Computación Numérica

Primer Parcial A - Febrero 2016

1. En una máquina controlada numéricamente, los enteros no negativos necesitan ser almacenados en un espacio de memoria.

   (a) ¿Cuál es el número mínimo de bits necesarios para representar todos los enteros entre 0 y 1300?

   (b) Para el mismo número de bits, si fueran enteros con signo y usara la representación sesgada ¿cuál sería el mayor positivo?

   (c) ¿Cómo representarías entonces -1000?

(a) Con m bits representamos 2^m enteros. Como empezamos representado el 0 podemos representar enteros positivos dentro del rango $[0, (2^m - 1)_{10}]$. Si tomamos $m = 10$

$$1300 \not< (2^m - 1) = 1024 - 1 = 1023$$

que no nos vale. Pero si tomamos $m = 11$

$$1300 < (2^m - 1) = 2048 - 1 = 2047$$

Solución:

$\boxed{\text{11 bits}}$

(b) El número de enteros con signo que podemos representar con m bits serían los mismos que el número de enteros sin signo. Pero los enteros representados serían los anteriores menos el $sesgo = 2^{m-1} = 2^{10} = 1024$, es decir, el rango de números a representar sería

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

y por lo tanto el valor máximo es

Solución:

$\boxed{1023}$

(c) Tendríamos que sumarle el *sesgo*

$$-1000 + sesgo = -1000 + 1024 = 24$$

Convertimos a binario dividiendo sucesivamente entre dos y nos quedamos con el último cociente y todos los restos empezando por el último obtenido:

Cociente	24	12	6	3	1
Resto		0	0	0	1

y el número es

Solución:

$\boxed{(00000011000)_2}$

2. Una máquina almacena números en punto flotante en 14 bits. El primer bit se usa para el signo del número, los seis siguientes para el exponente sesgado y los últimos siete bits para la magnitud de la mantisa. Si se sigue un criterio similar al de la norma IEEE 754, calcular en base 10 el número

signo	exponente	mantisa
1	000110	0110000

EXPONENTE

Como el número de bits es $m = 6$, el sesgo es $2^{m-1} - 1 = 2^5 - 1 = 31$. El exponente es $E = e - sesgo = e - 31$. Como

$$e = (000110)_2 = 2^2 + 2^1 = 6,$$

tenemos $E = 6 - 31 = -25$.

MANTISA

Teniendo en cuenta el bit escondido

$$\mathbf{1}.0110000 \times 2^{-25} = \left(1 + 2^{-2} + 2^{-3}\right) \times 2^{-25} = 4.0978 \times 10^{-8}$$

Y teniendo en cuenta el signo

Solución:

$$\boxed{-4.0978 \times 10^{-8}}$$

3. Sea el conjunto de números en punto flotante en base 2, con exponente de 3 bits y precisión 4 y que sigue normas análogas a la IEEE 754. Determinar cuántos números positivos normalizados se pueden representar.

EXPONENTES

Como el número de bits es $m = 3$, podemos representar $2^m = 2^3 = 8$ números

$$0, 1, 2, 3, 4, 5, 6, 7.$$

El primer valor y el último están reservados por lo tanto podemos representar

$$R, 1, 2, 3, 4, 5, 6, R.$$

Y como el sesgo es $2^{m-1} - 1 = 2^2 - 1 = 3$ los números representados serán estos menos el sesgo

$$R, -2, -1, 0, 1, 2, 3, R.$$

También podemos razonar el número de matisas de la siguiente forma: como el número de bits es $m = 3$, podemos representar $2^m = 2^3 = 8$ números distintos, pero como dos de ellos (000 y 111) están reservados, tenemos $8 - 2 = 6$ exponentes distintos.

MANTISAS

Como tenemos precisión 4, si tenemos en cuenta el bit escondido, la mantisa utiliza 3 bits, por lo que para cada exponente tenemos $2^m = 2^3 = 8$ mantisas diferentes. Como tenemos 6 exponentes distintos, el número de números normalizados representables en este sistema es

Solución:

$6\,\text{exponentes} \times 8\,\text{mantisas} = 48$ números positivos normalizados distintos

4. 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 $b_0 = 1$,

$$x = \pm\Big(1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots\Big) \times 2^e.$$

Si lo redondeamos hacia cero,

$$x^* = \pm\Big(1.b_1b_2\ldots b_{23}\Big) \times 2^e,$$

y el error relativo

$$\frac{|x-x^*|}{|x|} = \frac{\Big(0.\overbrace{00\ldots0}b_{24}b_{25}\ldots\Big)\times 2^e}{\Big(1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots\Big)\times 2^e} \leq \frac{0.\overbrace{00\ldots0}b_{24}b_{25}\ldots}{1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots} \leq$$

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

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.

5. ¿Con cuantos dígitos significativos aproxima $x^* = 100$ a $x = 99.99$? Entonces, ¿cómo deberíamos escribir x^*?

Se tiene

$$\frac{|x - x^*|}{|x|} = \frac{|99.99 - 100|}{99.99} = 1 \times 10^{-4} > 5 \times 10^{-3},$$

pero

$$\frac{|x - x^*|}{|x|} = 1 \times 10^{-4} < 5 \times 10^{-4},$$

y la solución es

Solución:

4 dígitos significativos y se escribe 100.0

Computación Numérica

Primer Parcial A - Febrero 2016

1. Un ingeniero que trabaja en el Ministerio de Defensa está escribiendo un programa que transforma números reales no negativos al formato entero. Para evitar problemas de overflow

   (a) ¿Cuál es el máximo entero no negativo que puede representar con un entero de 12 bits?

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

   (c) ¿Cómo representarías entonces -2000?

(a) Con m bits representamos 2^m enteros. Como empezamos representado el 0 podemos representar enteros positivos dentro del rango $[0, (2^m - 1)_{10}]$. Si tomamos $m = 12$

$$2^m - 1 = 2^{12} - 1 = 4096 - 1 = 4095$$

Solución:

$\boxed{4095}$

(b) El número de enteros con signo que podemos representar con m bits serían los mismos que el número de enteros sin signo. Pero los enteros representados serían los anteriores menos el $sesgo = 2^{m-1} = 2^{11} = 2048$, es decir, el rango de números a representar sería

$$[0 - 2048, 4095 - 2048] = [-2048, 2047]$$

y por lo tanto el valor máximo es

Solución:

$\boxed{2047}$

(c) Tendríamos que sumarle el *sesgo*

$$-2000 + sesgo = -2000 + 2048 = 48$$

Convertimos a binario dividiendo sucesivamente entre dos y nos quedamos con el último cociente y todos los restos empezando por el último obtenido:

Cociente	48	24	12	6	3	1
Resto		0	0	0	0	1

y el número es

Solución:

$\boxed{(000000110000)_2}$

2. Una máquina almacena números en punto flotante en 18 bits. El primer bit se usa para el signo del número, los siete siguientes para el exponente sesgado y los últimos diez bits para la magnitud de la mantisa. Si se sigue un criterio similar al de la norma IEEE 754:

   (a) ¿Cómo se define el ϵ de máquina y qué utilidad tiene?¿Cual sería el ϵ de máquina expresado en base 10?

   (b) ¿Cuales son el menor y el mayor real positivo que se almacena en forma desnormalizada? Dar el resultado en forma binaria

   (c) ¿Qué precisión tendría cada uno?

   (d) Dar el valor del menor positivo normalizado en forma decimal

(a) El ϵ de máquina es la distancia que existe entre el número 1 y el siguiente número representable en ese sistema.
El ϵ de máquina es una cota superior del error relativo que cometemos al almacenar cualquier número con este sistema.
Como el número de dígitos de la mantisa es 10. El número 1 se representa

$$1 \;=\; 1.0000000000 \times 2^0$$

y el siguiente número representable es

$$1+\epsilon \;=\; 1.0000000001 \times 2^0$$

Por lo que

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

Y en base 10,

Solución:

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

(b) En binario, el menor y el mayor real positivo no normalizados serían

	signo	exponente	mantisa
menor real positivo no normalizado	0	0000000	0000000001

	signo	exponente	mantisa
mayor real positivo no normalizado	0	0000000	1111111111

(c)

Solución:

La precisión del menor número desnormalizado es 1.

Solución:

La precisión del mayor número desnormalizado es 10

(d)

EXPONENTE

El exponente es el mínimo que se puede representar con siete bits en representación sesgada. Como el número de bits es $m = 7$, podemos representar $2^m = 2^7 = 128$ números

$$0, 1, 2, 3, \ldots, 124, 125, 126, 127.$$

El primer valor y el último están reservados por lo tanto podemos representar

$$R, 1, 2, 3, \ldots, 124, 125, 126, R.$$

Y como el sesgo es $2^{m-1} - 1 = 2^6 - 1 = 63$ los números representados serán estos menos el sesgo

$$R, -62, -61, -60, \ldots, 61, 62, 63, R.$$

Y el exponente que buscamos es el mínimo que en este caso es -62.

MANTISA

La menor mantisa es la que es todo ceros y teniendo en cuenta el bit escondido, que para los números normalizados es 1, el menor número normalizado es

$$1.0000000000 \times 2^{-62} = 1 \times 2^{-62} = 2^{-62} = 2.17 \times 10^{-19}$$

Solución:

2.17×10^{-19}

3. Mostrar que en la representación binaria de precisión doble de la norma IEEE 754 el número de dígitos decimales significativos es aproximadamente 16.

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

$$x = \pm\Big(1.b_1b_2\ldots b_{52}b_{53}b_{54}\ldots\Big)\times 2^e.$$

Si lo redondeamos hacia cero,

$$x^* = \pm\Big(1.b_1b_2\ldots b_{52}\Big)\times 2^e,$$

y el error relativo

$$\frac{|x-x^*|}{|x|} = \frac{\Big(0.\overbrace{00\ldots0}b_{53}b_{54}\ldots\Big)\times 2^e}{\Big(1.b_1b_2\ldots b_{52}b_{53}b_{54}\ldots\Big)\times 2^e} \leq \frac{0.\overbrace{00\ldots0}b_{24}b_{25}\ldots}{1.b_1b_2\ldots b_{52}b_{53}b_{54}\ldots} \leq$$

$$\leq \frac{0.\overbrace{00\ldots0}b_{53}b_{54}\ldots}{1} \leq \frac{0.\overbrace{00\ldots0}11\ldots}{1} \leq 0.\overbrace{00\ldots1} = 2^{-52} \approx 2.22\times 10^{-16}$$

Por lo tanto

Solución:

$$\boxed{\frac{|x-x^*|}{|x|} \leq 5\times 10^{-16}}$$

y el número de dígitos decimales significativos es al menos 16.

4. Redondear al par más cercano, si la precisión es 4, los siguientes números en base 2:
$n_1 = 1,111111$, $n_2 = 1,111001$, $n_3 = 1,010100$, $n_4 = 1,011100$

Los números redondeados correspondientes son
$n_1^* = 10,00$, $n_2^* = 1,111$, $n_3^* = 1,010$, $n_4^* = 1,100$