacer el cálculo

$$y = \ln\left(x - \sqrt{x^2 - 1}\right) \ ,$$

para $x = 30$, y que al calcular la raíz cuadrada tenemos un método que sólo nos da 6 decimales correctos. Estimad una cota del error que se comete al hacer el cálculo. Reescribir la función $y(x)$ para que el error cometido al hacer el cálculo sea menor.

7. Consideremos una barra de longitud l y sección rectangular, de anchura a y altura b, empotrada por uno de sus extremos. Si en el extremo libre se aplica una fuerza F perpendicular a la barra, la flexión s experimentada viene dada por la expresión

$$s = \frac{4l^3}{Eab^3}F \ ,$$

donde E es el módulo de Young del material que forma la barra. De un ensayo, se sabe que una fuerza F de 140 kp aplicada a una barra de hierro de 125 cm de longitud y sección cuadrada de 2.4 cm de lado le produce una flexión de 1.71 mm. Calculad el módulo de Young del hierro y una cota del error asociado, si se supone que las medidas tomadas tienen la precisión asociada con la última cifra decimal.

8. Calcular el error permitido en la inclinación de un cañón para asegurar que acierte en un objetivo rectangular de 40 m de longitud y 20 m de anchura, situado a la misma altura que el cañón, y cuyo centro está a una distancia de 3000 m de éste. La velocidad de salida del proyectil es de 600 m/s. Supóngase que todas las magnitudes son exactas y que las dimensiones del objetivo son pequeñas en relación a la distancia a la que se encuentra el cañón. Considérese también que se apunta al centro del objetivo y que está permitido tocar en cualquier lugar de éste.