Anotações MS211

Aulas ministradas pela
Prof$^{\underline{a}}$ Dr.$^{\underline{a}}$ Kelly Cristina Poldi

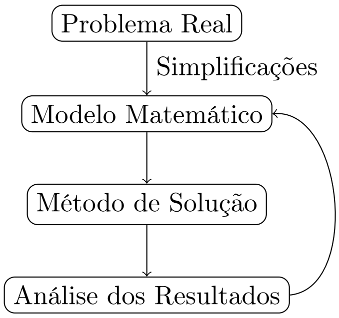
Eduardo M. F. de Souza

17 de março de 2020

Sumário

1 Erros em processos Numéricos

Em diversas áreas científicas, os métodos numéricos podem ser usados para resolução de problemas. A resolução de problemas envolve várias fases:



Uma vez resolvido um problema, pode ocorrer que a solução obtida não seja a esperada. Isto porque durante o processo de resolução pode ter ocorrido:

1. Erros de modelagem matemáticas;
2. Erro de parâmetros;
3. Erros associados aos sistema de numeração utilizado;
4. Erros resultantes das operações efetuadas;
5. Entre outros;

1.1 Representação dos Números

A representação de um número depende da base escolhida (ou disponível na máquina utilizada) e o número de dígitos usados na sua representação.

Exemplo: calcular a área de uma circunferência:

$$
\begin{aligned}
r &= 1000m \\
A &\cong 31400m^2 \\
A &\cong 31416m^2 \\
A &\cong 31415,92m^2
\end{aligned}
$$

Quanto maior o número de dígitos utilizados, maior será a precisão obtida! Além disso, um número pode ter representação finita em uma base e não finita em outra. Por exemplo:

$$(0,2)_{10} = (0,0011\overline{0011}\ldots)_2$$

1.2 Conversão de Números nos Sistemas decimal e binário

Sistema posicional:

$$
\begin{aligned}
(6)_{10} &= 6 \times 10^0 \\
(347)_{10} &= 3 \times 10^2 + 4 \times 10^1 + 7 \times 10^0
\end{aligned}
$$

$$
\begin{aligned}
(10111)_2 &= 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^1 + 1 \times 2^0 \\
&= (23)_{10}
\end{aligned}
$$

De forma geral, um número n na base β é:

$$
\begin{gathered}
(a_j aj - 1 a_{j-2} \cdots a_2 a_1 a_0)_\beta, \text{ com } 0 \leq a_k \leq \beta - 1, \text{ e } k = 1 \ldots j \\
(n)_\beta = a_j \beta^j + a_{j-1} \beta^{j-1} + \cdots + a_2 \beta^2 + a_1 \beta^1 + a_0 \beta^0
\end{gathered}
$$

1.2.1 Mudança da base decimal para a base binária

1. Parte inteira (Divisões sucessivas):

$$
\begin{aligned}
& (6)_{10} = (?)_2 \\
6 \bmod 2 = \boxed{0} \rightarrow \quad & (6)_{10} = (?\ldots\boxed{0})_2 \\
\lfloor 6 \div 2 \rfloor = 3 \quad & \\
3 \bmod 2 = \boxed{1} \rightarrow \quad & (6)_{10} = (?\ldots\boxed{1}0)_2 \\
\lfloor 3 \div 2 \rfloor = 1 \quad & \\
1 \bmod 2 = \boxed{1} \rightarrow \quad & (6)_{10} = (?\ldots\boxed{1}10)_2 \\
\lfloor 1 \div 2 \rfloor = \underline{0} \quad & \quad \textbf{(Pare!)} \\
& (6)_{10} = (110)_2
\end{aligned}
$$

2. Parte fracionária (Multiplicações sucessivas)

$$
\begin{aligned}
& (0,1875)_{10} = (?)_2 \\
0,1875 \times 2 = \boxed{0},375 \rightarrow \quad & (0,1875)_{10} = (0.\boxed{0}\ldots ?)_2 \\
0,375 \times 2 = \boxed{0},75 \rightarrow \quad & (0,1875)_{10} = (0.0\boxed{0}\ldots ?)_2 \\
0,5 \times 2 = \boxed{1},5 \rightarrow \quad & (0,1875)_{10} = (0.00\boxed{1}\ldots ?)_2 \\
0,5 \times 2 = \boxed{1},\underline{0} \rightarrow \quad & (0,1875)_{10} = (0.001\boxed{1}\ldots ?)_2 \\
\textbf{(Pare!)} \quad & (0,1875)_{10} = (0.0011)_2
\end{aligned}
$$

Obs: Alguns números não têm representação finita na base binária — um ciclo de multiplicações passa a ocorrer. O fato de um número não ter uma representação finita no sistema binário (usado nos computadores) pode acarretar na ocorrência de erros.

$$
\begin{aligned}
& (0,1)_{10} = (?)_2 \\
0,1 \times 2 = \boxed{0},2 \rightarrow \quad & (0,1)_{10} = (0.\boxed{0}\ldots ?)_2 \\
0,2 \times 2 = \boxed{0},4 \rightarrow \quad & (0,1)_{10} = (0.0\boxed{0}\ldots ?)_2 \\
0,4 \times 2 = \boxed{0},8 \rightarrow \quad & (0,1)_{10} = (0.00\boxed{0}\ldots ?)_2 \\
0,8 \times 2 = \boxed{1},6 \rightarrow \quad & (0,1)_{10} = (0.0001\boxed{1}\ldots ?)_2 \\
0,6 \times 2 = \boxed{1},2 \rightarrow \quad & (0,1)_{10} = (0.0001\boxed{1}\ldots ?)_2 \\
0,2 \times 2 = \boxed{0},4 \rightarrow \quad & (0,1)_{10} = (0.00011\boxed{0}\ldots ?)_2 \\
0,4 \times 2 = \boxed{0},8 \rightarrow \quad & (0,1)_{10} = (0.000110\boxed{0}\ldots ?)_2 \\
& \vdots \\
& (0,1)_{10} = (0,0001100110\overline{011}\ldots)_2
\end{aligned}
$$

1.2.2 Mudança da base binária para base decimal:

1. Parte inteira: expoentes positivos da direita para a esquerda

$$\begin{aligned}(10011)_2 &= 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 \\ &= 8 + 0 + 2 + 1 \\ &= (11)_{10}\end{aligned}$$

2. Parte fracionária: expoentes negativos da esquerda para direita

$$\begin{aligned}(0.101)_2 &= 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} \\ &= 0,5 + 0,125 \\ &= (0,625)_{10}\end{aligned}$$

Exercício: escreva $(1101.011)_2$ na base decimal:

$$\begin{aligned}(1101.011)_2 &= 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 0 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} \\ &= 8 + 4 + 0 + 1 + 0 + 0,25 + 0,125 \\ &= (13,375)_{10}\end{aligned}$$

1.3 Aritmética de Ponto Flutuante

Um número real pode ser representado no sistema de ponto flutuante como:

$$x = m \times \beta^e, \text{ em que:}$$
$$m = \pm\, 0, d_1 d_2 \ldots\ d_n,\ n \in \mathbb{N} \text{ (chamada mantissa)}$$

Em que:

- n é o número de dígitos na mantissa;
- β é a base do sistema;
- e é o expoente, $e \in \mathbb{Z},\ l \leq\ e \leq\ u$, sendo que l e u são inteiros fixos;
- d_j é o j-ésimo digito da mantissa, sendo que $0 \leq\ d_j \leq\ (\beta - 1),\ j = 1,\ \ldots\ ,\ n$;
- $d_1 \neq 0$;

A reunião de todos os números reais em ponto flutuante mais o zero constitui o sistema de ponto flutuante, denotado por:

$$F(\beta, n, l, u)$$

Obs: Nesse sistema, o menor número, em valor absoluto, é $(0,1) \times \beta^e$ e o maior valor, em valor absoluto, é $0,(\beta - 1)(\beta - 1)\ldots(\beta - 1) \times \beta^u$.

Considere o conjunto dos números reais $\mathbb{R}$ e o seguinte conjunto:

$$G = \{x \in \mathbb{R} : \beta^l \leq |x|\beta^u\}$$

Dado um número real x, três situações podem ocorrer (Exemplos com o caso $F(10, 5, -5, 5)$):

1. $x \in G$: nesse caso, pode-se encontrar valor aproximado de x, $\overline{x} \in F$.
   Exemplo: $x = 235.89 = 0.23589 \times 10^3$

2. $|x| < 0.1\beta^l$: x está na condição de *underflow*.
   Exemplo: $x = 0.325 \times 10^{-7}$

3. $|x| \geq 0.(\beta - 1)(\beta - 1)\ \ldots\ (\beta - 1) \times \beta^u$: x está na condição de *overflow*.
   Exemplo: $x = 0.875 \times 10^{\underline{9}}$ $(9 \geq u = 5)$

Exemplo: $F(\beta = 10, n = 4, l = -5, u = 5)$

- menor (abs): 0.1×10^{-5}
- maior (abs): 0.9999×10^5

$$G = \{x \in \mathbb{R}/10^{-6} \leq x \leq 99990\}$$

Truncamento: $x = 423.5\boxed{7} = 0.4235 \times 10^3$ Arredondamento: $x = 423.5\boxed{7} = 0.4236 \times 10^3$
Underflow: $x = 0.5 \times 10^{-9} \rightarrow -9 < l = -5$ Overflow: $x = 0.3 \times 10^8 \rightarrow 8 \geq u = 5$

1.4 Erros

Erro absoluto: $EA_x = |x - \overline{x}|$

Erro relativo: $ER_x = \frac{EA_x}{|\overline{x}|} = \frac{|x-\overline{x}|}{|\overline{x}|}$

1.4.1 Erros de Arredondamento e Truncamento

Considere um sistema de ponto flutuante com n dígitos e base 10. Podemos escrever x como:

$$x = f_x \times 10^e + g_x \times 10^{e-n}$$

onde

$$0.1 \leq f_x < 1$$

$$0 \leq g_x < 1$$

Exemplo: $x = 234,57$ e $n = 4$

$$x = 234,57 = 2 \times 10^2 + 3 \times 10^1 + 4 \times 10^0 + 5 \times^{-1} + 7 \times 10^{-2}$$

$$= (2 \times 10^{-1} + 3 \times 10^{-2} + 4 \times 10^{-3} + 5 \times 10^{-4}) \times 10^3 + (7 \times 10^-1) \times 10^{-1}$$

$$= 0.2345 \times 10^3 + 0.7 \times 10^{-1}$$

$$f_x = 0.2345 \text{ e } g_x = 0.7 \times 10^{-1}$$

$$e = 3 \text{ e } e - n = -1$$

Para representar x nesse sistema, podemos usar dois critérios:

- Truncamento: $\overline{x} = f_x \times 10^e$ e $g_x \times 10^{e-n}$ é desprezado;

$$|EA_x| < 10^{e-n}$$
$$|ER_x| < 10^{-n+1}$$

Ex: $\overline{x} = 0.2345 \times 10^3$

- Arredondamento: $\overline{x} = \begin{cases} f_x \times 10^e & \text{se } |g_x| < \frac{1}{2} \\ f_x \times 10^e + 10^{e-n} & \text{se } |g_x| \geq \frac{1}{2} \end{cases}$

$$|EA_x| < \frac{1}{2} \times 10^{e-n}$$
$$|ER_x| < \frac{1}{2} \times 10^{-n+1}$$

Ex: $\overline{x} = 0.23456 \times 10^3$

$$\begin{aligned} |EA_x| = |x - \overline{x} &= f_x \times 10^e + g_x \times 10^{e-n} - f_x \times 10^e \\ &= |g_x \times 10^{e-n} \\ |g_x| \times 10^{e-n} &< 10^{e-n} \end{aligned}$$

$|ER_x|$ fica de tarefa!

1.4.2 Teorema de Taylor

Suponha $f \in C^n[a,b]$, f^{n-1} existe em $[a,b]$ e $x_0 \in [a,b]$ $\forall x \in [a,b] \exists \xi(x)$ entree x e x_0 com $f(x) = P_n(x) + R_n(x)$, onde:

$$\begin{aligned} P_n(x) &= f(x_0) + f^{(1)}(x_0)(x - x_0) + f^{(2)}(x_0)\frac{(x-x_0)^2}{2!} + \ldots + f^{(n)}(x_0)\frac{(x-x_0)^n}{n!} \\ &= \sum_{k=0}^{n} f^{(k)}(x_0) \times \frac{(x-x_0)^k}{k!} \\ R_n(x) &= f^{(n+1)}(\xi(x)) \times \frac{(x-x_0)^{n+1}}{(n+1)!} \end{aligned}$$

$P_n(x)$ é chamado de polinômio de taylor de ordem n para f em torno de x_0 e $R_n(x)$ é o resto (erro de truncamento) associado à $P_n(x)$.

Obs: A série infinita obtida fazendo-se o limite de $P_n(x)$ quando $n \to \inf$ é chamada de série de Taylor de f em torno de x_0. No caso $x_0 = 0$, chamamos de polinômio de McLaurin (e série de McLaurin).

Por exemplo, para calcular o valor de $e^{0.5}$:
$f(x) = e^x$, $f^{(1)}(x) = e^x$, $f^{(2)}(x) = e^x$, ...

Série de Taylor ($x_0 = 0$)

$$= f(0) + f^{(1)}(0)(x-0) + f^{(2)}(0)\frac{(x-0)^2}{2!} + f^{(3)}(x)\frac{(x-0)^3}{3!} + \cdots + f^{(n)}(0)\frac{(x-0)^n}{n!}$$

$$= e^0 + e^0(x) + \frac{e^0(x)^2}{2!} + \frac{e^0(x)^3}{3!} + \cdots + \frac{e^0 x^n}{n!} + \cdots$$

$$= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

Para o cálculo de $e^{0.5}$, precisamos truncar a série, usando apenas um número finito de termos da série.

Por exemplo, usando os seis primeiros termos ($n = 5$), como aproximação:

$$e^x \approxeq 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}$$

$$e^{0.5} \approxeq 1 + 0.5 + \frac{0.5^2}{2} + \frac{0.5^3}{6} + \frac{0.5^4}{24} + \frac{0.5^5}{120}$$

$$= 1.5 + 0.125 + 0.0208333 + 0.002604166 + 0.00026041666$$

$$= 1.648697$$

O erro de truncamento é dado por:

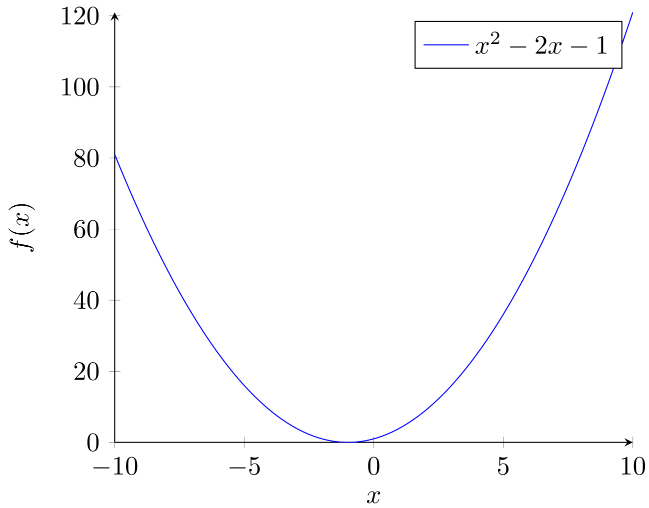
$$R_n(x) = f^{(n=1)}(\xi_x) \times \frac{(x - x_0)^{n_1}}{(n+1)!} = f^{(6)}(\alpha) \times \frac{(x-0)^6}{6!}$$

Neste caso: $\frac{\alpha^6}{6!}$, com $0 \leq \alpha \leq 0.5$ Estimativa para o erro, faço $\alpha = 0.5$

$$R_n(x) \leq \gamma = \frac{0.5^6}{720} = 0.217 \times 10^{-6}$$

1.5 Zeros reais de funções reais

Definição: um número real ξ é um zero da função $f(x)$ ou uma raíz da equação $f(x) = 0$ se $f(\xi) = 0$.



Para calcularmos raízes reais, os métodos (numéricos) geralmente são de duas fases:

1. Determinar uma aproximação inicial: localizar ou isolar as raízes (ou seja, determinar umintervalo que contenha a raiz);
2. Refinamento: melhorar essa aproximação usando métodos iterativos (obter aproximações dentro de uma precisão E fixada);

1.5.1 Fase 1: Localizar raízes de $f(x)$

Teorema: Seja $f(x)$ contínua em $[a, b]$. Se $f(a) \times f(b) < 0$, então pelo menos temos um ponto ξ tal que $f(\xi) = 0$

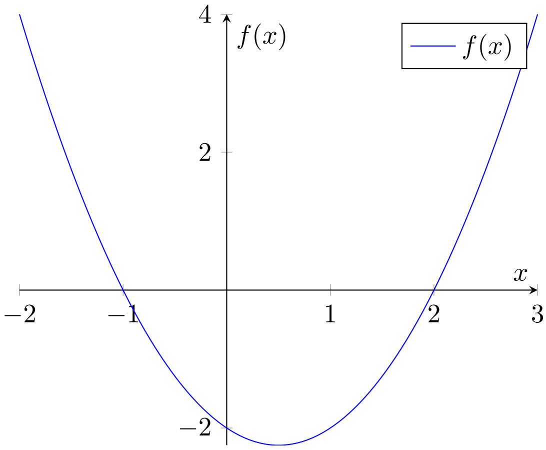
Método gráfico

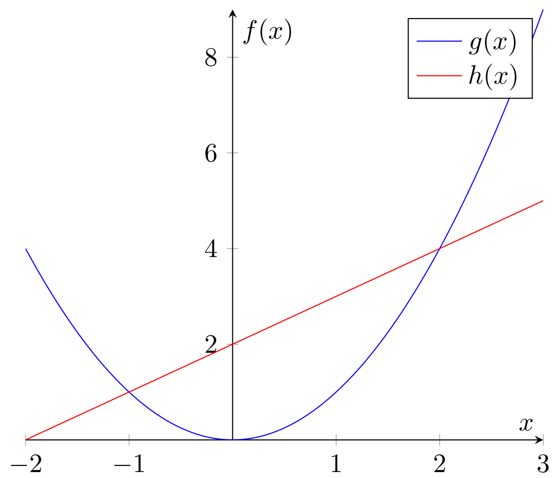
Pode ser utilizado para obter uma aproximação inicial para a raiz. Ela consiste em:

1. Construir o gráfico de $y = f(x)$ e obter sua intersecção com o eixo $\vec{Ox}$, **<u>ou</u>**
2. Escrever $f(x)$ na forma $g(x) = h(x)$ e obter a intersecção dos gráficos $g(x) - h(x) = 0 = f(x)$

Exemplo:

a $f(x) = x^2 - x - 2 = 0$





$$g(x) = h(x)$$
$$x^2 = x + 2$$

b $f(x) = x^2 - e^x = 0$

$$g(x) = h(x)$$
$$x^2 = x + 2$$

c $f(x) = \log x + x = 0$ (Tarefa!)

Fase 2: Refinamento

Métodos iterativos para se obter zeros de funções:

- ponto inicial;
- sequência de instruções (iterações);
- critério de parada;

$$|-f(x)| < E$$
$$|x_{k-1} - x_k| < E$$

Método da Bisseção

Seja f uma função contínua em $[a, b]$ e suponha $f(a) \times f(b) < 0$ (ou seja, existe raiz real). Para simplificar, vamos supor que existe uma única raíz real em $[a, b]$

O método da bisseção consistem em determinar uma sequência de intervalos $[a_i, b_i],\ i = 1,\ 2,\ \ldots$ de tal forma que $[a_i, b_i]$ sempre contenha a raíz.

De forma geral:

$$x_i = \frac{a_i + b_i}{2}$$
$$f(a_i) \times f(x_i) = \begin{cases} < 0, \quad \text{então} \begin{cases} a_{i+1} = a_i \\ b_{i+1} = x_i \end{cases} \\ \\ > 0, \quad \text{então} \begin{cases} a_{i+1} = x_i \\ b_{i+1} = b_i \end{cases} \end{cases}$$

Critério de Parada:

$$|b_k - a_k| < E$$

Estimativa para o número de iterações

Na iteração k, temos que:

$$\begin{aligned} b_k - a_k &= \frac{b_{k-1} - a_{k-1}}{2} \\ &= \frac{b_0 - a_0}{2^k} \end{aligned}$$

Queremos saber o valor de k tal que $B_k - a_k < E$, ou seja:

$$\frac{b_0 - a_0}{2^k} < E$$
$$\rightarrow \; 2^k E > b_0 - a_0$$
$$\rightarrow \; 2^k > \frac{b_0 - a_0}{E}$$
Aplicando log em ambos os lados
$$k \times \log 2 > \log (b_0 - a_0) - \log E$$
$$\boxed{k > \frac{\log (b_0 - a_0) - \log E}{\log 2}}$$

Exemplo: $f(x) = x^3 + 3x - 1 = 0$, com $[a, b] = [0, 1]$ e $E = 10^{-1} = 0,1$

$$f(a = a_1) = f(0) = -1 < 0 \text{ e } f(b = b_1) = f(1) = 3 > 0$$
$$f(a) \times f(b) = 0 \rightarrow \text{Ok!}$$
$$k = 1$$
$$x_1 = \frac{a_1 + b_1}{2} = \frac{0 + 1}{2} = 0,5$$
$$\rightarrow f(b_2 = 0,5) = 0,625 > 0$$
$$\xi \in [0, 0.5] \rightarrow |b_2 - a_2| = 0,5 > E$$
$$k = 2$$
$$x_2 = \frac{a_2 + b_2}{2} = \frac{0 + 0,5}{2} = 0,25$$
$$\rightarrow f(a_3 = 0,25) = -0,23437 < 0$$
$$\xi \in [0.25, 0.5] \rightarrow |b_3 - a_3| = 0,5 > E$$
$$k = 3$$
$$x_3 = \frac{a_3 + b_3}{2} = \frac{0,25 + 0,5}{2} = 0,375$$
$$\rightarrow f(b_4 = 0,375) = 0,1777 > 0$$
$$\xi \in [0.25, 0.375] \rightarrow |b_2 - a_2| < 0,125 > E$$
$$k = 4$$
$$x_4 = \frac{a_4 + b_4}{2} = \frac{0,25 + 0,375}{2} = 0,3125$$
$$\rightarrow f(a_5 = 0,3125) = -0,0319824 < 0$$
$$\xi \in [0.3125, 0.375] \rightarrow |b_5 - a_5| < 0,0625 < E \; \textbf{(Pare!)}$$

Estimativa para o número de iterações: $\begin{cases} a_0 = 0 \\ b_0 = 1 \end{cases}$, $E = 0,1$.

$$k > \frac{\log{(b_0 - a_0)} - \log E}{\log 2}$$
$$k > \frac{\log{(1 - 0)} - \log 0.1}{\log 2}$$