Problema 6 llista 1

Marco Praderio 1361525

Quina de les següents aproximacions de π, delimita millor la propagació de l'error:

$$\text{(a) } \pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9}\cdots\right) = 4\sum_{i=0}^{\infty}\frac{(-1)^i}{2i+1}$$

$$\text{(b) } \pi = 6\left(0.5 + \frac{0.5^3}{2\cdot 3} + \frac{3\cdot 0.5^5}{2\cdot 4\cdot 5} + \frac{3\cdot 5\cdot 0.5^7}{2\cdot 4\cdot 6\cdot 7} + \cdots\right) = 6\sum_{i=0}^{\infty}\frac{0.5^{2i+1}}{2i+1}\prod_{j=1}^{i}\frac{2j-1}{2j}$$

Abans de començar serà necessari fer algunes suposicions i observacions sobre la manera en la que és calculen la sèrie (a) i la sèrie (b).

- **Podem descartar errors de segon ordre:** Els de primer ordre resulten molt superiors.
- **L'error relatiu comés per representació és de ε i per operacions aritmètiques (suma, resta, producte i divisió) és de 2ε:** Aquestes són les dades presentades en un exemple fet a classe de teoria.
- **Podem reescriure la sèrie (a) sumant termes dos a dos per obtindre $\pi = 8\sum_{i=0}^{\infty}\frac{1}{(4i+1)(4i+3)}$:** Aquesta modificació, te dos avantatges. Primer porta a la definició de la successió $S_n^a = \sum_{i=0}^{n}\frac{1}{(4i+1)(4i+3)}$ la qual té la peculiar característica de res monòtona creixent la qual cosa fa que sigui més fàcil estudiar-la. Segona convergeix obviament el doble de ràpid que la sèrie $4\sum_{i=0}^{\infty}\frac{(-1)^i}{2i+1}$ i propaga menys l'error en quant la única diferència que te respecte a la sèrie anterior és que fem sumes de forma exacta[1].
- **Els únics errors de representació de la sèrie (a) són els que venen donats per la representació de les fraccions:** Aquesta suposició és molt realista en quant tots els enters més petits que el valor màxim representable són números màquina i, per tant, no tenen error de representació i els enters en el denominador de la sèrie (a) reescrita creixen de manera prou lenta (cuadràticament) com per no causar immediatament un overflow en la representació com a long int. Aquesta suposició no és però valida en el cas de la sèrie (b). Tot i que el valor 0.5 es pugui interpretar com la fracció $\frac{1}{2}$ i, per tant, com un 2 en el denominador el qual no tindria error de representació els nombres al numerador i al denominador de la sèrie (b) creixen de manera tan ràpida (semifactorialment) que enseguida (menys de 1000 iteracions) seria impossible guardar-los com a enters fins i tot en una variable del tipus long int. Ens topem amb dos possibles solucions per solucionar aquest problema. La primera consisteix en calcular les fraccions com a productori de fraccions amb numerador i denominador prou petits com per ser guardats en una variable del tipus long int. En aquest cas seguiríem tenint error de representació únicament en les fraccions numerador i denominador petits però com que que cada fracció s'obtindria com a producte de varies d'aquestes fraccions al final l'error s'acumularia i no sortiria a compte[2]. L'altra opció consisteix en guardar numerador i denominador com a doubles els quals tenen un rang de representació molt mes gran que els long int. Això implicaria no obstant que tant numerador com denominador en cadascuna de les fraccions de la sèrie (b) estarien sotmesos a errors de representació amb punt flotant. Tot i així aquest mètode és preferible al altre (sempre i quan els nombres a numerador i denominador no provoquin un overflow en la representació amb coma flotant[3]) en quant, d'aquesta manera, tindríem que el error relatiu de cada fracció de la sèrie és limitaria a 4ε

[1] podem considerar que la suma $\frac{1}{4n+1} - \frac{1}{4n+3} = \frac{2}{(4i+1)(4i+3)}$ està feta de manera exacta en quant els enters representables són nombres màquina i el seu producte (si representable) segueix sent un nombre màquina i, per tant, l'únic error fet en aquesta suma és el de representació de la fracció final que, per altra banda, ja existia en la representació de cadascuna de les dues fraccions inicials.

[2] L'error associat a cada fracció (suposant de manera exageradament optimista que no hi hagués error en el producte de fraccions) seria igual a la suma dels errors associats a cada una de les fraccions que la formen o sigui seria $n\varepsilon$ on n indica el nombre de fraccions amb numerador i denominador petits que formen cada fracció de la sèrie.

[3] Cosa que trigarà bastant més temps en passar però eventualment passarà igual que eventualment tindrem el mateix problema per representar els denominadors de la sèrie (a) de manera exacta i que només podem intentar allunyar construint variables que ocupin sempre mes i mes memòria per poder representar una major varietat de nombres.

(2ε per les representacions de numerador i denominador i 2ε per la operació de divisió[4]).

- **El calcul de les fraccions del cas (b) calculant numerador i denominador com a doubles i després dividint:** És important mantindre en ment que d'aquesta manera obtindrem donarà un error relatiu de 4ε per a cada fracció. En el cas (a) d'altra banda tindríem que el error relatiu de cada fracció de la seria vindrà donat únicament per la operació de divisió entre numerador i denominador[5] i, per tant, serà de 2ε tal i com hem especificat en la segona suposició.
- **Les sèries es calculen en l'ordre indicat tenint en compte la nova representació de la sèrie (a):** És important tindre en compte l'ordre amb que es suma en quant la suma en punt flotant no és associativa.

A partir d'aquestes hipòtesis podem estudiar com augmenta l'error relatiu per a cada fracció que sumem en la sèrie (a) i en el cas (b).
Denotem per S_N la suma dels N primers termes de la sèrie (a) o (b) i per E_N l'error associat a S_N. Si a més a més denotem per $\frac{x_{N+1}}{y_{N+1}}$ el terme $N+1$ de la suma de qualsevol de les dues sèries obtenim

$$\begin{aligned}\tilde{S}_{N+1} &= \left(\tilde{S}_N + \frac{\tilde{x_{n+1}}}{y_{n+1}}\right)(1+\delta_3) = \left(S_N(1+\delta_1) + \frac{x_{n+1}}{y_{n+1}}(1+\delta_2)\right)(1+\delta_3) \approx \\ &\approx \left(S_N + \frac{x_{n+1}}{y_{n+1}}\right)\left(1 + \frac{S_N}{S_N + \frac{x_{n+1}}{y_{n+1}}}\delta_1 + \frac{\frac{x_{n+1}}{y_{n+1}}}{S_N + \frac{x_{n+1}}{y_{n+1}}}\delta_2 + \delta_3\right) = \\ &= (S_{N+1})\left(1 + \frac{S_N}{S_{N+1}}\delta_1 + \frac{\frac{x_{n+1}}{y_{n+1}}}{S_{N+1}}\delta_2 + \delta_3\right)\end{aligned}$$

On $|\delta_1| < E_N$, $|\delta_3| < 2\varepsilon$ i, per lo dit anteriorment, $|\delta_2| < 2\varepsilon$ en el cas de la sèrie (a) i $|\delta_2| < 4\varepsilon$ en el cas de la sèrie (b).

Notem ara que si el nostre objectiu és comparar els errors propagats en la equació (a) i en la equació (b) no fa falta tindre en compte el nombre per el qual es multipliquen les sèries (8 en el cas (a) i 6 en el cas (b)) en quant, donat que una multiplicació senzillament suma els errors relatius i afegeix 2ε [6] i els errors de representació de 8 i 6 són els dos iguals[7] i, per tant, aquesta operació augmenta l'error de la mateixa forma tant per la sèrie (a) com per la sèrie (b) o sigui que considerar aquesta multiplicació no és rellevant per comparar les dues sèries.
Si ara denotem per S_N^a la suma dels N primers termes de la sèrie (a), per S_N^b la suma dels N primers termes de la sèrie (b) i anàlogament amb els errors relatius E_N^a i E_N^b tindrem, en el cas (a)

$$\begin{aligned}E_{N+1}^a &= \frac{S_N^a}{S_{N+1}^a}E_N^a + \frac{1}{(4N+1)(4N+3)}\frac{2\varepsilon}{S_{N+1}^a} + 2\varepsilon = \\ &= \frac{S_{N+1}^a}{S_{N+1}^a}E_N^a + \frac{1}{(4N+1)(4N+3)}\frac{2\varepsilon - E_N^a}{S_{N+1}^a} + 2\varepsilon = \\ &= E_N^a + \frac{1}{(4N+1)(4N+3)}\frac{2\varepsilon - E_N^a}{S_{N+1}^a} + 2\varepsilon\end{aligned} \tag{1}$$

[8] i en el cas (b)

$$\begin{aligned}E_{N+1}^b &= \frac{S_N^b}{S_{N+1}^b}E_N^b + \frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{4\varepsilon}{S_{N+1}^b} + 2\varepsilon \\ &= \frac{S_{N+1}^b}{S_{N+1}^b}E_N^b + \frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{4\varepsilon - E_N^b}{S_{N+1}^b} + 2\varepsilon = \\ &= E_N^b + \frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{4\varepsilon - E_N^b}{S_{N+1}^b} + 2\varepsilon\end{aligned} \tag{2}$$

Com podem observar, cada vegada que es suma un terme de la sèrie l'error relatiu augmenta de, com a mínim 2ε per tant podem afirmar que, a partir de la segona suma es complirà que E_N^b,$E_N^a > 4\varepsilon$ per a tot $N > 2$ és més podem

[4] $fl\left(\frac{x}{y}\right) = \frac{fl(x)}{fl(y)}(1+2\delta_3) = \frac{x(1+\delta_1)}{y(1+\delta_2)}(1+2\delta_3) \approx \frac{x}{y}(1+\delta_1)(1-\delta_2)(1+2\delta_3) \approx \frac{x}{y}(1+\delta_1-\delta_2+2\delta_3)$ i, com que $|\delta_3| < \varepsilon$ aleshores l'error relatiu de la fracció $\frac{x}{y}$ serà més petit que 4ε.

[5] Numerador i denominador no tindran error de representació en quant seran enters i, per tant, els podem suposar nombres màquina màquina.

[6] $fl(x\cdot y) = (fl(x)fl(y))(1+2\delta_3) = x(1+\delta_1)y(1+\delta_2)(1+2\delta_3) \approx xy(1+\delta_1+\delta_2+2\delta_3)$ on $|\delta_1|$,$|\delta_2|$, $|\delta_3| < \varepsilon$

[7] Sent nombre màquina ambdós errors de representació són 0.

[8] Tant S_N^a com S_N^b i tots els termes de les sèries són positius per a tot $N \in \mathbb{N}$ per tant no fa falta posar mòduls.

afirmar que E_N^a,$E_N^b >> \varepsilon$ bastant ràpidament i podem per tant aproximar $\frac{4\varepsilon - E_N^b}{S_{N+1}^b} \approx -\frac{E_N^b}{S_{N+1}^b}$ i $\frac{2\varepsilon - E_N^a}{S_{N+1}^a} \approx -\frac{E_N^a}{S_{N+1}^a}$. Notem ara que S_N^a i $S_N^b S$ tenen ambdues cotes superior i inferior i, a més a més, els termes de la sèrie (b) decreixen prou ràpid com perquè es compleixi en molt poc tems (menys de 10 iteracions) que

$$\frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{1}{S_{N+1}^b} < \frac{1}{(4N+1)(4N+3)S_{N+1}^a}$$

fent servir aquestes observacions i les equacions 1 i 2 podem concloure que, si $E_N^a = E_N^b = E_N$ aleshores

$$\begin{aligned} E_{N+1}^a - 2\varepsilon = E_N + \frac{1}{(4N+1)(4N+3)}\frac{2\varepsilon - E_N}{S_{N+1}^a} \approx E_N - \frac{E_N}{(4N+1)(4N+3)S_{N+1}^a} < \\ < E_N - \frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{E_N}{S_{N+1}^b} \approx E_N^b + \frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j}\frac{4\varepsilon - E_N^b}{S_{N+1}^b} = E_{N+1}^b - 2\varepsilon \end{aligned}$$

Per tant $E_{N+1}^b > E_{N+1}^a$ en altres paraules la sèrie (a) propaga menys l'error a cada iteració.
Tot i aixis, si l'objectiu és el de calcular π aleshores la sèrie b és molt preferible respecte de la sèrie (a) en quant, tot i que a cada suma l'error augmenta més en la sèrie (b) que en la sèrie (a) la sèrie (b) convergeix molt més ràpidament cap a π com podem deduir del fet que els termes de la sèrie (b) es fan petits molt més ràpidament que els termes de la sèrie (a). Aleshores per obtindre una bona aproximació de π són necessàries moltes menys iteracions de la sèrie (b) que de la sèrie (a), per tant, tot i que a cada iteració es propagui una miqueta més l'error a cada iteració de la sèrie la propagació total de l'error en (b) és menor que en (a) on és necessari fer més iteracions.