Tema 1 - Más sobre errores de truncamiento

Curso de Física Computacional

M. en C. Gustavo Contreras Mayén

Contenido

Contenido

Números de punto flotante

Un conjunto F de números de punto flotante está caracterizado por los siguientes parámetros:

1. La base del sistema β.
2. El número de dígitos n en la mantisa.
3. Un exponente $m \leq e \leq M$.

donde β, n, m, e son enteros.

Cada número en el sistema de punto flotante F tiene la forma

$$\pm(.d_1 d_2 \dots d_n)_\beta \beta^e$$

donde $d_i = 0, 1, \dots, \beta - 1$, $i = 1, 2, \dots, n$, $(.d_1 d_2 \dots d_n)_\beta$ es una β-fracción llamada *mantisa*, e es un entero llamado *exponente*.

El sistema F de punto flotante está *normalizado* si $d_1 \neq 0$, en general, todos los números flotantes se noramlizan con la excepción del cero, en el cual $d_1 = d_2 = \ldots = d_n = 0$.

El conjunto F es discreto y finito. Su cardinalidad (i.e. el número de elementos que lo constituyen, está dada por

$$2(\beta - 1)\beta^{n-1}(M - m + 1) + 1$$

Como consecuencia de la finitud de F para representar al conjunto de números reales $\mathbb{R}$, existirá una infinidad de números en $\mathbb{R}$ que no pueden representarse en forma exacta en F.

Sea x un número real denotemos por $fl(x)$ el número en F que es más cercano a x. La diferencia entre x y $fl(x)$ se llama *error de rendondeo*, éste depende de la magnitud de x y es por lo tanto medido relativo a x

$$\delta(x) = \frac{fl(x) - x}{x}, \qquad x \neq 0$$

luego $|\delta(x)|$ es el error relativo introducido en la representación de x en el sistema de punto flotante F; de la ecuación anterior obtenemos que

$$fl(x) = x(1 + \delta(x))$$

Hay dos formas generalmente usadas para convertir un número real x a un $n-\beta$ número flotante $fl(x)$: *redondeado* o *truncando*.

Cuando se redondea, $fl(x)$ se elige como el número de punto flotante normalizado más cercano a x, si hay empate se usa alguna regla especial, por ejemplo, se toma el de la derecha.

Si se trunca, $fl(x)$ se escoge como el número flotante normalizado más cercano entre 0 y x, esto es, se toman algunas d's y se desprecian otras.

A continuación encontraremos una cota para $\delta(x)$ que sea independiente de x. En el sistema de números de punto flotante F, el número cuyo valor absoluto es el más pequeño está dado por

$$+(.100\ldots0)_\beta\beta^m = \beta^{m-1}$$

el sucesor inmediato de β^{m-1} se encuentra sumando a éste el número

$$+(.000\ldots1)_\beta\beta^m = \beta^{m-n}.$$

Se concluye entonces, que la distancia entre dos números consecutivos en el intervalo $[\beta^{m-1}, \beta^m]$ es β^{m-n}.

En forma similar se demuestra que la distancia entre dos números consecutivos que pertenezcan a cualquier intervalo de la forma $[\beta^j, \beta^{j+1}] j = m, \ldots M - 1$ está dada por

$$\beta^{j+1-n}$$

Hemos demostrado un aspecto singular de la distribución de los números del sistema de punto flotante F:

- *estos no están igualmente espaciados a través de todo su rango, sino únicamente cuando se encuentran entre potencias sucesivas de la base* β*.*

Sunpongamos que $x \in [\beta^j, \beta^{j+1}]$ para alguna $j = m, \dots M-1$, si el número flotante $fl(x)$ que representa a x es seleccionado por redondeo, entonces de acuerdo con el resultado anterior, el error introducido es a lo más $(1/2)\beta^{j+1-n}$, si $fl(x)$ se selecciona por truncamiento el error es a lo más $\beta^{\mathrm{j}+1-n}$. Lo anterior nos da la *medida del error de redondeo absoluto*

$$|fl(x) - x| \leq \begin{cases} \frac{1}{2}\beta^{j+1-n} & \text{redondeo,} \\ \beta^{j+1-n} & \text{truncamiento} \end{cases}$$

El *error de redondeo relativo* $|\delta(x)|$ se obtiene dividiendo al error de redondeo absoluto por $|x|$.

Dado que $0 < \beta^j \leq |x|$, se tiene

$$|\delta(x)| \leq \begin{cases} \frac{1}{2}\beta^{1-n} & \text{redondeo,} \\ \beta^{1-n} & \text{truncamiento} \end{cases}$$

El épsilon de la máquina se define como

$$\epsilon = \begin{cases} \frac{1}{2}\beta^{1-n} & \text{redondeo,} \\ \beta^{1-n} & \text{truncamiento} \end{cases}$$

De lo anterior, tenemos que se cumple $|\delta(x)| \leq \epsilon$ para toda x.

La cota para $\delta(x)$ independiente de x es el épsilon de la máquina.

Dado $x \in \mathbb{R}$, su flotante $fl(x)$ se define como

$$fl(x) = x(1+\delta), \qquad |\delta| \leq \epsilon$$

La exactitud de la aritmética de punto flotante está entonces caracterizada por el épsilon de la máquina.

Supongamos que estamos en el intervalo $[1, \beta]$. Queremos calcular $1 \oplus \epsilon$, donde $\oplus$ denota la suma entre números que pertenecen a F

$$1 \oplus \epsilon = fl(1 + \epsilon)$$

Si el error introducido en la representación de $1 + \epsilon$ es por redondeo, entonces éste queda localizado en la mitad del intervalo $[1, 1 + \epsilon^{1-n}]$.

Ya que la distancia entre dos números consecutivos de F que pertenecen al intervalo $[1, \beta]$ es β^{1-n}, el número de punto flotante $1 \oplus \epsilon$ es tomado como $1 + \beta^{1-n} > 1$.

Sin embargo, si $0 < \epsilon_1 < \epsilon$ obtenemos por un procedimiento análogo al anterior que $1 \oplus \epsilon_1 = 1$.

Un resultado similar es encontrado cuando el error introducido en la representacion de $1 \oplus \epsilon$ es por truncamiento.

Contenido

Modelo de aritmética en F

La aritmética en el sistema numérico de punto flotante F permite aproximar a la del sistema de números reales $\mathbb{R}$.

Como notación emplearemos $\oplus, \ominus, \otimes, \oslash$ para indicar las aproximaciones a las operaciones aritméticas $+, -, \times, /$ de $\mathbb{R}$:

$$
\begin{aligned}
x \oplus y &= fl(x + y) \\
x \ominus y &= fl(x - y) \\
x \otimes y &= fl(x \times y) \\
x \oslash y &= fl(x/y)
\end{aligned}
$$

El modelo que asumiremos para la artimética en F es el siguiente:

$$fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \quad |\delta| \leq \epsilon, \quad \text{op} = +, -, *, /$$

Para efectuar operaciones en forma manual en este modelo aritmético, por cada operación $+, -, \times, /$ encontrada, hágala en aritmética exacta, normalice el resultado, trunque o redondee de acuerdo al número de dígitos permitido.

Contenido

Problema 1

Considera la siguiente suma finita

$$S_N^{(1)} = \sum_{n=1}^{2N} (-1)^n \frac{n}{n+1} \tag{1}$$

Si sumamos de manera separada los valores impares y los pares de x, tendremos dos sumas:

$$S_N^{(2)} = -\sum_{n=1}^{N} \frac{2n-1}{2n} + \sum_{n=1}^{N} \frac{2n}{2n+1} \tag{2}$$

Tercera suma

Podemos eliminar la diferencia mediante una combinación entre las dos sumas, quedando de la siguiente manera

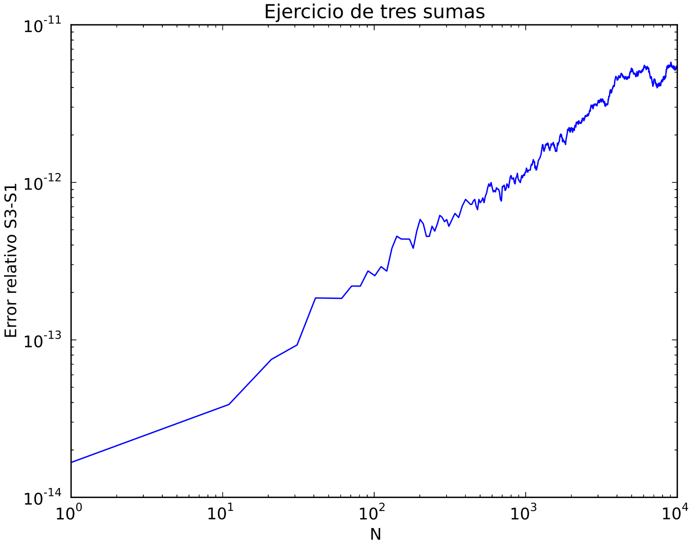
$$S_N^{(3)} = \sum_{n=1}^{N} \frac{1}{2n(2n+1)} \tag{3}$$

Sabemos que aunque el valor de las tres sumas $S_N^{(1)}$, $S_N^{(2)}$, $S_N^{(3)}$, es el mismo, pero el resultado númerico puede ser diferente.

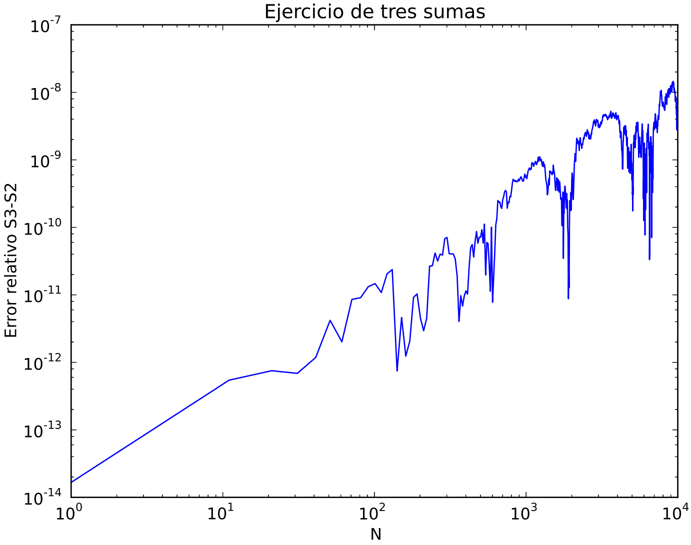
Ejercicio a resolver

1. Escribe un programa que calcule $S_N^{(1)}$, $S_N^{(2)}$, $S_N^{(3)}$.
2. Supongamos que $S_N^{(3)}$ es el valor exacto de la suma. Grafica el error relativo contra el número de términos en la suma (tip: usa una escala log-log). Comienza con $N = 1$ hasta $N = 1000000$. Describe la gráfica.
3. Identifica en tu gráfica una región en donde la tendencia es casi lineal, ¿qué representa ésta sección con respecto al error?

Error relativo entre $S3$ y $S1$



Error relativo entre $S3$ y $S2$



Problema 2

Aunque tengamos el apoyo de una buena computadora, el cálculo de la suma de una serie requiere reflexión y cuidado.
Considera la serie:

$$S^{(u)} = \sum_{n=1}^{N} \frac{1}{n}$$

que será una suma finita mientras N sea finito.
Cuando hacemos la suma de manera analítica, no importa si se hace de manera ascendente: desde $n = 1, 2, 3, \ldots, N-1, N$, o descendente: desde $n = N, N-1, N-2, \ldots, 3, 2, 1$

$$S^{(d)} = \sum_{n=N}^{1} \frac{1}{n}$$

Sin embargo, debido a los errores por redondeo, cuando calculamos de manera analítica, el valor de las sumas no es el mismo, $S^{(u)} \neq S^{(d)}$

1. Escribe un programa que calcule $S^{(u)}$ y $S^{(d)}$ como función de N.
2. Grafica (log-log) la diferencia relativa entre la suma relativa contra N.
3. Identifica en tu gráfica una región en donde la tendencia es casi lineal, ¿qué representa ésta sección con respecto al error?