

Classificação dos Erros. Representação dos números reais

MS211 – Cálculo Numérico – Turma C

Giuseppe Romanazzi

Agosto 2024

Conteúdo

- 1 Classificação dos erros
- 2 Representação dos números reais na aritmética finita, Erro de Representação
- 3 Erro Absoluto e Erro Relativo

Erros

- Erros de modelização
- Erros inerentes
- Erros de truncamento
- Erros de representação dos números (também chamado erro de arredondamento)

Erros de Modelização

A modelagem de um problema depende dos:

- dados disponíveis
- variáveis usadas t, a, b, N (variáveis representativas do problema)
- relações analisadas, exemplo $N = a * t + b$

Obtemos um **Erro de Modelização** se escolhemos variáveis ou dados do problema pouco significativas/representativas, ou se usamos relações erradas entre as variáveis para descrever o problema.

Estes erros ocorrem se interpretamos mal o problema, ou se não encontramos as relações certas entre as variáveis que concorrem no modelo.

Erros Inerentes

Os Erros Inerentes dividem-se em duas categorias

- Erros Sistemáticos
- Erros Fortuitos (random)

Erros Sistemáticos aparecem quando o instrumento de medição dos dados do problema apresenta algum erro de funcionamento ou de calibração. Este leva a medir erroneamente os dados (de input) do problema.

Erros Fortuitos são devidos a erros humanos não previsíveis, por exemplo devido a falta de atenção durante as medidas.

São também fortuitos os erros devidos a variações das condições ambientais (como mudança de tensão elétrica, mudança de umidade, temperatura etc.) que podem afetar as medidas dos dados de input do problema.

Erros de Truncamento

Para determinar a solução exata de um problema matemático como:

- determinar o valor de uma função num certo ponto
- obter a integral definida num intervalo
- a derivada de uma função num ponto
- resolver uma equação etc...

Precisamos de usar uma formula ou um método numérico que normalmente necessita infinitos passos computacionais que apresentam operações simples ($+$, $-$, $*$, $/$).

Erros de Truncamento

Por exemplo se quiséssemos determinar e^x para um dado $x \in \mathbb{R}$ podemos usar a formula da expansão em serie de Taylor do exponencial

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

Para aproximar e^x podemos decidir de **truncar** a soma: consideramos somente os primeiros três termos (mais pesados/representativos) da soma: $1 + x + \frac{x^2}{2}$ é uma aproximação de e^x .

A diferença $|e^x - (1 + x + \frac{x^2}{2})|$ é chamada erro de truncamento da expansão serie de Taylor de e^x usando três termos.

Erros de Truncamento

A diferença em valor absoluto

$$|V_E - V_A|$$

entre o valor exato V_E do problema e de uma sua aproximação V_A obtida truncando um processo (ou método) que determina a solução é chamada **Erro de Truncamento**.

Outro exemplo (erro de truncamento na aproximação da derivada):
considere a formula

$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(x) + \dots$ para obter a
derivada $f'(x)$:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}f''(x) - \frac{h^2}{3!}f'''(x) + \dots$$

Uma aproximação da derivada se pode obter truncando esta
formula usando os termos mais simples: Se aproximamos $f'(x)$
com $\frac{f(x+h)-f(x)}{h}$,

o erro de truncamento será

$$\left| f'(x) - \frac{f(x+h)-f(x)}{h} \right| = \left| \frac{h}{2}f''(x) + \frac{h^2}{6}f'''(x) + \dots \right|.$$

Sistema de Representação dos reais em virgula móvel (ponto flutuante, Floating Point)

Dado um numero x real a sua representação no sistema $FP(\beta, t, \ell, u)$ é $\bar{x}$ definido como segue

$$\bar{x} = \pm m\beta^e \quad (\text{ ou } fl(x) = \pm m\beta^e)$$

onde

- β é a base do sistema
- $m = 0.d_1d_2 \dots d_t$ é chamada mantissa do número $\bar{x}$
- t é o número de dígitos da mantissa
- e expoente que tem de ser contido no intervalo $[\ell, u]$,
 $e \in [\ell, u]$

O sistema de representação $FP(\beta, t, \ell, u)$ é chamado sistema de representação em virgula móvel, ou em ponto flutuante.

Propriedades:

- β, t são inteiros positivos;
- $0 < d_1 \leq \beta - 1$;
- $0 \leq d_i \leq \beta - 1$ para $i = 2, \dots, t$
- ℓ, u são inteiros com $\ell < u$; Normalmente ℓ é negativo e u é positivo;
- O expoente 'e' é um número inteiro no sistema com base β , tal que $e \in [\ell, u]$

Sistema Decimal

O sistema decimal usa $\beta = 10$

- $d_1 \in \{1, 2, 3, \dots, 9\}$
- $d_i \in \{0, 1, 2, 3, \dots, 9\}$, $i > 1$
- 'e' é inteiro e tal que $\ell \leq e \leq u$, por exemplo se $\ell = -99$, $u = 98$ então o expoente 'e' é um inteiro que satisfaz $-99 \leq e \leq 98$.

Sistema Binário

O sistema binário usa $\beta = 2$, $FP(2, t, \ell, u)$

- $d_1 = 1$
- $d_i \in \{0, 1\}$, $i > 1$
- $\ell \leq e \leq u$, por exemplo se $\ell = -11$ e $u = 10$,
 e é um binário com até dois dígitos entre -11 e 10 .
 $e \in \{-11, -10, -1, 0, 1, 10\}$

Transformação de binário a decimal

$$(b_n \ b_{n-1} \ \dots \ b_1 \ b_0)_2 = \sum_{j=0}^n b_j 2^j$$

Exemplos:

$$(110)_2 = 0 * 2^0 + 1 * 2^1 + 1 * 2^2 = 0 + 2 + 4 = 6$$

$$-(1101)_2 = -(1*2^0 + 0*2^1 + 1*2^2 + 1*2^3) = -(1+0+4+8) = -13$$

Pergunta: Dado um número em decimal consegue transformá-lo em binário?

Truncamento e Arredondamento em $FP(\beta, t, \ell, u)$

Quando o número x tem mantissa com mais de t dígitos, a sua aproximação (ou representação) $\bar{x}$ em $FP(\beta, t, \ell, u)$ é obtida ou por truncamento ou por arredondamento a t dígitos significativos.

Quando usamos somente o truncamento dentro o sistema FP , então indicaremos o sistema de representação com $FP(\beta, t, \ell, u, T)$ se sem vez usássemos o arredondamento o sistema é $FP(\beta, t, \ell, u, A)$

Exemplos de Representação em $FP(10, 4, -99, 99)$

<i>Numero</i>	<i>Truncamento</i>	<i>Arredondamento</i>	Representação exata
1.256	$0.1256 * 10^1$	$0.1256 * 10^1$	Sim
10.2053	$0.1020 * 10^2$	$0.1021 * 10^2$	Não
-238.15	$-0.2381 * 10^3$	$-0.2382 * 10^3$	Não
-2.71828	$-0.2718 * 10^1$	$-0.2718 * 10^1$	Não
$7 * 10^{-9}$	$0.7 * 10^{-8}$	$0.7 * 10^{-8}$	Sim
$7.1267 * 10^{-9}$	$0.7126 * 10^{-8}$	$0.7127 * 10^{-8}$	Não
$-3.1437 * 10^2$	$-0.3143 * 10^3$	$-0.3144 * 10^3$	Não
$-3.143 * 10^2$	$-0.3143 * 10^3$	$-0.3143 * 10^3$	Sim
$7 * 10^{-100}$	$0.7 * 10^{-99}$	$0.7 * 10^{-99}$	Sim
$7 * 10^{-101}$	<i>underflow</i>	<i>underflow</i>	Não
10^{99}	<i>overflow</i>	<i>overflow</i>	Não
-10^{99}	<i>overflow</i>	<i>overflow</i>	Não

Maior valor positivo representável é $0.9999 * 10^{99}$,

Menor valor positivo representável é $0.1 * 10^{-99}$,

Maior valor negativo representável é $-0.1 * 10^{-99}$,

Menor valor negativo representável é $-0.9999 * 10^{99}$

Erro de Representação dos números reais

Cada vez que um número real x não é representado exatamente no sistema de ponto flutuante $FP(\beta, t, l, u)$ tem um erro ao considerar a sua representação $fl(x)$ obtida após o truncamento ou arredondamento.

Por exemplo no caso anterior se usássemos $FP(10, 4, -99, 99)$ com Truncamento, o valor $x = 10.2053$ é representado como $fl(x) = 0.1020 \cdot 10^2$ na FP e portanto cometemos um erro.

O erro $|x - fl(x)|$ chama-se **erro de representação** do valor x na $FP(\beta, t, l, u)$.

Os números que são representados exatamente no sistema de ponto flutuante tem um erro de representação nulo porque $x = fl(x)$.

Para os números que não são representados, porque correspondem a situação de overflow ou underflow, não se pode definir o erro de representação.

Erros de representação na aritmética finita

$FP(10, 4, -99, 99)$

Número	Trunc.	Arred.	Erro trunc.	Erro arred.
1.256	$0.1256 * 10^1$	$0.1256 * 10^1$	0	0
10.2053	$0.1020 * 10^2$	$0.1021 * 10^2$	0.0053	0.0047
-238.15	$-0.2381 * 10^3$	$-0.2382 * 10^3$	0.05	0.05
-2.71828	$-0.2718 * 10^1$	$-0.2718 * 10^1$	0.00028	0.00028
$7 * 10^{-9}$	$0.7 * 10^{-8}$	$0.7 * 10^{-8}$	0	0
$7.1267 * 10^{-9}$	$0.7126 * 10^{-8}$	$0.7127 * 10^{-8}$	$0.0007 * 10^{-9}$	$0.0003 * 10^{-9}$
$7 * 10^{-100}$	$0.7 * 10^{-99}$	$0.7 * 10^{-99}$	0	0
$7 * 10^{-101}$	<i>underflow</i>	<i>underflow</i>	Não existe	Não existe
10^{99}	<i>overflow</i>	<i>overflow</i>	Não existe	Não existe
-10^{99}	<i>overflow</i>	<i>overflow</i>	Não existe	Não existe

Valor exato e aproximação de uma medida x

Seja x um valor exato de uma medida ou de um calculo matemático, ou de solução de um problema.

Quando modelamos o problema, ou truncamos um calculo ou usamos um computador para computar a solução: obtemos somente uma aproximação $\tilde{x}$ em vez que o valor exato x do problema.

Erro absoluto e erro relativo

Definição do Erro Absoluto

O valor absoluto da diferença entre o valor exato x e a sua aproximação $\bar{x}$ chama-se erro absoluto:

$$E_A(x) := |x - \bar{x}|$$

Propriedade: $|x - \bar{x}| \geq 0$,

O erro absoluto é sempre positivo ou nulo.

Os erros normalmente são consideradas como quantidades positivas ou nulas.

Não obstante isso, as vezes será importante saber se x for maior ou menor do valor aproximado $\bar{x}$ e portanto se analisará o erro sem valor absoluto.

Exemplo

Sabemos que as medidas $\bar{x} = 2112.9 \text{ cm}$ e $\bar{y} = 5.3 \text{ cm}$ aproximam respectivamente x e y a menos de um erro absoluto $E_A < 0.1 \text{ cm}$. Que valor podem tomar x e y ?

Resposta: $x \in (2112.8, 2113)$,
 $y \in (5.2, 5.4)$

Erro relativo

E' possível dizer qual é a melhor aproximação entre $\bar{x} \approx x$ e $\bar{y} \approx y$ se bem eles tem o mesmo erro absoluto?

A resposta é sim mas temos de analisar o erro relativo, ou seja escalar o erro absoluto respeito o tamanho do valor que queremos aproximar:

um erro de 0.1 cm tem mais peso (é mais significativo) no aproximar uma barra de comprimento 5.2 cm

que no aproximar uma barra de comprimento $2112.8 \text{ cm} = 21.128 \text{ m}$

Erro relativo

Definição do Erro Relativo

O valor absoluto da diferença entre o valor exato x e a sua aproximação $\bar{x}$ dividida por o valor absoluto do valor aproximado $|\bar{x}|$ chama-se erro relativo:

$$E_R(x) := \frac{|x - \bar{x}|}{|\bar{x}|}$$

Nalgum livro pode encontrar a definição do erro relativo dividindo por $|x|$.

Se $x = 2112.8$, $\bar{x} = 2112.9$ tem

$$E_R(x) = \frac{|x - \bar{x}|}{|\bar{x}|} = \frac{0.1}{2112.9} \approx 4.7 \cdot 10^{-5} = 0.000047$$

E' um erro relativo de 0.0047% respeito $\bar{x}$. Se $y = 5.2$, $\bar{y} = 5.3$

$$\text{tem } E_R(y) = \frac{|y - \bar{y}|}{|\bar{y}|} = \frac{0.1}{5.3} \approx 0.01887$$

E' um erro relativo de 1.887% respeito $\bar{y}$.