Representação dos números no computador
Precisão da máquina

Algoritmos Numéricos - Topico 1
Computação numérica e Erros

Prof. Cláudia G. Varassin DI/UFES
emails:claudia.varassin@ufes.br e galarda@inf.ufes.br

Setembro 2020

Sumário

1. Representação dos números no computador
2. Precisão da máquina

Representando os números no computador

A unidade fundamental para se representar uma informação em um computador é chamada de palavra.

A palavra é um agrupamento dígitos.
Para os valores numéricos, há uma representação para os números inteiros e uma outra para os números com parte fracionária.

Vamos, em um primeiro momento entender como seria se em cada "posição" da palavra fosse possível armazenar um dígito da base decimal (valores: de 0 a 9).

Mais adiante, vamos ver como realmente isso ocorre, isto é, quando cada posição desta palavra armazena uma informção simples bem simples (0:ZERO/ 1:UM) ou (LIGADO ou DESLIGADO).

A palavra é, na verdade, um agrupamento de dígitos binários, um agrupamento de bits, de BInary digiTS. Um agrupamento de Os e 1s que permite, via combinações, representar todo tipo de informação.

Representando os números inteiros no computador

Para armazenar os inteiros, é preciso armazenar o seu sinal e os dígitos do número:

sinal	Dígitos do número

Se na "palavra" houver 11 " espaços " então um será para o sinal e os outros 10 reservados para armazenar os dígitos do número.

Ex 1: valor K = 444 nesta palavara

+	0| 0| 0| 0| 0| 0| 0 | 4| 4| 4

Exe 2: O valor L= -1234567891 nesta palavra

-	1| 2| 3| 4| 5| 6| 7| 8 | 9| 1

Representando os números não inteiros no computador

Para armazenar os valores reais (inteiros e não inteiros) usa-se a representação em ponto flutuante normalizada

Nesta abordagem o número é expresso através de uma parte fracionária, chamada de mantissa e uma parte inteira que corresponde ao expoente do número mais o sinal do número.

$$x = \pm 0.d_1 d_2 \cdots d_p \times B^{\,e}$$

Exemplo:$x = 12.345$

$$x = +0.12345 \times 10^{+2}$$

Um número $x \in \mathbb{R}$ armazenado neste sistema tem seguinte formato:

$$x = \pm 0.d_1 d_2 \cdots d_p \times B^{\,e}$$

onde

$$\begin{array}{rcl} B & : & \text{base empregada (ex: base 10, base 2)} \\ d_{i's} & : & \text{dígitos da parte fracionária (mantissa)} \\ p & : & \text{qte de dígitos na mantissa,} \\ & & d_1 \neq 0, 0 \leq d_i \leq B-1, i = 2, \cdots, p \\ e & : & \text{expoente inteiro} \\ \pm & : & \text{sinal do número} \end{array}$$

Um sistema de ponto flutuante é comumennte representado por

$$F = F(B, p, e_1, e_2),$$

onde $p =$ é a quantidade de dígitos da mantissa.
onde $e_1, e_2 =$ menor e maior expoente possíveis de serem armazenados.

Outros exemplos

$$\pm 0.d_1 d_2 \cdots d_p \times B^{\,e}$$

a)
$y = 12345000000$

$$y = +0.12345 \times 10^{\,+11}$$

b)
$z = 0.00000012345$

$$z = +0.12345 \times 10^{\,-6}$$

Sistema de representação em ponto flutuante

Haverá na "palavra" um " espaço " reservado para o sinal, outro para o expoente do número (e o seu sinal) e o restante para para armazenar os dígitos do número.

Sinal	Expoente (com o sinal)	Dígitos do valor

Ex 1:

$x = 12.345 \rightarrow + 0.123450000\ldots.0 * 10^{(2)}$

+	+|0 | 0 | 0 | 0| 0| 0 | 2	1 | 2 | 3 | 4 | 5 | 0 | 0 | ... ... |0 |0

Mais exemplos:

Suponha uma máquina que opere na base $B = 10$, com mantissa de $p = 5$ dígitos, e expoentes $e_1 = -9, e_2 = +9$

$F(B = 10, p = 5, e_1 = -9, e_2 = +9)$

Os valores, abaixo, serão representados, nesta máquina, por:

Valor real	Valor armazenado(x_{maq})
x	$\pm 0.d_1d_2d_3d_4d_5 \times 10^e$
12.345	$+.12345 \times 10^{+2} \rightarrow 12.345$
12.3	$+.12300 \times 10^{+2} \rightarrow 12.3$
0.12345	$+.12345 \times 10^{0} \rightarrow 0.12345$

Nesta máquina ($F(B = 10, p = 5, e_1 = -9, e_2 = +9)$) os valores serão representados por

x	x_{maq}
x	$\pm 0.d_1 d_2 d_3 d_4 d_5 \times 10^{exp}$
-151.112	$-.15111 \times 10^{+3} \rightarrow -151.11$
$0.33333....$	$+.33333 \times 10^{0} \rightarrow 0.33333$
5432000	$+.54320 \times 10^{+7} \rightarrow 5432000$
5432000000	$+.54320 \times 10^{+10} \rightarrow$ OVERFLOW!!
0.000000012345	$+.12345 \times 10^{-7} \rightarrow$
0.000000000012345	$+.12345 \times 10^{-11} \rightarrow$ underFLOW!!

Nesta máquina $F(B = 10, p = 5, e_1 = -9, e_2 = +9)$

Nos positivos:

O menor valor positivo (x_{min}) : $+0.10000 \times 10^{-9}$

O maior valor positivo (x_{max}) :$+0.99999 \times 10^{9}$

Visualmente (Overflow e Underflow)

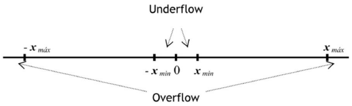


Figura 1.1. Regiões de *underflow* e *overflow*.

Simétrico nos negativos

O menor (em módulo) valor negativo ($-x_{min}$): -0.10000×10^{-9}

O maior (em módulo) valor negativo ($-x_{max}$):-0.99999×10^{9}

Sistema de Ponto Flutuante

- Intervalo I: $(-x_{max}, -x_{min}) \cup \{0\} \cup (x_{min}, x_{max})$.
- Existe apenas um quantidade finita de valores que podem ser representados na faixa "representável" (em I).
- A MAIORIA dos valores que precisam ser respresentados serão armazenados via uma aproximação!.
- Assim haverá (na maioria das vezes) ARREDONDAMENTOS, causando os erros de arredondamentos

Estratégias de arredondamento

- ESTRATÉGIA de arredondamento por CORTE (*chopping*) (ou ainda por falta):
  Despreza se todos os dígitos que não cabem na mantissa, a partir do $(p+1)$ *ésimo* dígito.
  ex: Se $F(B = 10, p = 5, e_1, e_2)$
  $12.3456 \rightarrow +.12345 \times 10^{+2} \rightarrow 12.345$

Estratégias de arredondamento

- ESTRATÉGIA de arredondamento por CORTE (*chopping*) (ou ainda por falta):
  Despreza se todos os dígitos que não cabem na mantissa, a partir do $(p+1)$ *ésimo* dígito.
  ex: Se $F(B=10, p=5, e_1, e_2)$
  $12.3456 \rightarrow +.12345 \times 10^{+2} \rightarrow 12.345$
- ESTRATÉGIA de arredondamento para o MAIS PRÓXIMO
  Verifica se o valor do $(p+1)$ *ésimo* dígito
  SE o $(p+1)$ *ésimo* $>=5$ soma se 1 ao p *ésimo* dígito e despreza-se todos os demais (a partir do $p+1$ *ésimo*)
  SENÃO despreza se todos os demais dígitos (a partir do $(p+1)$ *ésimo*)
  ex: $12.3458 \rightarrow +.12346 \times 10^{+2} \rightarrow 12.346$

Mais exemplos:

Suponha que se tenha uma máquina, que opere na base $B = 10$, com mantissa de $p = 2$ dígitos, e expoentes $e_1 = -9, e_2 = +9$ e arredondamento para o mais próximo.

$F(10, 2, -9, +9)$

Os valores serão representados, nesta máquina, via:

Valor real	Valor armazenado(x_{maq})
x	$\pm 0.d_1 d_2 \times 10^e$
1.72	$+.17 \times 10^{+1} \to 1.7$
1.713	$+.17 \times 10^{+1} \to 1.7$

Mais exemplos:

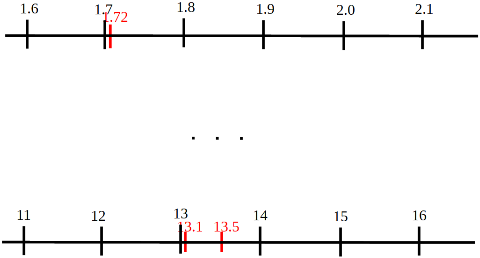
Suponha que se tenha uma máquina, que opere na base $B = 10$, com mantissa de $p = 2$ dígitos, e expoentes $e_1 = -9, e_2 = +9$ e arredondamento para o mais próximo.

$F(10, 2, -9, +9)$

Os valores serão representados, nesta máquina, via:

Valor real	Valor armazenado(x_{maq})
x	$\pm 0.d_1 d_2 \times 10^e$
1.72	$+.17 \times 10^{+1} \rightarrow 1.7$
1.713	$+.17 \times 10^{+1} \rightarrow 1.7$
1.78	$+.18 \times 10^{+1} \rightarrow 1.8$
1.7632	$+.18 \times 10^{+1} \rightarrow 1.8$
1.75	$+.18 \times 10^{+1} \rightarrow 1.8$

$$1.72 \rightarrow +.17 \times 10^{+1} \rightarrow 1.7$$
$$13.1 \rightarrow +.13 \times 10^{+1} \rightarrow 13$$



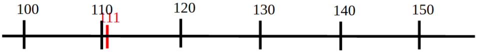
Continuando com $F(10, 2, -9, +9)$. Mais exemplos:

Valor real	Valor armazenado(x_{maq})
13.1	$+.13 \times 10^{+2} \rightarrow 13$
13.617	$+.14 \times 10^{+2} \rightarrow 14$
14.3	$+.14 \times 10^{+2} \rightarrow 14$
13.99	$+.14 \times 10^{+2} \rightarrow 14$

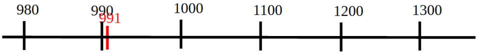
Continuando com $F(10, 2, -9, +9)$. Mais exemplos:

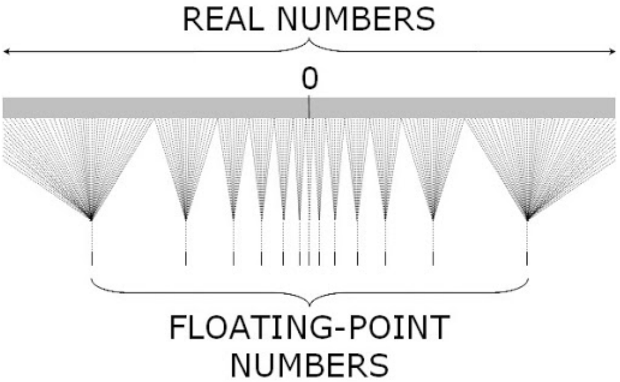
Valor real	Valor armazenado(x_{maq})
13.1	$+.13 \times 10^{+2} \to 13$
13.617	$+.14 \times 10^{+2} \to 14$
14.3	$+.14 \times 10^{+2} \to 14$
13.99	$+.14 \times 10^{+2} \to 14$
136.12	$+.14 \times 10^{+3} \to 140$
144.12	$+.14 \times 10^{+3} \to 140$
148.12	$+.14 \times 10^{+3} \to 150$
0.000036	$+.36 \times 10^{-4} \to 0.000036$
0.0000364	$+.36 \times 10^{-4} \to 0.000036$
0.000037	$+.37 \times 10^{-4} \to 0.000037$

$$110 \to +.11 \times 10^{+3} \to 110$$
$$111 \to +.11 \times 10^{+3} \to 110$$



. . .




REAL NUMBERS
0
FLOATING-POINT
NUMBERS

O erro na aproximação pela máquina
$F(B = 10, 2, e_1, e_2) \rightarrow \pm 0.d_1 d_2 \times 10^{\text{ e}}$

x=1.74

Erro verdadeiro (em módulo) é:

$$|E_v| = |ValorReal - ValorRepresentado| = 1.74 - 1.7 = 0.04$$

Erro relativo (em módulo):

$$|E_{rel}| = |\frac{0.4}{1.74}| = 0.0229 \Rightarrow 2,29\%$$

O erro na aproximação pela máquina

x=174

Erro verdadeiro (em módulo) é:

$$|E_v| = |ValorReal - ValorRepresentado| = |174 - 170| = 4$$

$$|E_{rel}| = |\frac{4}{174}| = 0.0229 \Rightarrow 2,29\%$$

O erro contido na aproximação pela máquina em

x=0.000 0174

Erro verdadeiro (em módulo) é:

$$|E_v| = |ValorReal - ValorRepresentado| = |0.0000174 - 0.000017| =$$

$$|E_{rel}| = |\frac{0.0000004}{0.0000174}| = 0.0229 \Rightarrow 2,29\%$$

Precisão da máquina

O sistema de ponto flutuante permite que a precisão seja a mesma independente da grandeza do número.

A precisão (ϵ) da máquina corresponde ao erro relativo máximo (em módulo) que a máquina pode cometer ao representar os valores reais.

$$|E_{rel}| = \left|\frac{ValorReal - ValorRepresentado}{ValorReal}\right| <= \epsilon$$

Torna se uma forma versátil para a representação de valores tão distintos! Fazendo uma analogia: se fosse necessário medir distâncias o sistema comporta-se como um micrômetro, uma trena, um teodolito e assim por diante dependendo da grandeza

RESUMINDO

(1) Os valores reais são representados no computador usando o sistema de ponto flutuante e ao serem armazenados sofrem erros de arredondamento.

(2) A precisão (ϵ) do computador corresponde ao erro relativo máximo (em módulo) que a máquina pode cometer ao representar os valores reais.

$$|E_{rel}| = |\frac{ValorReal - ValorRepresentado}{ValorReal}| <= \epsilon$$

Esta precisão depende do tamanho da mantissa.

Bibliografia Básica

[2] Métodos Numéricos para Engenharia, Steven C. Chapa e Raymond P. Canale, Ed. McGraw-Hill, 5ª Ed., 2008.

[1] Algoritmos Numéricos, Frederico F. Campos, Filho - 2ª Ed., Rio de Janeiro, LTC, 2007.

[3] Cálculo Numérico - Aspectos Teóricos e Computacionais, Márcia A. G. Ruggiero e Vera Lúcia da Rocha Lopes, Ed. Pearson Education, 2ª Ed., 1996.