Álgebra matricial numérica. Unidad 3. Operaciones vectoriales y matriciales en aritmética de punto flotante

Dr. Javier de Jesús Cortés Aguirre.
Facultad de Ciencias, UNAM.
Semestre: septiembre 2021 - enero 2022.

1. Aritmética de punto flotante.
   - Introducción
   - Desarrollo
     - Conjunto de elementos flotantes
     - Tipos de errores
   - Ejemplos ilustrativos
     - Técnicas de redondeo
     - Técnicas de truncamiento

2. Propiedades del conjunto flotante
   - Ejemplos ilustrativos

3. Unidad de redondeo

En esta sección explicaremos la aritmética con la que trabaja la computadora, llamada Aritmética de punto flotante y los tipos de error que pueden comenterse en la misma.

Conjunto de elementos flotantes

Definición

*(***Conjunto de elementos flotantes ó conjunto flotante)**
Se denomina conjunto de elementos flotantes, a aquel que consta de elementos de la forma:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e \tag{1}$$

con:
t : *precisión del conjunto flotante.*
β : *base del conjunto flotante.*
e : *exponente*
y donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, *para* $i = 2, 3, ..., t$
$-L \leq e \leq U$

Generalmente a este tipo de conjuntos de se les denota como

$$F = fl(\beta, t, L, U)$$

donde a esta notación se le conoce como notación estándar

Tipos de errores

Dado que los elementos con los que trabaja la computadora deben pertenecer a un cierto conjunto flotante, una pregunta natural es plantearnos que pasa si le damos a la computadora un elemento que está fuera del conjunto. Es decir, supongamos que la computadora trabaja con 2 dígitos de precisión ($t = 2$), entonces si en nuestro algoritmo capturamos el valor de $4/3 = 1.333333...$, la computadora deberá aproximarlo a un valor que solo tenga 2 dǵitos. Para ello existen dos formas de hacerlo, usando redondeo o usando truncamiento.

Supongamos que una computadora trabaja con el conjunto de elementos flotantes

$$F = fl(10, 3, 3, 3)$$

Es decir, un conjunto en base 10, una precisión de 3 dígitos y el exponente en un rango entre -3 y 3.

Primero ilustraremos algunos elementos que si pertenecen al conjunto (que si pueden ser generados por esa computadora) y luego lo que pasa al tomar elementos que no pertenezcan al conjunto, ahí utilizaremos las técnicas de redondeo y truncamiento.

- $x_1 = 5$
  Primero debemos colocarlo en la notación estándar indicada en la ecuación 1, así nos queda:

$$x_1 = 5 \times 10^0$$

  observemos que solo usamos un dígito de precisión, la base es 10 y el exponente es 0 (cae en el rango de -3 y 3), por tanto el elemento si pertence al conjunto flotante y la computadora no realizaría algún redondeo o truncamiento.

- $x_2 = 139$
  Al colocarlo en la notación estándar nos queda:

$$x_2 = 1.39 \times$$

- $x_2 = 139$
  Al colocarlo en la notación estándar nos queda:

$$x_2 = 1.39 \times 10^2$$

ahora, usamos tres dígitos de precisión (aún está dentro de la precisión del conjunto), la base sigue siendo 10 y el exponente es 2 (cae en el rango de -3 y 3), por tanto el elemento también pertence al conjunto flotante y la computadora no realizaría algún redondeo o truncamiento.

- $x_3 = 5670$
  A primera vista parece que el elemento no pertenece al conjunto ya que podría superar la precisión del mismo, pero si lo denotamos en notación estándar

$$x_3 = 5.67 \times$$

- $x_3 = 5670$
  A primera vista parece que el elemento no pertenece al conjunto ya que podría superar la precisión del mismo, pero si lo denotamos en notación estándar

$$x_3 = 5.67 \times 10^3$$

observamos que solo usamos 3 dígitos de precisión y que el cero de la última posición se colocó dentro del exponente, además este último sigue estando en el rango de -3 y 3, por tanto el elemento también pertence al conjunto flotante y la computadora no realizaría algún redondeo o truncamiento.

- $x_4 = 1237$
  En este elemento podemos notar que la precisión del conjunto queda superada, al denotarlo en notación estándar nos queda

$$x_4 = 1.237 \times 10^3$$

  efectivamente usamos ahora 4 dígitos de precisión y entonces el elemento no está dentro del conjunto flotante, por lo cual la computadora necesitará aproximarlo a un elemento de su conjunto utilizando redondeo o truncamiento.

Técnicas de redondeo

En el caso de que se realice un redondeo, este debe realizarse al elemento más cercano en el conjunto flotante, a esta técnica se le llama la técnica de redondeo al más cercano. En nuestro ejemplo

$$x_4 = 1237 = 1.237 \times 10^3$$

se encuentra entre dos elementos que si pertenecen al conjunto, los cuales son

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

Técnicas de redondeo

En el caso de que se realice un redondeo, este debe realizarse al elemento más cercano en el conjunto flotante, a esta técnica se le llama la técnica de redondeo al más cercano. En nuestro ejemplo

$$x_4 = 1237 = 1.237 \times 10^3$$

se encuentra entre dos elementos que si pertenecen al conjunto, los cuales son

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

entonces como el redondeo debe realizarse al elemento más cercano, debe efectuarse a x_b, es decir

$$x_4 = 1237 = 1.237 \times 10^3 \xrightarrow{R} x_b = 1240 = 1.24 \times 10^3$$

Tomemos ahora

$$x_5 = 1234 = 1.234 \times 10^3$$

se encuentra entre dos elementos que si pertenecen al conjunto, los cuales son

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

Tomemos ahora

$$x_5 = 1234 = 1.234 \times 10^3$$

se encuentra entre dos elementos que si pertenecen al conjunto, los cuales son

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

como el redondeo debe realizarse al elemento más cercano, debe efectuarse a x_a, es decir

$$x_5 = 1234 = 1.234 \times 10^3 \xrightarrow{R} x_a = 1230 = 1.23 \times 10^3$$

Ejemplos

- $x_6 = 1234.9999999999999999$
- $x_7 = 1235.0000000000000001$
- $x_8 = 1235$

Redondeo a los pares

En el caso de que el valor se encuentre justo en el punto medio de dos elementos flotantes se utiliza el redondeo a los pares, en este caso se redondea al elemento flotante cuyo último dígito significativo es par.

$$x_7 = 1235 = 1.235 \times 10^3 \xrightarrow{R}$$

Redondeo a los pares

En el caso de que el valor se encuentre justo en el punto medio de dos elementos flotantes se utiliza el redondeo a los pares, en este caso se redondea al elemento flotante cuyo último dígito significativo es par.

$$x_7 = 1235 = 1.235 \times 10^3 \xrightarrow{R} 1240 = 1.24 \times 10^3$$

$$x_8 = 1245 = 1.245 \times 10^3 \xrightarrow{R}$$

Redondeo a los pares

En el caso de que el valor se encuentre justo en el punto medio de dos elementos flotantes se utiliza el redondeo a los pares, en este caso se redondea al elemento flotante cuyo último dígito significativo es par.

$$x_7 = 1235 = 1.235 \times 10^3 \xrightarrow{R} 1240 = 1.24 \times 10^3$$

$$x_8 = 1245 = 1.245 \times 10^3 \xrightarrow{R} 1240 = 1.24 \times 10^3$$

Técnicas de truncamiento

En el caso de que se realice un truncamiento, este debe realizarse al elemento del conjunto flotante anterior a nuestro valor. En nuestro ejemplo

$$x_4 = 1237 = 1.237 \times 10^3$$

se encuentra entre los elementos

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

Técnicas de truncamiento

En el caso de que se realice un truncamiento, este debe realizarse al elemento del conjunto flotante anterior a nuestro valor. En nuestro ejemplo

$$x_4 = 1237 = 1.237 \times 10^3$$

se encuentra entre los elementos

$$x_a = 1230 = 1.23 \times 10^3 \text{ y } x_b = 1240 = 1.24 \times 10^3$$

entonces al realizar el truncamiento, nos queda

$$x_4 = 1237 = 1.237 \times 10^3 \xrightarrow{T} x_a = 1230 = 1.23 \times 10^3$$

Observe que al realizar la técnica de truncamiento el valor siempre se aproxima por el valor anterior en el conjunto flotante, sin embargo, al realizar la técnica de redondeo puede aproximarse por el anterior o por el posterior.

Tarea 1

Tarea

Investigue cual es el conjunto flotante que utilizan nuestros sistemas de cómputo que usamos en la vida diaria (computadoras, laptops, tablets, calculadoras, etc) y sus características. A este conjunto se le llama **doble precisión**

Propiedades del conjunto flotante

El conjunto de elementos flotantes tiene algunas propiedades que son de nuestro interés:

- Cardinalidad (número de elementos del conjunto)
- Elementos mínimo y máximo

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

Card=2

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

$$\text{Card}=2(\beta - 1)$$

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

$$\text{Card}=2(\beta - 1)\beta^{t-1}$$

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

$$\text{Card}{=}2(\beta - 1)\beta^{t-1}(U + L + 1)$$

Cardinalidad

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Veamos las posibilidades que hay para cada valor:

$$\text{Card}=2(\beta - 1)\beta^{t-1}(U + L + 1)+1$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.0$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.00$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.00...0$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.00...0 \times \beta^{-L}$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento mínimo positivo debemos colocar el menor valor para cada posición:

$$Min_+ = 1.00...0 \times \beta^{-L} = \beta^{-L}$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

$$Max_+ = (\beta - 1).$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

$$Max_+ = (\beta - 1).(\beta - 1)$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

$$Max_+ = (\beta - 1).(\beta - 1)(\beta - 1)$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

$$Max_+ = (\beta - 1).(\beta - 1)(\beta - 1)...(\beta - 1)$$

Elementos mínimo y máximo positivos

Tomemos la notación de los elementos flotantes:

$$x = \pm d_1.d_2 d_3 ... d_t \times \beta^e$$

donde:
$0 < d_1 < \beta$
$0 \leq d_i < \beta$, para $i = 2, 3, ..., t$
$-L \leq e \leq U$

Para calcular el elemento máximo positivo debemos colocar el mayor valor para cada posición:

$$Max_+ = (\beta - 1).(\beta - 1)(\beta - 1)...(\beta - 1) \times \beta^U$$

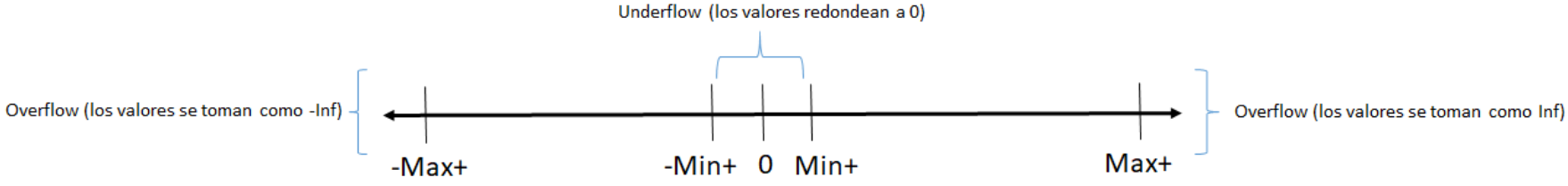


Figura: Underflow y Overflow

Ejemplo

Ejemplo

Para el conjunto flotante $fl(2, 2, 1, 1)$, calcule:

- *Todos los elementos del conjunto.*
- *Los elementos máximo y mínimo.*
- *Compare con las fórmulas que hemos deducido.*

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^0

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}
6. 1.1×2^{1}

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}
6. 1.1×2^{1}

Faltaría tomar a los negativos y al cero, así nos queda:

$$Card = 13$$

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}
6. 1.1×2^{1}

Faltaría tomar a los negativos y al cero, así nos queda:

$$Card = 13$$

De la lista de elementos: $Min_+ =$

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}
6. 1.1×2^{1}

Faltaría tomar a los negativos y al cero, así nos queda:

$$Card = 13$$

De la lista de elementos: $Min_+ = 1.0 \times 2^{-1}$, $Max_+ =$

Los elementos de este conjunto son de la forma:

$$\pm d_1.d_2 \times 2^e$$

así tenemos que los elementos del conjunto son:

1. 1.0×2^{-1}
2. 1.0×2^{0}
3. 1.0×2^{1}
4. 1.1×2^{-1}
5. 1.1×2^{0}
6. 1.1×2^{1}

Faltaría tomar a los negativos y al cero, así nos queda:

$$Card = 13$$

De la lista de elementos: $Min_+ = 1.0 \times 2^{-1}$, $Max_+ = 1.1 \times 2^{1}$

Verificamos con las fórmulas obtenidas:

$$Card = 2(\beta - 1)\beta^{t-1}(U + L + 1) + 1$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} \mathit{Card} &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \end{aligned}$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} \mathit{Card} &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \\ &= 13 \end{aligned}$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} Card &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \\ &= 13 \end{aligned}$$

$$Min_+ = 1.0 \times \beta^{-L}$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} Card &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \\ &= 13 \end{aligned}$$

$$\begin{aligned} Min_+ &= 1.0 \times \beta^{-L} \\ &= 1.0 \times 2^{-1} \end{aligned}$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} Card &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \\ &= 13 \end{aligned}$$

$$\begin{aligned} Min_+ &= 1.0 \times \beta^{-L} \\ &= 1.0 \times 2^{-1} \end{aligned}$$

$$Max_+ = (\beta-1).(\beta-1)(\beta-1)...(\beta-1) \times \beta^U$$

Verificamos con las fórmulas obtenidas:

$$\begin{aligned} Card &= 2(\beta-1)\beta^{t-1}(U+L+1)+1 \\ &= 2(1)(2^1)(3)+1 \\ &= 13 \end{aligned}$$

$$\begin{aligned} Min_+ &= 1.0 \times \beta^{-L} \\ &= 1.0 \times 2^{-1} \end{aligned}$$

$$\begin{aligned} Max_+ &= (\beta-1).(\beta-1)(\beta-1)...(\beta-1) \times \beta^{U} \\ &= 1.1 \times 2^1 \end{aligned}$$

Tarea 2

Tarea

Considera a los números de punto flotante $fl(10, 4, 4, 3)$.

1. *Encuentra la cardinalidad del conjunto.*
2. *Halla el elemento* α *positivo más chico.*
3. *Halla el elemento* ω *positivo más grande.*
4. *¿Cuántos elementos son enteros?*

Unidad de redondeo

Un valor muy importante en la Aritmética de Punto Flotante es la unidad de redondeo ya que con este valor podemos saber cual es el error que comete la computadora por efectos de redondeo.

Mantisa de un valor flotante

Definición

Mantisa de un valor flotante.

Dado un elemento flotante

$$x = \pm d_1.d_2 d_3 \ldots d_t \times \beta^e$$

se define a su mantisa como

$$m = \pm d_1.d_2 d_3 \ldots d_t$$

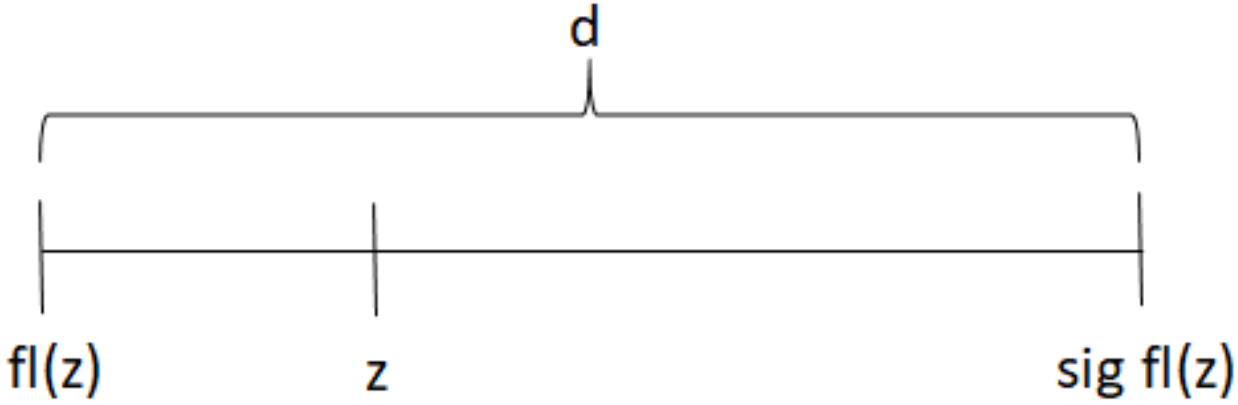
donde

$$1 \leq |m| < \beta$$

Tomemos un valor $z \in \mathbb{R}$ el cual se introduce a la computadora y debe redondearse a un valor $fl(z)$. Queremos estimar cual será el error entre z y $fl(z)$.

Tomemos un valor $z \in \mathbb{R}$ el cual se introduce a la computadora y debe redondearse a un valor $fl(z)$. Queremos estimar cual será el error entre z y $fl(z)$.

Observemos que z debe estar en el siguiente intervalo:



Una primera forma de estimar el error entre z y $fl(z)$ consiste en utilizar la distancia d del intervalo anterior.

Una primera forma de estimar el error entre z y $fl(z)$ consiste en utilizar la distancia d del intervalo anterior.

Notemos que, como z redondea a $fl(z)$ entonces a lo más podría estar en el punto medio del intervalo, es decir

$$|z - fl(z)| \leq \frac{d}{2}$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

así, podemos obtener el valor de d como

$$d \quad = \quad sig\ fl(z) - fl(z)$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

así, podemos obtener el valor de d como

$$\begin{aligned} d &= sig\ fl(z) - fl(z) \\ &= d_1.d_2d_3...(d_t + 1) \times \beta^e - d_1.d_2d_3...d_t \times \beta^e \end{aligned}$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

así, podemos obtener el valor de d como

$$\begin{aligned} d &= sig\ fl(z) - fl(z) \\ &= d_1.d_2d_3...(d_t + 1) \times \beta^e - d_1.d_2d_3...d_t \times \beta^e \\ &= 0.00...01 \times \beta^e \end{aligned}$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

así, podemos obtener el valor de d como

$$\begin{aligned} d &= sig\ fl(z) - fl(z) \\ &= d_1.d_2d_3...(d_t + 1) \times \beta^e - d_1.d_2d_3...d_t \times \beta^e \\ &= 0.00...01 \times \beta^e \\ &= \beta^{-(t-1)} \times \beta^e \end{aligned}$$

Donde, si tomamos a

$$fl(z) = d_1.d_2d_3...d_t \times \beta^e$$

entonces, sin pérdida de generalidad

$$sig\ fl(z) = d_1.d_2d_3...(d_t + 1) \times \beta^e$$

así, podemos obtener el valor de d como

$$\begin{aligned} d &= sig\ fl(z) - fl(z) \\ &= d_1.d_2d_3...(d_t + 1) \times \beta^e - d_1.d_2d_3...d_t \times \beta^e \\ &= 0.00...01 \times \beta^e \\ &= \beta^{-(t-1)} \times \beta^e \\ &= \beta^{e-(t-1)} \end{aligned}$$

Por tanto, como

$$|z - fl(z)| \leq \frac{d}{2}$$

Por tanto, como

$$|z - fl(z)| \leq \frac{d}{2}$$

entonces

$$|z - fl(z)| \leq \frac{1}{2}\beta^{e-(t-1)}$$

Por tanto, como

$$|z - fl(z)| \leq \frac{d}{2}$$

entonces

$$|z - fl(z)| \leq \frac{1}{2}\beta^{e-(t-1)}$$

observemos que esta cota de error está en función de t y e, sin embargo, el valor de e en el conjunto está entre $-L$ y U por lo cual es un valor variable.

En general, un valor $z \in \mathbb{R}$ puede escribirse como:

$$z = d_1.d_2 d_3 ... d_t d_{t+1} d_{t+2} ... \times \beta^e$$

es decir, puede tener una mantisa infinita.

En general, un valor $z \in \mathbb{R}$ puede escribirse como:

$$z = d_1.d_2d_3...d_td_{t+1}d_{t+2}... \times \beta^e$$

es decir, puede tener una mantisa infinita.

Sin embargo, la mantisa sigue cumpliendo con la propiedad

$$1 \leq |m| < \beta$$

En general, un valor $z \in \mathbb{R}$ puede escribirse como:

$$z = d_1.d_2d_3...d_td_{t+1}d_{t+2}... \times \beta^e$$

es decir, puede tener una mantisa infinita.

Sin embargo, la mantisa sigue cumpliendo con la propiedad

$$1 \leq |m| < \beta$$

entonces

$$|z| = |m \times \beta^e| = |m| \times \beta^e \geq$$

En general, un valor $z \in \mathbb{R}$ puede escribirse como:

$$z = d_1.d_2d_3...d_td_{t+1}d_{t+2}... \times \beta^e$$

es decir, puede tener una mantisa infinita.

Sin embargo, la mantisa sigue cumpliendo con la propiedad

$$1 \leq |m| < \beta$$

entonces

$$|z| = |m \times \beta^e| = |m| \times \beta^e \geq \beta^e$$

En general, un valor $z \in \mathbb{R}$ puede escribirse como:

$$z = d_1.d_2d_3...d_td_{t+1}d_{t+2}... \times \beta^e$$

es decir, puede tener una mantisa infinita.

Sin embargo, la mantisa sigue cumpliendo con la propiedad

$$1 \leq |m| < \beta$$

entonces

$$|z| = |m \times \beta^e| = |m| \times \beta^e \geq \beta^e$$

y asi

$$\frac{1}{|z|} \leq \beta^e$$

Por tanto, como sabiamos que

$$|z - fl(z)| \leq \frac{1}{2}\beta^{e-(t-1)}$$

Por tanto, como sabiamos que

$$|z - fl(z)| \leq \frac{1}{2}\beta^{e-(t-1)}$$

entonces

$$\frac{|z - fl(z)|}{|z|} \leq \frac{1}{2}\frac{\beta^{e-(t-1)}}{\beta^{e}}$$

Por tanto, como sabiamos que

$$|z - fl(z)| \leq \frac{1}{2}\beta^{e-(t-1)}$$

entonces

$$\frac{|z - fl(z)|}{|z|} \leq \frac{1}{2}\frac{\beta^{e-(t-1)}}{\beta^{e}}$$

$$\frac{|z - fl(z)|}{|z|} \leq \frac{1}{2}\beta^{-(t-1)}$$

Definición

Unidad de redondeo.

Se define a la unidad de redondeo de la CPU como una cota del error relativo entre z y $fl(z)$, la cual está dada por:

$$u = \frac{1}{2}\beta^{-(t-1)}$$

Observe también que, la expresión del error relativo entre z y $fl(z)$ induce una definición para $fl(z)$ en términos de z. Tomemos a δ como:

$$\delta = \frac{fl(z) - z}{z}$$

Observe también que, la expresión del error relativo entre z y $fl(z)$ induce una definición para $fl(z)$ en términos de z. Tomemos a δ como:

$$\delta = \frac{fl(z) - z}{z}$$

entonces

$$z\delta = fl(z) - z$$

Observe también que, la expresión del error relativo entre z y $fl(z)$ induce una definición para $fl(z)$ en términos de z. Tomemos a δ como:

$$\delta = \frac{fl(z) - z}{z}$$

entonces

$$z\delta = fl(z) - z$$

$$z\delta + z = fl(z)$$

Observe también que, la expresión del error relativo entre z y $fl(z)$ induce una definición para $fl(z)$ en términos de z. Tomemos a δ como:

$$\delta = \frac{fl(z) - z}{z}$$

entonces

$$z\delta = fl(z) - z$$

$$z\delta + z = fl(z)$$

$$fl(z) = z(1 + \delta)$$

donde

$$|\delta| \leq u$$

Tarea 3

Tarea

Tomando el hecho de que $fl(a) = a(1+\delta)$, *para algún* δ *con* $|\delta| \leq u$ *(con* u *la unidad de redondeo de la máquina), para cada uno de los casos siguientes:*

$$a) fl_{10}(e) = 2.71 \qquad b) fl_{10}(\sqrt{2}) = 1.4142$$

$$c) fl_{10}\left(\frac{1}{3}\right) = 0.3333333 \quad d) fl_{10}(\pi) = 3.1415926535$$

¿Cuánto vale δ*?, ¿cuál es la unidad de redondeo* u*? y ¿se cumple que* $|\delta| \leq u$*?*

Nota. Para calcular los valores de δ *toma los valores de* e, $\sqrt{2}$ *y* π *calculados con Matlab en formato largo (usa la instrucción format long para desplegar los valores en este formato).*