Càlcul numèric
Eduardo Pérez Motato. (31/03/2024)

Pràctica 1
NIU: 1709992

Exercici 1: Considerar la funció

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

Volem avaluar $f(x_0)$ per al valor $x_0 = 1.2 \times 10^{-5}$.

a) Escriure dos **programes en C**, un en **precisió simple** i un altre en **precisió doble** que avaluïn la funció $f(x)$.
Calcular per cadascun dels programes el valor $f(x_0)$.
Comparar i comentar els resultats.
Solució. A `Pr1Ex1a.c` creem dues funcions, `fsimp` amb precisió float i `fdoble` amb precisió doble. En avaluar x_0, en el cas del simple retorna 0, i en el cas del doble retorna ≈ 0.4999997, el que s'assembla més al valor real (ja que $\lim_{x\to 0} f(x) = \frac{1}{2}$). □

b) Reescriure la funció $f(x)$ fent servir fórmules trigonomètriques de forma que es redueixi l'error que es produeix fent servir la expressió (1).
Solució. A `Pr1Ex1b.c` reemplacem $1 - \cos x$ per $2\sin^2\left(\frac{x}{2}\right)$ en ambdues funcions, llavors, podem veure com les dues funcions abans creades tenen millor precisió. □

c) Discutir l'observat en aquest exercici.
Solució. Podem veure com, en aquest cas, $1 - \cos x$ perd precisió quan x tendeix a 0. Per allò hem de fer servir una substitució trigonomètrica que tindrà millor precisió. □

Exercici 2: Equació quadràtica.
La solució d'una equació quadràtica amb coeficients reals,

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

s'obté a partir de la expressiò

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

Suposant que $a > 0$ i $b^2 > 4ac$.

a) Escriure dos **programes en C**, un en **precisió simple** i un altre en **precisió doble** que calculin la solució d'una equació quadràtica mitjançant (2).
Solució. A `Pr1Ex2a.c` estan escrites dues funcions, `fquad` en precisió float i `quad` en precisió doble. Al `main`, hi ha un exemple amb $a = 1, b = 2$ i $c = 1$, donant el resultat correcte de -1 com a solució doble. □

b) Comprovar que si $b^2 >> 4ac$ una de les dues fórmules per al càlcul de les arrels amb (2) produeix resultats contaminats amb error de cancel·lació.
Solució. Com podem veure a `Pr1Ex2b.c`, on calculem les arrels de $a = 1, b = 40000$ i $c = 1$. En aquest cas, tenim $b^2 = 40000^2$ que és, òbviament, molt més gran que $4ac = 4$. En calcular, veiem que la funció amb precisió float dona 0 i -40000 com a solució, mentre en doble dona $\approx -2.5 \times 10^{-5}$ i -39999.999975, un resultat força més creïble. □

c) Proposa un procediment alternatiu per al càlcul de les arrels que eviti l'error de cancel·lació.
Solució. Sabem que (2) és equivalent a

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \underbrace{\frac{-b \mp \sqrt{b^2 - 4ac}}{-b \mp \sqrt{b^2 - 4ac}}}_{=1} = \frac{b^2 - (b^2 - 4ac)}{2a\left(-b \mp \sqrt{b^2 - 4ac}\right)}$$

i si continuem simplificant, tenim

$$\frac{2c}{-b \mp \sqrt{b^2 - 4ac}} \tag{3}$$

Ara, amb (3), canviem el programa anterior i fem els canvis a `Pr1Ex2c.c` per al cas on l'arrel és positiva, ja que és allí on hi ha error de cancel·lació. En cas de posar-ho també quan l'arrel es negativa crearia un altre error de cancel·lació fatal que produirà $-\infty$. Amb els canvis fets, al evaluar-ho dona un resultat més correcte en cadascun. □

d) Construir exemples numèrics on el càlcul de les arrels en simple i doble precisió proporcionin diferències significatives en exactitud fent servir (2) i el procediment que has proposat.
Solució. Un exemple numèric molt bo i il·lustratiu és el que hi ha a `Pr1Ex2b.c`. En el cas $a = 1, b = 40000, c = 1$, podem observar com (2) (implementat a `Pr1Ex2a.c` i `Pr1Ex2b.c`) dona, en el cas float, 0 i -40000. En el programa fet amb (3), en cas float, dona $-2.4999999 \times 10^{-5}$ i -40000. □

Exercici 3: Càlcul de la variància mostral.
En estadística la variància mostral de n nombres es defineix com

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\,,\ \text{on } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

Una fórmula alternativa equivalent que fa servir un nombre d'operacions similars és

$$s_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) \tag{5}$$

Aquesta última fórmula pot sofrir error de cancel·lació!

a) Escriure **programes en C en simple i doble precisió** que calculin la variància mostral amb les mateixes fórmules on l'input sigui un vector de nombres reals i l'output sigui la variància mostral.
Solució. A `Pr1Ex3a.c` es creen 4 funcions, dues amb (4) on una és simple i l'altre és double; i altres dues amb (5), una en simple i altre en double. Aquest programa demana el nombre de components del vector i després demana cadascun dels components. Finalment, calcula la variància mostral i mostra en cadascun dels casos l'output de la funció en qüestió. □

b) Considerar el vector $x = (10000, 10001, 10002)$ i calcular la variància amb els programes generats. Analitzar les discrepàncies.
Solució. Amb aquest vector, totes les funcions donen 1 menys la simple amb (5), que dona 0. Això és degut al fet que en aquest cas hi ha un error de cancel·lació. □

c) Construir dos exemples de vectors de dimensió gran (almenys 100 components) on aquestes discrepàncies siguin més evidents.
Solució. A `Pr1Ex3c.c` he posat els exemples dels vectors $(10000, 10001, \ldots, 10099)$ i ($\underbrace{10000}_{33\text{ vegades}}$, $\underbrace{10001}_{34\text{ vegades}}$, $\underbrace{10002}_{33\text{ vegades}}$). En el primer cas podem veure un error de cancel·lació de $\Delta x \approx 16.8384$ mentre al segón hi ha $\Delta x \approx 40.7070$, amb un error relatiu ≈ 61.0606. □

d) Discutir les diferències en els resultats.
Solució. Podem veure que, en el cas de la fórmula amb possible cancel·lació (5) en simple hi ha una alta probabilitat de tindre un error de cancel·lació, mentre ambdues fórmules són aproximadament igual d'eficients, llavors no dona benefici fer-la servir. □

Exercici 4: Suma d'una sèrie.
És conegut que la sèrie dels recíprocs dels quadrats dels nombres naturals convergeix i la seva suma és $\frac{\pi^2}{6}$

$$S = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \approx 1.644934066848226\ldots \tag{6}$$

Volem calcular aproximadament la suma S sumant termes (sumes parcials) de la sèrie i establirem dues estratègies per programar-les en C en **simple i doble precisió.**

a) Escriure programes C que calculin la suma dels termes de la sèrie S en ordre creixent fins a un terme màxim $(5000, 10000, \ldots)$ on les dades siguin el nombre de termes a sumar.
Solució. Suposant que quan es refereix en ordre creixent es refereix a k creixent.
A `Pr1Ex4a.c` es creen dues funcions, una en simple i l'altra en doble, que calculen en ordre creixent (6), sent més eficient la doble. □

b) Escriure programes en C (doble i simple precisió) que sumin els termes de la sèrie en ordre decreixent.
Solució. A `Pr1Ex4b.c` es creen dues funcions, una en simple i l'altra en doble, que calculen en ordre decreixent (6), sent més o menys igual d'eficients però la simple amb menor precisió. □

c) Comparar els resultats anteriors amb el valor exacte i justifica els diferents resultats.
Solució. Podem veure que el cas d'ordre decreixent de k millora el càlcul. Això a causa de les propietats dels punts flotants en fer la suma els bits més petits del nombre inferior es perden, en aquest cas quan k decreix, $\frac{1}{k^2}$ creix i això fa que al primer sumar els nombres més petits hi hagi menys error. □

d) Proporcionar una fórmula alternativa que es comporti millor que (6).
Solució. Buscant per internet vaig trobar la seguent igualtat

$$\frac{\pi}{6} = \sum_{k=1}^{\infty} \frac{-120 + 329k + 568k^2}{k(1+k)(1+2k)(1+4k)(3+4k)(5+4k)} \tag{7}$$

llavors, si fem servir això podem arrivar fàcilment a $\frac{\pi^2}{6}$ mitjançant (7) de la següent manera

$$\frac{\pi^2}{6} = 6\left(\sum_{k=1}^{\infty} \frac{-120 + 329k + 568k^2}{k(1+k)(1+2k)(1+4k)(3+4k)(5+4k)}\right)^2 \tag{8}$$

això programat en `Pr1Ex4d.c`, podem veure que en el cas doble (perqué en el simple no es pot demanar més precisió) l'error és de -2.22×10^{-16}, és a dir, un error molt petit. □

Exercici 5: Escriure conclusions sobre l'observat i après en aquesta pràctica. Extensió màxima de mitja pàgina.

Solució. A l'hora de fer càlculs hem de tindre molta cura amb la precisió de les funcions que utilitzem, tant en el tipus de dada rebuda i retornada, com els limits de precisió de la propia funció (com que per exemple si fem 1.2×10^{-5} el cosinus no dona un valor correcte i en canvi el sinus si). L'ordre de les operacions també afecta i és millor sumar o restar sempre, a ser possible, en el mateix ordre de magnitud. □