

TN-726 - Métodos Numéricos

Conceitos

Marcelo Zamith

e-mail: mzamith@ufrj.br

<https://www.dcc.ufrj.br/~marcelo/>

Universidade Federal Rural do Rio de Janeiro - DCC



Introdução

Introdução

- ▶ Métodos e técnicas **aproximativas**;
 - algoritmos - conjunto específico de problemas.
- ▶ solução de problemas físicos e matemáticos;
- ▶ o computador digital:
 - operações matemáticas repetitivas;
 - tempo computacional aceitável;

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;
 - limitação do Método Numérico empregado;

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;
 - limitação do Método Numérico empregado; e ←

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;
 - limitação do Método Numérico empregado; e ←
 - capacidade de representação do computador.

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;
 - limitação do Método Numérico empregado; e ←
 - capacidade de representação do computador. ←

Introdução

- ▶ Não é raro obtermos resultados com métodos numéricos distantes do resultado esperado, **solução exata**.
- ▶ Resultados numéricos dependem:
 - da precisão dos dados de entrada;
 - da representação dos dados;
 - limitação do Método Numérico empregado; e ←
 - capacidade de representação do computador. ←
- ▶ **Erro** é a diferença entre a solução exata e sua versão aproximada.

Erros Absolutos e Relativos

Erros Absolutos e Relativos

Introdução

- ▶ Precisão de aproximações.
- ▶ Valores aproximados por diversas razões:
 - divisões inexatas;
 - números irracionais;
 - abandono de casas decimais de mais baixa ordem;
 - etc....

Observação:

Computador digital trabalha sempre com um número fixo de algarismos.

Erros Absolutos e Relativos

Erro absoluto

Definição:

Erro absoluto: Quando se substitui um valor a por outro aproximado a' , onde $a \neq a'$. Assim, diz-se que o erro absoluto cometido é de:

$$EA = |a - a'|$$

Erros Absolutos e Relativos

Erro absoluto

Definição:

Erro absoluto: Quando se substitui um valor a por outro aproximado a' , onde $a \neq a'$. Assim, diz-se que o erro absoluto cometido é de:

$$EA = |a - a'|$$

- Mas o a não é conhecido, e aí?

Erros Absolutos e Relativos

Erro absoluto

Definição:

Cota superior de erro de um valor aproximado que substitui certo valor exato é um valor superior ao módulo da diferença de ambos.

Erros Absolutos e Relativos

Erro absoluto

Definição:

Cota superior de erro de um valor aproximado que substitui certo valor exato é um valor superior ao módulo da diferença de ambos.

$$\epsilon \geq EA \rightarrow \epsilon \geq |a' - a|$$

Erros Absolutos e Relativos

Erro relativo

Definição:

Erro relativo: Chama-se de erro relativo cometido sobre um valor a , quando este é aproximado por a' , ao quociente positivo:

$$ER = \frac{|EA|}{|a|}$$

Erros Absolutos e Relativos

Resumo

- ▶ **Erro absoluto** - $|EA|$: diferença entre o valor **exato** (x) de um número e seu valor **aproximado** ($\bar{x}$):
 - $EA = |x - \bar{x}|$, pode ou estar em módulo. O sinal é importante ?
- ▶ Em geral, apenas o valor **aproximado** ($\bar{x}$) é conhecido, tornando-se impossível obter o valor exato do erro absoluto.
- ▶ **Exemplo**: $2.71 < e' < 2.72$. Assim temos: $|EA_e| = |e - \bar{e}| < 0.1$
 - Define-se os limites superiores e inferiores.

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;
2. $\bar{y} = 5.3$ tal que $|EA_y| < 0.1$.

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

2. $\bar{y} = 5.3$ tal que $|EA_y| < 0.1$.

- Sabendo-se que os dois números têm o mesmo erro absoluto, pode-se afirmar que estão sendo representados com mesma precisão ?

Erros Absolutos e Relativos

Resumo

- ▶ Para avaliar a precisão de um cálculo independente da ordem de grandeza, utiliza-se o **erro relativo**.
- ▶ **Erro relativo** é definido como o erro absoluto dividido pelo valor aproximado.

$$ER_x = \frac{|EA_x|}{|\bar{x}|} = \frac{|x - \bar{x}|}{|\bar{x}|}$$

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

$$\blacksquare |ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0.1}{2112.9} \approx 4.7 \times 10^{-5}$$

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

$$\blacksquare |ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0.1}{2112.9} \approx 4.7 \times 10^{-5}$$

2. $\bar{y} = 5.3$ tal que $|EA_y| < 0.1$.

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

■ $|ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0.1}{2112.9} \approx 4.7 \times 10^{-5}$

2. $\bar{y} = 5.3$ tal que $|EA_y| < 0.1$.

■ $|ER_y| = \frac{|EA_y|}{|\bar{y}|} < \frac{0.1}{5.3} \approx 0.02$

Erros Absolutos e Relativos

Resumo

► Exemplos:

1. $\bar{x} = 2112.9$ tal que $|EA_x| < 0.1$;

$$\blacksquare |ER_x| = \frac{|EA_x|}{|\bar{x}|} < \frac{0.1}{2112.9} \approx 4.7 \times 10^{-5}$$

2. $\bar{y} = 5.3$ tal que $|EA_y| < 0.1$.

$$\blacksquare |ER_y| = \frac{|EA_y|}{|\bar{y}|} < \frac{0.1}{5.3} \approx 0.02$$

Conclusão:

O número x é representado com maior precisão que o número y ou o erro foi menor em x do que em y .

Erros Absolutos e Relativos

- ▶ Se o valor exato de x não é conhecido, então como achar os erros absolutos e relativos ?

Erros Absolutos e Relativos

- ▶ Se o valor exato de x não é conhecido, então como achar os erros absolutos e relativos ?
 - Utiliza-se a medida de erro entre as aproximações:

Erros Absolutos e Relativos

- ▶ Se o valor exato de x não é conhecido, então como achar os erros absolutos e relativos ?
 - Utiliza-se a medida de erro entre as aproximações:

$$\epsilon = \frac{|x_{novo} - x_{antigo}|}{|x_{novo}|}$$

Exemplos - Método Numérico

Métodos

Número de Euler

- ▶ Erro de representação do método numérico (contínua para discreto).
- ▶ Exemplo: Número de Euler:

Métodos

Número de Euler

- ▶ Erro de representação do método numérico (contínua para discreto).
- ▶ Exemplo: Número de Euler:

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x$$

Métodos

Número de Euler

- ▶ Erro de representação do método numérico (contínua para discreto).
- ▶ Exemplo: Número de Euler:

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x} \right)^x$$

$$e = \sum_{i=0}^{\infty} \frac{1}{i!}$$

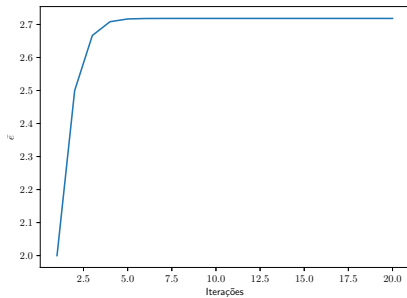
Métodos

Número de Euler

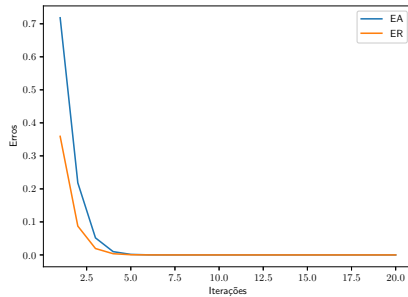
i	e'_i	$ e - e'_i $	$\frac{ e - e'_i }{ e'_i }$
1	2.000000000	0.718281828	0.359140914
2	2.500000000	0.218281828	0.087312731
3	2.666666667	0.051615162	0.019355686
4	2.708333333	0.009948495	0.003673291
5	2.716666667	0.001615162	0.000594538
6	2.718055556	0.000226273	0.000083248
7	2.718253968	0.000027860	0.000010249
8	2.718278770	0.000003059	0.000001125
9	2.718281526	0.000000303	0.000000111
10	2.718281801	0.000000027	0.000000010

Métodos

Número de Euler



Convergência



$|EA|$ e $|ER|$

Métodos

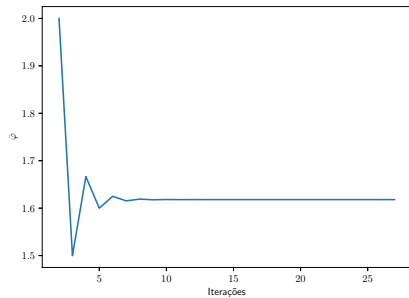
φ - número de ouro

$$\varphi = \frac{1 + \sqrt{5}}{2} \approx 1.6180339887499$$

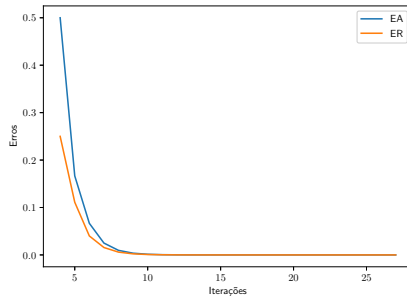
i	φ_{i-1}	$\bar{\varphi}_i$	$EA = \varphi_{i-1} - \bar{\varphi}_i $	$\frac{EA}{ \bar{\varphi}_i }$
1	1.0000000000	0.0000000000	-	-
2	2.0000000000	1.0000000000	1.00E+00	1.00E+00
3	1.5000000000	2.0000000000	1.00E+00	5.00E-01
4	1.6666666667	1.5000000000	5.00E-01	3.33E-01
5	1.6000000000	1.6666666667	1.67E-01	1.00E-01
...	...	...	...	...
27	1.6180339887	1.6180339888	1.00E-10	6.18E-11

Métodos

φ - número de ouro



Convergência



$|EA|$ e $|ER|$

Métodos

π

- ▶ Outro exemplo: π

Métodos

π

- ▶ Outro exemplo: π

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}$$

Métodos

π

- ▶ Outro exemplo: π

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}$$

$$\pi = 4 \times \left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots \right)$$

Métodos

π

- ▶ Outro exemplo: π

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}$$

$$\pi = 4 \times \left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots \right)$$

Métodos

π

- Precisão do método:

i	π_i	$EA = \pi_i - \pi_{i-1} $
2.01E+02	3.14656774718296	1.00E-02
2.00E+03	3.14209240368353	1.00E-03
2.00E+04	3.14164265108989	1.00E-04
2.00E+05	3.14159765356476	1.00E-05
2.00E+06	3.14159315358947	1.00E-06
2.00E+07	3.14159270358982	1.00E-07
2.00E+08	3.14159264858941	1.00E-08
2.00E+09	3.14159265308808	1.00E-09
2.00E+10	3.14159265363840	1.00E-10

Métodos

π

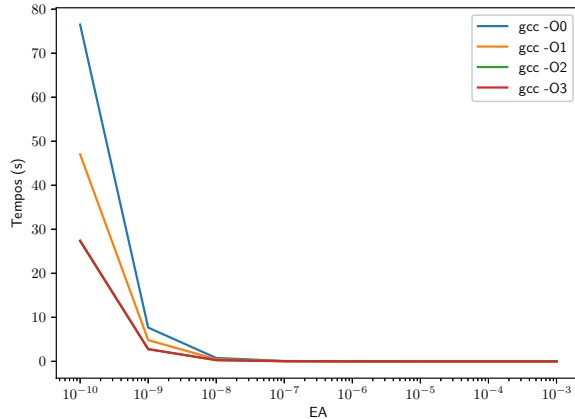
- Tempo de processamento (s) versus erro:

$ EA $	gcc -00	gcc -01	gcc -02	gcc -03	Aceleração
1.00E-02	5.00E-06	0.00E+00	0.00E+00	0.00E+00	0.000
1.00E-03	2.30E-05	4.00E-06	3.00E-06	3.00E-06	7.667
1.00E-04	8.00E-05	5.90E-05	2.90E-05	2.90E-05	2.759
1.00E-05	8.02E-04	5.05E-04	2.88E-04	2.89E-04	2.775
1.00E-06	8.06E-03	5.12E-03	2.89E-03	2.90E-03	2.780
1.00E-07	7.79E-02	4.85E-02	2.77E-02	2.77E-02	2.812
1.00E-08	7.68E-01	4.81E-01	2.77E-01	2.78E-01	2.760
1.00E-09	7.67E+00	4.82E+00	2.77E+00	2.77E+00	2.767
1.00E-10	7.65E+01	4.70E+01	2.73E+01	2.74E+01	2.788

Métodos

π

- Tempo de processamento (s) versus erro:



Trucamento e arredondamento

Erros de Arredondamento e truncamento

- ▶ **Truncamento:** Descarte dos dígitos finais de uma representação exata por limitações de representação em ponto flutuante.
 - $x = f_x \times 10^e + g_x \times 10^{e-t}$
 - onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$.
 - A parte $g_x \times 10^{e-t}$ não é considerada na mantissa;

Erros de Arredondamento e truncamento

► Exemplo: Se $t = 4$ e $x = 234,57$, então

1. $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1}$ ($g_x \times 10^{e-t}$ desprezado)

Erros de Arredondamento e truncamento

► Exemplo: Se $t = 4$ e $x = 234,57$, então

1. $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1}$ ($g_x \times 10^{e-t}$ desprezado)

2. $\bar{x} = 0,2345 \times 10^3$

Erros de Arredondamento e truncamento

► Exemplo: Se $t = 4$ e $x = 234,57$, então

1. $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1}$ ($g_x \times 10^{e-t}$ desprezado)

2. $\bar{x} = 0,2345 \times 10^3$

3. Logo, o $|ER_x| = 10^{-t+1}$

Erros de Arredondamento e truncamento

► Exemplo: Se $t = 4$ e $x = 234,57$, então

1. $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1}$ ($g_x \times 10^{e-t}$ desprezado)

2. $\bar{x} = 0,2345 \times 10^3$

3. Logo, o $|ER_x| = 10^{-t+1}$

$$|EA_x| = |x - \bar{x}| = |g_x| \times 10^{e-t} < \times 10^{e-t}, \text{ visto que } |g_x| < 1$$

Erros de Arredondamento e truncamento

► Exemplo: Se $t = 4$ e $x = 234,57$, então

1. $x = 0,2345 \times 10^3 + 0,7 \times 10^{-1}$ ($g_x \times 10^{e-t}$ desprezado)

2. $\bar{x} = 0,2345 \times 10^3$

3. Logo, o $|ER_x| = 10^{-t+1}$

$$|EA_x| = |x - \bar{x}| = |g_x| \times 10^{e-t} < \times 10^{e-t}, \text{ visto que } |g_x| < 1$$

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{10^{e-t}}{0,1 \times 10^e} = 10^{-t+1}$$

Erros de Arredondamento e truncamento

- **Arredondamento simétrico:** apresentação de uma aproximação do número. Neste caso, f_x é modificado de acordo com o valor de g_x (Utiliza-se o arredondamento simétrico):

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

Erros de Arredondamento e truncamento

Observações:

- ▶ arredondamento simétrico é normalmente utilizado porque minimiza o erro cometido;
- ▶ arredondamento por excesso utilizado no arredondamento dos erros de forma a obter um majorante.

Erros de Arredondamento e truncamento

- ▶ Se $|g_x| < \frac{1}{2}$ então tem-se:

Erros de Arredondamento e truncamento

- Se $|g_x| < \frac{1}{2}$ então tem-se:

Erro absoluto

$$|EA_x| = |x - \bar{x}| = |g_x| \times 10^{e-t} < \frac{1}{2} \times 10^{e-t}$$

Erros de Arredondamento e truncamento

- Se $|g_x| < \frac{1}{2}$ então tem-se:

Erro absoluto

$$|EA_x| = |x - \bar{x}| = |g_x| \times 10^{e-t} < \frac{1}{2} \times 10^{e-t}$$

Erro relativo

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} = \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} < \frac{0,5 \times 10^{e-t}}{0,1 \times 10^e} = \frac{1}{2} \times 10^{-t+1}$$

Erros de Arredondamento e truncamento

- ▶ Se $|g_x| \geq \frac{1}{2}$ então tem-se:

Erros de Arredondamento e truncamento

- ▶ Se $|g_x| \geq \frac{1}{2}$ então tem-se:

Erro absoluto

$$\begin{aligned}|EA_x| &= |x - \bar{x}| = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})| \\ &= |g_x \times 10^{e-t} - 10^{e-t}| = |g_x - 1| \times 10^{e-t} \leq \frac{1}{2} \times 10^{e-t}\end{aligned}$$

Erros de Arredondamento e truncamento

- ▶ Se $|g_x| \geq \frac{1}{2}$ então tem-se:

Erro absoluto

$$\begin{aligned}|EA_x| &= |x - \bar{x}| = |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})| \\ &= |g_x \times 10^{e-t} - 10^{e-t}| = |g_x - 1| \times 10^{e-t} \leq \frac{1}{2} \times 10^{e-t}\end{aligned}$$

Erro relativo

$$|ER_x| = \frac{|EA_x|}{|\bar{x}|} \leq \frac{\frac{1}{2} \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} < \frac{\frac{1}{2} \times 10^{e-t}}{|f_x \times 10^e|} = \frac{1}{2} \times 10^{-t+1}$$

Representação de ponto flutuante - base 10 e 2

Representação de ponto flutuante - base 10 e 2

- ▶ Computador fala palavra binária.
- ▶ Valores discretos e finitos.
- ▶ Por que falar do computador na matéria de métodos numéricos ?

Representação de ponto flutuante - base 10 e 2

- ▶ Computador fala palavra binária.
- ▶ Valores discretos e finitos.
- ▶ Por que falar do computador na matéria de métodos numéricos ?
 - O computador proporcionou outro nível de contas.

Representação de ponto flutuante - base 10 e 2

- ▶ Computador fala palavra binária.
- ▶ Valores discretos e finitos.
- ▶ Por que falar do computador na matéria de métodos numéricos ?
 - O computador proporcionou outro nível de contas.
 - Considerar os limites de representação.

Representação de ponto flutuante - base 10 e 2

- ▶ Computador fala palavra binária.
- ▶ Valores discretos e finitos.
- ▶ Por que falar do computador na matéria de métodos numéricos ?
 - O computador proporcionou outro nível de contas.
 - Considerar os limites de representação.

$$x = \underbrace{\sum_{i=0}^n a_i \beta^i}_{\text{parte inteira}} + \underbrace{\sum_{i=0}^{\infty} b_i \beta^{-i}}_{\text{parte fracionária}}$$

Representação de ponto flutuante - base 10 e 2

- Perguntas:

Representação de ponto flutuante - base 10 e 2

- Perguntas:

1. Por que chamar de ponto fixo ?

Representação de ponto flutuante - base 10 e 2

- Perguntas:

1. Por que chamar de ponto fixo ?
2. Por que chamar de ponto flutuante ?

Representação de ponto flutuante - base 10 e 2

► Perguntas:

1. Por que chamar de ponto fixo ?
2. Por que chamar de ponto flutuante ?
3. Como representar $(123.45)_{10}$ na base 10 e na base 2

Representação de ponto flutuante - base 10 e 2

► Perguntas:

1. Por que chamar de ponto fixo ?
2. Por que chamar de ponto flutuante ?
3. Como representar $(123.45)_{10}$ na base 10 e na base 2
4. O número $(0.11)_{10}$ pode ser representado na base 10 e também na base 2 ?

Representação de ponto flutuante - base 10 e 2

Como o computador trabalha com números $\mathbb{R}$?

$$x = \pm 0.d_1 d_2 d_3 \cdots d_t \times \beta^e$$

- ▶ β é a base;
- ▶ t é o número de dígitos da mantissa, com $d_1 \neq 0$ e $0 \leq d_j \leq \beta - 1$, para todo $j \in [1 : t]$; e,
- ▶ e é o expoente, $-m \leq e \leq M$.

Representação de ponto flutuante - base 10 e 2

Como o computador trabalha com números $\mathbb{R}$?

- ▶ Valores discretos.
- ▶ Representação do 0.
- ▶ Erro na representação dos números.

Representação de ponto flutuante - base 10 e 2

Como o computador trabalha com números $\mathbb{R}$?

- ▶ Valores discretos.
- ▶ Representação do 0.
- ▶ Erro na representação dos números. ←

Representação de ponto flutuante - base 10 e 2

Como o computador trabalha com números $\mathbb{R}$?

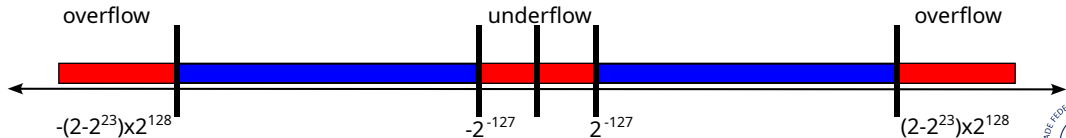
- ▶ Normalização dos valores.
- ▶ Conjunto de valores discretos, ainda que um valor de ponto flutuante.
- ▶ Padrão IEEE-754



Representação de ponto flutuante - base 10 e 2

Como o computador trabalha com números $\mathbb{R}$?

- ▶ valores negativos menores que $-(2 - 2^{23}) \times 2^{128}$ - *overflow*;
- ▶ valores maiores que -2^{-127} até zero - *underflow*;
- ▶ zero ←
- ▶ valores menores que 2^{-127} até zero - *underflow*;
- ▶ valores maiores que $(2 - 2^{23}) \times 2^{128}$ - *overflow*;



Aritmética de ponto flutuante

Aritmética de Ponto Flutuante

Definição: normalização

Um número real ($\mathbb{R}$) $x \neq 0$ é um ponto flutuante normalizado se pode ser escrito na forma:

$$x = \pm 0.d_1d_2d_3\cdots d_t \times \beta^e \quad (1)$$

onde:

- ▶ β é a base;
- ▶ t é o número de dígitos da mantissa, com $d_1 \neq 0$ e $0 \leq d_j \leq \beta - 1$, para todo $j \in [1 : t]$; e,
- ▶ e é o expoente, $-m \leq e \leq M$.

Aritmética de Ponto Flutuante

Definição: normalização

Denotamos por $F(\beta, t, m, M)$ o conjunto de todos os pontos flutuantes para β, t, m e M fixos com exceções do zero. Essa abordagem traz algumas limitações:

- ▶ valores muito grandes ou pequenos não conseguem ser representados;
- ▶ operações de multiplicação e divisão de grandes números pode gerar valores fracionados que não conseguem ser representados;

Aritmética de Ponto Flutuante

► **Exemplo:** Considere a máquina que opera no sistema: $F(\beta = 10, t = 3, m = -5, M = 5)$

- Os números são representados nesse sistema da seguinte forma:

- $0.d_1d_2d_3 \times 10^e$, $0 \leq d_j \leq 9$, $d_1 \neq 0$, $e \in [-5, 5]$;

- o menor número representado é: $m = 0.100 \times 10^{-5}$; e

- o maior número representado é: $M = 0.999 \times 10^5$.

- Considere os casos:

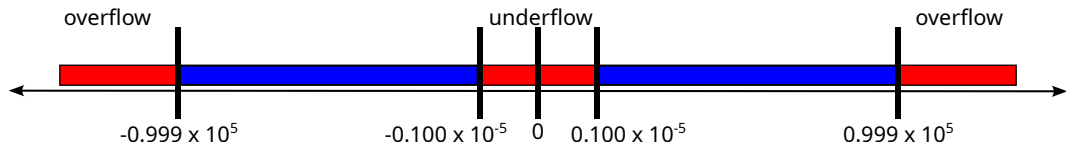
1. $x = 0.345 \times 10^{-7}$;

2. $x = 0.875 \times 10^9$; e

3. $x = 235.89$.

Aritmética de Ponto Flutuante

- ▶ $F(\beta = 10, t = 3, m = -5, M = 5)$



- ▶ valores negativos menores que -0.999×10^5 - *overflow*;
- ▶ valores maiores que -0.100×10^{-5} até zero - *underflow*;
- ▶ zero ←
- ▶ valores menores que 0.100×10^{-5} até zero - *underflow*;
- ▶ valores maiores que 0.999×10^5 - *overflow*;

Aritmética de Ponto Flutuante

- ▶ Operações Aritméticas em Ponto Flutuante:
 - **Adição e subtração:** Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número.
 - **Multiplicação e Divisão:** Realiza-se a operação nas mantissas e nos expoentes.
 - Os valores devem ser representados no sistema utilizado.
 - Os resultados devem ser truncados ou arredondados a cada operação.
 - As operações devem ser feitas com dois operados por vez.
 - O número deve ficar no formato normalizado a cada operação

Aritmética de Ponto Flutuante

- ▶ Problemas e efeitos:
 - **Cancelamento:** soma ou subtração de valores com ordens de grandeza muito diferentes.
 - **Propagação do erro:** sucessivas operações onde o erro é acumulado.
 - **Overflow:** multiplicação/divisão de grandes números positivo ou negativo.
 - **Underflow:** multiplicação/divisão de pequenos números positivo ou negativo.

Aritmética de Ponto Flutuante

► Exemplos:

1. Considerando um sistema de aritmética de ponto flutuante data por $F(\beta = 10, t = 4, m = -5, M = 5)$, Resolva:

■ $x = 0.1 \times 10^0$ e $y = 0.5 \times 10^4$, qual o resultado da operação $z = x + y$?

■ $x = 0.100 \times 10^{-5}$ e $y = 0.100 \times 10^{-5}$, qual o resultado da operação $z = x * y$?

■ $x = 0.999 \times 10^5$ e $y = 0.999 \times 10^5$, qual o resultado da operação $z = x * y$?

■ $S = \sum_{i=1}^4 (x_i + y_i)$, onde $x_i = 0.46709$ e $y_i = 3.5678$. Resolva considerando truncamento e arredondamento e compare com o resultado exato.

2. Implemente em os somatórios abaixo no computador usando a variável do tipo `float`, `double` e `long double` e imprima os respectivos resultados:

■ $S = \sum_{i=1}^{300000} x_i$ se $x_i = 0.5$

■ $S = \sum_{i=1}^{300000} x_i$ se $x_i = 0.11$

