Cálculo Numérico

Aritmética de Ponto Flutuante e Noções de Erro

Ana Paula

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

UFMG

Sumário

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Introdução

Introdução

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- O objetivo aqui é estudar métodos numéricos
- Logo, é importante entender como os números são representados no computador e como as operações aritméticas são realizadas
  - Limitações da representação finita
  - Determinar os casos em que erros ocorrem
- Noções de erro
  - Efeitos numéricos
  - Cancelamento
  - Propagação do erro

Introdução

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- O objetivo aqui é estudar métodos numéricos
- Logo, é importante entender como os números são representados no computador e como as operações aritméticas são realizadas
  - Limitações da representação finita
  - Determinar os casos em que erros ocorrem
- Noções de erro
  - Efeitos numéricos
  - Cancelamento
  - Propagação do erro

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Sistemas de Numeração

Sistema Decimal

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- O sistema decimal é normalmente adotado
  - Dez dígitos são utilizados para representar os números
    - base 10
  - Sistema posicional
- Qualquer número inteiro no sistema decimal pode ser representado como

$$\begin{aligned} N &= (a_n a_{n-1} \ldots a_1 a_0)_{10} \\ &= a_n \times 10^n + a_{n-1} \times 10^{n-1} + \cdots + a_1 \times 10^1 + a_0 \times 10^0 \end{aligned}$$

onde $a_i \in \{0, 1, \ldots, 8, 9\}$

Sistema Decimal

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Por exemplo

$$(21)_{10} = 2 \times 10^1 + 1 \times 10^0$$

$$(2001)_{10} = 2 \times 10^3 + 0 \times 10^2 + 0 \times 10^1 + 1 \times 10^0$$

Sistema Decimal

- Por exemplo

$$(21)_{10} = 2 \times 10^1 + 1 \times 10^0$$

$$(2001)_{10} = 2 \times 10^3 + 0 \times 10^2 + 0 \times 10^1 + 1 \times 10^0$$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Sistema Binário

- Os computadores adotam um sistema com dois estados
- Sistema binário
  - base 2
- Também é posicional
- Os números não negativos podem ser representados como

$$\begin{aligned} N &= (a_n a_{n-1} \ldots a_1 a_0)_1 \\ &= a_n \times 2^n + a_{n-1} \times 2^{n-1} + \cdots + a_1 \times 2^1 + a_0 \times 2^0 \end{aligned}$$

onde $a_i \in \{0, 1\}$

Sistema Binário

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Por exemplo

$$(101)_2 = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

$$(1001)_2 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

Sistema Binário

- Por exemplo

$$(101)_2 = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

$$(1001)_2 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Representação de Números Inteiros no Computador

Conversão de Bases

- Um número na base β pode ser convertido para base decimal como

$$(N)_{10} = a_n \times \beta^n + a_{n-1} \times \beta^{n-1} + \cdots + a_1 \times \beta^1 + a_0 \times \beta^0$$

onde a_i são os dígitos do número representado em na base β

Exemplo

- **Exemplo 1**
  Converta $(110)_2$ para a base decimal.

Exemplo

- **Exemplo 2**

CQ: Converta $(1001)_2$ para a base decimal.

Conversão de Bases

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Conversão da base decimal para a base β
  - Divisões sucessivas do número em base decimal por β até que o quociente seja igual a zero
  - O número na base β é formado pela concatenação em ordem inversa dos restos das divisões

Exemplo

- **Exemplo 3**
  Converta $(35)_{10}$ para a base 2.

Representação de Números Inteiros no Computad

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Para número não negativos, a representação é direta

$$33 \Rightarrow \boxed{0\ 0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- Para número inteiros com sinal, uma possibilidade é reservar 1 bit para indicar o sinal
  - $0 \Rightarrow$ positivo
  - $1 \Rightarrow$ negativo

$$-33 \Rightarrow \boxed{1}\boxed{0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- 32 bits são normalmente adotados para representação simples
  - Valores em $[0, 2^{32} - 1]$ podem ser representados quando o sinal não é considerado

Representação de Números Inteiros no Computad

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Para número não negativos, a representação é direta

$$33 \Rightarrow \boxed{0\ 0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- Para número inteiros com sinal, uma possibilidade é reservar 1 bit para indicar o sinal
  - $0 \Rightarrow$ positivo
  - $1 \Rightarrow$ negativo

$$-33 \Rightarrow \boxed{1 \mid 0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- 32 bits são normalmente adotados para representação simples
  - Valores em $[0, 2^{32} - 1]$ podem ser representados quando o sinal não é considerado

Representação de Números Inteiros no Computad

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Para número não negativos, a representação é direta

$$33 \Rightarrow \boxed{0\ 0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- Para número inteiros com sinal, uma possibilidade é reservar 1 bit para indicar o sinal
  - $0 \Rightarrow$ positivo
  - $1 \Rightarrow$ negativo

$$-33 \Rightarrow \boxed{1 \mid 0\ 1\ 0\ 0\ 0\ 0\ 1}$$

- 32 bits são normalmente adotados para representação simples
  - Valores em $[0, 2^{32} - 1]$ podem ser representados quando o sinal não é considerado

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Representação de Números Reais no Computador

Representação de Números Reais no Computador

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Um número real positivo x pode ser escrito como

$$x = \underbrace{\sum_{i=0}^{n} a_i B^i}_{x_{\text{int}}} + \underbrace{\sum_{i=1}^{\infty} b_i B^{-i}}_{x_{\text{frac}}}$$

onde a_i e b_i são, respectivamente, os coeficientes da parte inteira e fracionária do número x

- Por exemplo,

$$(123{,}45)_{10} = 1 \times 10^2 + 2 \times 10^1 + 3 \times 10^0 + 4 \times 10^{-1} + 5 \times 10^{-2}$$

Representação de Números Reais no Computador

- Se $b_i = 0$ para todo i maior que um valor inteiro, então diz-se que a fração termina
- Caso contrário, diz-se que a fração não termina
- Exemplos:

$$0{,}45 = 4 \times 10^{-1} + 5 \times 10^{-2} \quad \Rightarrow \text{termina}$$

$$0{,}666\ldots = 6 \times 10^{-1} + 6 \times 10^{-2} + 6 \times 10^{-3} + \ldots \quad \Rightarrow \text{não termina}$$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Mudança de Base

- Conversão de base 2 para decimal
  - Similar ao caso inteiro
- Conversão de base decimal para binária
  - Converte-se a parte inteira
    - Divisões sucessivas
  - Converte-se a parte fracionária
    - Multiplicações sucessivas

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Mudança de Base

- Conversão de base 2 para decimal
  - Similar ao caso inteiro
- Conversão de base decimal para binária
  - Converte-se a parte inteira
    - Divisões sucessivas
  - Converte-se a parte fracionária
    - Multiplicações sucessivas

Mudança de Base – Multiplicações sucessivas

- Seja $(x_{\text{frac}})_{10}$ a parte fracionária de $(x)_{10}$, a fração binária $(,b_1 b_2 \dots)_2$ é determinada como

$$
\begin{aligned}
c_0 &= x_{\text{frac}} \\
b_1 &= (2 \times c_0)_{\text{int}} & c_1 &= (2 \times c_0)_{\text{frac}} \\
b_2 &= (2 \times c_1)_{\text{int}} & c_2 &= (2 \times c_1)_{\text{frac}} \\
&\vdots & &\vdots
\end{aligned}
$$

  onde "int" representa a parte inteira do número e "frac" a parte fracionária

- O processo pode ser finalizado quando $c_i = 0$

Exemplo

- **Exemplo 4**
  Converta o número $(111{,}01)_2$ para a base 10.

Exemplo

- **Exemplo 5**
  Converta o número $(3{,}25)_{10}$ para a base 2.

Exemplo

- **Exemplo 6**
  Converta o número $(0{,}1)_{10}$ para a base 2.

Representação de Números Reais no Computador

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- O computador representa os números em sistema binário
- A representação é finita
  - Números como o $\pi = 3{,}1415\ldots$ são aproximados
- Existem duas formas de representar números reais no computador
  - Ponto fixo
  - **Ponto flutuante**

Representação em Ponto Fixo

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Neste sistema uma palavra (número) é representada por 3 campos
  - 1 bit para o sinal
  - bits que formam a parte inteira
  - bits que formam a parte fracionária
- Por exemplo, o número $(12{,}75)_{10}$ pode ser representado em um sistema com 32 bits (15 bits para a parte inteira e 16 bits para a parte fracionária) como

0	000000000001100	1100000000000000

- O sistema de ponto fixo limita muito a magnitude dos números que podem ser representados
- Essa representação é raramente adotada

Representação em Ponto Fixo

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Neste sistema uma palavra (número) é representada por 3 campos
  - 1 bit para o sinal
  - bits que formam a parte inteira
  - bits que formam a parte fracionária
- Por exemplo, o número $(12{,}75)_{10}$ pode ser representado em um sistema com 32 bits (15 bits para a parte inteira e 16 bits para a parte fracionária) como

0	000000000001100	1100000000000000

- O sistema de ponto fixo limita muito a magnitude dos números que podem ser representados
- Essa representação é raramente adotada

Representação em Ponto Fixo

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Neste sistema uma palavra (número) é representada por 3 campos
  - 1 bit para o sinal
  - bits que formam a parte inteira
  - bits que formam a parte fracionária
- Por exemplo, o número $(12{,}75)_{10}$ pode ser representado em um sistema com 32 bits (15 bits para a parte inteira e 16 bits para a parte fracionária) como

0	000000000001100	1100000000000000

- O sistema de ponto fixo limita muito a magnitude dos números que podem ser representados
- Essa representação é raramente adotada

Representação em Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- A representação em ponto flutuante é baseado na notação científica

$$x = \pm d \times \beta^e$$

onde d é a mantissa, β é a base do sistema de numeração e e é o expoente

- A mantissa é um número na forma

$$(0,d_1 d_2 \ldots d_t)_\beta$$

onde t é o número de dígitos e $d_i \in \{0, 1, \ldots, (\beta - 1)\}$, $i = 1, \ldots, t$

- O expoente e é definido no intervalo $[L, U]$
- Um número é dito normalizado quando $d_1 \neq 0$
  - Os sistemas apresentados no curso são normalizados (a menos que o contrário seja dito)

Representação em Ponto Flutuante

- Um sistema de ponto flutuante pode ser definido como

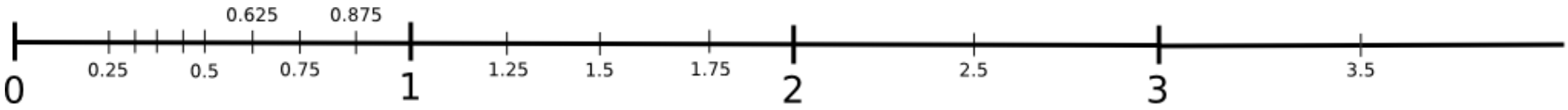
$$F(\beta, t, L, U)$$

onde

  - β é a base do sistema
  - t é o número de dígitos da mantissa
  - L é o menor valor para o expoente
  - U é o maior valor para o expoente

Representação em Ponto Flutuante

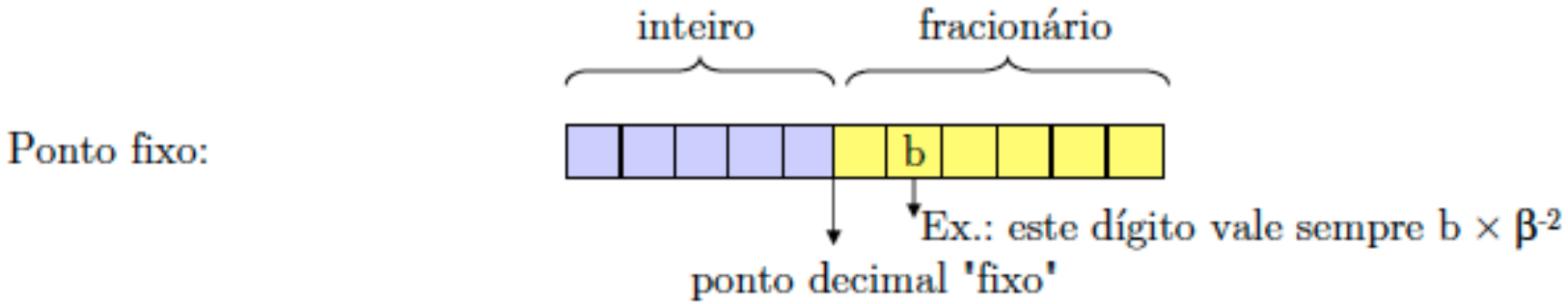
- Nota-se que os números em ponto flutuante são discretos
- Uma característica é a variação entre a discretização de números com magnitudes diferentes

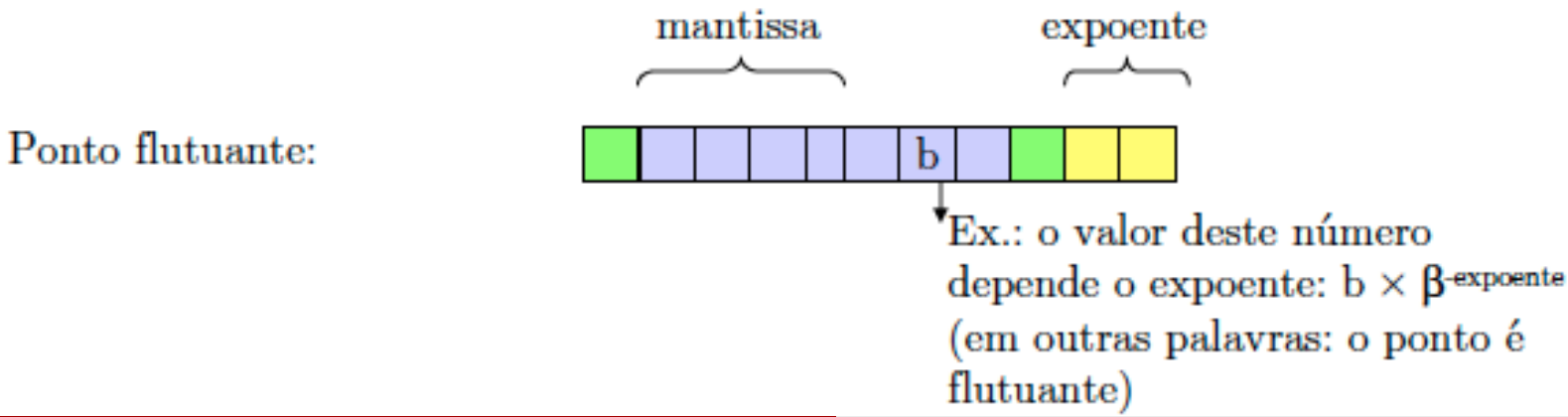


Ponto Fixo *versus* Ponto Flutuante

Diferenças: ponto fixo × flutuante

Suponha que temos 10 dígitos disponíveis:



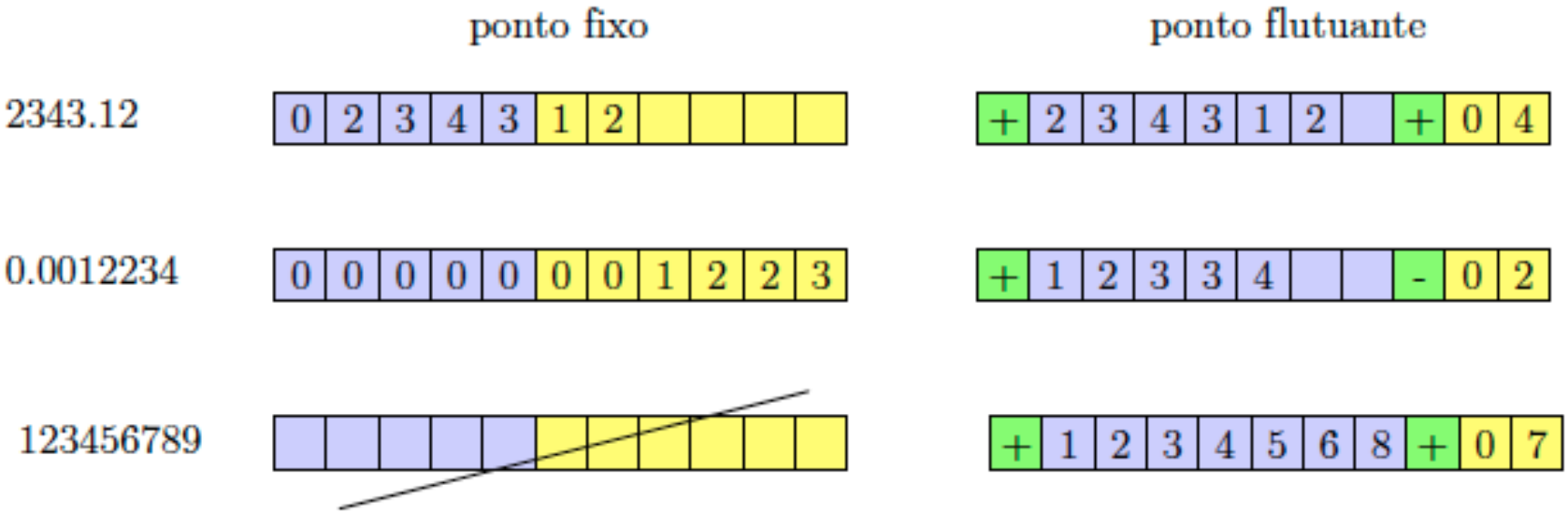


Ponto Fixo *versus* Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Exemplo: ponto fixo × flutuante

Base 10.



Qual a vantagem de ponto flutuante?

(Questão em aberto: arredondamento!)

Exemplo

- **Exemplo 7**
  Considerando o sistema $F(10, 3, -5, 5)$.
  Represente o número $1{,}23$ nesse sistema.

Exemplo

- **Exemplo 8**
  Considerando o sistema $F(10, 3, -5, 5)$.
  Qual o menor número em valor absoluto que esse sistema pode representar?

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Exemplo

- **Exemplo 9**

CQ: Considerando o sistema $F(10, 3, -5, 5)$.
Qual o maior número que esse sistema pode representar?

Representação em Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Sejam m e M, respectivamente, o menor e o maior valores absolutos representáveis no sistema $F(\beta, t, L, U)$
- Dado um número x, então
  - Se $m \leq |x| \leq M$, então o número pode ser representado no sistema
    - Os valores podem ser arredondados ou truncados
    - Truncamento: dígitos $d_{t+1}d_{t+2}\ldots$ são removidos
    - Arredondamento: na base 10, além de remover os dígitos $d_{t+1}d_{t+2}\ldots$, soma-se 1 ao dígito d_t se $d_{t+1} \geq 5$,
  - Se $|x| < m$, então o número não pode ser representado no sistema e diz-se que ocorre *underflow*
  - Se $|x| > M$, então o número não pode ser representado no sistema e diz-se que ocorre *overflow*

Exemplo

- **Exemplo 10**
  Considerando o sistema $F(2, 3, -1, 2)$ com truncamento. Represente o número $(0{,}38)_{10}$ nesse sistema.

Exemplo

- **Exemplo 11**
  Considerando o sistema $F(2, 3, -1, 2)$ com truncamento.
  Represente o número $(5{,}3)_{10}$ nesse sistema.

Exemplo

- **Exemplo 12**
  Considerando o sistema $F(2, 3, -1, 2)$ com truncamento. Represente o número $(0{,}15)_{10}$ nesse sistema.

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Operações Aritméticas em Ponto Flutuante

Operações Aritméticas em Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Adição/Subtração
  - Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número
- Multiplicação/Divisão
  - Realiza-se a operação nas mantissas e nos expoentes
- Os valores devem ser representados no sistema utilizado
- Os resultados devem ser truncados ou arredondados
  - Definição do sistema

Operações Aritméticas em Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Adição/Subtração
  - Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número
- Multiplicação/Divisão
  - Realiza-se a operação nas mantissas e nos expoentes
- Os valores devem ser representados no sistema utilizado
- Os resultados devem ser truncados ou arredondados
  - Definição do sistema

Operações Aritméticas em Ponto Flutuante

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Adição/Subtração
  - Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número
- Multiplicação/Divisão
  - Realiza-se a operação nas mantissas e nos expoentes
- Os valores devem ser representados no sistema utilizado
- Os resultados devem ser truncados ou arredondados
  - Definição do sistema

Exemplo

- **Exemplo 13**
  Seja o sistema $F(10, 2, L, U)$ com arredondamento; os limitantes do expoente são ignorados nesse exemplo. Some $4{,}32$ e $0{,}064$ nesse sistema.

Exemplo

- **Exemplo 14**
  Seja o sistema $F(10, 2, L, U)$ com arredondamento. Multiplique 1234 por $0{,}016$ nesse sistema.

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Noções Básicas Sobre Erros

Noções Básicas Sobre Erros

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Além disso, erros podem ser introduzidos ao representar números no computador
- Um número real x provavelmente será aproximado quando representado em ponto flutuante no computador
- É necessário definir medidas para calcular erros em aproximações
  - erro absoluto
  - erro relativo

Erro Absoluto

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Seja $\bar{x}$ uma aproximação de x, o erro absoluto é definido como

$$EA(\bar{x}) = |x - \bar{x}|$$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Exemplo

- **Exemplo 15**
  Seja o sistema $F(10, 4, L, U)$ com arredondamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?

- Solução:
  $\bar{x} = 0{,}1429 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1429| = 0{,}244$

Exemplo

- **Exemplo 15**
  Seja o sistema $F(10, 4, L, U)$ com arredondamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?

- Solução:
  $\bar{x} = 0{,}1429 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1429| = 0{,}244$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Exemplo

- **Exemplo 16**
  Seja o sistema $F(10, 4, L, U)$ com truncamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?

- Solução:
  $\bar{x} = 0{,}1428 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1428| = 0{,}756$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Exemplo

- **Exemplo 16**
  Seja o sistema $F(10, 4, L, U)$ com truncamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?

- Solução:
  $\bar{x} = 0{,}1428 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1428| = 0{,}756$

Erro de truncamento pode ser menor que erro de arredondamento?

Erro Relativo

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Seja $\bar{x}$ uma aproximação de x, o erro relativo é definido como

$$ER(\bar{x}) = \frac{|x - \bar{x}|}{|x|} = \frac{EA(\bar{x})}{|x|}$$

- dado $x \neq 0$.

Exemplo

- **Exemplo 17**
  Sejam $x_1 = 1000{,}5$, $\bar{x_1} = 1000{,}6$, $x_2 = 10{,}5$ e $\bar{x_2} = 10{,}6$. Nota-se que $EA(\bar{x_1}) = EA(\bar{x_2}) = 0{,}1$. Quais os erros relativos?

Noções Básicas Sobre Erros

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- O valor de x geralmente não é conhecido
- Na prática utiliza-se uma medida de erro entre aproximações

$$\frac{||x_{novo} - x_{antigo}||}{||x_{novo}||}$$

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Efeitos Numéricos

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Além dos erros causados pela representação no computador, existem certos efeitos numéricos que contribuem para aumentar os erros
  - Somar (ou subtrair) números com ordens de grandeza muito diferentes
  - Cancelamento
  - Propagação do erro

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Somar (ou subtrair) números com ordens de grandeza muito diferentes
  - As operações de soma e subtração podem não ter o efeito desejado
- Por exemplo, ao somar $0{,}1$ e 5000 num sistema $F(10, 4, L, U)$, obtém-se

$$\begin{aligned} 0{,}1 + 5000 &= 0{,}1000 \times 10^0 + 0{,}5000 \times 10^4 \\ &= 0{,}00001 \times 10^4 + 0{,}5000 \times 10^4 \\ &= 0{,}50001 \times 10^4 \\ &= 0{,}5000 \times 10^4 \text{ (arredondando ou truncando)} \end{aligned}$$

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Cancelamento
  - Ocorre quando dois números muito parecidos são subtraídos
  - Os expoentes devem ser igualados quando se calcula $x - y$
  - Quando x e y são similares, vários zeros aparecem no final da mantissa do resultado ao normalizá-lo
  - Ocorre assim uma perda de dígitos significativos

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Um grande número de operações elementares é normalmente utilizado em métodos numéricos para buscar a solução de um determinado problema
  - Assim, o erro cometido em uma operação isolada pode não ser muito significativo para a solução do problema tratado
  - Entretanto, é necessário analisar como esses erros se propagam
    - erro ilimitado: se acumulam a uma taxa crescente e a sequência de operações é considerada instável
    - erro limitado: se acumulam a uma taxa decrescente e a sequência de operações é considerada estável

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Por exemplo, considerando um sistema $F(10, 4, L, U)$ com truncamento, ao efetuar a operação

$$S = \sum_{i=1}^{4}(x_i + y_i); \qquad x_i = 0{,}46709 \text{ e } y_i = 3{,}5678$$

  - Para $i = 1$

$$(x_1 + y_1) = 0{,}4034 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |4{,}03569 - 4{,}034| = 0{,}00169$

  - Para $i = 2$

$$(x_1 + y_1) + (x_2 + y_2) = 0{,}8068 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |8{,}07138 - 8{,}068| = 0{,}00338$

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Por exemplo, considerando um sistema $F(10, 4, L, U)$ com truncamento, ao efetuar a operação

$$S = \sum_{i=1}^{4}(x_i + y_i); \qquad x_i = 0{,}46709 \text{ e } y_i = 3{,}5678$$

  - Para $i = 1$

$$(x_1 + y_1) = 0{,}4034 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |4{,}03569 - 4{,}034| = 0{,}00169$

  - Para $i = 2$

$$(x_1 + y_1) + (x_2 + y_2) = 0{,}8068 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |8{,}07138 - 8{,}068| = 0{,}00338$

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Por exemplo, considerando um sistema $F(10, 4, L, U)$ com truncamento, ao efetuar a operação

$$S = \sum_{i=1}^{4}(x_i + y_i); \qquad x_i = 0{,}46709 \text{ e } y_i = 3{,}5678$$

  - Para $i = 1$

$$(x_1 + y_1) = 0{,}4034 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |4{,}03569 - 4{,}034| = 0{,}00169$

  - Para $i = 2$

$$(x_1 + y_1) + (x_2 + y_2) = 0{,}8068 \times 10^1$$

    e o erro absoluto é dado por $EA(\bar{S}) = |8{,}07138 - 8{,}068| = 0{,}00338$

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Para $i = 3$

    $$(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = 0{,}1210 \times 10^2$$

    e o erro absoluto é dado por $EA(\bar{S}) = |12{,}10707 - 12{,}10| = 0{,}00707$
  - Para $i = 4$

    $$S = \sum_{i=1}^{4}(x_i + y_i) = 0{,}1613 \times 10^2$$

    e o erro absoluto é dado por $EA(\bar{S}) = |16{,}14267 - 16{,}13| = 0{,}01276$
  - Pode-se observar que o erro absoluto aumenta à medida em que as operações são realizadas

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Propagação dos erros
  - Para $i = 3$

    $$(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = 0{,}1210 \times 10^2$$

    e o erro absoluto é dado por $EA(\bar{S}) = |12{,}10707 - 12{,}10| = 0{,}00707$
  - Para $i = 4$

    $$S = \sum_{i=1}^{4}(x_i + y_i) = 0{,}1613 \times 10^2$$

    e o erro absoluto é dado por $EA(\bar{S}) = |16{,}14267 - 16{,}13| = 0{,}01276$
  - Pode-se observar que o erro absoluto aumenta à medida em que as operações são realizadas

Efeitos Numéricos

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- A implementação ou o uso incorreto de algoritmos e softwares científicos já foi responsável por alguns desastres
- Guerra do Golfo (1991)
- Uma bateria de mísseis Patriot ("Phased Array TRacking Intercept Of Target") americano, falhou ao rastrear e interceptar um míssil Scud do Iraque
- O míssil Scud acertou o acampamento americano
  - 28 soldados morreram e centenas ficaram feridos
- O tempo era medido em décimos de segundo
  - Uma dízima periódica em um sistema binário
- O acúmulo do erro durante o tempo em que o sistema estava operante levou à falha

Fontes

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

- Curso de Cálculo Numérico - UFJF

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

Dúvidas?