Computación Numérica

Primer Parcial - Marzo 2018

1. Una máquina almacena números en punto flotante en base 2 en 16 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los cinco siguientes para el exponente sesgado y los últimos diez bits para la mantisa.

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

   (b) Calcular el valor mínimo y máximo de los números positivos normalizados. Expresarlos en binario y en decimal. ¿Qué precisión tendría cada uno?

   (c) ¿Cuántos números positivos normalizados se pueden representar de forma exacta con este formato? Razonarlo.

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

$$2^m = 2^5 = 32$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \ldots, 30, 31]$$

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

$$[R, 1, \ldots, 30, R]$$

Pero los enteros representados serían los anteriores menos el $sesgo = 2^{m-1} - 1 = 2^{5-1} - 1 = 2^4 - 1 = 15$, es decir, el rango de números a representar sería

$$[R, 1 - 15, \ldots, 30 - 15, R] = [R, -14, \ldots, 15, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -14 \text{ y } e_{max} = 15}$$

(b)
El valor mínimo tiene mantisa y exponente mínimos y se representa en binario

signo	exponente	mantisa
0	00001	00000 00000

que se corresponde con

$$1.00000\,00000 \times 2^{-14} \longrightarrow 2^{-14} \approx 6.10 \times 10^{-5}$$

Solución:

$$\boxed{\text{Mínimo número normalizado: } 6.10 \times 10^{-5}}$$

El valor máximo tiene mantisa y exponente máximos. Se representa en binario

signo	exponente	mantisa
0	11110	11111 11111

que, teniendo en cuenta el bit escondido, tenemos una mantisa con 11 dígitos, y el número máximo es

$$1.11111\,11111 \times 2^{15}$$

que expresado en decimal es

$$\left(1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-7} + 2^{-8} + 2^{-9} + 2^{-10}\right) \times 2^{15} \approx 65504$$

Solución:

$$\boxed{\text{Máximo número normalizado: } 65504}$$

La precisión del numero mínimo es 11 porque es el número de dígitos almacenados más el bit escondido $10 + 1 = 11$(c) El número de mantisas posibles es 2^{10} (el bit escondido es siempre 1, no hay dos opciones). Y el número de exponentes posibles es $2^m - 2 = 2^5 - 2 = 32 - 2 = 30$, es decir, los posibles para 5 bits menos los reservados. En total $2^{10} \times 30 = 30720$

Solución:

$$\boxed{\text{30720 números normalizados}}$$

2. (a) Sea la ecuación

$$\frac{x^2}{2} - \ln(2+x) = 0$$

(b) Demostrar que en $[1,2]$ existe una única raíz.

(c) ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

(d) Aproximar la raíz haciendo dos iteraciones con el método de bisección en dicho intervalo.

(e) Dar una cota del error cometido al calcular esta raíz.

(a)

Sea $f(x) = x^2/2 - \ln(2+x) = 0$. Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[1,2]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es la suma de un polinomio y un logaritmo (el intervalo $[1,2]$ está incluído en el dominio de esta función logarítmica), todas ellas funciones continuas.
2. f tiene distinto signo en los extremos del intervalo:

$$f(1) = -0.6 \quad \text{y} \quad \text{f}(2) = 0.6$$

3. f es estrictamente creciente o decreciente en $[1,2]$. Es decir $f' > 0$ o $f' < 0$ en $(1,2)$:

$$f'(x) = x - \frac{1}{2+x} = \frac{x^2+2x-1}{2+x}.$$

Teniendo en cuenta que

$$x^2+2x-1 = 0 \Longrightarrow x_{1,2} = \frac{-2 \pm \sqrt{4+4}}{2} \Longrightarrow x_1 = -2.41 \quad x_2 = 0.41$$

podemos factorizar el numerador

$$f'(x) = \frac{(x+2.41)(x-0.41)}{(2+x)} = \frac{(+)(+)}{(+)} > 0$$

para todos los $x \in (1,2)$ del intervalo.Y $f'(x) > 0$ en $(1,2)$.

(b)

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

(c)

k	a	m	b	*f(a)*	*f(m)*	*f(b)*	*cota de error*
1	1	$(1+2)/2 = 1.5$	2	-0.6	-0.13	0.6	0.5
2	1.5	$(1.5+2)/2 = 1.75$	2	-0.13		0.6	0.25

Y podemos dar como raíz aproximada 1.75 (la solución exacta es 1.6007)

(d)

La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{2-1}{2^2} = \frac{1}{4} = 0.25$$