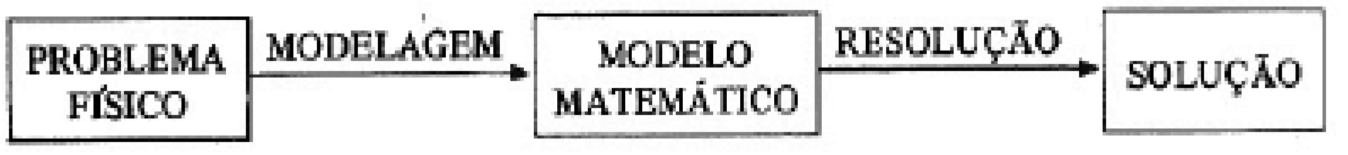
Matemática Computacional
Erros de representação

Prof. Wladimir A. Tavares
Universidade Federal do Ceará
Campus de Quixadá

Processo de Solução



- Modelagem é a fase de obtenção de um modelo matemático que descreve o comportamento do sistema físico em questão.
- Resolução é a fase de obtenção da solução do modelo matemático através da aplicação de métodos númericos.

Exemplo 1

$$d = d_0 + v_0 t + \frac{1}{2} a t^2$$

tempo	distância
3s	44,1 m
3,5s	60 m

Um variação de 16,7% do valor lido no cronômetro, gera uma variação de 36% na altura calculada

Exemplo 2

$$\Delta \ell = \ell_0 (\alpha t + \beta t^2)$$

onde:

$\Delta \ell$	—	variação do comprimento
ℓ_0	—	comprimento inicial
t	—	temperatura
α e β	—	constantes específicas para cada metal

$\ell_0 = 1$ m

$$\left.\begin{array}{l}\alpha = 0{,}001253 \\ \beta = 0{,}000068\end{array}\right\} \text{obtidos experimentalmente}$$

$$0{,}001252 < \alpha < 0{,}001254$$

$$0{,}000067 < \beta < 0{,}000069$$

Exemplo 2

$$\Delta\ell > 1 \cdot (0{,}001252 \cdot 10 + 0{,}000067 \cdot 10^2)$$

$$\Delta\ell < 1 \cdot (0{,}001254 \cdot 10 + 0{,}000069 \cdot 10^2)$$

logo:

$$0{,}019220 < \Delta\ell < 0{,}019440$$

ou, ainda,

$$\Delta\ell = 0{,}0193 \pm 10^{-4}$$

Uma imprecisão na sexta casa decimal de α e β implicam na imprecisão na quarta casa decimal na variação calculada.

Exemplo 3

Calcule a área de uma circunferência com raio 100m.

- $\pi = 3,14 \rightarrow A = 31400m^2$
- $\pi = 3,1416 \rightarrow A = 31416m^2$
- $\pi = 3,141592654 \rightarrow A = 31415,92654m^2$

Exemplo 4

Calcule $$S = \sum_{i=1}^{30000} x$$

- x = 0,5
  - Calculadora S = 15000
  - Computador S = 15000
- X = 0,11
  - Calculadora S = 3300
  - Computador S = 3300.9851074219

Conversão de Bases

$$\sum_{i=n}^{m} a_i 2^i$$

$$a_i \in \{0,1\}$$

$$n \leq 0$$

$$m > 0$$

Exemplos:

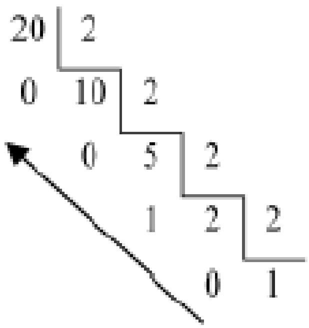
a)$(11101)_2$ = $1 \times 2^4 + 1 \times 2^3 + 1x2^2 + 0x2^1 + 1x2^0 = 16 + 8 + 4 + 0 + 1 = 29$

b)$(10001)_2$ = $1 \times 2^4 + 0 \times 2^3 + 0x2^2 + 0x2^1 + 1x2^0 = 16 + 0 + 0 + 1 = 17$

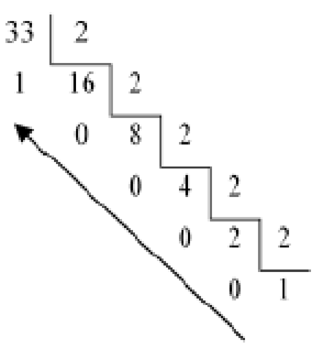
c)$(1,1)_2$ = $1 \times 2^0 + 1 \times 2^{-1} = 1,5$

d)$(10,001)_2$ = $1 \times 2^1 + 0 \times 2^0 + 0x2^{-1} + 0x2^{-2} + 1x2^{-3} = 2 + 0,125 = 2,125$

Conversão da base 10 para base 2



$(20)_{10} = (10100)_2$



$(33)_{10} = (100001)_2$

Conversão da base 10 para 2

0,8125 x 2 = 1,625
0,625 x 2 = 1,25
0,25 x 2 = 0,5
0,5 x 2 = 1,0

$(0{,}8125)_{10} = (0{,}1101)_2$

Exemplo 1.8

$$\begin{array}{r} 0{,}1875 \\ \times\ 2 \\ \hline 0{,}3750 \end{array} \qquad \begin{array}{r} 0{,}375 \\ \times\ 2 \\ \hline 0{,}750 \end{array} \qquad \begin{array}{r} 0{,}75 \\ \times\ 2 \\ \hline 1{,}50 \end{array} \qquad \begin{array}{r} 0{,}50 \\ \times\ 2 \\ \hline 1{,}00 \end{array}$$

$0{,}1875_{10} = 0{,}0011_2$

Exemplo 1.9

$$\begin{array}{r} 0{,}6 \\ \times\ 2 \\ \hline 1{,}2 \end{array} \qquad \begin{array}{r} 0{,}2 \\ \times\ 2 \\ \hline 0{,}4 \end{array} \qquad \begin{array}{r} 0{,}4 \\ \times\ 2 \\ \hline 0{,}8 \end{array} \qquad \begin{array}{r} 0{,}8 \\ \times\ 2 \\ \hline 1{,}6 \end{array} \qquad \begin{array}{r} 0{,}6 \\ \times\ 2 \\ \hline 1{,}2 \end{array}$$

... os produtos estão começando a se repetir

$0{,}6_{10} = 0{,}1001\ldots_2$

Conversão da base 10 para base 2

Exemplo 1.10

$13{,}25_{10} = 13_{10} + 0{,}25_{10}$

$$
\begin{array}{lllll}
13 & \underline{|\,2} & & & \\
1 & 6 & \underline{|\,2} & & \\
 & 0 & 3 & \underline{|\,2} & \\
 & & 1 & 1 &
\end{array}
\qquad\qquad
\begin{array}{r}
0{,}25 \\
\underline{\times\, 2} \\
0{,}50
\end{array}
\qquad
\begin{array}{r}
0{,}50 \\
\underline{\times\, 2} \\
1{,}00
\end{array}
$$

$13_{10} = 1101_2$

$0{,}25_{10} = 0{,}01_2$

$13{,}25_{10} = 1101_2 + 0{,}01_2 = 1101{,}01_2$

Conversão da base 2 para base 10

Represente os seguintes números da base 2 para base 10:

a) $(0,11)_2$

b) $(11,11)_2$

c) $(11,101)_2$

d) $(101,1001)_2$

e) $(1,1001)_2$

Conversão da base 10 para base 2

Representente os seguintes números da base 10 para base 2:

a) 0,1
b) 0,2
c) 3,5
d) 0,000015259
e) 0,000015289

Conversão da base 10 para base 2

Representante os seguintes números da base 10
para base 2:

a) 0,1

1.100110011001100110011001100110011001100110011001101 * 2^-4

a) 0,2

1.100110011001100110011001100110011001100110011001101 * 2^-3

a) 3,5

1.11 * 2^1

a) 0,000015259

1.0000000000000000111001111110110110100000101111100001 * 2^-16

a) 0,000015289

1.0000000010000000111000001010001101110110111000010101 * 2^-16

Aritmética de Ponto Flutuante

Um computador ou calculadora representa um número real no sistema denominado aritmética de ponto flutuante. Neste sistema, o número x será representado na forma:

$$\pm(d_1 d_2 \ldots d_t) x \beta^e$$

O modo de representa um sistema de ponto flutuante é F(β,t,m,M)

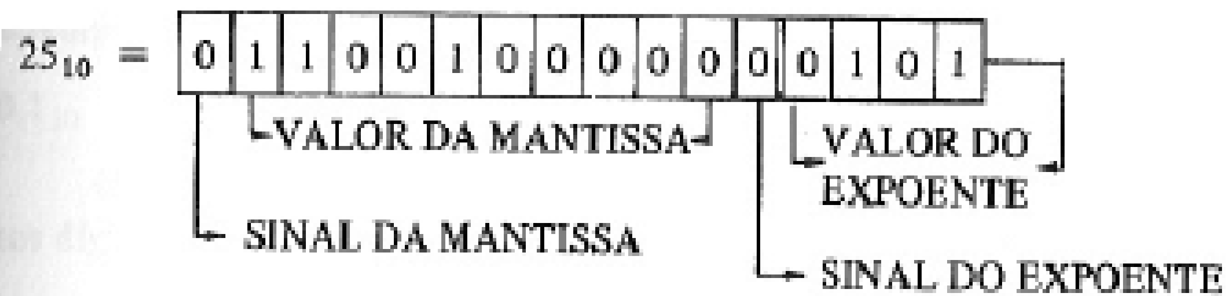
- β é a base que a máquina opera
- t é o número de dígitos da mantissa
- $0 \leq d_j \leq \beta - 1$, para j = 1,2,..,t
- $d_1 \neq 0$
- e é o expoente de β no intervalo [m,M]

Aritmética de Ponto Flutuante

Numa máquina de calcular cujo sistema de representação utilizado tenha $\beta = 2$, $t = 10$, $I = -15$ e $S = 15$, o número 25 na base decimal é, assim representado:

$$25_{10} = 11001_2 = 0{,}11001 \cdot 2^5 = 0{,}11001 \cdot 2^{101}$$

$$\left(\frac{1}{2^1} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{0}{2^4} + \frac{1}{2^5} + \frac{0}{2^6} + \frac{0}{2^7} + \frac{0}{2^8} + \frac{0}{2^9} + \frac{0}{2^{10}}\right) \cdot 2^{101}$$



Aritmética de Ponto Flutuante

O maior valor representado por esta máquina descrita no exemplo 1.13 seria:

0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

que, na base decimal, tem o seguinte valor:

$$0{,}1111111111 \cdot 2^{1111} = 32736_{10}$$

E o menor valor seria:

$$-0{,}1111111111 \cdot 2^{1111} = -32736_{10}$$

Logo, os números que podem ser representados nesta máquina estariam contidos no intervalo [– 32736 ; 32736].

Aritmética de Ponto Flutuante

Nesta máquina, ainda, o valor zero seria representado por:

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

O próximo número positivo representado seria:

0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1

$0{,}1 \cdot 2^{-15} = 0{,}000015259$

O subseqüente seria:

0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1

$0{,}1000000001 \cdot 2^{-15} = 0{,}000015289$

Através desses exemplos pode-se concluir que o conjunto dos números repre sentáveis neste sistema é um subconjunto dos números reais, dentro do interval mostrado anteriormente.

Aritmética de Ponto Flutuante

Considere o sistema de ponto flutuante F(10,3,-5,5)

- *Qual é o menor número positivo representado na máquina?*
- *Qual é o segundo menor número positivo representado na máquina?*
- *Qual é o maior número representado nesta máquina?*
- Como é possível representar $x = 0{,}245 \times 10^{-7}$?
- Como é possível representar $x = 0{,}875 \times 10^{9}$?

Aritmética de Ponto Flutuante

Considere o sistema de ponto flutuante F(10,3,-5,5)

- *Qual é o menor número positivo representado na máquina?*

$0,100 \times 10^{-5} = 0,000001$

- *Qual é o segundo menor número positivo representado na máquina?*

$0,101 \times 10^{-5} = 0,00000101$

- *Qual é o maior número representado nesta máquina?*

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

- Como é possível representar x = 0,245 x 10-7?

Neste caso, x < menor número do sistema. Nesta situação, a máquina acusa a ocorrência de **underflow**.

- Como é possível representar x = 0,875 x 109?

Neste caso, x > maior número do sistema. Nesta situação, a máquina acusa a ocorrência de **overflow**.

Aritmética de Ponto Flutuante

Um parâmetro que é muito utilizado para se avaliar a precisão de um determinado sistema de representação é o número de casas decimais exatas da mantissa e este valor é dado pelo valor decimal do último bit da mantissa, ou seja, o bit de maior significância. Logo:

$$\text{PRECISÃO} \leqslant \frac{1}{\beta^t}$$

Exemplo 1.18

Numa máquina com $\beta = 2$ e $t = 10$, a precisão da mantissa é da ordem de $\frac{1}{2^{10}} = 10^{-3}$. Logo, o número de dígitos significativos é 3.

Para concluir este item sobre erros de arredondamento, deve-se ressaltar a importância de se saber o número de dígitos significativos do sistema de representação da máquina que está sendo utilizada para que se tenha noção da precisão do resultado obtido.

Aritmética de Ponto Flutuante

	Sinal	Expoente	Mantissa
float	1 bit	8 bits	24 bits
double	1 bit	11 bits	53 bits

A precisão do sistema de ponto flutuante de 32 bits é

$$\frac{1}{2^{24}} = 0{,}59x10^{-7}$$

O número de dígitos significativos é 7.
A precisão do sistema de ponto flutuante de 64 bits é

$$\frac{1}{2^{53}} = 0{,}11x10^{-15}$$

O número de dígitos significativos é 15.

Aritmética de Ponto Flutuante

```
int main(){
   float eps = 1.0, eps1;
   do{
      eps = eps/2.0;
      eps1 = eps + 1.0;
   }while(eps1 > 1.0);
   printf("O seguinte valor vale zero \n");
   printf("%e \n", eps);
}
```

O seguinte valor vale zero

5.960464e-008

O número de dígitos significativos é 7 casas decimais.

Aritmética de Ponto Flutuante

```
int main(){
  double eps = 1.0, eps1;
  do{
    eps = eps/2.0;
    eps1 = eps + 1.0;
  }while(eps1 > 1.0);
  printf("O seguinte valor vale zero \n");
  printf("%e \n", eps);
}
```

O seguinte valor vale zero

1.110223e-016

O número de dígitos significativos é 15 casas decimais.

Aritmética de Ponto Flutuante

```
void fraiz(float a, float b, float c, float & x1, float & x2){
  float d = sqrt(b*b - 4*a*c);
  x1 = (-b + d)/(2*a);
  x2 = (-b - d)/(2*a);
}
void draiz(double a, double b, double c, double & x1, double & x2){
  double d = sqrt(b*b - 4*a*c);
  x1 = (-b + d)/(2*a);
  x2 = (-b - d)/(2*a);
}
int main(){
  float x1,x2;
  fraiz(1.0, 3000.001,3.0,x1,x2);
  printf("x1 %.6f x2 %.6f\n", x1, x2);
  double x11,x22;
  draiz(1.0, 3000.001,3.0,x11,x22);
  printf("x1 %.6f x2 %.6f\n", x11, x22);
}
```

Aritmética de Ponto Flutuante

```
x1 -0.000977 x2 -3000.000000
x1 -0.001000 x2 -3000.000000
```

Erros de Arredondamento e Truncamento

- Quando um número em sua forma normalizada possui mais dígitos significativos que o sistema pode suportar, será realizada uma aproximação para um valor que o sistema pode suportar com a perda de dígitos significativos.
- As duas formas de fazer isso são **arredondamento** e **truncamento**

Truncamento

- O **truncamento** consiste em, simplesmente, descartar os últimos dígitos significativos do número que estão fora do alcance do sistema.

$$F(10, 4, -2, 2)$$

$$x = 12,456 = 0,12456 * 10^2$$

$$\overline{x} = 0.1245 * 10^2$$

$$F(2, 3, -2, 2)$$

$$x = 101,111 = 0,101111 * 2^3$$

$$\overline{x} = 0.1011 * 23$$

Arrendodamento

- O **arredondamento** consiste em somar ½ * β^{-t} a mantissa e truncar o resultado.

$$F(10,3,5,5)$$

$$x = 0,123456$$

$$x = 0,123456 + 0,5 * 10^{-3} = 0,123456 + 0,0005 = 0,123956$$

$$\bar{x} = 0,123$$

$$F(2,3,5,5)$$

$$x = 0,10011$$

$$x = 0,10011 + 0,5 * 2^{-3} = 0,10011 + 0,0001 = 0,10101$$

$$\bar{x} = 0,101$$

Arredondamento

$$F(2,10,-15,15)$$

$$x = (0,1)_{10} = (0,00011001100\ldots)_2$$

$$x = (0,00011001100\ldots)_2 + 2^{-11}$$

$$x = (0,00011001100\ldots)_2 + (0,00000000001)_2$$

$$x = (0,000110011011100..)_2$$

$$\overline{x} = (0,0001100110)_2 = 0,099976$$

Arredondamento

$$F(2,10,-15,15)$$

$$x = (0{,}00001527)_{10} = (0{,}1000000000011\ldots * 2^{-15})_2$$

$$x = (0{,}1000000000011\ldots)_2 + 2^{-11}$$

$$x = (0{,}1000000000011\ldots)_2 + (0{,}0000000001)_2$$

$$x = (0{,}1000000001..)_2$$

$$\bar{x} = (0{,}1000000000 * 2^{-15})_2 = (0{,}00001529)_{10}$$

Exercício

Arredonde para o sistema F(10,2,-100,100):

a) 11,5749

b) 2220,0732

c) 0,0845

d) 0,0245 + 1,888

e) 0,654 x 0,018

Erro de Truncamento em Processos

- Esse erro surge cada vez que substituirmos um processo matemático infinito por um processo finito ou discreto.
- Exemplo: Serie de Taylor da função f(x) = e^x

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots, \text{ então}$$

$$e^1 = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots$$

Erro de Truncamento em Processos

- Teremos uma melhor aproximação quando n for muito grande mas tornaria o processamento muito alto. Nesse caso interrompermos os cálculos quando uma determinada precisão é atingida

Erro de Truncamento em Processos

```c
double exp(double x, double eps, int maxiter)
{
  int iter;
  double sol, solold, num, den, ea;
  iter = 1;
  num  = den = sol = ea = 1.0;
  printf("iter %d sol %e ea %e\n", iter, sol, ea);
  do{
    solold = sol;
    num *= x;
    den *= iter;
    sol += num/den;
    ea = fabs((sol-solold)/sol);
    iter++;
    printf("iter %d sol %e ea %e\n", iter, sol, ea);
  }while(ea > eps && iter < maxiter);
  return sol;
}
int main(){
  printf("%e\n", exp(0.5, 0.0001, 10) );
}
```

Erro de Truncamento em Processos

iter 1 sol 1.000000e+000 ea 1.000000e+000
iter 2 sol 1.500000e+000 ea 3.333333e-001
iter 3 sol 1.625000e+000 ea 7.692308e-002
iter 4 sol 1.645833e+000 ea 1.265823e-002
iter 5 sol 1.648438e+000 ea 1.579779e-003
iter 6 sol 1.648698e+000 ea 1.579529e-004
iter 7 sol 1.648720e+000 ea 1.316257e-005
1.648720e+000
e = 1.644821…

Erros absolutos e relativos

- **Erro absoluto**: Definimos como erro absoluto a diferença entre o valor exato de um número x e de seu valor aproximado

$$EA_x = x - \bar{x}.$$

- Em geral, apenas o valor aproximado é conhecido, e, neste caso, é impossível obter o valor exato do erro absoluto. Neste caso obtemos um limitante superior ou uma estimativa para o módulo do erro absoluto.
- Por exemplo, sabemos que pi pertence ao intervalo (3.14, 3.15), logo $|EA_{pi}| < 0.01$.

Erros absolutos e relativos

- **Erro relativo**: O erro relativo leva em consideração a ordem de grandeza do número aproximado, dessa forma ele é calculado como

$$ERx = EAx/\bar{x}$$

$$\bar{x} = 2112.9 \text{ com } |EAx| < 0.1$$

$$\bar{y} = 5.3 \text{ com } |EAy| < 0.1.$$

$$|ERx| < 4.7 * 10 - 5$$

$$|ERy| < 0.02.$$

Propagação de erros

Suponha que as operações abaixo sejam processadas em uma máquina com 4 dígitos significativos utilizando arredondamento.

$x_1 = 0,3491 \times 10^4$ e $x_2 = 0,2345 \times 10^0$

Calcule (x2 + x1) - x1

Alinhando os pontos decimais de cada número:

$$x_1 = 0,3491 \times 10^4$$

$$x_2 = 0,2345 \times 10^0 = 0,00002345$$

$$x_2 + x_1 = (0,3491 + 0,00002345) \times 10^4 = 0,34912345 \times 10^4 = 0,3491 \times 10^4$$

$$(x2 + x1) - x1 = (0,3491 - 0,3491) \times 10^4$$

$$(x2 + x1) - x1 = 0,0000$$

Propagação de erros

Suponha que as operações abaixo sejam processadas em uma máquina com 4 dígitos significativos utilizando arredondamento.

$x_1 = 0,3491 \times 10^4$ e $x_2 = 0,2345 \times 10^0$

Calcule x2 + (x1 - x1)

$$\mathrm{x}_1 = 0,3491 \times 10^4$$

$$x1 - x1 = (0,3491 + 0,3491) \times 10^4 = \overline{0,0000}$$

$$x_2 + (x_1 - x_1) = (0,2345 + 0,0000) \times 10^0$$

$$x_2 + (x_1 - x_1) = 0,2345$$

Propagação de erros

Considere a solução do sistema linear:

$$\begin{cases} 0{,}0030x1 + 30x2 = 5{,}0010 \\ x1 + 4x2 = 1 \end{cases}$$

A solução exata é x1 = 1/3 e x2 = 1/6

Multiplique a primeira equação por $\dfrac{-1}{0{,}0030}$

$$\begin{cases} -x1 - 10^4 x2 = -1{,}667x10^3 \\ x1 + 4x2 = 1 \end{cases}$$

Some a segunda equação com a primeira

$$(4 - 10^4)x2 = (1 - 1{,}667x10^3) \Leftrightarrow (0{,}4\text{x}10^1 - 0{,}1\text{x}10^5)\text{x}2 = (0{,}1\text{x}10^1 - 0{,}1667\text{x}10^4)$$

$$(0{,}00004\text{x}10^5 - 0{,}1\text{x}10^5)\text{x}2 = (0{,}0001\text{x}10^4 - 0{,}1667\text{x}10^4) \Leftrightarrow (-0{,}09996\text{x}10^5)\text{x}2 = (-0{,}1666\text{x}10^4)$$

$$(-0{,}9996\text{x}10^4)\text{x}2 = (-0{,}1666\text{x}10^4) \Leftrightarrow \text{x}2 = \frac{-0{,}1666\text{x}10^4}{-0{,}9996\text{x}10^4} = 0{,}1667$$

Propagação de erros

O valor de x1 pode ser obtido a partir da 1° equação:

$$-x1-10^4 x2=-1,667x10^3$$

$$-x1-0,1667x10^4=-1,667x10^3$$

$$-x1=-1,667x10^3+0,1667x10^4$$

$$-x1=-0,1667x10^4+0,1667x10^4$$

$$-x1=0,0000$$

Propagação de erros

$$\begin{cases} 0{,}0030x1 + 30x2 = 5{,}0010 \\ x1 + 4x2 = 1 \end{cases}$$

A solução exata é x1 = 1/3 e x2 = 1/6

Multiplique a segunda equação por - 0,003

$$\begin{cases} 0{,}0030x1 + 30x2 = 5{,}0010 \\ -\ 0{,}0030x1 - 0{,}012x2 = -\ 0{,}003 \end{cases}$$

Some a segunda equação com a primeira

$$(30 - 0{,}012)x2 = (5{,}0010 - 0{,}003) \Leftrightarrow (0{,}3\text{x}10^2 - 0{,}00012\text{x}10^2)\text{x}2 = (0{,}50010\text{x}10^1 - 0{,}0003\text{x}10^1)$$

$$(0{,}29988\text{x}10^2)\text{x}2 = (0{,}4998\text{x}10^1) \Leftrightarrow \text{x}2 = \frac{0{,}29988\text{x}10^2}{0{,}4998\text{x}10^1} = 1{,}6667$$

Propagação de erros

O valor de x1 pode ser obtido a partir da 2° equação:

$$-0{,}0030x1-0{,}012x2=-0{,}003$$

$$0{,}0030x1+0{,}012(0{,}16667)=0{,}003$$

$$x1=\frac{0{,}00099996}{0{,}003}=0{,}3333$$

Exemplo

Considere o sistema F(10,3,-5,5), determine

$$S = \sum_{i=1}^{10} 0,333$$

$$\frac{\begin{array}{c}0,333\\0,333\end{array}}{0,666} \quad \frac{\begin{array}{c}0,666\\0,333\end{array}}{0,999} \quad \frac{\begin{array}{c}0,999\\0,333\end{array}}{1,332} \quad \frac{\begin{array}{c}0,133\\0,033\end{array}}{0,166x10^{1}} \quad \frac{\begin{array}{c}0,166x10^{1}\\0,033x10^{1}\end{array}}{0,199x10^{1}}$$

$$\frac{\begin{array}{c}0,199x10^{1}\\0,033x10^{1}\end{array}}{0,232x10^{1}} \quad \frac{\begin{array}{c}0,232x10^{1}\\0,033x10^{1}\end{array}}{0,265x10^{1}} \quad \frac{\begin{array}{c}0,265x10^{1}\\0,033x10^{1}\end{array}}{0,298x10^{1}} \quad \frac{\begin{array}{c}0,298x10^{1}\\0,033x10^{1}\end{array}}{0,331x10^{1}}$$

$$E = 0,333x10^{1} - 0,331x10^{1} = 0,002x10^{1} = 0,02$$