Universidad Nacional de Ingeniería
Escuela Profesional de Matemática
Ciclo 2022-2

[[**Análisis y Modelamiento Numérico I - CM4F1**]
[**Los profesores**] UNI, 03 de octubre de 2022

Práctica Calificada 1

1. Considere representar el conjunto de números consecutivos

$$1, 2, 3, 4, \cdots, n$$

usando:

a) [1 *pto.*]32 bits sin signo
b) [1 *pto.*]32 bits con signo
c) [2 *ptos.*]punto flotante de precisión simple IEEE

Para cada una de estas representaciones, ¿cuál es el mayor n tal que cada número en el conjunto anterior puede ser representado, esto es, no hay espacios entre los números?

Solución:

a) El mayor sin signo es $2^{32} - 1$

b) El mayor con signo es $2^{31} - 1$

c) Primero, tenga en cuenta que 1 a $2^{24} - 1$ se pueden representar exactamente. ¿Por qué? Porque cuando se escribe uno de esos números en binario (sin signo), solo necesita 24 bits. Cuando intenta escribir cualquiera de estos números en formato IEEE, el bit 1 más significativo no se usa porque estamos representando los números como normalizados. Entonces tenemos espacio para hasta 23 bits, para la mantisa. En particular, el número 111111111111111111111111 que tiene 24 unos se puede representar exactamente, esto es $2^{24} - 1$. A continuación, ¿qué pasa con $2^{24} = (1000000000000000000000000)_2$? ¿Se puede escribir exactamente? Sí, porque es solo $1{,}0 \times 2^{24}$, y se puede representar exactamente en el formato IEEE, es decir, el valor del exponente es 24 y las cifras de la mantisa son todas nulas. Por tanto, todos los números del 1 al 2^{24} se pueden representar exactamente. Ahora, ¿qué pasa con

$$2^{24} + 1 = (1000000000000000000000001)_2 = (1{,}000000000000000000000001)_2 \times 2^{24}?$$

¿Se puede representar exactamente? No, no se puede. ¿Por que no? Porque solo hay espacio para 23 bits en la mantisa y esos bits serían los 23 ceros a la derecha del punto binario, por lo que no podríamos entrar en el bit 2^{-24}. Por lo tanto, el n más grande tal que $1, \cdots, n$ se puede representar exactamente en formato IEEE de precisión simple es 2^{24}.

2. Calcule las siguientes operaciones de coma flotante para 32 bits (represente todos los números en sistema binario):

a) [2 *ptos.*]$1313{,}3125 + 0{,}1015625$

b) [2 *ptos.*]1313,3125 × 0,1015625

Solución:

a) Expresamos los números en sistema binario

$$1313{,}3125 = 1{,}01001000010101 \times 2^{10}$$
$$0{,}1015625 = 1{,}101 \times 2^{-4}$$

normalizamos al mismo exponente:

$$\begin{array}{r} 101001000010101{,}0 \times 2^{-4} \\ 1{,}1010 \times 2^{-4} \\ - - - - - - - - - - - - \\ 101001000010110{,}1010 \times 2^{-4} \end{array}$$

renormalizado: $1{,}010010000101101010 \times 2^{10}$
nueva mantisa: 010010000101101010
exponente: $10 + 127 = 137 = 10001001$
respuesta: 01000100101001000010110101000000

b) Expresamos los números en sistema binario

$$1313{,}3125 = 1{,}01001000010101 \times 2^{10}$$
$$0{,}1015625 = 1{,}101 \times 2^{-4}$$

Signo: 0
Exponente temporal: $10 + -4 = 6$
Nueva mantisa:

$$\begin{array}{r} 1{,}01001000010101 \\ 1{,}101 \\ - - - - - - - - - - - - \\ 10{,}00010101100010001 \end{array}$$

Mantisa ajustada: 1,00001010110001000 1
Exponente ajustado: $7 + 127 = 134 = 10000110$
Mantisa final: 00001010110001000100000
Respuesta: 0 10000110 00001010110001000100000

3. Justificando su respuesta, determine el valor de verdad de las siguientes proposiciones:

a) [1 *pto.*] Si la cantidad de elementos del conjunto $\mathbb{F}$, con $\mathbb{F}(2, t, -1, 2)$ es 33, entonces los dígitos en la mantisa son 4.

b) [1 *pto.*] En un computadora de doble precisión su sistema de números puntos flotantes está distribuido en 11 bits para la mantisa y 52 para el exponente.

c) [1 *pto.*] Si $\hat{x} = 3212.5$, es la aproximación de un número x en alguna máquina y $|EA(x)| < 0.5$, entonces $x \in \langle 3212, 3213\rangle$.

d) [1 *pto.*] Sean x, y y z números en un computador con longitud de palabra de 32 bits y $\beta = 2$, entonces $ER(x(yz)) \leq 3 \times 2^{-24}$.

Solución:

a) (Falso) De la fórmula $2(\beta-1)\beta^{t-1}(U-L+1)$ tenemos $t=3$.

b) (Falso) Se tiene 52 bits para la mantisa y 11 bits para el exponente

c) (Verdadero) $-0{,}5 < x - \hat{x} < 0.5$, de donde $x \in \langle 3212, 3213\rangle$.

d) (Verdadero) Sabemos que existen δ_1, δ_2 y δ_3 tales que $fl(y) = y(1+\delta_1)$, $fl(x) = x(1+\delta_2)$ y $fl(x(fl(y)fl(z))) = x(fl(y)fl(z))(1+\delta_3)$, tal que $|\delta_i| \leq 2^{-24}$ para cada $i = 1, 2, 3$, luego

$$\begin{aligned}
fl(x(fl(y)fl(z))) &= x(yz)(1+\delta_1)(1+\delta_2)(1+\delta_3)\\
&= x(yz)(1+\delta_1+\delta_2+\delta_3+\delta_1\delta_2+\delta_1\delta_3+\delta_2\delta_3+\delta_1\delta_2\delta_3)\\
&\approx x(yz)(1+\delta_1+\delta_2+\delta_3)\\
&= x(yz)(1+\delta)
\end{aligned}$$

donde $|\delta| = |\delta_1+\delta_2+\delta_3| \leq |\delta_1| + |\delta_2| + |\delta_3| \leq 3 \times 2^{-24}$.

4. Sea la sucesión definida por:

$$x_n = \frac{Sen\left(\frac{1}{n}\right)}{\frac{1}{n}},\ \forall n \geq 1.$$

a) [1 *pto.*] Determine la tabla de los 10 primeras iteraciones usando 10 decimales.

b) [1 *pto.*] Para $f(x) = Sen(x)$, determine su desarrollo usando la fórmula de Taylor en torno de $x = 0$ hasta su segundo orden.

c) [1 *pto.*] Determine la rapidez de convergencia de la sucesión usando (*b*).

d) [1 *pto.*] Usando (*c*) indique la nueva sucesión a la que equivale su convergencia.

Solución:

a) [1 *pto.*] La tabla es:

n	1	2	3	4	5
x_n	0,8414709848	0,9588510772	0,9815840904	0,9896158370	0,9933466540
n	6	7	8	9	10
x_n	0,9953767962	0,9966021085	0,9973978671	0,9979436566	0,9983341665

b) [1 *pto.*] Aplicando la fórmula de Taylor en torno de $x = 0$ es:

$$f(x) = f(0) + \frac{f'(0)}{1!}(x-0) + \frac{f''(0)}{2!}(x-0)^2 + \frac{f'''(\xi(x))}{3!}(x-0)^3$$

Reemplazando:

$$Sen(x) = x - Cos(\xi(x))\frac{x^3}{3!},\ \xi(x) \in]0, x[$$

c) [1 *pto.*] De *b*) al despejar:

$$\frac{Sen(x) - x}{x^3} \leq \frac{Cos(\xi(x))}{6},\ \forall x \in]0, x[;$$

Tomano valor absoluto m.a.m. tenemos:

$$\frac{\left|\frac{Sen(x)}{x} - 1\right|}{x^2} = \frac{|Cos(\xi(x))|}{6} \leq \frac{\max\limits_{[0,x]} |Cos(x)|}{6} \leq \frac{1}{6},\ \forall x \in]0, x[.$$

Haciendo $x = \dfrac{1}{n}$ y asociamos a este último resultado con la definición de rapidez de convergencia, tenemos:

$$\frac{\left|\dfrac{Sen\left(\dfrac{1}{n}\right)}{\dfrac{1}{n}} - 1\right|}{\dfrac{1}{n^2}} \leq \frac{1}{6}.$$

d) [1 *pto.*] De *c)* concluimos que:

$$\frac{Sen\left(\frac{1}{n}\right)}{\frac{1}{n}} = 1 + O\left(\frac{1}{n^2}\right).$$

Es decir, que la sucesión $\dfrac{sen\left(\frac{1}{n}\right)}{\frac{1}{n}}$ tiene una convergencia equivalente a $\dfrac{1}{n^2}$.