Computación Numérica

Primer Parcial - Marzo 2019

1. Una máquina almacena números en punto flotante en base 2 en 10 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 tres bits para la mantisa.

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

   (b) Escribir **todos** los números **des**normalizados positivos en este sistema en binario (siguiendo la norma). Dar el valor en decimal de el máximo y el mínimo número desnormalizados.

   (c) Calcular en base 10 el valor del número representado en este formato:

signo	exponente	mantisa
1	111110	000

(a) El número de exponentes que podemos representar con m bits sería

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

que son

$$[000000, 000001, 000010, \ldots, 111101, 111110, 111111]$$

Si no tenemos en cuenta el signo son

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

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)

Num(bin)		Num(dec)
0 000000 001	0.001×2^{-30}	1.16×10^{-10}
0 000000 010	0.010×2^{-30}	
0 000000 011	0.011×2^{-30}	
0 000000 100	0.100×2^{-30}	
0 000000 101	0.101×2^{-30}	
0 000000 110	0.110×2^{-30}	
0 000000 111	0.111×2^{-30}	8.15×10^{-10}

El valor mínimo se corresponde con el valor

$$0.001 \times 2^{-30} \longrightarrow 2^{-3} \times 2^{-30} = 2^{-33} \approx 1.16 \times 10^{-10}$$

El máximo se corresponde con el valor

$$0.111 \times 2^{-30} \longrightarrow \left(2^{-1} + 2^{-2} + 2^{-3}\right) \times 2^{-30} \approx 8.15 \times 10^{-10}$$

Solución:

Mínimo número desnormalizado: 1.16×10^{-10}
Máximo número desnormalizado: 8.15×10^{-10}

(c) El número representado en este estandar como

signo	exponente	mantisa
1	111110	000

tiene exponente máximo (111111 está reservado), mantisa mínima y signo negativo. Por lo tanto, el número es

$$-1.000 \times 2^{31}$$

que expresado en decimal es

$$-2^{31} = -2147483648$$

Solución:

-2147483648

2. Sea la ecuación

$$x^3 - 4x + 1 = 0$$

(a) Demostrar que en $[0, 1]$ existe una única raíz.

(b) ¿Se puede calcular por el método de régula-falsi partiendo de dicho intervalo?

(c) Aproximar la raíz haciendo dos iteraciones con el método de regula-falsi en dicho intervalo.

(d) Dar una cota del error cometido al calcular esta raíz. ¿Es una buena cota?¿Por qué?

(a)

Sea

$$f(x) = x^3 - 4x + 1$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[0, 1]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio, que es una función continua.
2. f tiene distinto signo en los extremos del intervalo:

$$f(0) = 1 \quad \text{y} \quad \text{f}(1) = -2$$

3. f es estrictamente creciente o decreciente en $[0, 1]$. Es decir $f' > 0$ o $f' < 0$ en $(0, 1)$:

$$f'(x) = 3x^2 - 4.$$

Teniendo en cuenta que

$$3x^2 - 4 = 0 \Longrightarrow x_{1,2} = \pm\sqrt{\frac{4}{3}} \Longrightarrow x_1 = -1.15 \quad x_2 = 1.15$$

podemos factorizar el numerador

$$f'(x) = 3\,(x - 1.15)\,(x + 1.15) = (+)\,(-)\,(+) < 0$$

para todos los $x \in (0, 1)$ del intervalo.Y $f'(x) < 0$ en $(0, 1)$.

(b)

Si, porque se cumplen las condiciones necesarias, que son las condiciones 1. y 2. de la pregunta anterior.

(c)

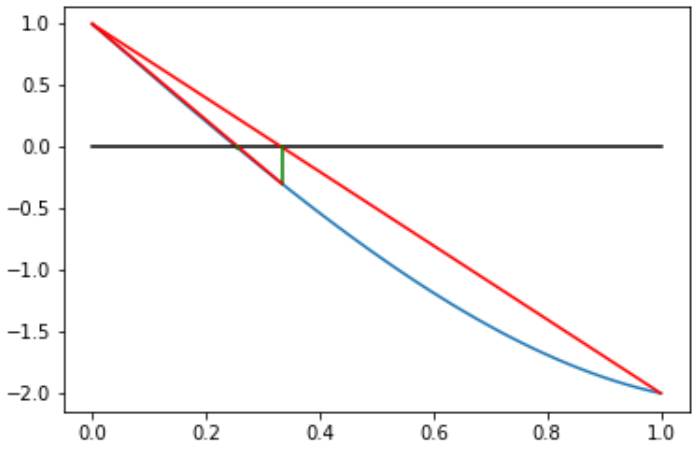
El método de Regula-Falsi calcula un nuevo punto en cada iteración con la fórmula

$$c = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{bf(b) - bf(a) - bf(b) + af(b)}{f(b)-f(a)} = \frac{a\,f(b) - b\,f(a)}{f(b)-f(a)}$$

y selecciona el intervalo siguiente de forma que los signos de los extremos sean distintos

k	a	$c = \frac{a\,f(b)-b\,f(a)}{f(b)-f(a)}$	b	$f(a)$	$f(c)$	$f(b)$	$cota\ de\ error = b-a$
1	0	0.3333	1	1	-0.296	-2	1
2	0	0.2571	0.3333	1		-0.296	0.3333

Y podemos dar como raíz aproximada 0.2571 (la solución exacta es 0.2541)



(d)

La cota de error es la longitud del intervalo que contiene la raíz

$$b_2 - a_2 = 0.3333$$

Regula-Falsi no suele dar buenas cotas de error porque la longitud del intervalo que contiene la raíz en muchos casos no se reduce significativamente con las iteraciones. En este caso el error absoluto es $E_a = 0.2571 - 0.2541 = 0.003$ que es mucho menor que la cota dada.