Computación Numérica

Primer Parcial A - Febrero 2015

1. Representamos un entero utilizando 11 bits:

   (a) ¿Cuál es el máximo entero no negativo que puede representar?

   (b) Y si fueran enteros con signo y usáramos la representación sesgada ¿cuál sería el mayor positivo?

(a) Con m bits representamos enteros en $[0, (2^m - 1)_{10}]$. Por lo tanto

$$[0, (2^m - 1)_{10}] = [0, (2^{11} - 1)_{10}] = [0, 2047],$$

Solución:

$$\boxed{2047}$$

(b) El enteros con signo que podemos representar con m bits serían los mismos que en el caso anterior menos el $sesgo = 2^{m-1} = 2^{10} = 1024$, es decir

$$[0 - 1024, 2047 - 1024] = [-1024, 1023]$$

y por lo tanto el valor máximo es

Solución:

$$\boxed{1023}$$

2. Sea el conjunto de números en punto flotante en base 2, con precisión 4 y $e_{max} = 7$ y que sigue normas análogas a la IEEE 754 pero con distinto número de bits para el exponente y la mantisa. Calcular:

   (a) Número de bits del exponente y la mantisa.

   (b) El valor mínimo y máximo normalizados en binario y en decimal.

   (c) El valor mínimo y máximo desnormalizados en binario y en decimal.

(a) Como seguimos la norma IEEE 754, tenemos

$$sesgo = 2^{m-1} - 1 = 7 \Rightarrow 2^{m-1} = 8 \Rightarrow m - 1 = 3 \Rightarrow m = 4.$$

Y si tenemos en cuenta que la precisión es 4 y uno de los bits es el bit escondido la representación sería:

signo	exponente	mantisa
s	$e_1e_2e_3e_4$	$m_1m_2m_3$

Solución:

bits exponente = 4, bits mantisa = 3

(b) La representación binaria del número más grande normalizado es

signo	exponente	mantisa
0	1110	111

que tendría el exponente máximo

$$(1.111) \times 2^7 = \left(1 + 2^{-1} + 2^{-2} + 2^{-3}\right) 2^7 = 240.$$

La representación binaria del número más pequeño normalizado es

signo	exponente	mantisa
0	0001	000

que sería tendría el exponente mínimo que es $1 - sesgo = 1 - 7 = -6$ y por lo tanto

$$(1.000) \times 2^{-6} = 2^{-6} = 0.015625$$

Solución:

$Max = 240$, $min = 0.015625$

(c) Si el número es desnormalizado el exponente es $e = 0000$ pero se le atribuye el valor del menor exponente posible, es decir, -6. Además, el bit escondido es cero. La representación binaria del número más grande normalizado es

signo	exponente	mantisa
0	0000	111

que sería tendría el exponente máximo

$$(0.111) \times 2^{-6} = \left(2^{-1} + 2^{-2} + 2^{-3}\right) 2^{-6} = 0.013671875.$$

La representación binaria del número más pequeño normalizado es

signo	exponente	mantisa
0	0000	001

por lo tanto

$$(0.001) \times 2^{-6} = 2^{-3} \times 2^{-6} = 2^{-9} = 0.001953125$$

Solución:

$$\boxed{Max = 0.013671875,\ min = 0.001953125}$$

3. Demostrar 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

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

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

Por lo tanto

Solución:

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

4. Redondear al par más cercano, si la precisión es 4, los siguientes números en base 2: $n_1 = 1,111111$, $n_2 = 1,111001$, $n_3 = 1,010100$, $n_4 = 1,010101$, $n_5 = 1,010001$.

	1.111111	1.111001	1.010100	1.011100	1.010101
Redondeado	10.00	1.111	1.010	1.100	1.011

5. El error relativo aproximado al final de una iteración para calcular la raíz de una ecuación es 0.07%. ¿Cuál es el mayor número de cifras significativas que podemos dar por buenas en la solución?

Se tiene

$$\frac{|x - x^*|}{|x|} = \frac{0.07}{100} = 7 \times 10^{-4} > 5 \times 10^{-4},$$

pero

$$\frac{|x - x^*|}{|x|} = \frac{0.07}{100} = 7 \times 10^{-4} < 5 \times 10^{-3},$$

y por lo tanto

Solución:

Tres dígitos significativos

Computación Numérica

Primer Parcial A - Febrero 2015

1. Si el número

signo	1 *bit*	`1`
exponente	5 *bits*	`10001`
mantisa	10 *bits*	`0110100000`

sigue la norma IEEE 754 para representación en punto flotante con 16 bits llamado de media precisión, calcular su representación en base 10.

El signo es -1. Como el número de bits es $m = 5$, el sesgo es $2^{m-1} - 1 = 2^4 - 1 = 15$. El exponente es $E = e - sesgo = e - 15$. Como

$$e = (10001)_2 = 2^4 + 2^0 = 17,$$

tenemos $E = 17 - 15 = 2$. Para calcular la mantisa hemos de tener en cuenta el bit escondido

$$\mathbf{1}.01101 \times 2^2 = \left(1 + 2^{-2} + 2^{-3} + 2^{-5}\right) \times 2^2 = 5.625.$$

Solución:

$\boxed{-5.625}$

2. Una máquina almacena números en punto flotante en media precisión de la norma IEEE 754:

   (a) ¿Cual sería el ϵ de máquina expresado en base 10?

   (b) ¿Cuál es el mayor entero que se puede almacenar de forma exacta? Escribirlo en decimal y en binario.

   (c) ¿Cuál sería la representación de 0, $+\infty$, $-\infty$?

   (d) Da un ejemplo de representación de `NaN`.

(a) El número de dígitos de la mantisa es 10. El número 1 se representa

$$1 \quad = \quad \texttt{1.0000000000} \times 2^{\texttt{0}}$$

y el siguiente número representable es

$$1 + \epsilon \quad = \quad \texttt{1.0000000001} \times 2^{\texttt{0}}$$

Por lo que

$$\epsilon \quad = \quad \mathtt{0.0000000001} \times 2^{\mathtt{0}} \quad = \quad 2^{-10}$$

Y en base 10,

Solución:

$$\boxed{9.77 \times 10^{-4}}$$

(b) El entero más grande es 2^p, donde la precisión es, teniendo en cuenta el bit escondido, $p = 10 + 1$. Thus

Solución:

$$\boxed{2^{11} = (2048)_{10} = (100000000000)_2}$$

(c) y (d)

	signo	exponente	mantisa
cero	0	00000	0000000000

	signo	exponente	mantisa
$+\infty$	0	11111	0000000000

	signo	exponente	mantisa
$-\infty$	1	11111	0000000000

	signo	exponente	mantisa
NaN	0	11111	0001000000

3. Representar 0.2 en media precisión. Dar el error absoluto en base 10.

Convertimos a binario:

Parte Fraccionaria	0.2	0.4	0.8	0.6	0.2	$\cdots$
Parte Entera	0	0	1	1	0	$\cdots$

Y tenemos que $(0.2)_{10} = (0.001100110011\ldots)_2$, es decir

$$(0.2)_{10} = 1.\overbrace{1001100110}01100\ldots \times 2^{-3}$$

que redondeado al para más cercano es

$$(0.2)_{10} \approx 1.1001100110\ldots \times 2^{-3}$$

El sesgo es 15, y $e = -3 + sesgo = -3 + 15 = 12$. Pasándolo a binario

Quotients	12	6	3	1
Remainders	0	0	1	

Y el exponente es, $(e)_2 = 1100$. Y teniendo en cuenta el bit escondido

signo	exponente	mantisa
0	01100	1001100110

El número redondeado, x^*, es

$$\left(1 + 2^{-1} + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9}\right) \times 2^{-3} \approx 0.1999951$$

Solución:

$$\boxed{|0.2 - x^*| = 4.88 \times 10^{-5}}$$

4. ¿Con cuantos dígitos significativos aproxima $x_1^* = 0.27351$ a $x_1 = 0.2736$? ¿Y $x_2^* = 1$ a $x_2 = 0.9999$? Entonces, ¿cómo deberíamos escribir x_2^*?

$$\frac{0.2736 - 0.27351}{0.2736} \approx 3.3 \times 10^{-4} \leq 5. \times 10^{-4},$$

Solución:

Cuatro dígitos significativos

$$\frac{0.9999 - 1}{0.9999} = 1. \times 10^{-4} \leq 5. \times 10^{-4},$$

Solución:

Cuatro dígitos significativos

Deberíamos escribir

Solución:

$$\boxed{x_2^* = 1.000}$$

Computación Numérica

Primer Parcial A - Febrero 2015

1. Si el número

signo	1 *bit*	`1`
exponente	5 *bits*	`10001`
mantisa	10 *bits*	`0110100000`

sigue la norma IEEE 754 para representación en punto flotante con 16 bits llamado de media precisión, calcular su representación en base 10.

El signo es -1. Como el número de bits es $m = 5$, el sesgo es $2^{m-1} - 1 = 2^4 - 1 = 15$. El exponente es $E = e - sesgo = e - 15$. Como

$$e = (10001)_2 = 2^4 + 2^0 = 17,$$

tenemos $E = 17 - 15 = 2$. Para calcular la mantisa hemos de tener en cuenta el bit escondido

$$\mathbf{1}.01101 \times 2^2 = \left(1 + 2^{-2} + 2^{-3} + 2^{-5}\right) \times 2^2 = 5.625.$$

Solución:

$\boxed{-5.625}$

2. Una máquina almacena números en punto flotante en media precisión de la norma IEEE 754:

   (a) ¿Cual sería el ϵ de máquina expresado en base 10?

   (b) ¿Cuál es el mayor entero que se puede almacenar de forma exacta? Escribirlo en decimal y en binario.

   (c) ¿Cuál sería la representación de 0, $+\infty$, $-\infty$?

   (d) Da un ejemplo de representación de `NaN`.

(a) El número de dígitos de la mantisa es 10. El número 1 se representa

$$1 \quad = \quad \texttt{1.0000000000} \times 2^{\texttt{0}}$$

y el siguiente número representable es

$$1 + \epsilon \quad = \quad \texttt{1.0000000001} \times 2^{\texttt{0}}$$

Por lo que

$$\epsilon \quad = \quad \mathtt{0.0000000001} \times 2^{\mathtt{0}} \quad = \quad 2^{-10}$$

Y en base 10,

Solución:

$$\boxed{9.77 \times 10^{-4}}$$

(b) El entero más grande es 2^p, donde la precisión es, teniendo en cuenta el bit escondido, $p = 10 + 1$. Thus

Solución:

$$\boxed{2^{11} = (2048)_{10} = (100000000000)_2}$$

(c) y (d)

	signo	exponente	mantisa
cero	0	00000	0000000000

	signo	exponente	mantisa
$+\infty$	0	11111	0000000000

	signo	exponente	mantisa
$-\infty$	1	11111	0000000000

	signo	exponente	mantisa
NaN	0	11111	0001000000

3. Representar 0.2 en media precisión. Dar el error absoluto en base 10.

Convertimos a binario:

Parte Fraccionaria	0.2	0.4	0.8	0.6	0.2	$\cdots$
Parte Entera	0	0	1	1	0	$\cdots$

Y tenemos que $(0.2)_{10} = (0.001100110011\ldots)_2$, es decir

$$(0.2)_{10} = 1.\overbrace{1001100110}01100\ldots \times 2^{-3}$$

que redondeado al para más cercano es

$$(0.2)_{10} \approx 1.1001100110\ldots \times 2^{-3}$$

El sesgo es 15, y $e = -3 + sesgo = -3 + 15 = 12$. Pasándolo a binario

Quotients	12	6	3	1
Remainders	0	0	1	

Y el exponente es, $(e)_2 = 1100$. Y teniendo en cuenta el bit escondido

signo	exponente	mantisa
0	01100	1001100110

El número redondeado, x^*, es

$$\left(1 + 2^{-1} + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9}\right) \times 2^{-3} \approx 0.1999951$$

Solución:

$$\boxed{|0.2 - x^*| = 4.88 \times 10^{-5}}$$

4. ¿Con cuantos dígitos significativos aproxima $x_1^* = 0.27351$ a $x_1 = 0.2736$? ¿Y $x_2^* = 1$ a $x_2 = 0.9999$? Entonces, ¿cómo deberíamos escribir x_2^*?

$$\frac{0.2736 - 0.27351}{0.2736} \approx 3.3 \times 10^{-4} \leq 5. \times 10^{-4},$$

Solución:

Cuatro dígitos significativos

$$\frac{0.9999 - 1}{0.9999} = 1. \times 10^{-4} \leq 5. \times 10^{-4},$$

Solución:

Cuatro dígitos significativos

Deberíamos escribir

Solución:

$$\boxed{x_2^* = 1.000}$$