CAPÍTULO 2

EJERCICIOS RESUELTOS: ARITMÉTICA DE ORDENADORES Y ANÁLISIS DE ERRORES

Ejercicios resueltos

Ejercicios 2.1 *Calcula la suma y la resta de los números* $a = 0.4523 \cdot 10^4$*, y* $b = 0.2115 \cdot 10^{-3}$*, con una aritmética flotante con mantisa de cuatro dígitos decimales, es decir, una aritmética de cuatro dígitos de precisión. ¿Se produce alguna diferencia cancelativa?*

Solución. El cálculo es fácil y directo

$$\begin{aligned} fl(a+b) &= 0.4523 \cdot 10^4 + 0.000\,2115 \cdot 10^0 \\ &= 0.4523 \cdot 10^4 + 0.0000\,000\,2115 \cdot 10^4 = 0.4523 \cdot 10^4, \\ fl(a-b) &= 0.4523 \cdot 10^4. \end{aligned}$$

Estos cálculos muestran claramente la pérdida de dígitos significativos en las operaciones de suma y resta en punto flotante. Observamos que en el caso de la resta no se ha producido una diferencia cancelativa, ya que el resultado tiene una exactitud igual a la precisión (4 dígitos) de la aritmética usada.

Ejercicios 2.2 *Usando aritmética de cuatro dígitos de precisión, sume la siguiente expresión*

$$0.1025 \cdot 10^4 + (-0.9123) \cdot 10^3 + (-0.9663) \cdot 10^2 + (-0.9315) \cdot 10^1$$

tanto ordenando los números de mayor a menor (en valor absoluto), como de menor a mayor. ¿Cuál de las dos posibilidades es más exacta? Justifique los resultados que encuentre.

Solución. La suma exacta s_E es

$$s_E = 1025 - 912.3 - 96.63 - 9.315 = 6.755\,.$$

Nuestra experiencia en exámenes nos ha mostrado que algunos alumnos contestan este ejercicio de forma incorrecta. Para sumar en orden de mayor a menor, que es el que aparece originalmente en dicha suma, primero igualan los exponentes de los números al mayor de ellos,

$$s = 0.1025 \cdot 10^4 - 0.0912\underline{3} \cdot 10^4 - 0.00966\underline{3} \cdot 10^4 - 0.0009\underline{315} \cdot 10^4,$$

donde los dígitos subrayados no entran dentro de la mantisa, por lo que los redondean,

$$s = 0.1025 \cdot 10^4 - 0.0912 \cdot 10^4 - 0.0097 \cdot 10^4 - 0.0009 \cdot 10^4.$$

y finalmente los suman con aritmética exacta obtendremos $s = 0.0007 \cdot 10^4$. Obviamente, esta respuesta es incorrecta ya que un ordenador realiza cada operación de forma separada, igualando exponentes y normalizando el resultado en cada paso. La propiedad asociativa de la suma no se cumple para la aritmética flotante.

La respuesta correcta requiere evaluar con el orden

$$(((0.1025 \cdot 10^4 + (-0.9123) \cdot 10^3) + (-0.9663) \cdot 10^2) + (-0.9315) \cdot 10^1),$$

y se obtiene mejor paso a paso como sigue

$$\begin{aligned}
s_1 &= 0.1025 \cdot 10^4, \\
s_2 &= s_1 - 0.0912 \cdot 10^4 = 0.0113 \cdot 10^4 = 0.1130 \cdot 10^3, \\
s_3 &= s_2 - 0.0966\underline{3} \cdot 10^3 = 0.1130 \cdot 10^3 - 0.0966 \cdot 10^3 \\
&= 0.0164 \cdot 10^3 = 0.1640 \cdot 10^2, \\
s_4 &= s_3 - 0.0931\underline{5} \cdot 10^2 = 0.1640 \cdot 10^2 - 0.0932 \cdot 10^2 \\
&= 0.0708 \cdot 10^2 = 0.7080 \cdot 10^1 = 7.080.
\end{aligned}$$

El error relativo cometido sumando estos números de mayor a menor es

$$\frac{s_4 - s_E}{s_E} = \frac{7.080 - 6.755}{6.755} = 0.048 \approx 5\%.$$

Si sumamos en orden de menor a mayor (en valor absoluto),

$$(((-0.9315) \cdot 10^1 + (-0.9663) \cdot 10^2) + (-0.9123) \cdot 10^3) + 0.1025 \cdot 10^4,$$

obtenemos, paso a paso,

$$
\begin{aligned}
s_1' &= -0.9315 \cdot 10^1, \\
s_2' &= s_1' - 0.9663 \cdot 10^2 = -0.0931\underline{5} \cdot 10^2 - 0.9663 \cdot 10^2 \\
&\approx -0.0932 \cdot 10^2 - 0.9663 \cdot 10^2 = -1.059\underline{5} \cdot 10^2 = -0.1060 \cdot 10^3, \\
s_3' &= s_2' - 0.9123 \cdot 10^3 = -0.1060 \cdot 10^3 - 0.9123 \cdot 10^3 \\
&= -1.018\underline{3} \cdot 10^3 = -0.1018 \cdot 10^4, \\
s_4' &= s_3' + 0.1025 \cdot 10^4 = -0.1018 \cdot 10^4 + 0.1025 \cdot 10^4 \\
&= 0.0007 \cdot 10^4 = 0.7000 \cdot 10^1 = 7.
\end{aligned}
$$

El error relativo cometido sumando los números de menor a mayor es

$$\frac{s_4' - s_E}{s_E} = \frac{7 - 6.755}{6.755} = 0.036 \approx 4\%,$$

que es algo menor que el obtenido sumando los números en el orden original (de mayor a menor).

Hemos observado que sumar los números de menor a mayor, en valor absoluto, conduce a una respuesta más exacta. Un análisis de error de la suma nos indica que para sumar números, todos del mismo signo, conviene hacerlo ordenándolos de menor a mayor módulo, ya que ello reduce la cota del error progresivo del resultado. En nuestro caso hemos observado que incluso cuando casi todos los sumandos son del mismo signo, también es recomendable esta regla.

Ejercicios 2.3 *Acote mediante propagación de errores hacia adelante el error relativo cometido en la operación flotante de suma de números reales. Aproxímelo utilizando el épsilon de la máquina.*

Solución. Para calcular la suma $x + y$ de dos números, habrá que representar éstos como números flotantes

$$fl(x) = x\,(1 + \delta_x), \qquad fl(y) = y\,(1 + \delta_y), \qquad |\delta_x|, |\delta_y| \leq u,$$

donde δ_x, y δ_y son sus errores relativos de redondeo, y u es la unidad de redondeo. El modelo estándar de la aritmética para acotar el error de la suma es

$$fl(\hat{x} + \hat{y}) = (\hat{x} + \hat{y})\,(1 + \delta_s), \qquad |\delta_s| \leq u, \quad \hat{x}, \hat{y} \in \mathbb{F}.$$

Introducir este modelo implica una fuente adicional de error,

$$fl(fl(x) + fl(y)) = (x\,(1+\delta_x) + y\,(1+\delta_y))\,(1+\delta_s),$$

$$= x\,(1+\delta_x)\,(1+\delta_s) + y\,(1+\delta_y)\,(1+\delta_s).$$

El error absoluto de la operación suma es igual a

$$fl(fl(x) + fl(y)) - (x+y) = (x+y)\,\delta_s + x\,\delta_x\,(1+\delta_s) + y\,\delta_y\,(1+\delta_s), \tag{2.1}$$

que podemos acotar como

$$|fl(fl(x) + fl(y)) - (x+y)| \leq (|x|+|y|)\,u + (|x|+|y|)\;u\,(1+u) = 2\,(|x|+|y|)\,u + O\big(u^2\big)\,,$$

que utilizando la unidad de redondeo ajustada, $\tilde{u} = 1.01\,u$, nos permite obtener

$$|fl(x+y) - (x+y)| \leq (|x|+|y|)\,2\,\tilde{u}.$$

Finalmente, podemos obtener la cota pedida para el error relativo de la operación suma

$$\frac{|fl(x+y) - (x+y)|}{|x+y|} \leq \frac{|x|+|y|}{|x+y|}\,2\,\tilde{u}, \tag{2.2}$$

que podemos escribir en función del épsilon de la máquina, ε,

$$\frac{|fl(x+y) - (x+y)|}{|x+y|} \leq \frac{|x|+|y|+2}{|x+y|}\,1.01\,\frac{\varepsilon}{2}.$$

Se podría haber calculado el error relativo exacto de la suma, que no es mucho más difícil. Tomando

$$fl(x+y) = (x+y)\,(1+\delta),$$

y comparando con la expresión (2.1), se tiene

$$\delta = \frac{(x+y)\,\delta_s + \delta_x\,(1+\delta_s) + \delta_y\,(1+\delta_s)}{x+y},$$

cuya cota ya hemos obtenido en (2.2).

Desde el punto de vista del análisis de errores, el factor 2 en la cota(2.2) es poco importante. Wilkinson [1, 2] nos indica que lo importante no es el valor exacto de la constante de error si no su orden de magnitud, en este caso $O(u) \equiv O(\epsilon)$, así como su dependencia respecto a los datos a través del número de condicionamiento, en este caso, podemos obviar el facto 2, y tomar

$$\kappa\{x+y\} = \frac{|x|+|y|}{|x+y|}.$$

De hecho, el orden de magnitud de este número no se altera si éste se multiplica por 2. Así que, a la hora de realizar una interpretación del resultado de un análisis de errores, estas constantes pueden no ser tenidas en cuenta, ya que el número de condición es grande o pequeño independiente de las mismas.

A partir de la cota obtenida, observamos que habrá un error relativo muy grande cuando $|x| + |y| \gg |x + y|$, que conduce a la condición $x \approx -y$, que equivale a que se produzca una diferencia cancelativa. En muchos casos el resultado es exacto dentro de la precisión de la aritmética, $fl(fl(x)+fl(y)) = (x+y)\,(1+\delta_s)$, $|\delta_s| \leq u$; algunos autores denominan en este caso a la diferencia cancelativa "benigna". Aún así se puede producir una pérdida de dígitos significativos en el resultado, que como sabemos, si el resultado es utilizado en cálculos posteriores, puede hacer que ésta se convierta en "maligna" o catastrófica.

En los ordenadores cuyos coprocesadores matemáticos utilizan un número suficiente de dígitos de reserva, se puede garantizar que el error relativo de la suma δ_s está acotado por $|\delta_s| \leq u$, lo que minimiza, pero no evita completamente, los efectos de las diferencias cancelativas.

Ejercicios 2.4 *Estime mediante propagación de errores hacia adelante el error relativo cometido en la operación de multiplicación flotante de números reales en función de los errores absolutos de los datos iniciales.*

Solución. El análisis de errores, es uno los temas más temidos por los alumnos a la hora de resolver los exámenes de este curso. En él aparecen una serie de ambigüedades, que a manos del inexperto, conducen a una serie de contradicciones que llevan al desconcierto, al rechazo de resultados correctos, y a la presentación como válida de otros incorrectos. Vamos a resolver este problema de varias maneras, en la línea de los resultados presentados por alumnos en cursos precedentes. Con ello pretendemos que al alumno aprenda de sus "errores" a la hora de realizar un análisis de errores correcto.

Algunos alumnos se amparan en el modelo estándar modificado de la aritmética que hemos presentado en este curso, que nos permite escribir[1], para $\hat{x}, \hat{y} \in \mathbb{F}$,

$$fl(\hat{x}\,\hat{y}) = \hat{x}\,\hat{y} + \Delta\hat{x}\hat{y} = \frac{\hat{x}\,\hat{y}}{1+\delta_{\hat{x}\hat{y}}}, \qquad |\delta_{\hat{x}\hat{y}}| \leq u, \quad |\Delta\hat{x}\hat{y}| \leq |\hat{x}\,\hat{y}|\,u.$$

Utilizando este modelo y considerando los errores absolutos en los datos iniciales directamente escriben[2]

$$fl(x\,y) = x\,y + \Delta xy = (x+\Delta x)\,(y+\Delta y), \qquad |\Delta x| \leq |x|\,u, \quad |\Delta y| \leq |y|\,u.$$

[1] Muchos alumnos olvidan esta importante hipótesis.

[2] Esta expresión es errónea porque x e y no son números flotantes, no tiene en cuenta la condición $fl(x), fl(y) \in \mathbb{F}$

Continuando con su análisis,

$$fl(x\,y) - x\,y = \Delta xy = x\,\Delta x + y\,\Delta y + \Delta x\,\Delta y,$$

el error relativo pedido es

$$\delta_{xy} = \frac{fl(x\,y) - x\,y}{x\,y} = \frac{\Delta xy}{x\,y} = \frac{\Delta x}{x} + \frac{\Delta y}{y} + \frac{\Delta x}{x}\,\frac{\Delta y}{y} = \delta_x + \delta_y + \delta_x\,\delta_y,$$

es decir, la suma de los errores relativos de los datos más el producto de éstos. Este resultado recuerda al obtenido en el análisis de errores experimentales que han estudiado en la asignatura de Física para la interpretación de experimentos. Acotando este resultado obtienen

$$|\delta_{xy}| \leq 2\,u + O\big(u^2\big) = 2\,\tilde{u}. \tag{2.3}$$

Aparentemente sólo hay dos fuentes de error, aunque las operaciones implicadas requieren tres pasos de normalización de números, para los operandos, x e y, y para el resultado del producto $x\,y$, ello lleva a los alumnos a pensar, correctamente, que este resultado está mal. Muchos no se dan cuenta del paso erróneo que han cometido al aplicar el modelo estándar modificado de la aritmética sin tener en cuenta sus hipótesis.

Un análisis correcto, siguiendo esta línea, nos lleva a escribir,

$$fl(fl(x)\,fl(y)) = fl(x)\,fl(y) + \Delta xy = (x + \Delta x)\,(y + \Delta y) + \Delta xy,$$

donde[3]

$$|\Delta xy| \leq |(x + \Delta x)\,(y + \Delta y)|\,u \leq |x\,y|\,u + O\big(u^2\big)\,,$$

por tanto,

$$\delta_{xy} = \frac{fl(fl(x)\,fl(y)) - x\,y}{x\,y} = \frac{\Delta x}{x} + \frac{\Delta y}{y} + \frac{\Delta x}{x}\,\frac{\Delta y}{y} + \frac{\Delta xy}{x\,y},$$

con lo que obtenemos como cota del error

$$|\delta_{xy}| \leq 3\,u + O\big(u^2\big) = 3\,\tilde{u}.$$

Esta cota, que es correcta, debe resultar mucho más razonable para el alumno, como ya hemos indicado previamente, porque se han realizado 3 normalizaciones de números flotantes para obtener el resultado del producto.

Otros alumnos, en exámenes, resuelven este problema de una forma diferente. Toman el error relativo para la operación de multiplicación a partir de los errores relativos de los datos,

$$\begin{aligned} fl(fl(x)\,fl(y)) &= x\,(1 + \delta_x)\,y\,(1 + \delta_y)\,(1 + \delta_m) \\ &= x\,y\,(1 + \delta_x)\,(1 + \delta_y)\,(1 + \delta_m), \qquad |\delta_x|, |\delta_y| \leq u, \end{aligned}$$

[3]Es importante no olvidar este punto.

donde utilizando el modelo estándar de la aritmética, correctamente, hacen $|\delta_m| \leq u$, con lo que operando

$$\begin{aligned} fl(fl(x)\, fl(y)) &= x\, y\, (1 + \delta_x + \delta_y + \delta_m + \delta_x\, \delta_y + \delta_x\, \delta_m + \delta_y\, \delta_m + \delta_x\, \delta_y\, \delta_m) \\ &= x\, y\, (1 + \delta_p), \end{aligned}$$

y acotando, obtienen el resultado correcto

$$|\delta_p| \leq |\delta_x + \delta_y + \delta_m| + O\big(u^2\big) \leq 3\, u + O\big(u^2\big)\,. \tag{2.4}$$

Sin embargo, como el enunciado pide escribir este resultado en función de los errores absolutos de los datos, introducen éstos,

$$\begin{aligned} fl(x) &= x + \Delta x, \qquad & |\Delta x| &\leq |x|\, u, \\ fl(y) &= y + \Delta y, \qquad & |\Delta y| &\leq |y|\, u, \end{aligned}$$

acabando con la expresión

$$|\delta_p| \leq \frac{|\Delta_x|}{|x|} + \frac{|\Delta_y|}{|y|} + |\delta_m| + O\big(u^2\big)\,.$$

donde necesitan conocer el error relativo δ_m del producto en función del error absoluto de los datos. Al estimar esta cota, muchos alumnos cometen errores.

Algunos alumnos aplican el resultado aparentemente razonable,

$$|\delta_m| \leq \frac{|\Delta_x|\, |\Delta_y|}{|x|\, |y|} = O\big(u^2\big)\,,$$

que conduce al resultado incorrecto

$$|\delta_{xy}| \equiv |\delta_p| \leq 2\, u + O\big(u^2\big)\,.$$

Sin embargo, este resultado está en contradicción con (2.4). Ello les hace dudar del análisis realizado, muchas veces sin ser capaces de encontrar el paso erróneo.

Otros alumnos, sin embargo, calculan una cota para $|\delta_m|$ siguiendo un razonamiento similar al usado al principio de este problema, obteniendo

$$|\delta_m| \leq 2\, u + O\big(u^2\big)\,,$$

lo que conduce al resultado, también incorrecto,

$$|\delta_{xy}| \equiv |\delta_p| \leq 4\, u + O\big(u^2\big)\,. \tag{2.5}$$

Este resultado también está en contradicción con (2.4). De nuevo, éstos también dudan sobre el resultado obtenido sin encontrar la fuente de su error.

El hecho de que el análisis de errores, a veces, en manos del inexperto, conduzca a expresiones como (2.3), (2.4) y (2.5), que son contradictorias entre sí, hace que muchos alumnos le tengan gran temor, sobre todo en los exámenes, conduciendo a resultados pobres en éstos. El alumno debe aplicar su intuición, que le indica que la constante de error debe ser similar al número de operaciones de normalización de números flotantes realizadas.

Ejercicios 2.5 *La operación de suma de números flotantes no cumple con la propiedad asociativa, aunque sí es conmutativa, es decir, el orden de los factores, si hay más de dos, altera el resultado y, por tanto, el error de éste. Demostrar que si se suman varios números positivos empezando por el menor y en orden creciente se minimiza la pérdida de dígitos significativos en el resultado*[4].

Solución. Calculemos mediante propagación de errores hacia adelante el error cometido al sumar n números x_i,

$$s = x_1 + x_2 + \cdots + x_n.$$

Introduzcamos las sumas parciales s_i que nos indican el orden en que se realizan las sumas

$$s_2 = x_1 + x_2, \qquad s_3 = s_2 + x_3, \qquad \ldots, \qquad s_n = s_{n-1} + x_n.$$

Estudiemos como se propagan los errores relativos en estas sumas parciales. Es importante que el lector note que en el enunciado se plantea el estudio de la suma de números flotantes, y no de números reales. Operando y despreciando los productos de errores relativos $\epsilon_i\,\epsilon_j$ como infinitésimos de orden superior,

$$\begin{aligned}
fl(s_2) &= (x_1 + x_2)\,(1 + \epsilon_2),\\
fl(s_3) &= (fl(s_2) + x_3)\,(1 + \epsilon_3)\\
&= x_3\,(1 + \epsilon_3) + (x_1 + x_2)\,(1 + \epsilon_2)\,(1 + \epsilon_3)\\
&= x_3\,(1 + \epsilon_3) + (x_1 + x_2)\,(1 + \epsilon_2 + \epsilon_3)\\
&= s_3 + (x_1 + x_2)\,(\epsilon_2 + \epsilon_3) + x_3\,\epsilon_3,
\end{aligned}$$

y siguiendo con el mismo procedimiento

$$fl(s_4) = (fl(s_3) + x_4)\,(1 + \epsilon_4)$$

[4]Este ejercicio ya ha sido resuelto en el contenido teórico del segundo tema.

$$= s_4 + (x_1 + x_2)(\epsilon_2 + \epsilon_3 + \epsilon_4) + x_3(\epsilon_3 + \epsilon_4) + x_4\,\epsilon_4.$$

La fórmula general que se obtiene es

$$fl(s_n) = s_n + (x_1 + x_2)\sum_{i=2}^{n}\epsilon_i + x_3\sum_{i=3}^{n}\epsilon_i + \cdots + x_n\,\epsilon_n.$$

Acotando $|\epsilon_i| \leq u$, la unidad de redondeo, tenemos finalmente

$$fl(s_n) \leq s_n + (x_1 + x_2)(n-1)\,u + x_3(n-2)\,u + \cdots + x_n\,u + O\left(u^2\right),$$

donde no aparecen errores absolutos porque todos los números son positivos, y hemos incluido un término cuadrático para recordar los errores que hemos despreciado en pasos anteriores.

En la expresión obtenida se observa que el error los primeros sumandos afectan más al resultado que los últimos. Por ello, si sumamos primero los números más pequeños, que coinciden con los de menor módulo, haremos que el error de redondeo de la suma sea menor.

Si los números a sumar no son todos positivos (o todos negativos), el orden que minimiza el error es el que minimiza la sumas parciales $|s_i|$. Encontrar este orden es difícil, como ya indicamos en la teoría de este tema.

Ejercicios 2.6 *Evalúe (con 5 dígitos tras la coma decimal) la función e^x cuando $x = 5$, y $x = -5$, utilizando desarrollos en serie de Taylor. Si la convergencia del desarrollo en serie de Taylor es muy lenta, proponga un método más preciso para dicha evaluación.*

Solución. El resultado exacto (redondeado a 5 dígitos tras la coma decimal) que se obtiene utilizando aritmética IEEE de doble precisión es

$$e^5 = 148.41316, \qquad e^{-5} = 0.0067379.$$

El desarrollo en serie de Taylor de la exponencial es

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + O\left(x^4\right). \tag{2.6}$$

Esta serie converge para $x = 5$ como se prueba fácilmente mediante el criterio de Cauchy, ya que, denotando por t_n el término n-ésimo de la serie,

$$\frac{t_{n+1}}{t_n} = \frac{x^{n+1}/(n+1)!}{x^n/n!} = \frac{x}{n+1} = \frac{5}{n+1} \leq 1,$$

que se cumple para $n \geq 4$.

Definamos (con $0! = 1$) la sucesión de sumas parciales

$$s(n) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

Entonces obtenemos, operando con cinco dígitos decimales,

$$s(0) = 1.00000, \qquad s(1) = 1 + 5 = 6.00000,$$

$$s(2) = 6 + \frac{5^2}{2} = 18.50000,$$

$$s(3) = s(2) + \frac{5^3}{3!} = 39.33333,$$

$$s(4) = s(3) + \frac{5^4}{4!} = 65.37500,$$

$$s(5) = s(4) + \frac{5^5}{5!} = 91.41667,$$

$$s(6) = s(5) + \frac{5^6}{6!} = 113.11806,$$

$$s(7) = s(6) + \frac{5^7}{7!} = 128.61905,$$

$$s(8) = s(7) + \frac{5^8}{8!} = 138.30717,$$

$$s(9) = s(8) + \frac{5^9}{9!} = 143.68946,$$

$$s(10) = s(9) + \frac{5^{10}}{10!} = 146.38060,$$

$$s(11) = s(10) + \frac{5^{11}}{11!} = 147.60385,$$

$$s(12) = s(11) + \frac{5^{12}}{12!} = 148.11354,$$

$$s(13) = s(12) + \frac{5^{13}}{13!} = 148.30957,$$

$$\ldots$$

Comparando con la solución exacta, el error relativo cometido hasta ahora es

$$\frac{e^5 - s(13)}{e^5} = 0.0007\,.$$

Como vemos, la serie de Taylor permite calcular el valor de la exponencial para $x > 0$ con gran precisión, aunque requiere un gran número de operaciones aritméticas.

La serie (2.6) para $x = -5$ es una serie alternada que converge, ya que la serie de los valores absolutos de sus términos converge, como ya se ha probado anteriormente. Sin embargo, la convergencia de una serie alternada suele ser extremadamente lenta. Realicemos algunos cálculos

$$s(0) = 1.00000, \qquad s(1) = 1 - 5 = -4.00000,$$

$$s(2) = s(1) + \frac{5^2}{2} = 8.50000,$$

$$s(3) = s(2) - \frac{5^3}{3!} = -12.33333,$$

$$\dots$$

Para calcular el valor pedido es mejor utilizar

$$e^{-x} = \frac{1}{e^x} = \frac{1}{1 + x + \frac{x^2}{2} + \cdots},$$

que en nuestro caso da

$$e^{-5} \approx \frac{1}{s(13)} = \frac{1}{148.30957} = 0.0067427$$

cuyo error relativo es

$$\frac{e^{-5} - 1/s(13)}{e^{-5}} = 0.0007,$$

que es el mismo que el que obtuvimos previamente para e^5.

Ejercicios 2.7 *Dada*

$$\phi(x) = \sum_{k=1}^{\infty} \frac{1}{k\,(k+x)}.$$

Demuestre que $\phi(1) = 1$.

Solución. Factorizando la expresión a sumar

$$\phi(x) = \sum_{k=1}^{\infty} \frac{1}{k\,(k+x)} = \sum_{k=1}^{\infty} \left(\frac{1}{k} - \frac{1}{k+x}\right) \frac{1}{x} = \frac{1}{x} \left(\sum_{k=1}^{\infty} \frac{1}{k} - \sum_{k=1}^{\infty} \frac{1}{k+x}\right)$$

que para $x = 1$

$$\phi(1) = \left(\sum_{k=1}^{\infty} \frac{1}{k} - \sum_{k=1}^{\infty} \frac{1}{k+1}\right) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots - \frac{1}{2} - \frac{1}{3} - \cdots = 1.$$

Es muy importante que note que hemos podido hacer esta suma término a término porque la suma original es absolutamente convergente, como se demuestra fácilmente. En caso contrario no se podría haber realizado una suma término a término.

Ejercicios 2.8 *¿Cuál es el número de condicionamiento de $f(x) = e^x$ para $x < 0$? Compare este número de condicionamiento con los que resultan de la evaluación de $f(x) = e^x$ por medio de desarrollos de Taylor.*

Solución. El número de condicionamiento de $f(x) = e^x$ para $x < 0$, con $|x - x^*|$ pequeño es

$$\max \left| \frac{f(x) - f(x^*)}{f(x)} : \frac{x - x^*}{x} \right| \approx \left| \frac{f'(x)}{f(x)} \, x \right| = |x|,$$

lo que indica que el número de condicionamiento aumenta linealmente con $|x|$.

Si escribimos el desarrollo en serie de Taylor de la exponencial

$$f(x) = e^x = \sum_{n=0}^{\infty} f_n(x), \qquad f_n(x) = \frac{x^n}{n!},$$

y calculamos el número de condicionamiento de un término general de dicha serie $f_n(x)$, obtenemos aproximadamente

$$\left| \frac{n \, \dfrac{x^{n-1}}{n!}}{\dfrac{x^n}{n!}} \, x \right| = n,$$

que aumenta a medida que aumenta el orden n del término de la serie. Más aún, para $x < 0$, la serie de Taylor de e^x es una serie alternada para la que

$$\left| \frac{f_{n+1}}{f_n} \right| = \left| \frac{x^{n+1}/(n+1)!}{x^n/n!} \right| = \frac{x}{n+1},$$

por lo que, aunque la serie es convergente, para $|x|$ grande se requieren un gran número de términos.

Ejercicios 2.9 *Determine el número de condicionamiento para la evaluación de la función e^x para $x < 0$. Para los valores de x para los que este problema está mal condicionado, cómo evaluaría la exponencial (utilice desarrollo en serie de Taylor).*

Solución. Dado que $f(x) = e^x = f'(x)$, su número de condicionamiento es

$$\left| \frac{f(x + \Delta x) - f(x)}{f(x)} \, \frac{x}{\Delta x} \right| \approx \left| \frac{f'(x)}{f(x)} \, x \right| = |x|.$$

El número de condicionamiento crece conforme x crece.

Podemos evaluar e^x mediante su desarrollo de Taylor

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots,$$

que es convergente para toda $x \in \mathbb{R}$. Para $x < 0$ tenemos una serie de términos alternados, donde para $x \ll 0$ el valor absoluto de cada término crece indefinidamente. Por lo tanto, su evaluación numérica es difícil ya que se trata de una serie de convergencia lenta que requiere el cálculo de un gran número de términos para evaluar e^x con suficiente precisión $\forall x < 0$.

Para determinar el número de términos que tenemos que calcular, definamos la suma parcial de la serie

$$s_n = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!}.$$

El error cometido al aproximar e^x por s_n es

$$e^x - s_n = \frac{x^{n+1}}{(n+1)!} + O\big(x^{n+2}\big) = \frac{x^{n+1}}{(n+1)!}e^{\xi},$$

donde $0 \geq \xi \geq x$ y hemos aplicado el teorema del valor medio. Por tanto,

$$|e^x - s_n| \leq \left|\frac{x^{n+1}}{(n+1)!}\right|$$

que podemos hacer tan pequeño como deseemos haciendo n suficientemente grande dado que el factorial crece más rápido que cualquier potencia. Para obtener una precisión inferior al épsilon de la máquina hay que calcular s_n sucesivamente hasta que $s_n = s_{n-1}$.

Sin embargo, es más eficiente computacionalmente aproximar la exponencial de la siguiente forma

$$e^x = \frac{1}{e^{-x}} = \frac{1}{1 - x + \frac{x^2}{2} - \frac{x^3}{3!} + \cdots}$$

Para $x < 0$, todos los términos del denominador son positivos (de hecho, la exponencial es una función no negativa). Es mejor calcular la sucesión

$$s_n = \frac{1}{\sum_{i=0}^{n}(-1)^i x^i/i!},$$

donde hemos usado el convenio habitual $0! = 1$.

Ejercicios 2.10 *Cómo se debe evaluar la función*

$$f(x) = x - \sqrt{x^2 - \alpha}$$

para $\alpha \ll x$, de forma tal que se eviten diferencias cancelativas.

Solución. Propondremos dos maneras de resolver este problema. Por un lado, podemos desarrollar la raíz cuadrada mediante serie de Taylor,

$$\begin{aligned} f(x) &= x\left(1-\sqrt{1-\frac{\alpha}{x^2}}\right) \\ &= x\left(1-\left(1-\frac{\alpha}{2\,x^2}+O\left(\frac{\alpha^2}{x^4}\right)\right)\right) \\ &= \frac{\alpha}{2\,x}+O\left(\frac{\alpha^2}{x^3}\right). \end{aligned}$$

Por otro lado, sin utilizar Taylor, podemos aplicar de forma exacta

$$\begin{aligned} f(x) = x-\sqrt{x^2-\alpha} &= \frac{\left(x-\sqrt{x^2-\alpha}\right)\left(x+\sqrt{x^2-\alpha}\right)}{x+\sqrt{x^2-\alpha}} \\ &= \frac{\alpha}{x+\sqrt{x^2-\alpha}}. \end{aligned}$$

Aunque las dos expresiones que hemos obtenido son diferentes, la segunda expresión tiende a la primera cuando $x \gg \alpha$. Aunque la primera expresión es aproximada y la segunda exacta, la primera tiene la ventaja de que es computacionalmente más eficiente, y en la mayoría de los casos el error es despreciable cuando $x \gg \alpha$.

Ejercicios 2.11 *Calcule*

$$f(x) = \frac{x-\sin x}{\tan x}$$

para $x = 0.000001$*, con una exactitud de cuatro cifras decimales.*

Solución. Para calcular

$$f(x) = \frac{x-\sin x}{\tan x}$$

con $x = 10^{-6}$ utilizaremos la calculadora de Windows (que trabaja hasta con 16 dígitos decimales). El resultado es

$$f(10^{-6}) = \frac{2\cdot 10^{-19}}{10^{-6}} = 2\cdot 10^{-13}.$$

¿Cuántos dígitos significativos tiene este resultado? La mejor manera de determinarlos, dado que x es muy pequeño, es utilizar la serie de Taylor de $f(x)$ y cuantificar el error cometido mediante el teorema del resto de Taylor.

El desarrollo de Taylor del numerador es

$$x-\sin x = -\frac{x^3}{3!}+\frac{x^5}{5!}+O\left(x^7\right)$$

y el del denominador

$$\tan x = x + O\big(x^3\big)\,.$$

Para $x = 10^{-6}$,

$$x - \sin x = 10^{-18}\,(\frac{1}{3!} + O\big(10^{-12}\big)), \qquad \tan x = 10^{-6} + O\big(x^{-18}\big) = 10^{-6}$$

por lo que podemos aproximar, con más de cuatro cifras de exactitud,

$$f(x) \approx \frac{x^3/3!}{x} = \frac{x^2}{6} = 0.16667 \cdot 10^{-12}$$

ya que el siguiente término del desarrollo de Taylor es $O\big(10^{-24}\big)$.

Como podemos ver, la solución obtenida con la calculadora es bastante mala y tiene un error relativo muy alto

$$\left|\frac{2 - 1.6667}{1.6667}\right| = 0.2 \approx 20\,\%.$$

Ejercicios 2.12 *Dadas $f(x) = e^x$ y $g(x) = x$ en el intervalo $[0, 1]$. ¿Para qué valores de ξ se satisfacen las siguientes condiciones?*

1. $\int_0^1 f(x)\,dx = f(\xi)$,
2. $\int_0^1 g(x)\,dx = g(\xi)$,
3. $\int_0^1 f(x)\,g(x)\,dx = f(\xi)\,\int_0^1 g(x)\,dx$.

Solución. Dado que $f(x) = e^x$ y $g(x) = x$ son funciones continuas, tenemos que

$$\int_0^1 e^x\,dx = e^x|_0^1 = e - 1 = f(\xi) = e^\xi$$

$$\xi = \ln(e - 1) = 0.541,$$

$$\int_0^1 x\,dx = \frac{x^2}{2}\bigg|_0^1 = \frac{1}{2} = g(\xi) = \xi$$

$$\xi = \frac{1}{2},$$

$$\int_0^1 f(x)\,g(x)\,dx = f(\xi)\,\int_0^1 g(x)\,dx,$$

$$\int_0^1 x\,e^x\,dx = x\,e^x|_0^1 - \int_0^1 e^x\,dx = e - e^x|_0^1 = 1$$

$$f(\xi)\,\int_0^1 x\,dx = f(\xi)\,\frac{1}{2} = e^\xi\,\frac{1}{2} = 1,$$

$$\xi = \ln 2 = 0.693.$$

La tercera relación presentada en el enunciado del problema sólo es cierta por que $g(x) = x$ en $[0, 1]$ tiene el mismo signo que $f(x) = e^x$ en dicho intervalo. En caso contrario, dicha relación no sería verdad, por ejemplo, para el intervalo $[-1, 1]$,

$$\int_{-1}^{1} f(x)\, g(x)\, dx = \int_{-1}^{1} x\, e^x\, dx = x\, e^x|_{-1}^{1} - \int_{-1}^{1} e^x\, dx =$$

$$e + e^{-1} - e^x|_0^1 = e + e^{-1} - e + e^{-1} = \frac{2}{e},$$

$$f(\xi) \int_{-1}^{1} g(x)\, dx = \quad f(\xi) \int_{-1}^{1} x\, dx = f(\xi)\, 0 = 0,$$

$$\text{y } \frac{2}{e} \neq 0.$$

Ejercicios 2.13 *Resuelva el sistema de dos ecuaciones lineales*

$$0.780\, x + 0.563\, y = 0.217,$$

$$0.457\, x + 0.330\, y = 0.127,$$

con cuatro y con tres cifras significativas, y compare los resultados con los de la solución exacta. Justifique los resultados obtenidos. Nota: si utiliza una calculadora, redondee los resultados intermedios.

Solución. Para resolver el sistema lineal

$$0.780\, x + 0.563\, y = 0.217,$$

$$0.457\, x + 0.330\, y = 0.127,$$

primero operamos con cuatro cifras significativas. Despejando de la primera ecuación

$$x = 0.2782 - 0.7218\, y,$$

y sustituyendo en la segunda

$$0.457\, (0.2782 - 0.7218\, y) + 0.330\, y = 0.127,$$

$$0.1271 - 0.3299\, y + 0.330\, y = 0.127,$$

$$0.0001\, y = -0.0001,$$

con lo que, finalmente,

$$y = -1, \qquad x = 1.$$

Seguidamente operaremos con tres cifras significativas. Despejando de nuevo de la primera ecuación

$$x = 0.278 - 0.722\,y,$$

y sustituyendo en la segunda

$$0.457\,(0.278 - 0.722\,y) + 0.330\,y = 0.127,$$

$$0.127 - 0.330\,y + 0.330\,y = 0.127,$$

$$0.000\,y = 0.000,$$

con lo que el valor de y está indeterminado, y el sistema no se puede resolver.

Para calcular el valor exacto de la solución utilizamos la regla de Cramer. Calculemos el determinante

$$\mathrm{Det} = \begin{vmatrix} 0.780 & 0.563 \\ 0.457 & 0.330 \end{vmatrix} = 1.09 \times 10^{-6},$$

y la solución para x

$$x = \frac{1}{\mathrm{Det}} \begin{vmatrix} 0.217 & 0.563 \\ 0.127 & 0.330 \end{vmatrix} = \frac{1.09 \times 10^{-6}}{\mathrm{Det}} = 1,$$

y para y

$$y = \frac{1}{\mathrm{Det}} \begin{vmatrix} 0.780 & 0.217 \\ 0.457 & 0.127 \end{vmatrix} = \frac{-1.09 \times 10^{-6}}{\mathrm{Det}} = -1.$$

Estos resultados indican que este problema está mal condicionado, aunque los podemos justificar debido a que el determinante (Det) es muy próximo a cero. Pero además, podemos estudiar sus autovalores. Para ello, escribimos el sistema como

$$A\,\vec{x} = \vec{b}, \qquad \vec{x} = A^{-1}\,\vec{b},$$

donde

$$\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} 0.217 \\ 0.127 \end{pmatrix}, \qquad A = \begin{pmatrix} 0.780 & 0.563 \\ 0.457 & 0.330 \end{pmatrix}.$$

Calculemos los autovalores de A,

$$|A - \lambda\,I| = 0 = \begin{vmatrix} 0.780 - \lambda & 0.563 \\ 0.457 & 0.330 - \lambda \end{vmatrix}$$

es decir,

$$2.574 \times 10^{-1} - 1.11\,\lambda + \lambda^2 - 0.257291 = 0,$$

$$\lambda^2 - 1.11\,\lambda - 0.000109 = 0,$$

$$\lambda = \frac{1.11}{2} \pm \sqrt{\left(\frac{1.11}{2}\right)^2 + 0.000109} = 0.555 \pm 0.555098190,$$

con lo que los autovalores son

$$\lambda_+ = 1.110098190, \qquad \lambda_- = -0.000098190.$$

Dado que los dos autovalores tienen magnitudes muy dispares, el problema está mal condicionado.

Ejercicios 2.14 *Dada la ecuación diferencial ordinaria*

$$\frac{d^2y}{dx^2} - y = 0, \qquad y(0) = a, \quad \frac{dy}{dx}(0) = b.$$

¿Para qué valores iniciales es el problema estable o está físicamente bien condicionado?

Solución. La ecuación diferencial ordinaria

$$y'' - y = 0, \qquad y(0) = a, \quad y'(0) = b,$$

es fácil de resolver suponiendo una solución de la forma

$$y(x) = A\,e^x + B\,e^{-x} = C\,\sinh x + D\,\cosh x,$$

$$y(0) = D = a,$$

$$y'(x) = C\,\cosh x + D\,\sinh x,$$

$$y(0) = C = b,$$

y por tanto

$$y(x) = b\,\sinh x + a\,\cosh x = b\,\frac{e^x - e^{-x}}{2} + a\,\frac{e^x + e^{-x}}{2}$$

$$= \frac{a+b}{2}\,e^x + \frac{a-b}{2}\,e^{-x}.$$

Con objeto de estudiar el condicionamiento con respecto a las condiciones iniciales, introducimos un pequeño error en a y en b,

$$y(0) = a\,(1+\epsilon_a), \qquad y'(0) = b\,(1+\epsilon_b),$$

con lo que la solución del problema perturbado es

$$y_P(x) = \frac{a\,(1+\epsilon_a) + b\,(1+\epsilon_b)}{2}\,e^x + \frac{a\,(1+\epsilon_a) - b\,(1+\epsilon_b)}{2}\,e^{-x}.$$

Comparando las dos soluciones obtenidas

$$y_P - y = \frac{a\,\epsilon_a + b\,\epsilon_b}{2}\,e^x + \frac{a\,\epsilon_a - b\,\epsilon_b}{2}\,e^{-x}.$$

Para $x \gg 0$,

$$y_P - y \approx \frac{a\,\epsilon_a + b\,\epsilon_b}{2}\,e^x, \qquad y \approx \frac{a+b}{2}\,e^x,$$

y el error relativo toma la forma

$$\frac{y_P - y}{y} \approx= \frac{a\,\epsilon_a + b\,\epsilon_b}{a+b}.$$

Está expresión se hará "muy grande" si $a+b=0$ y $\epsilon_a \neq \epsilon_b$, y en ese caso el problema está mal condicionado. Sin embargo, si $\epsilon_a = \epsilon_b = \epsilon$,

$$\frac{y_P - y}{y} \approx= \frac{a+b}{a+b}\,\epsilon = \epsilon,$$

y el problema está bien condicionado.

Otra manera de comprobar que para $a+b=0$ el problema considerado está mal condicionado es teniendo en cuenta que la solución exacta en dicho caso es

$$y = a\,e^{-x} = -b\,e^{-x},$$

que tiende a cero cuando $x \to \infty$, aunque cualquier perturbación que cause $a+b \neq 0$ hará que la solución se vuelva no acotada para $x \to \infty$.

El análisis presentado en esta solución también se podría haber realizado expresando la ecuación original de segundo grado como dos ecuaciones de primer grado. Para ello, se reescribe

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) = y,$$

con lo que se define y obtiene

$$\frac{dy}{dx} = z, \qquad z(0) = b,$$

$$\frac{dz}{dx} = y, \qquad y(0) = a.$$

Ejercicios 2.15 *Realice el análisis de errores de la operación $x_0^{2^n}$ mediante el siguiente algoritmo que parte de x_0, y procede así*

$$x_1 = x_0^2, \quad x_2 = x_1^2, \quad \ldots \quad x_n = x_{n-1}^2.$$

Solución. En el análisis se representan los números flotantes que se van calculando por y_i, de forma que se obtiene la sucesión $\{y_i\}$ donde

$$x_0 \longrightarrow y_0 = fl(x_0) = x_0\,(1+\delta_0),$$

$$x_1 \longrightarrow y_1 = fl(x_1) = fl(y_0^2) = y_0^2\,(1+\delta_1),$$

$$\cdots\cdots$$

$$x_n \longrightarrow y_n = fl(y_{n-1}^2) = y_{n-1}^2\,(1+\delta_n),$$

por lo que

$$y_n = y_0^{2^n}\,(1+\delta_1)^{2^{n-1}}\ldots(1+\delta_{n-1})^2\,(1+\delta_n),$$

que acotando todos los $|\delta_i| < u$, donde $u = \varepsilon/2$ (es la mitad del épsilon de la máquina), nos da

$$|y_n| \leq |x_n|\,(1+u)^p,$$

donde $p = 1 + 2 + \cdots + 2^n$ es un progresión geométrica. Operando del modo usual

$$p = 2^0 + 2^1 + \cdots + 2^n, \qquad 2\,p = 2^1 + \cdots + 2^{n+1},$$

$$2\,p - p = p = 2^{n+1} - 2^0 = 2^{n+1} - 1.$$

De esta forma los errores absoluto y relativo son (considerando x_0 positivo para evitar los valores absolutos)

$$|y_n - x_n| \leq x_0^{2^n}\,\left((1+u)^p - 1\right),$$

$$\left|\frac{y_n - x_n}{x_n}\right| \leq (1+u)^p - 1,$$

respectivamente, donde $p = 2^{n+1} - 1$.

Como $u \ll 1$, podemos utilizar desarrollo en serie de Taylor de la función potencia (o la regla del binomio de Newton) para acotar el error relativo como

$$1 + p\,u + O\big(u^2\big) - 1 = p\,u + O\big(u^2\big) = \left(2^{n+1} - 1\right)\,u + O\big(u^2\big)\,,$$

es decir, los errores crecen proporcionalmente al número de operaciones. Estos resultados son exactos si se utiliza la unidad de reondeo ajustada, que para $p\,u \ll 1$, nos permite escribir

$$\left|\frac{y_n - x_n}{x_n}\right| \leq p\,\tilde{u} \approx 2^{n+1}\,\tilde{u}.$$

Se deja al lector la presentación de este problema utilizando la notación de θ_n y γ_n presentado en el texto del tema.

Ejercicios 2.16 *Considere el siguiente sistema de ecuaciones lineales*

$$2\,x + 6\,y = 8,$$

$$2\,x + 6.00001\,y = 8.00001.$$

Resuélvalo con (1) la regla de Cramer y 6 cifras significativas (es decir $0.\star\star\star\star\star\star$*), y (2) la regla de Cramer y 4 cifras significativas. Explique sus resultados. ¿Es un problema bien condicionado? Justifique su respuesta.*

Solución. Cuando se trata de resolver este problema con la regla de Cramer y 6 dígitos significativos, observamos que el determinante de la matriz A del sistema es

$$\det(A) = 12.0000 - 12.0000 = 0.000000,$$

con lo que

$$x = \frac{\begin{vmatrix} 8.00000 & 6.00000 \\ 8.00001 & 6.00001 \end{vmatrix}}{\begin{vmatrix} 2 & 6 \\ 2 & 6.00001 \end{vmatrix}} = \frac{48.0000 - 48.0000}{0.000000} = \frac{0.000000}{0.000000},$$

operación que no está definida.

Cuando se trata de resolver el problema con 4 dígitos significativos obtenemos, similarmente, $\det(A) = 0.0000$. Luego tampoco podemos utilizar la regla de Cramer.

Obviamente este problema está mal condicionado ya que un pequeño cambio, por ejemplo, del orden de 10^{-5} en cualquiera de los coeficientes, trabajando con 6 dígitos signficativos, hace que la regla de Cramer nos de una posible solución. Es decir, un pequeño cambio afecta mucho al resultado.

Ejercicios 2.17 *Considere la transformación discreta*

$$x_{i+1} = a\,x_i + b.$$

Se define el "punto fijo", x_F*, de esta transformación como el número tal que* $x_F = x_{i+1} = x_i$*, es decir,* $x_F = a\,x_F + b$*, y por lo tanto,*

$$x_F = \frac{b}{1-a}.$$

Determine los valores de a para los cuales esta transformación (1) converge al punto fijo, y (2) diverge (o no converge) al punto fijo, cuando $i \to \infty$. Justifique y dé una interpretación geométrica de sus resultados.

Solución. La definición de la convergencia de la sucesión $\{x_n\} \to x_F$, requiere que $\forall \epsilon > 0$, $\exists n_0$, tal que $\forall n > n_0$, $|x_n - x_F| < \epsilon$. Como el alumno estudió en cursos anteriores, se puede aplicar el criterio de Cauchy,

$$\frac{|x_{n+1} - x_n|}{|x_n - x_{n-1}|} = |a| < 1,$$

que es la condición buscada.

También podemos analizar la convergencia de la sucesión directamente. Iterando la sucesión

$$x_{n+1} = a\, x_n + b = a^2\, x_{n-1} + a\, b + b = a^3\, x_{n-2} + a^2\, b + a\, b + b,$$

luego

$$\begin{aligned} x_n &= a\, x_{n-1} + b \\ &= a^2\, x_{n-2} + a\, b + b \\ &= a^3\, x_{n-2} + (a^2 + a + 1)\, b \\ &\cdots \\ &= a^n\, x_0 + (a^{n-1} + a^{n-2} + \cdots a + 1)\, b, \end{aligned}$$

donde aparece una progresión geométrica fácilmente sumable,

$$s = a^n + a^{n-1} + \cdots a + 1, \qquad (1 - a)\, s = 1 - a^n,$$

con lo que obtenemos

$$x_n = a^n\, x_0 + \frac{1 - a^n}{1 - a}\, b.$$

Con el término general de la sucesión en forma explícita podemos estudiar directamente su convergencia. Para $|a| < 1$, es decir, $-1 < a < 1$, este límite converge ya que

$$\lim_{n\to\infty} a^n = 0, \qquad \lim_{n\to\infty} x_n = x_F,$$

y para $|a| > 1$ diverge. Sin embargo, los casos límites deben estudiarse con cuidado. Para $a = 1$, tenemos que

$$\frac{1 - a^n}{1 - a} = 1 + a + \cdots + a^n = n + 1, \qquad \lim_{n\to\infty} a^n = 1,$$

con lo que

$$\lim_{n\to\infty} x_n = x_0 + b \lim_{n\to\infty} (n+1),$$

que diverge. Para $a = -1$, también diverge ya que la sucesión de signos alternados $\{(-1)^n\}$ no tiene límite.

En resumen, la sucesión $\{x_n\}$ converge sólo cuando la pendiente $|y'(x)| = |a|$ de la recta $y(x) = a\,x + b$, es menor que la unidad. En el tema 7 estudiaremos en más detalle la convergencia de sucesiones de este tipo y su uso para el cálculo de ceros de funciones.

Ejercicios 2.18 *Haga una análisis de los errores del producto de n números y calcule su error relativo. ¿Cuál es la relación (si es que la hay) entre el error relativo del producto y los errores de redondeo de los valores x_i? ¿Por qué?*

Solución. Para calcular el producto

$$\prod_{i=1}^{n} x_i,$$

operaremos paso a paso

$$p_2 = x_1\,x_2, \qquad p_3 = p_1\,x_3, \qquad \ldots, \qquad p_n = p_{n-1}\,x_n,$$

que si operamos utilizando el modelo estándar de la aritmética flotante nos da

$$fl(p_2) = fl(fl(x_1)\,fl(x_2)) = x_1\,x_2\,(1+\delta_1)\,(1+\delta_2)\,(1+\delta_{p2}), \qquad |\delta_1|, |\delta_2|, |\delta_{p2}| < u,$$

donde u es la unidad de redondeo, δ_i los errores relativos debidos a la representación flotante del número x_i, y δ_{pi} los debidos al $(i-1)$-ésimo producto. Acotando observamos que

$$|fl(p_2)| \leq |x_1\,x_2|\,(1+u)^3.$$

Procediendo de esta manera obtenemos también

$$|fl(p_3)| \leq |x_1\,x_2\,x_3|\,(1+u)^5,$$

y así sucesivamente,

$$|fl(p_n)| \leq |x_1\,x_2\,x_3\,\cdots\,x_n|\,(1+u)^{2\,n-1},$$

con lo que el error relativo en función de los errores de los datos toma la siguiente expresión, cuya cota también presentamos.

$$e_r(p_n) = \frac{fl(p_n) - p_n}{p_n} = \prod_{i=1}^{n}(1+\delta_i)\,\prod_{j=2}^{n}(1+\delta_{pi}) \leq (1+u)^{2\,n-1} - 1 \leq 1 + (2\,n-1)\,\tilde{u},$$

donde $\tilde{u}$ es la unidad de redondeo ajustada.

Ejercicios 2.19 *Calcule $f(x) = 1 - \cos x$, con aritmética flotante de seis dígitos para el número $x = 0.000010$. Explique su resultado. ¿Puede obtener un valor más exacto? ¿Cómo? ¿Por qué?*

Solución. Para $x = 0.000010 = 10^{-5}$, podemos aproximar el coseno por su desarrollo de Taylor alrededor de 0 obteniendo

$$\cos x \approx 1 - \frac{x^2}{2} = 1 - 0.5 \times 10^{-10} = 1.00000,$$

donde hemos redondeado el resultado a 6 dígitos significativos. De esta forma, obtenemos para el valor de la función, con la misma precisión,

$$f(x) = 1 - \cos x = 0.000000,$$

cuando la respuesta exacta (con 10 dígitos de precisión) es

$$f(x) = 1 - \cos x = \frac{x^2}{2} + O\big(x^4\big) = 0.5 \times 10^{-10}.$$

La respuesta que obtuvimos antes tenía un error relativo infinito.

Ejercicios 2.20 *Estime el número de condición de $f(x) = \sqrt{x+1} - \sqrt{x}$ para $x = 10^4$. Calcule $f(12345)$ con aritmética de seis dígitos. Explique sus resultados. ¿Puede obtener un valor más exacto? ¿Cómo? ¿Por qué?*

Solución. Para estimar el número de condición de una función utilizamos la expresión que lo define

$$\kappa\{f(x)\} = \left| \frac{\dfrac{f(x+\Delta x) - f(x)}{f(x)}}{\dfrac{x + \Delta x - x}{x}} \right| = \left| \frac{f(x+\Delta x) - f(x)}{f(x)} \, \frac{x}{\Delta x} \right|$$

eligiendo un valor para Δx conveniente. Por ejemplo, para $\Delta_x = 1$, obtenemos como estimación

$$\kappa\{f(x)\} == \left| \frac{\sqrt{10^4+2} - \sqrt{10^4+1} - \sqrt{10^4+1} + \sqrt{10^4}}{\sqrt{10^4+1} - \sqrt{10^4}} \, \frac{10^4}{1} \right| = \frac{|100.01 - 2 \times 100.005 + 100|}{|100.005 - 100|} \, 10^4 \approx 0.5,$$

que indica que esta función no está mal condicionada. La diferencia cancelativa que contiene no es catastrófica. El alumno puede comprobar, dado que $f(x) \in \mathbf{C}^2$, que el valor exacto del número de condición es 0.5, luego nuestra estimación ha sido bastante buena.

Operando con 6 decimales,

$$f(12345) = \sqrt{12346} - \sqrt{12345} = 111.113 - 111.108 = 0.005.$$

Para obtener un valor aún más exacto operando con 6 dígitos significativos, podemos realizar la operación exacta

$$\frac{\sqrt{x+1}-\sqrt{x}}{\sqrt{x+1}+\sqrt{x}}\,(\sqrt{x+1}+\sqrt{x}) = \frac{x+1-x}{\sqrt{x+1}+\sqrt{x}} = \frac{1}{\sqrt{x+1}+\sqrt{x}},$$

que conduce al resultado

$$f(12345) = \frac{1}{111.113+111.108} = \frac{1}{122.221} = 0.00450002,$$

que es un valor mucho más exacto, ya que redondeando el resultado en aritmética exacta a 6 dígitos significativos obtenemos

$$f(12345) = 0.00450003.$$

Ejercicios 2.21 *Estime el error en la evaluación de*

$$f(x) = \cos x \cdot \exp(10\,x^2),$$

para $x = 2$*, si el error absoluto en* x *es* 10^{-6}*.*

Solución. Dado que el error absoluto en x, sea $x+\epsilon$, con $|\epsilon| = 10^{-6}$, es muy pequeño respecto al tamaño de la función, podemos aplicar Taylor para aproximar,

$$f(x+\epsilon) = f(x) + \epsilon\, f'(x) + O\big(\epsilon^2\big)\,,$$

que nos da

$$f(x+10^{-6}) = f(x)+10^{-6}\,(20\,x\,\cos x-\sin x)\,e^{10\,x^2} = e^{10\,x^2}\,(\cos x+10^{-6}\,(20\,x\,\cos x-\sin x)+O\big(10^{-12}\big)),$$

con lo que el error absoluto cometido en la evaluación de $f(x)$ es aproximadamente

$$10^{-6}\,(20\,x\,\cos x - \sin x)\,e^{10\,x^2}\Big|_{x=2} \approx -4.1\times 10^{12},$$

que es muy grande debido a la exponencial del cuadrado de x que crece muy rápido.

Ejercicios 2.22 *Utilice una mantisa de cuatro cifras decimales para calcular las raíces de*

$$x^2 + 0.4002\times 10^0\,x + 0.8\times 10^{-4} = 0.$$

Explique sus resultados. ¿Puede mejorar estas raíces? ¿Cómo? ¿Por qué?

Solución. Utilizando la fórmula estándar para las raíces de una ecuación cuadrática, $a\,x^2 + b\,x + c = 0$, que se escribe

$$x_\pm = \frac{-b \pm \sqrt{b^2 - 4\,a\,c}}{2\,a},$$

y operando paso a paso, obtenemos

$$b^2 - 4\,a\,c = 0.16016 - 0.00032 = 0.1602 - 0.0003 = 0.1599,$$

que nos da como raíces

$$x_\pm = \frac{1}{2}\,(-0.4002 \pm \sqrt{0.1599}) = \frac{1}{2}\,(-0.4002 \pm 0.3999),$$

$$x_+ = -\frac{0.0003}{2} = -0.0002, \qquad x_- = -\frac{0.8001}{2} = -0.4000,$$

que es el resultado exacto redondeado a dicha precisión.

No es necesario mejorar el resultado obtenido, pues no es posible hacerlo.

Ejercicios 2.23 *Calcular los números de condición asociados a las siguientes operaciones:* $x/y, x - y, \sqrt{x}, e^x$. *Determinar los errores relativos.*

Solución. El número de condición de una función real de 2 variables reales, $f(x, y)$, se define fácilmente como[5]

$$\kappa\{f(x,y)\} = \frac{\left|\dfrac{f(x+\Delta x, y+\Delta y) - f(x,y)}{f(x,y)}\right|}{\dfrac{\|(\Delta x, \Delta y)\|}{\|(x,y)\|}},$$

donde para medir el tamaño del vector de variables independientes se deben utilizar normas de vectores, como por ejemplo, la norma 1 o la norma infinito,

$$\|(x,y)\|_1 = |x| + |y|, \qquad \|(x,y)\|_\infty = \max\{|x|, |y|\}.$$

De esta forma obtenemos para la división, usando norma infinito,

$$\begin{aligned}
\kappa\left\{\frac{x}{y}\right\} &= \left|\frac{\frac{x+\Delta x}{y+\Delta y} - \frac{x}{y}}{\frac{x}{y}}\right| \frac{\|(x,y)\|_\infty}{\|(\Delta x, \Delta y)\|_\infty} \\
&= \left|\frac{y\,\Delta x - x\,\Delta y}{(y+\Delta y)\,x}\right| \frac{\|(x,y)\|_\infty}{\|(\Delta x, \Delta y)\|_\infty} \\
&\leq \frac{\|(x,y)\|_\infty\,\|(\Delta x, \Delta y)\|_\infty}{|(y+\Delta y)\,x|} \frac{\|(x,y)\|_\infty}{\|(\Delta x, \Delta y)\|_\infty},
\end{aligned}$$

[5]En el próximo capítulo estudiaremos en detalle las normas de vectores y matrices.

donde suponiendo $|\Delta y| < \epsilon\,|y|$, con $\epsilon \ll 1$, podemos aproximar

$$\kappa \left\{ \frac{x}{y} \right\} \lesssim \|(x,y)\|_\infty .$$

Ejercicios 2.24 Examen 21/Marzo/1996. *Estudie la estabilidad de Hadamard de los siguientes problemas y determine cuando están bien condicionados:*

1. *La ecuación diferencial*

$$\frac{d^2y}{dx^2} - y = 0, \qquad y(0) = a, \quad \frac{dy}{dx}(0) = b.$$

2. *La ecuación algebraica no lineal* $f(x) = x^2 - 1 = 0$.

3. *La ecuación algebraica lineal*

$$x + y = 1,$$

$$0.99999\,x + y = 1.$$

Solución.

1. La solución general de la ecuación diferencial

$$\frac{d^2y}{dx^2} = y, \qquad y(0) = a, \quad \frac{dy}{dx}(0) = b.$$

se escribe como

$$y(x) = A\,e^x + B\,e^{-x},$$

y aplicando las condiciones iniciales,

$$y(0) = A + B = a,$$

$$y'(0) = A - B = b,$$

se obtiene finalmente

$$y(x) = \frac{a+b}{2}\,e^x + \frac{a-b}{2}\,e^{-x}.$$

Por lo tanto, la solución existe y es única, pero el problema no está bien condicionado físicamente a no ser que $a = -b$, ya que sino

$$\lim_{x\to\infty} y(x) = \text{sign}\,(a+b)\,\infty.$$

Si suponemos $a = -b$ para que el problema esté bien condicionado física y matemáticamente, la solución será

$$y = a\,e^{-x} = -b\,e^{-x}.$$

Estudiemos la estabilidad de Hadamard en ese caso, es decir, el problema perturbado

$$y(x) = A\,e^{x} + B\,e^{-x}, \qquad y(0) = a + \epsilon_1, \quad \frac{dy}{dx}(0) = b + \epsilon_2 = -a + \epsilon_2.$$

La solución de este problema perturbado es

$$y(x) = \frac{\epsilon_1 + \epsilon_2}{2}\,e^{x} + \left(a + \frac{\epsilon_1 - \epsilon_2}{2}\right)\,e^{-x}.$$

Por tanto, este problema no está bien condicionado en el sentido de Hadamard ya que a no ser que $\epsilon_1 + \epsilon_2 = 0$,

$$\lim_{x\to\infty} y(x) = \text{sign}\,(\epsilon_1 + \epsilon_2)\,\infty.$$

2. La ecuación algebraica no lineal $f(x) = x^2 - 1 = 0$, tiene dos raíces $x = \pm 1$, por tanto la solución existe y es única. Si se considera el polinomio más general $a\,x^2 + b\,x + c = 0$, está claro que la ecuación considerada corresponde a $a = 1, b = 0, c = -1$. Para estudiar su estabilidad en el sentido de Hadamard, supongamos pequeños errores en estos coeficientes,

$$a = 1 + \epsilon_1, \qquad b = 0 + \epsilon_2, \qquad c = -1 + \epsilon_3,$$

por lo que el polinomio perturbado será

$$(1 + \epsilon_1)\,x^2 + \epsilon_2\,x - 1 + \epsilon_3 = 0.$$

Las raíces del polinomio perturbado son

$$x = -\frac{\epsilon_2}{2\,(1 + \epsilon_1)} \pm \sqrt{\frac{\epsilon_2^2}{4\,(1 + \epsilon_1)^2} + 1 - \epsilon_3},$$

por lo que se ve que si ϵ_1, ϵ_2 y ϵ_3 son pequeños, las raíces del polinomio perturbado son próximas a la solución del problema no perturbado, es decir, ± 1. Por tanto, este problema está bien condicionado en el sentido de Hadamard.

3. La ecuación algebraica lineal

$$x + y = 1,$$

$$0.99999\,x + y = 1.$$

tiene como solución única $x = 0$ e $y = 1$. El problema perturbado asociado es

$$(1 + \epsilon_1)\, x + (1 + \epsilon_2)\, y = 1 + \epsilon_3,$$

$$(0.99999 + \epsilon_4)\, x + (1 + \epsilon_5)\, y = 1 + \epsilon_6.$$

El determinante de su matriz de coeficientes es

$$\det = (1 + \epsilon_1)\,(1 + \epsilon_5) - (1 + \epsilon_2)\,(0.99999 + \epsilon_4),$$

que es nulo para $\epsilon_1 = \epsilon_5 = \epsilon_2 = 0$ y $\epsilon_4 = 0.00001$, es decir, para esos valores no existe solución (el problema es incompatible si $\epsilon_3 \neq \epsilon_6$) o existen infinitas soluciones (si $\epsilon_3 = \epsilon_6$). Si el determinante no es nulo, la solución es

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{\det} \begin{pmatrix} 1 + \epsilon_5 & -(1 + \epsilon_2) \\ -(0.99999 + \epsilon_4) & 1 + \epsilon_1 \end{pmatrix} \cdot \begin{pmatrix} 1 + \epsilon_3 \\ 1 + \epsilon_6 \end{pmatrix}.$$

El condicionamiento de un sistema lineal se mide mediante su número de condicionamiento, definido como

$$\kappa(A) = \|A\|\,\|A^{-1}\|.$$

Utilizando normas uno e infinito para determinarlo, tenemos que

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}| = \max 1.99999, 2 = 2,$$

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}| = \max 2, 1.99999 = 2,$$

$$A^{-1} = \frac{1}{1 - 0.99999} \begin{pmatrix} 1 & -1 \\ -0.99999 & 1 \end{pmatrix} = \begin{pmatrix} 10^5 & -10^5 \\ -99999 & 10^5 \end{pmatrix},$$

$$\|A^{-1}\|_1 = \max 199999, 2 \cdot 10^5 = 2 \cdot 10^5,$$

$$\|A^{-1}\|_\infty = 200000.$$

Por lo que, finalmente

$$\kappa(A) = \|A\|_1\,\|A^{-1}\|_1 = \|A\|_\infty\,\|A^{-1}\|_\infty = 4 \cdot 10^5.$$

Por lo tanto, este problema está extremadamente mal condicionado. Para ordenadores con menos de seis cifras significativas, la segunda ecuación es la misma que la primera y, por tanto, existen infinitas soluciones.

BIBLIOGRAFÍA

[1] James H. Wilkinson, "*Rounding Errors in Algebraic Processes*," Prentice Hall, Englewoods Cliffs, New Jersey, USA (1963).

[2] Nicholas J. Higham, "*Accuracy and Stability of Numerical Algorithms*," SIAM, Philadelphia (1996).