PRÁCTICA 1: ERRORES

Grupo 1
Júlia Albero Pes, NIU:1566550
Víctor Ballester Ribó, NIU:1570866

Métodos numéricos
Grado en Matemáticas
Universitat Autònoma de Barcelona
15 de Marzo de 2021

Problema 1

Los resultados que hemos obtenido al calcular la función

$$f(x) = \begin{cases} \dfrac{1-\cos(x)}{x^2} & \text{si} \quad x \neq 0 \\ \dfrac{1}{2} & \text{si} \quad x = 0 \end{cases} \tag{1}$$

en el punto $x_0 = 1,2 \times 10^{-5}$ en precisión simple y en precisión doble se muestran en la tabla siguiente:

	Valor en precisión simple	Valor en precisión doble
$f(x_0)$	0	0.4999997329749008

Tabla 1: Resultados obtenidos al evaluar $f(x)$ en el punto $x_0 = 1,2 \times 10^{-5}$ con simple y doble precisión.

Claramente vemos que el programa en precisión simple esta fallando. El problema viene al evaluar la función $\cos(x)$ en un número relativamente cercano a 0. Sabemos que un número guardado en precisión simple tiene 7 u 8 cifras significativas. Ahora bien si calculamos el valor de $\cos(x_0)$ con al menos 12 cifras exactas, por ejemplo, tenemos que $\cos(x_0) \approx 0.999999999928 \times 10^0$. Pero inmediatamente vemos que éste número aproximado a 8 cifras significativas es exactamente 1, por lo que el ordenador, en precisión simple, intuye que $\cos(x_0) = 1$. Y por lo tanto:

$$f(x_0) = \frac{1-\cos(x_0)}{x_0^2} \approx \frac{1-1}{x_0^2} = 0$$

Y de aquí sale el 0 obtenido al evaluar la función. Con precisión doble éste fenómeno no pasa porque un número guardado en precisión doble tiene 15 o 16 cifras significativas con lo cual viendo la expresión anterior de $\cos(x_0)$ con 12 cifras se intuye que, en doble precisión, $\cos(x_0) \neq 1$.

Para reducir el error utilizando la expresión (1), emplearemos la relación trigonométrica siguiente:

$$1 - \cos(x) = 2\sin(x/2)^2.$$

Substituyendo ésta expresión en (1) obtenemos la nueva función

$$g(x) = \begin{cases} \dfrac{2\sin(x/2)^2}{x^2} & \text{si} \quad x \neq 0 \\ \dfrac{1}{2} & \text{si} \quad x = 0 \end{cases} \tag{2}$$

que es matemáticamente equivalente a $f(x)$.

Los resultados obtenidos al evaluar $g(x)$ en $x = x_0$ se muestran en la tabla siguiente:

	Algoritmo en precisión simple	Algoritmo en precisión doble
$f(x_0)$	0.5	0.499999999994

Tabla 2: Resultados obtenidos al evaluar $g(x)$ en el punto $x_0 = 1,2 \times 10^{-5}$ con simple y doble precisión.

Observando las tablas 1 i 2 vemos una clara diferencia entre los resultados, especialmente en precisión simple. La razón por la que $g(x_0) = 0.5 \neq 0$ es la siguiente. Fijémonos que $\sin(x_0/2) \approx 5.99999999996400 \times 10^{-6}$. Por lo tanto, trabajando en precisión simple, tenemos que $\sin(x_0/2) = 6 \times 10^{-6}$. Haciendo los cálculos obtenemos:

$$\frac{2 \times \sin(x_0/2) \times \sin(x_0/2)}{x_0 \times x_0} = \frac{2 \times 6 \times 10^{-6} \times 6 \times 10^{-6}}{1.2 \times 10^{-5} \times 1.2 \times 10^{-5}} = \frac{7.2 \times 10^{-11}}{1.44 \times 10^{-10}} = 0.5.$$

En éste caso no tenemos error de cancelación como lo teníamos en el caso de $f(x_0)$.

Notemos que también obtenemos una diferencia, más pequeña por eso, en los valores de $f(x_0)$ i $g(x_0)$ en precisión doble. Esto también se debe al mismo factor de cancelación que hemos comentado. En efecto, tenemos que $\cos(x_0) = 0.9999999999280000$, en precisión doble. Es importante mencionar que a partir del último 0, los números que aparezcan no son relativos de la propia función, ya que en precisión doble trabajamos solo con 16 cifras significativas. Es por eso que al restar éste último número a 1 obtenemos solamente las primeras cifras correctas y las demás son inciertas. Y por lo tanto, una vez dividido por x^2, obtenemos los primeros dígitos correctos (0.499999) pero los dígitos a continuación incorrectos. Observamos que utilizando la función $g(x)$ este problema no ocurre.

Problema 2

La solución de una ecuación cuadrática con coeficientes reales,

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

se obtiene a partir de la expresión

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3}$$

Suponiendo que $a > 0$ y $b^2 > 4ac$, como se menciona en el enunciado, podríamos tener errores de cancelación en una de las soluciones cuando $b^2 \gg 4ac$. En efecto, si $b^2 \gg 4ac$ tenemos que $\frac{4ac}{b^2} \ll 1$. Ahora bien si $\frac{4ac}{b^2}$ es del orden de 10^{-8}, quiere decir que b^2 es del orden de 10^8 veces mas grande que $4ac$. En precisión simple esto se traduce a la igualdad $b^2 - 4ac = b^2$, ya que la primera cifra de b^2 y la primera cifra de $4ac$ están separadas por más de 8 posiciones (que es el numero de cifras significativas con el que trabajamos en precisión simple) en el valor $b^2 - 4ac$. Es por eso que una de las soluciones seria 0.

$$x = \frac{-b \pm |b|}{2a} = \begin{cases} \dfrac{-b+b}{2a} = 0 \quad \text{y} \quad \dfrac{-b-b}{2a} = \dfrac{-b}{a} & \text{si} \quad b > 0 \\ \dfrac{-b+(-b)}{2a} = \dfrac{-b}{a} \quad \text{y} \quad \dfrac{-b-(-b)}{2a} = 0 & \text{si} \quad b < 0 \end{cases} \tag{4}$$

Sin embargo, podemos prescindir de este error multiplicando y dividiendo la solución que sufre éste error de cancelación por el conjugado. Para el caso $b > 0$, tenemos que:

$$\frac{(-b+\sqrt{b^2-4ac})}{2a} \cdot \frac{(-b-\sqrt{b^2-4ac})}{(-b-\sqrt{b^2-4ac})} = \frac{b^2 - b^2 + 4ac}{2a \cdot (-b-\sqrt{b^2-4ac})} = \frac{4ac}{2a \cdot (-b-\sqrt{b^2-4ac})} = \frac{-2c}{b+\sqrt{b^2-4ac}}. \tag{5}$$

Notemos que para valores $b^2 \gg 4ac$ tendremos $\dfrac{-2c}{b+\sqrt{b^2-4ac}} \approx \dfrac{-c}{b}$.

Si $b < 0$, procedemos de forma análoga a (5).

$$\frac{(-b-\sqrt{b^2-4ac})}{2a} \cdot \frac{(-b+\sqrt{b^2-4ac})}{(-b+\sqrt{b^2-4ac})} = \frac{b^2 - b^2 + 4ac}{2a \cdot (-b+\sqrt{b^2-4ac})} = \frac{4ac}{2a \cdot (-b+\sqrt{b^2-4ac})} = \frac{-2c}{b-\sqrt{b^2-4ac}}. \tag{6}$$

Veamos ahora un ejemplo de cómo se comportan estos dos algoritmos para calcular las raíces del polinomio $x^2 + 100000x + 3$ en precisión simple y doble. Los resultados se exponen en la siguiente tabla:

	Precisión simple		Precisión doble	
	solución 1	solución 2	solución 1	solución 2
Método ordinario	0	-100000	$-3.000000288011506 \times 10^{-5}$	-99999.99997
Método propuesto	$-2.9999999 \times 10^{-5}$	-100000	$-3.0000000009 \times 10^{-5}$	-99999.99997

Tabla 3: Soluciones del polinomio $x^2 + 100000x + 3$ mediante los dos métodos explicados.

En este caso, en precisión simple, es decir trabajando con 8 cifras significativas, tenemos que:

$$b^2 - 4ac = 100000^2 - 4 \cdot 3 = 10^{10} - 12 \approx 10^{10}.$$

Y por consiguiente

$$\frac{-b+\sqrt{b^2-4ac}}{2a} \approx \frac{-100000 + \sqrt{10^{10}}}{2} = \frac{-100000 + 100000}{2} = 0.$$

Sin embargo mediante el método propuesto éste valor se convierte en $-2.9999999 \times 10^{-5}$, que tiene un error mucho menor que el calculado con el método ordinario.

Por otro lado, observamos que de forma similar al problema 1, en precisión doble también observamos cierta incertidumbre en los últimos dígitos de la solución 1 calculada con el método ordinario. Esto se debe al mismo error de cancelación con el cual obtenemos la solución 0 en precisión simple.

Problema 3

En estadística la varianza muestral de n números se define como

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}), \quad \text{donde} \quad \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{7}$$

O alternativamente como:

$$s_n^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]. \tag{8}$$

Denotaremos estos dos métodos para calcular la varianza muestral como V_1 y V_2, respectivamente. En lo que sigue mostramos como se comportan las fórmulas de V_1 i V_2 para varios ejemplos de muestras y en las distintas precisiones.

Para $n = 3$, consideramos el vector $x = \{10000, 10001, 10002\}$. Los resultados de los cálculos se muestran en la tabla siguiente:

	Precisión simple	Precisión doble
V_1	1	1
V_2	0	1

Tabla 4: Comportamiento de la varianza muestral mediante las diferentes fórmulas.

Observamos que mientras la fórmula V_1 en precisión simple da 1, V_2 da 0. Esto se debe a que el cálculo en la segunda fórmula utiliza valores de orden de 10^8 (procedentes de 10000^2) en donde también aparecen dígitos no nulos en las últimas cifras de los números (la de las unidades). Concretamente, si hacemos el cálculo de $\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$ de la fórmula (8) con los datos propuestos vemos que da

$$\frac{1}{3}\left(\sum_{i=1}^{3} x_i\right)^2 = \frac{1}{3}(10000 + 10001 + 10002)^2 = 300060003. \tag{9}$$

Calculemos ahora $\sum_{i=1}^{n} x_i^2$:

$$\sum_{i=1}^{3} x_i^2 = 10000^2 + 10001^2 + 10002^2 = 300060005. \tag{10}$$

Fijémonos que los valores de (9) y (10) son valores iguales en precisión simple ya que solo cambia la novena cifra significativa. Es decir, el programa los lee como 3.0006×10^8 en ambos casos, lo que hace que al restarlos de 0.

Esto no ocurre en V_1 ya que cuando elevamos al cuadrado, que es la operación "peligrosa", ya hemos restado previamente los números, reduciendo así su orden de magnitud. De ésta forma, el programa no tiene problemas de cancelación. Es decir como $\bar{x} = \frac{1}{3}\sum_{i=1}^{n} x_i = 10001$, entonces

$$V_1 = \frac{1}{2}\sum_{i=1}^{3}(x_i - \bar{x}) = \frac{1}{2}\left[(10000 - 10001)^2 + (10001 - 10001)^2 + (10002 - 10001)^2\right] = \frac{1}{2}\left[1 + 0 + 1\right] = 1.$$

Veamos mejor estas discrepancias usando los siguientes vectores de 200 componentes:

$$x_1 = \{1000000, 1000001, \ldots, 1000199\} \quad \text{y} \quad x_2 = \{1, 1.000001, 1.000002, \ldots, 1.000199\}.$$

	Precisión simple		Precisión doble	
	Vector x_1	Vector x_2	Vector x_1	Vector x_2
V_1	3350.1887	3.3502197×10^{-9}	3350	$3.350000000000287 \times 10^{-9}$
V_2	843076.19	$-1.5335466 \times 10^{-7}$	3350	$3.350000478024609 \times 10^{-9}$

Tabla 5: Discrepancias de la varianza muestral mediante x_1 y x_2

Observamos que en el caso del vector x_1, utilizando V_2 obtenemos un valor mucho más grande del que deberíamos. De nuevo, esto se debe a la pérdida de números significativos al elevar al cuadrado los términos de la fórmula V_2. Un caso más interesante es el que ocurre con el vector x_2. Resulta que aplicando el método V_2 obtenemos una varianza negativa, que no puede ser de ninguna manera! En efecto, si calculamos la suma $\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$ aplicada al vector x_2 obtendremos, en precisión simple, el valor de 200.03986. Por otro lado, el valor de $\sum_{i=1}^{n} x_i^2$ da 200.03983, de donde se desprende que la diferencia entre estos dos valores es negativa. Esto se debe, una vez más, a la pérdida de cifras significativas al evaluar los números al cuadrado.

Problema 4

En la tabla siguiente se muestra el valor de las sumas parciales S_N de la serie numérica

$$S = \sum_{k=1}^{\infty} \frac{1}{k^2}, \tag{11}$$

calculadas en precisión simple y doble y en orden creciente y decreciente para varios valores de N.

	Valor de S_N en precisión simple		
	$N = 5000$	$N = 10000$	$N = 30000$
Orden creciente	1.6447253	1.6447253	1.6447253
Orden decreciente	1.6447341	1.6448340	1.6449008

	Valor de S_N en precisión doble		
	$N = 5000$	$N = 10000$	$N = 30000$
Orden creciente	1.644734086846901	1.644834071848065	1.644900734070444
Orden decreciente	1.644734086846893	1.644834071848060	1.644900734070442

Tabla 6: Cálculo de la suma S_N para varios valores de N en simple y doble precisión

El valor exacto de la suma de (11) sabemos que es $S = \pi^2/6 = 1.644934066848226....$ De nuevo, en precisión doble no hay nada que comentar; ambas sumas (en orden creciente y decreciente) coinciden, excepto en los últimos dígitos, dónde pueden variar varios dígitos debido a las aproximaciones de los números en punto flotante. El problema está al sumar la serie utilizando precisión simple. Fijémonos que, sumando la serie en orden ascendente, un vez llegado al término 5000 de la suma tenemos que realizar la operación

$$S_{5000} = S_{4999} + \frac{1}{5000^2} = 1.64... + 4 \times 10^{-8}.$$

Ahora bien como trabajamos en precisión simple tenemos que ésta última suma da exactamente S_{4999}, ya que sólo se almacenan 8 dígitos significativos y justamente el dígito 4 de 4×10^{-8} está en la posición de la novena cifra significativa en la suma $1.64... + 4 \times 10^{-8}$. Claramente éste error también pasará al sumar los términos $5001, 5002, \ldots, 10000$. Es por eso que el valor de S_N en precisión simple y orden creciente para $N = 5000$, $N = 10000$ i $N = 30000$ da exactamente el mismo resultado.

Por otro lado si hacemos la suma en orden decreciente, éste error de aproximación no se percibe por el siguiente motivo. Al empezar la suma S_N des de N, tenemos que, como los primeros dígitos son $1/N^2$, $1/(N-1)^2, \ldots$ y éstos tienen la misma magnitud, no perdemos los primeros dígitos significativos de éstos valores que sí que perdíamos sumando la serie en orden ascendente. A medida que vayamos sumando términos más grandes, iremos perdiendo los últimos dígitos significativos del valor de la suma acumulada. Pero éstas pérdidas las haremos de forma "progresiva"de manera que no se percibirá tanto el error final. Con el método anterior, ésta "progresividad"no la teníamos ya que a partir de un cierto número las sumas parciales permanecían constantes.

Para calcular una fórmula alternativa que se comporte mejor que (11) utilizaremos la serie de potencias de la función $\arcsin x$:

$$\arcsin x = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{(2k-1)!!}{(2k)!!} \frac{x^{2k+1}}{2k+1}.$$

Notemos que $\arcsin(1/2) = \pi/6$ y por lo tanto tenemos que:

$$\frac{\pi}{6} = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{(2k-1)!!}{(2k)!!} \frac{1}{2k+1} \frac{1}{2^{2k+1}} \implies \frac{\pi^2}{6} = 6\left(\frac{1}{2} + \sum_{k=1}^{\infty} \frac{(2k-1)!!}{(2k)!!} \frac{1}{2k+1} \frac{1}{2^{2k+1}}\right)^2. \tag{12}$$

Denotamos por $R_N := \frac{1}{2} + \sum_{k=1}^{N} \frac{(2k-1)!!}{(2k)!!} \frac{1}{2k+1} \frac{1}{2^{2k+1}}$. En la tabla siguiente se muestran las aproximaciones parciales de $\pi^2/6$ obtenidas a partir de la fórmula (12) haciendo la suma de R_N en orden decreciente.

	$N = 4$	$N = 9$	$N = 21$
Valor de $6 \cdot (R_N)^2$ en precisión simple	1.6448488	1.6449342	1.6449342
Valor de $6 \cdot (R_N)^2$ en precisión doble	1.644848740989538	1.644934034679181	1.644934066848226

Tabla 7: Cálculo de la suma $6 \cdot (R_N)^2$ para varios valores de N en simple y doble precisión

Notemos la diferencia de velocidad de convergencia de éste método en comparación con el anterior. En particular, en precisión doble, con solamente 21 términos somos capaces de conseguir 16 cifras (el máximo posible) exactas de $\pi^2/6$ mientras que con el método anterior y utilizando 30000 términos sólo eramos capaces de calcular 5 cifras exactas de éste mismo valor.

Conclusiones

La conclusión principal de ésta practica es que los números guardados en punto flotante no son exactos. El ordenador tiene una memoria finita y es por eso que tiene que decidir cuántos dígitos significativos guardar, según en qué precisión se esté trabajando. Este motivo es la causa principal de muchos errores observados a lo largo de la práctica.

Entre ellos está el de no poder sumar (o restar) un número muy grande con un número muy pequeño ya que si la primera cifra significativa del primer número y la primera cifra significativa del segundo están separadas por más de 8 posiciones (trabajando en precisión simple), entonces el valor de la suma será el mismo que el número grande. Este es uno de los varios casos que nos hemos ido encontrando a lo largo de esta práctica.

Otro caso peculiar y sorprendente es que el problema 4 nos ha dejado muy claro que las operaciones en punto flotante no son asociativas. En efecto, empezando a sumar los términos de una serie parcial desde el inicio hemos obtenido diferentes resultados que empezando a sumar los términos desde el final.

Observando los resultados de cada problema, podemos concluir que con precisión doble éstos errores, no ocurren tan a menudo, al menos en la magnitud trabajada a lo largo de la práctica. Es por eso que concluimos que en general vale la pena, si es posible, trabajar siempre en precisión doble para evitar estos errores, que a veces pueden pasar desapercibidos.