Segunda relación de problemas — Técnicas Numéricas
Profesor Francisco R. Villatoro — 23 de Octubre de 1998

SOLUCIONES

Soluciones de los ejercicios de la segunda relación de problemas.

1. Suponga que $f(x) \in \mathrm{C}^\infty$. Use el desarrollo en serie de Taylor y evalúe $f(x+h)$ y $f(x-h)$. A partir de estos desarrollos, calcule $f'(x)$ y $f''(x)$ y establezca una cota de los errores que comete. Nota: $f'(x)$ y $f''(x)$ tienen que expresarse sólo en función de $f(x)$, $f(x+h)$ y $f(x-h)$.

   Solución. Desarrollando en serie de Taylor

$$f(x+h) = f(x) + h\,f'(x) + \frac{h^2}{2}\,f''(x) + \frac{h^3}{3!}\,f'''(x) + O\left(h^4\right)$$
$$f(x-h) = f(x) - h\,f'(x) + \frac{h^2}{2}\,f''(x) - \frac{h^3}{3!}\,f'''(x) + O\left(h^4\right).$$

   Que nos permiten obtener las siguientes expresiones para la primera derivada

   Hacia adelante
$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h^2}{2}\,f''(\xi), \qquad \text{con } \xi \in (x, x+h),$$
   Hacia atrás
$$f'(x) = \frac{f(x) - f(x-h)}{h} + \frac{h^2}{2}\,f''(\xi), \qquad \text{con } \xi \in (x-h, x),$$
   Centrada
$$f'(x) = \frac{f(x+h) - f(x-h)}{2\,h} - \frac{h^3}{3}\,f'''(\xi), \qquad \text{con } \xi \in (x-h, x+h).$$

   Para la segunda derivada podemos obtener la siguiente expresión centrada

$$f''(x) = \frac{f(x+h) - 2\,f(x) + f(x-h)}{h^2} - \frac{h^2}{12}\,f''''(\xi),$$
$$\text{con } \xi \in (x-h, x+h).$$

2. Dados los resultados del problema 1, evalúe $f'(1)$ y $f''(1)$ para las siguientes funciones

   a) $f_1(x) = \sin(x)$,

   b) $f_2(x) = 10000\,\sin(x)$,

   c) $f_3(x) = \tan(x)$.

   ¿Cómo afecta el valor de h a sus resultados? Nota: utilice calculadora y presente los resultados con cuatro dígitos de precisión.

   Solución. Las tres funciones son analíticas, excepto $f(x) = \tan(x)$ que no lo es para $x = (2\,\pi + 1)\,\pi/2$, con $n \in \mathbb{N}$. Nota: hemos supuesto que los naturales incluyen el cero.

   Calculando previamente las primeras y segundas derivadas exactas, obtenemos

$$
\begin{aligned}
&f_1(x) = \sin x, \qquad f_1'(x) = \cos x, \qquad f_1''(x) = -f_1(x);\\
&f_2(x) = 10^4\,\sin x, \qquad f_2'(x) = 10^4\,\cos x, \qquad f_2''(x) = -f_2(x);\\
&f_3(x) = \tan x, \qquad f_3'(x) = \frac{1}{\cos x} = 1 + f_3(x)^2,\\
&f_3''(x) = 2\,f_3(x)\,f_3'(x) = 2\,\tan x\,(1 + \tan^2 x).
\end{aligned}
$$

   Los resultados exactas (operando en radianes) hasta cuatro dígitos significativos son

$$
\begin{aligned}
&f_1'(1) = 0{,}5403, \qquad f_2'(1) = 5403, \qquad f_3'(1) = 3{,}426,\\
&f_1'(1) = -0{,}8415, \qquad f_2'(1) = -8415, \qquad f_3'(1) = 10{,}67\,.
\end{aligned}
$$

   Nota: todas las operaciones aritméticas han sido realizadas con una calculadora HP-67 que opera con 10 dígitos significativos.

   Utilizando las expresiones derivadas en el problema 1, con un h extremadamente pequeño, por ejemplo $h = 10^{-8}$, obtenemos para la fórmula en diferencias hacia adelante

$$
\begin{aligned}
f_1'(1) &= \frac{\sin(1 + 10^{-8}) - \sin(1)}{10^{-8}} = 0{,}54,\\
f_2'(1) &= 10^4\,f_1'(1) = 5400,\\
f_1'(1) &= \frac{\tan(1 + 10^{-8}) - \tan(1)}{10^{-8}} = 3{,}5;
\end{aligned}
$$

para la fórmula centrada (omitimos los resultados de la fórmula hacia atrás para la primera derivada),

$$
\begin{aligned}
f_1'(1) &= \frac{\sin(1+10^{-8}) - \sin(1-10^{-8})}{2\,10^{-8}} = 0{,}545,\\
f_2'(1) &= 10^4\, f_1'(1) = 5450,\\
f_1'(1) &= \frac{\tan(1+10^{-8}) - \tan(1-10^{-8})}{2\,10^{-8}} = 3{,}45;
\end{aligned}
$$

y, para la fórmula centrada para la segunda derivada

$$
\begin{aligned}
f_1''(1) &= \frac{\sin(1+10^{-8}) - \sin(1) + \sin(1-10^{-8})}{10^{-16}} = 0.,\\
f_2''(1) &= 10^4\, f_1'(1) = 0,\\
f_3''(1) &= \frac{\tan(1+10^{-8}) - \tan(1) + \tan(1-10^{-8})}{10^{-16}} = 10^{-7}.
\end{aligned}
$$

Para obtener mejores resultados podemos elegir un h pequeño, pero no tan pequeño, por ejemplo $h = 10^{-4}$, con lo que obtenemos para la fórmula en diferencias centradas (omitimos los resultados de las fórmulas hacia adelante y hacia atrás para la primera derivada),

$$
\begin{aligned}
f_1'(1) &= \frac{\sin(1+10^{-4}) - \sin(1-10^{-4})}{2\,10^{-4}} = 0{,}5403,\\
f_2'(1) &= 10^4\, f_1'(1) = 5403,\\
f_1'(1) &= \frac{\tan(1+10^{-4}) - \tan(1-10^{-4})}{2\,10^{-4}} = 3{,}426;
\end{aligned}
$$

y, para la fórmula centrada para la segunda derivada

$$
\begin{aligned}
f_1''(1) &= \frac{\sin(1+10^{-4}) - \sin(1) + \sin(1-10^{-4})}{10^{-8}} = -0{,}9000,\\
f_2''(1) &= 10^4\, f_1'(1) = -9000,\\
f_3''(1) &= \frac{\tan(1+10^{-4}) - \tan(1) + \tan(1-10^{-4})}{10^{-8}} = 10{,}7.
\end{aligned}
$$

De los resultados obtenidos queda claro que las primeras derivadas se pueden evaluar muy precisamente con pequeños h, mientras que este no es el caso para las segundas derivadas debido a errores numéricos de redondeo (por diferencias cancelativas y divisiones por números pequeños).

3. Escriba un programa en Matlab para el cálculo del épsilon de su máquina. ¿Coincide con `eps`? Nota: Se denomina épsilon de la máquina ε al al error absoluto que representa el último dígito representable en la mantisa, es decir, se define como

$$\varepsilon = min\{\epsilon : fl(1+\epsilon) \neq 1\}.$$

Solución. Recordemos que el épsilon del a máquina es el número flotante más pequeño tal que $1+\varepsilon = 1$. En Matlab la variable `eps` contiene dicho valor. Calculémoslo

```
epsilon = 1;
while (1+epsilon > 1),
  epsilon=epsilon/2;
end;
epsilon = epsilon*2,
```

En Matlab en un ordenador PC-compatible `eps = 2.2204e-016`, que coincide con 2^{-52}. Matlab también nos permite calcular el número flotante más grande `realmax = 1.7977e+308` $= 2^{1024}$ y el número flotante positivo más pequeño `realmin = 2.2251e-308` $= 2^{-1022}$.

4. Estime mediante propagación de errores hacia atrás el error relativo cometido en la operación de suma de números flotantes. Aproximelo utilizando el épsilon de la máquina.

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

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

que contienen un error relativo de redondeo y luego realizar la operación de suma, que también tiene un error asociado,

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

Introduciendo errores en los datos iniciales, tenemos

$$fl(x+y) = x\,(1+\delta_1) + y\,(1+\delta_2) = x + y + \delta_1\,x + \delta_2\,y,$$

por lo que el error relativo de la suma es

$$\frac{fl(x+y)-(x+y)}{x+y}=\frac{\delta_1\,x+\delta_2\,y}{x+y}.$$

Este error lo podemos escribir como

$$fl(x+y)=(x+y)\,(1+\delta),\qquad \delta=\frac{fl(x+y)-(x+y)}{x+y}.$$

Si como es usual $\delta_x=\delta_y=\delta_s=\varepsilon$, son iguales al épsilon de la máquina, tenemos que

$$fl(x+y)=(x+y)\,(1+\varepsilon)^2,\qquad \delta=(1+\varepsilon)^2-1\approx 2\,\varepsilon,$$

es decir, el error relativo en la suma ese igual al doble del epsilon de la máquina.

5. Estime mediante propagación de errores hacia atrás el error relativo cometido en la operación de multiplicación de números flotantes en función de los errores absolutos de los datos iniciales.

   Solución. De manera del todo similar a lo presentado en este tema podemos determinar el error relativo para la operación de multiplicación de la siguiente forma

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

   Suponiendo errores absolutos en los datos iniciales

$$fl(x\,y)=(x+\epsilon_x)\,(y+\epsilon_y)=x\,y+x\,\epsilon_x+y\,\epsilon_y+\epsilon_x\,\epsilon_y$$

   con lo que el error relativo es

$$\frac{fl(x\,y)-x\,y}{x\,y}=\frac{\epsilon_x}{x}+\frac{\epsilon_y}{y}+\frac{\epsilon_x}{x}\,\frac{\epsilon_y}{y},$$

   es decir, la suma de los errores relativos de los datos más el producto de éstos. Nótese que el error relativo en la variable x es

$$\frac{fl(x)-x}{x}=\frac{x+\epsilon_x-x}{x}=\frac{\epsilon_x}{x}.$$

6. La operación de suma de números flotantes no cumple con las propiedades asociativa y conmutativa, es decir, el orden de los factores altera el resultado y, por tanto, el error de éste. Demostrar que si se suman los números empezando por el menor y en orden creciente se minimiza la pérdida de dígitos significativos en el resultado.

   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

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

   Estudiemos como se propagan los errores relativos en estas sumas parciales. Operando y despreciando los productos $\epsilon_i\,\epsilon_j$ 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

$$\begin{aligned} fl(s_4) &= (fl(s_3) + x_4)\,(1 + \epsilon_4) \\ &= s_4 + (x_1 + x_2)\,(\epsilon_2 + \epsilon_3 + \epsilon_4) + x_3\,(\epsilon_3 + \epsilon_4) + x_4\,\epsilon_4. \end{aligned}$$

   La fórmula general que se obtiene es

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

   Haciendo $\epsilon_i = \epsilon$, tenemos finalmente

$$fl(s_n) = s_n + (x_1 + x_2)\,(n-1)\,\epsilon + x_3\,(n-2)\,\epsilon + \cdots + x_n\,\epsilon,$$

   por lo que el error en los primeros sumandos es mayor que en los últimos. Por ello, si sumamos primeros los números de módulo menor haremos que el error de redondeo de la suma se minimice.

7. Escriba un algoritmo (y una función en Matlab) para el cálculo de las dos raíces de una ecuación de segundo grado que evite las diferencias cancelativas cuando $b^2 \gg 4\,a\,c$.

   Solución. Dado que no conocemos a priori el signo de b, en lugar de usar una selección entre los dos posibles signos, podemos usar la expresión

$$x_1 = -\frac{b + \text{sign}\,(b)\,\sqrt{b^2 - 4\,a\,c}}{2\,a},$$

   que es independiente del signo. Aprovechando que $a\,x^2 + b\,x + c = a\,(x - x_1)\,(x - x_2)$ indica que $a\,x_1\,x_2 = c$, podemos calcular la otra raíz como

$$x_2 = \frac{c}{a\,x_1}.$$

   Esta expresión, como es fácil de verificar, coincide con la dada previamente en el capítulo.

   Podemos escribir una función en Matlab que calcule las raíces:

```
function [x1, x2] = raices2(a, b, c)
%% [x1,x2] = raices2(a,b,c)
%% Calcula las dos raices de    a x^2 + b x + c
%% (evita diferencias cancelativas)
%%   NOTA: funciona para vectores de ecuaciones
%%
 x1 = - (b + sign(b).*sqrt(b.^2 - 4*a.*c)) ./ (2*a);
 x2 = c ./ (a.*x1);
```

8. 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

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

   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 y que 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 calcular 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

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

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 epsilon de la máquina habrá que calcular s_n sucesivamente hasta que $s_n = s_{n-1}$.

Sin embargo, es más eficiente computacionalmente aproximar la exponencial de las 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 secuencia

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

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

9. 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.

   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.

10. Realizar un análisis de propagación de errores hacia atrás para la suma de n pares de productos dada por

$$s = a_1\, b_1 + a_2\, b_2 + \cdots + a_n\, b_n.$$

   Suponga que los datos son números flotantes y por tanto sin error de redondeo. Además suponga, para simplificar, que el error relativo de las operaciones de suma y producto es el mismo.

   Solución. Calcularemos s de la siguiente forma iterativa

$$\begin{aligned} fl(s_1) &= a_1\, b_1\, (1+\delta), \\ fl(s_2) &= (fl(s_1) + a_2\, b_2\, (1+\delta))\ (1+\delta) \\ &= (a_1\, b_1 + a_2\, b_2)\, (1+\delta)^2, \\ fl(s_3) &= (fl(s_2) + a_3\, b_3\, (1+\delta))\ (1+\delta) \\ &= (a_1\, b_1 + a_2\, b_2)\, (1+\delta)^3 + a_3\, b_3\, (1+\delta)^2, \end{aligned}$$

   que conduce a la expresión general

$$\hat{s} = fl(s) = (a_1\, b_1 + a_2\, b_2)\, (1+\delta)^n + a_3\, b_3\, (1+\delta)^{n-1} + \cdots + a_n\, b_n\, (1+\delta),$$

   que eliminando infinitésimos de orden superior se puede escribir como

$$\hat{s} = fl(s) = (a_1\, b_1 + a_2\, b_2)\, (1+n\,\delta) + a_3\, b_3\, (1+(n-1)\,\delta) + \cdots + a_n\, b_n\, (1+\delta).$$

   De esta forma en un análisis de error hacia atrás buscaremos datos iniciales tales que el cálculo de la expresión de s para ellos coincida con

$\hat{s}$. Por simetría, estos valores serán

$$\hat{a}_1 = a_1\,(1 + \frac{n\,\delta}{2}), \qquad \hat{a}_2 = a_2\,(1 + \frac{n\,\delta}{2}),$$
$$\hat{a}_3 = a_3\,(1 + \frac{(n-1)\,\delta}{2}), \qquad \cdots \qquad \hat{a}_n = a_n\,(1 + \frac{\delta}{2}),$$

y de forma del todo similar para las $\hat{b}_i$. Como vemos la contribución del error de los primeros sumandos es mayor que la de los últimos.

Dada la importancia de expresiones del tipo de s, muchos coprocesadores numéricos tienen la implementan directamente en hardware. Para que el resultado de s tenga un error menor que el epsilon de la máquina es necesario que $n\,\delta/2$ sea de dicho orden, por lo que estos procesadores tienen que retener internamente un número suficiente de bits adicionales para la mantisa que lo garanticen. Seleccionar el valor de n, y con él el número de bits adicionales, es un compromiso importante en el diseño de dichos coprocesadores.

11. Cómo se debe evaluar la función

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

para $\alpha < x$, 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,

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

Por otro lado, podemos aplicar

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

Aunque las dos expresiones que hemos obtenido son diferentes, la segunda expresión tiende a la primera cuando $x \gg \alpha$. Aunque la primera expresión pueda parecer 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 será despreciable.

12. Con una mantisa de cuatro dígitos decimales, 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 como de menor a mayor (en valor absoluto). Justifique los resultados que encuentre.

Solución. La suma exacta s_E es

$$s_E = 1025 - 912{,}3 - 96{,}63 - 9{,}315 = 6{,}755\,.$$

Para sumar en orden de mayor a menor (que es el que aparece originalmente en dicha suma, primero igualaremos los exponentes de los números al mayor de ellos, es decir,

$$s = 0{,}1025 \cdot 10^4 - 0{,}0912\underline{3} \cdot 10^4 - 0{,}00966\underline{3} \cdot 10^4 - 0{,}0009\underline{315} \cdot 10^4.$$

Los dígitos subrayados no entran dentro de la mantisa, por lo que los redondearemos,

$$s = 0{,}1025 \cdot 10^4 - 0{,}0912 \cdot 10^4 - 0{,}0097 \cdot 10^4 - 0{,}0009 \cdot 10^4.$$

Si sumamos estos números obtendremos $s = 0{,}0007 \cdot 10^4$, sin embargo, esta respuesta es incorrecta ya que un ordenador realiza cada operación de resta de forma separada, igualando exponentes y normalizando el resultado en cada paso. La respuesta correcta reza como sigue

$$\begin{aligned}
&s_1 = 0{,}1025 \cdot 10^4,\\
&s_2 = s_1 - 0{,}0912 \cdot 10^4 = 0{,}0113 \cdot 10^4 = 0{,}1130 \cdot 10^3,\\
&s_3 = s_2 - 0{,}0966\underline{3} \cdot 10^3 \approx s_2 - 0{,}0966 \cdot 10^3 = 0{,}0164 \cdot 10^3 = 0{,}1640 \cdot 10^2,\\
&s_4 = s_3 - 0{,}0931\underline{5} \cdot 10^2 \approx s_3 - 0{,}0932 \cdot 10^2 = 0{,}0708 \cdot 10^2 = 0{,}7080 \cdot 10^1 = 7{,}080.
\end{aligned}$$

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

$$\frac{s_4 - s_E}{s_E} = \frac{7{,}080 - 6{,}755}{6{,}755} = 0{,}048 \approx 5\,\%,$$

que es un error bastante alto.

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

$$\begin{aligned}
s'_1 &= -0{,}9315 \cdot 10^1, \\
s'_2 &= s_1 - 0{,}9663 \cdot 10^2 = -0{,}0931\underline{5} \cdot 10^2 - 0{,}9663 \cdot 10^2 \\
&\approx -0{,}0932 \cdot 10^2 - 0{,}9663 \cdot 10^2 = -1{,}0595 \cdot 10^2 = -0{,}1060 \cdot 10^3, \\
s'_3 &= s'_2 - 0{,}9123 \cdot 10^3 = -0{,}1060 \cdot 10^3 - 0{,}9123 \cdot 10^3 \\
&= -1{,}0183 \cdot 10^3 = -0{,}1018 \cdot 10^4, \\
s'_4 &= s'_3 + 0{,}1025 \cdot 10^4 = -0{,}1018 \cdot 10^4 + 0{,}1025 \cdot 10^4 \\
&= 0{,}0007 \cdot 10^4 = 0{,}7000 \cdot 10^1 = 7.
\end{aligned}$$

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).

13. Evalúe (con 5 dígitos tras la coma decimal) la función e^x cuando $x = 5$ y $x = -5$ utilizando

    a) Desarrollos en serie de Taylor.

    b) Si la convergencia del desarrollo en serie de Taylor es muy lenta, proponga un método (o métodos) más precisos para dicha evaluación.

Solución. El resultado exacto (con 5 dígitos tras la coma decimal) que nos nuestra calculadora es

$$e^5 = 148{,}41316, \qquad e^{-5} = 0{,}0067380.$$

El desarrollo en serie de Taylor de la exponencial es

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

Esta serie converge para $x = 5$ como se prueba fácilmente mediante el criterio de Cauchy, ya que

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

para $n \geq 4$, donde por t_n hemos denotado el término n-ésimo de la serie.

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

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

Entonces obtenemos, operando con cinco dígitos decimales,

$$\begin{aligned}
&s(0) = 1{,}00000, \qquad\qquad s(1) = 1 + 5 = 6{,}00000,\\
&s(2) = 6 + \frac{5^2}{2} = 18{,}50000,\\
&s(3) = s(2) + \frac{5^3}{3!} = 39{,}33333,\\
&s(4) = s(3) + \frac{5^4}{4!} = 65{,}37500,\\
&s(5) = s(4) + \frac{5^5}{5!} = 91{,}41667,\\
&s(6) = s(5) + \frac{5^6}{6!} = 113{,}11806,\\
&s(7) = s(6) + \frac{5^7}{7!} = 128{,}61905,\\
&s(8) = s(7) + \frac{5^8}{8!} = 138{,}30717,\\
&s(9) = s(8) + \frac{5^9}{9!} = 143{,}68946,\\
&s(10) = s(9) + \frac{5^{10}}{10!} = 146{,}38060,\\
&s(11) = s(10) + \frac{5^{11}}{11!} = 147{,}60385,\\
&s(12) = s(11) + \frac{5^{12}}{12!} = 148{,}11354,
\end{aligned}$$

$$s(13) = s(12) + \frac{5^{13}}{13!} = 148{,}30957,$$
$$\dots$$

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 (1) 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

$$\frac{e^{-5} - 1/s(13)}{e^5} = 0{,}0007,$$

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

14. 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 $x = 1$

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

15. 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

$$\tan x = x + O\left(x^3\right).$$

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

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

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 $O(10^{-24})$.

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

$$\left|\frac{2 - 1{,}6667}{1{,}6667}\right| = 0{,}2 \approx 20\,\%.$$

16. ¿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

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

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

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

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.

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

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

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

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

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

$$\int_0^1 e^x\,dx = e^x|_0^1 = e - 1 = f(\xi) = e^\xi$$
$$\xi = \ln(e-1) = 0{,}541,$$

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

$$\int_0^1 f(x)\,g(x)\,dx = f(\xi)\int_0^1 g(x)\,dx,$$
$$\int_0^1 x\,e^x\,dx = x\,e^x|_0^1 - \int_0^1 e^x\,dx = e - e^x|_0^1 = 1$$
$$f(\xi)\int_0^1 x\,dx = f(\xi)\,\frac{1}{2} = e^\xi\,\frac{1}{2} = 1,$$
$$\xi = \ln 2 = 0{,}693.$$

La tercera relación presentada en el enunciado del problema sólo es verdad 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 $[-1, 1]$,

$$\int_{-1}^{1} f(x)\, g(x)\, dx = \int_{-1}^{1} x\, e^x\, dx = x\, e^x|_{-1}^{1} - \int_{-1}^{1} e^x\, dx =$$
$$e + e^{-1} - e^x|_0^1 = e + e^{-1} - e + e^{-1} = \frac{2}{e},$$
$$f(\xi) \int_{-1}^{1} g(x)\, dx = \quad f(\xi) \int_{-1}^{1} x\, dx = f(\xi)\, 0 = 0,$$
$$\text{y } \frac{2}{e} \neq 0.$$

18. 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

$$0{,}780\, x + 0{,}563\, y = 0{,}217,$$
$$0{,}457\, x + 0{,}330\, y = 0{,}127,$$

primeramente operaremos 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 utilizaremos 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. Por un lado los podemos justificar debido a que el determinante (Det) es muy próximo a cero. Por otro lado, podemos estudiar sus autovalores. Escribamos 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.

19. 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

$$\begin{aligned} y(x) &= A\, e^x + B\, e^{-x} = C\, \sinh x + D\, \cosh x, \\ y(0) &= D = a, \\ y'(x) &= C\, \cosh x + D\, \sinh x, \\ y(0) &= C = b, \end{aligned}$$

y por tanto

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

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

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

$$\begin{aligned} \frac{dy}{dx} &= z, \qquad & z(0) &= b, \\ \frac{dz}{dx} &= y, \qquad & y(0) &= a. \end{aligned}$$

20. Ejemplo. Supongamos que queremos calcular las integrales

$$E_n = \int_0^1 x^n \, e^{x-1} \, dx, \qquad n = 1, 2, \ldots .$$

Usando integración por partes

$$\int_0^1 x^n \, e^{x-1} \, dx = x^n \, e^{x-1} \Big|_0^1 - \int_0^1 n \, x^{n-1} \, e^{x-1} \, dx$$

podemos obtener el siguiente algoritmo numérico iterativo

$$E_n = 1 - n \, E_{n-1}, \qquad n = 2, \ldots ,$$

donde $E_1 = 1/e$. Si usamos una aritmética flotante con 6 dígitos decimales, obtenemos los siguientes resultados tras 9 iteraciones

$$\begin{aligned} E_1 &\approx 0{,}367879, \qquad & E_2 &\approx 0{,}264242, \\ E_3 &\approx 0{,}207274, \qquad & E_4 &\approx 0{,}170904, \\ E_5 &\approx 0{,}145480, \qquad & E_6 &\approx 0{,}127120, \\ E_7 &\approx 0{,}110160, \qquad & E_8 &\approx 0{,}118720, \\ E_9 &\approx -0{,}0684800. \end{aligned}$$

Con lo que E_9 es negativo, lo que es un resultado sorprendente, ya que el integrando $x^n \, e^{x-1}$ es positivo en todo el intervalo $(0, 1)$ y por tanto debe ser $E_n > 0$

La causa de este error tan grande ha sido la propagación del error cometido en E_1 al aproximar $1/e$ con 6 dígitos decimales. Este error se ha propagado (ha sido multiplicado por $(-2)(-3)\ldots(-9) = 9!$, por lo que el error en E_1 de aproximadamente $4{,}4 \times 10^{-7}$ se ha convertido en un error de $9! \, 4{,}4 \times 10^{-7} \approx 0{,}16$. El valor correcto de E_9 (con dos dígitos decimales) es $0{,}16 - 0{,}06848 = 0{,}092$.

Para evitar la inestabilidad numérica de nuestro algoritmo debemos buscar otra algoritmo que sea estable. Por ejemplo, la iteración hacia atrás

$$E_{n-1} = \frac{1 - E_n}{n}, \qquad n = \ldots, 3, 2,$$

tiene la ventaja que el error de redondeo original decrece conforme vamos iterando. Para obtener un valor inicial para esta iteración podemos usar,

$$E_n = \int_0^1 x^n \, e^{x-1} \, dx \leq \int_0^1 x^n \, dx = \frac{1}{n+1}.$$

Con lo que observamos que E_n tiende a cero conforme $n \Rightarrow \infty$. Si consideramos $E_{20} \approx 0$ y calculamos E_9 con el algoritmo estable

$$\begin{aligned}
E_{20} &\approx 0{,}0000000, & E_{19} &\approx 0{,}0500000, \\
E_{18} &\approx 0{,}0500000, & E_{17} &\approx 0{,}0527778, \\
E_{16} &\approx 0{,}0557190, & E_{15} &\approx 0{,}0590176, \\
E_{14} &\approx 0{,}0627322, & E_{13} &\approx 0{,}0669477, \\
E_{12} &\approx 0{,}0717733, & E_{11} &\approx 0{,}0773523, \\
E_{10} &\approx 0{,}0838771, & E_9 &\approx 0{,}0916123.
\end{aligned}$$

A partir de E_{15} el error inicial en E_{20} se ha disipado gracias a la estabilidad del algoritmo y, por tanto, los valores obtenidos para $E_{15}, \ldots E_9$ son exactos hasta los 6 dígitos de precisión con los que han sido calculados, salvo errores de redondeo en el último dígito.