Computación Numérica

Primer Parcial - Mayo 2021

1. Una máquina almacena números en punto flotante en base 2 en 13 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los seis siguientes para el exponente sesgado y los últimos seis bits para la mantisa.

   a) Calcular los exponentes máximo y mínimo y dar su valor en base 10.

   b) Calcular el valor normalizado positivo mínimo. Expresarlo en binario (siguiendo la norma) y en decimal. ¿Qué precisión tendría?

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

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

$$2^m = 2^6 = 64$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \ldots, 62, 63]$$

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

$$[R, 1, \ldots, 62, R]$$

Pero los enteros representados serían los anteriores menos el sesgo $= 2^{m-1} - 1 = 2^{6-1} - 1 = 2^5 - 1 = 31$, es decir, el rango de números a representar sería

$$[R, 1 - 31, \ldots, 62 - 31, R] = [R, -30, \ldots, 31, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -30 \text{ y } e_{max} = 31}$$

(b) El valor mínimo normalizado tiene exponente mínimo `000001` (el valor `000000` está reservado) y mantisa mínima y se representa en binario

signo	exponente	mantisa
0	000 001	000 000

Teniendo en cuenta que el bit escondido es 1, se corresponde con

$$1{,}000000 \times 2^{-30} \longrightarrow 2^{-30} \approx 9{,}31 \times 10^{-10}$$

Solución:

$$\boxed{\text{Mínimo número normalizado: } 9{,}31 \times 10^{-10}}$$

La precisión del numero mínimo es 7 porque hay que contar los números de la mantisa que son los representados (6) más el bit escondido (1).

Solución:

$$\boxed{p_{min} = 7}$$

(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

$$(263)_{10} = (100000111)_2 \longrightarrow 1{,}000001\,11 \times 2^8$$

Si truncamos, el número sería 1,000001. Pero como después tenemos dos dígitos más que son 1, está más cerca del número siguiente, que redondeando al par más cercano sería 1,000010 siendo el uno a la izquierda de la coma el bit escondido.

EXPONENTE: Como sesgo = 31 representaremos

$$8 + 31 = 39$$

Cociente	39	19	9	4	2	1	
Resto	1	1	1	0	0	1	$\leftarrow$

y el exponente en binario es $(100111)_2$ y la representación completa será

signo	exponente	mantisa
0	100 111	000 010

Hemos representado 263 con

$$1{,}000\,010 \times 2^8$$

por lo que el error es

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

2. Calcular $\sqrt[3]{3}$ usando únicamente sumas, restas, multiplicaciones y divisiones:

   a) Hacer tres iteraciones por el método de bisección partiendo del intervalo $[1, 2]$

   b) Hacer dos iteraciones por el método de Newton con $x_0 = 1$.

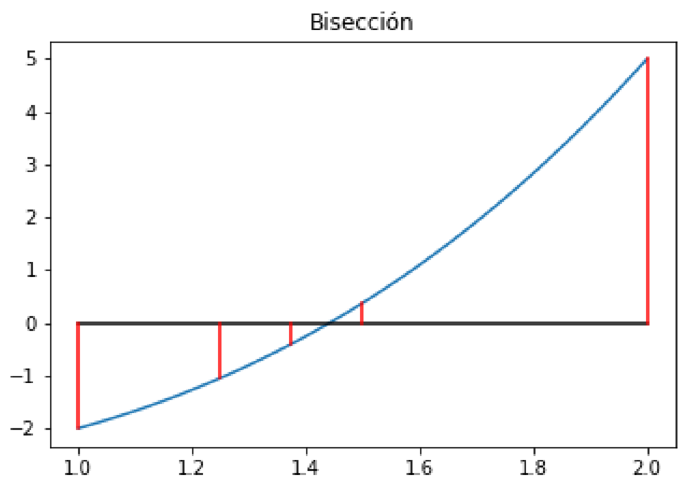
   c) Hacer dos iteraciones por el método de la secante con $x_0 = 2$ y $x_1 = 1$.

Queremos resolver la ecuación

$$x = \sqrt[3]{3} \Longleftrightarrow x^3 = 3 \Longleftrightarrow x^3 - 3 = 0.$$

Y la ecuación a resolver es $f(x) = 0$ donde $f(x) = x^3 - 3$.
(a) Bisección. 3 iteraciones.

k	a	c	b	$f(a)$	$f(c)$	$f(b)$	cota de error
1	1	$(1+2)/2 = 1{,}5$	2	-2	0,375	5	0,5
2	1	$(1+1{,}5)/2 = 1{,}25$	1,5	-2	$-1{,}05$	0,375	0,25
3	1,25	$(1{,}25+1{,}5)/2 = \boxed{1.375}$	1.5	$-1{,}05$		0,375	0,125



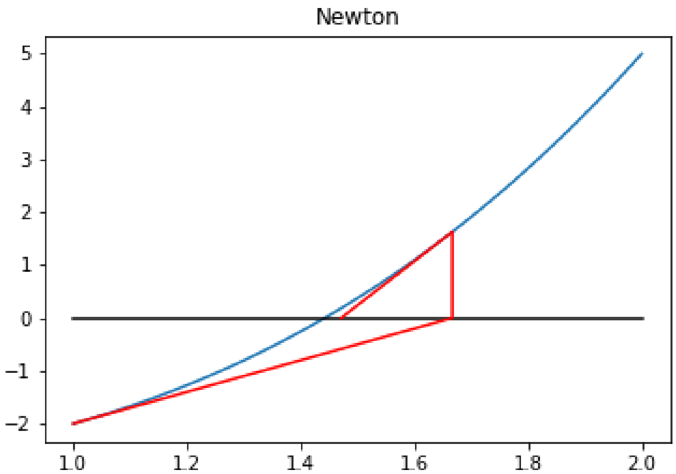
(b) Newton. 2 iteraciones.

Como $f(x) = x^3 - 3$ se tiene que $f'(x) = 3x^2$ y para iterar usaremos

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^3 - 3}{3x_k^2}$$

$$x_1 = x_0 - \frac{x_0^3 - 3}{3x_0^2} = 1 - \frac{1-3}{3(1)} = 1 + \frac{2}{3} = \frac{5}{3} = 1{,}667$$

$$x_2 = x_1 - \frac{x_2^3 - 3}{3x_2^2} = 5/3 - \frac{(5/3)^3 - 3}{3(5/3)^2} = 1{,}471$$



(c) Secante. 2 iteraciones.

La fórmula para iterar por la secante es

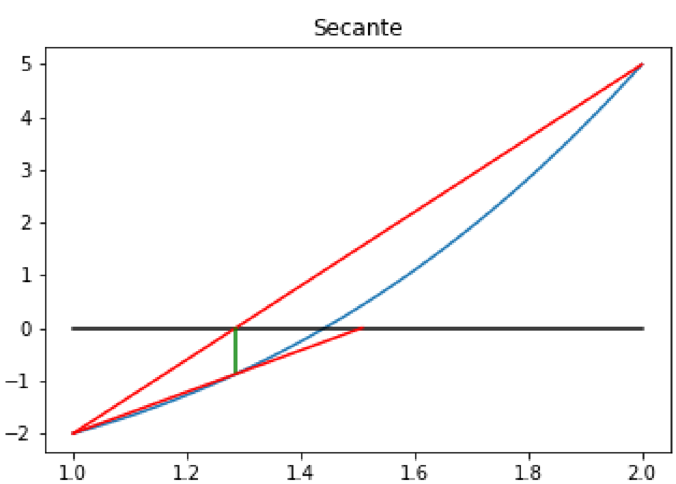
$$x_{k+2} = x_{k+1} - f\left(x_{k+1}\right) \frac{x_{k+1} - x_k}{f\left(x_{k+1}\right) - f\left(x_k\right)}$$

Por lo tanto, si $x_1 = 1$ y $x_0 = 2$

$$x_2 = x_1 - f\left(x_1\right) \frac{x_1 - x_0}{f\left(x_1\right) - f\left(x_0\right)} = 1 - f\left(1\right) \frac{1 - 2}{f\left(1\right) - f\left(2\right)} = 1{,}286$$

Si $x_2 = 1{,}286$ y $x_1 = 1$

$$x_3 = x_2 - f\left(x_2\right) \frac{x_2 - x_1}{f\left(x_2\right) - f\left(x_1\right)} = 1{,}286 - f\left(1{,}286\right) \frac{1{,}286 - 1}{f\left(1{,}286\right) - f\left(1\right)} = 1{,}508$$



3. Un isótopo radiacivo se descompone de acuerdo con la ecuación

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

Se han reunido los siguientes datos

t	0	2	9
Q	10	6	1

Calcular el valor de c y K utilizando el criterio de los mínimos cuadrados. Operar con tres cifras decimales.

Así que vamos a linealizar la función a ajustar. Si tenemos en cuenta que

$$\ln(AB) = \ln A + \ln B \quad \ln A^B = B \ln A \quad \ln e = 1$$

entonces

$$Q = c\,e^{Kt} \quad \Longrightarrow \quad \ln Q = \ln(c\,e^{Kt}) \quad \Longrightarrow \quad \ln Q = \ln c + \ln(e^{Kt}) \quad \Longrightarrow$$

$$\Longrightarrow \quad \ln Q = \ln c + Kt\,\ln e \quad \Longrightarrow \quad \ln Q = \ln c + K\,t$$

Y si llamamos

$$y_k = \ln Q_k, \quad x_k = t_k, \quad a_0 = \ln c, \quad a_1 = K$$

tenemos

$$\ln Q_k \approx \ln c + K\,t_k \quad \Longrightarrow \quad y_k \approx a_0 + a_1\,x_k$$

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

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

Y haciendo las operaciones

	$x_k = t_k$	Q_k	$y_k = \ln Q_k$	x_k^2	$x_k\,y_k$	
	0	10	2,303	1	0	
	2	6	1,792	4	3,584	
	9	1	0	81	0	
$\sum$	11		4,095	86	3,584	

Sustituyendo los datos y operando

$$\begin{array}{rcrcl} 3\,a_0 & + & 11\,a_1 & = & 4{,}095 \\ 11\,a_0 & + & 86\,a_1 & = & 3{,}584 \end{array}$$

Y la solución de este sistema es $a_0 = -0{,}255 \quad a_1 = 2{,}30$. Como

$$a_0 = \ln c, \quad a_1 = K \quad \Longrightarrow \quad c = e^{\,a_0} \approx 10 \quad K = a_1 \approx 2{,}3$$

La curva ajustada es

$$Q(t) = 10\,e^{-2{,}3t}$$