Cálculo Numérico

Um Livro Colaborativo

23 de agosto de 2016

Organizadores

Dagoberto Adriano Rizzotto Justo - UFRGS

Esequia Sauter - UFRGS

Fabio Souto de Azevedo - UFRGS

Leonardo Fernandes Guidi - UFRGS

Matheus Correia dos Santos - UFRGS

Pedro Henrique de Almeida Konzen - UFRGS

Licença

Este trabalho está licenciado sob a Licença Creative Commons Atribui
CompartilhaIgual 3.0 Não Adaptada. Para ver uma cópia desta licença, visite http://creativecommons.org/licenses/by-sa/3.0/ ou envie uma carta para Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Nota dos organizadores

Este livro vem sendo construído de forma colaborativa desde 2011. Nosso intuito é melhorá-lo, expandi-lo e adaptá-lo às necessidades de um curso de cálculo numérico em nível de graduação.
Caso queira colaborar, tenha encontrado erros, tenha sugestões ou reclamações, entre em contato conosco pelo endereço de e-mail:

`livro_colaborativo@googlegroups.com`

Alternativamente, abra um chamado no repositório GitHub do projeto:

https://github.com/livroscolaborativos/CalculoNumerico

Prefácio

Este livro busca abordar os tópicos de um curso de introdução ao cálculo numérico moderno oferecido a estudantes de matemática, física, engenharias e outros. A ênfase é colocada na formulação de problemas, implementação em computador da resolução e interpretação de resultados. Pressupõe-se que o estudante domine conhecimentos e habilidades típicas desenvolvidas em cursos de graduação de cálculo, álgebra linear e equações diferenciais. Conhecimentos prévios em linguagem de computadores é fortemente recomendável, embora apenas técnicas elementares de programação sejam real-

mente necessárias.
Ao longo do livro, fazemos ênfase na utilização do **software** livre `Scilab` para a implementação dos métodos numéricos abordados. Recomendamos que o leitor tenha à sua disposição um computador com o `Scilab` instalado. Não é necessário estar familiarizado com a linguagem `Scilab`, mas recomendamos a leitura do Apêndice **??**, no qual apresentamos uma rápida introdução a este pacote computacional. Alternativamente, existem algumas soluções em nuvem que fornecem acesso ao Scilab via internet. Por exemplo, a plataforma virtual rollApp (`https://www.rollapp.com/app/scilab`).

Capítulo 1

Introdução

Cálculo numérico é a disciplina que estuda as técnicas para a solução aproximada de problemas matemáticos. Estas técnicas são de natureza analítica e computacional. As principais preocupações normalmente envolvem exatidão e perfórmance.

Aliado ao aumento contínuo da capacidade de computação dispo-

nível, o desenvolvimento de métodos numéricos tornou a simulação computacional de modelos matemáticos uma prática usual nas mais diversas áreas científicas e tecnológicas. As então chamadas simulações numéricas são constituídas de um arranjo de vários esquemas numéricos dedicados a resolver problemas específicos como, por exemplo: resolver equações algébricas, resolver sistemas lineares, interpolar e ajustar pontos, calcular derivadas e integrais, resolver equações diferenciais ordinárias, etc.. Neste livro, abordamos o desenvolvimento, a implementação, utilização e aspectos teóricos de métodos numéricos para a resolução desses problemas.

Os problemas que discutiremos não formam apenas um conjunto de métodos fundamentais, mas são, também, problemas de interesse na engenharia e na matemática aplicada. A necessidade de aplicar aproximações numéricas decorre do fato de que esses problemas podem se mostrar intratáveis se dispomos apenas de meios puramente analíticos, como aqueles estudados nos cursos de cálculo e álgebra linear. Por exemplo, o teorema de Abel-Ruffini nos garante que não existe uma fórmula algébrica, isto é, envolvendo apenas opera-

ções aritméticas e radicais, para calcular as raízes de uma equação polinomial de qualquer grau, mas apenas casos particulares:

- Simplesmente isolar a incógnita para encontrar a raiz de uma equação do primeiro grau;
- Fórmula de Bhaskara para encontrar raízes de uma equação do segundo grau;
- Fórmula de Cardano para encontrar raízes de uma equação do terceiro grau;
- Existe expressão para equações de quarto grau;
- Casos simplificados de equações de grau maior que 4 onde alguns coeficientes são nulos também podem ser resolvidos.

Equações não polinomiais podem ser ainda mais complicadas de resolver exatamente, por exemplo:

$$\cos(x) = x \qquad \text{e} \qquad xe^x = 10$$

Para resolver o problema de valor inicial

$$y' + xy = x,$$

$$y(0) = 2,$$

podemos usar o método de fator integrante e obtemos $y = 1+e^{-x^2/2}$.
Já o cálculo da solução exata para o problema

$$y' + xy = e^{-y},$$

$$y(0) = 2,$$

não é possível.
Da mesma forma, resolvemos a integral

$$\int_1^2 xe^{-x^2}dx$$

pelo método da substituição e obtemos $\frac{1}{2}(e^{-1} - e^{-2})$. Porém a integral

$$\int_1^2 e^{-x^2}dx$$

não pode ser resolvida analiticamente.
A maioria dos modelos de fenômenos reais chegam em problemas matemáticos onde a solução analítica é difícil (ou impossível) de ser encontrada, mesmo quando provamos que ela existe. Nesse curso propomos calcular aproximações numéricas para esses problemas, que apesar de, em geral, serem diferentes da solução exata, mostraremos que elas podem ser bem próximas.
Para entender a construção de aproximações é necessário estudar um pouco como funciona a aritmética de computador e erros de arredondamento. Como computadores, em geral, usam uma base binária para representar números, começaremos falando em mudança de base.

Capítulo 2

Aritmética de máquina

2.1 Sistema de numeração e mudança de base

Usualmente, utilizamos o sistema de numeração decimal para representar números. Esse é um sistema de numeração posicional onde a posição do dígito indica a potência de 10 que o dígito está representando.

Exemplo 1. O número 293 é decomposto como

$$\begin{aligned} 293 &= 2 \text{ centenas} + 9 \text{ dezenas} + 3 \text{ unidades} \\ &= 2 \cdot 10^2 + 9 \cdot 10^1 + 3 \cdot 10^0. \end{aligned}$$

O sistema de numeração posicional também pode ser usado com outras bases. Vejamos a seguinte definição.

Definição 1 (Sistema de numeração de base b)**.** *Dado um número natural $b > 1$ e o conjunto de símbolos $\{,,-,\mathbf{0},\mathbf{1},\mathbf{2},\ldots,\boldsymbol{b-1}\}$[a], a sequência de símbolos*

$$(d_n d_{n-1} \cdots d_1 d_0,d_{-1}d_{-2}\cdots)_b$$

representa o número positivo

$$d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots$$

Para representar números negativos usamos o símbolo $-$ a esquerda do numeral.

---

[a] Para $b > 10$, veja a Observação 1

Observação 1 ($b \geq 10$). Para sistemas de numeração com base $b \geq 10$ é usual utilizar as seguintes notações:

- No sistema de numeração decimal ($b = 10$), costumamos representar o número sem os parênteses e o subíndice, ou seja,

  $$\pm d_n d_{n-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots := \pm(d_n d_{n-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots)_{10}$$

- Se $b > 10$, usamos as letras $A, B, C, \cdots$ para completar os símbolos: $A = 10$, $B = 11$, $C = 12$, $D = 13$, $E = 14$, $F = 15$.

Exemplo 2 (Sistema binário)**.** O sistema de numeração em base dois é chamado de binário e os algarismos binários são conhecidos como *bits*, do inglês **binary digits**. Um *bit* pode assumir dois

valores distintos: 0 ou 1. Por exemplo:

$$\begin{aligned} x &= (1001{,}101)_2 \\ &= 1\cdot 2^3 + 0\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0 + 1\cdot 2^{-1} + 0\cdot 2^{-2} + 1\cdot 2^{-3} \\ &= 8 + 0 + 0 + 1 + 0{,}5 + 0 + 0{,}125 = 9{,}625 \end{aligned}$$

Ou seja, $(1001{,}101)_2$ é igual a 9,625 no sistema decimal.

Exemplo 3 (Sistema quaternário)**.** No sistema quaternário a base b é igual a 4. Por exemplo:

$$(301{,}2)_4 = 3\cdot 4^2 + 0\cdot 4^1 + 1\cdot 4^0 + 2\cdot 4^{-1} = 49{,}5$$

Exemplo 4 (Sistema octal)**.** No sistema octal a base é $b = 8$ e utilizamos os símbolos em $\{0, 1, 2, 3, 4, 5, 6, 7\}$. Por exemplo:

$$\begin{aligned} (1357{,}24)_8 &= 1\cdot 8^3 + 3\cdot 8^2 + 5\cdot 8^1 + 7\cdot 8^0 + 2\cdot 8^{-1} + 4\cdot 8^{-2} \\ &= 512 + 192 + 40 + 7 + 0{,}25 + 0{,}0625 = 751{,}3125 \end{aligned}$$

Exemplo 5 (Sistema hexadecimal)**.** O sistema de numeração cuja a base é $b = 16$ é chamado de sistema hexadecimal. O conjunto de símbolos necessários é $S = \{“,”, -, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E$
Convertendo o número $(E2AC)_{16}$ para a base 10 temos

$$\begin{aligned}(E2AC)_{16} &= 14 \cdot 16^3 + 2 \cdot 16^2 + 10 \cdot 16^1 + 12 \cdot 16^0 \\ &= 57344 + 512 + 160 + 12 = 58028\end{aligned}$$

Exemplo 6 (Scilab)**.** O `Scilab` oferece algumas funções para a conversão de números inteiros em dada base para a base decimal. Por exemplo, temos:

```
-->bin2dec('1001')
 ans  =
    9.
-->hex2dec('451')
 ans  =
    1105.
-->oct2dec('157')
```

```
 ans  =
    111.
-->base2dec('BEBA',16)
 ans  =
    48826.
```

A partir da Definição 1 acabamos de mostrar vários exemplos de conversão de números de uma sistema de numeração de base b para o sistema decimal. Agora, vamos estudar como fazer o processo inverso. Isto é, dado um número decimal $(X)_{10}$ queremos escrevê-lo em uma outra base b, i.e., queremos obter a seguinte representação:

$$\begin{aligned}(X)_{10} &= (d_n d_{n-1} \cdots d_0, d_{-1} \cdots)_b \\ &= d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots\end{aligned}$$

Separando as partes inteira e fracionária de X, i.e. $X = X^{\mathrm{i}} + X^{\mathrm{f}}$, temos:

$$X^{\mathrm{i}} = d_n \cdot b^n + \cdots + d_{n-1} b^{n-1} \cdot + d_1 \cdot b^1 + d_0 \cdot b^0 \qquad \text{e} \qquad X^{\mathrm{f}} = \frac{d_{-1}}{b^1} + \frac{d_{-2}}{b^2} + \cdots$$

Nosso objetivo é determinar os algarismos $\{d_n, d_{n-1}, ...\}$.
Primeiramente, vejamos como tratar a parte inteira X^{i}. Calculando sua divisão por b, temos:

$$\frac{X^{\mathrm{i}}}{b} = \frac{d_0}{b} + d_1 + d_2 b^1 \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}.$$

Observe que d_0 é o resto da divisão de X^{i} por b, pois $d_1 + d_2 b^1 \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ é inteiro e $\frac{d_0}{b}$ é uma fração (lembramos que $d_0 < b$). Da mesma forma, o resto da divisão de $d_1 + d_2 b^1 \cdots + d_{n-1} \cdot b^{n-2} + d_n \cdot b^{n-1}$ por b é d_1. Repetimos o processo até encontrar os símbolos $d_0, d_1, d_2, \ldots$.

Exemplo 7 (Conversão da parte inteira)**.** Vamos escrever o número 125 na base 6. Para tanto, fazemos sucessivas divisões por 6 como segue:

$$\begin{aligned} 125 &= 20 \cdot 6 + 5 \quad \text{(125 dividido por 6 é igual a 20 e resta 5)} \\ &= (3 \cdot 6 + 2) \cdot 6 + 5 = 3 \cdot 6^2 + 2 \cdot 6 + 5, \end{aligned}$$

logo $125 = (325)_6$.
Estes cálculos podem ser feitos no `Scilab` com o auxílio das funções `modulo` e `int`. A primeira calcula o resto da divisão entre dois números, enquanto que a segunda retorna a parte inteira de um número dado. No nosso exemplo, temos:

```
-->q = 125, d0 = modulo(q,6)
-->q = int(q/6), d1 = modulo(q,6)
-->q = int(q/6), d2 = modulo(q,6)
```

Verifique!

Exemplo 8 (Scilab)**.** O `Scilab` oferece algumas funções para a conversão de números inteiros em dada base para a base decimal. Assim, temos:

```
-->bin2dec('1001')
 ans  =
    9.
-->hex2dec('451')
```

```
 ans  =
    1105.
-->oct2dec('157')
 ans  =
    111.
-->base2dec('BEBA',16)
 ans  =
    48826.
```

Vamos converter a parte fracionária de um número decimal em uma dada base b. Usando a notação $X = X^{\mathrm{i}} + X^{\mathrm{f}}$ para as partes inteira e fracionária, respectivamente, temos:

$$bX^{\mathrm{f}} = d_{-1} + \frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$$

Observe que a parte inteira desse produto é d_{-1} e $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ é a parte fracionária. Quando multiplicamos $\frac{d_{-2}}{b} + \frac{d_{-3}}{b^2} + \cdots$ por b novamente, encontramos d_{-2}. Repetimos o processo até encontrar todos os símbolos.

Exemplo 9 (Conversão da parte fracionária)**.** Escrever o número $125{,}58\overline{3}$ na base 6. Do exemplo anterior temos que $125 = (325)_6$. Assim, nos resta converter a parte fracionária. Para tanto, fazemos sucessivas multiplicações por 6 como segue:]

$$\begin{aligned} 0{,}58\overline{3} &= 3{,}5 \cdot 6^{-1} \quad (0{,}58\overline{3} \text{ multiplicado por 6 é igual a } 3{,}5) \\ &= 3 \cdot 6^{-1} + 0{,}5 \cdot 6^{-1} \\ &= 3 \cdot 6^{-1} + (3 \cdot 6^{-1}) \cdot 6^{-1} \\ &= 3 \cdot 6^{-1} + 3 \cdot 6^{-2}, \end{aligned}$$

logo $0{,}58\overline{3} = (0{,}33)_6$. As contas feitas aqui, também podem ser feitas no `Scilab`. Você sabe como?

Uma maneira de converter um número dado numa base b_1 para uma base b_2 é fazer em duas partes: primeiro converter o número dado na base b_2 para base decimal e depois converter para a base b_1.

Exercícios

E 2.1.1. Converta para base decimal cada um dos seguintes números:

a) $(100)_2$ c) $(100)_b$ e) $(AA)_{16}$ g) $(3{,}12)_5$

b) $(100)_3$ d) $(12)_5$ f) $(7{,}1)_8$

E 2.1.2. Escreva os números abaixo na base decimal.

a) $(25{,}13)_8$

b) $(101{,}1)_2$

c) $(12F{,}4)_{16}$

d) $(11{,}2)_3$

E 2.1.3. Escreva cada número decimal na base b.

a) $7{,}\overline{6}$ na base $b = 5$

b) $29{,}1\overline{6}$ na base $b = 6$

E 2.1.4. Escreva cada número dado para a base b.

a) $(45{,}1)_8$ para a base $b = 2$

b) $(21{,}2)_8$ para a base $b = 16$

c) $(1001{,}101)_2$ para a base $b = 8$

d) $(1001{,}101)_2$ para a base $b = 16$

E 2.1.5. Escreva o número $x = 5{,}5$ em base binária.

E 2.1.6. Escreva o número $x = 17{,}109375$ em base hexadecimal (16).

E 2.1.7. Quantos algarismos são necessários para representar o número 937163832173947 em base binária? E em base 7? Dica: Qual é o menor e o maior inteiro que pode ser escrito em dada base com N algarismos?

E 2.1.8. Escreva $x = (12.4)_8$ em base decimal e binária.

2.2 Representação de números

Os computadores, em geral, usam a base binária para representar os números, onde as posições, chamadas de bits, assume as condições "verdadeiro" ou "falso", ou seja, 0 ou 1. Cada computador tem um número de bits fixo e, portanto, representa uma quantidade finita de números. Os demais números são tomados por proximidade àqueles conhecidos, gerando erros de arredondamento. Por exemplo, em aritmética de computador, o número 2 tem representação exata, logo $2^2 = 4$, mas $\sqrt{3}$ não tem representação finita, logo $(\sqrt{3})^2 \neq 3$. Veja isso no `Scilab`:

```
-->2^2 == 4
 ans  =
  T
-->sqrt(3)^2 == 3
 ans  =
  F
```

2.2.1 Números inteiros

Tipicamente um número inteiro é armazenado num computador como uma sequência de dígitos binários de comprimento fixo denominado **registro**.

Representação sem sinal

Um registro com n bits da forma

d_{n-1}	d_{n-2}	$\cdots$	d_1	d_0

representa o número $(d_{n-1}d_{n-2}...d_1d_0)_2$.
Assim é possível representar números inteiros entre

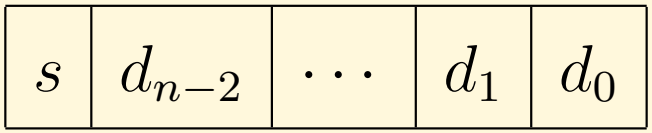
$$\begin{aligned}
(111...111)_2 &= 2^{n-1} + 2^{n-2} + \cdots + 2^1 + 2^0 = 2^n - 1. \\
\vdots \qquad &= \\
(000...011)_2 &= (3)_{10} \\
(000...010)_2 &= (2)_{10} \\
(000...001)_2 &= (1)_{10} \\
(000...000)_2 &= (0)_{10}
\end{aligned}$$

Exemplo 10. No `Scilab`,

```
-->uint8( bin2dec('00000011') )
   ans = 3
-->uint8( bin2dec('11111110') )
   ans = 254
```

Representação com bit de sinal

O bit mais significativo (o primeiro à esquerda) representa o sinal: por convenção, 0 significa positivo e 1 significa negativo. Um registro com n bits da forma



representa o número $(-1)^s(d_{n-2}...d_1d_0)_2$. Assim é possível representar números inteiros entre -2^{n-1} e 2^{n-1}, com duas representações para o zero: $(1000...000)_2$ e $(00000...000)_2$.

Exemplo 11. Em um registro com 8 bits, teremos os números

$$
\begin{aligned}
(11111111)_2 &= -(2^6 + \cdots + 2 + 1) = -127 \\
&\vdots \\
(10000001)_2 &= -1 \\
(10000000)_2 &= -0 \\
(01111111)_2 &= 2^6 + \cdots + 2 + 1 = 127 \\
&\vdots \\
(00000010)_2 &= 2 \\
(00000001)_2 &= 1 \\
(00000000)_2 &= 0
\end{aligned}
$$

Representação complemento de dois

O bit mais significativo (o primeiro à esquerda) representa o coeficiente de -2^{n-1}. Um registro com n bits da forma

d_{n-1}	d_{n-2}	$\cdots$	d_1	d_0

representa o número $-d_{n-1}2^{n-1} + (d_{n-2}...d_1d_0)_2$.
Note que todo registro começando com 1 será um número negativo.

Exemplo 12. O registro com 8 bits [01000011] representa o número

$$-0(2^7) + (1000011)_2 = 64 + 2 + 1 = 67.$$

O registro com 8 bits [10111101] representa o número

$$-1(2^7) + (0111101)_2 = -128 + 32 + 16 + 8 + 4 + 1 = -67.$$

Note que podemos obter a representação de -67 invertendo os dígitos de 67 em binário e somando 1.

Exemplo 13. Em um registro com 8 bits, teremos os números

$$
\begin{aligned}
(11111111)_2 &= -2^7 + 2^6 + \cdots + 2 + 1 = -1 \\
&\vdots \\
(10000001)_2 &= -2^7 + 1 = -127 \\
(10000000)_2 &= -2^7 = -128 \\
(01111111)_2 &= 2^6 + \cdots + 2 + 1 = 127 \\
&\vdots \\
(00000010)_2 &= 2 \\
(00000001)_2 &= 1 \\
(00000000)_2 &= 0
\end{aligned}
$$

Exemplo 14. No `Scilab`,

```
-->int8( bin2dec('00000011') )
```

```
   ans = 3
-->int8( bin2dec('11111110') )
   ans = -2
```

2.2.2 Sistema de ponto fixo

O sistema de ponto fixo representa as partes inteira e fracionária do número com uma quantidade fixas de dígitos.

Exemplo 15. Em um computador de 32 bits que usa o sistema de ponto fixo, o registro

d_{31}	d_{30}	d_{29}	$\cdots$	d_1	d_0

pode representar o número

- $(-1)^{d_{31}}(d_{30}d_{29}\cdots d_{17}d_{16}, d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal for representado por um dígito. Observe que nesse caso o zero possui duas representações possíveis:

  $$10000000000000000000000000000000$$

  e

  $$00000000000000000000000000000000$$

- $(d_{30}d_{29}\cdots d_{17}d_{16})_2 - d_{31}(2^{15} - 2^{-16}) + (0{,}d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal do número estiver representado por uma implementação em complemento de um. Observe que o zero também possui duas representações possíveis:

  $$11111111111111111111111111111111$$

  e

  $$00000000000000000000000000000000$$

- $(d_{30}d_{29}\cdots d_{17}d_{16})_2 - d_{31}2^{15} + (0{,}d_{15}d_{14}\cdots d_1d_0)_2$ se o sinal do número estiver representado por uma implementação em complemento de dois. Nesse caso o zero é unicamente representado por

  $$00000000000000000000000000000000$$

Observe que 16 dígitos são usados para representar a parte fracionária, 15 são para representar a parte inteira e um dígito, o d_{31}, está relacionado ao sinal do número.

2.2.3 Normalização

Os números $h = 6.626 \times 10^{-34}$ e $N_A = 6.0221 \times 10^{23}$ não podem ser armazenados na máquina em ponto fixo do exemplo anterior.

Entretanto, a constante

$$h = 6626 \times 10^{-37}$$
$$h = 6.626 \times 10^{-34}$$
$$h = 0.6626 \times 10^{-33}$$
$$h = 0.006626 \times 10^{-31}$$

pode ser escrita de várias formas diferentes. Para termos uma **representação única** definimos como notação normalizada a segunda opção ($1 \leq m < 10$) que apresenta apenas um dígito diferente de zero a esquerda do ponto decimal ($m = 6.626$).

Definição 2. *Definimos que*

$$x = (-1)^s (M)_b \times b^E,$$

está na ***notação normalizada***[a] *quando* $1 \leq (M)_b < b$, *onde*

- s *é o* ***sinal*** *(*0 *para positivo e* 1 *para negativo),*
- E *é o* ***expoente****,*
- b *é a base (por ex.* 2, 8, 10 *ou* 16*),*
- $(M)_b$ *é o significando. O* ***significando*** *(também chamado de mantissa ou coeficiente) contém os dígitos significativos do número.*

---

[a] Em algumas referências é usado $(0.1)_b \leq (M)_b < 1$.

Exemplo 16. Os números abaixo estão em notação normalizada:

$$x_1 = (-1.011101)_2 \times 2^{(100)_2}$$

$$x_2 = (-2.325)_{10} \times 10^1$$

Exemplo 17. Represente os números $0{,}00\overline{51}$ e $1205{,}41\overline{54}$ em um sistema de ponto fixo de 4 dígitos para a parte inteira e 4 dígitos para a parte fracionária. Depois represente os mesmos números utilizando notação normalizada com 7 dígitos significativos.

Solução. As representações dos números $0{,}00\overline{51}$ e $1205{,}41\overline{54}$ no sistema de ponto fixo são 0,0051 e 1205,4154, respectivamente. Em notação normalizada, as representações são $5{,}151515 \cdot 10^{-3}$ e $1{,}205415 \cdot 10^3$, respectivamente. $\Diamond$

Observação 2. No `Scilab`, a representação em ponto flutuante com n dígitos é dada na forma $\pm d_1{,}d_2 d_3 \ldots d_n \times 10^E$. Consulte sobre o comando `format`!

2.2.4 Sistema de ponto flutuante

O sistema de ponto flutuante não possui quantidade fixa de dígitos para as partes inteira e fracionária do número.
Podemos definir uma máquina F em ponto flutuante de dois modos:

$$F(\beta,|M|,|E|,BIAS) \text{ ou } F(\beta,|M|,E_{MIN},E_{MAX})$$

onde

- β é a base (em geral 2 ou 10),
- $|M|$ é o número de dígitos da mantissa,
- $|E|$ é o número de dígitos do expoente,
- $BIAS$ é um valor de deslocamento do expoente (veja a seguir),
- E_{MIN} é o menor expoente,
- E_{MAX} é o maior expoente.

Considere uma máquina com um registro de 64 bits e base $\beta = 2$. Pelo padrão IEEE754, 1 bit é usado para o sinal, 11 bits para o expoente e 52 bits são usados para o significando tal que

s	c_{10}	c_9	$\cdots$	c_0	m_1	m_2	$\cdots$	m_{51}	m_{52}

represente o número (o $BIAS = 1023$ por definição)

$$x = (-1)^s M \times 2^{c-BIAS},$$

onde a **característica** é representada por

$$c = (c_{10}c_9 \cdots c_1c_0)_2 = c_{10}2^{10} + \cdots + c_1 2^1 + c_0 2^0$$

e o significando por

$$M = (1.m_1m_2 \cdots m_{51}m_{52})_2.$$

Em base 2 não é necessário armazenar o primeiro dígito (por quê?).

Por exemplo, o registro

$$[0|100\ 0000\ 0000|1010\ 0000\ 0000...0000\ 0000]$$

representa o número

$$(-1)^0(1+2^{-1}+2^{-3})\times 2^{1024-1023} = (1+0.5+0.125)2 = 3.25.$$

O expoente deslocado

Uma maneira de representar os expoentes inteiros é deslocar todos eles uma mesma quantidade. Desta forma permitimos a representação de números negativos e a ordem deles continua crescente. O expoente é representado por um inteiro sem sinal do qual é deslocado o **BIAS**.
Tendo $|E|$ dígitos para representar o expoente, geralmente o $BIAS$ é predefinido de tal forma a dividir a tabela ao meio de tal forma que o expoente *um* seja representado pelo sequência [100...000].

Exemplo 18. Com 64 bits, pelo padrão $IEEE754$, temos que $|E| := 11$. Assim $(100\ 0000\ 0000)_2 = 2^{10} = 1024$. Como queremos que esta sequência represente o 1, definimos $BIAS := 1023$, pois

$$1024 - BIAS = 1.$$

Com 32 bits, temos $|E| := 8$ e $BIAS := 127$. E com 128 bits, temos $|E| := 15$ e $BIAS := 16383$.

Com 11 bits temos

$$
\begin{aligned}
[111\ 1111\ 1111] &= reservado \\
[111\ 1111\ 1110] &= 2046 - BIAS = 1023_{10} = E_{MAX} \\
&\vdots = \\
[100\ 0000\ 0001] &= 2^{10} + 1 - BIAS = 2_{10} \\
[100\ 0000\ 0000] &= 2^{10} \quad\ - BIAS = 1_{10} \\
[011\ 1111\ 1111] &= 1023 \quad - BIAS = 0_{10} \\
[011\ 1111\ 1110] &= 1022 \quad - BIAS = -1_{10} \\
&\vdots = \\
[000\ 0000\ 0001] &= 1 - BIAS = -1022 = E_{MIN} \\
[000\ 0000\ 0000] &= reservado
\end{aligned}
$$

O maior expoente é dado por $E_{MAX} = 1023$ e o menor expoente é dado por $E_{MIN} = -1022$.

O menor número representável positivo é dado pelo registro

$$[0|000\ 0000\ 0001|0000\ 0000\ 0000...0000\ 0000]$$

quando $s = 0$, $c = 1$ e $M = (1.000...000)_2$, ou seja,

$$MINR = (1 + 0)_2 \times 2^{1-1023} \approx 0.2225 \times 10^{-307}.$$

O maior número representável é dado por

$$[0|111\ 1111\ 1110|1111\ 1111\ ...1111\ 1111]$$

quando $s = 0$, $c = 2046$ e $M = (1.1111\ 1111....1111)_2 = 2 - 2^{-52}$, ou seja,

$$MAXR = (2 - 2^{-52}) \times 2^{2046-1023} \approx 2^{1024} \approx 0.17977 \times 10^{309}.$$

Casos especiais

O **zero** é um caso especial representado pelo registro

$$[0|000\ 0000\ 0000|0000\ 0000\ 0000...0000\ 0000]$$

Os expoentes **reservados** são usados para casos especiais:

- $c = [0000...0000]$ é usado para representar o zero (se $m = 0$) e os números subnormais (se $m \neq 0$).

- $c = [1111...1111]$ é usado para representar o infinito (se $m = 0$) e NaN (se $m \neq 0$).

Os números subnormais[1] tem a forma

$$x = (-1)^s(0.m_1m_2 \cdots m_{51}m_{52})_2 \times 2^{1-BIAS}.$$

[1]Note que poderíamos definir números um pouco menores que o $MINR$.

Observação 3. O menor número positivo, o maior número e o menor número subnormal representáveis no `Scilab` são:

```
-->MINR=number_properties('tiny')
-->MAXR=number_properties('huge')
-->number_properties('tiniest')
```

Outras informações sobre a representação em ponto flutuante podem ser obtidas com `help number_properties`.

2.2.5 A precisão e o epsilon de máquina

A **precisão** p de uma máquina é o número de dígitos significativos usado para representar um número. Note que $p = |M| + 1$ em binário e $p = |M|$ para outras bases.
O **epsilon de máquina**, $\epsilon_{mach} = \epsilon$, é definido como o menor número representável tal que $1 + \epsilon$ seja diferente de 1.

Exemplo 19. Com 64 bits, temos que o epsilon será dado por

$$\begin{array}{rl} 1 \rightarrow & (1.0000\ 0000....0000)_2 \times 2^0 \\ \epsilon \rightarrow & +(0.0000\ 0000....0001)_2 \times 2^0 = 2^{-52} \\ \hline & (1.0000\ 0000....0001)_2 \times 2^0 \neq 1 \end{array}$$

Assim $\epsilon = 2^{-52}$.

2.2.6 A distribuição dos números

Utilizando uma máquina em ponto flutuante temos um número finito de números que podemos representar.

Um número muito pequeno geralmente é aproximado por zero (underflow) e um número muito grande (overflow) geralmente faz o cálculo parar. Além disso, os números não estão uniformemente espaçados no eixo real. Números pequenos estão bem próximos enquanto que números com expoentes grandes estão bem distantes.

Se tentarmos armazenar um número que não é representável, devemos utilizar o número mais próximo, gerando os erros de arredondamento.

Por simplicidade, a partir daqui nós adotaremos $b = 10$.

Observação 4. O chamado modo de exceção de ponto flutuante é controlado pela função `ieee`. O padrão do `Scilab` é `ieee(0)`. Estude os seguintes resultados das seguintes operações usando os diferentes modos de exceção:

```
-->2*number_properties('huge'), 1/2^999, 1/0, 1/-0
```

Exercícios

E 2.2.1. Explique a diferença entre o sistema de ponto fixo e ponto flutuante.

2.3 Tipos de Erros

Em geral, os números não são representados de forma exata nos computadores. Isto nos leva ao chamado erro de arredondamento. Quando resolvemos problemas com técnicas numéricas estamos sujeitos a este e outros tipos de erros. Nesta seção, veremos quais são estes erros e como controlá-los, quando possível.
Quando fazemos aproximações numéricas, os erros são gerados de várias formas, sendo as principais delas as seguintes:

1. **Incerteza dos dados**: equipamentos de medição possuem precisão finita, acarretando erros nas medidas físicas.

2. **Erros de Arredondamento**: são aqueles relacionados com as limitações que existem na forma representar números de máquina.

3. **Erros de Truncamento**: surgem quando aproximamos um procedimento formado por uma sequência infinita de passos

através de um procedimento finito. Por exemplo, a definição de integral é dada por uma soma infinita e a aproximamos por um soma finita. O erro de truncamento deve ser analisado para cada método empregado.

Uma questão fundamental é a quantificação dos erros que estamos sujeitos ao computar a solução de um dado problema. Para tanto, precisamos definir medidas de erros (ou de exatidão). As medidas de erro mais utilizadas são o **erro absoluto** e o **erro relativo**.

Definição 3 (Erro absoluto e relativo)**.** *Seja x um número real e $\overline{x}$ sua aproximação. O* ***erro absoluto*** *da aproximação $\overline{x}$ é definido como*

$$|x - \overline{x}|.$$

O ***erro relativo*** *da aproximação $\overline{x}$ é definido como*

$$\frac{|x - \overline{x}|}{|x|}, \quad x \neq 0.$$

Observação 5. Observe que o erro relativo é adimensional e, muitas vezes, é dado em porcentagem. Mais precisamente, o erro relativo em porcentagem da aproximação $\overline{x}$ é dado por

$$\frac{|x-\bar{x}|}{|x|} \times 100\%.$$

Exemplo 20. Sejam $x = 123456{,}789$ e sua aproximação $\bar{x} = 123000$. O erro absoluto é

$$|x-\bar{x}| = |123456{,}789 - 123000| = 456{,}789$$

e o erro relativo é

$$\frac{|x-\bar{x}|}{|x|} = \frac{456{,}789}{123456{,}789} \approx 0{,}00369999 \text{ ou } 0{,}36\%$$

Exemplo 21. Sejam $y = 1{,}23456789$ e $\bar{y} = 1{,}13$. O erro absoluto é

$$|y-\bar{y}| = |1{,}23456789 - 1{,}13| = 0{,}10456789$$

que parece pequeno se compararmos com o exemplo anterior. Entretanto o erro relativo é

$$\frac{|y-\bar{y}|}{|y|} = \frac{0{,}10456789}{1{,}23456789} \approx 0{,}08469999 \text{ ou } 8{,}4\%$$

Note que o erro relativo leva em consideração a escala do problema.

Exemplo 22. Observe os erros absolutos e relativos em cada caso

x	$\bar{x}$	erro absoluto	erro relativo
$0{,}\overline{3} \cdot 10^{-2}$	$0{,}3 \cdot 10^{-2}$	$0{,}\overline{3} \cdot 10^{-3}$	$\frac{0{,}\overline{3}\cdot 10^{-3}}{0{,}\overline{3}\cdot 10^{-2}} = 10^{-1} = 10\%$
$0{,}\overline{3}$	$0{,}3$	$0{,}\overline{3} \cdot 10^{-1}$	$\frac{0{,}\overline{3}\cdot 10^{-1}}{0{,}\overline{3}} = 10^{-1} = 10\%$
$0{,}\overline{3} \cdot 10^{2}$	$0{,}3 \cdot 10^{2}$	$0{,}\overline{3} \cdot 10^{1}$	$\frac{0{,}\overline{3}\cdot 10^{1}}{0{,}\overline{3}\cdot 10^{2}} = 10^{-1} = 10\%$

Outra forma de medir a exatidão de uma aproximação numérica é contar o **número de dígitos significativos corretos** em relação ao valor exato.

Definição 4 (Número de dígitos significativos corretos)**.** *A aproximação $\overline{x}$ de um número x tem s* ***dígitos significativos corretos*** *quando*[a]

$$\frac{|x-\overline{x}|}{|x|} < 5 \times 10^{-s}.$$

---

[a]Esta definição é apresentada em [**?**]. Não existe uma definição única na literatura para o conceito de dígitos significativos corretos, embora não precisamente equivalentes, elas transmitem o mesmo conceito. Uma maneira de interpretar essa regra é: calcula-se o erro relativo na forma normalizada e a partir da ordem do expoente temos o número de dígitos significativos corretos. Como queremos o expoente, podemos estimar s por

$$DIGSE(x,\bar{x}) = s \approx int \left| \log_{10} \frac{|x-\overline{x}|}{|x|} \right|.$$

Exemplo 23. Vejamos os seguintes casos:

a) A aproximação de $x = 0{,}333333$ por $\overline{x} = 0{,}333$ tem 3 dígitos significativos corretos, pois

$$\frac{|x - \overline{x}|}{|x|} = \frac{0{,}000333}{0{,}333333} \approx 0{,}000999 \leq 5 \times 10^{-\mathbf{3}}.$$

b) Considere as aproximações $\bar{x}_1 = 0{,}666$ e $\bar{x}_2 = 0{,}667$ de $x = 0{,}666888$. Os erros relativos são

$$\frac{|x - \bar{x}_1|}{|x|} = \frac{|0{,}666888 - 0{,}666|}{0{,}666888} \approx 0{,}00133... < 5 \times 10^{-3}.$$

$$\frac{|x - \bar{x}_2|}{|x|} = \frac{|0{,}666888 - 0{,}667|}{0{,}666888} \approx 0{,}000167... < 5 \times 10^{-4}.$$

Note que $\bar{x}_1$ possui 3 dígitos significativos corretos e $\bar{x}_2$ possui 4 dígitos significativos (o quarto dígito é o dígito 0 que não aparece a direita, i.e, $\bar{x}_2 = 0.6670$. Isto também leva a conclusão que x_2 aproxima melhor o valor de x do que x_1 pois está mais próximo de x.

c) $\overline{x} = 9{,}999$ aproxima $x = 10$ com 4 dígitos significativos corretos, pois

$$\frac{|x - \overline{x}|}{|x|} = \frac{|10 - 9{,}999|}{10} \approx 0{,}0000999... < 5 \times 10^{-4}.$$

d) Considere as aproximações $\overline{x}_1 = 1{,}49$ e $\overline{x}_2 = 1{,}5$ de $x = 1$. Da definição, temos que 1,49 aproxima 1 com um dígito significativo correto (verifique), equanto 1,5 tem zero dígito significativo correto, pois:

$$\frac{|1 - 1{,}5|}{|1|} = 5 \times 10^{-1} < 5 \times 10^{0}.$$

2.3.1 Erros de arredondamento

Os erros de arredondamento são aqueles gerados quando aproximamos um número real por um número com representação finita. Existem várias formas de arredondar

$$x = \pm d_0,d_1d_2\ldots d_{k-1}d_kd_{k+1}\ldots d_n \times 10^e$$

usando k dígitos significativos. As duas principais são as seguintes:

1. **Arredondamento por truncamento** (ou corte): aproximamos x por

   $$\bar{x} = \pm d_0,d_1d_2\ldots d_k \times 10^e$$

   simplesmente descartando os dígitos d_j com $j > k$.

2. **Arredondamento por proximidade**: se $d_{k+1} < 5$ aproximamos x por

   $$\bar{x} = \pm d_0,d_1d_2\ldots d_k \times 10^e$$

senão aproximamos x por[2]

$$\bar{x} = \pm(d_0,d_1d_2\ldots d_k + 10^{-k}) \times 10^e$$

Observação 6. Observe que o arredondamento pode mudar todos os dígitos e o expoente da representação em ponto flutuante de um número dado.

Exemplo 24. Represente os números $x_1 = 0{,}567$, $x_2 = 0{,}233$, $x_3 = -0{,}675$ e $x_4 = 0{,}314159265\ldots \times 10^1$ com dois dígitos significativos por truncamento e arredondamento.

Solução. Vejamos cada caso:

---

[2]Note que essas duas opções são equivalentes a somar 5 no dígito a direita do corte e depois arredondar por corte, ou seja, arredondar por corte

$$\pm(d_0,d_1d_2\ldots d_kd_{k+1} + 5 \times 10^{-(k+1)}) \times 10^e$$

- Por truncamento:

$$x_1 = 0{,}56, \quad x_2 = 0{,}23, \quad x_3 = -0{,}67 \quad \text{e} \quad x_4 = 3{,}1.$$

No `Scilab`, podemos obter a representação de $x_3 = -0{,}675$ fazendo (verifique):

```
-->format('e',8)
-->int(-0.675*1e2)/1e2
```

- Por arredondamento:

$$x_1 = 0{,}57; \quad x_2 = 0{,}23; \quad x_3 = -0{,}68 \quad \text{e} \quad x_4 = 3{,}1.$$

No `Scilab`, a representação de números por arredondamento é o padrão. Assim, para obtermos a representação desejada de $x_3 = 0{,}675$ fazemos: podemos obter a representação de $x_3 = -0{,}675$ fazemos (verifique):

```
-->format('e',8)
-->-0.675
```

$\Diamond$

Exemplo 25. O arredondamento de $0{,}9999 \times 10^{-1}$ com 3 dígitos significativos é $0{,}1 \times 10^{0}$.

Exercícios

E 2.3.1. Calcule os erros absoluto e relativo das aproximações $\bar{x}$ para x.

a) $x = \pi = 3{,}14159265358979\cdots$ e $\bar{x} = 3{,}141$

b) $x = 1{,}00001$ e $\bar{x} = 1$

c) $x = 100001$ e $\bar{x} = 100000$

E 2.3.2. Arredonde os seguintes números para cinco algarismos significativos corretos:

a) 1,7888544

b) 1788,8544

c) 0,0017888544

d) 0,004596632

e) $2{,}1754999 \times 10^{-10}$

f) $2{,}1754999 \times 10^{10}$

E 2.3.3. Verifique quantos são os dígitos significativos corretos em cada aproximação $\bar{x}$ para x.

a) $x = 2{,}5834$ e $\bar{x} = 2{,}6$

b) $x = 100$ e $\bar{x} = 99$

E 2.3.4. Represente os números 3276; 42,55 e 0,00003331 com três dígitos significativos por truncamento e arredondamento.

E 2.3.5. Resolva a equação $0{,}1x - 0{,}01 = 12$ usando arredondamento com três dígitos significativos em cada passo e compare com o resultado analítico

E 2.3.6. Calcule o erro relativo e absoluto envolvido nas seguintes aproximações e expresse as respostas com três algarismos significativos corretos.

a) $x = 3{,}1415926535898$ e $\tilde{x} = 3{,}141593$

b) $x = \frac{1}{7}$ e $\tilde{x} = 1{,}43 \times 10^{-1}$

2.4 Erros nas operações elementares

O erro presente relativo nas operações elementares de adição, subtração, multiplicação e divisão é da ordem do epsilon de máquina. Se estivermos usando uma máquina com 64 bits, temos que $\epsilon = 2^{-52} \approx 2{,}22E16$.

Este erro é bem pequeno! Assumindo que x e y são representados com todos dígitos corretos, temos aproximadamente 15 dígitos significativos corretos quando fizemos uma das operações $x+y$, $x-y$, $x \times y$ ou x/y.

Mesmo que fizéssemos, por exemplo, 1000 operações elementares em ponto flutuante sucessivas, teríamos no pior dos casos acumulado todos esses erros e perdido 3 casas decimais ($1000 \times 10^{-15} \approx 10^{-12}$).

Entretanto, quando subtraímos números muito próximos, os problemas aumentam.

2.5 Cancelamento catastrófico

Quando fazemos subtrações com números muito próximos entre si ocorre o cancelamento catastrófico, onde podemos perder vários dígitos de precisão em uma única subtração.

Exemplo 26. Efetue a operação

$$0{,}987624687925 - 0{,}987624 = 0{,}687925 \times 10^{-6}$$

usando arredondamento com seis dígitos significativos e observe a diferença se comparado com resultado sem arredondamento.

Solução. Os números arredondados com seis dígitos para a mantissa resultam na seguinte diferença

$$0{,}987625 - 0{,}987624 = 0{,}100000 \times 10^{-5}$$

Observe que os erros relativos entre os números exatos e aproximados no lado esquerdo são bem pequenos,

$$\frac{|0{,}987624687925 - 0{,}987625|}{|0{,}987624687925|} = 0{,}00003159$$

e

$$\frac{|0{,}987624 - 0{,}987624|}{|0{,}987624|} = 0\%,$$

enquanto no lado direito o erro relativo é enorme:

$$\frac{|0{,}100000 \times 10^{-5} - 0{,}687925 \times 10^{-6}|}{0{,}687925 \times 10^{-6}} = 45{,}36\%.$$

$\Diamond$

Exemplo 27. Considere o problema de encontrar as raízes da equação de segundo grau

$$x^2 + 300x - 0{,}014 = 0,$$

usando seis dígitos significativos.
Aplicando a fórmula de Bhaskara com $a = 0{,}100000 \times 10^1$, $b =$

$0{,}300000 \times 10^3$ e $c = 0{,}140000 \times 10^{-1}$, temos o discriminante:

$$\begin{aligned}
\Delta &= b^2 - 4 \cdot a \cdot c \\
&= 0{,}300000 \times 10^3 \times 0{,}300000 \times 10^3 \\
&+ 0{,}400000 \times 10^1 \times 0{,}100000 \times 10^1 \times 0{,}140000 \times 10^{-1} \\
&= 0{,}900000 \times 10^5 + 0{,}560000 \times 10^{-1} \\
&= 0{,}900001 \times 10^5
\end{aligned}$$

e as raízes:

$$\begin{aligned}
x_1, x_2 &= \frac{-0{,}300000 \times 10^3 \pm \sqrt{\Delta}}{0{,}200000 \times 10^1} \\
&= \frac{-0{,}300000 \times 10^3 \pm \sqrt{0{,}900001 \times 10^5}}{0{,}200000 \times 10^1} \\
&= \frac{-0{,}300000 \times 10^3 \pm 0{,}300000 \times 10^3}{0{,}200000 \times 10^1}
\end{aligned}$$

Então, as duas raízes são:

$$\tilde{x}_1 = \frac{-0{,}300000 \times 10^3 - 0{,}300000 \times 10^3}{0{,}200000 \times 10^1}$$
$$= -\frac{0{,}600000 \times 10^3}{0{,}200000 \times 10^1} = -0{,}300000 \times 10^3$$

e

$$\tilde{x}_2 = \frac{-0{,}300000 \times 10^3 + 0{,}300000 \times 10^3}{0{,}200000 \times 10^1} = 0{,}000000 \times 10^0$$

Agora, os valores das raízes com seis dígitos significativos deveriam ser

$$x_1 = -0{,}300000 \times 10^3 \quad \text{e} \quad x_2 = 0{,}466667 \times 10^{-4}.$$

Observe que um raiz saiu com seis dígitos significativos corretos, mas a outra não possui nenhum dígito significativo correto.

Observação 7. No exemplo anterior b^2 é muito maior que $4ac$, ou seja, $b \approx \sqrt{b^2 - 4ac}$, logo a diferença

$$-b + \sqrt{b^2 - 4ac}$$

estará próxima de zero. Uma maneira padrão de evitar o cancelamento catastrófico é usar procedimentos analíticos para eliminar essa diferença. Abaixo veremos alguns exemplos.

Exemplo 28. Para eliminar o cancelamento catastrófico do exemplo anterior, usamos a seguinte expansão em série de Taylor em torno da origem

$$\sqrt{1 - x} = 1 - \frac{1}{2}x + O(x^2).$$

Substituindo na fórmula de Bhaskara, temos:

$$\begin{aligned} x &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\ &= \frac{-b \pm b\sqrt{1 - \frac{4ac}{b^2}}}{2a} \\ &\approx \frac{-b \pm b\left(1 - \frac{4ac}{2b^2}\right)}{2a} \end{aligned}$$

Observe que $\frac{4ac}{b^2}$ é um número pequeno e por isso a expansão faz sentido. Voltamos no exemplo anterior e calculamos as duas raízes

com o nova expressão

$$\begin{aligned}
\tilde{x}_1 &= \frac{-b - b + \frac{4ac}{2b}}{2a} = -\frac{b}{a} + \frac{c}{b} \\
&= -\frac{0{,}300000 \times 10^3}{0{,}100000 \times 10^1} - \frac{0{,}140000 \times 10^{-1}}{0{,}300000 \times 10^3} \\
&= -0{,}300000 \times 10^3 - 0{,}466667 \times 10^{-4} \\
&= -0{,}300000 \times 10^3
\end{aligned}$$

$$\begin{aligned}
\tilde{x}_2 &= \frac{-b + b - \frac{4ac}{2b}}{2a} \\
&= -\frac{4ac}{4ab} \\
&= -\frac{c}{b} = -\frac{-0{,}140000 \times 10^{-1}}{0{,}300000 \times 10^3} = 0{,}466667 \times 10^{-4}
\end{aligned}$$

Observe que o efeito catastrófico foi eliminado.

2.6 Condicionamento de um problema

Geralmente podemos pensar um problema como um mapeamento f onde a partir de valores de entrada x devemos encontrar a saída, a solução y, ou seja, $f : x \to y$, ou simplesmente

$$y = f(x) \tag{2.1}$$

Entretanto, a entrada do problema x normalmente terá erros (por exemplo, erros na coleta dos dados ou erros na representação dos dados devido a arredondamentos). Assim, ao invés de usar x estamos usando x^* para resolver o problema e encontrar a solução y^*, ou seja, estamos resolvendo

$$y^* = f(x^*) \tag{2.2}$$

Estamos interessados em saber se os erros cometidos na entrada $\Delta x = x - x^*$ influenciaram na saída do problema $\Delta y = y - y^*$.

No caso mais simples, temos que $x \in \mathbb{R}$ e $y \in \mathbb{R}$. Assumindo que f seja diferenciável, a partir da série de Taylor

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x \tag{2.3}$$

obtemos (subtraindo $f(x)$ dos dois lados)

$$\Delta y = f(x + \Delta x) - f(x) \approx f'(x)\Delta x \tag{2.4}$$

Para relacionarmos os erros relativos, dividimos o lado esquerdo por y, o lado direito por $f(x) = y$ e obtemos

$$\frac{\Delta y}{y} \approx \frac{f'(x)}{f(x)} \frac{x\Delta x}{x} \tag{2.5}$$

sugerindo a definição de número de condicionamento de um problema.

Definição 5. *Seja f uma função diferenciável. O* ***número de condicionamento*** *de um problema é definido como*

$$\kappa_f(x) := \left| \frac{xf'(x)}{f(x)} \right| \tag{2.6}$$

e fornece uma estimativa de quanto os erros relativos na entrada $\left|\frac{\Delta x}{x}\right|$ serão amplificados na saída $\left|\frac{\Delta y}{y}\right|$.

De modo geral, quando f depende de várias variáveis, podemos obter

$$\delta_f = |f(x_1,x_2,...,x_n) - f(\bar{x}_1, \bar{x}_2,...,\bar{x}_n)| \approx \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(x_1, x_2,...,x_n) \right| \delta_{x_i}$$

Uma matriz de números de condicionamento também poderia ser obtida como em [**?**].

Exemplo 29. Considere o problema de calcular $\sqrt{x}$ em $x = 2$. Se usarmos $x^* = 1{,}999$, quanto será o erro relativo na saída? O erro relativo na entrada é

$$\left|\frac{\Delta x}{x}\right| = \left|\frac{2 - 1{,}999}{2}\right| = 0{,}0005 \tag{2.7}$$

O número de condicionamento do problema calcular a raiz é

$$\kappa_f(x) := \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\frac{1}{2\sqrt{x}}}{\sqrt{x}}\right| = \frac{1}{2} \tag{2.8}$$

Ou seja, os erros na entrada serão diminuídos pela metade. De fato, usando $y = \sqrt{2} = 1{,}4142136...$ e $y^* = \sqrt{1{,}999} = 1{,}41386...$, obtemos

$$\frac{\Delta y}{y} = \frac{\sqrt{2} - \sqrt{1{,}999}}{\sqrt{2}} \approx 0{,}000250031... \tag{2.9}$$

Exemplo 30. Considere a função $f(x) = \frac{10}{1-x^2}$ e $x^* = 0{,}9995$ com um erro absoluto na entrada de 0,0001.

Calculando $y^* = f(x^*)$ temos

$$y^* = \frac{10}{1-(0{,}9995)^2} \approx 10002{,}500625157739705173 \tag{2.10}$$

Mas qual é a estimativa de erro nessa resposta? Quantos dígitos significativos temos nessa resposta?
Sabendo que $f'(x) = -10/(1-x^2)^2$, o número de condicionamento é

$$\kappa_f(x) := \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{2x^2}{1-x^2}\right| \tag{2.11}$$

o que nos fornece para $x^* = 0{,}9995$,

$$\kappa_f(0{,}9995) \approx 1998{,}5 \tag{2.12}$$

Como o erro relativo na entrada é

$$\left|\frac{\Delta x}{x}\right| = \left|\frac{0{,}0001}{0{,}9995}\right| \approx 0{,}00010005... \tag{2.13}$$

temos que o erro na saída será aproximadamente

$$\left|\frac{\Delta y}{y}\right| \approx \kappa_f(x) \left|\frac{\Delta x}{x}\right| \approx 1998{,}5 \times 0{,}00010005... \approx 0{,}1999 \tag{2.14}$$

ou seja um erro relativo de aproximadamente 19,99%.
Note que se usarmos $x_1 = 0{,}9994$ e $x_2 = 0{,}9996$ (ambos no intervalo do erro absoluto da entrada) encontramos

$$y_1^* \approx 8335{,}83 \tag{2.15}$$

$$y_2^* \approx 12520{,}50 \tag{2.16}$$

confirmando a estimativa de 19,99%.

Exemplo 31. Seja $f(x) = x\exp(x)$. Calcule o erro absoluto em se calcular $f(x)$ sabendo que $x = 2 \pm 0{,}05$.

Solução. Temos que $x \approx 2$ com erro absoluto de $\delta_x = 0{,}05$. Neste caso, calculamos δ_f, i.e. o erro absoluto em se calcular $f(x)$, por:

$$\delta_f = |f'(x)|\delta_x.$$

Como $f'(x) = (1+x)e^x$, temos:

$$\begin{aligned}\delta_f &= |(1+x)e^x| \cdot \delta_x \\ &= |3e^2| \cdot 0{,}05 = 1{,}1084.\end{aligned}$$

Portanto, o erro absoluto em se calcular $f(x)$ quando $x = 2 \pm 0{,}05$ é de 1,084. $\Diamond$

Exemplo 32. Calcule o erro relativo ao medir $f(x,y) = \frac{x^2+1}{x^2}e^{2y}$ sabendo que $x \approx 3$ é conhecido com 10% de erro e $y \approx 2$ é conhecido com 3% de erro.

Solução. Calculamos as derivadas parciais de f:

$$\frac{\partial f}{\partial x} = \frac{2x^3 - (2x^3 + 2x)}{x^4}e^{2y} = -\frac{2e^{2y}}{x^3}$$

e

$$\frac{\partial f}{\partial y} = 2\frac{x^2+1}{x^2}e^{2y}$$

Calculamos o erro absoluto em termos do erro relativo:

$$\frac{\delta_x}{|x|} = 0{,}1 \Rightarrow \delta_x = 3 \cdot 0{,}1 = 0{,}3$$

$$\frac{\delta_y}{|y|} = 0{,}03 \Rightarrow \delta_y = 2 \cdot 0{,}03 = 0{,}06$$

Aplicando a expressão para estimar o erro em f temos

$$\begin{aligned}\delta_f &= \left|\frac{\partial f}{\partial x}\right| \delta_x + \left|\frac{\partial f}{\partial y}\right| \delta_y \\ &= \frac{2e^4}{27} \cdot 0{,}3 + 2\frac{9+1}{9} e^4 \cdot 0{,}06 = 8{,}493045557\end{aligned}$$

Portanto, o erro relativo ao calcular f é estimado por

$$\frac{\delta f}{|f|} = \frac{8{,}493045557}{\frac{9+1}{9} e^4} = 14\%$$

$\Diamond$

Exemplo 33. No exemplo anterior, reduza o erro relativo em x pela metade e calcule o erro relativo em f. Depois, repita o processo reduzindo o erro relativo em y pela metade.

Solução. Na primeira situação temos $x = 3$ com erro relativo de 5% e $\delta_x = 0{,}05 \cdot 3 = 0{,}15$. Calculamos $\delta_f = 7{,}886399450$ e o erro relativo em f de 13%. Na segunda situação, temos $y = 2$ com erro de 1,5% e $\delta_y = 2 \cdot 0{,}015 = 0{,}03$. Calculamos $\delta_f = 4{,}853168892$ e o erro relativo em f de 8%. Observe que mesma o erro relativo em x sendo maior, o erro em y é mais significante na função. $\Diamond$

Exemplo 34. Considere um triângulo retângulo onde a hipotenusa e um dos catetos são conhecidos a menos de um erro: hipotenusa $a = 3 \pm 0{,}01$ metros e cateto $b = 2 \pm 0{,}01$ metros. Calcule o erro absoluto ao calcular a área dessa triângulo.

Solução. Primeiro vamos encontrar a expressão para a área em função da hipotenusa a e um cateto b. A tamanho de segundo cateto c é dado pelo teorema de Pitágoras, $a^2 = b^2 + c^2$, ou seja,

$c = \sqrt{a^2 - b^2}$. Portanto a área é

$$A = \frac{bc}{2} = \frac{b\sqrt{a^2 - b^2}}{2}.$$

Agora calculamos as derivadas

$$\frac{\partial A}{\partial a} = \frac{ab}{2\sqrt{a^2 - b^2}},$$

$$\frac{\partial A}{\partial b} = \frac{\sqrt{a^2 - b^2}}{2} - \frac{b^2}{2\sqrt{a^2 - b^2}},$$

e substituindo na estimativa para o erro δ_A em termos de $\delta_a = 0{,}01$ e $\delta_b = 0{,}01$:

$$\begin{aligned} \delta_A &\approx \left|\frac{\partial A}{\partial a}\right| \delta_a + \left|\frac{\partial A}{\partial b}\right| \delta_b \\ &\approx \frac{3\sqrt{5}}{5} \cdot 0{,}01 + \frac{\sqrt{5}}{10} \cdot 0{,}01 = 0{,}01565247584 \end{aligned}$$

Em termos do erro relativo temos erro na hipotenusa de $\frac{0{,}01}{3} \approx 0{,}333\%$, erro no cateto de $\frac{0{,}01}{2} = 0{,}5\%$ e erro na área de

$$\frac{0{,}01565247584}{\frac{2\sqrt{3^2-2^2}}{2}} = 0{,}7\%$$

$\Diamond$

Exercícios

E 2.6.1. Considere que a variável $x \approx 2$ é conhecida com um erro relativo de 1% e a variável $y \approx 10$ com um erro relativo de 10%. Calcule o erro relativo associado a z quando:

$$z = \frac{y^4}{1+y^4} e^x.$$

Suponha que você precise conhecer o valor de z com um erro de 0,5%. Como engenheiro, você propõe uma melhoria na medição da variável x ou y? Explique.

E 2.6.2. A corrente I em ampères e a tensão V em volts em uma lâmpada se relacionam conforme a seguinte expressão:

$$I = \left(\frac{V}{V_0}\right)^{\alpha},$$

onde α é um número entre 0 e 1 e V_0 é tensão nominal em volts. Sabendo que $V_0 = 220 \pm 3\%$ e $\alpha = -,8 \pm 4\%$, calcule a corrente e o

erro relativo associado quando a tensão vale $220 \pm 1\%$.
Obs:. Este problema pode ser resolvido de duas formas distintas: usando a expressão aproximada para a propagação de erro e inspecionando os valores máximos e mínimos que a expressão pode assumir. Pratique os dois métodos.

E 2.6.3. A corrente I em ampères e a tensão V em volts em uma lâmpada se relacionam conforme a seguinte expressão:

$$I = \left(\frac{V}{V_0}\right)^{\alpha}$$

Onde α é um número entre 0 e 1 e V_0 é a tensão nominal em volts. Sabendo que $V_0 = 220 \pm 3\%$ e $\alpha = 0{,}8 \pm 4\%$ Calcule a corrente e o erro relativo associado quando a tensão vale $220 \pm 1\%$. **Dica:** lembre que $x^{\alpha} = e^{\alpha \ln(x)}$

2.7 Mais exemplos

Exemplo 35. Considere o seguinte processo iterativo:

$$\begin{cases} x_0 = \frac{1}{3} \\ x_{n+1} = 4x_n - 1, \qquad n \in \mathbb{N} \end{cases}.$$

Observe que $x_0 = \frac{1}{3}$, $x_1 = 4 \cdot \frac{1}{3} - 1 = \frac{1}{3}$, $x_2 = \frac{1}{3}$, ou seja, temos uma sequência constante igual a $\frac{1}{3}$. No entanto, ao calcularmos no computador, usando o sistema de numeração 'double', a sequencia obtida não é constante e, de fato, diverge. Faça o teste no `Scilab`, colocando:

```
-->x = 1/3
```

e itere algumas vezes a linha de comando:

```
-->x = 4*x-1
```

Para compreender o que acontece, devemos levar em consideração que o número $\frac{1}{3} = 0{,}\overline{3}$ possui um representação infinita tanto na base decimal quanto na base binária. Logo, sua representação de máquina incliu um erro de arredondamento. Seja ϵ a diferença entre o valor exato de $\frac{1}{3}$ e sua representação de máquina, isto é, $\tilde{x}_0 = \frac{1}{3} + \epsilon$. A sequência efetivamente calculada no computador é:

$$\begin{aligned}
\tilde{x}_0 &= \frac{1}{3} + \epsilon \\
\tilde{x}_1 &= 4x_0 - 1 = 4\left(\frac{1}{3} + \epsilon\right) - 1 = \frac{1}{3} + 4\epsilon \\
\tilde{x}_2 &= 4x_1 - 1 = 4\left(\frac{1}{3} + 4\epsilon\right) - 1 = \frac{1}{3} + 4^2\epsilon \\
&\vdots \\
\tilde{x}_n &= \frac{1}{3} + 4^n\epsilon
\end{aligned}$$

Portanto o limite da sequência diverge,

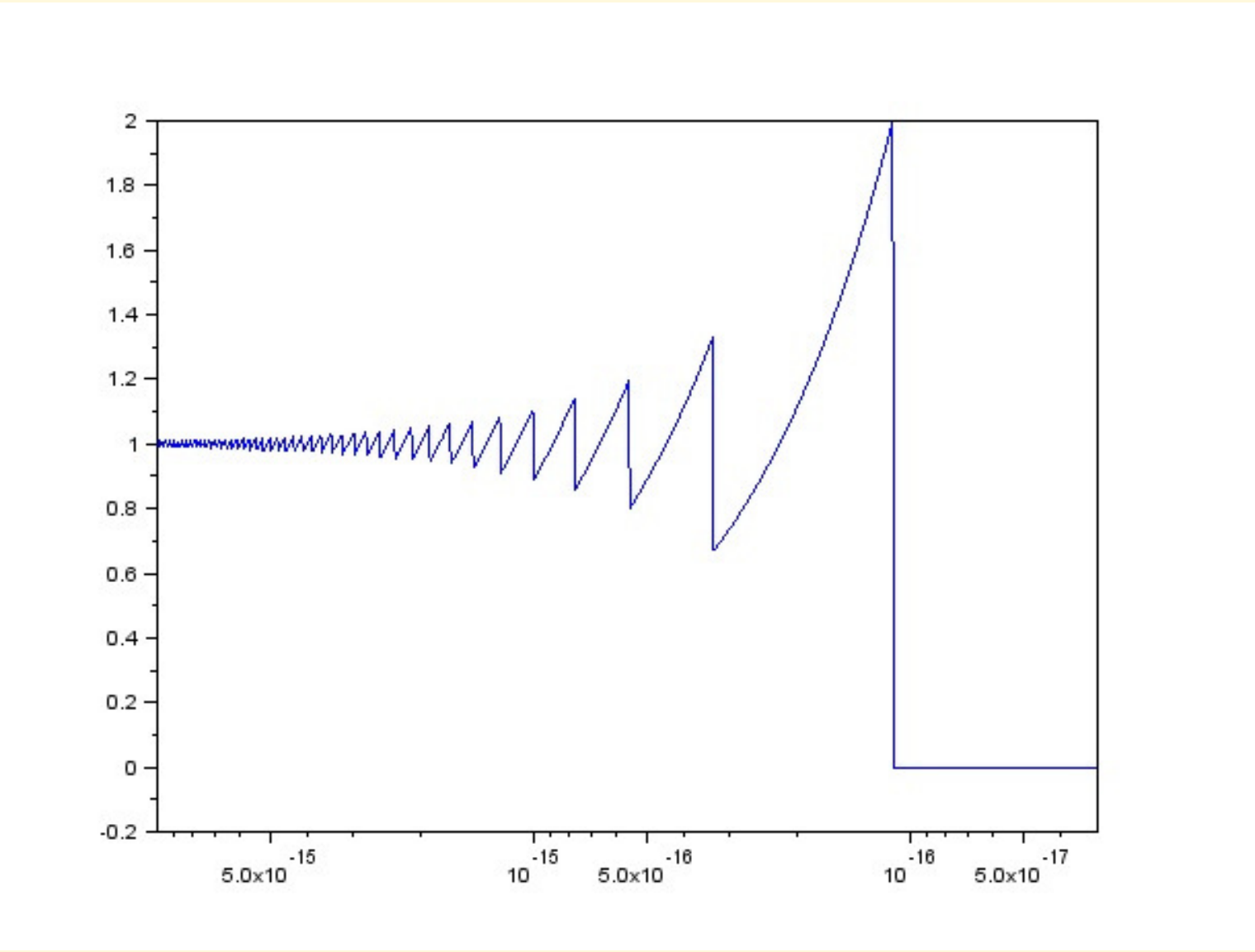
$$\lim_{x\to\infty} |\tilde{x}_n| = \infty$$

Qual o número de condicionamento desse problema?

Exemplo 36. Observe a seguinte identidade

$$f(x) = \frac{(1+x)-1}{x} = 1$$

Calcule o valor da expressão à esquerda para $x = 10^{-12}$, $x = 10^{-13}$, $x = 10^{-14}$, $x = 10^{-15}$, $x = 10^{-16}$ e $x = 10^{-17}$. Observe que quando x se aproxima do ϵ de máquina a expressão perde o significado. Veja abaixo o gráfico de $f(x)$ em escala logarítmica.



Exemplo 37. Neste exemplo, estamos interessados em compreen-

der mais detalhadamente o comportamento da expressão

$$\left(1+\frac{1}{n}\right)^n \tag{2.17}$$

quando n é um número grande ao computá-la em sistemas de numeral de ponto flutuante com acurácia finita. Um resultado bem conhecido do cálculo nos diz que o limite de (2.17) quando n tende a infinito é o número de Euler:

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e = 2{,}718281828459... \tag{2.18}$$

Sabemos também que a sequência produzida por (2.17) é crescente, isto é:

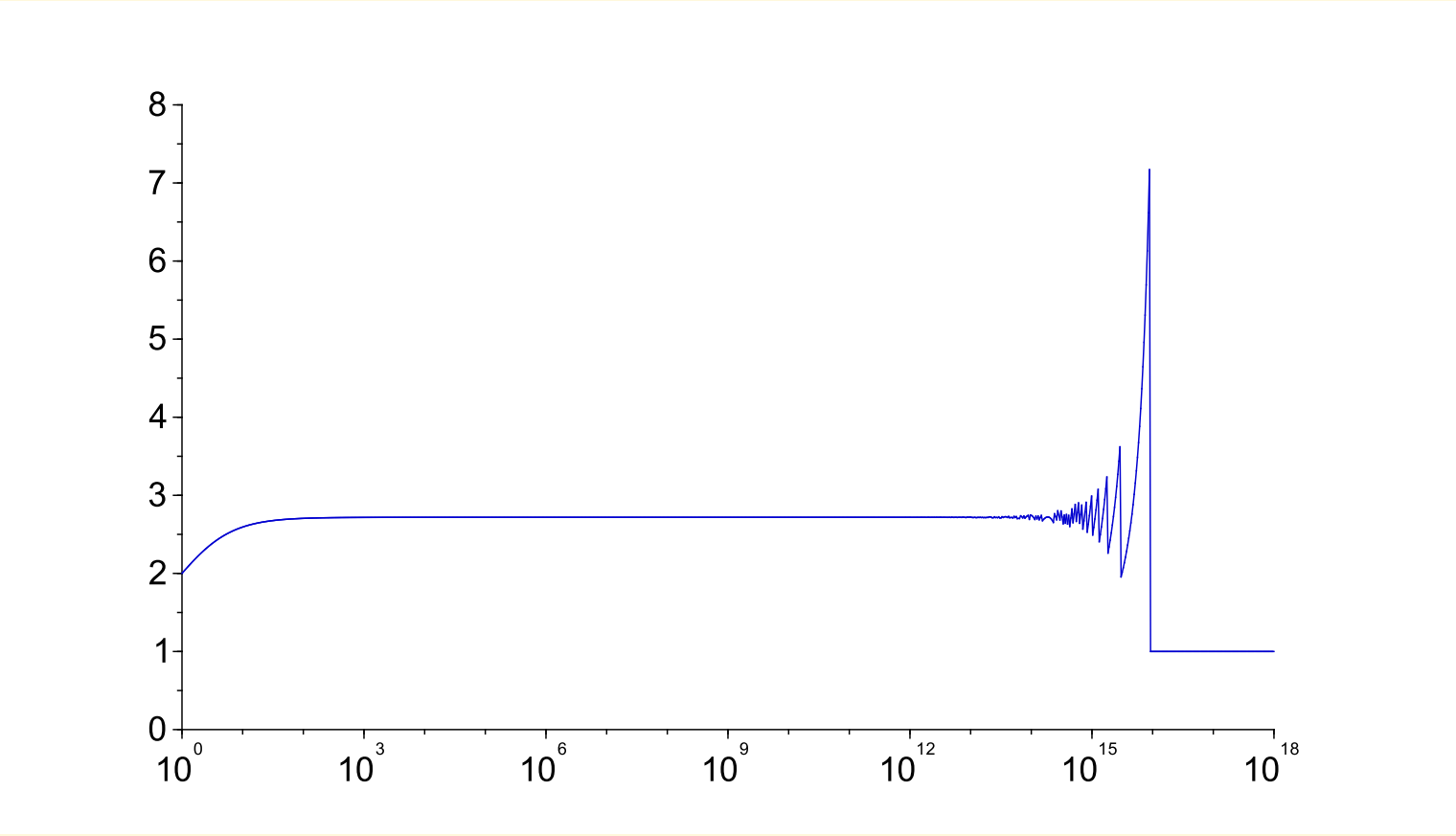
$$\left(1+\frac{1}{1}\right)^1 < \left(1+\frac{1}{2}\right)^2 < \left(1+\frac{1}{3}\right)^3 < \cdots$$

No entanto, quando calculamos essa expressão no `Scilab`, nos de-

frontamos com o seguinte resultado:

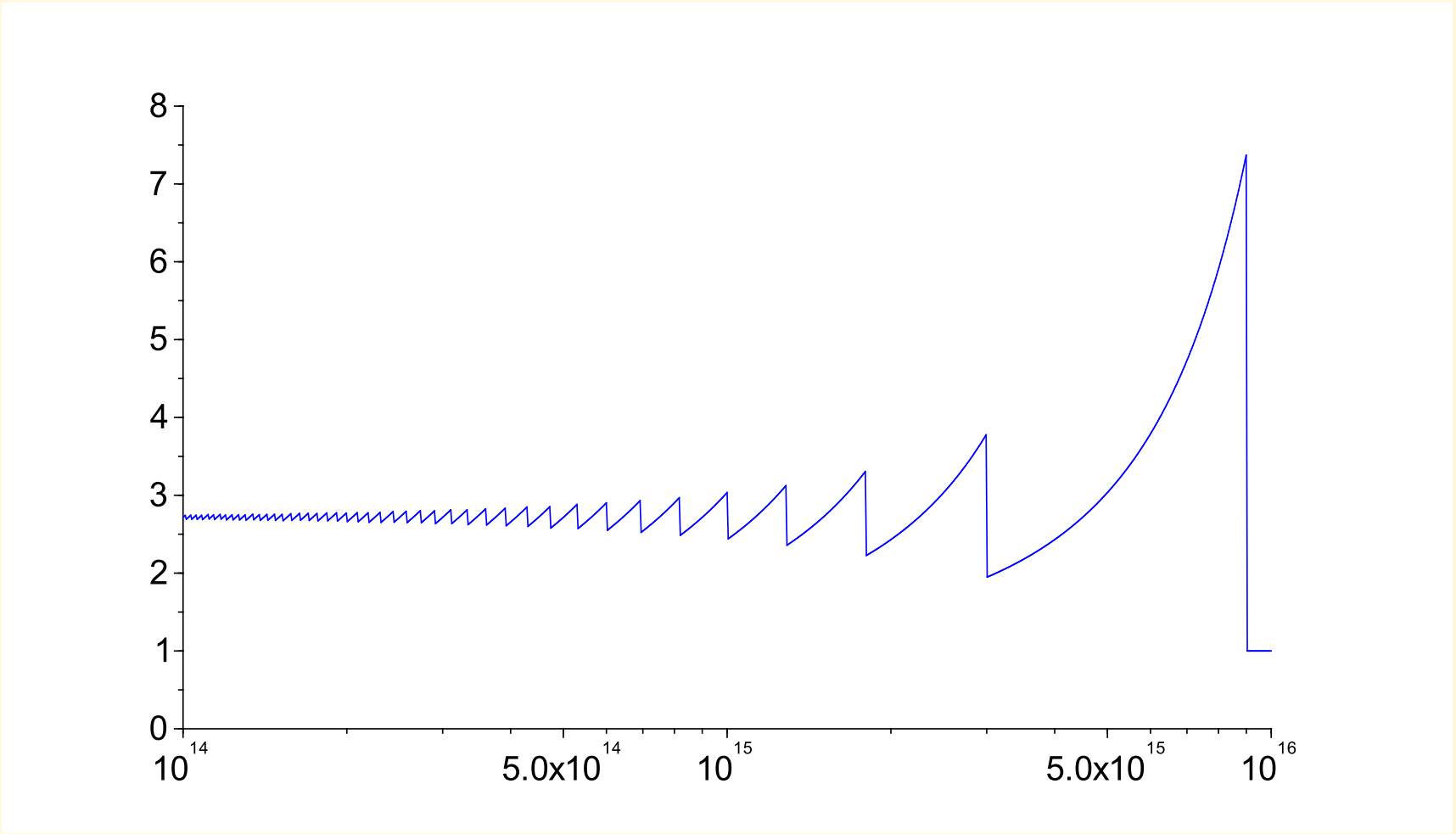
n	$\left(1+\frac{1}{n}\right)^n$		n	$\left(1+\frac{1}{n}\right)^n$
1	2,0000000000000		10^2	2,7048138294215
2	2,2500000000000		10^4	2,7181459268249
3	2,3703703703704		10^6	2,7182804690957
4	2,4414062500000		10^8	2,7182817983391
5	2,4883200000000		10^{10}	2,7182820532348
6	2,5216263717421		10^{12}	2,7185234960372
7	2,5464996970407		10^{14}	2,7161100340870
8	2,5657845139503		10^{16}	1,0000000000000
9	2,5811747917132		10^{18}	1,0000000000000

Podemos resumir esses dados no seguinte gráfico de $\left(1+\frac{1}{n}\right)^n$ em função de n:



Observe que quando x se torna grande, da ordem de 10^{15}, o gráfico da função deixa de se crescente e apresenta oscilações. Observe também que a expressão se torna identicamente igual a 1 depois de um certo limiar. Tais fenômenos não são intrínsecos da função $f(x) = (1+1/x)^x$, mas **oriundas de erros de arredondamento**,

isto é, são resultados numéricos espúrios. A fim de pôr o comportamento numérico de tal expressão, apresentamos abaixo o gráfico da mesma função, porém restrito à região entre 10^{14} e 10^{16}.



..

Para compreendermos melhor por que existe um limiar N que, quando atingido torna a expressão do exemplo acima identicamente

igual a 1, observamos a sequência de operações realizadas pelo computador:

$$x \;\rightarrow\; 1/x \;\rightarrow\; 1+1/x \;\rightarrow\; (1+1/x)^x \tag{2.19}$$

Devido ao limite de precisão da representação de números em ponto flutuante, existe um menor número representável que é maior do que 1. Este número é 1+`eps`, onde `eps` é chamado de **épsilon de máquina** e é o menor número que somado a 1 produz um resultado superior a 1 no sistema de numeração usado. O épsilon de máquina no sistema de numeração **double** vale aproximadamente $2{,}22 \times 10^{-16}$. No `Scilab`, o epsilon de máquina é a constante `eps`. Observe que:

```
-->1+%eps
 ans  =
    1.0000000000000002220446
```

Quando somamos a 1 um número positivo inferior ao épsilon de máquina, obtemos o número 1. Dessa forma, o resultado obtido

pela operação de ponto flutuante $1 + x$ para $0 < x < 2{,}22 \times 10^{-16}$ é 1.

Portanto, quando realizamos a sequência de operações dada em (2.19), toda informação contida no número x é perdida na soma com 1 quando $1/x$ é menor que o épsilon de máquina, o que ocorre quando $x > 5 \times 10^{15}$. Assim $(1 + 1/x)$ é aproximado para 1 e a última operação se resume a 1^x, o que é igual a 1 mesmo quando x é grande.

Um erro comum é acreditar que o perda de significância se deve ao fato de $1/x$ ser muito pequeno para ser representado e é aproximando para 0. Isto é falso, o sistema de ponto de flutuante permite representar números de magnitude muito inferior ao épsilon de máquina. O problema surge da limitação no tamanho da mantissa. Observe como a seguinte sequência de operações não perde significância para números positivos x muito menores que o épsilon de máquina:

$$x \;\rightarrow\; 1/x \;\rightarrow\; 1/(1/x) \tag{2.20}$$

compare o desempenho numérico desta sequência de operações para

valores pequenos de x com o da seguinte sequência:

$$x \ \rightarrow \ 1+x \ \rightarrow \ (1+x)-1. \tag{2.21}$$

Finalmente, notamos que quando tentamos calcular $\left(1+\frac{1}{n}\right)^n$ para n grande, existe perda de significância no cálculo de $1+1/n$. Para entendermos isso melhor, vejamos o que acontece no `Scilab` quando $n = 7 \times 10^{13}$:

```
-->n=7e13
 n  =
     7.000000000000000D+13

-->1/n
 ans  =
     1.428571428571428435D-14

-->y=1+1/n
 y  =
```

```
    1.00000000000014211D+00
```

Observe a perda de informação ao deslocar a mantissa de $1/n$. Para evidenciar o fenômenos, observamos o que acontece quando tentamos recalcular n subtraindo 1 de $1 + 1/n$ e invertendo o resultado:

```
-->y-1
 ans  =
     1.42108547152020372D-14

-->1/(y-1)
 ans  =
     7.03687441776640000D+13
```

Exemplo 38 (Analogia da balança)**.** Observe a seguinte comparação interessante que pode ser feita para ilustrar os sistemas de numeração com ponto fixo e flutuante: o sistema de ponto fixo é como uma balança cujas marcas estão igualmente espaçadas; o sistema de ponto flutuante é como uma balança cuja distância entre

as marcas é proporcional à massa medida. Assim, podemos ter uma balança de ponto fixo cujas marcas estão sempre distanciadas de 100g (100g, 200g, 300g, ..., 1Kg, 1,1Kg,...) e outra balança de ponto flutuante cujas marcas estão distanciadas sempre de aproximadamente um décimo do valor lido (100g, 110g, 121g, 133g, ..., 1Kg, 1,1Kg, 1,21Kg, ...) A balança de ponto fixo apresenta uma resolução baixa para pequenas medidas, porém uma resolução alta para grandes medidas. A balança de ponto flutuante distribui a resolução de forma proporcional ao longo da escala.

Seguindo nesta analogia, o fenômeno de perda de significância pode ser interpretado como a seguir: imagine que você deseje obter o peso de um gato (aproximadamente 4Kg). Dois processos estão disponíveis: colocar o gato diretamente na balança ou medir seu peso com o gato e, depois, sem o gato. Na balança de ponto flutuante, a incerteza associada na medida do peso do gato (sozinho) é aproximadamente 10% de 4Kg, isto é, 400g. Já a incerteza associada à medida da uma pessoa (aproximadamente 70Kg) com o gato é de 10% do peso total, isto é, aproximadamente 7Kg. Esta incerteza é

da mesma ordem de grandeza da medida a ser realizada, tornado o processo impossível de ser realizado, já que teríamos uma incerteza da ordem de 14Kg (devido à dupla medição) sobre uma grandeza de 4Kg.

Exercícios

E 2.7.1. Considere as expressões:

$$\frac{\exp(1/\mu)}{1+\exp(1/\mu)}$$

e

$$\frac{1}{\exp(-1/\mu)+1}$$

com $\mu > 0$. Verifique que elas são idênticas como funções reais. Teste no computador cada uma delas para $\mu = 0{,}1$, $\mu = 0{,}01$ e $\mu = 0{,}001$. Qual dessas expressões é mais adequada quando μ é um número pequeno? Por quê?

E 2.7.2. Encontre expressões alternativas para calcular o valor das seguintes funções quando x é próximo de zero.

a) $f(x) = \frac{1-\cos(x)}{x^2}$

b) $g(x) = \sqrt{1+x} - 1$

c) $h(x) = \sqrt{x+10^6} - 10^3$

d) $i(x) = \sqrt{1+e^x} - \sqrt{2}$ Dica: Faça $y = e^x - 1$

E 2.7.3. Use uma identidade trigonométrica adequada para mostrar que:

$$\frac{1-\cos(x)}{x^2} = \frac{1}{2}\left(\frac{\sin(x/2)}{x/2}\right)^2.$$

Analise o desempenho destas duas expressões no computador quando x vale 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}, 10^{-200} e 0. Discuta o resultado. **Dica:** Para $|x| < 10^{-5}$, $f(x)$ pode ser aproximada por $1/2 - x^2/24$ com erro de truncamento inferior a 10^{-22}.

E 2.7.4. Reescreva as expressões:

$$\sqrt{e^{2x}+1} - e^x \qquad \text{e} \qquad \sqrt{e^{2x}+x^2} - e^x$$

de modo que seja possível calcular seus valores para $x = 100$ utilizando a aritmética de ponto flutuante ("Double") no computador.

E 2.7.5. Na teoria da relatividade restrita, a energia cinética de uma partícula e sua velocidade se relacionam pela seguinte fórmula:

$$E = mc^2 \left(\frac{1}{\sqrt{1 - (v/c)^2}} - 1 \right),$$

onde E é a energia cinética da partícula, m é a massa de repouso, v o módulo da velocidade e c a velocidade da luz no vácuo dada por $c = 299792458m/s$. Considere que a massa de repouso $m = 9{,}10938291 \times 10^{-31}Kg$ do elétron seja conhecida com erro relativo de 10^{-9}. Qual é o valor da energia e o erro relativo associado a essa grandeza quando $v = 0{,}1c$, $v = 0{,}5c$, $v = 0{,}99c$ e $v = 0{,}999c$ sendo que a incerteza relativa na medida da velocidade é 10^{-5}?

E 2.7.6. Deseja-se medir a concentração de dois diferentes oxidantes no ar. Três sensores eletroquímicos estão disponíveis para a

medida e apresentam a seguintes respostas:

$$v_1 = 270[A] + 30[B], \quad v_2 = 140[A] + 20[B] \ \text{ e } \ v_3 = 15[A] + 200[B]$$

as tensões v_1, v_2 e v_3 são dadas em mV e as concentrações em $milimol/l$.

a) Encontre uma expressão para os valores de $[A]$ e $[B]$ em termos de v_1 e v_2 e, depois, em termos de v_1 e v_3. Dica: Se $ad \neq bc$, então a matriz A dada por

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

é inversível e sua inversa é dada por

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

b) Sabendo que incerteza relativa associada às sensibilidades dos sensores 1 e 2 é de 2% e que a incerteza relativa associada às sensibilidades do sensor 3 é 10%, verifique a incerteza associada à medida feita com o par $1-2$ e o par $1-3$. Use $[A] = [B] = 10 milimol/l$. Dica: Você deve diferenciar as grandezas $[A]$ e $[B]$ em relação aos valores das tensões.

]

Capítulo 3

Solução de equações de uma variável

Neste capítulo buscaremos aproximações numéricas para a solução de **equações de uma variável real**. Observamos que obter uma solução para uma tal dada equação é equivalente a encontrar um

zero de uma função apropriada. Com isso, iniciamos este capítulo discutindo sobre condições de existência e unicidade de raízes de funções de uma variável real. Então, apresentamos o **método da bisseção** como uma primeira abordagem numérica para a solução de tais equações.
Em seguida, exploramos uma outra abordagem via **iteração do ponto fixo**. Desta, obtemos o **método de Newton**, para o qual discutimos sua aplicação e convergência. Por fim, apresentamos o **método das secantes** como uma das possíveis variações do método de Newton.

3.1 Existência e unicidade

O **teorema de Bolzano**[1] nos fornece condições suficientes para a existência do zero de uma função. Este é uma aplicação direta do **teorema do valor intermediário**.

Teorema 1 (Teorema de Bolzano)**.** *Se* $f : [a, b] \to \mathbb{R}$, $y = f(x)$, *é uma função contínua tal que* $f(a) \cdot f(b) < 0$, *então existe* $x^* \in (a, b)$ *tal que* $f(x^*) = 0$.

Demonstração. O resultado é uma consequência imediata do teorema do valor intermediário que estabelece que dada uma função contínua $f : [a, b] \to \mathbb{R}$, $y = f(x)$, tal que $f(a) < f(b)$ (ou $f(b) < f(a)$), então para qualquer $d \in (f(a), f(b))$ (ou $k \in (f(b), f(a))$) existe $x^* \in (a, b)$ tal que $f(x^*) = k$. Ou seja, nestas notações, se $f(a) \cdot f(b) < 0$, então $f(a) < 0 < f(b)$ (ou $f(b) < 0 < f(a)$). Logo,

[1] Bernhard Placidus Johann Gonzal Nepomuk Bolzano, 1781 - 1848, matemático do Reino da Boêmia.

tomando $k = 0$, temos que existe $x^* \in (a, b)$ tal que $f(x^*) = k = 0$. □

Em outras palavras, se $f(x)$ é uma função contínua em um dado intervalo no qual ela troca de sinal, então ela têm pelo menos um zero neste intervalo (veja a Figura 3.1).

Exemplo 39. Mostre que existe pelo menos uma solução da equação $e^x = x + 2$ no intervalo $(-2,0)$.

Solução. Primeiramente, observamos que resolver a equação $e^x = x + 2$ é equivalente a resolver $f(x) = 0$ com $f(x) = e^x - x - 2$. Agora, como $f(-2) = e^{-2} > 0$ e $f(0) = -2 < 0$, temos do teorema de Bolzano que existe pelo menos um zero de $f(x)$ no intervalo $(-2, 0)$. E, portanto, existe pelo menos uma solução da equação dada no intervalo $(-2, 0)$.
Podemos usar o `Scilab` para estudarmos esta função. Por exemplo, podemos definir a função $f(x)$ e computá-la nos extremos do intervalo dado com os seguintes comandos:

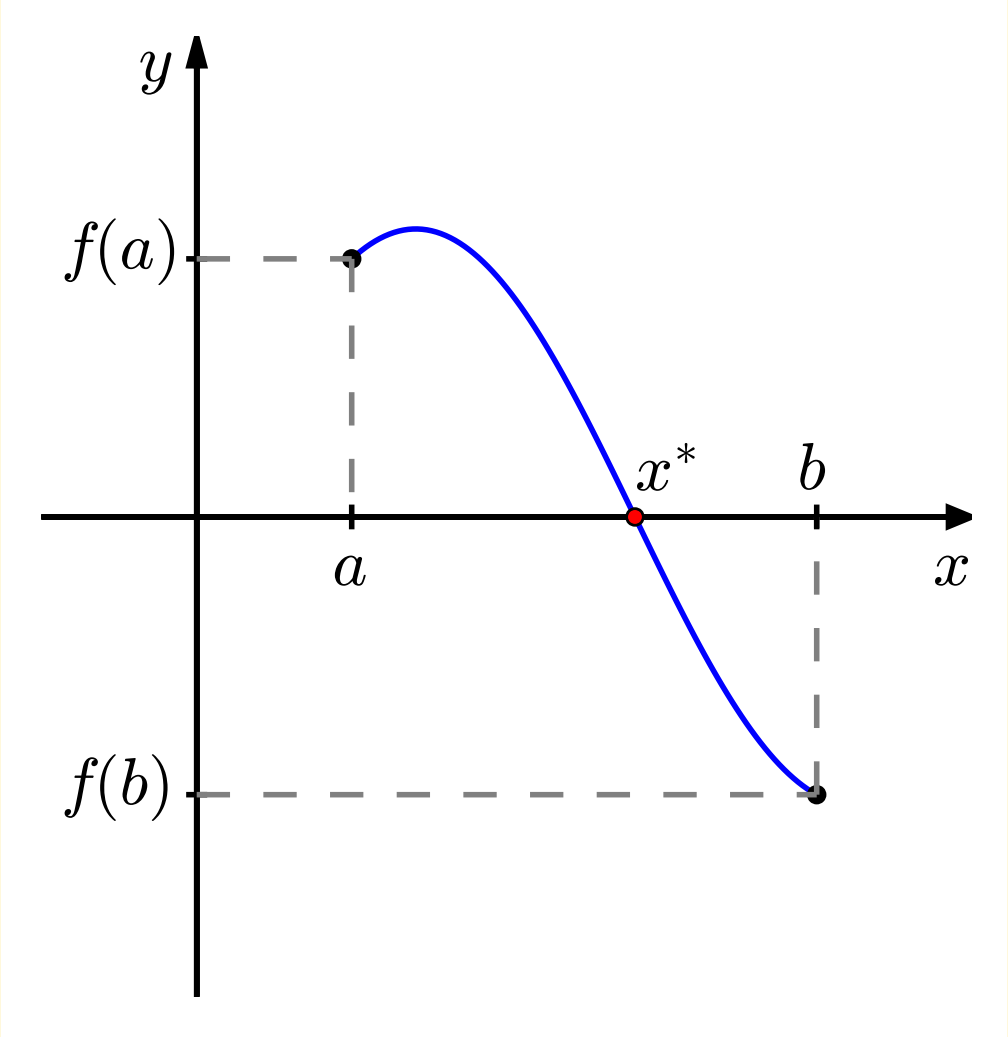


Figura 3.1: Teorema de Bolzano.

```
-->deff('y=f(x)','y=exp(x)-x-2')
-->f(-2),f(0)
 ans  =
    0.1353353
 ans  =
  - 1.
```

Alternativamente (e com maior precisão), podemos verificar diretamente o sinal da função nos pontos desejados com comando `sign`:

```
-->sign(f(-2)),sign(f(0))
 ans  =
    1.
 ans  =
  - 1.
```

$\Diamond$

Quando procuramos aproximações para zeros de funções, é aconselhável isolar cada raiz em um intervalo. Desta forma, gostaríamos de

poder garantir a existência e a unicidade da raiz dentro de um dado intervalo. A seguinte proposição nos fornece condições suficientes para tanto.

Proposição 1. *Se $f : [a, b] \to \mathbb{R}$ é um função diferenciável, $f(a) \cdot f(b) < 0$ e $f'(x) > 0$ (ou $f'(x) < 0$) para todo $x \in (a, b)$, então existe um único $x^* \in (a, b)$ tal que $f(x^*) = 0$.*

Em outras palavras, para garantirmos que exista um único zero de uma dada função diferenciável num intervalo, é suficiente que ela troque de sinal e seja monótona neste intervalo.

Exemplo 40. No Exemplo 39, mostramos que existe pelo menos um zero de $f(x) = e^x - x - 2$ no intervalo $(-2{,}0)$, pois $f(x)$ é contínua e $f(-2) \cdot f(0) < 0$. Agora, observamos que, além disso, $f'(x) = e^x - 1$ e, portanto, $f'(x) < 0$ para todo $x \in (-2{,}0)$. Logo, da Proposição 1, temos garantida a existência de um único zero no intervalo dado.

Podemos inspecionar o comportamento da função $f(x) = e^x - x -$

2 e de sua derivada fazendo seus gráficos no Scilab. Para tanto, podemos fazer o seguinte teste:

```
-->x = linspace(-2,0,50);
-->deff('y = f(x)','y=exp(x)-x-2')  // define f
-->plot(x,f(x));xgrid               // grafico de f
-->deff('y = fl(x)','y=exp(x)-1')   // a derivada
-->plot(x,fl(x));xgrid              // grafico de f'
```

A discussão feita nesta seção, especialmente o teorema de Bolzano, nos fornece os fundamentos para o método da bisseção, o qual discutimos na próxima seção.

Exercícios

E 3.1.1. Mostre que $\cos x = x$ tem solução no intervalo $[0, \pi/2]$.

E 3.1.2. Mostre que $\cos x = x$ tem uma única solução no intervalo $[0, \pi/2]$.

E 3.1.3. Mostre que a equação:

$$\ln(x) + x^3 - \frac{1}{x} = 10$$

possui uma única solução positiva.

E 3.1.4. Use o teorema de Bolzano para mostrar que o erro absoluto ao aproximar o zero da função $f(x) = e^x - x - 2$ por $\overline{x} = -1{,}841$ é menor que 10^{-3}.

E 3.1.5. Mostre que o erro absoluto associado à aproximação $\overline{x} =$

1,962 para a solução exata x^* de:

$$e^x + \sin(x) + x = 10$$

é menor que 10^{-4}.

E 3.1.6. Mostre que a equação

$$\ln(x) + x - \frac{1}{x} = v$$

possui uma solução para cada v real e que esta solução é única.