Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Semestre Otoño 2007
Cálculo Numérico
MA33A-02

Profesor: Gonzalo Hernández.
Auxiliar: Gonzalo Ríos.
Fecha: 14 de Marzo

Auxiliar 1: Representación Numérica

Resumen Materia

1. **Representación de números enteros:** $p = \sum_{i=0}^{n} \alpha_k b^k$ donde $b = 2, 8, 10$ y $\alpha_k \in \{0, ..., b-1\}$
2. **Representación de números reales en** $[0,1]$ **en codificación diádica:** $p = \sum_{i=1}^{n} \alpha_k 2^{-k}$ donde $\alpha_k \in \{0,1\}$
3. **Representación de números reales:** $x = \pm m \cdot b^z$ donde $b \in \mathbb{N}$, $z \in \mathbb{Z}$, $m \in [0,1]$
4. **Representación de punto flotante:** $x = \pm 0.m_1 m_2 ... m_t \cdot b^z$ donde $b \in \mathbb{N}$, $t \in \mathbb{N} \setminus \{\infty\}$, $z \in \mathbb{Z}$, $m_i \in \{0, ..., b-1\}$ .El número t es el número de cifras significativas.
5. **Representación de punto flotante actual:** $x = (-1)^s 2^{(z-c)}(1+m)$ donde $s \in \{0,1\}$ $c = 1023$ $z \in \left\{0, 1...2^{11}-1\right\}$ $m = m_1 m_2 ... m_{52}$ con $m_i \in \{0,1\}$

Ejercicios

1. Determinar un algoritmo para obtener la codificación binaria de un número natural,otro para obtener la codificación diádica de un número entre $[0,1]$.
2. Usando los dos algoritmos anteriores, explique en forma algoritmica como codificar un número real cualquiera. Aplique el algoritmo con $x = 236.892$ en codificación FL de 32 bits.
3. Sea una codificación binaria floating point tipo inicial, donde el número real máximo representable es $2^{63} - 2^{54}$.Explique cuantos bits son necesarios para implementar esta codificación y como se distribuyen.
4. Un grupo de ingenieros de Beauchef, que trabajan en la NASA recibieron en su receptor atómico el siguiente código binario: 0110101011111100011111.Los ingenieros, quienes obtuvieron excelentes promedios en Calculo Numérico, creen que se trata de algún mensaje de extraterrestres, los cuales desesperados por saber que dicen, proponen la siguiente codificación de punto flotante: una modificación de la codificación floating point inicial de 19 bits con los siguientes parámetros: 1 bit signo, 1 bit signo exponente, 7 bits exponente y 11 bits para la mantisa.
   (a) Decodifique el mensaje de los extraterrestes
   (b) Los ingenieros creen saber lo que dicen, por lo que deciden mandan un mensaje utilizando la codificación anterior. El mensaje es 56.715, que es la respuesta a la pregunta fundamental del Universo. Los ingenieros desean saber cual es la pregunta fundamental del Universo. Codifique el mensaje.
   (c) Luego de esperar un rato, y sin obtener respuesta alguna, uno de los ingenieros, el más superticioso de todos, cree que si manda algunos mensajes claves, los extraterrestes los tomarán en cuenta. Deciden mandar los mensajes más importantes de su codificación: el máximo, el mínimo y el más cercano a cero. Uno de los fisicos del lugar no sabía cuales eran estos números. Digale porfavor al fisico cuales son aquellos números.
   (d) Nuevamente sin respuesta, los ingenieros deciden ocupar la codificación actual floating point para mandar el siguiente mensaje: 0.97542, que corresponde al cuociente entre el día solar y el día sideral. (Cualquier coincidencia con la realidad, es solo coincidencia). Codifique el mensaje para ser enviado.
   (e) A los 32 segundos de ser enviado el mensaje, se obtiene respuesta de los extraterrestes. El mensaje obtenido es el siguiente: 00000001000111010111000100111 0...00. Decodifique el mensaje
   (f) Al analizar el mensaje mandado por los extraterrestes, los ingenieros llegan a la conclusión que fue un mensaje de guerra. Suponiendo que el mensaje fue recibido por los extraterrestes, y enviado inmediatamente de vuelta, y sabiendo que la velocidad de las ondas electromagnéticas es de $c = 300000 \left[\frac{km}{seg}\right]$, ¿cuanto tiempo de vida nos queda? ¿que tan lejos se encuentran los extraterrestes?

Respuestas

1. Algoritmos

   (a) Binaria

```
For k=0 To n
    if (x es par){
        a[k]=0;
        x=x/2;
    } else{
        a[k]=1;
        x=(x-1)/2;
    }
End k
```

1. (b) Diádica

```
x=(int)x*(2^n);
For k=n To 1
    if (x es par){
        a[k]=0;
        x=x/2;
    } else{
        a[k]=1;
        x=(x-1)/2;
    }
}
```

2. Codificación número real

   (a) Algoritmo

      i. Determinar la notación cientifica con la mantisa normalizada, es decir, $m \in [0,1]$

      ii. Cambiar base 10 a base 2 dividiendo por potencia de 2 más cercana a la potencia de 10. Esta es $\lceil \log_2(10^z) \rceil$

      iii. Recalcular la mantisa como $m^* = m * \frac{10^z}{2^{\lceil \log_2(10^z) \rceil}}$. Si $m^* < 0.5,$ entonces $\lceil \log_2(10^z) \rceil = \lceil \log_2(10^z) \rceil - 1,\ m^* = 2 \times m^*$

      iv. Determinar codificación binaria para el exponente y codificación diádica para la mantisa.

   (b) Aplicando el algoritmo a $x = 236.892$

      i. $x = 236.892 = x = 0.236892 \times 10^3$

      ii. $\lceil \log_2(10^3) \rceil = \lceil 9.965\,784\,284\,662\,087\,043\,6 \rceil = 10 \Longrightarrow x = 0.236892 \times \frac{10^3}{2^{10}} \times 2^{10}$

      iii. $m^* = 0.236892 \times \frac{10^3}{2^{10}} = 0.231\,339\,843\,75 \ < 0.5 \ \Longrightarrow \ m^* = 2 \times 0.231\,339\,843\,75 = 0.462\,679\,687\,5 \Longrightarrow \ m^* = 2 \times 0.462\,679\,687\,5 = 0.925\,359\,375\,, \quad e = 8$

   (c) Codificación binaria

      i. Exponente: $e = 8 \Longrightarrow 0001000$

      ii. Mantisa: $m = 0.925\,359\,375 \Longrightarrow m * 2^{23} = 0.925\,359\,375 * 2^{23} \ = 7762477.056 \Longrightarrow 7762477$

      iii. Codificar la mantisa

$7762\,477$	$\frac{7762477-1}{2} = 3881\,238$	$\frac{3881\,238}{2} = 1940\,619$	$\frac{1940\,619-1}{2} = 970\,309$	$\frac{970\,309-1}{2} = 485\,154$		
impar	par	impar	impar	par		
1	0	1	1	0		
$\frac{485\,154}{2} = 242\,577$	$\frac{242\,577-1}{2} = 121\,288$	$\frac{121\,288}{2} = 60\,644$	$\frac{60\,644}{2} = 30\,322$	$\frac{30\,322}{2} = 15\,161$		
impar	par	par	par	impar		
1	0	0	0	1		
$\frac{15\,161-1}{2} = 7580$	$\frac{7580}{2} = 3790$	$\frac{3790}{2} = 1895$	$\frac{1895-1}{2} = 947$	$\frac{947-1}{2} = 473$	$\frac{473-1}{2} = 236$	
par	par	impar	impar	impar	par	
0	0	1	1	1	0	
$\frac{236}{2} = 118$	$\frac{118}{2} = 59$	$\frac{59-1}{2} = 29$	$\frac{29-1}{2} = 14$	$\frac{14}{2} = 7$	$\frac{7-1}{2} = 3$	$\frac{3-1}{2} = 1$
par	impar	impar	par	impar	impar	impar
0	1	1	0	1	1	1

      iv. Codificación: 0 0 0001000 11101100111001000101101

      v. $1\times2^{-1}+1\times2^{-2}+1\times2^{-3}+0\times2^{-4}+1\times2^{-5}+1\times2^{-6}+0\times2^{-7}+0\times2^{-8}+1\times2^{-9}+1\times2^{-10}+1\times2^{-11}+0\times2^{-12}+0\times2^{-13}+1\times2^{-14}+0\times2^{-15}+0\times2^{-16}+0\times2^{-17}+1\times2^{-18}+0\times2^{-19}+1\times2^{-20}+1\times2^{-21}+0\times2^{-22}+1\times2^{-23} = 0.925\,359\,368\,324\,279\,785\,16$

      vi. $fl(x) = 0.925\,359\,368\,324\,279\,785\,16 \times 2^8 = 236.\,891\,998\,291\,015\,62$

3. Sistema de codificación binaria

   Como el número real máximo representable en una codificacion binaria siempre es $\frac{2^m-1}{2^m}\times 2^{(2^e-1)}$, donde m es la cantidad de bits de la mantisa, y e la cantidad de bits del exponente, esto equivale a $2^{2^e-1} - 2^{(2^e-1)-m}$, queda un sistema:

   (a) $2^{63} - 2^{54} = 2^{2^e-1} - 2^{(2^e-1)-m} \implies \begin{array}{l} 2^e - 1 = 63 \\ (2^e-1) - m = 54 \end{array} \implies \begin{array}{l} e = 6 \\ m = 9 \end{array}$

4. floating point

   (a)

Signo Mantisa	Signo Exponente	Exponente	Mantisa
0	1	1010101	11110001111

      i. Signo Mantisa: $+$
      ii. Signo Exponente: $-$
      iii. Exponente:$1\times 2^6 + 0\times 2^5 + 1\times 2^4 + 0\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 = 85$
      iv. Mantisa: $1\times 2^{-1}+1\times 2^{-2}+1\times 2^{-3}+1\times 2^{-4}+0\times 2^{-5}+0\times 2^{-6}+0\times 2^{-7}+1\times 2^{-8}+1\times 2^{-9}+1\times 2^{-10}+1\times 2^{-11} = 0.944\,824\,218\,75$
      v. Número: $+0.944\,824\,218\,75\times 2^{-85} = 2.\,442\,313\,362\,564\,252\,549\,6\times 10^{-26}$

   (b) $56.715 = 0.56715\times 10^2 = 0.56715\times 10^2\times\frac{2^6}{2^6} = 0.56715\times\frac{10^2}{2^6}\times 2^6 = 0.56715\times 1.\,562\,5\times 2^6 = 0.886\,171\,875\ \times 2^6$

      i. Signo Mantisa: $+$
      ii. Signo Exponente: $+$
      iii. Exponente: $6 = 0\times 2^7 + 0\times 2^6 + 0\times 2^5 + 0\times 2^4 + 0\times 2^3 + 1\times 2^2 + 1\times 2^1 + 0\times 2^0$
      iv. Mantisa:

	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}
	1	1	1	0	0
0.886 171 875	0.386 171 875	0.136 171 875	0.011 171 875	0.011 171 875	0.011 171 875
2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	2^{-11}
0	1	0	1	1	0
0.011 171 875	0.003 359 375	0.003 359 375	0.001 406 25	0.000 429 687 5	0.000 429 687 5

      v.

Signo Mantisa	Signo Exponente	Exponente	Mantisa
0	0	0000110	11100010110

   (c) Maximos y Minimos

      i. Real Máximo: 0 0 1111111 11111111111 $= \frac{2^{11}-1}{2^{11}}\times 2^{127} = 0.999\,511\,718\,75\times 2^{127}$
      ii. Real Mínimo: 1 0 1111111 11111111111 $= -\frac{2^{11}-1}{2^{11}}\times 2^{127} = -0.999\,511\,718\,75\times 2^{127}$
      iii. Real Más cercano a cero positivo: 0 1 1111111 00000000001 $= 2^{-11}\times 2^{-127} = 2^{-138} = 2.\,869\,859\,254\,937\,225\,361\,3\times 10^{-42}$
      iv. Real Más cercano a cero negativo: 1 1 1111111 00000000001 $= -2^{-11}\times 2^{-127} = -2^{-138} = -2.\,869\,859\,254\,937\,225\,361\,3\times 10^{-42}$

   (d) $0.975421 = 2^{z-c}(1+m)$

      i. Escogemos z de manera que $z - 1023 = -1 \Rightarrow z = 1022$
      ii. Luego: $m = 2\times 0.975421 - 1 = 0.950842$
      iii. La codificación binaria de $z = 1022$ es: 01111111110
      iv. La codificación diádica de $m = 0.950842$ es: 1111001101101010011000011001110110101001110010011001

   (e) Decodificar 0 0000001000 1110101110001001100...00

      i. El número real es: $x = (-1)^s 2^{z-1023}(1+m)$
      ii. El signo es $s = 0$
      iii. El exponente es: $z = 00000001000 = 2^3 = 8$. Luego $z - c = 8 - 1023 = -1015$
      iv. La mantisa es: $m = 1110101110001001110...00 = \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^9 + \left(\frac{1}{2}\right)^{13} + \left(\frac{1}{2}\right)^{16} + \left(\frac{1}{2}\right)^{17} + \left(\frac{1}{2}\right)^{18} = 0.920\,070\,648\,193\,359\,375$
      v. Finalmente: $x = (-1)^0 2^{-1015}(1 + 0.92007)$
      vi. Nos van a atacar el 1 de Septiembre de 2007 a las 10:15 horas. Que el físico calcule lo demás.