Computación Numérica

Primer Parcial A - Febrero 2017

1. Una máquina almacena números en punto flotante en base 2 en 12 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los siete siguientes para el exponente sesgado y los últimos cuatro bits para la mantisa.

   (a) Calcular los exponentes máximo y mínimo.

   (b) Calcular el valor mínimo y máximo desnormalizados. Expresarlos en binario y en decimal. ¿Qué precisión tendría cada uno?

   (c) ¿Cuántos números positivos desnormalizados se pueden representar con este formato? Razonarlo.

   (d) Calcular el número siguiente en base 10.

signo	exponente	mantisa
1	0001110	0100

   (e) Calcular el número 271 en este formato. Redondear al par más cercano. ¿Qué error cometemos al redondearlo (dar el resultado en decimal)?

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

$$2^m = 2^7 = 128$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \ldots, 126, 127]$$

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

$$[R, 1, \ldots, 126, R]$$

Pero los enteros representados serían los anteriores menos el $sesgo = 2^{m-1} - 1 = 2^{7-1} - 1 = 2^6 - 1 = 63$, es decir, el rango de números a representar sería

$$[R, 1-63, \ldots, 126-63, R] = [R, -62, \ldots, 63, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -62 \;\text{ y }\; e_{max} = 63}$$

(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 tiene mantisa mínima y se representa en binario

signo	exponente	mantisa
0	0000000	0001

que se corresponde con

$$0.0001 \times 2^{-62} \longrightarrow 2^{-4} \times 2^{-62} = 2^{-4-62} = 2^{-66} \approx 1.36 \times 10^{-20}$$

Solución:

$\boxed{\text{Mínimo número desnormalizado: } 1.36 \times 10^{-20}}$

El valor máximo tiene mantisa máxima. Se representa en binario

signo	exponente	mantisa
0	0000000	1111

que se corresponde con

$$0.1111 \times 2^{-62} \longrightarrow \left(2^{-1} + 2^{-2} + 2^{-3} + 2^{-4}\right) \times 2^{-62} \approx 2.03 \times 10^{-19}$$

Solución:

$\boxed{\text{Máximo número desnormalizado: } 2.03 \times 10^{-19}}$

La precisión del numero mínimo es 1 porque los ceros a la izquierda no cuentan a efecto de precisión y la del número máximo es 4

Solución:

$\boxed{p_{min} = 1 \text{ y } p_{max} = 4}$

(c) Como exponente es siempre -62 el número de números desnormalizados representables es el número de mantisas posibles. Cómo sólo podemos variar los elementos de 4 bits (el otro sería el bit escondido que es siempre 0) serán en total $2^4 = 16$ mantisas distintas. Y si no consideramos el 0 un número desnormalizado, serán $16 - 1 = 15$ mantisas distintas y por lo tanto

Solución:

$\boxed{\text{15 números desnormalizados}}$

(d)

signo	exponente	mantisa
1	1000010	0100

SIGNO: El número es negativo porque el bit del signo es 1.

EXPONENTE: Vimos en el apartado (a) que el $sesgo = 63$.

$$e = (1000010)_2 \longrightarrow 2^6 + 2^1 = 66,$$

y el exponente es $66 - sesgo = 66 - 63 = 3$.

MANTISA: Teniendo en cuenta el bit escondido

$$\mathbf{1}.0100 \times 2^3 \longrightarrow \left(1 + 2^{-2}\right) \times 2^3 = 10$$

Y teniendo en cuenta el signo

Solución:

$$\boxed{-10}$$

(e) Calcular el número 271 en este formato.

MANTISA:

$$\begin{array}{lccccccccc} \text{Cociente} & 271 & 135 & 67 & 33 & 16 & 8 & 4 & 2 & 1 \\ \text{Resto} & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & \swarrow \end{array}$$

Por lo que tenemos

$$(271)_{10} = (100001111)_2 \longrightarrow 1.00001111 \times 2^8$$

Redondeando al par más cercano la mantisa es 1.0001 siendo el primero el bit escondido.

EXPONENTE: Como $sesgo = 63$ representaremos

$$8 + 63 = 71$$

$$\begin{array}{lccccccc} \text{Cociente} & 71 & 35 & 17 & 8 & 4 & 2 & 1 \\ \text{Resto} & 1 & 1 & 1 & 0 & 0 & 0 & \swarrow \end{array}$$

y el exponente en binario es $(1000111)_2$ y la representación completa será

signo	exponente	mantisa
0	1000111	0001

Hemos representado 271 con

$$1.0001 \times 2^8$$

por lo que el error es

$$\left|271 - (1 + 2^{-4}) \times 2^8\right| = |271 - 272| = 1$$

2. $\boxed{\text{Mostrar que en la representación binaria de precisión simple de la norma IEEE 754 el número de dígitos decimales significativos es aproximadamente 7.}}$

Podemos escribir cualquier número binario, con $b_0 = 1$,

$$x = \pm\Big(1.b_1 b_2 \ldots b_{23} b_{24} b_{25} \ldots\Big) \times 2^e.$$

Si lo redondeamos hacia cero,

$$x^* = \pm\Big(1.b_1 b_2 \ldots b_{23}\Big) \times 2^e,$$

y el error relativo

$$\frac{|x - x^*|}{|x|} = \frac{\Big(0.\overbrace{00\ldots0}\,b_{24} b_{25}\ldots\Big) \times 2^e}{\Big(1.b_1 b_2 \ldots b_{23} b_{24} b_{25} \ldots\Big) \times 2^e} \leq \frac{0.\overbrace{00\ldots0}\,b_{24} b_{25}\ldots}{1.b_1 b_2 \ldots b_{23} b_{24} b_{25} \ldots} \leq$$

$$\leq \frac{0.\overbrace{00\ldots0}\,b_{24} b_{25}\ldots}{1} \leq \frac{0.\overbrace{00\ldots0}\,11\ldots}{1} \leq 0.\overbrace{00\ldots1} = 2^{-23} \approx 1.1921 \times 10^{-7}$$

Por lo tanto

Solución:

$$\boxed{\frac{|x - x^*|}{|x|} \leq 5 \times 10^{-7}}$$

y el número de dígitos decimales significativos es al menos 7.

3. $\boxed{\text{El error relativo aproximado al final de una iteración para calcular la raíz de una ecuación es 0.09\%. ¿Cuál es el mayor número de cifras significativas que podemos dar por buenas en la solución?}}$

Se tiene

$$E_r = \frac{0.09}{100} = 0.0009 = 9 \times 10^{-4} = 0.9 \times 10^{-3} < 5 \times 10^{-3}$$

pero

$$E_r = 9 \times 10^{-4} \not< 5 \times 10^{-4}$$

y la solución es

Solución:

$\boxed{\text{3 dígitos significativos}}$

Computación Numérica

Primer Parcial A - Febrero 2017

1. Si disponemos de 8 bits para almacenar números enteros en binario

   (a) Si lo usamos para representar sólo números positvos ¿cómo representaríamos 18 en este sistema de 8 bits?

   (b) Si lo usamos para representar enteros con signo con representación sesgada ¿Cuál es el mayor entero que podemos representar? ¿Y el mínimo?

(a)

Cociente	18	9	4	2	1
Resto	0	1	0	0	$\swarrow$

Solución:

$$\boxed{(18)_{10} = (00010010)_2}$$

(b) El número de enteros con signo que podemos representar con m bits serían $2^m = 2^8 = 256$ y como empiezan en 0 acabarían en 255. Pero los enteros representados serían los anteriores menos el

$$sesgo = 2^{m-1} = 2^7 = 128$$

, es decir, el rango de números a representar sería

$$[0 - 128, \ldots, 255 - 128] = [-128, \ldots, 127]$$

y por lo tanto el valor mínimo y máximo son

Solución:

$$\boxed{min = -128, \qquad max = 127}$$

2. 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 ocho siguientes para el exponente sesgado y los últimos seis bits para la mantisa.

   (a) Calcular los exponentes máximo y mínimo.

   (b) ¿Cual sería el ϵ de máquina expresado en base 10? ¿Cuál es el mayor entero que se puede almacenar de forma exacta?

   (c) Calcular el número 271 en este formato. Redondear por truncamiento hacia el cero. ¿Qué error absoluto cometemos al redondearlo (el error, en decimal)?

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

$$2^m = 2^8 = 256$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \ldots, 254, 255]$$

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

$$[R, 1, \ldots, 254, R]$$

Pero los enteros representados serían los anteriores menos el

$$sesgo = 2^{m-1} - 1 = 2^{8-1} - 1 = 2^7 - 1 = 127,$$

es decir, el rango de números a representar sería

$$[R, 1 - 127, \ldots, 254 - 127, R] = [R, -126, \ldots, 127, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -126 \text{ y } e_{max} = 127}$$

(b) El número de dígitos de la mantisa es 6. El número 1 se representa

$$1 \quad = \quad \mathsf{1.000000} \times \mathsf{2}^{\mathsf{0}}$$

y el siguiente número representable es

$$1 + \epsilon \quad = \quad \mathsf{1.000001} \times \mathsf{2}^{\mathsf{0}}$$

Por lo que

$$\epsilon \quad = \quad \mathsf{0.000001} \times \mathsf{2}^{\mathsf{0}} \quad = \quad 2^{-6}$$

Y en base 10,

Solución:

$$\boxed{\epsilon = 0.015625}$$

(c) El entero más grande es 2^p, donde la precisión es, teniendo en cuenta el bit escondido, $p = 6 + 1$. Por lo tanto

Solución:

$$\boxed{2^7 = 128}$$

(e) Calcular el número 271 en este formato.

MANTISA:

$$\begin{array}{lccccccccc} \text{Cociente} & 271 & 135 & 67 & 33 & 16 & 8 & 4 & 2 & 1 \\ \text{Resto} & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \swarrow \end{array}$$

Por lo que tenemos

$$(271)_{10} = (100001111)_2 \longrightarrow 1.00001111 \times 2^8$$

Redondeando por truncamiento hacia el cero la mantisa es 1.000011 siendo el primero el bit escondido.

EXPONENTE: Como $sesgo = 127$ representaremos

$$8 + 127 = 135$$

$$\begin{array}{lcccccccc} \text{Cociente} & 135 & 67 & 33 & 16 & 8 & 4 & 2 & 1 \\ \text{Resto} & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \swarrow \end{array}$$

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

signo	exponente	mantisa
0	10000111	000011

Hemos representado 271 con

$$1.000011 \times 2^8$$

por lo que el error es

$$\left|271 - (1 + 2^{-5} + 2^{-6}) \times 2^8\right| = |271 - 268| \approx 3$$

3. Mostrar que en la representación binaria de precisión simple de la norma IEEE 754 el número de dígitos decimales significativos es aproximadamente 7.

Podemos escribir cualquier número binario, con $b_0 = 1$,

$$x = \pm\Big(1.b_1 b_2 \dots b_{23} b_{24} b_{25} \dots\Big) \times 2^e.$$

Si lo redondeamos hacia cero,

$$x^* = \pm\Big(1.b_1 b_2 \dots b_{23}\Big) \times 2^e,$$

y el error relativo

$$\frac{|x - x^*|}{|x|} = \frac{\Big(0.\overbrace{00\dots0}\,b_{24} b_{25} \dots\Big) \times 2^e}{\Big(1.b_1 b_2 \dots b_{23} b_{24} b_{25} \dots\Big) \times 2^e} \le \frac{0.\overbrace{00\dots0}\,b_{24} b_{25} \dots}{1.b_1 b_2 \dots b_{23} b_{24} b_{25} \dots} \le$$

$$\le \frac{0.\overbrace{00\dots0}\,b_{24} b_{25} \dots}{1} \le \frac{0.\overbrace{00\dots0}\,11\dots}{1} \le 0.\overbrace{00\dots1} = 2^{-23} \approx 1.1921 \times 10^{-7}$$

Por lo tanto

Solución:

$$\boxed{\frac{|x - x^*|}{|x|} \le 5 \times 10^{-7}}$$

y el número de dígitos decimales significativos es al menos 7.

4. Truncar y redondear al par más cercano, si la precisión es 3, los siguientes números:

número	$1,111000$	$1,101000$	$1,111001$	$1,110001$

número	$1,111000$	$1,101000$	$1,111001$	$1,110001$
número redondeado	10,0	1,10	10,0	1,11