

# MÓDULO 01 | ARITMÉTICA DE PONTO FLUTUANTE E ESTUDO SOBRE ERROS

Prof. Paulo F. C. Tilles



Departamento de Matemática

17 de Abril de 2022

## Representação de números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Representação de números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Representação de números

- ▶ Conjunto finito  $\rightarrow$  Conjunto discreto
- ▶ Nem todos os pontos de um intervalo  $[a, b]$  podem ser representados
- ▶ Implicação: resultados de operações simples geralmente contêm erros

### Representação de números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Interação usuário/computador

- ▶ Dados de entrada enviados pelo usuário na base decimal
- ▶ Conversão: operações efetuadas pelo computador na base binária
- ▶ Erros aproximação/arredondamento: números podem apresentar representação finita em uma base e não-finita em outra

## Representação de números inteiros

Representação de número inteiro  $n \neq 0$ : sequência de números inteiros  $n_i$

$$n = \pm (n_{-k} n_{-k+1} \dots n_{-1} n_0) = \pm (n_0 \beta^0 + n_{-1} \beta^1 \dots + n_{-k} \beta^k)$$

- ▶ Base fixa  $\beta$ : inteiro maior ou igual à 2 ( $\beta \in \mathbb{Z} : \beta \geq 2$ )
- ▶ Números inteiros  $n_i$  satisfazem as condições

$$\begin{array}{l|l} 0 \leq n_i < \beta & \text{Nenhum número pode ser maior que a base} \\ n_{-k} \neq 0 & \text{Primeiro número da sequência não nulo} \end{array}$$

**Exemplo:**  $\beta = 10$ ,  $n = 2019$

$$2019 = 9 \times 10^0 + 1 \times 10^1 + 0 \times 10^2 + 2 \times 10^3$$

Representação de  
números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Representação de números reais: ponto fixo

Representação de número real  $x \neq 0$ : sequência de números inteiros  $x_i$  separados por um ponto (.)

$$x = \pm (x_{-k} \dots x_{-1} x_0 . x_1 \dots x_n) = \pm \sum_{i=-k}^n x_i \beta^{-i}$$

►  $x_i$ : número inteiro satisfazendo  $0 \leq x_i < \beta$ .

**Exemplo:**  $\beta = 10$ ,  $x = 692.84$

$$\begin{aligned} 692.84 &= \sum_{i=-2}^2 x_i 10^{-i} = x_{-2} x_{-1} x_0 . x_1 x_2 \\ &= 6 \times 10^2 + 9 \times 10^1 + 2 \times 10^0 + 8 \times 10^{-1} + 4 \times 10^{-2} \end{aligned}$$

## Representação de números reais: ponto flutuante

Representação de número real  $x \neq 0$ :

$$x = \pm d \times \beta^e$$

- ▶ Mantissa  $d$ : número em ponto fixo

$$d = \sum_{i=1}^t d_i \beta^{-i}$$

- ▶ Expoente  $e$ : define limites de representação dos números no sistema;  $e \in [e_{\min}, e_{\max}]$ .
- ▶ Ponto flutuante normalizado:  $d_1 \neq 0$
- ▶ Representação:  $\beta^{-1} \leq d < 1$ .
- ▶  $t$ : número de dígitos significativos (precisão do sistema)

### Representação de números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Exemplos

1) Números em ponto flutuante normalizados ( $\beta = 10$ )

$$\begin{aligned}65.168 &= 0.65168 \times 10^2 \\ 0.000571 &= 0.571 \times 10^{-3}\end{aligned}$$

2) Limites de representação de números:

( $\beta = 10 | t = 3 | -2 \leq e \leq 2$ ).

$$\begin{aligned}\text{Maior número representável} &= 0.999 \times 10^2 \\ \text{Menor número representável} &= 0.1 \times 10^{-2}\end{aligned}$$

## Underflow

Alocação de um número menor que o menor número admitível no sistema de representação.

- ▶ Exemplo anterior:  $0.00003 = 0.3 \times 10^{-4}$
- ▶ Resultado: computador substitui o valor por 0.

## Overflow

Alocação de um número maior que o maior número admitível no sistema de representação.

- ▶ Exemplo anterior:  $100 = 0.1 \times 10^3$
- ▶ Resultado: computador trava o programa por não saber como proceder.

## Problema insolúvel

Determinar raiz da função  $f(x) = x^3 - 3$  com  $(\beta = 10 | t = 10)$

$$f(0.1442249570 \times 10^1) = -0.2 \times 10^{-8}$$

$$f(0.1442249571 \times 10^1) = 0.4 \times 10^{-8}$$

- ▶ Não existe número representável no sistema entre os dois, portanto o problema não tem solução
- ▶ Resultado: programa roda indefinidamente



## Bases

- ▶ Base decimal ( $\beta = 10$ ): base de uso diário/comum.
- ▶ Base duodecimal ( $\beta = 12$ ): descrição tempo (horas), comprimento em pés.
- ▶ Base binária ( $\beta = 2$ ): computadores.

## Fluxo informação: usuário $\rightleftarrows$ computador

- 1) Usuário passa informações para computador na base decimal.
- 2) Computador aloca informação convertida para a base binária.
- 3) Computador executa cálculos na base binária.
- 4) Computador converte informação alocada para base decimal e envia ao usuário.

## Mudança de bases | Exemplos

1) **1101**: da base binária para a base decimal.

- Multiplicação de cada algarismo na base binária por potências crescentes de 2

$$\begin{aligned}(1101)_2 &= 1 \times 2^0 + 0 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 \\ &= 1 + 0 + 4 + 8 = 13\end{aligned}$$

2) **0.110**: da base binária para a base decimal.

- Multiplicação de cada algarismo - após o ponto - na base binária por potências decrescentes de 2

$$\begin{aligned}(0.110)_2 &= 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} \\ &= \frac{1}{2} + \frac{1}{4} + 0 = \frac{3}{4} = 0.75\end{aligned}$$

3) **13**: da base decimal para a base binária.

- Divisões da parte inteira por 2: divisão dos quocientes até que o ultimo seja igual a 1

$$\begin{array}{r} 13 \quad | \quad 2 \\ 1 \quad | \quad 6 \quad | \quad 2 \\ \quad | \quad 0 \quad | \quad 3 \quad | \quad 2 \\ \quad \quad | \quad 1 \quad | \quad 1 \end{array}$$

- O número na base binária será o ultimo quociente e todos os restos das divisões anteriores (da direita para a esquerda).

$$13 = (1101)_2$$

4) **0.75**: da base decimal para a base binária.

- Multiplicações da parte decimal por 2: multiplicação da parte decimal restante

$$0.75 \times 2 = 1.5$$

$$0.50 \times 2 = 1.0$$

$$0.00 \times 2 = 0.0$$

- O número na base binária será a sequência de inteiros obtida em cada multiplicação.

$$0.75 = (0.110)_2$$

5) **3.8**: da base decimal para a base binária.

- ▶ Parte inteira:  $3 = (11)_2$
- ▶ Parte decimal:

$$0.8 \times 2 = 1.6$$

$$0.6 \times 2 = 1.2$$

$$0.2 \times 2 = 0.4$$

$$0.4 \times 2 = 0.8$$

$$0.8 \times 2 = 1.6$$

⋮

- ▶ O número na base decimal 3.8 não tem expressão exata na base binária

$$3.8 = (11.110011001100\dots)_2$$

## Erros de aproximação/comparativos

Hipótese: valor  $p^*$  é uma aproximação do número  $p$

### Erro real

$$E(p, p^*) = p - p^*$$

### Erro absoluto

$$EA(p, p^*) = |p - p^*|$$

### Erro relativo

$$ER(p, p^*) = \frac{|p - p^*|}{|p|}$$

- Utilizados como critérios de parada para métodos numéricos iterativos.

## Erro de arredondamento

- ▶ Simétrico: considera o número dentro da representação que mais se aproxima do valor real
- ▶ Fornece resultado mais acurado.

## Erro de truncamento

- ▶ Despreza a parte que extrapola o número de dígitos significativos
- ▶ Utilizado em computadores devido à velocidade superior na operação

Representação ( $\beta = 10 | t = 5$ ):

$$6584.691 \xrightarrow{\text{arred.}} 0.65847 \times 10^4$$

$$6584.691 \xrightarrow{\text{trunc.}} 0.65846 \times 10^4$$

## Erros nas operações

- ▶ Contribuição dos erros nas parcelas para o erro no resultado final da operação
- ▶ Ordem dos fatores altera o resultado: em aritmética de ponto flutuante com precisão finita não são válidas mais as propriedades associativas e distributivas

## Exemplos

Representação ( $\beta = 10 | t = 3$ ) com arredondamento

### 1) Soma de três números

$$(11.4 + 3.18) + 5.05 = 14.6 + 5.05 = 19.7$$

$$11.4 + (3.18 + 5.05) = 11.4 + 8.23 = 19.6$$



## 2) Soma de $1/3$ comparada com multiplicação por inteiro

$$0.333 + 0.333 = 0.666 \quad | \quad 2 \times 0.333 = 0.666$$

$$0.666 + 0.333 = 0.999 \quad | \quad 3 \times 0.333 = 0.999$$

$$0.999 + 0.333 = 1.33 \quad | \quad 4 \times 0.333 = 1.33$$

$$1.33 + 0.333 = 1.66 \quad | \quad 5 \times 0.333 = 1.67$$

$$1.66 + 0.333 = 1.99 \quad | \quad 6 \times 0.333 = 2.00$$

$$1.99 + 0.333 = 2.32 \quad | \quad 7 \times 0.333 = 2.33$$

$$2.32 + 0.333 = 2.65 \quad | \quad 8 \times 0.333 = 2.66$$

$$2.65 + 0.333 = 2.98 \quad | \quad 9 \times 0.333 = 3.00$$

$$2.98 + 0.333 = 3.31 \quad | \quad 10 \times 0.333 = 3.33$$

Representação de  
números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

# Números | Aritmética de ponto flutuante

3) Cálculo do polinômio  $P(x)$  no ponto  $x = 5.24$ , onde

$$P(x) = x^3 - 6x^2 + 4x - 0.1$$

Cálculo exato:

$$\begin{aligned} P(5.24) &= 143.8777824 - 164.7456 + 20.96 - 0.1 \\ &= -0.00776 \end{aligned}$$

Cálculo aproximado 1: 5 multiplicações + 3 somas

$$\begin{aligned} P(5.24) &= 5.24 \times 27.5 - 6 \times 27.5 + 4 \times 5.24 - 0.1 \\ &= 144 - 165 + 21.0 - 0.1 \\ &= -0.1 \quad (\text{soma da esquerda para a direita}) \\ &= 0 \quad (\text{soma da direita para a esquerda}) \\ ER_1 &\approx 13.89 \quad (1389\%) \\ ER_2 &= 1.0 \quad (100\%) \end{aligned}$$

# Números | Aritmética de ponto flutuante

## Cálculo aproximado 2: 2 multiplicações + 3 somas

$$P(x) = x(x(x - 6) + 4) - 0.1$$

$$\begin{aligned}P(5.24) &= 5.24(5.24(5.24 - 6) + 4) - 0.1 \\&= 5.24(-3.98 + 4) - 0.1 \\&= 0.105 - 0.1 = 0.005\end{aligned}$$

$$ER \approx 1.644 \quad (164.4\%)$$

## Comentários

- ▶ Quanto maior o número de operações aritméticas, maior é a tendência do erro se amplificar.
- ▶ Ponto central do cálculo numérico: desenvolver algoritmos capazes de minimizar os efeitos da aritmética discreta na execução de um grande número de operações.

## Definição

Erros comumente observados em cálculo numérico resultantes da implementação inadequada de algoritmos

- ▶ Cancelamento
- ▶ Propagação de erros
- ▶ Instabilidade Numérica
- ▶ Mal condicionamento

## Cancelamento

Perda de precisão numérica quando dois números muito próximos são subtraídos.

Representação de  
números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

**Exemplo:** cálculo de  $\sqrt{9876} - \sqrt{9875}$

$$\begin{aligned}\sqrt{9876} &= 0.9937806599 \times 10^2 \\ \sqrt{9875} &= 0.9937303457 \times 10^2 \\ \sqrt{9876} - \sqrt{9875} &= 0.0000503142 \times 10^2 \\ &= 0.5031420000 \times 10^{-2}\end{aligned}$$

**Minimização do efeito:** reescrever a operação de subtração de outra maneira

$$\begin{aligned}\sqrt{x} - \sqrt{y} &= \frac{x - y}{\sqrt{x} + \sqrt{y}} \\ \sqrt{9876} - \sqrt{9875} &= \frac{1}{\sqrt{9876} + \sqrt{9875}} \\ &= 0.5031418679 \times 10^{-2}\end{aligned}$$

## Propagação de erros

Efeito de cancelamento presente em somas quando uma soma parcial é muito grande (em valor absoluto) quando comparada com o resultado final.

- ▶ Só ocorre quando existe inversão de sinal nos termos da soma.
- ▶ Dado um conjunto  $\{a_k\}$  de números reais, o cálculo da soma total  $s_n = \sum_{k=1}^n a_k$  é realizado a partir das somas parciais:

$$s_1 = a_1, \quad s_k = s_{k-1} + a_k, \quad k = 2, \dots, n.$$

**Exemplo:** Cálculo de  $e^{-5.25}$  através da expansão em série

$$e^{-x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!}$$

# Números | Efeitos numéricos

Cálculo: representação ( $\beta = 10 | t = 5$ ) com arredondamento

Ordem $k$	Termo $(-1)^k x^k / k!$	Soma parcial $s_k$
0	$+0.10000 \times 10^1$	$+0.10000 \times 10^1$
1	$-0.52500 \times 10^1$	$-0.42500 \times 10^1$
2	$+0.13781 \times 10^2$	$+0.95310 \times 10^1$
3	$-0.24117 \times 10^2$	$-0.14586 \times 10^2$
4	$+0.31654 \times 10^2$	$+0.17068 \times 10^2$
5	$-0.33236 \times 10^2$	$-0.16168 \times 10^2$
6	$+0.29082 \times 10^2$	$+0.12914 \times 10^2$
7	$-0.21811 \times 10^2$	$-0.88970 \times 10^1$
8	$+0.14314 \times 10^2$	$+0.54170 \times 10^1$
9	$-0.83497 \times 10^1$	$-0.29327 \times 10^1$
10	$+0.43836 \times 10^1$	$+0.14509 \times 10^1$
11	$-0.20922 \times 10^1$	$-0.64130 \times 10^0$
12	$+0.91532 \times 10^0$	$+0.27402 \times 10^0$
13	$-0.36965 \times 10^0$	$-0.95630 \times 10^{-1}$

# Números | Efeitos numéricos

Ordem $k$	Termo $(-1)^k x^k/k!$	Soma parcial $s_k$
14	$+0.13862 \times 10^0$	$+0.42990 \times 10^{-1}$
15	$-0.48517 \times 10^{-1}$	$-0.55270 \times 10^{-2}$
16	$+0.15919 \times 10^{-1}$	$+0.10392 \times 10^{-1}$
17	$-0.49163 \times 10^{-2}$	$+0.54757 \times 10^{-2}$
18	$+0.14339 \times 10^{-2}$	$+0.69096 \times 10^{-2}$
19	$-0.39622 \times 10^{-3}$	$+0.65134 \times 10^{-2}$
20	$+0.10401 \times 10^{-3}$	$+0.66174 \times 10^{-2}$
21	$-0.26002 \times 10^{-4}$	$+0.65914 \times 10^{-2}$
22	$+0.62049 \times 10^{-5}$	$+0.66038 \times 10^{-2}$
23	$-0.14163 \times 10^{-5}$	$+0.66024 \times 10^{-2}$
24	$+0.30983 \times 10^{-6}$	$+0.66027 \times 10^{-2}$
25	$-0.65063 \times 10^{-7}$	$+0.66026 \times 10^{-2}$

Valor correto:  $e^{-5.25} = 0.52475 \times 10^{-2}$  ( $ER \approx 0.258$ )



# Números | Efeitos numéricos

- ▶ A aproximação de 5 dígitos significativos dos termos de ordem mais altas (azul) excluem contribuições da mesma ordem de grandeza do resultado final

$$k = 3 \quad | \quad -0.14586 \times 10^2 = -1458.60000 \times 10^{-2}$$

$$k = 4 \quad | \quad +0.17068 \times 10^2 = +1706.80000 \times 10^{-2}$$

$$k = 5 \quad | \quad -0.16168 \times 10^2 = -1616.80000 \times 10^{-2}$$

$$k = 6 \quad | \quad +0.12914 \times 10^2 = +1291.40000 \times 10^{-2}$$

---

$$k = 23 \quad | \quad +0.66026 \times 10^{-2} = -0000.66026 \times 10^{-2}$$

Solução para o problema: cálculo de  $e^{-5.25} = 1/e^{5.25}$  através da expansão em série

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

# Números | Efeitos numéricos

$k$	$x^k/k!$	$s_k$	$s_k^{-1}$
0	$0.10000 \times 10^1$	$0.10000 \times 10^1$	$0.10000 \times 10^1$
1	$0.52500 \times 10^1$	$0.62500 \times 10^1$	$0.16000 \times 10^0$
2	$0.13781 \times 10^2$	$0.20031 \times 10^2$	$0.49923 \times 10^{-1}$
3	$0.24117 \times 10^2$	$0.44148 \times 10^2$	$0.22651 \times 10^{-1}$
4	$0.31654 \times 10^2$	$0.75802 \times 10^2$	$0.13192 \times 10^{-1}$
5	$0.33236 \times 10^2$	$0.10904 \times 10^3$	$0.91709 \times 10^{-2}$
6	$0.29082 \times 10^2$	$0.13812 \times 10^3$	$0.72401 \times 10^{-2}$
7	$0.21811 \times 10^2$	$0.15993 \times 10^3$	$0.62527 \times 10^{-2}$
8	$0.14314 \times 10^2$	$0.17424 \times 10^3$	$0.57392 \times 10^{-2}$
9	$0.83497 \times 10^1$	$0.18259 \times 10^3$	$0.54768 \times 10^{-2}$
10	$0.43836 \times 10^1$	$0.18697 \times 10^3$	$0.53485 \times 10^{-2}$
11	$0.20922 \times 10^1$	$0.18906 \times 10^3$	$0.52893 \times 10^{-2}$
12	$0.91532 \times 10^0$	$0.18998 \times 10^3$	$0.52637 \times 10^{-2}$
13	$0.36965 \times 10^0$	$0.19035 \times 10^3$	$0.52535 \times 10^{-2}$

$k$	$x^k/k!$	$s_k$	$s_k^{-1}$
14	$0.13862 \times 10^0$	$0.19049 \times 10^3$	$0.52496 \times 10^{-2}$
15	$0.48517 \times 10^{-1}$	$0.19054 \times 10^3$	$0.52482 \times 10^{-2}$
16	$0.15919 \times 10^{-1}$	$0.19056 \times 10^3$	$0.52477 \times 10^{-2}$

Valor correto:  $e^{-5.25} = 0.52475 \times 10^{-2}$  ( $ER \approx 0.381 \times 10^{-4}$ )

## Erros em cálculo numérico

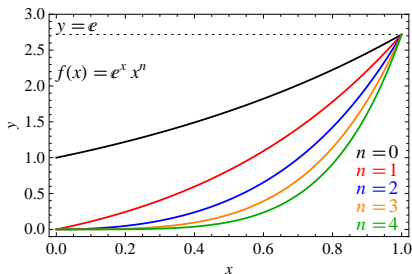
- ▶ Todo cálculo intermediário introduz erros de arredondamento/truncamento que afetam resultados subsequentes, independentemente do nível de precisão utilizado.
- ▶ Ponto central é como os erros influenciam o resultado final, i.e., se e como eles se somam ou se cancelam.
- ▶ **Algoritmo estável:** erros intermediários têm efeito desprezível no resultado final

## Instabilidade numérica

Efeito numérico em que os erros intermediários de um algoritmo afetam consideravelmente o resultado final.

**Exemplo:** cálculo iterativo da integral

$$I_n = \frac{1}{e} \int_0^1 dx e^x x^n$$



Método recursivo:

$$I_0 = 1 - \frac{1}{e} \approx 0.63212056$$

$$I_n = 1 - nI_{n-1}, \quad n = 1, 2, \dots$$

$$(I_{n+1} < I_n)$$

Cálculo numérico:  $t = 8$

$$I_0 = 0.63212056 \ (ER \sim 10^{-9}) \mid I_7 = 0.11237760 \ (ER \sim 10^{-5})$$

$$I_1 = 0.36787944 \ (ER \sim 10^{-9}) \mid I_8 = 0.10097920 \ (ER \sim 10^{-4})$$

$$I_2 = 0.26424112 \ (ER \sim 10^{-9}) \mid I_9 = 0.09118720 \ (ER \approx 0.005)$$

$$I_3 = 0.20727664 \ (ER \sim 10^{-8}) \mid I_{10} = 0.0881280 \ (ER \approx 0.051)$$

$$I_4 = 0.17089344 \ (ER \sim 10^{-7}) \mid I_{11} = 0.030591998 \ (ER \approx 0.6)$$

$$I_5 = 0.14553280 \ (ER \sim 10^{-7}) \mid I_{12} = 0.63289603 \ (ER \approx 7.8)$$

$$I_6 = 0.12680320 \ (ER \sim 10^{-6}) \mid I_{13} = -7.2276483 \ (ER \sim 10^2)$$

- Origem do erro: instabilidade numérica do algoritmo recursivo utilizado.

## Demonstração da instabilidade

- ▶  $I_n$ : valor exato da integral
- ▶  $\tilde{I}_n$ : valor calculado assumindo erro propagado de  $I_0$
- ▶  $\varepsilon_0$ : erro no valor de  $I_0$

$$\tilde{I}_0 = I_0 + \varepsilon_0$$

$$\tilde{I}_n = 1 - n \tilde{I}_{n-1}, \quad n = 1, 2, \dots$$

- ▶  $\varepsilon_n$ : erro cometido no cálculo da  $n$ -ésima integral

$$\varepsilon_n \equiv \tilde{I}_n - I_n = -n \varepsilon_{n-1}, \quad n = 1, 2, \dots$$

- ▶ Cálculo recursivo

$$\varepsilon_1 = -\varepsilon_0 = (-1) \varepsilon_0$$

$$\varepsilon_2 = -2\varepsilon_1 = 2\varepsilon_0$$

$$\varepsilon_3 = -3\varepsilon_2 = (-1) 3! \varepsilon_0$$

$$\varepsilon_4 = -4\varepsilon_3 = 4! \varepsilon_0 \quad \rightarrow \quad \varepsilon_n = (-1)^n n! \varepsilon_0$$

## Algoritmo estável

- ▶ Estabilidade na direção decrescente de  $I_n$

$$I_{n-1} = \frac{1 - I_n}{n}$$

- ▶ Como  $I_n \rightarrow 0$  conforme  $n \rightarrow \infty$ , basta assumir  $I_{\bar{n}} = 0$  para algum  $\bar{n}$  suficientemente grande

$$\tilde{I}_{\bar{n}} = I_{\bar{n}} + \varepsilon$$

$$\tilde{I}_{\bar{n}-1} = \frac{1 - I_{\bar{n}}}{\bar{n}} - \frac{\varepsilon}{\bar{n}}$$

$$\tilde{I}_{\bar{n}-2} = \frac{\bar{n}(1 + I_{\bar{n}}) - 1}{\bar{n}(\bar{n} - 1)} + \frac{\varepsilon}{\bar{n}(\bar{n} - 1)}$$

$$\tilde{I}_{\bar{n}-3} = \frac{\bar{n}^2 - \bar{n}(2 + I_{\bar{n}}) + 1}{\bar{n}(\bar{n} - 1)(\bar{n} - 2)} - \frac{\varepsilon}{\bar{n}(\bar{n} - 1)(\bar{n} - 2)}$$

$$r_{\bar{n}-n} = (-1)^n \frac{(\bar{n} - n)!}{\bar{n}!} \varepsilon$$

► Erro decrescente:  $r_n \rightarrow 0$  conforme  $n \rightarrow 0$

Cálculo numérico:  $t = 8$

$$I_{20} = 0.000000000 \quad (ER = 1)$$

$$I_{19} = 0.500000000 \quad (ER \approx 0.048)$$

$$I_{18} = 0.500000000 \quad (ER \approx 0.002)$$

$$I_{17} = 0.052777778 \quad (ER \sim 10^{-4})$$

$$I_{16} = 0.055718954 \quad (ER \sim 10^{-6})$$

$$I_{15} = 0.059017565 \quad (ER \sim 10^{-7})$$

⋮

$$I_0 = 0.63212056 \quad (ER \sim 10^{-9})$$



## Mal condicionamento

Uso do cálculo numérico na resolução de problemas segue três passos:

1. Entrada de dados: define o problema matemático e os seus parâmetros.
2. Processamento de dados: algoritmo realiza os cálculos.
3. Saída de dados: resultado obtido pelo algoritmo.

**Problema bem posto:** resultado depende continuamente dos dados, onde pequenas variações nos parâmetros geram pequenas variações no resultado.

**Problema mal posto:** resultado não depende continuamente dos dados.

- ▶ Criticalidade: pequenas alterações mudam qualitativamente o resultado final.

Representação de  
números

Introdução

Sistema de números discretos

Mudança de base

Erros

Aritmética de ponto flutuante

Efeitos numéricos

## Exemplo: Sistemas lineares

### Sistema linear original

$$x + y = 2, \quad x + 1.01y = 2.01$$

- ▶ Solução:  $x = y = 1$ .

### Sistema linear alterado 1

$$x + y = 2, \quad x + 1.01y = 2.02$$

- ▶ Solução:  $(x, y) = (0, 2)$ .

### Sistema linear alterado 2

$$x + y = 2, \quad x + 1.0y = 2.01$$

- ▶ Não existe solução: retas paralelas

**Exemplo:** Problema de valor inicial

$$\ddot{y} = y, \quad y(0) = 1, \quad \dot{y}(0) = 2\delta - 1$$

Solução:

$$y(t) = \delta e^t + (1 - \delta) e^{-t}$$

Limite 1:  $|\delta| > 0$

$$\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} \delta e^t = \begin{cases} +\infty, & \delta > 0, \\ -\infty, & \delta < 0. \end{cases}$$

Limite 2:  $\delta = 0$

$$\lim_{t \rightarrow \infty} y(t) = \lim_{t \rightarrow \infty} e^{-t} = 0$$