# Problema 6 llista 1

Marco Praderio 1361525

**Quina de les següents aproximacions de $\pi$, 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 suposicións i observacions sobre la manera en la que és calculen la sèrie (a) i la sèrie (b).

- **L'error relatiu comés per representació és de $\varepsilon$ i per operacions aritmètiques (suma i producte) és de $2\varepsilon$:** Aquestes són les dades presentades en un exemple fet a classe de teoria.
- **L'error en el càlcul de les sèries es genera en el càlcul de les fraccions:** Aquesta suposició és molt realista en quant tots els enters més petits que el valor màxim represenable són números màquina i, per tant, no tenen error de representacio. A més a més el valor 0.5 es pot interpretar com la fracció $\frac{1}{2}$ i, per tant, com un 2 en el denominador o sigui que tanpoc tindria error de representació. Per altra banda, depenent de com es calculi cada terme de la opció (b) si multiplicant tots els termes en el númerador i en el denominador i després dividint-los entre si o bé calculant el terme com el productòri mostrat en la segona igualtat de la opció (b). En el primer cas tindriem el desaventatge de que, com que els termes al numerador i els termes al denominador creixen amb una velocitat semifacorial (en realitat el numerador creix lleugerament més lentament en quant hi ha una potencia de 0.5 multiplicant-lo) aviat es produiria un overflow si intentem guradar-los com a enters i ens veurem per tant obligats a guardar-los com a punts flotants la cual cosa generarà un error de representació. Tot i així aquest mètode és preferible al següent (sempre i quan els nombres a numerador i denominador no provoquin un overflow en la representació amb coma flotant) en cuant el el següent cas tindriem error de representació per a cada fracció i com que es presenten $n$ fraccions multiplicades, totes elles amb error de representació, arribariem a un error relatiu de $n$ vegades el error de representació per cada terme de la suma (suposant que la operació suma es faci sense error). D'altra banda, en el primer cas, tindriem que el error relatiu de cada terme de la suma és limitaria a $4\varepsilon$ ($2\varepsilon$ per les representacions de numerador i denominador i $2\varepsilon$ per la operació producte[1]).
- **El calcul de les fraccions del cas (b) es realitza aplicant el primer mètode descrit en la suposició anterior:** És important mantindre en ment que aquest mètode ens donarà un error relatiu de $4\varepsilon$ per a cada fracció. En el cas (a) tindriem que el error relatiu de cada fracció de la suma serà donat únicament per la operació de divisió entre numerador i denominador[2] i, per tant, serà de $2\varepsilon$.
- **Les séries es calculen en l'ordre indicat:** Si es calculessin en un altre ordre (començant desde termes més avançats i després tornant enrere) es produirien errors de cancelació en la sèrie (a) que dispararien el error relatiu en comptes, si es calculen en ordre de dreta cap a esquerra no es produeixen errors de cancelació en quant cap terme de la sèrie és similar en mòdul a la suma de tots els anteriors.
- **Podem descartar errors de següent ordre:** Els de primer ordre resulten molt superiors.

A partir d'aquestes hipòtesis podem estudiar com augmenta l'error relatiu per a cada fracció que sumem en el cas (a) i en el cas (b).

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

[2] numerador i denominador no tindràn error de representació en quant seràn enters i, per tant els podem suposar nombres màquina màquina

Denotem per $S_N$ la suma dels $N$ primers teres de la série (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 súma de qualsevol de les dues séries obtenim

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

On $|\delta_1| < E_N$, $|\delta_3| < 2\varepsilon$ i, per lo que hem dit en les suposicions, $|\delta_2| < 2\varepsilon$ 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 (4 en el cas (a) i 6 en el cas (b)) en quant, donat que una multiplicació senzillament suma els errors relatius i afegeix $2\varepsilon$ [3] i els errors de representació de 4 i 6 són els dos iguals[4] i, per tant, considerar aquesta multiplicació no és comparativament útil.
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)

$$E_{N+1}^a = \frac{S_N^a}{S_{N+1}^a}E_N^a + \frac{1}{2N+1}\frac{2\varepsilon}{S_{N+1}^a} + 2\varepsilon$$

[5] i en el cas (b)

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

I com que $\prod_{j=1}^{N}\frac{2j-1}{2j} < 1$ aleshores per a $N > 1$

$$\frac{4}{S_{N+1}^b}\frac{0.5^{2N+1}}{2N+1}\prod_{j=1}^{N}\frac{2j-1}{2j} < \frac{4}{S_{N+1}^b}\frac{0.5^{2N+1}}{2N+1} < \frac{4\cdot 6}{\pi}\frac{0.5^{2N+1}}{2N+1} = \frac{12}{\pi}\frac{0.5^{2N}}{2N+1} =$$

$$= \frac{3}{2\pi}\frac{2\cdot 0.5^{2N-2}}{2N+1} < \frac{3}{2}\frac{2\cdot 0.5^{2N-2}}{2N+1} < \frac{3}{2}\frac{2}{2N+1} < \frac{1}{S_N^a}\frac{2}{2N+1}$$

[6] i $\frac{S_N^b}{S_{N+1}^b} < 1$ mentres que $\frac{S_N^a}{S_{N+1}^a}$ a vegades és més gran que 1 i d'altres més petit podem concloure que el augment del error propagat per cada suma de la sèrie decreix exponencialment més ràpid en el cas de la sèrie (b) respecte al cas de la sèrie (a). Per tant podem concloure que la sèrie (b) és millor a l'hora d'evitar la propagació d'error.

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

[4] sent nombre màquina ambdos errors de representació són 0

[5] tant $S_N^a$ com $S_N^b$ són positius per a tot $N \in \mathbb{N}$

[6] $S_N^b < \frac{\pi}{6}$ i $S_N^a > \frac{2}{3}\frac{1}{4}$ com podem veure fàcilment a partir de fet que la sèrie (b) és creixent acotada per $\frac{\pi}{6}$ mentres que la sèrie (a) te el seu mínim per $N = 2$ en $\frac{2}{3}$.