ALN – Errores

In. Co.
Facultad de Ingeniería
Universidad de la República

Temario

- Introducción
- Representación de punto fijo
- Representación de punto flotante
  - Normalización
  - IEEE 754
- Errores de la Representación

Introducción

- Objetivo de los métodos numéricos: llegar a la solución “correcta” (con un nivel de error aceptable) en forma “eficiente” (en un tiempo razonable usando los recursos disponibles).

Introducción

- Estudio de la propagación de errores
- Estudio del número de operaciones

Introducción

Problema físico

↓

Modelo teórico

↓

Problema matemático

↓

Modelo numérico

↓

Solución numérica del problema matemático

Introducción

- Errores de truncamiento
  - Debidos a las aproximaciones realizadas al resolver el modelo numérico
- Errores de redondeo
  - Debidos al número finito de dígitos con los que se trabaja en la computadora

Representación

- Problema: representar números reales en una computadora.
  - R es un cuerpo infinito y denso mientras que la computadora debe representar los números con un número finito de dígitos.
  - Alternativas más comunes:
    - Representación de punto fijo
    - Representación de punto flotante

Representación de punto fijo

- Representa los números reales como enteros divididos por cierto factor de escala
  - Ej: entero: 12340 escala:1000 número: 12.340
- Algunos procesadores simples o microcontroladores no poseen unidades de punto flotante
- Suelen utilizarse dos tipos:
  - Decimal
  - Binario

Representación de punto fijo

- Los tamaños usuales para representar los enteros sin signo son:
  - El byte (0 a 255)
  - La palabra de 2 bytes (16 bits, 0 a 2^{16} -1)
  - La palabra de 4 bytes (32 bits, 0 a 2^{32} -1)

Tipo	Sin signo
1 byte	255
2 bytes	65.535
4 bytes	4.294.967.295
8 bytes	18.446.744.073.709.551.615

Rep. de punto flotante
Introducción

- Necesidad de representar números reales y enteros con un rango de representación mayor que el que ofrece el punto fijo.
- Notación científica utilizada en física, química y matemática.
  - $n = \pm f * 10^{exp}$
- Se compone de tres partes
  - Signo
  - Mantisa (f) .
  - Un entero positivo o negativo denominado exponente (exp).

Rep. de punto flotante
Introducción

- La representación en punto flotante es la versión para computadoras de la notación, científica utilizando base 2
  - $n = \pm f * 2^{exp}$
  - Ejemplo n = 00011010 * 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 ($M = 2^{q-1}$)
    - $e = exp + 2^{q-1}$
    - exp varía entre -2^{q-1} y $2^{q-1}-1$
  - f es la mantisa, representada con p bits en binario.
  - 1 + p + q = n (bits)

Rep. de punto flotante
Normalización

- Número de punto flotante normalizado
  - El dígito más significativo de la mantisa es diferente de cero o lo que es equivalente, la mantisa es máxima.
  - 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

- Todos los números normalizados tienen un 1 en el bit más significativo
- Se define una representación que omite este bit y solo representa la porción después de la coma.

Rep. de punto flotante
Normalización

- Esta representación consiste en un 1 implícito, una coma implícita y luego la mantisa
- $n = \pm (1,f) * 2^{exp}$
- Solo se representa de manera física
  - El signo
  - La mantisa f
  - El exponente exp

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

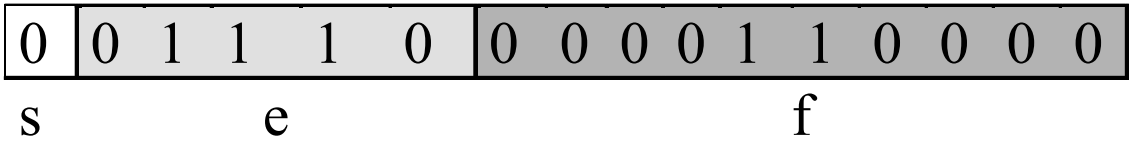
Rep. de punto flotante
Normalización

s = 0

f = 000011

e = 01110

Representación:



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 una de estas partes.

  - Signo
    - 1 bit de signo (0 positivo, 1 negativo)
  - Mantisa
    - Se representa como un binario puro

Rep. de punto flotante
Estándar IEEE 754

- Exponente
  - Se representa utilizando exceso a M
  - M se calcula como $2^{n-1}–1$ (Utilizan un cálculo diferente del habitual 2^{n-1}).
    - $M = 2^{8-1} – 1 = 127$ para simple precisión
    - $M = 2^{11-1} – 1 = 1023$ para doble precisión)
- Los números deberán estar normalizados
- Casos particulares
  - Cero
  - Desnormalizados
  - 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

- Representación para Cero
  - No puede ser normalizado.
  - Representación particular
    - $0 = \pm f * 2^{exp = e}$
    - e = 00..00
      - Simple precisión 8 bits
      - Doble precisión 11 bits
    - f = 00..00
      - Simple precisión 23 bits
      - Doble precisión 56 bits

Rep. de punto flotante
Estándar IEEE 754

- Representación para Infinito
  - Se reserva el máximo numero normalizado representable, para representar el infinito.
    - F = 0000..00
      - 23 bits en simple precisión
      - 56 bits en doble precisión
    - e = 1111...11
      - 8 bits en simple
      - 11 bits en doble precisión
    - s
      - 1 infinito negativo
      - O infinito positivo

Rep. de punto flotante
Estándar IEEE 754

- Representación para Not a Number
  - f mantisa en binario (distinta de 000.....00)
    - 23 bits en simple precisión
    - 52 bits en doble precisión
  - e = 1111...11
    - 8 bits en simple
    - 11 bits en doble precisión
  - s
    - 1 Nan
    - 0 Nan

Rep. de punto flotante
Estándar IEEE 754

- Representación para Desnormalizados
  - 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.
  - 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.
  - En simple precisión el menor número normalizado representable es n = $\pm$ 1,0 * 2^{-126}
    - f = 00..00 23 ceros
    - e = 00000001

Rep. de punto flotante
Estándar IEEE 754

- Números Desnormalizados
  - $n = \pm (0,f) * 2^{-126}$ (simple precisión)
    - e = 000…000
    - Tienen un exponente de cero y una mantisa dada por los siguientes 23 o 52 bits.
    - Se distinguirán de los números normalizados porque los primeros no pueden tener un exponente cero
    - El bit implícito a la izquierda se convierte ahora en cero.

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:

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
Estándar IEEE 754

- 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
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

Errores de la representación

- Mediante la representación en punto flotante se representan los números reales.
- Pero existen algunas diferencias importantes.

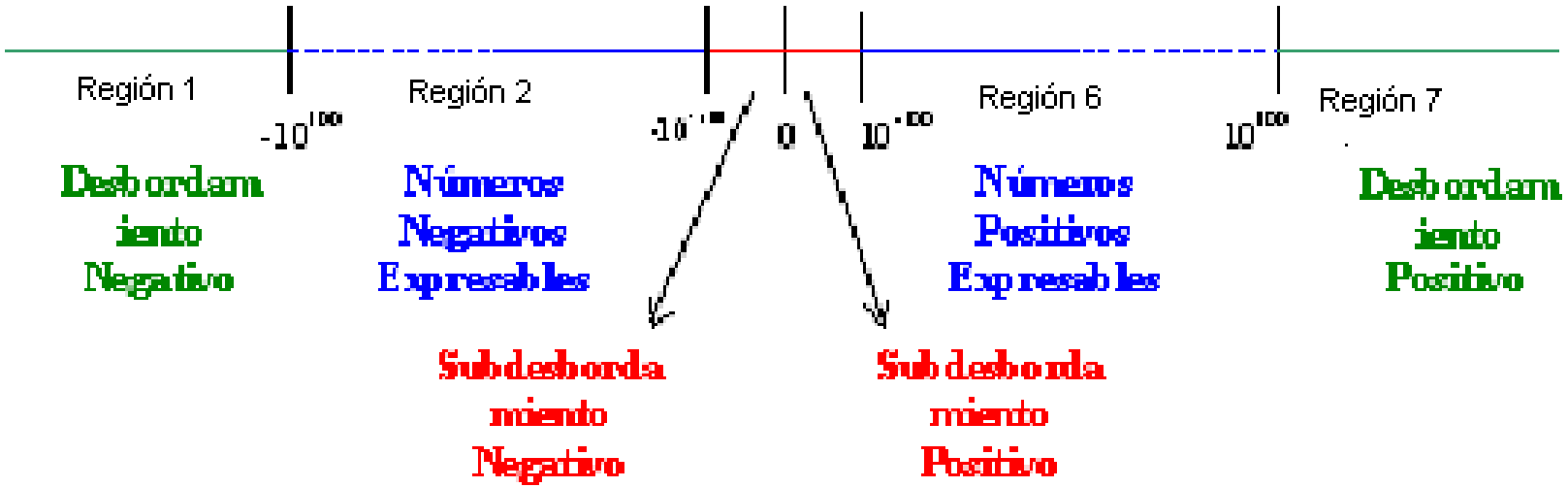
Errores de la representación

- 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.

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 $0.999 * 10^{99}$

Errores de la representación



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 (underflow).

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.

Errores de la representación

- Si el resultado de una operación no se puede expresar en la representación elegida, se debe aproximar este resultado a un número representable.
- Las dos alternativas para aproximar el resultado son
  - Redondeo
  - Truncamiento.

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: $Ex = x - \bar{x}$

Errores de la representación

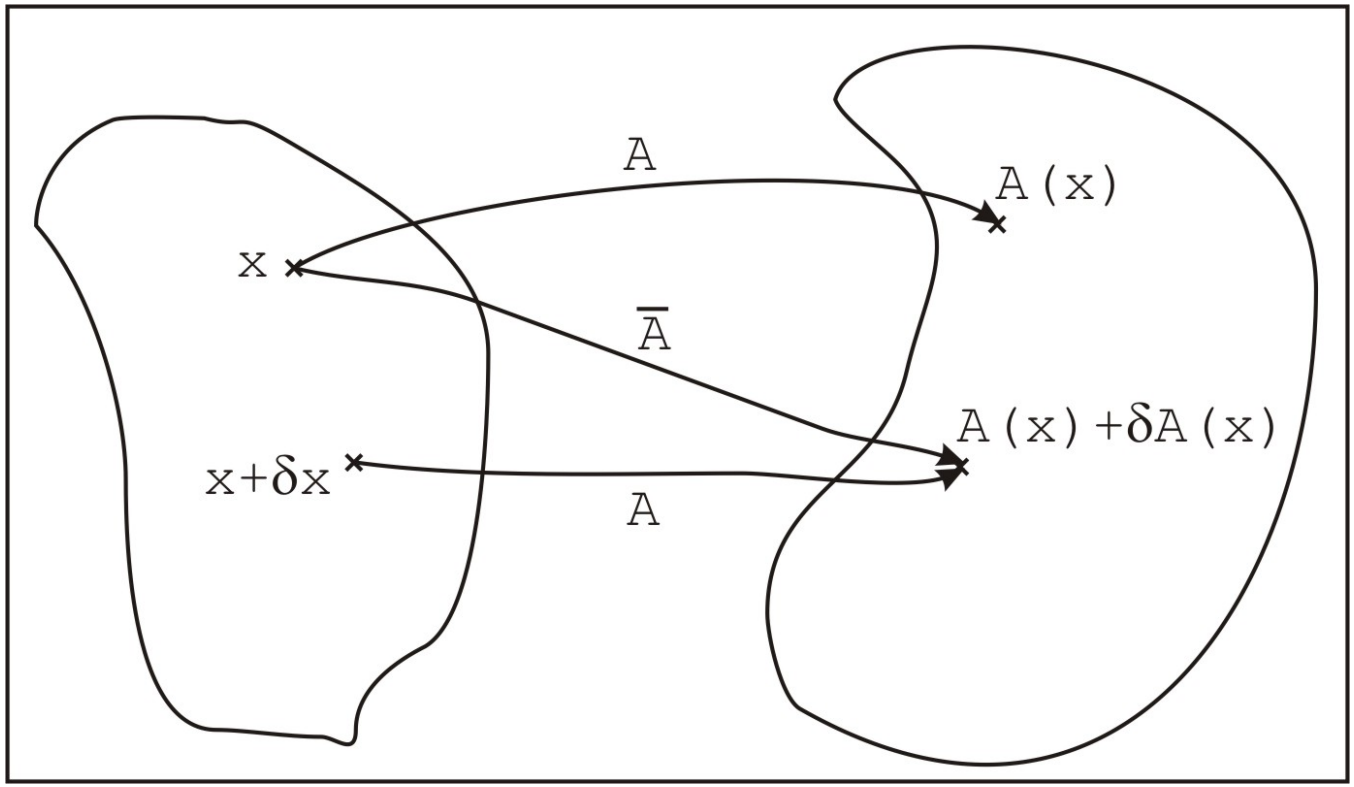
Errores de la representación

- Epsilon de la máquina (machine's epsilon):
  - Es una cota del error relativo que se comete al redondear un número en determinada representación de punto flotante.
  - También se puede expresar como donde p es la precisión (número de digitos de la mantisa) y b es la base (típicamente base 2).
    - Ej: Simple precisión, p = 24 bits, eps=2^{-24} =5.96e-08

Errores de la representación

- Not a Number (NaN)
- Overflow $x > \max_{z \in FP} |z|$
- Underflow $x < \min_{z \in FP} |z|$
- Cancelación catastrófica
  - Restar dos números grandes pero cercanos
- Shift out
  - Sumar un número pequeño a uno grande

Número de condición de un algoritmo



A(x) es el valor exacto calculado con el algoritmo A.

$\overline{A}(x)$ es el valor calculado con una máquina.

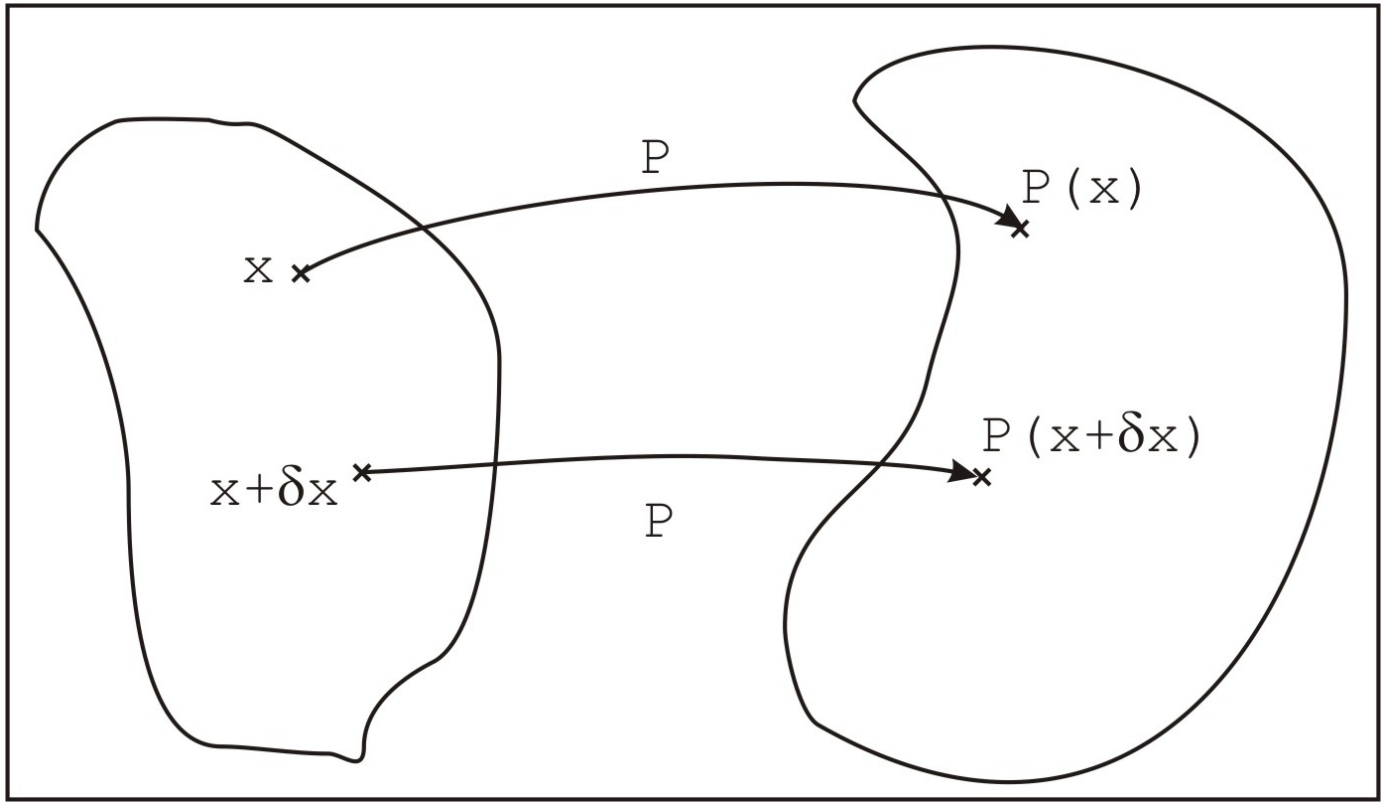
Número de condición de un algoritmo

$$C_{A(x)} = \frac{\|\delta x\|}{\|x\|\varepsilon_{mach}} = \frac{\left\|A^{-1}(\overline{A}(x)) - x\right\|}{\|x\|\varepsilon_{mach}}$$

Número de condición de un algoritmo

- Reflejar los errores “hacia atrás”, hacia los datos
- Incide la máquina por el ε_{Mach}

Número de condición del problema



P(x) es la solución exacta del problema con dato x.

$P(x+\delta x)$ es la solución exacta del problema con dato $x+\delta x$

Número de condición del problema

$$C_{P(x)} = \max_{\frac{|\delta x|}{|x|} \leq \varepsilon} \frac{\frac{\|P(x+\delta x) - P(x)\|}{\|P(x)\|}}{\frac{\|\delta x\|}{\|x\|}}$$

Número de condición del problema

- Estimar la sensibilidad relativa de los resultados respecto a los datos
- Independiente de los errores numéricos
- Independiente del algoritmo utilizado