Guía de ejercicios # 6 - Punto Fijo

Organización de Computadoras 2017

UNQ

1 Introducción

El sistema numérico de punto fijo es una generalización de los sistemas enteros que permite representar números fraccionarios. En un sistema de punto fijo, se destinan una cierta cantidad de bits a la parte entera y el resto a la parte fraccionaria, considerando que existe un punto (o coma) que los separa, tal como se utiliza la coma en el sistema decimal. Sin embargo, en los sistemas binarios de punto fijo, el punto no está representado explícitamente, sino que se asume una posición determinada.

Por ejemplo, se puede construir un sistema binario sin signo con punto fijo de 8 bits, en el que 5 bits sean para la parte entera y 3 bits para la parte fraccionaria.

Notación: En adelante, a la notación utilizada para sistemas enteros $BSS(n)$, $SM(n)$ se agrega un dato adicional que hace referencia a la cantidad de bits fraccionarios:

- $BSS(n, m)$: Sistema de punto fijo en *Binario Sin Signo* con n bits en total, de los cuales m son fraccionarios.
- $SM(n, m)$: Sistema de punto fijo en *Signo-Magnitud* con n bits en total, de los cuales 1 es de signo, n-1 son de parte entera y m de parte fraccionaria.

Los bits fraccionarios, al igual que los enteros, tienen un peso, pero está asociado a una potencia de 2 negativa, así, el primer bit (de izquierda a derecha) fraccionario tiene un peso de 2^{-1}, el segundo de 2^{-2}, y así sucesivamente hasta el último bit fraccionario. Entonces, para interpretar una cadena en punto fijo, se debe interpretar por un lado la parte entera en el sistema correspondiente y por otro lado la parte fraccionaria:

$I_{BSS(8,3)}(10011101) = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = 16 + 2 + 1 + 0,5 + 0,125 = 19,625$

También puede interpretarse ignorando inicialmente el hecho de que exista una parte fraccionaria y luego dividiendo el resultado entre 2^m donde m es la cantidad de bits de la parte fraccionaria:

$$I_{BSS(8,3)}(10011101) =$$

$$\frac{1 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0}{2^3} =$$

$$\frac{128 + 16 + 8 + 4 + 1}{8} = \frac{157}{8} = 19,625$$

1.1 Completar la siguiente tabla:

Cadena	$BSS(2)$	$BSS(2,1)$	$BSS(2,2)$
00			
01		$1 * 2^{-1}$	
10	$1 * 2^1$		
11			$1 * 2^{-1} + 1 * 2^{-2}$

1.2 Completar la siguiente tabla:

Cadena	$BSS(3)$	$BSS(3,1)$	$BSS(3,2)$
000			
001		$1 * 2^{-1}$	
010	$1 * 2^1$		
011			$1 * 2^{-1} + 1 * 2^{-2}$
100			
101		$1 * 2^1 + 1 * 2^{-1}$	
110	$1 * 2^2 + 1 * 2^1$		
111			$1 * 2^0 + 1 * 2^{-1} + 1 * 2^{-2}$

2 Interpretación

2.1 Suponer un sistema de punto fijo $BSS(7,3)$. Interprete las cadenas:

a. `0000001`

b. `0101011`

c. `0010110`

d. `1000000`

e. `1000001`

2.2 Suponer un sistema de punto fijo $BSS(10,4)$. Interprete las cadenas:

a. `0100000000`

b. `0101010101`

c. `1000000000`

d. `1111111000`

e. `1111111111`

f. `1010101010`

g. `0111111111`

h. `0110011000`

2.3 Interprete las cadenas del ejercicio anterior en un sistema $BSS(10,3)$.

2.4 Interprete las cadenas del ejercicio anterior en un sistema $SM(10,3)$.

3 Rango y Resolución

Al igual que en los sistemas enteros (BSS, SM y CA_2), el rango en Punto Fijo está determinado por el intervalo de números representables. Por ejemplo, el rango del sistema $BSS(2,1)$ está dado por:
Mínimo:

$$\mathcal{I}_{bss(2,1)}(00) = 0$$

Máximo:

$$\mathcal{I}_{bss(2,1)}(11) = 2^0 + 2^{-1} = 1 + 0,5 = 1,5$$

Y el del sistema $SM(4,2)$ por:
Mínimo:

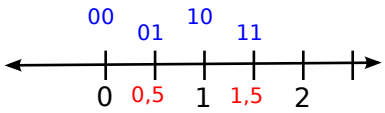
$$\mathcal{I}_{sm(4,2)}(1111) = -(2^0+2^{-1}+2^{-2}) = -(1+0,5+0,5) = -1,75$$

Máximo:

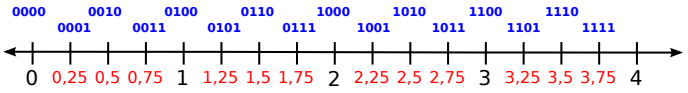
$$\mathcal{I}_{sm(4,2)}(0111) = 2^0 + 2^{-1} + 2^{-2} = 1 + 0,5 + 0,5 = 1,75$$

Sin embargo, el rango se refiere a todos los números representables en el intervalo. En los sistemas enteros esto es trivial (todos los enteros del intervalo), pero en punto fijo, para determinar exactamente qué números están representados dentro del intervalo es necesario el concepto de **resolución**: *la resolución es la distancia entre dos números representables consecutivos.*

Por ejemplo, en el sistema $BSS(2,1)$ la resolución es 0,5 y por lo tanto los números representables son:



Y en el sistema $BSS(4,2)$ la resolución es 0,25, y los números representables son:



3.1 Suponer un sistema $BSS(10,4)$.

(a) ¿Cuántos números se pueden representar?

(b) ¿Cuál es la resolución del sistema?

(c) ¿Cuáles son el máximo y el mínimo número representables?

(d) ¿Cuáles son el máximo y el mínimo número representable en el intervalo (0,1)? (Donde (0,1) es el intervalo desde el 0 hasta el 1, **ambos excluidos**).

3.2 Responder las preguntas anteriores para un sistema $BSS(8,3)$.

3.3 Responder las preguntas anteriores para un sistema $SM(8,3)$.

3.4 Responder las preguntas anteriores para un sistema $SM(10,4)$.

3.5 Calcule la resolución de los siguientes sistemas:

(a) $BSS(8,5)$

(b) $BSS(2,1)$

(c) $BSS(6,4)$

(d) $BSS(10000,1)$

4 Representación y error

Para representar un número en punto fijo, se puede "correr la coma" y representar el número entero correspondiente. Por ejemplo, representar 3,14 en $BSS(7,4)$

- $3,14 * 2^4 = 50,24$
- Redondeo: $50,24 \approx 50$
- $\mathcal{R}_{bss(7)}50 = 0110010$

Sin embargo, la interpretación en $BSS(7,4)$ de 0110010 **no es 3,14**:

$$\mathcal{I}_{bss(7,4)}(0110010) = 2^1 + 2^0 + 2^{-3} = 2 + 1 + 0,125 = 3,125$$

Esto ocurre porque no todos los números del intervalo son representables en el sistema, por lo tanto puede haber un **error de representación**, que es *el valor absoluto de la diferencia entre el número que se deseaba representar y el número que efectivamente se logró representar*. En este ejemplo, la diferencia es de $|3,14 - 3,125| = 0,015$

4.1 Suponer un sistema $BSS(8,4)$. Represente los siguientes números:

(a) 10,2
(b) 0,125
(c) 0,099
(d) 3,75
(e) 20,9

4.2 Suponer un sistema $BSS(10,4)$. Represente los siguientes números:

(a) 1,2
(b) 1,25
(c) 35
(d) 1,0625
(e) 13,763
(f) 1,4

Si alguno no se puede representar, justifique. Calcule el error absoluto y relativo en cada caso.

4.3 Suponer un sistema $SM(8,4)$. Represente los siguientes números:

(a) 1,1
(b) 0,125
(c) 0,099
(d) 4,75
(e) 19,99

4.4 Calcule el error absoluto al representar los siguientes números en $BSS(9,4)$.

(a) 1,1
(b) 0,125
(c) 0,099
(d) 4,75
(e) 19,99

4.5 Represente los siguientes números en $SM(10,4)$.

(a) 24,0
(b) 1,25
(c) -15.25
(d) 1,0625
(e) -13,763
(f) 1,4

Si alguno no se puede representar, justifique. Calcule el error absoluto y relativo en cada caso.

4.6 Suponer un sistema $BSS(4,1)$. ¿Cuál es el máximo error absoluto que puede ocurrir al respresentar un valor dentro del rango? ¿Cuál es el rango?

4.7 Suponer un sistema $BSS(4,2)$. ¿Cuál es el máximo error absoluto que puede ocurrir al respresentar un valor dentro del rango? ¿Cuál es el rango?

4.8 Suponer un sistema $BSS(4,3)$. ¿Cuál es el máximo error absoluto que puede ocurrir al respresentar un valor dentro del rango? ¿Cuál es el rango?

4.9 Suponer un sistema $BSS(4,4)$. ¿Cuál es el máximo error absoluto que puede ocurrir al respresentar un valor dentro del rango? ¿Cuál es el rango?

4.10 Suponer un sistema $SM(4,3)$. ¿Cuál es el máximo error absoluto que puede ocurrir al respresentar un valor dentro del rango? ¿Cuál es el rango?

4.11 Supongamos que se desea utilizar un sistema de punto fijo $SM(X,Y)$ para representar números entre -10 y 10. Se pretende además que el error absoluto sea menor a 0.2. ¿Cuales son los mínimos X,Y que satisfacen estos requerimientos?

4.12 Se necesita un sistema de punto fijo que permita las siguientes cosas:

- Representar al número -17
- Representar al número 42
- Que el error absoluto máximo sea menor a 0.05

Diseñe el sistema con la mínima cantidad de bits.