Cálculo Numérico

Um Livro Colaborativo

6 de julho de 2016

Autores

Lista de autores[1]:

Esequia Sauter - UFRGS

Fabio Souto de Azevedo - UFRGS

Pedro Henrique de Almeida Konzen - UFRGS

[1] em ordem alfabética

Licença

Nota dos autores

Este livro vem sendo construído de forma colaborativa desde 2011. Nosso intuito é de melhorá-lo, expandi-lo e adaptá-lo às necessidades de um curso semestral de cálculo numérico em nível de graduação.

Caso queira colaborar, entre em contato conosco pelo endereço de e-mail:

`livro_colaborativo@googlegroups.com`

Apresentação

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 resolução de problemas, implementação em computador 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 realmente necessárias.

Sumário

Capítulo 1

Introdução

Cálculo numérico é uma disciplina que compreende o estudo de métodos para a computação eficiente da solução de problemas matemáticos. Aliado ao avanço tecnológico dos computadores, o desenvolvimento de métodos numéricos tornou a simulação computacional de modelos matemáticos uma prática cotidiana 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. Estes 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 operações 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

$$\begin{cases} y' + xy = x, \\ y(0) = 2, \end{cases}$$

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

$$\begin{cases} y' + xy = e^{-y}, \\ y(0) = 2, \end{cases}$$

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 das modelagem 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 decomposto em centenas, dezenas e unidades:*

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

Assim, vemos que as centenas, dezenas e unidades são potências de 10.

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 a coleção de símbolos $\{\text{“,”}, -, 0, 1, 2, \ldots, b-1\}$[1], a sequência de dígitos:*

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

representa o número positivo

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

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

[1] Para sistemas de numeração com base $b > 10$, veja a Observação 1

- *No sistema de numeração decimal, i.e.* $b = 10$*, representamos o número:*

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

simplesmente por:

$$\pm d_n d_{n-1} \ldots d_1 d_0, d_{-1} d_{-2} \ldots$$

Ou seja, não usamos parênteses, nem o subíndice indicando a base.

- *Em sistemas de numeração com base* $b > 10$*, usamos as letras* A*,* B*,* C*, etc., para denotar os símbolos:* $A = 10$*,* $B = 11$*,* $C = 12$*, etc..*

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 apenas dois valores distintos:* 0 *ou* 1*. Por exemplo:*

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

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 quaternário a base é* $b = 8$*. Por exemplo:*

$$\begin{aligned}(1357,24)_8 &= 1\cdot 8^3 + 3\cdot 8^2 + 5\cdot 8^1 + 7\cdot 8^0 + 2\cdot 8^{-1} + 4\cdot 8^{-2} \\ &= 512 + 192 + 40 + 7 + 0,25 + 0,0625 = (751,3125)_{10}\end{aligned}$$

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* $S = \{\text{“, ”}, -, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$*. O número* $(E2AC)_{16}$ *no sistema decimal é igual a:*

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

Exercício 1. *Escreva os números abaixo na base decimal*

a) $(25, 13)_8$

b) $(101, 1)_2$

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

d) $(11, 2)_3$

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 queremos escrevê-lo em uma outra base b. Para tanto, consideramos um número decimal X_{10} representado na base b:

$$\begin{aligned} X_{10} &= (d_n d_{n-1} \cdots d_0, d_{-1} \cdots)_b \\ &= d_n \cdot b^n + d_{n-1} \cdot b^{n-1} + \cdots + d_1 \cdot b^1 + d_0 \cdot b^0 + d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \cdots \end{aligned}$$

Separando as partes inteira e parte fracionária de X, i.e. $X = X^i + X^f$, temos:

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

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:

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

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 6 (Conversão da parte inteira)**.** *Vamos escrever o número 125 na base* 6*. Para encontrar* d_0*, dividimos* 125 *por 6:*

$$\begin{array}{r|l} 125 & 6 \\ \underline{12} & 20 \\ 05 & \\ \underline{00} & \\ 5 & \end{array}$$

e encontramos $d_0 = 5$. *Dividindo o quociente por* 6 *para encontrar* d_1*:*

$$\begin{array}{r|l} 20 & 6 \\ \underline{18} & 3 \\ 2 & \end{array}$$

e obtemos $d_1 = 2$. *Observe que o quociente agora é menor que* 6*, ou seja, uma sucessiva divisão por* 6 *teria resto igual ao próprio quociente. Assim, concluímos que:*

$$125 = (325)_6$$

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!

Agora, para convertermos a parte fracionária X^f na base b, i.e. para encontrar os símbolos d_{-1}, d_{-2}, etc, multiplicamos a parte fracionária de X por b:

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

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 7 (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, Multiplicando-a por* 6*:*

$$\begin{array}{r} 0,58\overline{3} \\ \times\ 6 \\ \hline 3,49\overline{9} \end{array}$$

e obtemos $d_{-1} = 3$. *Agora, multiplicamos* $0,49\overline{9} = 0,5$ *por* 6*:*

$$\begin{array}{r} 0,5 \\ \times\ 6 \\ \hline 3,0 \end{array}$$

e obtemos $d_{-2} = 3$. *Portanto:*

$$125,58\overline{3} = (325,33)_6$$

As contas feitas aqui, também podem ser feitas no Scilab. Você sabe como?

Exercício 2. *Escreva cada número decimal na base b*

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

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

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

Exercício 3. *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* $b = 16$

2.2 Aritmética de Máquina

Os computadores, em geral, usam uma 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 Representação de números inteiros

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

Representação sem signal

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

$$\begin{aligned}(111...111)_2 &= 2^{n-1}+2^{n-2}+\cdots+2^1+2^0 = 2^n-1.\\ \vdots &= \\ (000...000)_2 &= 0\end{aligned}$$

Observação 2. *No Scilab, consulte sobre os comandos:* `uint8`, `uint16` *e* `uint32`.

Representação com bit de sinal

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

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

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 8. *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 9. *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 10. *Em um registro com* 8 *bits, teremos os números*

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

Observação 3. *No Scilab, consulte sobre os comandos:* `int8`, `int16` *e* `int32`.

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. Por exemplo, 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, nestes casos, 16 dígitos são foram usados para representar a parte fracionária e 15 para representar a parte inteira e um dígito, o d_{31}, está relacionado ao sinal do número.

2.2.3 Sistema de ponto flutuante

O sistema de ponto flutuante não possui quantidade fixa de dígitos para as partes inteira e fracionária do número.

Exemplo 11. *Um computador de 64 bits que usa o sistema de ponto flutuante com um dígito para o sinal, o registro:*

s	c_{10}	c_9	$\cdots$	c_0	m_{-1}	m_{-2}	$\cdots$	m_{-50}	m_{-51}

representa o número

$$(-1)^s 2^{c-1023}(1+m),$$

onde:

$$c = c_{10}2^{10} + c_9 2^9 + \cdots + c_1 2^1 + c_0 2^0$$

e

$$m = m_{-1}2^{-1} + m_{-2}2^{-2} + \cdots + m_{-50}2^{-50} + m_{-51}2^{-51}.$$

Os dígitos $c_{10}, c_9 \cdots, c_0$ *compõem a característica e* $m_{-1}, m_{-2} \cdots, m_{-51}$ *compõem a mantissa. Por exemplo, o número:*

0 10000000000 1000000000000000000000000000000000000000000000000000

representa o número:

$$(-1)^0 2^{1024-1023}(1 + 2^{-1}) = 3$$

O menor número positivo nesse sistema acontece quando $s = 0$, $c = 1$ *e* $m = 0$, *ou seja:*

$$2^{-1022}(1 + 0) \approx 0,2225 \times 10^{-307}.$$

O maior número acontece quando $s = 0$, $c = 2046$ *e* $m = 1 - 2^{-52}$, *ou seja:*

$$2^{1023}(2 - 2^{-52}) \approx 0,17977 \times 10^{309}.$$

$\Diamond$

Observação 4. *O menor número positivo e o maior número positivo representados no Scilab são, respectivamente:*

```
-->number_properties('tiny')
-->number_properties('huge')
```

Um número muito pequeno (**underflow**) geralmente é aproximado por zero e um número muito grande (**overflow**) geralmente faz o cálculo parar. Os demais números são aproximados pelo mais próximo, gerando os erros de arredondamento.

Observação 5. *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
```

Definição 2. *(Forma normalizada de ponto flutuante) Seja* x *um número real não nulo.*[2] *A representação de* x *na forma normalizada em ponto flutuante na base* $b \in \mathbb{N}$, $b > 1$, *e com* n **dígitos significativos**[3] *é dada por:*

$$fl(x) = 0, d_1 d_2 d_3 \ldots d_n \times b^E$$

onde, $d_i \in \{0, 1, \ldots, b-1\}$, $d_1 \neq 0$ *e* $E \in \mathbb{Z}$. *Os dígitos* d_i, $i = 1, 2, \ldots, n$, *formam o significando e* E *o expoente da representação de* x.

[2]Se $x = 0$, então $fl(x) = 0,000 \ldots 0 \times b^0$ será considerada sua representação em ponto flutuante.

[3]Na literatura, também encontra-se outros significados para o termo dígitos (algarismos) significativos. Veja, por exemplo, [3]

Exemplo 12. *O número* $0,05$ *é representado na forma normalizada de ponto flutuante na base* 2 *e com um dígito significativo por* $0,1 \times 2^{-1}$.

Observação 6. *Salvo especificado ao contrário, quando nos referirmos à representação em ponto flutuante de um número dado, estaremos nos referindo à representação deste número na forma normalizada de ponto flutuante na base dez.*

Exercício 4. *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 num sistema de ponto flutuante com 7 dígitos significativos.*

Solution. 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. No sistema de ponto flutuante, as representações são $0,5151515 \cdot 10^{-2}$ e $0,1205415 \cdot 10^{4}$, respectivamente. ◊

Observação 7. *Consulte sobre o comando* `format` *no Scilab.*

2.3 Origem e Definição de Erros

Quando fazemos aproximações numéricas, os erros são gerados de várias formas, sendo as principais delas as seguintes:

1. Dados de entrada: equipamentos de medição possuem precisão finita, acarretando erros nas medidas físicas.

2. Erros de Truncamento: ocorrem quando aproximamos um procedimento formado por uma sequência infinita de passos através de um outro procedimento finito. Por exemplo, a definição de integral é dada por uma soma infinita e, como veremos na terceira área, aproximarmo-la por um soma finita. Esse é um assunto que discutiremos várias vezes no curso, pois o tratamento do erro de truncamento é feito para cada método numérico.

3. Erros de Arredondamento: são aqueles relacionados com as limitações que existem na forma representar números de máquina. Sobre esse tópico dedicamos a subseção (2.3.1).

Definição 3. *Seja* x *um número real e* $\overline{x}$ *sua aproximação. O* **erro absoluto** *da aproximação* $\overline{x}$ *é definido como sendo o número:*

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

O **erro relativo** *da aproximação $\overline{x}$ é definido como sendo o número:*

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

Observação 8. *Observe que o erro relativo é adimensional e, muitas vezes, é dado em porcentagem. Ou seja, o erro relativo, em porcentagem, da aproximação $\overline{x}$ é definido por:*

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

Exemplo 13. *Se $x=\frac{1}{3}$ e $\overline{x}=0,333$, então o erro absoluto é*

$$|x-\overline{x}|=|0,\overline{3}-0,333|=0,000\overline{3}=0,\overline{3}\cdot 10^{-3}$$

e o erro relativo é

$$\frac{|x-\overline{x}|}{|x|}=\frac{0,\overline{3}\cdot 10^{-3}}{0,\overline{3}}=10^{-3}=0,1\%$$

Exemplo 14. *Observe os erros absolutos e relativos em cada caso*

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

Exercício 5. *Calcule os erros absoluto e relativo das aproximações $\overline{x}$ para x*

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

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

c) $x=100001$ *e* $\overline{x}=100000$

Definição 4. *A aproximação $\overline{x}$ de um número $x=\pm 0,d_{-1}d_{-2}d_{-3}\cdots\times 10^{m}$ possui s* **dígitos significativos corretos** *se o erro absoluto $|x-\overline{x}|$ satisfazer*[4]

$$|x-\overline{x}|\leq 0,5\times 10^{m-s}$$

[4]Observação: Não existe uma definição única na literatura para o conceito de dígitos significativos corretos, embora não precisamente equivalentes, transmitem a mesmo conceito.

Exemplo 15. *Veja os seguintes casos:*

a) *Considere* $x = 0,\overline{3}$, $\overline{x} = 0,333$ *e o erro absoluto* $\delta = |x - \overline{x}| = 0,\overline{3} \times 10^{-3} = 0,\overline{3} \times 10^{0-3}$. *Essa aproximação tem 3 dígitos significativos corretos.*

b) *Agora, considere* $x = 10,00\overline{1} = 0,1000\overline{1} \times 10^2$, $\overline{x} = 9,99933 = 0,999933 \times 10^1$ *e o erro absoluto* $\delta = |x - \overline{x}| = 0,178\overline{1} \times 10^{-2} = 0,178\overline{1} \times 10^{2-4}$. *Essa aproximação possui todos os dígitos diferentes se comparamos um a um, mas tem 4 dígitos significativos corretos.*

Exercício 6. *Verifique quantos são os dígitos significativos corretos em cada aproximação* $\overline{x}$ *para* x.

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

b) $x = 100$ *e* $\overline{x} = 99$

2.3.1 Erros de Arredondamento

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

Exemplo 16. *O número* $\frac{1}{3} = 0,\overline{3}$ *possui um representação infinita tanto na base decimal quanto na base binária. Logo, quando representamos ele no computador geramos um erro de arredondamento que denotaremos por* ϵ. *Agora considere a seguinte sequência:*

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

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* $\frac{1}{3}$. *Se calcularmos no computador essa sequência, temos que incluir os erros de arredondamento, ou seja,*

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

Portanto o limite da sequência diverge,

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

Faça o teste no scilab, colocando:

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

e itere algumas vezes a linha de comando:

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

Existem várias formas de aproximar um número em ponto flutuante $\pm 0, d_1 d_2 d_3 \ldots d_{k-1} d_k d_{k+1} \ldots d_n$ usando k dígitos significativos. As duas principais são as seguintes:

1. Por truncamento: aproximamos o número dado por:

$$\pm 0, d_1 d_2 d_3 \ldots d_k \times 10^e$$

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

2. Por arredondamento: aproximamos o número dado por:

$$\pm 0, \tilde{d}_1 \tilde{d}_2 \tilde{d}_3 \ldots \tilde{d}_k \times 10^{\tilde{e}}$$

que é a aproximação por truncamento do número:

$$\pm 0, d_1 d_2 d_3 \ldots d_k d_{k+1} \times 10^e \pm 0,5 \times 10^{e-k}$$

Exemplo 17. *Represente os números* $0,567$*;* $0,233$*;* $-0,6785$ *e* $\pi = 0,314159265... \times 10^1$ *com dois dígitos significativos por truncamento e arredondamento.*
Truncamento: $0,56$*;* $0,23$*;* $-0,67$ *e* $\pi = 0,31 \times 10^1 = 3,1$
Arredondamento: $0,57$*;* $0,23$*;* $-0,68$ *e* $\pi = 0,31 \times 10^1 = 3,1$

Observação 9. *Observe que o arredondamento pode mudar todos os dígitos e o expoente da representação em ponto flutuante de um número dado.*

Exemplo 18. *O arredondamento de* $0,9999 \times 10^{-1}$ *com* 3 *dígitos significatios é* $0,1 \times 10^0$.

Exercício 7. *Represente os números* 3276*;* $42,55$ *e* $0,00003331$ *com três dígitos significativos por truncamento e arredondamento.*

Exercício 8. *Resolva a equação* $0,1x - 0,01 = 12$ *usando arredondamento com três dígitos significativos em cada passo e compare com o resultado analítico*

2.4 Propagação de Erros

Dado uma função diferenciável f, considere $\overline{x}$ uma aproximação para x e $f(\overline{x})$ uma aproximação para $f(x)$. Sabendo o erro $\delta_x = |x - \overline{x}|$, queremos estimar o erro $\delta_f = |f(x) - f(\overline{x})|$. Pelo teorema do valor médio, existe ϵ contido no intervalo aberto formado por x e $\overline{x}$ tal que

$$f(x) - f(\overline{x}) = f'(\epsilon)(x - \overline{x}).$$

Como não conhecemos o valor de ϵ, supomos que a derivada $f'(\epsilon)$ é limitada por M ($|f'(\epsilon)| \leq M$) no intervalo fechado formado por x e $\overline{x}$ e obtemos

$$|f(x) - f(\overline{x})| \leq M|x - \overline{x}|.$$

Se $f'(x)$ não varia muito rápido nesse intervalo e supondo δ_x pequeno, aproximamos $M \approx |f'(x)|$ e temos:

$$|f(x) - f(\overline{x})| \approx |f'(x)||x - \overline{x}|,$$

ou

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

De modo geral, quando f depende de várias variáveis, a seguinte estimativa vale:

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

Exercício Resolvido 1. *Seja $f(x) = x\exp(x)$. Calcule o erro absoluto em se calcular $f(x)$ sabendo que $x = 2 \pm 0,05$.*

Solution. 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:

$$\begin{aligned} \delta_f &= |(1 + x)e^x| \cdot \delta_x \\ &= |3e^2| \cdot 0,05 = 1,084. \end{aligned}$$

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

Exercício Resolvido 2. *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.*

Solution. 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

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

$\Diamond$

Exercício Resolvido 3. *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.*

Solution. 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. $\Diamond$

Exercício Resolvido 4. *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.*

Solution. 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

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

$\Diamond$

Exercício 9. *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* $x^{\alpha} = e^{\alpha \ln(x)}$

2.5 Cancelamento Catastrófico

Operações aritméticas entre números com representação finita pode fazer com que o resultado seja dominado pelos erros de arredondamento. Em geral, esse efeito, denominado cancelamento catastrófico, acontece quando fazemos a diferença de números muito próximos entre si.

Exemplo 19. *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.

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,

$$\frac{|0,987624687925 - 0,987625|}{|0,987624687925|} = 0,00003159\% \qquad e \qquad \frac{0,987624 - 0,987624}{0,987624} = 0\%,$$

enquanto no lado direito o erro relativo é enorme,

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

Exemplo 20. *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:

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

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

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

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

Observação 10. *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 21. *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

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

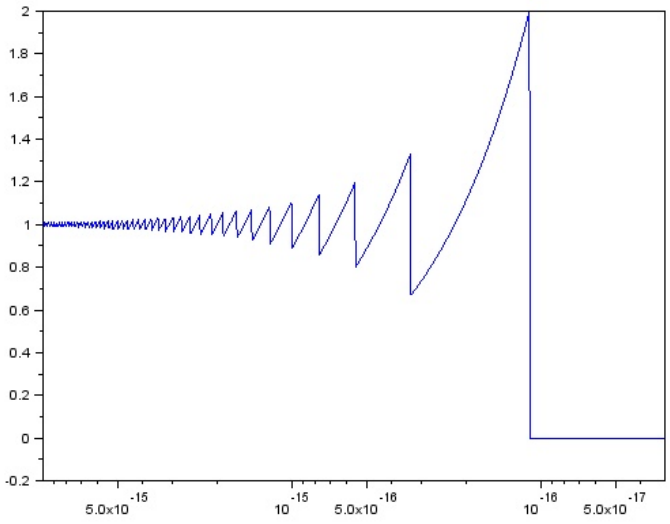
$$\begin{aligned} \tilde{x}_2 &= \frac{-b + b - \frac{4ac}{2b}}{2a} \\ &= -\frac{4ac}{4ab} \\ &= -\frac{c}{b} = -\frac{-0,140000 \times 10^{-1}}{0,300000 \times 10^3} = 0,466667 \times 10^{-4} \end{aligned}$$

Observe que o efeito catastrófico foi eliminado.

Exemplo 22. *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.



Exercício 10. *Considere a expressão*

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

para x pequeno. Verifique que

$$\lim_{x \to 0} f(x) = 0,5$$

Depois calcule no scilab $f(x)$ para $x = 10^{-5}$, $x = 10^{-6}$, $x = 10^{-7}$, $x = 10^{-8}$, $x = 10^{-9}$ e $x = 10^{-10}$. Finalmente, faça uma aproximação analítica que elimine o efeito catastrófico.

Exemplo 23. *Neste exemplo, estamos interessados em compreender mais detalhadamente o comportamento da expressão*

$$\left(1 + \frac{1}{n}\right)^n \tag{2.1}$$

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 (2.1) quando n tende a infinito é o número de Euler:

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = e = 2,718281828459... \tag{2.2}$$

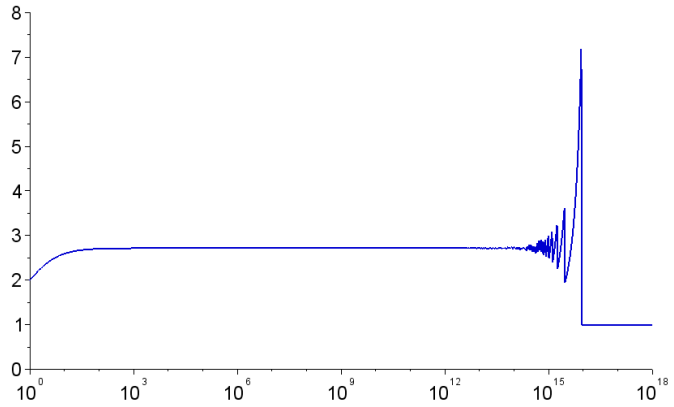
Sabemos também que a sequência produzida por (2.1) é crescente, isto é:

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

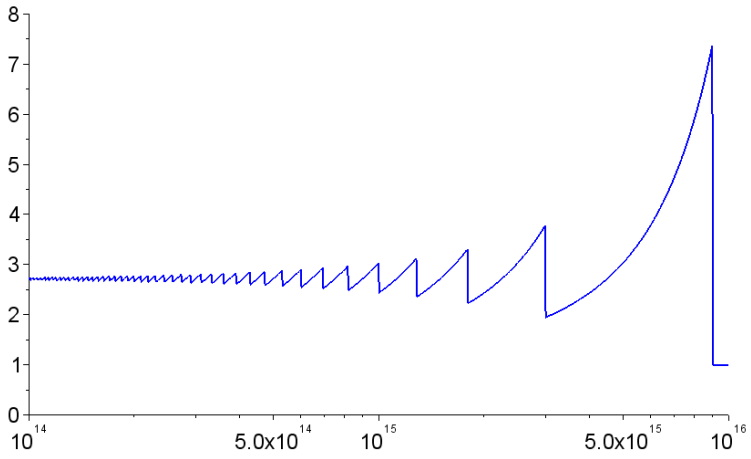
No entanto, quando calculamos essa expressão no Scilab, nos defrontamos com o seguinte resultado:

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
10	2,5937424601000		10^{20}	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 compreender por que existe um limiar N que, quando atingido torna a expressão identicamente igual a 1, observe a sequência de operações realizadas pelo computador:

$$x \;\rightarrow\; 1/x \;\rightarrow\; 1+1/x \;\rightarrow\; (1+1/x)^x \tag{2.3}$$

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 pode ser obtido pelo comando:

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

A quantidade dada por `%eps` *é chamada 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 $2,22 \times 10^{-16}$. 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 (2.3), 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:

$$x \ \rightarrow \ 1/x \ \rightarrow \ 1/(1/x) \tag{2.4}$$

compare o desempenho numérico desta sequência de operações para valores pequenos de x com o da seguinte sequência:

$$x \ \rightarrow \ 1+x \ \rightarrow \ (1+x)-1. \tag{2.5}$$

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 entender isso, observe o que acontece quando $n = 7 \times 10^{13}$:

```
-->n=7e13
 n  =
     7.000000000000000000D+13

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

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

    1.000000000000014211D+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  =
     1.421085471520200372D-14

-->1/(y-1)
 ans  =
     7.036874417766400000D+13
```

Exemplo 24 (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* $4Kg$*.*

Referências Bibliográficas

[1] Cecill and free sofware. `http://www.cecill.info`. Acessado em 30 de julho de 2015.

[2] M. Baudin. Introduction to scilab. `http://forge.scilab.org/index.php/p/docintrotoscilab/`. Acessado em 30 de julho de 2015.

[3] R.L. Burden and J.D. Faires. *Análise Numérica*. Cengage Learning, 8 edition, 2013.

[4] J. P. Demailly. *Analyse Numérique et Équations Differentielles*. EDP Sciences, Grenoble, nouvelle Édition edition, 2006.

[5] Walter Gautschi and Gabriele Inglese. Lower bounds for the condition number of vandermonde matrices. *Numerische Mathematik*, 52(3):241–250, 1987/1988.

[6] R. Rannacher. Einführung in die numerische mathematik (numerik 0). `http://numerik.uni-hd.de/~lehre/notes/num0/numerik0.pdf`. Acessado em 10.08.2014.