Profesor: Gonzalo Hernández.
Auxiliar: Gonzalo Ríos.
Fecha: 15 de Abril
Pauta Control 1

1. **Codificación binaria floating point tipo inicial**

   (a) Como el número real máximo representable en una codificacion binaria siempre es $\frac{2^m-1}{2^m} \times 2^{\left(2^e-1\right)}$, donde m es la cantidad de bits de la mantisa, y e la cantidad de bits del exponente, esto equivale a $2^{2^e-1} - 2^{\left(2^e-1\right)-m}$, queda un sistema:

   $$2^{63} - 2^{54} = 2^{2^e-1} - 2^{\left(2^e-1\right)-m} \Longrightarrow \begin{array}{l} 2^e - 1 = 63 \\ \left(2^e-1\right) - m = 54 \end{array} \Longrightarrow \begin{array}{l} e = 6 \\ m = 9 \end{array}$$

   Sumando el bit para el signo de la mantisa y el bit del signo del exponente, la codificación necesita $2 + 6 + 9 = 17$ bits.

   (b) **Como son 17 bits, entonces**

   i. Numero en codificacion binaria: $a = a_1a_2a_3...a_{17} = 11011101110111011$

   ii. Número real:

   A. Signo Mantisa: $-$

   B. Signo Exponente: $-$

   C. Exponente: $0 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 29$

   D. Mantisa: $1{\times}2^{-1}{+}1{\times}2^{-2}{+}0{\times}2^{-3}{+}1{\times}2^{-4}{+}1{\times}2^{-5}{+}1{\times}2^{-6}{+}0{\times}2^{-7}{+}1{\times}2^{-8}{+}1{\times}2^{-9} = \frac{443}{512} = 0.865\,234\,375$

   E. Numero: $a = -\frac{443}{512} \times 2^{-29} = -443 \times 2^{-38} = -1.\,611\,624\,611\,541\,628\,837\,6 \times 10^{-9}$

   (c) **Polinomio de Taylor**

   i. $p_4(x) = 1 + x^2 + \frac{1}{2}x^4$

   ii. Se debe calcular con la aritmetica finita de 5 cifras con redondeo, entonces $\frac{0.062\,5}{2} = 0.031\,25$, cumple la aritmética. Luego $p_4(0.5) = 1 + (0.5)^2 + \frac{1}{2}(0.5)^4 = 1 + 0.25 + \frac{0.062\,5}{2} = 1.25 + 0.031\,25 = 1.28125$, reduciendo a 5 cifras con redondeo, queda $p_4(0.5) = 1.2813$

   iii. En representacion de punto flotante quedan:

   A. $\boxed{f(0.5)}$ $1.\,284\,0 = 0.642 \times 2^1$

	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
	1	0	1	0	0	1	0	0	0
0.642	0.142		0.017			0.001375			

   $f(0.5) = 00000001101001000$

   B. $\boxed{p_4(0.5)}$ $1.2813 = 0.64065 \times 2^1$

	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
	1	0	1	0	0	1	0	0	0
0.64065	0.14065		0.015 65			0.000 025			

   $p_4(0.5) = 00000001101001000$

   C. $E_{rel} = \frac{|f(0.5)-p_4(0.5)|}{|f(0.5)|} = \frac{0.0027}{1.\,284\,0} = 2.\,102\,803\,738\,317\,757\,009\,3 \times 10^{-3}$

   La representación de $f(0.5)$ y de $p_4(0.5)$ en esta decodificación es exactamente la misma, por lo que quiere decir que la aproximación, en terminos del punto flotante, es muy buena.

2. **Serie, suma parcial exacta y suma parcial aproximada**

   (a) $E_{abs} = |S - Z_n| = |S - Z_n + S_n - S_n| = |S - S_n + S_n - Z_n| \leq |S - S_n| + |S_n - Z_n|$
   El primer término corresponde al error de truncacion que tiende a cero cuando $n \to \infty$ ya que la suma parcial tiende a la serir. El segundo término corresponde al error de redondeo, que aumenta cuando $n \to \infty$, ya que al aumentar el número de operaciones, se cometen más errores de redondeo.

   (b) $E_{abs} = |S - Z_n| \leq |S - S_n| + |S_n - Z_n| \leq \int_n^\infty \frac{1}{x^2}dx + C \cdot n = -\frac{1}{x}\Big|_n^\infty + C \cdot n = \frac{1}{n} + C \cdot n = g(n)$

(c) $\min g(n) = \frac{1}{n} + C \cdot n \iff \frac{dg}{dn}(\bar{n}) = 0 \wedge \frac{d^2g}{dn^2}(\bar{n}) > 0$

$\frac{dg}{dn} = -\frac{1}{n^2} + C = 0 \Longrightarrow \bar{n}_1 = \frac{1}{\sqrt{C}} \wedge \bar{n}_2 = -\frac{1}{\sqrt{C}}$

$\frac{d^2g}{dn^2} = 2\frac{1}{n^3} > 0 \Longrightarrow \boxed{\bar{n} = \frac{1}{\sqrt{C}}}$

$n = \frac{1}{\sqrt{4\times10^{-16}}} = 5 \times 10^7$

$E_{abs} \leq \frac{1}{n} + C \cdot n = \sqrt{C} + \frac{C}{\sqrt{C}} = 2\sqrt{C} = 2\sqrt{4 \times 10^{-16}} = 4 \times 10^{-8}$

Lo que implica que tendrá 7 dígitos decimales correctos.

(d) $\phi(x,y) = g(x) \times g(y) = \left(\frac{1}{x} + cx\right)\left(\frac{1}{y} + cy\right) = \frac{1}{xy} + c^2xy + c\left(\frac{x}{y} + \frac{y}{x}\right)$

$\varepsilon_\phi = \frac{x}{\phi(x,y)}\frac{\partial\phi}{\partial x}\varepsilon_x + \frac{y}{\phi(x,y)}\frac{\partial\phi}{\partial y}\varepsilon_y$

$\frac{\partial\phi}{\partial x} = \frac{-1}{x^2y} + c^2y + c\left(\frac{1}{y} - \frac{y}{x^2}\right)$

$x\frac{\partial\phi}{\partial x} = c^2xy + c\left(\frac{x}{y} - \frac{y}{x}\right) - \frac{1}{xy} = \frac{c^2x^2y^2 + c\left(x^2 - y^2\right) - 1}{xy}$

$\phi(x,y) = \frac{c^2x^2y^2 + c\left(x^2+y^2\right)+1}{xy}$

$\frac{x}{\phi(x,y)}\frac{\partial\phi}{\partial x} = \frac{c^2x^2y^2 + c\left(x^2-y^2\right)-1}{c^2x^2y^2 + c(x^2+y^2)+1} = \frac{c^2x^2y^2 + c\left(x^2-y^2\right)-1+(2-2)+(2y^2-2y^2)}{c^2x^2y^2 + c(x^2+y^2)+1} = \frac{c^2x^2y^2 + c\left(x^2+y^2\right)+1-2-2y^2}{c^2x^2y^2 + c(x^2+y^2)+1}$

$= 1 - \frac{2+2y^2}{c^2x^2y^2 + c(x^2+y^2)+1} = 1 - 2\frac{1+y^2}{c^2x^2y^2 + c(x^2+y^2)+1} = 1 - \frac{2}{c^2x^2y^2 + c(x^2+y^2)+1} - 2\frac{y^2}{c^2x^2y^2 + c(x^2+y^2)+1} =$

$1 - \frac{2}{c^2x^2y^2 + c(x^2+y^2)+1} - 2\frac{1}{c^2x^2 + c\left(\frac{x^2+y^2}{y^2}\right) + \frac{1}{y^2}}$

entonces $\frac{2}{c^2x^2y^2 + c(x^2+y^2)+1}$ es acotada, ya que $c > 0$ lo que implica que $c^2x^2y^2 + c\left(x^2+y^2\right)+1 > 0 \forall x, y$ Ahora si $\frac{1}{c^2x^2 + c\left(\frac{x^2+y^2}{y^2}\right) + \frac{1}{y^2}}$ no fuera acotada, entonces $\exists x, y$ tal que $c^2x^2 + c\left(\frac{x^2+y^2}{y^2}\right) + \frac{1}{y^2} = 0$, pero $c^2x^2 \geq 0$, $c\left(\frac{x^2+y^2}{y^2}\right) \geq 0$ y $\frac{1}{y^2} \geq 0$, entonces la única opcion sería que las tres fueran igual a cero, y eligiendo $x = 0$ para que se cumpla la primera igualdad, la segunda se convierte en c, que por enunciado es > 0. Esto implica que $\frac{1}{c^2x^2 + c\left(\frac{x^2+y^2}{y^2}\right) + \frac{1}{y^2}}$ es acotada, y la suma de expresiones acotadas es acotada,luego se puede acotar $\frac{x}{\phi(x,y)}\frac{\partial\phi}{\partial x}$. De forma análoga se acota $\frac{y}{\phi(x,y)}\frac{\partial\phi}{\partial y}$.

3. **SEL**

(a) **Métodos Directos**

i. $A^0 = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 4 \end{bmatrix}$

$A^1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{2}{4} & 1 & 0 \\ -\frac{0}{4} & 0 & 1 \end{bmatrix}\begin{bmatrix} 4 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 2 & 4 \end{bmatrix}$

$A^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{bmatrix}\begin{bmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & \frac{8}{3} \end{bmatrix}$

$U = \begin{bmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & \frac{8}{3} \end{bmatrix} \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{3} & 1 \end{bmatrix}$

Definiendo $U\vec{x} = \vec{y}$ el sistema queda

$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{3} & 1 \end{bmatrix}\vec{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \Longrightarrow \vec{y} = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{2}{3} \end{bmatrix}$

Resolviendo $U\vec{x} = \vec{y}$

$\begin{bmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & \frac{8}{3} \end{bmatrix}\vec{x} = \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{2}{3} \end{bmatrix} \Longrightarrow \vec{x} = \begin{bmatrix} \frac{1}{4} \\ 0 \\ \frac{1}{4} \end{bmatrix}$

ii. Es definida positiva si sus subdeterminantes son positivas

$|4| > 0$

$\begin{vmatrix} 4 & 2 \\ 0 & 3 \end{vmatrix} = 12 > 0$

$\begin{vmatrix} 4 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & \frac{8}{3} \end{vmatrix} = 32 > 0 \Longrightarrow$ es definida positiva

Entonces su factorizacion de Cholesky es

$$\begin{bmatrix}4&2&0\\2&4&2\\0&2&4\end{bmatrix}=\begin{bmatrix}2&0&0\\1&\sqrt{3}&0\\0&\frac{2}{3}\sqrt{3}&\frac{2}{3}\sqrt{2}\sqrt{3}\end{bmatrix}\begin{bmatrix}2&1&0\\0&\sqrt{3}&\frac{2}{3}\sqrt{3}\\0&0&\frac{2}{3}\sqrt{2}\sqrt{3}\end{bmatrix}$$

(b) **Métodos Iterativos**

i. Gauss-Seidel

$$M=-(D+L)^{-1}U=-\begin{bmatrix}4&0&0\\2&4&0\\0&2&4\end{bmatrix}^{-1}\begin{bmatrix}0&2&0\\0&0&2\\0&0&0\end{bmatrix}=\begin{bmatrix}0&-\frac{1}{2}&0\\0&\frac{1}{4}&-\frac{1}{2}\\0&-\frac{1}{8}&\frac{1}{4}\end{bmatrix}$$

$$\vec{N}=(D+L)^{-1}\vec{b}=\begin{bmatrix}4&0&0\\2&4&0\\0&2&4\end{bmatrix}^{-1}\begin{bmatrix}1\\1\\1\end{bmatrix}=\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}$$

$$\vec{x}^{k+1}=M\vec{x}^{k}+\vec{N}=\begin{bmatrix}0&-\frac{1}{2}&0\\0&\frac{1}{4}&-\frac{1}{2}\\0&-\frac{1}{8}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}x_1^k\\x_2^k\\x_3^k\end{bmatrix}+\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}=\begin{bmatrix}-\frac{1}{2}x_2^k+\frac{1}{4}\\\frac{1}{4}x_2^k-\frac{1}{2}x_3^k+\frac{1}{8}\\-\frac{1}{8}x_2^k+\frac{1}{4}x_3^k+\frac{3}{16}\end{bmatrix}$$

A. $\vec{x}^0=\begin{bmatrix}0\\0\\0\end{bmatrix}$

B. $\vec{x}^1=\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}$

C. $\vec{x}^2=\begin{bmatrix}\frac{3}{16}\\\frac{1}{16}\\\frac{7}{32}\end{bmatrix}$

D. $\vec{x}^3=\begin{bmatrix}\frac{7}{32}\\\frac{1}{32}\\\frac{15}{64}\end{bmatrix}$

E. $\vec{x}^4=\begin{bmatrix}\frac{15}{64}\\\frac{1}{64}\\\frac{31}{128}\end{bmatrix}=\begin{bmatrix}0.234\,375\\0.015\,625\\0.242\,187\,5\end{bmatrix}$

F. $\vec{x}^5=\begin{bmatrix}\frac{31}{128}\\\frac{1}{128}\\\frac{63}{256}\end{bmatrix}=\begin{bmatrix}0.242\,187\,5\\0.007\,812\,5\\0.246\,093\,75\end{bmatrix}$

Solo 4 iteraciones eran necesarias, es decir, hasta $\vec{x}^4$.

ii. $\rho(T_G)=\max\limits_{i=1\ldots n}|\lambda_i|$ donde λ_i es un valor propio de T_G

$\det(T_G-I\lambda)=0$

$$\begin{vmatrix}-\lambda&-\frac{1}{2}&0\\0&\frac{1}{4}-\lambda&-\frac{1}{2}\\0&-\frac{1}{8}&\frac{1}{4}-\lambda\end{vmatrix}=\tfrac{1}{2}\lambda^2-\lambda^3=0\Longrightarrow\lambda_1=\lambda_2=0\wedge\lambda_3=\tfrac{1}{2}\Longrightarrow\rho(T_G)=\tfrac{1}{2}$$

$\overline{\omega}=\frac{2}{1+\sqrt{1-\rho(T_G)}}=\frac{2}{1+\sqrt{1-\frac{1}{2}}}=\frac{2}{\frac{1}{2}\sqrt{2}+1}=4-2\sqrt{2}=2\left(2-\sqrt{2}\right)=1.171\,572\,875\,253\,809\,902\,4$

iii. Sor

$$\vec{x}^{k+1}=(1-\overline{\omega})\,\vec{x}^{k}+\overline{\omega}\left(M\vec{x}^{k}+\vec{N}\right)=(1-1.171\,5)\begin{bmatrix}x_1^k\\x_2^k\\x_3^k\end{bmatrix}+1.171\,5\left[\begin{bmatrix}0&-\frac{1}{2}&0\\0&\frac{1}{4}&-\frac{1}{2}\\0&-\frac{1}{8}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}x_1^k\\x_2^k\\x_3^k\end{bmatrix}+\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}\right]=$$

$$-0.1715\begin{bmatrix}x_1^k\\x_2^k\\x_3^k\end{bmatrix}+1.171\,5\begin{bmatrix}-\frac{1}{2}x_2^k+\frac{1}{4}\\\frac{1}{4}x_2^k-\frac{1}{2}x_3^k+\frac{1}{8}\\-\frac{1}{8}x_2^k+\frac{1}{4}x_3^k+\frac{3}{16}\end{bmatrix}=\begin{bmatrix}-0.171\,5x_1^k-0.585\,75x_2^k+0.292\,875\\0.121\,375\,x_2^k-0.585\,75x_3^k+0.146\,437\,5\\-0.146\,437\,5x_2^k+0.121\,375\,x_3^k+0.219\,656\,25\end{bmatrix}$$

A. $\vec{x}^0=\begin{bmatrix}0\\0\\0\end{bmatrix}$

B. $\vec{x}^1=(1-1.171\,5)\begin{bmatrix}0\\0\\0\end{bmatrix}+1.171\,5\left[\begin{bmatrix}0&-\frac{1}{2}&0\\0&\frac{1}{4}&-\frac{1}{2}\\0&-\frac{1}{8}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}0\\0\\0\end{bmatrix}+\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}\right]=1.171\,5\begin{bmatrix}\frac{1}{4}\\\frac{1}{8}\\\frac{3}{16}\end{bmatrix}=$

$$\begin{bmatrix}0.292\,875\\0.146\,437\,5\\0.219\,656\,25\end{bmatrix}$$

C. $\vec{x}^2 = (1 - 1.171\,5) \begin{bmatrix} 0.292\,875 \\ 0.146\,437\,5 \\ 0.219\,656\,25 \end{bmatrix} + 1.171\,5 \left[\begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{4} & -\frac{1}{2} \\ 0 & -\frac{1}{8} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0.292\,875 \\ 0.146\,437\,5 \\ 0.219\,656\,25 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} \\ \frac{1}{8} \\ \frac{3}{16} \end{bmatrix} \right] =$

$$\begin{bmatrix} 0.156\,871\,171\,875 \\ 0.035\,547\,703\,125 \\ 0.224\,873\,085\,937\,5 \end{bmatrix}$$

D. $\vec{x}^3 = (1 - 1.171\,5) \begin{bmatrix} 0.156\,871\,171\,875 \\ 0.035\,547\,703\,125 \\ 0.224\,873\,085\,937\,5 \end{bmatrix} + 1.171\,5 \left[\begin{bmatrix} 0 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{4} & -\frac{1}{2} \\ 0 & -\frac{1}{8} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0.156\,871\,171\,875 \\ 0.035\,547\,703\,125 \\ 0.224\,873\,085\,937\,5 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} \\ \frac{1}{8} \\ \frac{3}{16} \end{bmatrix} \right] =$

$$\begin{bmatrix} 0.245\,149\,526\,917\,968\,75 \\ 0.019\,032\,692\,378\,906\,25 \\ 0.241\,744\,704\,029\,296\,875 \end{bmatrix}$$

Hasta acá es necesario, es decir, 3 iteraciones