Computación I
Representación Interna

Curso 2025

Facultad de Ingeniería

Universidad de la República

Rep. de punto flotante
Temario

- Introducción
- Normalización
- Estándard IEEE 754
- Aritmética de Punto Flotante
- Errores de la Representación
  - Truncamiento
  - Redondeo

Rep. de punto flotante

Introducción

Necesidad de representar números reales con un rango de representación mayor que el que ofrece el punto fijo.

Notación “científica”...

$$n = \pm f * 10^{exp}$$

Se compone de tres partes:

- Signo
- Mantisa (f)
- Un entero positivo o negativo denominado exponente (exp).

Rep. de punto flotante
Introducción

Ejemplos en base 10:

$3.14 = 0.314 \times 10^{1} = 3.14 \times 10^{0}$

$0.000001 = 0.1 \times 10^{-5} = 1.0 \times 10^{-6}$

$1941 = 0.1941 \times 10^{4} = 1.941 \times 10^{3}$

Rep. de punto flotante
Introducción

La representación en punto flotante es la versión para computadoras de la notación científica utilizando base 2

$$n = \pm f * 2^{exp}$$

Ejemplo n = $1,0011010 * 2^{7}$

Solo se representa de manera física

- El signo
- La mantisa f
- El exponente exp

Rep. de punto flotante
Introducción

Representación utilizando n bits

- s es el bit de signo (0 positivo, 1 negativo)
- e es el exponente exp, representado con q bits en exceso a M ($M = 2^{q-1}-1$). $e = exp + 2^{q-1}-1$
- f es la mantisa, representada con p bits en binario.
- $1 + p + q = n$ (bits)

Rep. de punto flotante
Normalización

- Representación ambigua

Existen varias representaciones para un mismo número

  - $0.000001 = 0.1 * 2^{-5} = 1.0 * 2^{-6}$

- Representación normalizada

Versión mas restringida.

Rep. de punto flotante
Normalización

- Número de punto flotante normalizado:

El bit más significativo de la mantisa es un 1.

Los números normalizados proporcionan la máxima precisión posible para los números de punto flotante.

Rep. de punto flotante
Normalización

Consideremos el número binario n = 11010 * 2^0

Si lo representamos con 8 bits de mantisa, la mantisa es 00011010

No está normalizado: los tres primeros dígitos (dígitos más significativos) de la mantisa son 0s.

Rep. de punto flotante
Normalización

Para normalizarlo se desplaza la coma tres posiciones hacia la derecha y se descartan así los primeros ceros para obtener 11010000.

Los cambios realizados multiplicaron al número por 2^3

Para mantener el mismo valor debo restarle 3 al exponente $11010000 * 2^{-3}$

Rep. de punto flotante
Normalización

Todos los números normalizados tienen un 1 en el bit más significativo.

Se define una representación que omite este bit y solo almacena los dígitos después de la coma.

Rep. de punto flotante
Normalización

Esta representación consiste en un 1 implícito, una coma implícita y luego la mantisa

$$n = \pm 1,\mathbf{f} * 2^{exp}$$

Solo se representa de manera física

- El signo
- La mantisa f
- El exponente exp en exceso a M

Rep. de punto flotante
Normalización

Representar $67 * 2^{-7}$ en punto flotante de 16 bits:

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Cálculos

Mantisa:

- $67_{10} = 1000011_2 \rightarrow 1{,}000011 * 2^6$
- $67*2^{-7} = 1{,}000011 * 2^{-1}$

Exponente: 5 bits, representación en exceso a M,

- $M = 2^{5-1} - 1 = 15$
- $-1_{10} \rightarrow -1_{10} + 15_{10} = 14_{10} = 1110_2$

Rep. de punto flotante

Normalización

Representar $67 * 2^{-7}$ en punto flotante de 16 bits:

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Signo = 0

Mantisa (10 bits): 1,0000110000

Exponente (5 bits) 01110

0	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0

Rep. de punto flotante
Normalización

Representar $67 * 2^{-7}$ en punto flotante de 16 bits:

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Cálculos (otra forma)

Mantisa:

- $67_{10} = 1000011_2$

Exponente: 5 bits, representación en exceso a M,

- $M = 2^{5-1} - 1 = 15$
- $-7_{10} \rightarrow -7_{10} + 15_{10} = 8_{10} = 1000_2$

Rep. de punto flotante
Normalización

Representar $67 * 2^{-7}$ en punto flotante de 16 bits:

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

s = 0

f = 1000011,0

e = 1000

Para normalizar corro la coma 6 lugares a la izquierda:

f = 1,000011

Como dividí el número por 2^6 debo aumentar en 6 el exponente:

e= 1000 + 110 = 1110

Rep. de punto flotante

Normalización

Representar $67 * 2^{-7}$ en punto flotante de 16 bits:

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Signo = 0

Mantisa (10 bits): 1,0000110000

Exponente (5 bits) 01110

0	0	1	1	1	0	0	0	0	0	1	1	0	0	0	0

Rep. de punto flotante
Otra forma de verlo

Como un desplazamiento dentro de una ventana:

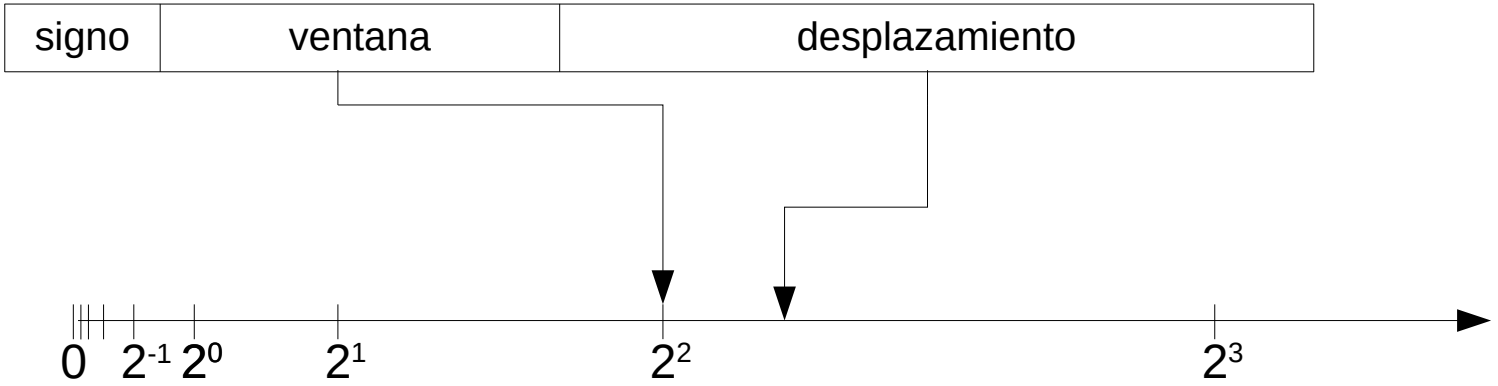
- El exponente define una ventana entre dos potencias de 2 contiguas
- La mantisa nos da una posición relativa dentro de esa ventana

signo	ventana	desplazamiento

Rep. de punto flotante
Otra forma de verlo

Si la mantisa es de m bits, cada ventana se divide en 2^m partes iguales.

- Cuanto más grande el exponente más grandes son estas partes… menos precisión.



Rep. de punto flotante

Normalización (otra forma)

Representar $67 * 2^{-7}$ en punto flotante de 16 bits

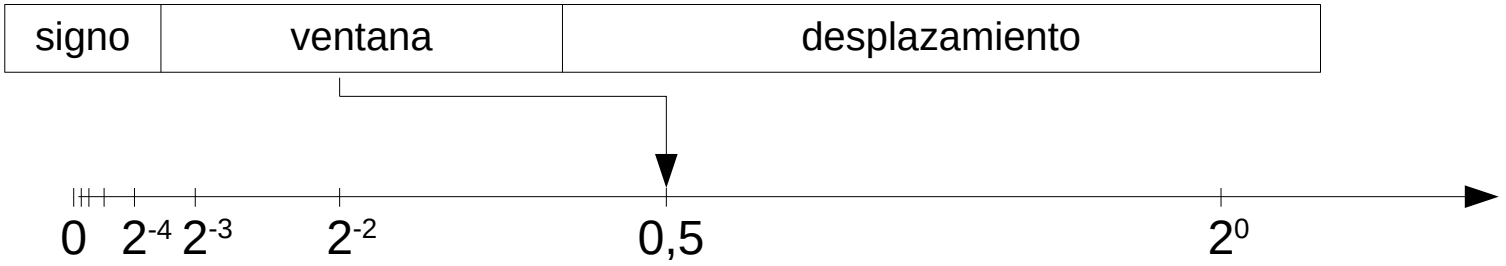
1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Cálculos

$67*2^{-7}{}_{10} = 0{,}5234375_{10}$

Está en la ventana que va de 2^{-1} a 2^{0} (exponente -1)

- $M = 2^{5-1} - 1 = 15$
- $-1_{10} \rightarrow -1_{10} + 15_{10} = 14_{10} = 1110_2$



Rep. de punto flotante
Normalización (otra forma)

Representar $67 * 2^{-7}$ en punto flotante de 16 bits

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Falta desplazarnos desde el 0,5 hasta 0,5234375…

Tenemos 10 bits, así que la ventana está dividida en 2^{10} números.

- Cada división mide $(1-0,5)/2^{10}=0,000488281$

Rep. de punto flotante

Normalización (otra forma)

Representar $67 * 2^{-7}$ en punto flotante de 16 bits

1 bit de signo | exponente de 5 bits | mantisa de 10 bits

Para calcular el desplazamiento dentro de la ventana puedo calcular

$$(0{,}5234375 - 0{,}5)/0{,}000488281 = 48$$

Obtengo la mantisa convirtiendo 48 a binario

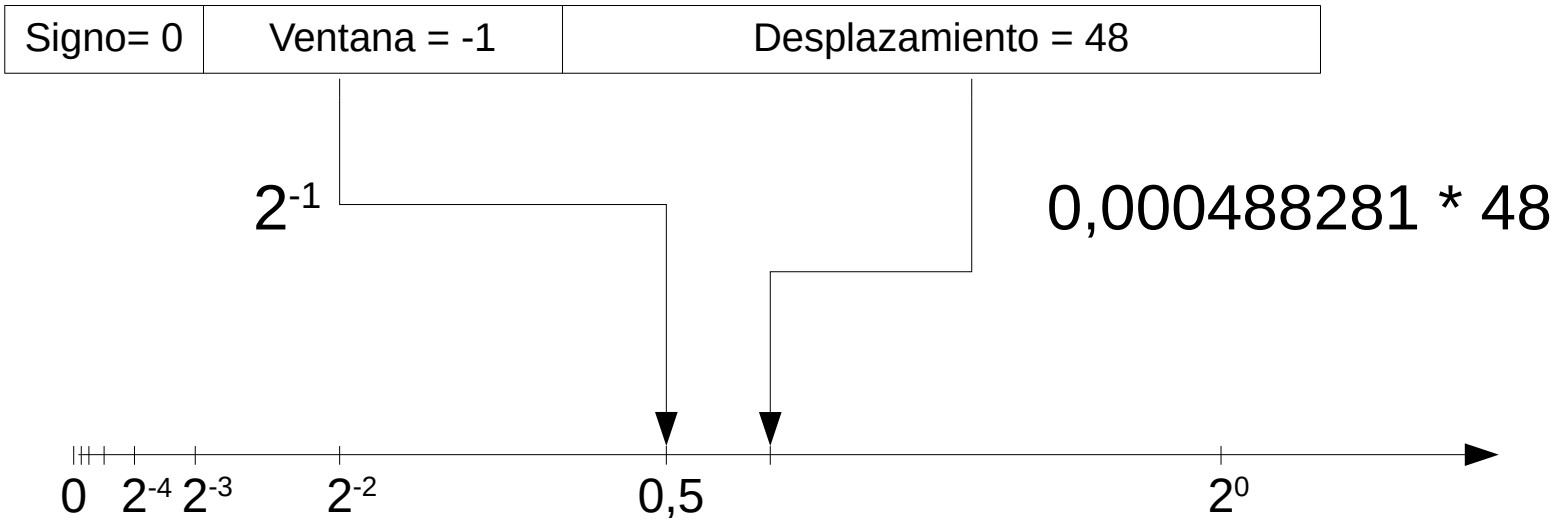
$48_{10} = 110000_2$

Rep. de punto flotante

Normalización (otra forma)

Representar $67 * 2^{-7}$ en punto flotante de 16 bits

1 bit de signo | exponente de 5 bits | mantisa de 10 bits



Rep. de punto flotante
Estándar IEEE 754

Existen infinitas formas de representar un número en punto flotante

- Cantidad de bits
- Representación del exponente
- Representación de la mantisa

Esto dificulta el intercambio de información entre distintas computadoras con arquitecturas diferentes.

Rep. de punto flotante
Estándar IEEE 754

Se establece el estándar IEEE 754

- Define el formato y las operaciones a utilizar.
- Los números representados en punto flotante se podrán intercambiar entre distintas arquitecturas.
- Es la representación de reales más común actualmente.

Rep. de punto flotante
Estándar IEEE 754

Primero se definen tres formatos

	s (bits)	e (bits)	F (bits)	Total (bytes)
simple precisión	1	8	23	4
doble precisión	1	11	52	8
precisión extendida	1	15	64	10

Estos tres formatos definen la cantidad de bits a utilizar en cada parte (mantisa, signo y exponente).

Rep. de punto flotante
Estándar IEEE 754

Luego se define como se representará cada parte:

- Signo
  - 1 bit (0 positivo, 1 negativo)
- Mantisa
  - Se representa como un binario puro
- Exponente
  - Se representa utilizando exceso a M
  - M se calcula como $2^{n-1} - 1$
  - $M = 2^{8-1} - 1 = 127$ para simple precisión
  - $M = 2^{11-1} - 1 = 1023$ para doble precisión)

Rep. de punto flotante
Estándar IEEE 754

Los números deberán estar normalizados

4 casos particulares:

- Desnormalizados
- Cero (desnormalizado)
- Infinitos
- Not a Number

Rep. de punto flotante
Estándar IEEE 754

Representación para Normalizados

$$n = \pm (1,f) * 2^{exp\ = e - M}$$

En simple precisión:

- e exponente exp en exceso a M con 8 bits (e≠0 y e ≠255)
- M = 127
- f mantisa en binario de 23 bits

En doble precisión:

- e exponente exp en exceso a M con 11 bits (e≠0 y e ≠2047)
- M = 1023
- f mantisa en binario de 52 bits

Rep. de punto flotante
Estándar IEEE 754

Surge un problema cuando el resultado de un cálculo tiene una magnitud menor que el número normalizado de punto flotante más pequeño que se puede representar en este sistema.

En particular no hay representación para el 0.

Por esta razón se crean los números desnormalizados.

Rep. de punto flotante
Estándar IEEE 754

Números Desnormalizados

Sirven para operar con números menores que el menor número normalizado representable:

- f = 00..00 23 ceros
- e = 00000001
- n = $\pm$ 1,0 * 2^{-126} en simple precisión

Rep. de punto flotante
Estándar IEEE 754

Representación

- Tienen un exponente de cero y una mantisa dada por los siguientes 23 o 52 bits.
- El bit implícito a la izquierda se convierte ahora en cero.
- Se distinguirán de los números normalizados porque los primeros no pueden tener un exponente cero

$$n = \pm (0,f) * 2^{e - 126} \quad (e \text{ es siempre cero})$$

Rep. de punto flotante
Estándar IEEE 754

- Representación para Cero

  (Como número desnormalizado…)

$$0 = \pm\ 0{,}0 * 2^{-126}$$

  - e = 00..00
    - Simple precisión 8 bits
    - Doble precisión 11 bits
  - f = 00..00
    - Simple precisión 23 bits
    - Doble precisión 52 bits

Rep. de punto flotante
Estándar IEEE 754

- Representación para Infinito
- Máximo (y mínimo) número normalizado representable.

$$\pm Inf = \pm\ 1{,}0 * 2^{128}$$

  - e = 11..11
    - Simple precisión 8 bits
    - Doble precisión 11 bits
  - f = 00..00
    - Simple precisión 23 bits
    - Doble precisión 52 bits

Rep. de punto flotante
Estándar IEEE 754

- Representación para Not a Number
- Sirve para representar el resultado de una operación no definida (ej. división por 0)

$$\pm NaN = \pm\ 1,x * 2^{128}$$

  - e = 11..11
    - Simple precisión 8 bits
    - Doble precisión 11 bits
  - f = x distinto de 00..00
    - Simple precisión 23 bits
    - Doble precisión 52 bits

Rep. de punto flotante
Estándar IEEE 754

Resumen de Representaciones

Número en Pto. Flotante	**e (Exponente)**	**f (Mantisa)**
Normalizados	0 < Exp < Max	Cualquier combinación de 1's y 0's
Desnormalizados	0000.....0	Cualquier combinación de 1's y 0's distinta de 0000.....0
Cero	0000.....0	0000.....0
Infinito	1111.....1	0000.....0
Not a Number	1111.....1	Cualquier combinación de 1's y 0's distinta de 0000.....0

Rep. de punto flotante
Estándar IEEE 754

Ejemplos utilizando simple precisión:

Normalizados: $n = \pm (1,f) * 2^{e-127}$

s	e	f	
0	10000000	00000000000000000000000	$= +1 \times 2^{128-127} \times 1.0$
1	10000001	10100000000000000000000	$= -1 \times 2^{129-127} \times 1.101$

Desnormalizados: $n = \pm(0,f)*2^{-126}$

s	e	f	
0	00000000	10000000000000000000000	$= +1 \times 2^{-126} \times 0.1$

Rep. de punto flotante
Estándar IEEE 754

Ejemplos utilizando simple precisión:

s	e	f	
0	00000000	00000000000000000000000	= 0
1	00000000	00000000000000000000000	= -0
0	11111111	00000000000000000000000	= Inf
1	11111111	00000000000000000000000	= -Inf
0	11111111	11001011100101100010010	= NaN
1	11111111	11010100101011101110011	= NaN

Rep. de punto flotante
Aritmética de Punto Flotante

Sumas y Restas

Para sumar o restar dos números en punto flotante es necesario que los exponentes sean iguales.

La operación de suma o resta se realiza del siguiente modo:

- Alinear las mantisas:
  - Se desplaza hacia la derecha la mantisa que tiene el exponente más pequeño tantos lugares como la diferencia entre los exponentes.
- Sumar o restar las mantisas.
- Normalizar el resultado

Rep. de punto flotante
Aritmética de Punto Flotante

Multiplicación

Para multiplicar dos números en punto flotante no es necesario alinear las mantisas.

La operación de multiplicar dos números expresados en punto flotante normalizados implica:

- Sumar los exponentes.
- Multiplicar las mantisas.
- Normalizar el resultado.

Rep. de punto flotante
Aritmética de Punto Flotante

División

Para dividir dos números en punto flotante no es necesario alinear las mantisas

Para llevar a cabo la división en punto flotante:

- Dividir la mantisa del numerador por la mantisa del denominador.
- Restar los exponentes.
- Normalizar el resultado.

Rep. de punto flotante
Errores de la representación

Los números reales son infinitos y densos.

Es imposible representarlos con una cantidad finita de bits.

Los errores son inevitables...

Rep. de punto flotante

Errores de la representación

Cada resultado de una operación debe ser aproximado por un número representable.

Las dos alternativas para aproximar el resultado son:

Redondeo

Truncamiento.

Rep. de punto flotante
Errores de la representación

Redondeo

Se elige como aproximación el número más cercano representable.

Truncamiento

Se corta el número cuando se exceda de la cantidad de dígitos representables

Rep. de punto flotante

Errores de la representación

Sea x = 2/3 = 0.6666666.... el número que se quiere representar

Utilizando redondeo = 0.667

Utilizando truncamiento = 0.666

Rep. de punto flotante
Errores de la representación

Consideremos una representación que

Utilice tres dígitos y signo para la mantisa

Donde el valor absoluto de la mantisa esté comprendido entre $0.1 \leq |f| < 1$ o cero

Utilice un exponente de dos dígitos y signo.

Trabajaremos en base 10 para simplificar los cálculos.

Rep. de punto flotante
Errores de la representación

Dividamos la recta real en siete regiones.

Números negativos menores que $-0.999 * 10^{99}$

Números negativos entre $-0.999 * 10^{99}$ y $-0.100 * 10^{-99}$

Números negativos entre $-0.100 * 10^{-99}$ y 0

Cero

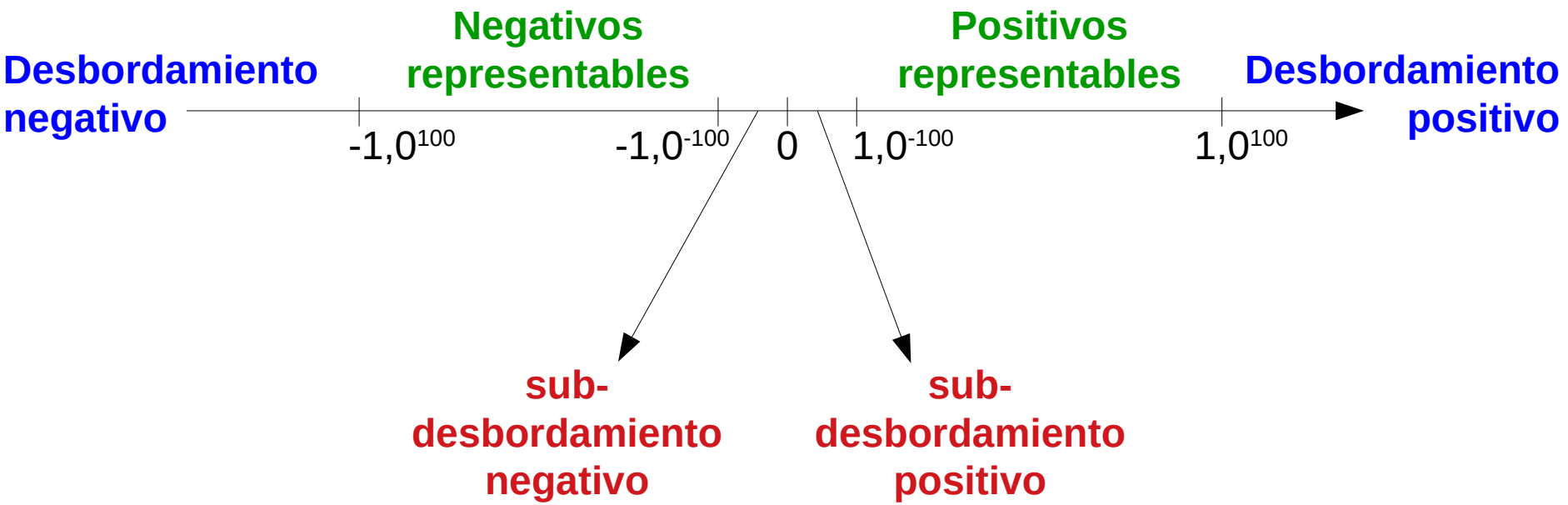
Números positivos entre 0 y $0.100 * 10^{-99}$

Números positivos entre $0.100 * 10^{-99}$ y $0.999 * 10^{99}$

Números positivos mayores que $0.999 * 10^{99}$

Rep. de punto flotante

Errores de la representación



Rep. de punto flotante

Errores de la representación

Diferencias entre el conjunto de los números representables en punto flotante y los números reales

Los primeros no pueden representar ningún número en las regiones 1,3,5 o 7.

- Si una operación aritmética diera como resultado un número en la región 1 o 7 se produciría un error de desbordamiento y el resultado sería incorrecto.
- La razón es la naturaleza finita de la representación.
- De manera similar no se puede representar ningún resultado de las zonas 3 o 5. Esto se llama error de subdesbordamiento (en inglés: underflow).

Rep. de punto flotante
Errores de la representación

Diferencias en densidad

- Mientras que los reales son densos, los números en punto flotante no lo son
- Con la representación elegida se pueden representar exactamente 179000 números positivos, 179000 números negativos y el 0, dando un total de 358201.
- Es posible que el resultado de alguna operación no caiga dentro de estos números aunque sí pertenezca a la región 2 o 6.

Rep. de punto flotante

Errores de la representación

Al aproximar un número se comete un error.

Llamaremos Error Absoluto a la diferencia entre el número que se quiere representar y el número efectivamente representado.

Error absoluto: $Ex = (x - \bar{x})$

Rep. de punto flotante
Errores de la representación

El espacio entre números adyacentes expresables no es constante a lo largo de las regiones 2 o 6.

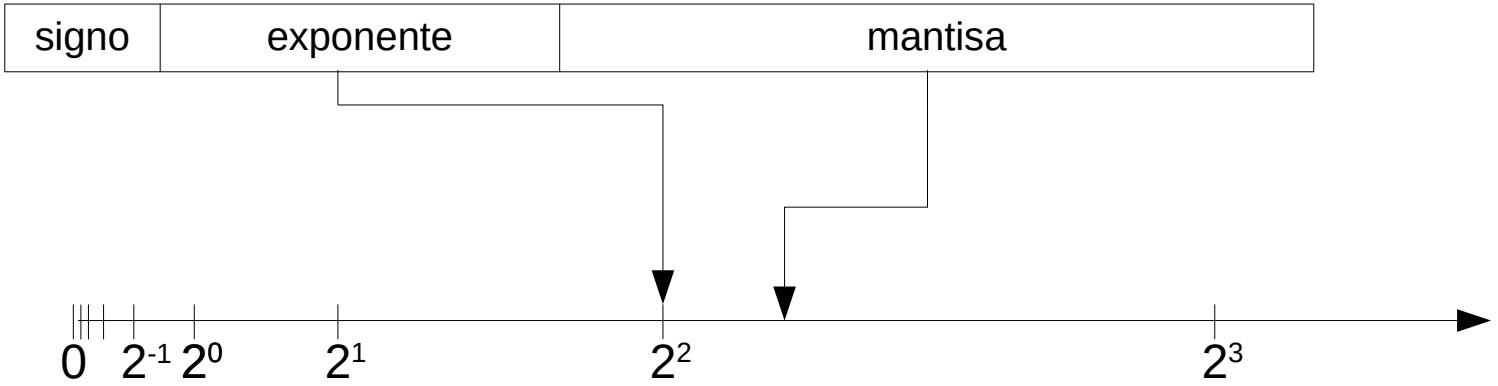
La separación entre $0.998 * 10^{99}$ y $0.999 * 10^{99}$ es muchísimo mayor que la separación entre

$+0.998 * 10^{0}$ y $0.999 * 10^{0}$

Rep. de punto flotante
Errores de la representación

Si la mantisa es de m bits, hay 2^m números entre un exponente y otro

- Cuanto más grande el exponente más separados están estos números en la recta real… menos precisión.



Rep. de punto flotante

Errores de la representación

Por esta razón se define el Error Relativo debido a la aproximación.

Error relativo: $ex = (x - \bar{x})/|x|$