Capítulo 1

Errores. Algoritmos: convergencia y estabilidad

1.1. Errores

1.1.1. Introducción

El análisis de los errores en los procedimientos numéricos es fundamental. Los datos de entrada rara vez son exactos debido, entre otros aspectos, a los siguientes:

1. *Basados en experimentos o estimados:* En los experimentos influyen factores aleatorios, bien porque las condiciones varían, bien porque no se consideran todos los factores que afectan al experimento.

2. *Precisión de los aparatos de medida:* Los aparatos utilizados tienen una precisión de medida determinada y finita.

3. *Proceso numérico:* Los métodos empleados y los procesos numéricos también introducen a su vez errores de varios tipos.

A continuación se exponen algunos ejemplo que nos indican la importancia del análisis de los errores en los procesos numéricos.

Ejemplo 1.1 *Supongamos que estamos buscando las raices de la ecuación*

$$x^2 + 0{,}4002 \cdot x + 0{,}00008 = 0,$$

para ello utilizamos la conocida fórmula de segundo grado

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}.$$

Si utilizamos 4 cifras decimales con redondeo por exceso, entonces

$$\begin{aligned} x &= \frac{-0{,}4002 \pm \sqrt{(0{,}4002)^2 - 4 \cdot 1 \cdot 0{,}00008}}{2} \\ &= \frac{-0{,}4002 \pm \sqrt{0{,}1602 - 0{,}0003}}{2} \\ &= \frac{-0{,}4002 \pm \sqrt{0{,}1599}}{2} \end{aligned}$$

obteniéndose los siguientes valores para las raíces

$$x_1^{(4)} = \frac{-0{,}4002 + 0{,}3999}{2} = -0{,}00015$$

y

$$x_2^{(4)} = \frac{-0{,}4002 - 0{,}3999}{2} = -0{,}40005$$

Si ahora utilizamos 8 de cifras decimales el resultado es el siguiente:

$$\begin{aligned} x &= \frac{-0{,}40020000 \pm \sqrt{(0{,}40020000)^2 - 4 \cdot 1 \cdot 0{,}00008000}}{2} \\ &= \frac{-0{,}40020000 \pm \sqrt{0{,}16010004 - 0{,}00032000}}{2} \\ &= \frac{-0{,}40020000 \pm \sqrt{0{,}15984004}}{2} \end{aligned}$$

siendo en este caso los valores obtenidos para las raíces

$$x_1^{(8)} = \frac{-0{,}40020000 + 0{,}39980000}{2} = -0{,}0002$$

y

$$x_2^{(8)} = \frac{-0{,}40020000 - 0{,}39980000}{2} = -0{,}4$$

El cálculo con 4 dígitos produce un error del $25\,\%$ *en el valor de* x_1 *que no parece el más adecuado*

$$\frac{x_1^{(8)} - x_1^{(4)}}{x_1^{(8)}} = \frac{(-0{,}0002) - (-0{,}00015)}{-0{,}0002} = 0{,}25$$

Este error aparece debido a la representación finita de los datos en un ordenador.

Ejemplo 1.2 *Otro ejemplo de la importancia de los errores en los procesos numéricos es el siguiente. Si consideramos la serie de Taylor de la función* $\operatorname{sen}(x)$

$$\operatorname{sen}(x) = x - \frac{1}{6}x^3 + \frac{1}{120}x^5 + O\left(x^7\right)$$

esta puede considerarse válida para cualquier ángulo finito cercano a 0 *y siendo el error de truncamiento menor en valor absoluto que el primer término despreciado. Sin embargo, no podemos emplear esta fórmula para calcular el seno de un ángulo grande, por ejemplo si tomamos* $x = 9$ *radianes, al sustituir en el polinomio de Taylor anterior*

$$9 - \frac{1}{6}9^3 + \frac{1}{120}9^5 = 9 - 121{,}5 + 492{,}075 = 379,575$$

que es un resultado absurdo puesto que el valor del seno no puede ser nunca superior a 1*. En esta ocasión, aunque aumentemos el número de dígitos de precisión, el resultado seguirá siendo incorrecto. Este error se debe al truncamiento de una serie infinita, cambiando dicha serie por una cantidad finita de términos.*

Ejemplo 1.3 *Otros errores aparecen debido a las particularidades del propio problema. Por ejemplo el sistema lineal*

$$\left.\begin{aligned} 5x - 331y &= 3{,}5 \\ 6x - 397y &= 5{,}2 \end{aligned}\right\}$$

tiene como solución exacta

$$\begin{aligned} x &= 331{,}7 \\ y &= 5{,}000 \end{aligned}$$

En primer lugar, si cambiamos el valor 5,2 *del término independiente de la segunda ecuación por* 5,1*, es decir, aplicamos una variación del* 2 % *en dicho coeficiente, la nueva solución sería*

$$\begin{aligned} x &= 298{,}6 \\ y &= 2{,}5 \end{aligned}$$

que constituyen variaciones del 10 % *y* 50 % *en las variables* x *e* y*, respectivamented respecto de la solución anterior, lo cual, sobre todo en el caso del valor de* y*, parece una variación excesiva (*25 *veces mayor que la variación en el término). Luego pequeños cambios en el problema, producen grandes variaciones en la solución.*

Por otra parte, si ahora sustituimos los siguientes valores

$$\begin{aligned} x &= 358{,}173 \\ y &= 5{,}4 \end{aligned}$$

en el sistema original, se obtiene que el valor redondeado de los primeros miembros es exactamente igual a los segundos miembros

$$\begin{aligned} 5 * 358{,}173 - 331 * 5{,}4 &= 1790{,}865 - 1787{,}4 = 3{,}465 \approx 3{,}5 \\ 6 * 358{,}173 - 397 * 5{,}4 &= 2149{,}038 - 2143{,}8 = 5{,}238 \approx 5{,}2 \end{aligned}$$

por tanto estos valores podrían considerarse las soluciones buscadas; sin embargo, como se aprecia, difieren bastante de la solución exacta del problema

$$\Delta x = (358{,}173 - 331{,}7) = 26{,}473 \Rightarrow 7{,}98\,\% \text{ de variación}$$

$$\Delta y = (5{,}4 - 5{,}0) = 0{,}4 \Rightarrow 8\,\% \text{ de variación}$$

Estos problemas no aparecen debido a la aritmética utilizada, en este caso el determinante del sistema es muy pequeño, debido a que las rectas son casi paralelas y este hecho influye de forma notable en la obtención de las soluciones.

Ejemplo 1.4 *Como ejemplo final de la influencia de los errores en los procesos numéricos, consideremos la integral*

$$\int_{e^{-4}}^{1} \frac{dx}{x}$$

que se calcula de forma inmedianta integrando directamente y usando la regla de Barrow

$$\int_{e^{-4}}^{1} \frac{dx}{x} = \ln x|_{e^{-4}}^{1} = \ln 1 - \ln e^{-4} = -\ln e^{-4} = 4.$$

Sin embargo, utilizando por ejemplo una fórmula de integración numérica trapezoidal para obtener una aproximación $\widehat{I}$ *de esta integral, obtendremos, según el número de intervalos utilizados la siguiente tabla*

10 *intervalos*	$\widehat{I} = 5{,}3064$
40 *intervalos*	$\widehat{I} = 4{,}1327$
100 *intervalos*	$\widehat{I} = 4{,}0002$

Necesitamos muchos intervalos y por tanto muchos cálculos para obtener el valor de la integral. En este caso el problema deriva de la naturaleza de la función $\frac{1}{x}$ *del integrando, que toma valores muy grandes, para valores pequeños de* x *dentro del intervalo de integración* $\left[e^{-4}, 1\right]$ *y este hecho influye en el proceso numérico.*

A partir de estos ejemplos queda claro que antes de realizar cualquier procedimiento numérico habría que efectuar un análisis de errores.

1.1.2. Errores relativos y absolutos

Por lo general, el valor exacto de una determinada variable x es desconocido y solamente tendremos a nuestro alcance un valor aproximado x^*. Con esta idea definimos los siguientes conceptos.

Definición 1.1 *Definimos el* ***error absoluto*** *de una variable x, a la diferencia entre su valor aproximado y su valor exacto, es decir*

$$e_a(x) = \Delta x = x - x^*$$

Si $e_a(x) > 0$, entonces se dice que se comete un error por defecto, *mientras que si $e_a(x) < 0$, se dirá que el error es* por exceso.

Definición 1.2 *Definimos el* ***error relativo*** *de una variable x, al cociente entre su error absoluto y el valor exacto, es decir*

$$e_r(x) = \frac{e_a(x)}{x}$$

siempre, claro está, que $x \neq 0$.

En la práctica, como el valor exacto x es desconocido, y si $x^* \neq 0$, se suele estimar este error mediante la expresión similar

$$e_r(x) \simeq \frac{e_a(x)}{x^*}$$

Definición 1.3 *Se define el* ***error porcentual*** *como $e_p(x) = e_r(x) \cdot 100$.*

Debido al desconocimiento del valor exacto de x, todos estos errores también estarán indeterminados y como mucho se conocerá una cota de los mismos; es decir, a lo sumo conoceremos un número $\varepsilon_a(x)$ con $|e_a(x)| \leq \varepsilon_a(x)$ para el error absoluto; o bien $|e_r(x)| \leq \varepsilon_r(x)$ para el error relativo.

Para expresar los datos que contienen un determinado error, se suele utilizar la siguiente notación

$$x = x^* \pm \varepsilon_a(x) x = x^*(1 \pm \varepsilon_r(x)) \tag{1.1}$$

Nota 1 *Muchos de los métodos numéricos que veremos son de tipo iterativo; en los que se consiguen aproximaciones a las soluciones exactas de los problemas, utilizando como punto de partida aproximaciones anteriores. En este caso, los errores se calculan como la diferencia entre dos aproximaciones sucesivas*

$$\varepsilon_a \simeq x_{n+1} - x_n$$

$$\varepsilon_r \simeq \frac{x_{n+1} - x_n}{x_{n+1}}$$

$$\varepsilon_r \cdot 100 \simeq \frac{x_{n+1} - x_n}{x_{n+1}} \cdot 100$$

1.1.3. Fuentes de error

Los errores pueden producirse por diversos motivos, aunque en particular estamos interesados en tres fuentes de error: errores en los datos de entrada, errores de redondeo durante los cálculos y errores de truncamiento por el método empleado.

1. **Error en los datos de entrada:** Pueden ser debidos a diversas causas entre las que destacamos:

   a) Mediciones Incorrectas por fallos en la realización de la medida o la precisión del apartado de medida empleado.

   b) Finitud de la representación digital de un dato. Este tipo de error aparece al utilizar los ordenadores y calculadoras, que tienen una capacidad finita de almacenamiento para el número de cifras decimales de un valor.

2. **Error de redondeo durante el cálculo:** Los errores no solamente provienen de los datos de entrada, sino que también aparecen del redondeo en los resultados intermedios.

3. **Error de truncamiento del método empleado:** Al resolver un problema matemático mediante métodos numéricos y aunque los cálculos se hagan de forma exacta, obtenemos sólo una aproximación numérica del resultado exacto (por ejemplo al aproximar una integral mediante una suma finita o una función por su polinomio de Taylor). El error producido depende del método numérico empleado y recibe el nombre de *error de truncamiento*. Para algunos métodos es posible obtener expresiones de este error.

   Al elegir un método, no solamente hay que tener en cuenta el error de truncamiento, sino también los errores de redondeo producido por las operaciones que el método comporta.

1.1.4. Números de punto flotante y error de redondeo

Puesto que las fuentes de error de los datos de entrada debido a mediciones incorrectas o precisión de los aparatos de medida no son fácilmente controlables, estos errores no serán estudiados. Los errores debidos a los datos de entrada por su representación digital en los ordenadores, tampoco son evitables, debido en parte a la capacidad finita de los ordenadores, sin embargo se puede controlar y es conveniente que los entendamos. Para ello es necesario estudiar cómo se almacenan estos datos en el ordenador.

Representación de números en base b

El método de representación de números que empleamos normalmente es el llamado sistema de numeración en base 10 o sistema decimal. Este sistema es de lo llamados posicionales puesto que el valor de un dígito varía según la posición que este ocupa dentro del número. El valor se obtiene mutiplicando el dígito correspondiente por la base, elevada a la posición; así un número cualquiera en el sistema decimal se puede representar como sigue

$$x = d_n \times 10^n + \ldots + d_1 \times 10 + d_0 \times 10^0 + d_{-1} \times 10^{-1} + \ldots$$

donde $d_k \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ y que escribimos como

$$x = d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots$$

Por ejemplo, el número decimal 2435,567, se puede expresar en base decimal como

$$\begin{aligned} 435{,}567 &= 4 \times 10^2 + 3 \times 10^1 + 5 \times 10^0 + 5 \times 10^{-1} + 6 \times 10^{-2} + 7 \times 10^{-3} \\ &= 4 \times 10^2 + 3 \times 10^1 + 5 + \frac{5}{10} + \frac{6}{10^2} + \frac{7}{10^3} \end{aligned}$$

No obstante y aunque la base decimal es la usada normalemente, es interesante conocer que también es posible una representación diferente, empleando para ello otra base distinta a la base 10. En general, es posible emplear cualquier número entero positivo que cumpla $b \geq 2$ como base de representación, siguiendo para ello el esquema descrito para la base 10 de la siguiente forma

$$x = d_n b^n + \ldots + d_1 b + d_0 b^0 + d_{-1} b^{-1} + \ldots = \sum_{k=-\infty}^{n} d_k b^k$$

donde en este caso $d_k \in \mathbb{Z}$ *y* $0 \leq d_k < b$.

Nota 2 *Si* $b = 16$*, es decir, la base hexadecimal, se utilizan como dígitos a los siguientes:* $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E$ *De esta forma el número hexadecimal* $(FFFF)_{16}$ *es equivalentes a* 65535 *en la base decimal, puesto que*

$$\begin{aligned}(FFFF)_{16} &= F \cdot 16^3 + F \cdot 16^2 + F \cdot 16^1 + F \cdot 16^0 \\ &= 15 \cdot 16^3 + 15 \cdot 16^2 + 15 \cdot 16^1 + 15 \cdot 16^0 \\ &= 61440 + +3840 + 240 + 15 \\ &= 65535\end{aligned}$$

Siendo esta base, junto con la binaria donde $b = 2$*, la más utilizada en los ordenadores.*

Definición 1.4 *Dado un número* $x \in \mathbb{R}$ *cuya representación en base* b *es* $x = \sum_{k=-\infty}^{n} d_k b^k$*, definimos:*

1. Parte entera de x en base b
$$x_e = \sum_{k=0}^{n} d_k b^{-k}$$
2. Parte decimal de x en base b
$$x_d = \sum_{k=-\infty}^{-1} d_k b^{-k}$$

Notar que

$$x = x_e + x_d$$

Cuando la base de representación no es decimal, se suele indicar mediante un subíndice, así la *representación digital* de x en la base b se expresa como

$$x = (d_n d_{n-1} \cdots d_1 d_0 . d_{-1} d_{-2} \cdots)_b$$

Los elementos d_k son los **dígitos** o **cifras** del número x en la base b.

La representación digital de un número x en una base dada es única, salvo para los números de la forma

$$\frac{k}{b^n} \quad k, n \in \mathbb{Z} \text{ y } 1 \leq k < b$$

donde hay dos posibles representaciones. Por ejemplo, en base 10 esas dos representaciones son

$$\frac{132}{100} = 1{,}319999999\ldots = 1{,}31\widehat{9} = 1{,}32.$$

si bien, para estos casos es preferible utilizar la representación finita.

Para la representación de números en los ordenadores se suele utilizar la base 2 (binaria) o la base 16 (hexadecimal), por ejemplo

$$\begin{aligned} 1001{,}11101_2 &= 1\times 2^3+0\times 2^2+0\times 2^1+1\times 2^0+1\times 2^{-1}+1\times 2^{-2}+ \\ &\quad 1\times 2^{-3}+0\times 2^{-4}+1\times 2^{-5} \\ &\approx 9{,}90625_{10}. \end{aligned}$$

Puede ocurrir que el número de cifras decimales pase de finito a infinito al cambiar de base, por ejemplo, para el número decimal $\frac{1}{10}$

$$\frac{1}{10} = (0{,}1)_{10} = (0{,}0001100110011001100110011.....)_2$$

1.1.5. Cambios de base

A continuación estudiaremos los mecanismos de cambio de base. En particular estudiaremos como se realiza un cambio de representación en base 10 a otra base b cualquiera y viceversa.

Cambio de base b a base 10

Sea $x \in \mathbb{R}$ y supongamos que su representación en la base b es

$$x = (d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} \ldots)_b$$

Si queremos obtener la representación de x en la base decimal, sólo hay que utilizar la definición de su representación en base b como sumatorio de potencias de esta base

$$x = (d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} \ldots)_b = \sum_{k=-\infty}^{n} d_k b^k$$

y realizar las operaciones correspondientes utilizando la base 10.

Ejemplo 1.5 *Sea $x = (123{,}456)_8$, encuentra la representación de x en la base 10.*

Solución: Como x está expresado en la base $b = 8$, para obtener la representación de x en la base 10, ponemos

$$\begin{aligned} x &= (123{,}456)_8 = 1\times 8^2+2\times 8^1+3\times 8^0+4\times 8^{-1}+5\times 8^{-2}+6\times 8^{-4} \\ &= 64+16+3+\frac{4}{8}+\frac{5}{64}+\frac{6}{512} \\ &= \frac{21399}{256} = 83{,}58984375 \end{aligned}$$

Cambio de base 10 a base b

Supongamos ahora que para $x \in \mathbb{R}$ es conocida su representación en base 10 y queremos obtener su representación en otra base $b \neq 10$. Para realizar este *cambio de base* en primer lugar hay que separar la parte entera (x_e) de la parte decimal (x_d) y tratar ambas partes de forma diferente. Supongamos por comodidad que $x > 0$ (en caso contrario se cambiaría el signo). Encontrar su representación en base b es encontrar una sucesión de dígitos $\{d_k\}$ con $0 \leq d_k < b$ tales que

$$x = (d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} \ldots)_b = \sum_{k=-\infty}^{n} d_k b^k$$

Como se ha indicado en el párrafo anterior, primero separamos x en la suma de su parte entera y su parte decimal

$$x = x_e + x_d$$

Empezaremos por la parte entera x_e. Si $x_e < b$, entonces la representación de x_e en base b sería $d_0 = x_e$ y $d_k = 0$ si $k \neq 0$, es decir

$$x_e = (x_e)_b$$

Supongamos ahora que $x_e \geq b$, entonces podremos realizar la división euclidea (división con resto) entre x_e y b

$$\frac{x_e}{b} = c_0 + \frac{r_0}{b}$$

donde c_0 es el cociente y r_0 el resto de esta división y donde $x_e, c_0, b, r_0 \in \mathbb{Z}$. Además teniendo en cuenta que x_e y b son positivos entonces se cumple

$$0 \leq r_0 < b$$

Si realizamos esta misma operación, pero utilizando ahora la representación de x como sumatorio de potencias de b se obtiene

$$\frac{x_e}{b} = \frac{(d_n d_{n-1} \dots d_1 d_0)_b}{b} = \frac{1}{b} \sum_{k=0}^{n} d_k b^k = \sum_{k=0}^{n} d_k b^{k-1}$$

y separando los términos positivos de los negativos del sumatorio

$$\frac{x_e}{b} = \sum_{k=0}^{n} d_k b^{k-1} = \sum_{k=1}^{n} d_k b^{k-1} + d_0 b^{-1}$$

Por claridad hacemos el cambio $j = k - 1$ en el sumatorio

$$\frac{x_e}{b} = \sum_{j=0}^{n-1} d_{j+1} b^j + \frac{d_0}{b} = \sum_{j=0}^{n-1} d_j^* b^j + \frac{d_0}{b}$$

donde se ha puesto $d_j^* = d_{j+1}$. Comparando este resultado con el obtenido mediante la división euclidea

$$\frac{x_e}{b} = \sum_{j=0}^{n-1} d_j^* b^j + \frac{d_0}{b} = c_0 + \frac{r_0}{b}$$

es fácil establecer las siguientes relaciones

$$\begin{aligned} \sum_{j=0}^{n-1} d_j^* b^j &= c_0 \\ \frac{d_0}{b} &= \frac{r_0}{b} \end{aligned}$$

y de aquí se deduce que

$$d_0 = r_0$$

que es el primer dígito de la representación de x_e en la base b. El proceso se repite con c_0 para obtener d_1, el siguiente dígito de x_e y un nuevo cociente c_1. Se construye así una sucesión de divisiones de la forma

$$\frac{c_{k-1}}{b} = c_k + \frac{r_k}{b} \quad k \geq 1$$

hasta encontrar un cociente c_k que cumpla $c_k < b$, en ese caso c_k será el último dígito de x_e en la base b y el proceso terminará. El resto r_k de cada división es el dígito asociado a la potencia b^k en la representación de x en base b.

Para expresar la parte decimal de x en base b, en primer lugar la multiplicamos por b

$$b x_d = b \sum_{k=-\infty}^{-1} d_k b^k = \sum_{k=-\infty}^{-1} d_k b^{k+1}$$

que podemos descomponer como

$$bx_d = d_{-1} + \sum_{k=-\infty}^{-2} d_k b^{k+1}$$

Como antes y para mayor claridad hacemos el cambio $j = k + 1$ en el sumatorio

$$bx_d = d_{-1} + \sum_{j=-\infty}^{-1} d_{j-1} b^j = d_{-1} + \sum_{j=-\infty}^{-1} d_j^* b^j$$

y donde se ha puesto $d_j^* = d_{j-1}$. En el sumatorio no hay potencias positivas de b y por tanto es un número decimal. Si llamamos x_d^1 a ese sumatorio,

$$bx_d = d_{-1} + x_d^1$$

y bx_d es un número real cuya parte entera, d_{-1}, es el primer dígito de x_d en la base b. El proceso se repite con x_d^1, la parte decimal de bx_d, para obtener el siguiente dígito de x_d y así sucesivamente de forma recursiva. Se obtiene así una sucesión de la forma

$$bx_d^{k-1} = d_{-k} + x_d^k$$

que nos permiten obtener cada uno de los dígitos d_{-k} del número x_d en la base b. Este proceso termina en alguno de los siguientes supuestos, bien cuando se encuentra un k tal que $x_d^k = 0$ (el número tiene una cantidad finita de cifras decimales en la base b), bien cuando se encuentran dos valores j y k con $j \neq k$ tal que $x_d^j = x_d^k$ (el número tiene una parte decimal periódica en la base b), o bien cuando tengamos una cantidad de dígitos suficiente.

Ejemplo 1.6 *Dado el siguiente número en base* 10, $x = 1234{,}56$. *Expresa* x *en base* 7.

Solución: En primer lugar separamos la parte entera y decimal de x

$$x = x_e + x_d = 1234 + 0{,}56$$

Representaremos cada una de ellas en la base 7. Las sucesivas divisiones de x_e por 7, proporcionan

$$\begin{aligned}
\frac{1234}{7} &= 176 + \frac{2}{7} \Rightarrow d_0 = 2 \\
\frac{176}{7} &= 25 + \frac{1}{7} \Rightarrow d_1 = 1 \\
\frac{25}{7} &= 3 + \frac{4}{7} \Rightarrow d_2 = 4 \\
3 &< 7 \Rightarrow d_3 = 3
\end{aligned}$$

por tanto

$$x_e = (3412)_7$$

Para la parte decimal, multiplicamos sucesivamente por 7 según la regla descrita

$$\begin{aligned}
0{,}56 \times 7 &= 3{,}92 \Rightarrow d_{-1} = 3 \\
0{,}92 \times 7 &= 6{,}44 \Rightarrow d_{-2} = 6 \\
0{,}44 \times 7 &= 3{,}08 \Rightarrow d_{-3} = 3 \\
0{,}08 \times 7 &= 0{,}56 \Rightarrow d_{-4} = 0 \\
0{,}56 \times 7 &= 3{,}92 \Rightarrow d_{-5} = 3
\end{aligned}$$

vemos que se repite el número inicial y por tanto el número será periódico en base 7

$$x_d = (0{,}3630363....)_7$$

La representación completa de x en la base 7 se obtiene sumando ambas partes

$$x = x_e + x_d = (3412{,}36303630.....)_7$$

1.1.6. Notación científica normalizada

Cualquier número real puede expresarse mediante la llamada *notación científica normalizada.* Esta notación consiste en desplazar el punto decimal de forma que todos los dígitos aparezcan a su derecha, multiplicando por la correspondiente potencia de 10 y siendo el primer dígito después de la coma decimal distinto de 0. Por ejemplo

$$\begin{aligned} 732{,}5051 &= 0{,}7325051 \times 10^3 \\ -0{,}005612 &= -0{,}5612 \times 10^{-2} \end{aligned}$$

serían representaciones de números en notación científica normalizada, mientras que

$$\begin{aligned} 732{,}5057 &= 7{,}325057 \times 10^2 \\ -0{,}005612 &= -5612 \times 10^{-6} \end{aligned}$$

no serían representaciones normalizadas de dichos números.

Un número real cualquiera $x \neq 0$ puede representarse en la forma

$$x = (-1)^s \times m \times 10^e$$

donde m, llamada **mantisa,** verifica

$$\frac{1}{10} \leq m < 1$$

donde e, el **exponente,** es un número entero y s, el **signo,** será 0 si x es positivo o 1 en caso contrario. Tanto m como e se representan en la misma base.

También es posible utilizar este tipo de notación en cualquier base $b \geq 2$, en este caso para $x \neq 0$ se podrá representar como

$$x = (-1)^s \times m \times b^e$$

donde ahora m cumple la relación

$$\frac{1}{b} \leq m < 1$$

1.1.7. Representación en punto flotante y error de representación

Mientras que en los cálculos matemáticos habituales es posible utilizar números con infinitas cifras decimales, como π o $\sqrt{2}$, la representación de números en un ordenador tiene la desventaja de utilizar una cantidad finita de cifras que está fijada con antelación. Un ordenador únicamente emplea un subconjunto relativamente pequeño de números reales para representarlos a todos, de hecho este subconjunto solamente contiene números racionales, tanto positivos como negativos. Veremos que esta limitación se traduce al realizar operaciones aritméticas en la aparición de los llamados errores de redondeo.

Suponemos ahora que realizamos los cálculos en un ordenador que representa datos con k dígitos en la base b y notación científica normalizada. En este caso la representación en el ordenador de un valor x con más de k dígitos no nulos, no será igual a x. A esta representación se le llama *representación en punto flotante* de x o abreviadamente $\mathrm{fl}\,(x)$

$$\mathrm{fl}\,(x) \quad = \quad \pm 0.d_1 d_2 \ldots d_k \times b^e$$

con $e, d_j \in \mathbb{Z}$, siendo $0 \leq d_j < b$ $(j = 1, \ldots, k)$ y $d_1 \neq 0$.

Supongamos por comodidad que $x > 0$ (en caso contrario se le cambia el signo) entonces el paso de x a $\mathrm{fl}\,(x)$ se puede hacer por *corte o truncamiento*, eliminando los dígitos de x a partir de d_k, es decir, si x se representa en notación científica normalizada como

$$x = (0.d_1 d_2 \ldots d_k d_{k+1} \ldots)_b \times b^e$$

donde $d_1 \neq 0$, entonces su representación en punto flotante por truncamiento sería

$$\mathrm{fl}_C\,(x) = (0.d_1 d_2 \ldots d_k)_b \times b^e$$

y como $x > 0$ ocurre

$$\mathrm{fl}_C\,(x) \leq x \tag{1.2}$$

También es posible elegir la representación en punto flotante de x mediante el llamado *redondeo por exceso*

$$\mathrm{fl}_E\,(x) = ((0.d_1 d_2 \ldots d_k)_b + b^{-k}) \times b^e$$

es decir, eliminamos el exceso de dígitos a partir de la cifra k, pero aumentamos en una unidad el último dígito de x. En este caso y como $x > 0$ ocurrirá

$$\mathrm{fl}_E\,(x) \geq x \tag{1.3}$$

Otra forma de representar x es mediante *redondeo*, eligiendo $\mathrm{fl}_R\,(x)$ de forma que el error $|x - \mathrm{fl}_R\,(x)|$ sea mínimo, es decir

$$|x - \mathrm{fl}_R\,(x)| = \text{mín}\left\{|x - \mathrm{fl}_E\,(x)|\,, |x - \mathrm{fl}_C\,(x)|\right\} \tag{1.4}$$

Si esta opción da lugar a dos posibles redondeos, se elige el que tiene el mayor valor absoluto. Concretamente para la base 10, si

$$x = (0.d_1 d_2 \ldots d_k d_{k+1} \ldots)_{10} \times 10^e$$

entonces

$$\mathrm{fl}_R\,(x) = \begin{cases} (0.d_1 d_2 \ldots d_k)_{10} \times 10^e & d_{k+1} < 5 \\ ((0.d_1 d_2 \ldots d_k)_{10} + 10^{-k}) \times 10^e & d_{k+1} \geq 5 \end{cases}$$

Las dos representaciones coincidirán cuando

$$|x - \mathrm{fl}_E\,(x)| = |x - \mathrm{fl}_C\,(x)|$$

y teniendo en cuenta que para $x > 0$, se cumplen las relaciones 1.2 y 1.3, quitamos el valor absoluto para obtener

$$\mathrm{fl}_E\,(x) - x = x - \mathrm{fl}_C\,(x)$$

de donde

$$x = \frac{\mathrm{fl}_E\,(x) + \mathrm{fl}_C\,(x)}{2} \tag{1.5}$$

Cotas de error

A continuación vamos a determinar cotas para los errores absoluto y relativo que se cometen al reemplazar x por su versión cortada ($\mathrm{fl}_C(x)$) o redondeada ($\mathrm{fl}_R(x)$).

Para la aproximación por corte tendremos

$$\begin{aligned} |x - \mathrm{fl}_C(x)| &= |(0.d_1d_2\ldots d_kd_{k+1}\ldots)_b \times b^e - (0.d_1d_2\ldots d_k)_b \times b^e| \\ &= |(0{,}00\ldots 0d_{k+1}\ldots)_b \times b^e| \\ &= \left|(0.d_{k+1}\cdots)_b \times b^{-k} \times b^e\right| \\ &= \left|(0.d_{k+1}\cdots)_b \times b^{e-k}\right| \end{aligned}$$

y teniendo en cuenta que

$$|(0.d_{k+1}\cdots)_b| < 1$$

se obtiene la siguiente cota para el error absoluto

$$|x - \mathrm{fl}_C(x)| \leq b^{e-k}$$

mientras que para el error relativo será

$$\left|\frac{x - \mathrm{fl}_C(x)}{x}\right| \leq \frac{b^{e-k}}{m \times b^e} \leq \frac{1}{m}b^{-k} \leq b^{1-k}$$

donde se ha tenido en cuenta la desigualdad $\frac{1}{b} \leq m < 1$.

Se obtienen las mismas cotas de error si utilizamos la aproximación mediante redondeo por exceso, ya que

$$\begin{aligned} |x - \mathrm{fl}_E(x)| &= \left|(0.d_1d_2\ldots d_kd_{k+1}\ldots)_b \times b^e - ((0.d_1d_2\ldots d_k)_b + b^{-k}) \times b^e\right| \\ &= \left|(0{,}00\ldots 0d_{k+1}\ldots)_b \times b^e - b^{-k}b^e\right| \\ &= \left|(0.d_{k+1}\cdots)_b \times b^{-k} \times b^e - b^{e-k}\right| \\ &= \left|(0.d_{k+1}\cdots)_b \times b^{e-k} - b^{e-k}\right| \\ &= \left|(0.d_{k+1}\cdots - 1)_b \times b^{e-k}\right| \end{aligned}$$

y teniendo en cuenta que

$$|(0.d_{k+1}\cdots - 1)_b| < 1$$

la cota para el error absoluto es

$$|x - \mathrm{fl}_E(x)| < b^{e-k}$$

mientras que para el error relativo

$$\left|\frac{x - \mathrm{fl}_E(x)}{x}\right| \leq \frac{b^{e-k}}{m \times b^e} \leq \frac{1}{m}b^{-k} \leq b^{1-k}$$

que son idénticas a las encontradas para la representación mediante truncamiento.

Para el caso de redondeo vamos a obtener unas cotas ligeramente distintas. Teniendo en cuenta la expresión 1.4

$$\begin{aligned} |x - \mathrm{fl}_R(x)| &\leq |x - \mathrm{fl}_E(x)| \\ &\text{y} \\ |x - \mathrm{fl}_R(x)| &\leq |x - \mathrm{fl}_C(x)| \end{aligned}$$

como $x > 0$

$$|x - \mathrm{fl}_R(x)| \leq \mathrm{fl}_E(x) - x$$

$$|x - \mathrm{fl}_R(x)| \leq x - \mathrm{fl}_C(x)$$

Si sumamos ambas desigualdades

$$2|x - \mathrm{fl}_R(x)| \leq (\mathrm{fl}_E(x) - x) + (x - \mathrm{fl}_C(x)) = \mathrm{fl}_E(x) - \mathrm{fl}_C(x)$$

por tanto

$$|x - \mathrm{fl}_R(x)| \leq \frac{1}{2}(\mathrm{fl}_E(x) - \mathrm{fl}_C(x))$$

Y utilizando las definiciones de $\mathrm{fl}_E(x)$ y $\mathrm{fl}_C(x)$

$$\begin{aligned} \mathrm{fl}_E(x) - \mathrm{fl}_C(x) &= \left(0.d_1d_2\ldots d_k)_b + b^{-k}\right) \times b^e - (0.d_1d_2\ldots d_k) \times b^e = \\ &= (0.d_1d_2\ldots d_k)_b \times b^e + b^{-k+e} - (0.d_1d_2\ldots d_k)_b \times b^e \\ &= b^{-k+e} \end{aligned}$$

y por tanto

$$|x - \mathrm{fl}_R(x)| \leq \frac{1}{2} \times b^{e-k}$$

En este caso la cota del error relativo es

$$\left|\frac{x - \mathrm{fl}_R(x)}{x}\right| \leq \frac{1}{2}\frac{b^{e-k}}{m \times b^e} \leq \frac{1}{2}\frac{1}{m}b^{-k} \leq \frac{1}{2}b^{1-k}$$

El valor $\frac{1}{2}b^{1-k}$ es el llamado *epsilon de máquina* (eps).

Aunque los errores de redondeo suelen ser pequeños, debido a la precisión de los ordenadores, estos pueden ser críticos por dos razones:

- Algunos métodos requieren una gran cantidad de operaciones para conseguir una respuesta, normalmente estas operaciones dependen de operaciones anteriores, de forma que aunque un error individual de redondeo pueda ser insignificante, el efecto de acumulación en el transcurso de una gran cantidad de cálculos puede ser muy significativo.
- El efecto de redondeo puede ser muy importante cuando se llevan a cabo operaciones algebráicas que emplean números muy pequeños y números muy grandes al mismo tiempo.

A continuación veremos algunos ejemplos de los efectos que provocan estos errores de redondeo

Ejemplo 1.7 *En el siguiente ejemplo vamos a usar operaciones con* 8 *cifras decimales y truncamiento. Dados los siguientes valores*

$$\begin{aligned} x &= 0{,}23371258 \times 10^{-4} \\ y &= 0{,}33678429 \times 10^{2} \\ z &= -0{,}33677811 \times 10^{2} \end{aligned}$$

comprobaremos que en el caso de aritmética finita la propiedad asociativa de los números reales no se cumple, es decir, vamos a comprobar que

$$(\mathrm{fl}(x) + \mathrm{fl}(y)) + \mathrm{fl}(z) \neq \mathrm{fl}(x) + (\mathrm{fl}(y) + \mathrm{fl}(z))$$

Para el miembro de la izquierda, realizamos en primer lugar la operación entre paréntesis $\text{fl}(x) + \text{fl}(y)$

$$\text{fl}(x) + \text{fl}(y) = 0{,}33678452 \times 10^2$$

y a continuación sumamos con $\text{fl}(z)$

$$(\text{fl}(x) + \text{fl}(y)) + \text{fl}(z) = 0{,}64100000 \times 10^{-3}$$

Mientras que para el miembros de la derecha, realizamos en primer lugar la suma dentro del paréntesis, $\text{fl}(y) + \text{fl}(z)$

$$\text{fl}(y) + \text{fl}(z) = 0{,}61800000 \times 10^{-3}$$

y después el resultado lo sumamos con $\text{fl}(x)$

$$\text{fl}(x) + (\text{fl}(y) + \text{fl}(z)) = 0{,}641371258 \times 10^{-3}$$

Como podemos comprobar este resultado es distinto al hallado anteriormente, además coincide con el valor exacto de la operación. En este caso el orden en el que se realicen las operaciones es determinante para encontrar el valor exacto. El problema aparece porque los números empleados no son del mismo orden de magnitud, ya que tanto y *como* z *son* 10^6 *veces más grandes que* x*.*

Ejemplo 1.8 *Dados los siguientes valores*

$$x = 0{,}3721478693$$

$$y = 0{,}3720230572$$

Vamos a realizar la operación $x - y$*, de forma exacta y utilizando redondeo a* 5 *cifras.*

$$x - y = 0{,}3721478693 - 0{,}3720230572 = 0{,}0001248121$$

$$\text{fl}(x) - \text{fl}(y) = 0{,}37215 - 0{,}37202 = 0{,}00013$$

El error relativo al utilizar la aritmética de cinco cifras es

$$\epsilon_r(\text{fl}(x) - \text{fl}(y)) = \frac{(\text{fl}(x) - \text{fl}(y)) - (x - y)}{(x - y)} = \frac{0{,}00013 - 0{,}0001248121}{0{,}0001248121} \simeq 0{,}41566 \times 10^{-2}$$

En este caso el problema aparece porque estamos restando cantidades casi iguales.

La pérdida de precisión debido a los errores de redondeo puede en algunos casos evitarse, bien realizando las operaciones en el orden adecuado, como en el primer ejemplo, bien reformulando el problema, como veremos a continuación.

Algunas de las técnicas generales para amortiguar el efecto de los errores en los cálculos, incluyen la racionalización de expresiones, como por ejemplo, dada la expresión

$$y = \sqrt{x^2 + 1} - 1$$

Para evaluar y en valores de x cercanos a 0, es mejor utilizar una versión racionalizada de la anterior

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

Veamos un ejemplo de aplicación.

Ejemplo 1.9 *Como sabemos sobradamente, la fórmula exacta para resolver la ecuación de segundo grado es*

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

que nos proporciona las siguientes soluciones

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

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Vamos a aplicar estas dos expresiones para encontrar las soluciones de la ecuación

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

utilizando aritmética de cuatro cifras y redondeo.

El valor de la raíz cuadrada es

$$\sqrt{b^2 - 4ac} = \sqrt{(0{,}6210 \times 10^2)^2 - (0{,}4000 \times 10^1)(0{,}1000 \times 10^1)(0{,}1000 \times 10^1)} =$$

$$= \sqrt{(0{,}3856 \times 10^4) - (0{,}4000 \times 10^1)} = 0{,}6206 \times 10^2$$

y los valores aproximados de las raíces serían

$$\mathrm{fl}_R(x_1) = \frac{-0{,}6210 \times 10^2 + 0{,}6206 \times 10^2}{0{,}2000 \times 10^1} = -0{,}0200$$

y

$$\mathrm{fl}_R(x_2) = \frac{-0{,}6210 \times 10^2 - 0{,}6206 \times 10^2}{0{,}2000 \times 10^1} = -62{,}10$$

Por otro lado, las soluciones exactas de la ecuación son

$$x_1 = -0{,}01610723$$

$$x_2 = -62{,}08390$$

por lo que los errores relativos son

$$\varepsilon_r(x_1) = \frac{|x_1 - \mathrm{fl}_R(x_1)|}{|x_1|} = \frac{|-0{,}01611 + 0{,}0200|}{|-0{,}01611|} = 0{,}24$$

y

$$\varepsilon_r(x_2) = \frac{|x_2 - \mathrm{fl}_R(x_2)|}{|x_2|} = \frac{|-62{,}08 + 62{,}10|}{|-62{,}08|} = 3{,}2 \times 10^{-4}$$

respectivamente. En el caso de x_1 el error relativo es demasiado grande, de un 25 %. El problema aparece porque en el numerador se realiza una resta de dos números casi iguales. Podemos tratar de resolver este problema racionalizando el numerador en la expresión de x_1

$$x_1 = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{-2c}{b + \sqrt{b^2 - 4ac}}$$

Empleando ahora esta fórmula obtenemos

$$\mathrm{fl}(x_1) = \frac{-0{,}2000 \times 10^1}{0{,}6210 \times 10^2 + 0{,}6206 \times 10^2} = -0{,}01610$$

cuyo error relativo es

$$\varepsilon_r(x_1) = \frac{|x_1 - \text{fl}_R(x_1)|}{|x_1|} = \frac{|-0{,}01611 + +0{,}01610|}{|-0{,}01611|} = 6{,}2 \times 10^{-4}$$

que ahora sí que es aceptable.

Sin embargo, si aplicamos esta técnica de racionalización a x_2, la otra raíz, obtenemos la expresión

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}$$

pero para la que el valor aproximado obtenido

$$\text{fl}(x_2) = \frac{-0{,}2000 \times 10^1}{0{,}6210 \times 10^2 - 0{,}6206 \times 10^2} = -50{,}00$$

tiene un error relativo muy grande

$$\varepsilon_r(x_2) = 1{,}9 \times 10^{-1}$$

esto se produce de nuevo porque ahora es en el denominador donde estamos restando dos números casi iguales.

También es posible utilizar desarrollos alternativos, por ejemplo, la expresión

$$y = x - \text{sen}(x)$$

puede dar problemas para valores de x cercanos a 0 ya que se estarían restando cantidades similares, en este caso podemos sustituir el valor de $\text{sen}(x)$ por su desarrolo de Taylor

$$y = x - \text{sen}(x) = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots$$

y después utilizar el esquema de Horner para disminuir el número de operaciones en cada cálculo

$$\begin{aligned} y &= \frac{x^3}{3!}\left(1 - \frac{x^2}{4 \cdot 5!} + \frac{x^4}{4 \cdot 5 \cdot 6 \cdot 7} + \cdots\right) \\ &= \frac{x^3}{3!}\left(1 - \frac{x^2}{4 \cdot 5}\left(1 - \frac{x^2}{6 \cdot 7} + \cdots\right)\right) \\ &= \frac{x^3}{6}\left(1 - \frac{x^2}{20}\left(1 - \frac{x^2}{42} + \cdots\right)\right) \end{aligned}$$

Ejemplo 1.10 *Calcula el valor del polinomio $P_3(x) = x^3 - 6{,}1x^2 + 3{,}2x + 1{,}5$ en $x_0 = 4{,}71$, utilizando aritmética de tres cifras.*

Realizaremos las operaciones de forma exacta, por corte y por redondeo:

	x	x^2	x^3	$6{,}1x^2$	$3{,}2x$
Exacta	4,71	22,1841	104,48711	135,32301	15,072
Corte	$0{,}471 \times 10^1$	$0{,}221 \times 10^2$	$0{,}104 \times 10^3$	$0{,}134 \times 10^3$	$0{,}150 \times 10^2$
Redondeo	$0{,}471 \times 10^1$	$0{,}222 \times 10^2$	$0{,}105 \times 10^3$	$0{,}135 \times 10^3$	$0{,}151 \times 10^2$

El valor del polinomio será

$$\textit{Exacta} \quad : \quad P_3(x) = -0{,}142638990 \times 10^2$$

$$\textit{Corte} \quad : \quad P_3(x) = -0{,}135 \times 10^2$$

$$\textit{Redondeo} \quad : \quad P_3(x) = -0{,}134 \times 10^2$$

y sus respectivas errores relativos:

$$Corte \quad : \quad \varepsilon_r\left(P_3\left(\mathrm{fl}_C\left(x\right)\right)\right) = \left|\frac{-0{,}142638990 \times 10^2 + 0{,}135 \times 10^2}{-0{,}142638990 \times 10^2}\right| \approx 0{,}05$$

$$Redondeo: \ \varepsilon_r\left(P_3\left(\mathrm{fl}_R\left(x\right)\right)\right) = \left|\frac{-0{,}142638990 \times 10^2 + 0{,}134 \times 10^2}{-0{,}142638990 \times 10^2}\right| \approx 0{,}06$$

Como opción rescribimos el polinomio mediante el esquema de Horner como

$$P_3\left(x\right) = \left(\left(x - 6{,}1x\right)x + 3{,}2\right)x + 1{,}5$$

y en este caso se obtiene

$$Corte \quad : \quad P_3\left(x\right) = -0{,}142 \times 10^2$$

$$Redondeo \quad : \quad P_3\left(x\right) = -0{,}143 \times 10^2$$

siendo ahora los errores relativos

$$Corte \quad : \quad \varepsilon_r\left(P_3\left(\mathrm{fl}_C\left(x\right)\right)\right) = \left|\frac{-0{,}142638990 \times 10^2 + 0{,}142 \times 10^2}{-0{,}142638990 \times 10^2}\right| \approx 0{,}0045$$

$$Redondeo: \ \varepsilon_r\left(P_3\left(\mathrm{fl}_R\left(x\right)\right)\right) = \left|\frac{-0{,}142638990 \times 10^2 + 0{,}143 \times 10^2}{-0{,}142638990 \times 10^2}\right| \approx 0{,}0025$$

En ambos casos se ha disminuido el error relativo de forma considerable, esto es debido a que se realizan un menor número de operaciones.

1.1.8. Estimación y acotación de errores

Tratamos de conocer el efecto que, sobre el resultado final de un problema numérico, produce cada uno de los diferentes tipos de error que pueden tener lugar.

Veremos qué ocurre para los tres tipos básicos de error: los de los datos, los de los cálculos intermedios y los errores de truncamiento por el método empleado. **El error total sobre el resultado final será la suma de las contribuciones de los tres tipos de error.**

Propagación de los errores de los datos

Al efectuar operaciones aritméticas ($+$, $-$, $\times$, $/$) sobre dos datos afectados de error x_1 y x_2, se obtienen las siguientes cotas para los errores absolutos

$$\begin{aligned} \epsilon_a(x_1 + x_2) &= \epsilon_a(x_1) + \epsilon_a(x_2) \\ \epsilon_a(x_1 - x_2) &= \epsilon_a(x_1) + \epsilon_a(x_2) \end{aligned}$$

y si los errores de x_1 y x_2 son pequeños, las siguientes cotas aproximadas para los errores relativos

$$\begin{aligned} \epsilon_r(x_1 x_2) &\simeq \epsilon_r(x_1) + \epsilon_r(x_2) \\ \epsilon_r(x_1/x_2) &\simeq \epsilon_r(x_1) + \epsilon_r(x_2) \end{aligned}$$

Si el problema consiste en calcular el valor de una determinada función, $y = f(x)$, a partir de un valor x que está afectado de error, obtenemos la siguiente *fórmula aproximada de propagación del error*

$$e_a(y) \simeq f\,'(x) \cdot e_a(x)$$

que es una consecuencia directa del teorema del valor medio para funciones de una variable. A partir de esta fórmula deducimos una cota para el error absoluto de y en función de una cota del error absoluto de x, dando lugar a la *fórmula aproximada de propagación del error maximal* siguiente

$$\epsilon_a(y) \simeq |f\,'(x)|\,\epsilon_a(x)$$

En el caso de que tengamos que calcular el valor utilizando una función de varias variables, $y = f(x_1, \dots x_n)$, disponemos de la *fórmula aproximada de propagación del error,* usando las derivadas parciales de f

$$e_a(y) \simeq \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \dots, x_n) e_a(x_i)$$

y si se conocen las cotas de $e_a(x_i)$, podremos acotar $e_a(y)$, mediante la *fórmula aproximada de propagación del error maximal*

$$\epsilon_a(y) \simeq \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(x_1, \dots, x_n)\right| \epsilon_a(x_i)$$

adecuada cuando n, el número de datos afectados de error, no es grande.

Ejemplo 1.11 *Si calculamos $y = x_1 + \ldots + x_n$, a partir de n datos, entonces*

$$e_a(y) \simeq \sum_{i=1}^{n} e_a(x_i)$$

si cada variable tiene una cota de error ϵ, la fórmula del error maximal nos da $\epsilon_a(y) \simeq n\epsilon$.

Propagación de los errores en los cálculos

Aunque los valores de las variables empleadas no tengan error e incluso aunque puedan representarse en punto flotante de forma exacta, es decir, que estos valores sean números máquina, el resultado de esta operación puede no ser un número máquina y por tanto no pueda expresarse de forma exacta. En estos casos habrá un error debido, no a los datos empleados, sino a los cálculos.

Ejemplo 1.12 *Supongamos que estamos utilizando aritmética de tres cifras y consideremos dos números máquina, es decir, que puedan expresarse de forma exacta con tres cifras decimales, $a = 0{,}401 \times 10^1$ y $b = 0{,}514 \times 10^1$. El cálculo de $a \cdot b$ utilizando aritmética exacta nos proporciona el siguiente valor*

$$a \cdot b = \left(0{,}401 \times 10^1\right) \cdot \left(0{,}514 \times 10^1\right) = 0{,}206114 \times 10^2$$

mientras que con aritmética de tres cifras

$$\mathrm{fl}\,(a \cdot b) = 0{,}206 \times 10^2$$

y el error relativo que se produce es

$$e_r\,(a \cdot b) = \left|\frac{a \cdot b - \mathrm{fl}\,(a \cdot b)}{a \cdot b}\right| = \left|\frac{0{,}206114 \times 10^2 - 0{,}206 \times 10^2}{0{,}206114 \times 10^2}\right| = 0{,}00055309$$

El estudio de la propagación de los errores en los cálculos, se estudia en dos fases: análisis del error hacia atrás y propagación de los errores imputados a los datos.

Análisis del error hacia atrás Partiendo de datos iniciales exactos y debido a la acumulación de errores durante las operaciones, se obtiene un resultado afectado de error. La idea básica del *análisis del error hacia atrás* consiste en estudiar las modificaciones que habría que efectuar sobre los datos de entrada, de forma que, si suponemos que las operaciones están exentas de error se obtuviese el mismo error en los resultados.

Este estudio se lleva a cabo aplicando sucesivamente la fórmula

$$\mathrm{fl}\,(a \star b) = (a \star b)\,(1 + \delta_\star)$$

con $|\delta_\star| \leq \epsilon_\star$, para cada una de las operaciones aritméticas $\star \in \{+, -, \times, /\}$, que componen el proceso de cálculo, donde $\epsilon_\star$ es una cota conocida del error relativo en la operación $\star$ y fl es el redondeo por exceso o truncamiento. Además para cada función $g\,(x)$, que intervenga en los cálculos

$$\mathrm{fl}\,(g\,(x)) = g\,(x)\,(1 + \delta_g)$$

con $|\delta_g| \leq \epsilon_g$, y ϵ_g indica una cota del error relativo en la evaluación de $g\,(x)$.

Ejemplo 1.13 *Vamos a estudiar la propagación del error en los cálculos para una suma de tres sumandos*

$$a + b + c$$

En este caso podemos separar la operación en dos partes

$$\begin{aligned} \eta &= a + b \\ y &= \eta + c \end{aligned}$$

y aplicamos la propagación del error en los cálculos para cada una de las sumas

$$\begin{aligned} \mathrm{fl}\,(\eta) &= (a + b)\,(1 + \varepsilon_1) \\ \mathrm{fl}\,(y) &= (\eta + c)\,(1 + \varepsilon_2) = ((a + b)\,(1 + \varepsilon_1) + c)\,(1 + \varepsilon_2) \end{aligned}$$

Si desarrollamos

$$\begin{aligned} \mathrm{fl}\,(a + b + c) &= (a + b + a\varepsilon_1 + b\varepsilon_1 + c)\,(1 + \varepsilon_2) \\ &= a + b + c + a\varepsilon_1 + b\varepsilon_1 + a\varepsilon_2 + b\varepsilon_2 + c\varepsilon_2 + a\varepsilon_1\varepsilon_2 + b\varepsilon_1\varepsilon_2 \\ &= (a + b + c) + (a + b + c)\,\varepsilon_2 + (a + b)\,\varepsilon_1 + (a + b)\,\varepsilon_1\varepsilon_2 \end{aligned}$$

sacamos factor común $(a + b + c)$

$$= (a + b + c)\left(1 + \varepsilon_2 + \frac{a + b}{a + b + c}\varepsilon_1\,(1 + \varepsilon_2)\right)$$

por tanto

$$\varepsilon_4\,(y) = \varepsilon_2 + \frac{a + b}{a + b + c}\varepsilon_1\,(1 + \varepsilon_2) \simeq \frac{a + b}{a + b + c}\varepsilon_1 + \varepsilon_2$$

donde se han descartado los términos de orden superior (en este caso 2) $\varepsilon_1\varepsilon_2$.

Propagación de los errores imputados a los datos Después de la reducción anterior, se aplica la fórmula de propagación del error maximal a las cotas de los errores imputados a los datos, considerando ahora que los cálculos se hacen sin error.

Ejemplo 1.14 *Vamos a estudiar la propagación del error en los cálculos y en los datos para la suma de tres sumandos del ejercicio anterior. Allí habíamos obtenido una expresión aproximada para la suma de la forma*

$$\mathrm{fl}\,(a + b + c) = (a + b + c)\,(1 + \varepsilon_4)$$

donde

$$\varepsilon_4 = \frac{a+b}{a+b+c}\varepsilon_1 + \varepsilon_2$$

Realizamos ahora la suma sin error pero suponiendo que ahora los datos están afectados de error.

$$\begin{aligned}
\mathrm{fl}\,(a+b+c) &= (a^*+b^*+c^*)\,(1+\varepsilon_4)\\
&= (a\,(1+\varepsilon_1)+b\,(1+\varepsilon_2)+c\,(1+\varepsilon_3))\,(1+\varepsilon_4)\\
&= (a+b+c+a\varepsilon_1+b\varepsilon_2+c\varepsilon_3)\,(1+\varepsilon_4)\\
&= (a+b+c)\,(1+\varepsilon_4)+(a\varepsilon_1+b\varepsilon_2+c\varepsilon_3)+a\varepsilon_1\varepsilon_4+b\varepsilon_2\varepsilon_4+c\varepsilon_3\varepsilon_4\\
&\simeq (a+b+c)\,(1+\varepsilon_4)+(a\varepsilon_1+b\varepsilon_2+c\varepsilon_3)
\end{aligned}$$

donde hemos descartado los términos de orden superior $\varepsilon_j\varepsilon_4$ para $j=1,2,3$.

Errores de Truncamiento

Dependerán de cada método numérico y su estudio se efectuará de manera específica para los diferentes métodos que se vayan presentando.

1.1.9. Tratamiento intervalar y estadístico

La fórmula de propagación de errores se puede utilizar como una aproximación de primer orden para la acotación del error. El cálculo del error utilizando intervalos permite acotaciones más rigurosas; aunque desafortunadamente, estas acotaciones no son realistas cuando el número de cálculos es grande; es decir, son acotaciones desproporcionadas y poco probables del error real. Para obtener unas aproximaciones más cercanas a la realidad resulta muchas veces conveniente analizar el error desde el punto de vista estadístico. Introducimos a continuación ambos tratamientos.

Análisis de errores con intervalos

Este análisis hace uso de las operaciones entre intervalos, con el objeto de determinar un intervalo, en el cual estará con toda seguridad el resultado final.

Definición 1.5 *Un **intervalo (cerrado y acotado)** I en $\mathbb{R}$ es un conjunto de números reales de la forma*

$$[a,b] = \{x \in \mathbb{R} : a \le x \le b\}$$

Llamaremos $I(\mathbb{R})$ al conjunto de los intervalos de $\mathbb{R}$.

Se definen a continuación las operaciones más sencillas con intervalos. Definimos primero las *operaciones aritméticas entre intervalos:* Si $I_1, I_2 \in I(\mathbb{R})$,

$$I_1 \star I_2 = \{y \in R \mid y = x_1 \star x_2;\ x_1 \in I_1,\ x_2 \in I_2\}$$

para cualquier operación aritmética $\star$ entre números reales. El intervalo resultante será de la forma

$$I_3 = [\text{mín}\,\{I_1 \star I_2\},\ \ \text{máx}\,\{I_1 \star I_2\}]$$

Para ciertas funciones entre números reales, podemos definir *funciones entre intervalos* de la siguiente manera: dada la función $f : \mathbb{R} \to \mathbb{R}$, definimos

$$\begin{aligned}
f &: I(\mathbb{R}) \longrightarrow I(\mathbb{R})\\
f(I) &= \{y \mid y = f(x),\ x \in I\}
\end{aligned}$$

este tipo de operaciones solamente tiene sentido cuando $f(I)$ es de nuevo un intervalo de $\mathbb{R}$; por ejemplo si $f(x)$ es continua, esta condición siempre se cumple. El intervalo obtenido al aplicar la función f será de la forma

$$I^* = f(I) = [\text{mín}_{x\in I}\{f(x)\},\ \ \text{máx}_{x\in I}\{f(x)\}]$$

De esta forma, si $f(x) = \operatorname{sen} x$ con $x \in I = [0, \pi]$, entonces $f(I) = [-1, 1]$.

Usar operaciones y funciones entre intervalos permite llevar a cabo un seguimiento, a lo largo de la cadena de cálculos, de la región de la recta real donde se encuentran los sucesivos resultados parciales y también el resultado final del proceso. El problema fundamental de este tipo de análisis, como se ha comentado anteriormente, es que normalmente se sobrestima el valor del error.

Ejemplo 1.15 *Vamos a determinar, operando mediante intervalos, el error máximo para* $y = x_1x_2^2$*, siendo*

$$x_1 = 2{,}0 \pm 0{,}1 \qquad x_2 = 3{,}0 \pm 0{,}2$$

En este caso los intervalos utilizados son

$$x_1 \in I_1 = [1{,}9, 2{,}1]$$

$$x_2 \in I_2 = [2{,}8, 3{,}2]$$

y las operaciones con intervalos serán

$$I_1 \times I_2 \times I_2$$

$$x_2^2 \in I_2 \times I_2 \qquad [2{,}8, 3{,}2] \times [2{,}8, 3{,}2] = [7{,}84, 10{,}24]$$

$$x_1x_2^2 \in I_1 \times (I_2 \times I_2) \quad [1{,}9, 2{,}1] \times [7{,}84, 10{,}24] = [14{,}896, 21{,}504]$$

Luego la solución se encuentra en el intervalor $I_3 = [14{,}896, 21{,}504]$ y por tanto tomando el punto intermedio del intervalo

$$x_1x_2^2 = 18{,}2 \pm 3{,}304$$

Análisis de errores estadístico

Introducimos aquí el *tratamiento estadístico del error,* como una forma más conveniente de tratar un proceso numérico en el que están implicadas muchas operaciones.

Para realizar este análisis se supone que los errores en los datos de entrada son *variables aleatorias independientes* con una cierta *función de distribución dada.*

Por ejemplo, para obtener una estimación estadística de los errores de redondeo, si llamamos ε al error de redondear un número a k cifras decimales. Entonces ε toma valores no nulos sobre el intervalo $[-\text{eps}, \text{eps}]$, donde $\text{eps} = \frac{1}{2}10^{-k}$ es el epsilon de máquina utilizando en este caso la representación decimal ($2^{-(k+1)}$ en binario). Si suponemos que cada valor sobre este intervalo es igualmente probable, *la función de densidad* $f(\varepsilon)$ de ε corresponde a una *distribución uniforme*

$$f(\varepsilon) = \begin{cases} \dfrac{1}{2\epsilon} \text{ si } \varepsilon \in [-\text{eps}, \text{eps}] \\ \\ 0 \text{ si } \varepsilon \notin [-\text{eps}, \text{eps}] \end{cases}$$

y la *función de distribución*

$$F(\varepsilon) = \int_{-\infty}^{\varepsilon} f(t)dt = \int_{-\text{eps}}^{\varepsilon} \frac{1}{2\,\text{eps}}dt = \frac{\varepsilon + \text{eps}}{2\,\text{eps}}$$

que nos da la probabilidad de que ε tome valores más pequeños o iguales que ε. Si μ_ε es el valor esperado y σ_ε la varianza de esta distribución, entonces

$$\mu_\varepsilon = E(\varepsilon) = \int_{-\infty}^{\infty} t f(t) dt = \int_{-\text{eps}}^{\text{eps}} \frac{t}{2\,\text{eps}} dt = 0$$

$$\sigma_\varepsilon^2 = E\left(\varepsilon^2\right) - E(\varepsilon)^2 = \int_{-\infty}^{\infty} t^2 f(t) dt - 0 = \int_{-\text{eps}}^{\text{eps}} \frac{t^2}{2\,\text{eps}} dt = \frac{1}{3}\,\text{eps}^2$$

La mayoría de las variables aleatorias ε que se manejan en la práctica siguen aproximadamente una *distribución normal*; esto es, tienen una función de densidad de probabilidad aproximadamente igual a

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\varepsilon-\mu)^2/(2\sigma^2)}$$

donde μ es la media y σ^2 es la varianza y siendo la función de distribución

$$F(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\varepsilon} e^{-(t-\mu)^2/(2\sigma^2)} dt$$

La probabilidad $P(x)$ de que ε tome valores entre $\mu - x$ y $\mu + x$ viene dada por

$$F(\mu + x) - F(\mu - x) = \int_{\mu-x}^{\mu+x} f(t) dt = \frac{2}{\sqrt{\pi}} \int_0^{\frac{x}{\sqrt{2}\sigma}} e^{-t^2} dt = \text{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)$$

donde

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

recibe el nombre de *función de error*.

El resultado x de un algoritmo que está sujeto a errores aleatorios es también una variable aleatoria con un valor esperado μ_x y varianza σ_x^2. La propagación de los errores en los cálculos está descrita mediante las siguientes fórmulas para variables aleatorias independientes

$$\begin{aligned} \mu_{\alpha x \pm \beta y} &= a\mu_x \pm b\mu_y \\ \sigma^2_{ax \pm by} &= \alpha^2\sigma_x^2 + \beta^2\sigma_y^2 \end{aligned} \tag{1.6}$$

La primera de las fórmulas se obtiene por la linealidad de la esperanza de una variable aleatoria. La segunda fórmula está basada en la relación $\mu_{xy} = \mu_x\mu_y$, que se cumple si x e y son independientes.

Del mismo modo, para variables independientes se obtiene

$$\begin{aligned} \mu_{xy} &= \mu_x\mu_y \\ \sigma^2_{xy} &= \sigma_x^2\sigma_y^2 + \mu_x^2\sigma_y^2 + \mu_y^2\sigma_x^2 \end{aligned} \tag{1.7}$$

Ejemplo 1.16 *Para calcular $y = a^2 - b^2$, suponiendo que los errores siguen una distribución uniforme ($\mu_\varepsilon = 0$, $\sigma_\varepsilon^2 = \frac{1}{3}\,\text{eps}^2$) y que los valores a y b son exactos, tendremos*

$$E(a) = a \quad \sigma_a^2 = 0$$

$$E(b) = b \quad \sigma_b^2 = 0$$

y utilizando las ecuaciones 1.6 y 1.7 obtenemos

$$\eta_1 = a^2(1+\varepsilon_1) \;\Big|\; E(\eta_1) = a^2 \;\Big|\; \sigma^2_{\eta_1} = a^4 \tfrac{1}{3}\,\text{eps}^2$$

$$\eta_2 = b^2(1+\varepsilon_2) \;\Big|\; E(\eta_2) = b^2 \;\Big|\; \sigma^2_{\eta_2} = b^4 \tfrac{1}{3}\,\text{eps}^2$$

Mientras que para el resultado final $y = a^2 - b^2$

$$\begin{aligned} y &= (\eta_1 - \eta_2)(1+\varepsilon_3) \\ \mu_y &= E(y) = E(\eta_1-\eta_2)\,E(1+\varepsilon_3) = a^2 - b^2 \\ \sigma^2_y &= \sigma^2_{\eta_1-\eta_2}\sigma^2_{1+\varepsilon_3} + \mu^2_{\eta_1-\eta_2}\sigma^2_{1+\varepsilon_3} + \mu^2_{1+\varepsilon_3}\sigma^2_{\eta_1-\eta_2} \\ &= \left(\sigma^2_{\eta_1} + \sigma^2_{\eta_2}\right)\left(\frac{1}{3}\,\text{eps}^2\right) + \left(a^2-b^2\right)^2\left(\frac{1}{3}\,\text{eps}^2\right) + 1\left(\sigma^2_{\eta_1} + \sigma^2_{\eta_2}\right) \\ &= \left(a^4+b^4\right)\left(\frac{1}{3}\,\text{eps}^2\right)^2 + \left[\left(a^2-b^2\right)^2 + a^4 + b^4\right]\left(\frac{1}{3}\,\text{eps}^2\right) \end{aligned}$$

y descartando los términos de orden eps^4 *comparados con* eps^2

$$\sigma^2_y = \left(\left(a^2-b^2\right)^2 + a^4 + b^4\right)\left(\frac{1}{3}\,\text{eps}^2\right)$$

Para $a = 0{,}3237$ *y* $b = 0{,}3134$ *y* $\text{eps} = 5\times10^{-4}$ *entonces*

$$\sigma_y = 0{,}144\left(\frac{1}{3}\,\text{eps}^2\right) = 0{,}0000415$$

que es similar en magnitud al verdadero error

$$\varepsilon_r(y) = 0{,}00001787$$

para aritmética de 4 cifras.

El análisis estadístico de los errores de redondeo es muy complicado si las variables aleatorias incluidas no son independientes, sin embargo, es bastante sencillo cuando se hacen las siguientes simplificaciones:

1. Los errores en los cálculos de cada operación aritmética $(+,-,\times,/)$ son variables aleatorias independientes.
2. Al calcular las varianzas, todos los términos con orden mayor que el más pequeño son ignorados.
3. Todas las varianzas son suficientemente pequeñas, de forma que para cada operación elemental ocurre

$$E(x \star y) = E(x) \star E(y)$$

Resumiendo, si los valores esperados μ_x se sustituyen por los valores estimados de x, y se incluyen las varianzas relativas

$$\varepsilon_x = \frac{\sigma^2_x}{\mu^2_x} \approx \frac{\sigma^2_x}{x^2}$$

entonces

$$z = \text{fl}(x \pm y) \qquad \varepsilon^2_z = \left(\frac{x}{z}\right)^2\varepsilon^2_x + \left(\frac{y}{z}\right)^2\varepsilon^2_y + \frac{1}{3}\,\text{eps}^2$$

$$z = \text{fl}(x \times y) \qquad \varepsilon^2_z = \varepsilon^2_x + \varepsilon^2_y + \frac{1}{3}\,\text{eps}^2$$

$$z = \text{fl}(x/y) \qquad \varepsilon^2_z = \varepsilon^2_x + \varepsilon^2_y + \frac{1}{3}\,\text{eps}^2$$

Volviendo a la teoría general del tratamiento estadístico de errores, consideremos $e_1, \ldots, e_n$ variables aleatorias con medias $\mu_1, \ldots, \mu_n$ y desviaciones típicas $\sigma_1, \ldots, \sigma_n$, y sean $d_1, \ldots, d_n \in \mathbb{Z}$, se cumplen las propiedades siguientes:

1. La combinación lineal $e = \sum_{i=1}^{n} d_i e_i$ es una variable aleatoria con media

$$\mu = \sum_{i=1}^{n} d_i \mu_i$$

2. Si $e_1, \ldots, e_n$ son linealmente independientes (es decir, el valor que toma una variable no condiciona el valor de cualquiera de las restantes), e tiene por desviación típica σ con

$$\sigma^2 = \sum_{i=1}^{n} d_i^2 \sigma_i^2$$

   Si además, $e_1, \ldots, e_n$ son normales, e también lo es.

3. Si $e_1, \ldots, e_n$ son linealmente independientes y tienen la misma función de distribución, la función de distribución de e se aproxima a una función de distribución normal (cuando $n \to \infty$ por el Teorema central del límite).

Si queremos calcular $y = f(x_1, \ldots, x_n)$ y suponemos que $e_a(x_i) = x_i^* - x_i$, son variables aleatorias independientes con media nula y desviación típica $\sigma(e_a(x_i))$ o simplemente $\sigma(x_i)$; el error en y, $e_a(y) = y^* - y$, viene dado por la fórmula aproximada de propagación del error y teniendo en cuenta las propiedades anteriores, es una variable aleatoria con media nula y desviación estándar $\sigma(e_a(y))$, o simplemente $\sigma(y)$, dada por la *fórmula aproximada de propagación del error estándar*

$$\sigma(y) \simeq \left(\sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i}(x_1, \ldots, x_{n)} \right]^2 \sigma(x_i)^2 \right)^{1/2}$$

Notar que la aproximación de $e_a(y)$ tiene una distribución normal si las variables $e_a(x_i)$ la tienen y aproximadamente normal, si éstas tienen la misma función de distribución y n es grande.

1.2. Algoritmos

Un algoritmo es un procedimiento que describe, sin niguna ambigüedad, una sucesión finita de pasos a realizar en un orden específico con el fin de resolver un determinado problema, es una lista de instrucciones para efectuar paso por paso algún proceso.

La palabra algoritmo tiene su origen en el matemático Al-Jwarizmi (Mohamed ben Musa) que en el año 880 escribió un libro, en el cual, por primera vez en la historia se expresaban métodos precisos para efectuar las cuatro operaciones aritméticas más comunes.

Un algoritmo no es un método de resolución de un problema particular con unos datos particulares, sino el método de resolución de todos los problemas de un mismo tipo, sean cuales sean los datos de partida.

El concepto de algoritmo no es exclusivo de las Matemáticas, por ejemplo una receta de un libro de cocina también puede considerarse como un algoritmo, la preparación de un plato complicado se divide en pasos simples, comprensibles para cualquier persona con experiencia en cocina. Otro ejemplo de un algoritmo para un proceso sería, por ejemplo, el que se describe a continuación para el caso de la avería de en una rueda de un coche; que tendría aproximadamente los siguientes pasos:

1. Aflojar los tornillos.
2. Levantar el coche con el gato.
3. Quitar los tornillos.
4. Quitar la rueda.

5. Poner la rueda nueva.
6. Poner los tornillos.
7. Apretar los tornillos.
8. Bajar el coche con el gato.
9. Quitar el gato.

El objetivo de un algoritmo numérico será el de implementar un procedimiento numérico para resolver o aproximar la solución de un determinado problema. En general, un algoritmo no es más que el fin de un proceso de **divide y vencerás**.

Las **características básicas** de un buen algoritmo son:

1. Debe ser lo más general y lo más corto posible dentro de una lógica que permita su comprensión.
2. Cada paso debe realizar un proceso concreto y debe ser lo más independiente posible de los demás pasos.
3. Debe ser determinístico ("No debe dejar nada al azar").
4. Los resultados deben depender únicamente de las condiciones iniciales y no del contexto.

Nos interesamos primordialmente en elegir métodos que produzcan resultados confiables en su precisión. Cuando sea posible exigiremos al algoritmo, que con cambios pequeños en los datos iniciales se produzcan cambios pequeños en los resultados finales. Un algoritmo con esta propiedad se dice *estable,* y siendo *inestable* si no se cumple ese criterio. Algunos algoritmos serán estables para cierto grupo de datos iniciales, pero no para todos. Se tratará, siempre que se pueda, de caracterizar las propiedades de estabilidad de los algoritmos.

1.2.1. Cálculos estables e inestables. Condicionamiento

Presentamos en esta sección la distinción entre los procesos numéricos que son *estables* y los que no lo son. También introducimos un concepto estrechamente relacionado con la estabilidad: *problema bien condicionado* y *mal condicionado.*

Inestabilidad numérica

De manera informal decimos que un proceso numérico es **inestable** cuando los pequeños errores que se producen en alguna de sus etapas se agrandan en etapas posteriores y degradan seriamente la exactitud del cálculo en su conjunto.

Un ejemplo nos permitirá explicar este concepto. Consideramos la sucesión de números reales definida mediante

$$P_n = \left(\frac{1}{3}\right)^n \qquad n > 0$$

Podemos generar cualquier término de la sucesión mediante el siguiente proceso iterativo

$$P_0 = 1, \qquad P_n = \left(\frac{1}{3}\right) P_{n-1} \text{ si } n \geq 1$$

Para el término 50 obtenemos

$$P_{50} = 1{,}392955569098535e - 24$$

y teniendo en cuenta que la sucesión tiene límite 0, parece que este algoritmo lleva hasta la solución. Sin embargo, si utilizamos la siguiente sucesión

$$P_0 = 1, \qquad P_1 = \frac{1}{3}, \qquad P_n = \left(\frac{10}{3}\right) P_{n-1} - P_{n-2} \text{ si } n \geq 2$$

que también cumple la sucesión P_n y la usamos para obtener el valor P_{50}, obtenemos

$$P_{50} = 6181308{,}259101972$$

que a todas luces es bastante diferente al obtenido con el algoritmo anterior, además está claro que estamos muy lejos del valor 0, luego podemos decir que este algoritmo es inestable.

El que un proceso sea numéricamente estable o inestable debería decidirse en base a los errores relativos. Así, si hay errores grandes en un cálculo, la situación puede ser del todo aceptable si los resultados son grandes.

Otro ejemplo de inestabilidad numérica lo proporciona el cálculo de los números

$$y_n = \int_0^1 x^n e^x dx \quad (n \geq 0) \tag{1.8}$$

Si aplicamos la integración por partes en la integral que define a y_{n+1} se obtiene la siguiente relación de recurrencia

$$y_{n+1} = e - (n+1)\, y_n \tag{1.9}$$

De lo anterior y del hecho evidente de que $y_0 = e - 1$ obtenemos y_1

$$y_1 = e - y_0 = e - (e-1) = 1$$

Utilizando la relación 1.9 e iniciando con $y_1 = 1$, generamos y_2, y_3, $\ldots, y_{20}$. Algunos de estos valores son por ejemplo

$$y_2 = 0{,}718281828459045$$

$$y_{11} = 0{,}210265160231056$$

$$y_{19} = 6{,}599099498065924$$

Estos valores no pueden ser correctos. De hecho es obvio, como se deduce de 1.8, que la sucesión y_n satisface la relación

$$y_1 > y_2 > \ldots > 0$$

puesto que el integrando es positivo y el valor y_{n+1} se obtiene al quitar una cantidad positiva. Además se puede comprobar que

$$\lim_{n\to\infty} y_n = 0$$

(Notar que para $0 < x < 1$, la expresión x^n decrece monótonamente a 0)

Condicionamiento

Las palabras *condición y condicionamiento* se usan de manera informal para indicar la sensibilidad del algoritmo respecto a pequeños cambios en las condiciones iniciales. Es decir, que cambios experimenta la solución de un problema si modificamos las condiciones iniciales de entrada. Un problema estará **mal condicionado** si pequeños cambios en los datos pueden dar lugar a grandes cambios en las respuestas.

Para ciertos tipos de problemas se puede definir un *número de condición.* Si el número es grande significa que se tiene un problema mal condicionado.

Definición 1.6 *Definimos* ***número condición*** *como una cantidad asociada a un problema que caracteriza el condicionamiento del mismo.*

Por ejemplo si queremos evaluar una función f en un punto x. ¿Qué ocurre si pertubarmos ligeramente el valor de x? ¿Qué efecto tiene sobre $f(x)$? Podemos recurrir al teorema del valor medio y escribir:

$$f(x+h) - f(x) = f'(\xi)h \approx hf'(x)$$

de manera que si $f'(x)$ no es grande, el efecto del cambio en x sobre $f(x)$ será pequeño.

Si ahora nos centramos en el error relativo, al cambiar x en una cantidad h, tenemos que h/x es el tamaño relativo de la perturbación y la perturbación relativa sobre $f(x)$ será

$$\frac{f(x+h)-f(x)}{f(x)} \approx \frac{hf'(x)}{f(x)} = \left[\frac{xf'(x)}{f(x)}\right]\left(\frac{h}{x}\right)$$

por tanto, el factor $(xf'(x)/f(x))$, puede servir como número condición para el problema.

Ejemplo 1.17 *¿Cuál es el número condición para la evaluación de la función arco seno? Si* $f(x) = \operatorname{arcsen} x$, *entonces*

$$\frac{xf'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\operatorname{arcsen} x}$$

Para valores de x *próximos a* 1, $\operatorname{arcsen} x \approx \pi/2$, *y el número condición tiende a* ∞, *por tanto, pequeños errores relativos en* x *pueden conducir a grandes errores relativos en* $\operatorname{arcsen} x$, *cerca de* $x = 1$.

Consideremos ahora el problema de localizar el cero de una función f de clase $\mathcal{C}^2$. Supongamos que r es una raíz simple de f ($f'(r) \neq 0$). Si perturbamos la función f mediante otra función g de clase $\mathcal{C}^2$

$$F = f + \varepsilon g$$

podríamos preguntarnos ¿dónde se localizará la nueva raíz? Si suponemos que la nueva raíz es $r + h$, veremos ahora una fórmula aproximada para el valor de h. Como $r + h$ es la raíz de la función perturbada entonces $F(r + h) = 0$

$$f(r+h) + \varepsilon g(r+h) = 0$$

Como f y g son de clase $\mathcal{C}^2$, podemos utilzar el teorema de Taylor para expresar $F(r + h)$

$$\left[f(r) + hf'(r) + \frac{1}{2}h^2 f''(\xi)\right] + \varepsilon\left[g(r) + hg'(r) + \frac{1}{2}h^2 g(\eta)\right] = 0$$

Si descartamos los términos en h^2 y utilizamos el hecho de que $f(r) = 0$ obtenemos

$$h \approx -\varepsilon\frac{g(r)}{f'(r)+\varepsilon g'(r)} \approx -\varepsilon\frac{g(r)}{f'(r)}$$

Para ilustrar este análisis consideremos el siguiente ejemplo:

Ejemplo 1.18 *Consideremos el polinomio y la función perturbación siguientes*

$$\begin{aligned} f(x) &= \prod_{k=1}^{20}(x-k) = (x-1)(x-2)\ldots(x-20) \\ g(x) &= x^{20} \end{aligned}$$

Obviamente las raíces de $f(x)$, *son los enteros* $1, 2, \ldots, 20$. *¿Cómo se altera la raíz* $r = 20$ *cuando* f *se perturba a* $f + \varepsilon g$? *La respuesta la obtenemos aplicando la fórmula anterior para los datos actuales*

$$h \approx -\varepsilon\frac{g(20)}{f'(20)} = -\varepsilon\frac{20^{20}}{19!} \approx -\varepsilon 10^9$$

y por tanto un cambio de magnitud ε *en el coeficiente de* x^{20} *de* $f(x)$ *puede provocar una perturbación en la raíz* 20 *del orden de* $\varepsilon 10^9$. *Las raíces de este polinomio son muy sensibles a las perturbaciones de los coeficientes.*

Otro tipo de número de condición aparece asociado con la solución de sistemas lineales de la forma $Ax = b$, de manera muy breve, el número de condición de la matriz A se denota con $\kappa(A)$ y se define como el producto de ciertas magnitudes de A y de su inversa A^{-1}, es decir

$$\kappa(A) = \|A\| \cdot \left\|A^{-1}\right\|$$

donde $\|\cdot\|$ es una norma matricial. Si la solución de $Ax = b$ es poco sensible a cambios pequeños en el lado derecho b, entonces pequeñas perturbaciones en b sólo ocasionarán leves perturbaciones en los cálculos de las soluciones de x. En este caso se dice que A está bien condicionada. Esta situación corresponde al caso en el que el número condición $\kappa(A)$ es pequeño. Por otra parte, si el número de condición es grande, entonces A está mal condicionada y cualquier solución numérica de $Ax = b$ debería manejarse con cierta cautela, se verá adelante con más detalle el tema de resolución de sistemas de ecuaciones lineales.

Convergencia

Definición 1.7 *Si E_n es el error después de n operaciones subsecuentes*

$$\begin{aligned} \text{Si } |E_n| &\approx C \cdot n \cdot \varepsilon, C \text{ constante} \implies \text{Crecimiento de error lineal} \\ \text{Si } |E_n| &\approx k^n \cdot \varepsilon,\ k > 1 \implies \text{Crecimiento de error exponencial} \end{aligned}$$

El crecimiento lineal es normalmente evitable, y es aceptable si C y ε son pequeños. Pero hay que evitar la propagación del error de forma exponencial que crece muy rápidamente.

Definición 1.8 *Un algoritmo* ***converge*** *si genera una serie de pasos que nos llevan a la solución o se acerca a esta todo lo que se quiera, es decir, si el error tiende a* 0 *cuando el número de pasos tiende a* ∞.

Desde el punto de vista computacional, un algoritmo es lineal si el tiempo que necesita es proporcional al número de operaciones (cuadrático, cúbico, exponencial, etc.....).

Un algoritmo puede considerarse como una sucesión de valores y podemos aplicar los órdenes de convergencia.