Computación Numérica

Primer Parcial - Abril 2021

1. $\boxed{\text{}}$ Una máquina almacena números en punto flotante en base 2 en 15 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los nueve siguientes para el exponente sesgado y los últimos cinco bits para la mantisa.

   (a) Calcular los exponentes máximo y mínimo.

   (b) Calcular el valor mínimo desnormalizado. Expresarlo en binario y en decimal. ¿Qué precisión tendría?

   (c) Calcular el número 263 en este formato. Redondear al par más cercano. ¿Qué error cometemos al redondearlo?

(a) El número de enteros que podriamos representar con m bits sería

$$2^m = 2^9 = 512$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \dots, , 510, 511]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \dots, 510, R]$$

Pero los enteros representados serían los anteriores menos el sesgo $= 2^{m-1} - 1 = 2^{9-1} - 1 = 2^8 - 1 = 255$, es decir, el rango de números a representar sería

$$[R, 1 - 255, \dots, 510 - 255, R] = [R, -254, \dots, 255, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -254 \ \text{ y } \ e_{max} = 255}$$

(b) Los números desnormalizados tienen 0 como bit escondido y el exponente es el mínimo de los normalizados. Sabemos que un número es desnormalizado porque todos los bits de su exponente son 0.

El valor mínimo desnormalizado tiene mantisa mínima y se representa en binario

signo	exponente	mantisa
0	0000 00000	00001

que se corresponde con

$$0.00001 \times 2^{-254} \longrightarrow 2^{-5} \times 2^{-254} = 2^{-5-254} = 2^{-259} \approx 1.08 \times 10^{-78}$$

Solución:

$$\boxed{\text{Mínimo número desnormalizado: } 1.08 \times 10^{-78}}$$

La precisión del numero mínimo es 1 porque los ceros a la izquierda no cuentan a efecto de precisión.

Solución:

$$\boxed{p_{min} = 1}$$

(c) Calcular el número 263 en este formato.

MANTISA:

Cociente	263	131	65	32	16	8	4	2	1	
Resto	1	1	1	0	0	0	0	0	1	$\leftarrow$

Por lo que tenemos

$$(271)_{10} = (100000111)_2 \longrightarrow 1.00000\,111 \times 2^8$$

Si truncamos, el número sería 1.00000. Pero como después tenemos tres dígitos que son 1, está más cerca del número siguiente, que redondeando al par más cercano sería 1.00001 siendo el uno a la izquierda de la coma el bit escondido.

EXPONENTE: Como sesgo = 255 representaremos

$$8 + 255 = 263$$

Cociente	263	131	65	32	16	8	4	2	1	
Resto	1	1	1	0	0	0	0	0	1	$\leftarrow$

y el exponente en binario es $(10000111)_2$ y la representación completa será

signo	exponente	mantisa
0	1000 00111	00001

Hemos representado 263 con

$$1.00001 \times 2^8$$

por lo que el error es

$$\left|263 - (1 + 2^{-5}) \times 2^8\right| = |263 - 264| = 1$$

2. Sea una función f que cumple las condiciones del teorema de Bolzano en el intervalo $[0.5, 2]$ y tiene una raíz en dicho intervalo. ¿Cuántos pasos se necesitan para aproximar la raíz con un error menor que 10^{-7} usando bisección partiendo del intervalo $[0.5, 2]$?

Si α es la raíz obtenida por el método de bisección con intervalo inicial $[a, b]$, el error en la iteración n es $e_n = |\alpha - x_n|$ está acotado por la longitud del intervalo en la iteración n. Es decir

$$e = |\alpha - m_n| < \frac{b-a}{2^n}.$$

Buscamos que $e < 10^{-7}$. Una condición suficiente es que se cumpla

$$\frac{b-a}{2^n} < 10^{-7}.$$

Trabajaremos con esta desigualdad y aplicaremos las siguientes propiedades:

(a) Si $a < b$ y $c > 0 \Longrightarrow a\,c < b\,c$

(b) Si f es una función estrictamente creciente se tiene que

$$x < y \Longrightarrow f(x) < f(y)$$

(c) $\log A^B = B \log A$ (log es un logaritmo en cualquier base)

Reescribimos la desigualdad utilizando los datos.

$$\frac{b-a}{2^n} < 10^{-7} \Longleftrightarrow \frac{2-0.5}{2^n} < 10^{-7} \Longleftrightarrow \frac{1.5}{2^n} < 10^{-7}$$

Teniendo en cuenta la propiedad (a) y multiplicando ambos miembros de la desigualdad primero por 2^n y luego 10^7 tenemos que

$$\frac{1.5}{2^n} < 10^{-7} \Longleftrightarrow \frac{1.5}{2^n} \times 2^n \times 10^7 < 10^{-7} \times 10^7 \times 2^n \Longleftrightarrow 1.5 \times 10^7 < 2^n$$

Como $f(x) = \ln(x)$ es una función estrictamente creciente, aplicando la propiedad (b) tenemos que

$$\ln\left(1.5 \times 10^7\right) < \ln\left(2^n\right)$$

y teniendo en cuenta las propiedad (c)

$$\ln\left(1.5 \times 10^7\right) < n \ln 2.$$

Como $\ln 2 > 0$, aplicando la propiedad (a) con $c = 1/\ln 2$

$$\frac{\ln\left(1.5 \times 10^7\right)}{\ln 2} < n$$

$$23.82 < n$$

Solución:

$\boxed{\text{Si hacemos } n = 24 \text{ iteraciones garantizamos que el error es menor que } 10^{-7}.}$

3. Dados los puntos

x_k	1.3	1.4	1.5	1.6	1.7
$y_k = \ln x_k$	0.2624	0.3365	0.4055	0.4700	0.5306

aproximar el valor en el punto $x = 1.68$ utilizando interpolación lineal a trozos. Calcular una cota del error. Operar con cuatro cifras decimales.

La interpolación lineal a trozos consiste en unir los nodos con segmentos de recta y aproximar el valor de la función con su valor sobre estas rectas. Queremos aproximar la función en $x = 1.68$. Primero tenemos que escoger los nodos que usaremos para nuestro segmento de recta. Tomaremos los dos nodos más próximos a este valor que son 1.6 y 1.7.

$$x_0 = 1.6 \quad x_1 = 1.7$$

x_k	1.6	1.7
$y_k = \ln x_k$	0.4700	0.5306

Forma de Lagrange. Podemos interpolar utilizando la forma de Lagrange y entonces el polinomio de interpolación lineal en el intervalo $[1.6, 1.7]$ tiene la forma

$$P_1(x) = y_0\, L_0(x) + y_1\, L_1(x)$$

$$P_1(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0}$$

$$P_1(x) = (0.4700) \frac{x - 1.7}{1.6 - 1.7} + (0.5306) \frac{x - 1.6}{1.7 - 1.6}$$

Y el valor en $x = 1.68$ es

$$P_1(1.68) = (0.4700) \frac{1.68 - 1.7}{1.6 - 1.7} + (0.5306) \frac{1.68 - 1.6}{1.7 - 1.6} = 0.5185$$

Forma de Newton. La tabla de diferencias divididas es

$$\begin{array}{ccc}
x & y & \\
1.6 & \boxed{0.4700}^{\,c_0} & \\
 & & \dfrac{0.5306 - 0.4700}{1.7 - 1.6} = \boxed{0.6060}^{\,c_1} \\
1.7 & 0.5306 &
\end{array}$$

Y el polinomio de interpolación en la forma de Newton

$$P_1\,(x) = c_0 + c_1\,(x - x_0)$$

Y el valor en $x = 1.68$ es

$$P_1\,(1.68) = 0.4700 + 0.6060\,(1.68 - 1.6) = 0.5185$$

Solución:

$$\boxed{P_1\,(1.68) = 0.5185}$$

El error de interpolación viene dado por la fórmula

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x-x_0)\dots(x-x_n)}{(n+1)!} \quad c \in (x_0, x_n)$$

Para dos nodos, esta fórmula es

$$E(x) = f(x) - P_1(x) = f''(c)\frac{(x-x_0)(x-x_1)}{2!} \quad c \in (x_0, x_1)$$

El valor c es desconocido, aunque sabemos que está en el intervalo de interpolación, en nuestro caso $c \in (1.6, 1.7)$, y tenemos que encontrar una cota para ese valor. Si $f(x) = \ln x$ se tiene

$$f'(x) = \frac{1}{x} \qquad f''(x) = \frac{-1}{x^2} \qquad |f''(x)| = \frac{1}{x^2}$$

y como $\dfrac{1}{x^2}$ es decreciente, si $x \in (a, b)$ se tiene que

$$|f''(x)| < \frac{1}{a^2}$$

Y como $c \in (1.6, 1.7)$ en nuestro caso

$$|f''(c)| < \frac{1}{1.6^2} = 0.3906$$

Y por lo tanto podemos dar como cota del error

$$|E(1.68)| < 0.3906\frac{|(1.68-1.6)(1.68-1.7)|}{2!} = 0.0003$$

Solución:

$$\boxed{|E(1.68)| < 0.0003}$$

4. En una reacción química, la cantidad de producto de la reacción evoluciona con el tiempo de acuerdo con la fórmula

$$Q\,(t) = 10(1 - c\,e^{Kt}) \quad \text{con} \quad c > 0$$

Se han reunido los siguientes datos

t	0	2	4
Q	2	7	9

Calcular el valor de c y K utilizando el criterio de los mínimos cuadrados.

Vamos a linealizar la función a ajustar.

$$Q = 10\left(1 - c\,e^{Kt}\right) \quad \Longrightarrow \quad \frac{Q}{10} = 1 - c\,e^{Kt} \quad \Longrightarrow \quad c\,e^{Kt} = 1 - \frac{Q}{10} \quad \Longrightarrow$$

$$\Longrightarrow \quad 1 - \frac{Q}{10} = c\,e^{Kt} \quad \Longrightarrow \quad \ln\left(1 - \frac{Q}{10}\right) = \ln\left(c\,e^{Kt}\right) \quad \Longrightarrow$$

$$\Longrightarrow \quad \ln\frac{10-Q}{10} = \ln c + \ln e^{Kt} \quad \Longrightarrow \quad \ln\frac{10-Q}{10} = \ln c + K\,t\ln e \quad \Longrightarrow$$

$$\ln\frac{10-Q}{10} = \ln c + K\,t$$

Y si llamamos

$$y_k = \ln\frac{10-Q_k}{10}, \quad x_k = t_k, \quad a_0 = \ln c, \quad a_1 = k$$

tenemos

$$\ln\frac{10-Q_k}{10} \approx \ln c + K\,t_k \quad \Longrightarrow \quad y_k \approx a_0 + a_1\,x_k$$

el problema es ahora ajustar una recta de regresión mínimo cuadrática

$$P_1(x) = a_0 + a_1\,x$$

a los datos transformados (x_k, y_k), $\quad k = 1, \ldots, 3$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^{3} 1 & \sum_{k=1}^{3} x_k \\ \sum_{k=1}^{3} x_k & \sum_{k=1}^{3} x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{3} y_k \\ \sum_{k=1}^{3} x_k y_k \end{pmatrix}$$

Si calculamos los elementos de estas matrices

	$x_k = t_k$	Q_k	$y_k = \ln((10-Q_k)/10)$	x_k^2	$x_k\,y_k$
	0	2	-0.2231	0	0.
	2	7	-1.2040	4	-2.4078
	4	9	-2.3026	16	-9.2104
$\sum$	6		-3.7297	20	-11.61842

Sustituyendo los datos y operando

$$\begin{array}{rcrcr} 3a_0 & + & 6a_1 & = & -3.7297 \\ 6a_0 & + & 20a_1 & = & -11.6184 \end{array}$$

Resolvemos el sistema por Gauss: la segunda ecuación $e_2 \to e_2 - 2e_1$

$$\begin{array}{rcrcr} 3a_0 & + & 6a_1 & = & -3.7297 \\ & & 8a_1 & = & -4.1590 \end{array}$$

y por sutitución reversiva

$$\begin{array}{rcccr} a_1 & = & -4.1590/8 & = & -0.5199 \\ a_0 & = & (-3.7297 - 6a_1)/3 & = & -0.2034 \end{array}$$

Como

$$a_0 = \ln c \quad a_1 = K \quad \Longrightarrow c = e^{a_0} \approx 0.82 \quad K = a_1 \approx -0.52$$

La curva ajustada es

$$\boxed{P\,(t) = 10\,(1 - 0.82\,e^{-0.52t})}$$