UNIVALI

Escola do Mar, Ciência & Tecnologia

Cálculo Numérico

Curso de Ciência da Computação
Campus Kobrasol

Prof. Denise Prado Kronbauer
denise.kronbauer@univali.br
denipk@gmail.com

Ementa:

- Noções básicas de erros.
- Sistemas de equações lineares.
- Equações não lineares.

M1

- Interpolação polinomial.
- Integração numérica.

M2

- Derivação numérica.
- Métodos numéricos para soluções de EDO.
- Noções de métodos numéricos para soluções de EDP.

M3

Frequência e Média Final

Frequência: De acordo com as normas da UNIVALI, cada aluno precisa de, no mínimo, 75% de frequência para ser aprovado.

Média Final: A média final (MF) deverá ser igual ou superior a 6,0. O aluno que não atingir a média final 6,0 estará reprovado, pois não existem exames finais.

$$MF = \frac{M_1 + M_2 + M_3}{3}$$

Justificativa de Faltas

O aluno que se ausentar nas atividades avaliativas ou solicitar processo de justificativa de falta, ficará sujeito ao deferimento pela coordenação do curso.

Caso aprovado o processo, todas as avaliações atrasadas serão realizadas individualmente no dia **09.11 (sábado), das 8h às 11:30h**.

Calendário das Aulas e Avaliações

Encontro	Data
1	06 de agosto
2	13 de agosto
3	20 de agosto
4	**27 de agosto (1 – M1)**
5	03 de setembro
6	10 de setembro
7	**17 de setembro (2 – M1)**
8	24 de setembro
9	01 de Outubro

Encontro	Data
10	**08 de outubro (1 – M2)**
11	22 de outubro
12	**29 de outubro (2 – M2)**
13	05 de novembro
14	**12 de novembro (1 – M3)**
15	19 de novembro
16	26 de novembro
17	03 de dezembro
18	**10 de dezembro (2 – M3)**

Avaliações atrasadas: 09.11 (sábado), das 8h às 11:30h.

Referências bibliográficas

- FRANCO, Neide Bertoldi. Cálculo numérico. São Paulo, SP: Pearson Prentice Hall, 2006. xii, 505 p. ISBN 8576050870.
- BURDEN, Richard L; FAIRES, J. Douglas. Analise numérica. São Paulo, SP: Pioneira Thomson Learning, c2003. 736 p il ISBN 852210297X (broch.).
- CHAPRA, Steven C; CANALE, Raymond P. Métodos numéricos para engenharia. 5. ed. São Paulo, SP: McGraw-Hill, 2008. xxi, 809 p. ISBN 978-85-86804-87-8.

Referências bibliográficas

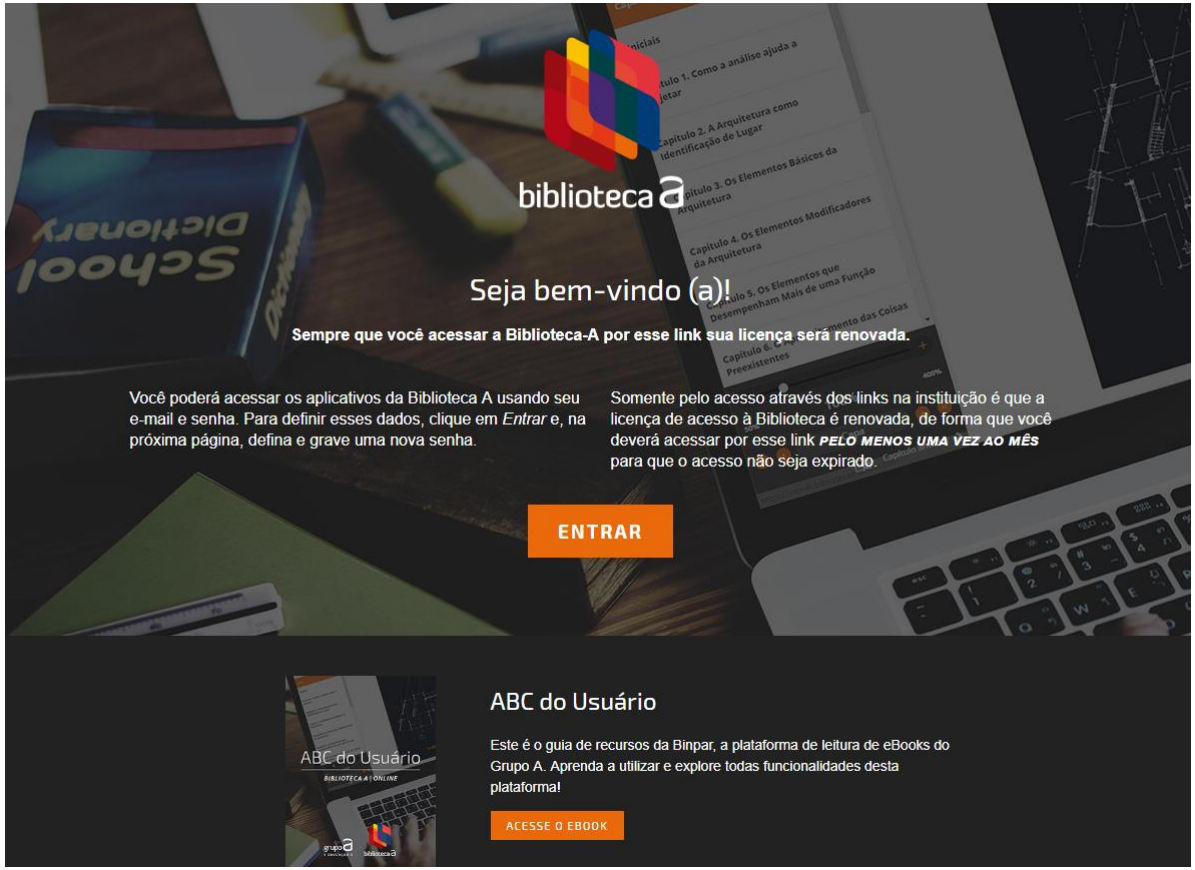
- CLAUDIO, Dalcidio Moraes; MARINS, Jussara Maria. Calculo numérico computacional: teoria e pratica. 3. ed. São Paulo, SP: Atlas, 2000. 464p ISBN 8522410437 (broch.).
- SPERANDIO, Decio; MENDES, Joao Teixeira. Calculo numérico: características matemáticas e computacionais dos métodos numéricos. São Paulo, SP: Prentice Hall, 2003. 354 p. ISBN 8587918745 (broch.).
- BARROSO, Leônidas Conceição. Cálculo numérico: (com aplicações). 2. ed. São Paulo, SP: Harba, c1987. xii, 367 p. ISBN 8529400895.

Referências bibliográficas

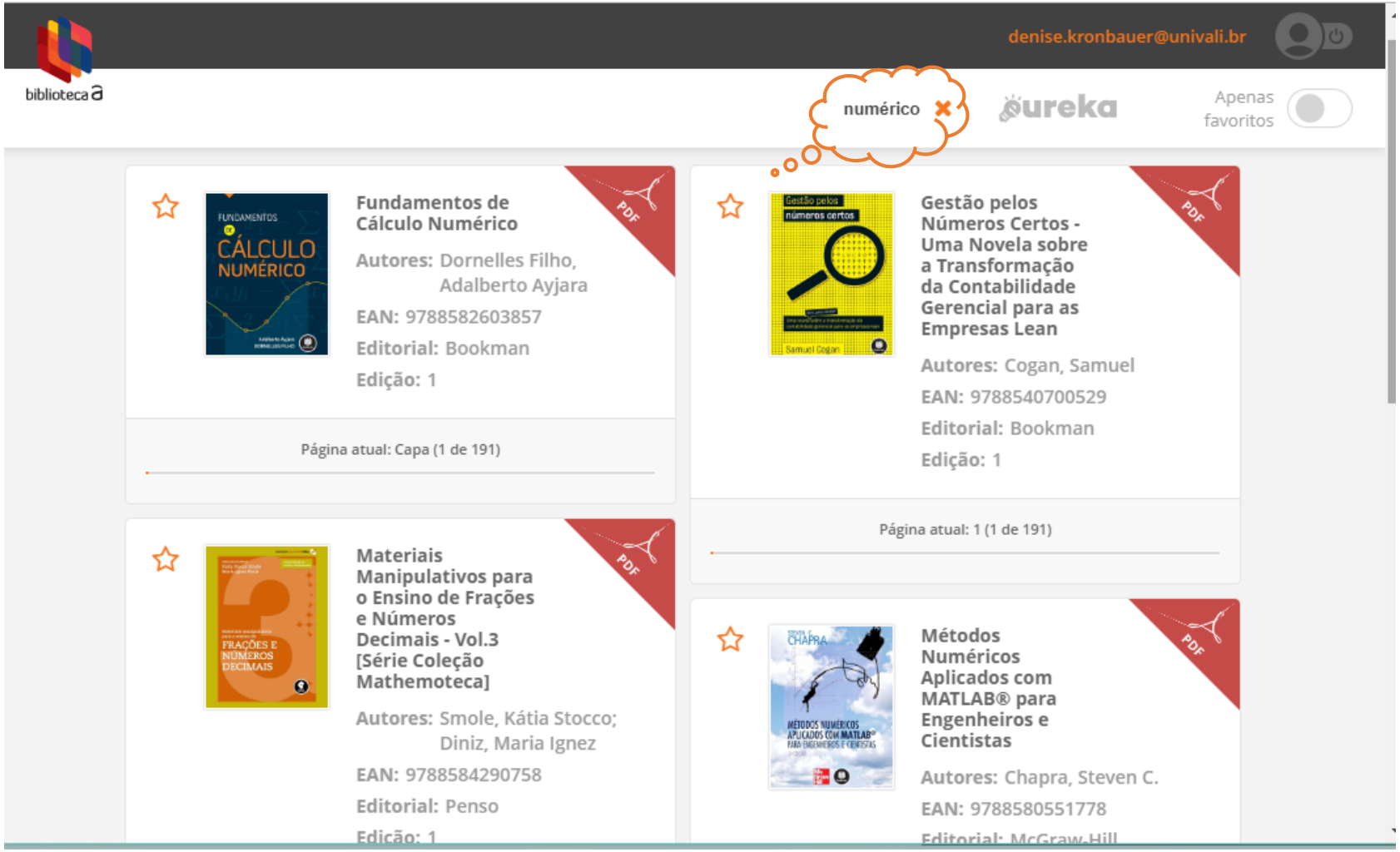
- RUGGIERO, Márcia A. Gomes; LOPES, Vera Lúcia da Rocha. Cálculo numérico: aspectos teóricos e computacionais. 2. ed. São Paulo, SP: Makron Books, c1998, 1997. xvi, 406 p. ISBN 8534602042.
- STARK, Peter A. Introducao aos metodos numericos. Rio de Janeiro, RJ: Interciência, 1979. 340p.
- DIEGUEZ, Jose Paulo P. (Jose Paulo do Prado). Métodos numéricos computacionais para a engenharia. Rio de Janeiro, RJ: Ambito Cultural, c1994. 2 v. ISBN Broch.
- CAMPOS, Rui J. A. Calculo numerico basico. 1978. 127p ISBN Broch.

Novidade: Biblioteca Digital

Univali → Minha Univali → Intranet → Biblioteca Digital



Novidade: Biblioteca Digital



UNIVALI

Escola do Mar, Ciência & Tecnologia

Cálculo Numérico

Curso de Ciência da Computação
Campus Kobrasol

Prof. Denise Prado Kronbauer
denise.kronbauer@univali.br
denipk@gmail.com

Foi com o surgimento do computador na década de 40 que a importância da modelagem começou a ser notada, uma vez que, por meio do processamento eletrônico de dados, as técnicas numéricas se tornaram viáveis.

O cálculo numérico tem sua importância centrada no fato de que, mesmo quando a solução analítica é difícil de ser obtida, as técnicas numéricas podem ser empregadas sem maiores dificuldades.

Por exemplo, uma solução da equação

$$x^6 - 20x^5 - 110x^4 + 50x^3 - 5x^2 + 70x - 100 = 0$$

pode, sem grandes dificuldades, ser obtida por meio do cálculo numérico, mesmo quando é impossível encontrar uma solução analítica.

No cálculo numérico serão apresentados os principais métodos numéricos, como utilizá-los, suas restrições e suas vantagens, desde aspectos teóricos até a implementação computacional.

Método Numérico: É um conjunto de procedimentos utilizados para transformar um modelo matemático num problema numérico ou um conjunto de procedimentos usados para resolver um problema numérico. A escolha do método mais eficiente para resolver um problema numérico deve envolver os aspectos:

i. Precisão desejada para os resultados;

ii. Capacidade do método em conduzir aos resultados desejados (velocidade de convergência);

iii. Esforço computacional despendido (tempo de processamento, economia de memória necessária para a resolução).

Algoritmo: É a descrição sequencial dos passos que caracterizam um método numérico. São as operações por meio das quais o conjunto de dados de entrada é transformado em dados de saída, ou seja, o algoritmo consiste de uma sequência de n passos, cada um envolvendo um número finito de operações, para chegar em valores ao menos "próximos" daqueles que são procurados.

Um algoritmo poderá depender ou não de dados iniciais (entradas), mas terá obrigatoriamente de produzir pelo menos um resultado final (saída).

Iterações ou Aproximação Sucessiva: Iteração significa a repetição de um processo. Partimos de uma solução aproximada inicial, repetimos as mesmas ações/processos para refinar a solução inicial.

Para evitar que trabalhemos sem fim, devemos determinar se a iteração converge (nem sempre ocorre) e/ou quais são as condições de parada.

Tipos de Erros: Na busca da solução do modelo matemático por meio de cálculo numérico, os erros surgem de várias fontes. As principais fontes de erros são as seguintes:

i. Erro nos dados de entrada;

ii. Erros no estabelecimento do modelo matemático;

iii. Erros de arredondamento durante a computação;

iv. Erros de truncamento, e

v. Erros humanos e de máquinas.

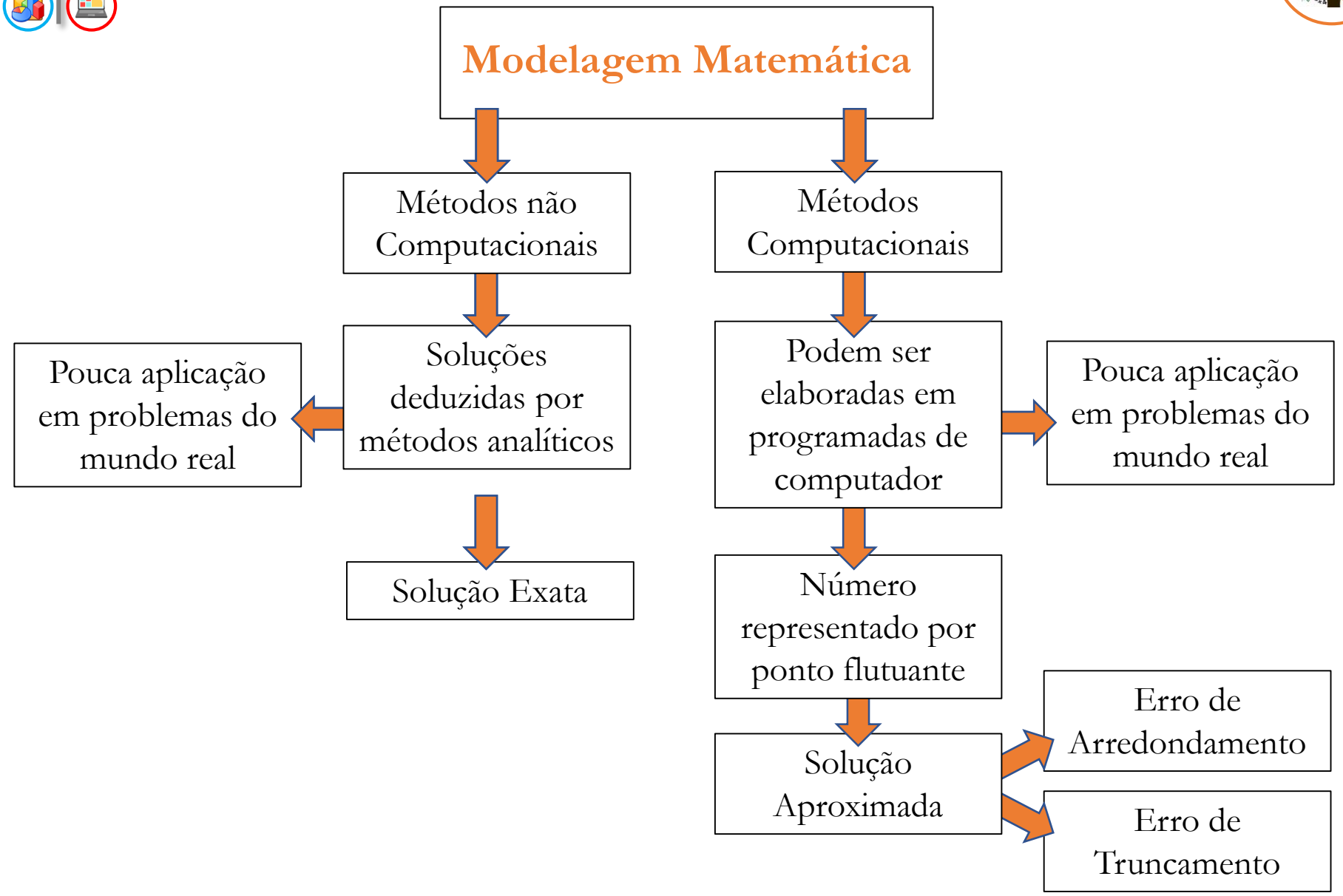
Modelagem Matemática: É a área da matemática que transforma os fenômenos ou as situações reais em linguagem numérica. Para isso, utiliza símbolos, formas e padrões adotados ao longo de anos de evolução desses algoritmos.

A modelagem matemática que conhecemos hoje é, na verdade, consequência de centenas de anos de evolução do processo algébrico. Quando falamos em modelar um problema, estamos nos referindo à tradução de uma situação real para a linguagem matemática formal, que se utiliza de letras, números e métodos numéricos para a solução de problemas reais ou fictícios.

Etapas de Construção de um Modelo

1. **Situação Problema:** Nesta fase, o modelo matemático transforma-se em modelo numérico programável. O uso dos métodos numéricos é importante para evitar erros e determinar a precisão necessária para o modelo.
2. **Coleta de Dados:** Esta fase consiste em juntar as informações conhecidas sobre o problema: valores conhecidos, objetivos a serem alcançados, fórmulas, etc.
3. **Modelagem Matemática:** Aqui, obtemos um modelo matemático inicial que atenda à situação-problema.

4. **Verificação do Modelo:** Um bom modelo matemático atende à situação-problema e qualquer situação similar.
5. **Programação do Modelo:** Nesta fase, confrontamo-nos com uma situação real ou hipotética que precisará ser transformada em modelo matemático.
6. **Aplicação do Modelo:** Após programado, verificamos o funcionamento do modelo.
7. **Verificação e Otimização de Erros:** Nesta fase, os erros devem ser corrigidos ou, na ausência deles, pode-se verificar, nos métodos numéricos utilizados, possíveis formas de fazer a máquina ser mais eficiente nos cálculos.


Modelagem Matemática
Métodos não Computacionais
Métodos Computacionais
Soluções deduzidas por métodos analíticos
Pouca aplicação em problemas do mundo real
Solução Exata
Podem ser elaboradas em programadas de computador
Pouca aplicação em problemas do mundo real
Número representado por ponto flutuante
Solução Aproximada
Erro de Arredondamento
Erro de Truncamento

Atualmente, o sistema numérico mais utilizado é o sistema decimal, que tem importante aplicação por utilizar o valor posicional, ou seja, um mesmo símbolo pode possuir diferentes valores de acordo com a posição que ocupa no número.

Nos computadores e processadores digitais em geral, a linguagem utilizada é a binária. Dessa forma, para um processador, todo e qualquer símbolo inserido e processado é lido como uma combinação de 0 e 1.

Observe o que acontece na interação entre o usuário e o computador: os dados de entrada são enviados ao computador pelo usuário no sistema decimal (10 dígitos possíveis: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9): toda esta informação é convertida para o sistema binário (2 dígitos possíveis: 0, 1), e as operações todas serão efetuadas neste sistema.

Os resultados finais serão convertidos para o sistema decimal e, finalmente, serão transmitidos ao usuário. Todo este processo de conversão é uma fonte de erros que afetam o resultado final dos cálculos.

Veremos a seguir a conversão do sistema decimal para o binário e vice-versa, para depois abordar o sistema da aritmética de ponto flutuante que as máquinas utilizam para representar os números.

De maneira geral, um número real x na base β é representado por:

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta^1 + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} + \cdots + b_m\beta^{-m}$$

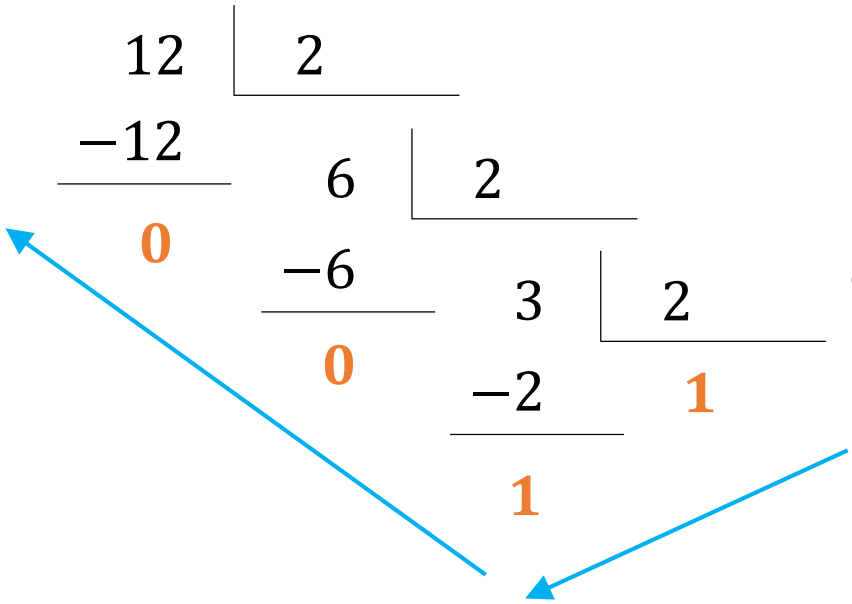
Onde $a_i, i = 0,1,2 \ldots, n$ e $b_j, j = 1,2, \ldots, m$

Para fazer a conversão da **base decimal para binária**, consideremos x na base 10 e dividimos sucessivamente a parte inteira de x por 2, até que o último quociente seja igual a 1.

O número é composto pelo último quociente obtido que é igual a 1 e com os restos das divisões lidos em sentido inverso àquele que foram obtidos.

Para converter uma parte fracionária, multiplicamos sucessivamente a parte decimal por 2. O número na base 2 será então obtido tomando-se a parte inteira do resultado de cada multiplicação.

Vejamos a representação do número inteiro 12 na base 2:



O número será representado na base binária pelos restos das divisões efetuadas, do último para o primeiro:

$$(12)_{10} = (1100)_2$$

Vejamos a representação do número inteiro 0,75 na base 2:

$$0{,}75 \times 2 = \mathbf{1}{,}50$$

$$0{,}50 \times 2 = \mathbf{1}{,}00$$

$$0{,}00 \times 2 = 0{,}00$$

O número será representado na base binária pela parte inteira do resultado de cada multiplicação:

$$(0{,}75)_{10} = (0{,}11)_2$$

Exercícios: Dado o número x na base decimal, converta no seu equivalente na base binária:

a) 23

b) 347

c) 23,625

d) 0,6

e) 0,125

f) 0,1

g) 2012

Respostas: base decimal → base binária:

a) 10111

b) 101011011

c) 10111,101

d) 0,10011001 ...

e) 0,001

f) $0,0\overline{0011}$

g) 11111011100

Para proceder à conversão da **base binária para decimal**,

o procedimento é multiplicar cada algarismo do número na base 2 por potências decrescentes de 2, da esquerda para a direita e somar as parcelas.

Se o número na base binária for decimal, o procedimento é multiplicar cada algarismo do número na base 2, após o ponto, por potências decrescentes de 2, da esquerda para a direita, e somar todas as parcelas.

Vejamos a representação do número 1101 que está na base 2, para a base 10:

$$(1101)_2 = (1 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0)$$

$$= 8 + 4 + 0 + 1 = 13$$

$$(1101)_2 = (13)_{10}$$

Vejamos a representação do número 0,110 que está na base 2, para a base 10:

$$(0{,}110)_2 = (1 \times 2^{-1}) + (1 \times 2^{-2}) + (0 \times 2^{-3})$$

$$= \frac{1}{2} + \frac{1}{4} + 0 = \frac{3}{4}$$

$$(0{,}110)_2 = \left(\frac{3}{4}\right)_{10} = (0{,}75)_{10}$$

Exercícios: Dado o número x na base binária, converta no seu equivalente na base decimal:

a) 110111

b) 0,01011

c) 11,0101

d) 10101101

e) 101010001

Respostas: base binária → base decimal:

a) 55

b) 0,34375

c) 3,3125

d) 173

e) 337

As máquinas utilizam o sistema de **aritmética de ponto flutuante** para representar os números e executar as operações. Um número real na base β, em aritmética de ponto flutuante de t dígitos, tem a seguinte forma geral:

$$r = \pm(.d_1 d_2 \ \dots \ d_t) \times \beta^e$$

Onde: β é a base em que a máquina opera;

t é o número de dígitos na mantissa $(.d_1\ d_2\ \dots\ d_t)$;

$0 \leq d_j \leq (\beta - 1), j = 1, \dots, t; d_1 \neq 0,$

e o expoente e é um número inteiro entre m e M.

Um sistema de ponto flutuante F depende das variáveis $\beta, t, m\ e\ M$

e pode ser representado pela função:

$$F = (\beta, t, m, M)$$

Onde a precisão da máquina com o sistema F

é definida pelo número de dígitos da mantissa t.

Exemplos: Seja o sistema de aritmética de ponto flutuante $F = (10{,}3, -4{,}4)$.

Represente os números x neste sistema:

a) $x = -279{,}15$

b) $x = 1{,}35$

c) $x = 0{,}024712$

d) $x = 10{,}093$

Respostas: $F = (10{,}3, -4{,}4)$

a) $x = -279{,}15 \rightarrow -0{,}279 \times 10^{3}$

b) $x = 1{,}35 \rightarrow 0{,}135 \times 10^{1}$

c) $x = 0{,}024712 \rightarrow 0{,}247 \times 10^{-1}$

d) $x = 10{,}093 \rightarrow 0{,}101 \times 10^{2}$

Exemplos: Considere uma máquina que opera no sistema:

$$\beta = 10;\ t = 3;\ e \in [-5,5]$$

Verifique o maior e o menor número, em valor absoluto,

que podem ser representados nesta máquina:

Os números serão representados na seguinte forma nesse sistema:

$$x = 0, d_1 d_2 d_3 \times 10^e;\ 0 \leq d_j \leq 9;\ d_1 \neq 0;\ e \in [-5,5]$$

O menor número, em valor absoluto, representado nessa máquina é:

$$menor = 0{,}100 \times 10^{-5} = 10^{-6}$$

E o maior número, em valor absoluto, a ser representado é:

$$maior = 0{,}999 \times 10^{5} = 99900$$

Exemplos: Considere $F = (2,2,-1,2)$. Quais os números que podem ser representados por esta máquina na base decimal?

$$r = 0,d_1 d_2 \times 2^e; \quad 0 \leq d_j \leq 1;\ d_1 \neq 0;\ e \in [-1,2]$$

$$r = \pm 0,10 \times 2^e; \quad ou\ ainda, \quad r = \pm 0,11 \times 2^e$$

sendo $-1 \leq e \leq 2$

Convertendo para decimal: $0,10 \rightarrow \frac{1}{2}$ e $0,11 \rightarrow \frac{3}{4}$

	2^{-1}	2^0	2^1	2^2
$0,10 = \frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2
$0,11 = \frac{3}{4}$	$\frac{3}{8}$	$\frac{3}{4}$	$\frac{3}{2}$	3

Em valores absolutos, os números que podem ser representados nesta máquina, são:

$$\frac{1}{4}, \quad \frac{3}{8}, \quad \frac{1}{2}, \quad \frac{3}{4}, \quad 1, \quad \frac{3}{2}, \quad 2, \quad 3$$

	2^{-1}	2^{0}	2^{1}	2^{2}
$0{,}10 = \frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2
$0{,}11 = \frac{3}{4}$	$\frac{3}{8}$	$\frac{3}{4}$	$\frac{3}{2}$	3

O número total de elementos de uma aritmética de ponto flutuante é dado por:

$$n = 2(\beta - 1).\beta^{t-1}.(e_{máx} - e_{mín} + 1) + 1$$

Para contabilizar os números negativos

Para o número zero

Para o exemplo anterior temos que o número de elementos é 17.

(8 positivos, 8 negativos e o zero).

Underflow e Overflow

O conjunto de números reais é infinito, entretanto, a sua representação em um sistema de ponto flutuante é limitada, pois é um sistema finito, logo, não há a representação exata da totalidade dos números reais.

Essa limitação tem duas origens:

- A faixa dos expoentes é limitada $(e_{min} \leq e \leq e_{máx})$;
- A mantissa pode representar um número finito de números $(\beta^{t-1} \leq m \leq \beta^{t-1})$.

A primeira limitação leva aos fenômenos chamados de **overflow** e **underflow**.

A segunda leva aos **erros de arredondamento**, que veremos a seguir.

Quando a faixa dos expoentes é limitada ($e_{min} \leq e \leq e_{máx}$):

Sempre que uma operação aritmética produz um número com expoente superior ao expoente máximo, tem-se o fenômeno de "**overflow**".

De forma similar, operações que resultem em expoente inferior ao expoente mínimo tem-se o fenômeno de "**underflow**".

No caso do exemplo dado, pode-se observar qual as regiões que ocorrem o overflow e o underflow. Neste caso, considera-se a parte positiva e negativa da aritmética do exemplo.

Observe que, se o expoente for maior que 3 ou menor que -1, não tem-se representação no conjunto formado pela aritmética de ponto flutuante. No primeiro caso, tem-se o overflow, no segundo caso, tem-se o underflow.

Exemplos: Considere uma aritmética de ponto flutuante $F = (10{,}2, -5{,}5)$.

Sejam $x = 875$ e $y = 3172$. Calcular $x.y$:

Exemplos: Considere uma aritmética de ponto flutuante $F = (10{,}2, -5{,}5)$. Sejam $x = 875$ e $y = 3172$. Calcular $x.y$:

Primeiro devemos arredondar os números e armazená-los no formato indicado. A operação de multiplicação é efetuada usando dois dígitos:

$$\left.\begin{array}{l} x = 0{,}88 \times 10^3 \\ y = 0{,}32 \times 10^4 \end{array}\right\} \rightarrow x.y = 0{,}28 \times 10^7$$

Como o expoente é maior que 5, resulta em **overflow**.

Underflow e Overflow

Exemplos: Considere uma aritmética de ponto flutuante $F = (10{,}2, -5{,}5)$.

Sejam $x = 0{,}0064$ e $y = 7312$. Calcular $x \div y$:

Exemplos: Considere uma aritmética de ponto flutuante $F = (10{,}2, -5{,}5)$. Sejam $x = 0{,}0064$ e $y = 7312$. Calcular $x \div y$:

Primeiro devemos arredondar os números e armazená-los no formato indicado. A operação de divisão é efetuada usando dois dígitos:

$$\left.\begin{array}{l} x = 0{,}64 \times 10^{-2} \\ y = 0{,}73 \times 10^{4} \end{array}\right\} \rightarrow x \div y = 0{,}88 \times 10^{-6}$$

Como o expoente é inferior a -5, resulta em **underflow**.

Exercícios

1. Considere o sistema $F(10,3,-2,2)$. Represente neste sistema os números $x_1 = 0{,}35;\ x_2 = -5{,}172;\ x_3 = 0{,}0123;\ x_4 = 5391{,}3\ e\ x_5 = 0{,}0003$.

2. Considere $F(2,3,-1,2)$. Verificar todos os elementos possíveis de F.
3. Considere o sistema $F(3,2,-1,2)$:
   a) Qual o menor número decimal representável nessa máquina?
   b) Qual o maior número decimal representável nessa máquina?
   c) Quantos números reais positivos podemos representar nesse sistema?

Respostas

1. $x_1 = 0{,}35 \rightarrow 0{,}350 \times 10^0$

$x_2 = -5{,}172 \rightarrow -0{,}517 \times 10^1$

$x_3 = 0{,}0123 \rightarrow 0{,}123 \times 10^{-1}$

$x_4 = 5391{,}3 \rightarrow 0{,}539 \times 10^4\ (Overflow)$

$x_5 = 0{,}0003 \rightarrow 0{,}300 \times 10^{-3}\ (underflow)$

Respostas

2. Os números positivos são: $\left(\frac{1}{4}, \frac{5}{16}, \frac{3}{8}, \frac{7}{16}, \frac{1}{2}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}, 1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4}, 2, \frac{5}{2}, 3, \frac{7}{2}\right)$.

3. $F(3{,}2, -1{,}2)$:

a) $-0{,}10 \times 3^{-1} = \left(-\frac{1}{9}\right)_{10}$

b) $0{,}22 \times 3^{2} = (8)_{10}$

c) 24 números positivos, 24 números negativos + zero = 49 elementos

UNIVALI

Escola do Mar, Ciência & Tecnologia

Cálculo Numérico

Curso de Ciência da Computação
Campus Kobrasol

Prof. Denise Prado Kronbauer
denise.kronbauer@univali.br
denipk@gmail.com

Conforme visto anteriormente, ocorrem limitações quanto à representação do conjunto dos números reais num sistema de ponto flutuante. Uma das limitações refere-se à faixa dos expoentes que é limitada, levando aos fenômenos chamados de overflow e underflow.

Outra limitação ocorre devido à mantissa poder representar um número finito de números, o que leva aos **erros de arredondamento ou truncamento**.

No **arredondamento**, o último algarismo pode se alterar de acordo com o primeiro dígito após a mantissa. Os critérios para arredondamento são os seguintes:

- Se o primeiro algarismo após a mantissa for maior ou igual a 5, acrescentamos uma unidade ao último algarismo da mantissa. Por exemplo, se tivermos uma mantissa de 4 algarismos, o número $53{,}237$ passa a ser $0{,}5324 \times 10^2$;
- Se o primeiro algarismo após a mantissa for menor que 5, devemos manter inalterado o último algarismo da mantissa. Por exemplo, para uma mantissa de 4 algarismos, o número $42{,}873$ passa a ser $0{,}4287 \times 10^2$.

No **truncamento**, ocorre uma pausa mais brusca na separação da mantissa e da parte a ser desprezada, uma vez que, simplesmente, se abre mão de todo número após a mantissa sem qualquer tipo de mudança nos algarismos da mantissa.

Esse tipo de "pausa" para o ponto flutuante é bem mais simples para o processador, por isso permite um processamento mais rápido. Porém, cria erros maiores, pois diminui significativamente a precisão numérica.

Considere o número 0,999999..., sendo esse uma dízima periódica igual a 1. No arredondamento, qualquer que fosse o tamanho da mantissa, isso iria gerar o aumento da casa decimal à esquerda e, por fim, o número se tornaria o próprio 1. Já no truncamento isso não iria ocorrer, pois qualquer que fosse o tamanho da mantissa, o número continuaria uma sequência de algarismos 9, apenas determinados pelo tamanho da mantissa.

Exemplo: Seja uma máquina que opere com apenas 6 dígitos na mantissa, ou seja, que seja capaz de armazenar números no formato

$$m = \pm 0, d_1 d_2 d_3 d_4 d_5 d_6 \times 10^e$$

Como armazenaríamos o número abaixo nesta máquina?

$$(0{,}11)_{10} = (0{,}0001110000101000111101\overline{1110000101000111101}\ldots)_2$$

Como $(0{,}11)_{10}$ não tem representação binária finita, teremos neste caso:

$$(0{,}11)_{10} \rightarrow (0{,}000111)_2 \rightarrow (0{,}109375)_{10}$$

Exemplo: Considere a representação binária de 0,6 e 0,7:

$$(0{,}6)_{10} = (0{,}100110011001\ldots)_2$$

$$(0{,}7)_{10} = (0{,}1011001100110\ldots)_2$$

Se esses dois números forem representados na aritmética $F(2,2,-1,2)$ eles serão representados igualmente por $0{,}10 \times 2^0$. Esse número equivale a $0{,}5$ em decimal. Portanto, tanto o $0{,}6$ quanto o $0{,}7$ serão considerados $0{,}5$.

Exemplo: Considere uma aritmética de ponto flutuante: $F(10,2,-5,5)$

Sejam $x = 4{,}32$ e $y = 0{,}064$. Calcular $x + y$.

As operações soma e subtração na aritmética de ponto flutuante requer o alinhamento dos pontos decimais dos dois números, e ainda, os números devem ter a potência do maior deles.

$$\left.\begin{array}{l} x = 0{,}43 \times 10^1 \\ y = 0{,}0064 \times 10^1 \rightarrow y = 0{,}01 \times 10^1 \end{array}\right\} x + y = 0{,}44 \times 10^1$$

Resultado com dois dígitos na mantissa: $0{,}44 \times 10^1$

Exemplo: Considere uma aritmética de ponto flutuante: $F(10{,}2, -5{,}5)$

Sejam $x = 372$ e $y = 371$. Calcular $x - y$.

A aritmética de ponto flutuante requer o alinhamento dos pontos decimais dos dois números.

$$\left.\begin{array}{l} x = 0{,}37 \times 10^3 \\ y = 0{,}37 \times 10^3 \end{array}\right\} x - y = 0{,}00 \times 10^3$$

Resultado com dois dígitos na mantissa: $0{,}00 \times 10^3$

A partir do momento em que se calcula um resultado por aproximação, é preciso saber como estimar ou delimitar o erro cometido na aproximação. Sem isso, a aproximação obtida não tem significado.

A análise dos resultados obtidos através de um método numérico representa uma etapa fundamental no processo das soluções numéricas.

Para se estimar ou delimitar o erro, recorre-se a dois conceitos:

erro absoluto e **erro relativo.**

Erro Absoluto (EA_x)

O erro absoluto é definido como a diferença gerada pelo número da forma decimal e na forma de ponto flutuante. Dessa forma, temos:

$$EA_x = |x - \tilde{x}|$$

Onde:

- EA_x = erro absoluto de x
- x = valor efetivo do número
- $\tilde{x}$ = valor do número em ponto flutuante

Exemplo: Dado o número $534{,}278$, de valor efetivo, considerando uma mantissa de 4 algarismos teremos que:

$$x = 534{,}278 \quad \rightarrow \quad x = 0{,}534278 \times 10^3$$

$$\tilde{x} = 0{,}5342 \times 10^3$$

$$EA_x = |x - \tilde{x}| = |0{,}534278 \times 10^3 - 0{,}5342 \times 10^3|$$

$$EA_x = 0{,}000078 \times 10^3 = 0{,}78 \times 10^{-1}$$

Erro Relativo (ER_x)

O erro relativo tem esse nome justamente por depender diretamente do número envolvido. A expressão para o erro relativo é dada por:

$$ER_x = \left|\frac{EA_x}{x}\right| = \left|\frac{x - \tilde{x}}{x}\right|$$

Onde:

- EA_x = erro absoluto de x
- x = valor efetivo do número
- $\tilde{x}$ = valor do número em ponto flutuante

Exemplo: Seguindo o exemplo anterior, o erro relativo referente ao número 534,278, em ponto flutuante com a mesma mantissa, será:

$$ER_x = \left|\frac{EA_x}{x}\right| = \left|\frac{x - \tilde{x}}{x}\right|$$

$$ER_x = \left|\frac{0{,}78 \times 10^{-1}}{0{,}534278 \times 10^{3}}\right| = 1{,}460 \times 10^{-4}$$

$$ER_x = 0{,}1460 \times 10^{-3}$$