- LIMITE DE UMA SEQÜÊNCIA DE REAIS OU COMPLEXOS

* DEFINIÇÃO: SEJA

$$\{x_n\}_{n=1}^{\infty}$$

UMA SEQÜÊNCIA INFINITA DE REAIS OU COMPLEXOS.

ESTA SEQÜÊNCIA TEM O LIMITE x (i.e., CONVERGE PARA x), SE, PARA QUALQUER $\varepsilon > 0$, EXISTE UM NÚMERO POSITIVO INTEIRO $N(\varepsilon)$, TAL QUE $|x_n - x| < \varepsilon$, SEMPRE QUE $n > N(\varepsilon)$.

* NOTAÇÃO: $\lim_{n\to\infty} x_n = x$ OU $x_n \to x$ QUANDO $n \to \infty$

- CONVERGÊNCIA E CONTINUIDADE

* TEOREMA: SE f É DEFINIDA EM $X \subset \mathbb{R}$ E $x_0 \in X$, ENTÃO AS SEGUINTES AFIRMAÇÕES SÃO EQUIVALENTES:

A) f É CONTÍNUA EM x_0

B) SE $\{x_n\}_{n=1}^{\infty}$ É QUALQUER SEQÜÊNCIA EM X CONVERGENTE PARA x_0, ENTÃO

$$\lim_{n\to\infty} f(x_n) = f(x_0)$$

OBSERVAÇÃO: ASSUMIREMOS QUE TODAS AS FUNÇÕES CONSIDERADAS EM ANÁLISE NUMÉRICA SERÃO CONTÍNUAS.

– DERIVADA

* DEFINIÇÃO: SEJA f UMA FUNÇÃO DEFINIDA NUM INTERVALO ABERTO CONTENDO x_0.

f É **DIFERENCIÁVEL** EM x_0, SE EXISTE

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

$f'(x_0)$ É CHAMADA DE **DERIVADA** DE f EM x_0.

f É DITA DIFERENCIÁVEL EM $X \subset \mathbb{R}$ SE TEM DERIVADA PARA CADA $x \in X$.

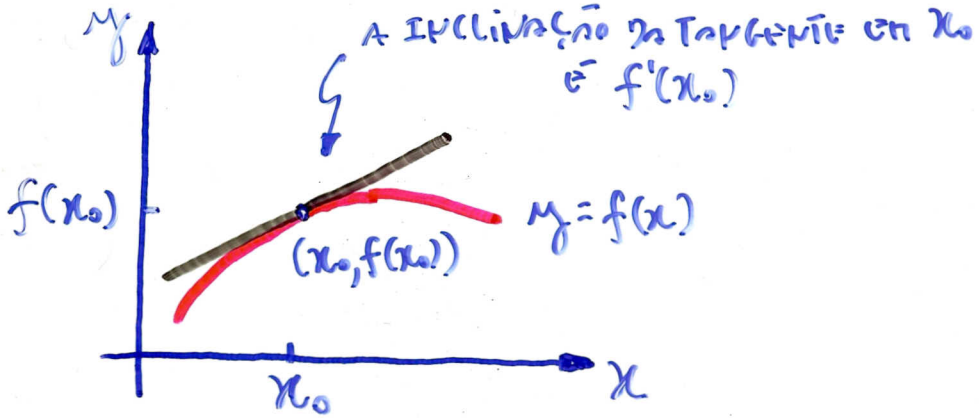
* TEOREMA: SE f É DIFERENCIÁVEL EM x_0, ENTÃO f É CONTÍNUA EM x_0.

* NOTAÇÃO: $C^n(X) \equiv$ CONJUNTO DE TODAS AS FUNÇÕES QUE TÊM n DERIVADAS CONTÍNUAS EM $X \subset \mathbb{R}$

$C^\infty(X) \equiv$ CONJUNTO DAS FUNÇÕES QUE TÊM DERIVADAS DE TODAS AS ORDENS EM $X \subset \mathbb{R}$

- TEOREMAS DE ROLLE, VALOR MÉDIO E VALOR EXTREMO

OBSERVAÇÃO: A DERIVADA DE f EM x_0 É O VALOR DA INCLINAÇÃO DA LINHA TANGENTE À CURVA REPRESENTATIVA DE f, NO PONTO $(x_0, f(x_0))$

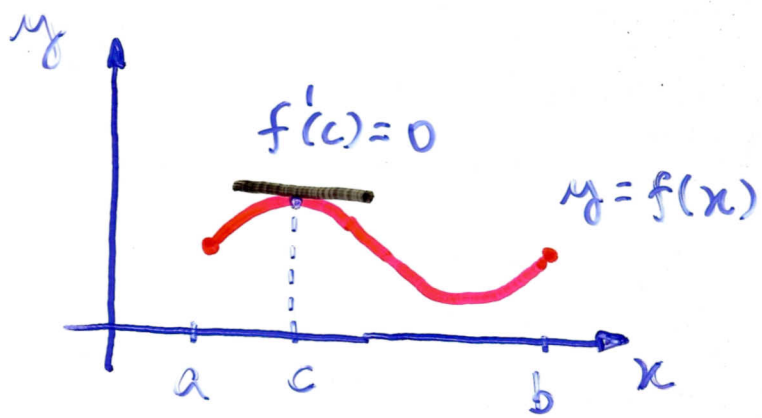


* TEOREMA DE ROLLE:

SEJA $f \in C[a,b]$ E DIFERENCIÁVEL EM (a,b).

SE $f(a) = f(b)$, EXISTE UM NÚMERO c EM (a,b) PARA O QUAL $f'(c) = 0$.

OBSERVAÇÃO: PODE EXISTIR MAIS DE UM PONTO COM TAL PROPRIEDADE

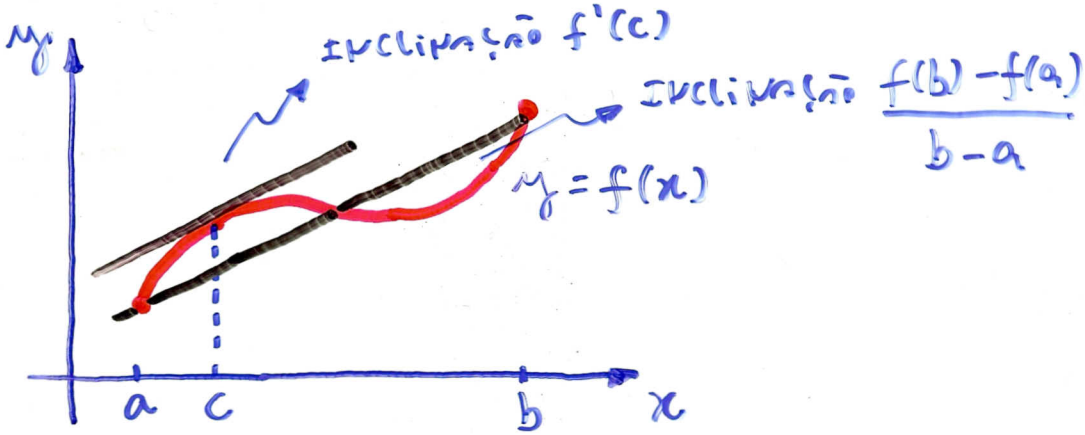


GENERALIZAÇÃO: SE $f \in C[a,b]$ E f É DIFERENCIÁVEL n VEZES EM (a,b), E ADEMAIS SE $f(x) = 0$ NOS $n+1$ VALORES DISTINTOS $x_0, \ldots, x_n$ EM $[a,b]$, ENTÃO EXISTE c EM (a,b) TAL QUE $f^{(n)}(c) = 0$

* TEOREMA DO VALOR MÉDIO (DERIVADAS):

SE $f \in C[a,b]$ E f É DIFERENCIÁVEL EM (a,b), EXISTE UM NÚMERO c NO INTERVALO (a,b) TAL QUE

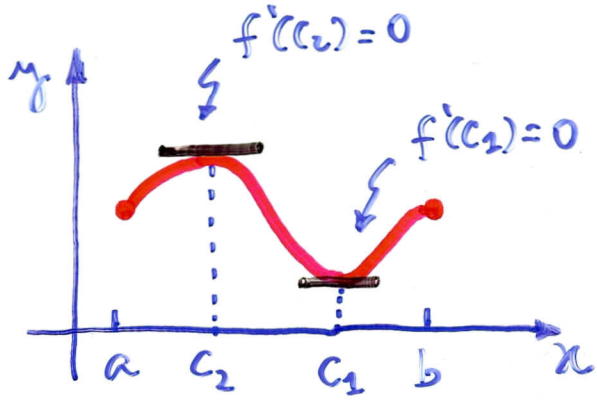
$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



* TEOREMA DO VALOR EXTREMO:

SE $f \in C[a,b]$, ENTÃO EXISTEM c_1 E c_2 EM $[a,b]$ TAIS QUE $f(c_1) \leq f(x) \leq f(c_2)$ PARA TODO $x \in [a,b]$.

ADEMAIS, SE f É DIFERENCIÁVEL EM (a,b), OS NÚMEROS c_1 E c_2 PODEM EXISTIR TANTO NOS EXTREMOS DE $[a,b]$ COMO ONDE $f' = 0$.



– INTEGRAIS

* DEFINIÇÃO: **INTEGRAL DE RIEMANN**

A INTEGRAL DE RIEMANN DE f EM $[a,b]$ É O SEGUINTE LIMITE, DESDE QUE ELE EXISTA:

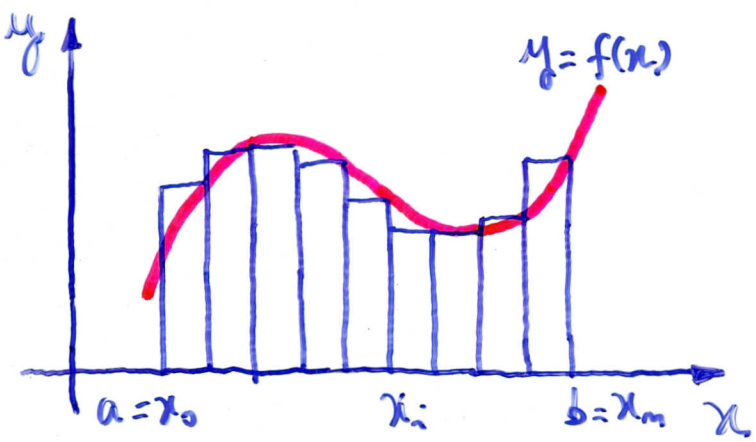
$$\int_a^b f(x)dx = \lim_{\max \Delta x_i \to 0} \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

ONDE $a = x_0 \leq x_1 \leq \dots \leq x_n = b$, E ONDE $\Delta x_i = x_i - x_{i-1}$, PARA CADA $i = 1, 2, \dots, n$, COM ξ_i ESCOLHIDO ARBITRARIAMENTE EM $[x_{i-1}, x_i]$

OBSERVAÇÕES: i) TODA FUNÇÃO CONTÍNUA EM $[a,b]$ ADMITE INTEGRAL DE RIEMANN NESSE INTERVALO.

ii) PARA MAIOR CONVENIÊNCIA COMPUTACIONAL, PODEMOS ESCOLHER OS PONTOS x_i ESPAÇADOS EM **INTERVALOS IGUAIS** DE $[a,b]$, E ESCOLHER $\xi_i = x_i$ PARA CADA $i = 1, 2, \dots, n$

NESSE CASO, $$\int_a^b f(x)dx = \lim_{n \to \infty} \frac{b-a}{n} \sum_{i=1}^{n} f(x_i)$$



$$x_i = a + \left(\frac{b-a}{n}\right) i$$

$$\therefore \Delta x_i = \frac{b-a}{n}$$

$$\lim_{\max \Delta x_i \to 0} \Delta x_i = \lim_{n \to \infty} \left(\frac{b-a}{n}\right)$$

- Generalização do Teorema do Valor Médio (Integrais)

* Teorema do Valor Médio Ponderado para Integrais:

Supondo que exista $f \in C[a,b]$, que exista a integral de Riemann de g em $[a,b]$, e que $g(x)$ não troque de sinal em $[a,b]$, então existe um número c em (a,b) tal que

$$\int_a^b f(x)g(x)\,dx = f(c)\int_a^b g(x)\,dx$$

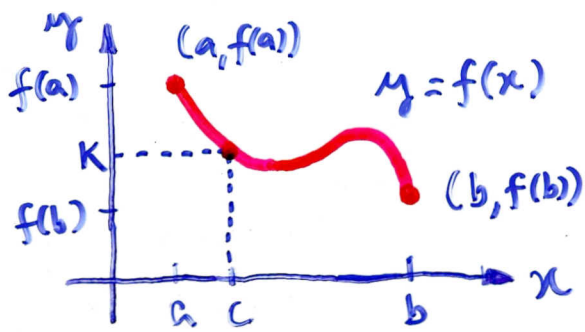
Observação: Quando $g(x)=1$, este teorema se torna o tradicional Teorema do Valor Médio para Integrais, que dá o valor médio de f no intervalo $[a,b]$ como sendo

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx$$

(Neste caso, c é o ponto – indeterminado – tal que o valor da função ali é o seu valor médio no intervalo $[a,b]$)

* Teorema do Valor Intermediário:

Se $f \in C[a,b]$ e K é qualquer número entre $f(a)$ e $f(b)$, existe um número c em (a,b) para o qual $f(c)=K$.



- Polinômios de Taylor

* TEOREMA DE TAYLOR:

Supondo que $f \in C^n[a,b]$, que $f^{(n+1)}$ existe em $[a,b]$ e que $x_0 \in [a,b]$, existe um número $\xi(x)$ entre x_0 e x, para todo $x \in [a,b]$, tal que

$$f(x) = P_n(x) + R_n(x)$$

onde

$$P_n(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

$$= \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

e

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}$$

$P_n(x)$ é chamado **Polinômio de Taylor** de n-ésimo grau para $f(x)$, centrado em x_0

$R_n(x)$ é chamado **resto** ou **erro de truncamento** associado a $P_n(x)$

A série infinita obtida tomando-se $\lim_{n\to\infty} P_n(x)$ é chamada **série de Taylor** para $f(x)$, centrada em x_0

Quando $x_0 = 0$, o polinômio e a série de Taylor também são conhecidos como de **Maclaurin**

2. ERROS DE ARREDONDAMENTO E ARITMÉTICA COMPUTACIONAL

- ARITMÉTICA COMPUTACIONAL

  ⇒ ARITMÉTICA DE DÍGITOS FINITOS

  i.e., CADA NÚMERO REPRESENTÁVEL TEM APENAS UM NÚMERO FIXO E FINITO DE DÍGITOS

  ⇒ APENAS UM SUBCONJUNTO DOS NÚMEROS RACIONAIS PODE SER REPRESENTADO EXATAMENTE

  ⇒ O CÁLCULO COM OS DEMAIS NÚMEROS LEVA A ERROS DE ARREDONDAMENTO

- REPRESENTAÇÃO DE NÚMEROS

  ⇒ DEPENDE DA BASE DISPONÍVEL NA MÁQUINA EM USO, E DO NÚMERO MÁXIMO DE DÍGITOS USADO NA REPRESENTAÇÃO

  COMPUTADORES OPERAM NORMALMENTE EM BASE BINÁRIA

  ⇒ DADOS DE ENTRADA SÃO ENVIADOS AO COMPUTADOR EM SISTEMA DECIMAL E CONVERTIDOS AO SISTEMA BINÁRIO.

  AS OPERAÇÕES SÃO EFETUADAS NO SISTEMA BINÁRIO E OS RESULTADOS SÃO CONVERTIDOS AO SISTEMA DECIMAL PARA TRANSMISSÃO AO USUÁRIO

- Conversão entre os Sistemas Binário e Decimal

* Conversão de Números Inteiros

Representação numa Base β:

Número $N = (a_j a_{j-1} \ldots a_2 a_1 a_0)_\beta$

$$= a_j\beta^j + a_{j-1}\beta^{j-1} + \ldots + a_2\beta^2 + a_1\beta^1 + a_0\beta^0$$

onde $0 \leq a_k \leq \beta - 1$, $k = 0, \ldots, j$

e.g., $(10111)_2 = (1\times 2^4) + (0\times 2^3) + (1\times 2^2) + (1\times 2^1) + (1\times 2^0)$

$(347)_{10} = (3\times 10^2) + (4\times 10^1) + (7\times 10^0)$

1. Conversão do sistema Binário para o Decimal

e.g., $(10111)_2 = 1\times 2^4 + 0\times 2^3 + 1\times 2^2 + 1\times 2^1 + 1\times 2^0$

$= (1\times 2^1) + (1\times 2^4) + (0\times 2^3) + (1\times 2^4) + 1$

Colocando 2 em evidência:

$(10111)_2 = 2\times(1 + 2\times(1 + 2\times(0 + 2\times 1))) + 1$

$= 22 + 1 = 23$

$\therefore (10111)_2 = (23)_{10}$

No caso geral: Para obter a representação decimal, denotaremos por b_0, do número binário $(a_j a_{j-1} \cdots a_2 a_1 a_0)_2$, fazemos

$$b_j = a_j$$

$$b_{j-1} = a_{j-1} + 2b_j$$

$$b_{j-2} = a_{j-2} + 2b_{j-1}$$

$$\vdots$$

$$b_1 = a_1 + 2b_2$$

$$b_0 = a_0 + 2b_1$$

No exemplo: Para $(10111)_2$

$$b_4 = a_4 = 1$$

$$b_3 = a_3 + 2b_4 = 0 + 2 \times 1 = 2$$

$$b_2 = a_2 + 2b_3 = 1 + 2 \times 2 = 5$$

$$b_1 = a_1 + 2b_2 = 1 + 2 \times 5 = 11$$

$$b_0 = a_0 + 2b_1 = 1 + 2 \times 11 = 23$$

2. Conversão do Sistema Decimal para o Binário

e.g., $N_0 = 347$

Seja $(a_j a_{j-1} \ldots a_2 a_0)_2$ a representação binária procurada

$$N_0 = 347 = a_j 2^j + a_{j-1} 2^{j-1} + \ldots + a_2 2^2 + a_1 2 + a_0$$

$$\therefore 2 \times 173 + 1 = 2 \times (a_j 2^{j-1} + a_{j-1} 2^{j-2} + \ldots + a_2 2 + a_1) + a_0$$

$$\therefore a_0 = 1$$

Repetindo o processo para $N_1 = 173$:

$$N_1 = 173 = a_j 2^{j-1} + a_{j-1} 2^{j-2} + \ldots + a_2 2 + a_1$$

$$\therefore 2 \times 86 + 1 = 2 \times (a_j 2^{j-2} + a_{j-1} 2^{j-3} + \ldots + a_2) + a_1$$

$$\therefore a_1 = 1$$

Continuando, obtemos

$N_2 = 86 = 2 \times 43 + 0 \Rightarrow a_2 = 0$

$N_3 = 43 = 2 \times 21 + 1 \Rightarrow a_3 = 1$

$N_4 = 21 = 2 \times 10 + 1 \Rightarrow a_4 = 1$

$N_5 = 10 = 2 \times 5 + 0 \Rightarrow a_5 = 0$

$N_6 = 5 = 2 \times 2 + 1 \Rightarrow a_6 = 1$

$N_7 = 2 = 2 \times 1 + 0 \Rightarrow a_7 = 0$

$N_8 = 1 = 2 \times 0 + 1 \Rightarrow a_8 = 1$

$$\therefore 347 = (101011011)_2$$

No caso geral: Para obter a representação binária, denotada por $(a_j a_{j-1} \cdots a_0)_2$, do número decimal N, implementamos o seguinte algoritmo, que obtém, a cada k, o dígito binário a_k.

Passo 0: $k = 0$

$N_k = N$

Passo 1: Obtenha q_k e r_k tais que

$N_k = 2 \times q_k + r_k$

Faça $a_k = r_k$

Passo 2: Se $q_k = 0$, pare

senão, faça $N_{k+1} = q_k$

Faça $k = k + 1$

e volte ao Passo 1

* Conversão de Números Fracionários

1. Conversão do Sistema Decimal para o Binário

Um número real entre 0 e 1 pode ter uma representação finita no sistema decimal e infinita no binário

⇒ Usaremos a representação binária

$$(0.d_1d_2....)_2 = 0\times 2^0 + d_1\times 2^{-1} + d_2\times 2^{-2} + ...$$

e.g., $N=(0.125)_{10}$

Multiplicando os dois lados por 2:

$$2\times 0.125 = 0.25 = 0 + 0.25 = d_1 + d_2 2^{-1} + ...$$

⇒ d_1 representa a parte inteira de 2×0.125, que é zero.

As demais parcelas representam a parte fracionária, que é 0.25

Repetindo o processo para 0.25:

$$0.5 = d_2 + d_3 2^{-2} + d_4 2^{-2} + ... \Rightarrow d_2 = 0$$

Da mesma forma, para 0.5:

$$1.0 = d_3 + d_4 2^{-2} + d_5 2^{-2} + ... \Rightarrow d_3 = 1$$

Como a parte fracionária é zero, o processo pára e $(0.125)_{10} = (0.001)_2$

No caso geral: Para obter a representação binária, denotada por $(0.d_1d_2\ldots d_j\ldots)_2$, do número decimal π entre 0 e 1, implementamos o algoritmo seguinte que obtém, a cada k, o dígito binário d_k.

Passo 0: $k = 1$

$\pi_1 = \pi$

Passo 1: Calcule $2\pi_k$

Se $2\pi_k \geq 1$, faça $d_k = 1$

Senão, faça $d_k = 0$

Passo 2: Faça $\pi_{k+1} = 2\pi_k - d_k$

Se $\pi_{k+1} = 0$, pare.

Senão, faça $k = k+1$

e volte ao passo 1

Observações:

i) O algoritmo pode não parar após um número finito de passos, quando o número tem representação binária <u>infinita</u>.

ii) O fato de alguns números terem representação binária infinita pode acarretar erros aparentemente inexplicáveis nos cálculos efetuados em computadores.

e.g., Para $\pi = (0.1)_{10}$

$k=1$ $\pi_1 = 0.1$, $2\pi_1 = 0.2 < 1 \therefore d_1 = 0$, $\pi_2 = 0.2$

$k=2$ $2\pi_2 = 0.4 < 1 \therefore d_2 = 0$, $\pi_3 = 0.4$

$k=3$ $2\pi_3 = 0.8 < 1 \therefore d_3 = 0$, $\pi_4 = 0.8$

$k=4$ $2\pi_4 = 1.6 > 1 \therefore d_4 = 1$, $\pi_5 = 0.6$

$k=5$ $2\pi_5 = 1.2 > 1 \therefore d_5 = 1$, $\pi_6 = 0.2 = \pi_2$

$\Rightarrow$ os resultados para k de 2 a 5 se repetem

i.e., $\pi_2 = \pi_6 = \pi_{10} = \ldots$

$\therefore (0.1)_{10} = (0.0001100\overline{1100}\ldots)_2$

onde o traço indica a parte que se repete

$\Rightarrow$ o número 0.1 tem representação infinita na base 2.

2. Conversão do sistema Binário para o Decimal

⇒ Seguimos um algoritmo semelhante ao anterior

Seja a representação decimal

$$(0.b_2b_2\ldots b_j)_{10} = b_2 10^{-2} + b_2 10^{-2} + \ldots + b_j 10^{-j}$$

e.g., $\Pi = (0.000111)_2 = b_2 10^{-2} + b_2 10^{-2} + \ldots + b_j 10^{-j}$

Multiplicando os dois lados por 10 [na base Binária, i.e., $(1010)_2$]:

$$(0.000111)_2 \times (1010)_2 = (1.11 \times 2^{-1})_2 \times (1.01 \times 2^3)_2 =$$

$$= (1.11 \times 1.01)_2 \times 2^{-2} = (1.00011)_2 = b_2 + b_2 10^{-2} + \ldots + b_j 10^{-j+1}$$

⇒ b_2 representa a parte inteira do resultado,

$\underline{b_2 = 1}$

Repetindo o processo para a parte Fracionária, $\Pi_2 = (0.00011)_2$:

$$(0.00011)_2 \times (1010)_2 = (1.1 \times 1.01) \times 2^{-2} = (0.1111)_2 =$$

$$= b_2 + b_3 10^{-2} + \ldots + b_j 10^{-j+2}$$

⇒ $\underline{b_2 = 0}$ e $\Pi_3 = 0.1111$

Continuando o processo, chegamos a

$$(0.000111)_2 = (0.109375)_{10}$$

No caso geral: Para obter a representação decimal, denotada por $(0.b_1 b_2 \ldots b_j)_{10}$. Do número binário R entre 0 e 1, implementamos o algoritmo seguinte, que obtém, a cada k, o dígito decimal b_k.

Passo 0: $k = 1$
$R_1 = R$

Passo 1: Calcule $W_k = (1010)_2 \times R_k$
Seja Z_k a parte inteira de W_k
b_k é a conversão de Z_k para a base 10

Passo 2: Faça $R_{k+1} = W_k - Z_k$
Se $R_{k+1} = 0$, Pare.
Senão, Faça $k = k+1$
e Volte ao Passo 1.

- DIFICULDADES DA COMPUTAÇÃO COM PRECISÃO FINITA

* REPRESENTAÇÃO EM PONTO FLUTUANTE:

⇒ UM NÚMERO REAL É REPRESENTADO COMO

$$\pm (0.d_1 d_2 \ldots d_k) \times \beta^{\ell}$$

ONDE: $\beta \equiv$ BASE EM QUE A MÁQUINA OPERA

$k \equiv$ NÚMERO DE DÍGITOS NA MANTISSA OU PARTE FRACIONÁRIA

$\ell \equiv$ EXPOENTE INTEIRO

$0 \leq d_i \leq (\beta - 1)$, PARA $i = 1, \ldots, k$

COM $d_1 \neq 0$

e.g., REPRESENTAÇÃO DECIMAL EM PONTO FLUTUANTE:

$$\pm (0.d_1 d_2 \ldots d_k) \times 10^n \quad 0 \leq d_i \leq 9, \; d_1 \neq 0$$

⇒ DADO UM NÚMERO REAL POSITIVO NO FORMATO

$$y = 0.d_1 d_2 \ldots d_k d_{k+1} \ldots \times 10^n$$ (DECIMAL NORMALIZADO)

A SUA REPRESENTAÇÃO EM PONTO FLUTUANTE É OBTIDA LIMITANDO-SE A MANTISSA DE y A k DÍGITOS DECIMAIS.

POSSIBILIDADES: i) TRUNCAMENTO

ii) ARREDONDAMENTO

i) TRUNCAMENTO

⇒ CORTAMOS OS DÍGITOS $d_{k+1} d_{k+2} \ldots$, PARA OBTER

$$fl(y) = 0.d_1 d_2 \ldots d_k \times 10^n$$

ii) ARREDONDAMENTO

⇒ ADICIONAMOS $5 \times 10^{n-(k+1)}$ A y, E TRUNCAMOS O RESULTADO, PARA OBTER

$$fl(y) = 0.\delta_1 \delta_2 \ldots \delta_k \times 10^n$$

⇒ ADICIONAMOS 1 A d_k, SE $d_{k+1} \geq 5$ (ARREDONDAMOS PARA CIMA)

SE NÃO, CORTAMOS TODOS OS DÍGITOS SEGUINTES AOS k PRIMEIROS (ARREDONDAMOS PARA BAIXO)

e.g., $\pi = 3.14159265\ldots$ TEM INFINITOS DÍGITOS

NOTAÇÃO DECIMAL NORMALIZADA: $\pi = 0.314159265\ldots \times 10^1$

NOTAÇÃO EM PONTO FLUTUANTE COM 5 DÍGITOS:

- TRUNCAMENTO: $fl(\pi) = 0.31415 \times 10^1 = 3,1415$

- ARREDONDAMENTO:

$$\pi + 5 \times 10^{1-(5+1)} = \pi + 5 \times 10^{-5} =$$

$$= 3,14159265\ldots + 0.00005 =$$

$$= 3.14164265$$

$$\therefore fl(\pi) = 0.31416 \times 10^1 = 3,1416$$

- ERROS DE ARREDONDAMENTO E TRUNCAMENTO

≡ ERROS RESULTANTES DA SUBSTITUIÇÃO DO NÚMERO REAL POR SUA NOTAÇÃO EM PONTO FLUTUANTE≡

* DEFINIÇÃO: ERRO ABSOLUTO E ERRO RELATIVO

SE p^* É UMA APROXIMAÇÃO PARA p,

O ERRO ABSOLUTO É $|p - p^*|$

O ERRO RELATIVO É $\frac{|p - p^*|}{|p|}$, COM $p \neq 0$

OU $\frac{|p - p^*|}{|p^*|}$, COM $p^* \neq 0$

OBSERVAÇÃO: COMO UMA MEDIDA DE PRECISÃO, O ERRO RELATIVO É MAIS SIGNIFICATIVO, POIS LEVA EM CONTA A MAGNITUDE DOS VALORES

* DEFINIÇÃO: ALGARISMOS SIGNIFICATIVOS

O NÚMERO p^* SE APROXIMA DO VALOR p COM t ALGARISMOS SIGNIFICATIVOS, SE t É O MAIOR VALOR INTEIRO NÃO NEGATIVO PARA O QUAL

$$\frac{|p - p^*|}{|p|} < 5 \times 10^{-t}$$

e.g., $p = 0{,}5$, $p^* = 0{,}5002$ $\Rightarrow$ $\frac{|p - p^*|}{|p|} = 0.0004 < 5 \times 10^{-4}$

$\Rightarrow$ $t = 4$ É O MAIOR INTEIRO NÃO-NEGATIVO PARA O QUAL $|p - p^*|/|p| < 5 \times 10^{-t}$

$\Rightarrow$ p^* APROXIMA p COM 4 ALGARISMOS SIGNIFICATIVOS

– LIMITES PARA OS ERROS DE ARREDONDAMENTO E TRUNCAMENTO

SEJA y UM NÚMERO REAL A SER EXPRESSO EM NOTAÇÃO DE PONTO FLUTUANTE DECIMAL COM k DÍGITOS.

EXPRESSEMOS y COMO

$$y = f \times 10^{n} + g \times 10^{n-k}$$

ONDE $\quad 0.1 \leq f < 1$

$$0 \leq g < 1$$

e.g., $y = 234.57$ COM $k = 4$ DÍGITOS:

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

$$\therefore f = 0.2345 \text{ E } g = 0.7$$

$\Rightarrow$ O FATOR $g \times 10^{n-k}$ REPRESENTA A PARTE DE y QUE NÃO PODE SER INCORPORADA À MANTISSA COM k DÍGITOS

$\Rightarrow$ O ERRO COMETIDO DEPENDE DA APROXIMAÇÃO UTILIZADA: POR TRUNCAMENTO OU ARREDONDAMENTO.

i) TRUNCAMENTO

$\Rightarrow$ O FATOR $g \times 10^{n-k}$ É DESPREZADO E $y^{*} = f \times 10^{n}$

ERRO ABSOLUTO: $E_A = |y - y^{*}| = |g| \times 10^{n-k} < 10^{n-k}$

Pois $|g| < 1$

ERRO RELATIVO: $E_R = \dfrac{|E_A|}{|y^{*}|} = \dfrac{|g| \times 10^{n-k}}{|f| \times 10^{n}} < \dfrac{10^{n-k}}{0.1 \times 10^{n}} = 10^{-k+1}$

Pois $f \geq 0.1$

ii) ARREDONDAMENTO

⇒ O FATOR $g \times 10^{m-k}$ É LEVADO EM CONTA, MODIFICANDO $f \times 10^m$

$$y^* = \begin{cases} f \times 10^m, & \text{SE } |g| < 1/2 \\ f \times 10^m + 10^{m-k}, & \text{SE } |g| \geq 1/2 \end{cases}$$

* SE $|g| < 1/2$:

ERRO ABSOLUTO: $\varepsilon_A = |y - y^*| = |g| \times 10^{m-k} < \frac{1}{2} \times 10^{m-k}$

ERRO RELATIVO: $\varepsilon_R = \frac{|\varepsilon_A|}{|y^*|} = \frac{|g| \times 10^{m-k}}{|f| \times 10^m} < \frac{0.5 \times 10^{m-k}}{0.1 \times 10^m} = \frac{1}{2} \times 10^{-k+1}$

* SE $|g| \geq 1/2$:

ERRO ABSOLUTO: $\varepsilon_A = |y - y^*| = |(f \times 10^m + g \times 10^{m-k}) - (f \times 10^m + 10^{m-k})|$

$= |g \times 10^{m-k} - 10^{m-k}| = |(g-1)| \times 10^{m-k} \leq$

$\leq \frac{1}{2} \times 10^{m-k}$

ERRO RELATIVO: $\varepsilon_R = \frac{|\varepsilon_A|}{|y^*|} \leq \frac{1/2 \times 10^{m-k}}{|f \times 10^m + 10^{m-k}|} < \frac{1/2 \times 10^{m-k}}{|f| \times 10^m} <$

$< \frac{1/2 \times 10^{m-k}}{0.1 \times 10^m} = \frac{1}{2} \times 10^{-k+1}$

⇒ EM AMBOS OS CASOS, $|\varepsilon_A| \leq \frac{1}{2} \times 10^{m-k}$ E $|\varepsilon_R| < \frac{1}{2} \times 10^{-k+1}$

⇒ O ARREDONDAMENTO ACARRETA ERROS MENORES, AO CUSTO DE UM TEMPO MAIOR DE EXECUÇÃO.

- ARITMÉTICA COMPUTACIONAL

OPERAÇÕES: Dados x e y reais

Adição: $x \oplus y = fl(fl(x) + fl(y))$

Subtração: $x \ominus y = fl(fl(x) - fl(y))$

Multiplicação: $x \otimes y = fl(fl(x) \times fl(y))$

Divisão: $x \oslash y = fl(fl(x) \div fl(y))$

$\Rightarrow$ OPERAÇÕES ARITMÉTICAS EXATAS com os valores de x e y em PONTO FLUTUANTE e CONVERSÃO do RESULTADO EXATO para PONTO FLUTUANTE

OBSERVAÇÕES: i) Numa SEQÜÊNCIA de OPERAÇÕES, o ERRO em uma delas se PROPAGA ao longo das OPERAÇÕES SUBSEQÜENTES.

ii) O ERRO TOTAL em uma OPERAÇÃO é COMPOSTO pelo ERRO nas PARCELAS ou FATORES e pelo ERRO no RESULTADO da OPERAÇÃO.

- ERROS ABSOLUTO e RELATIVO nas OPERAÇÕES ARITMÉTICAS:

Sejamos Parcelas ou Fatores com erro

e.g., $x = x^* + E_A(x)$

$y = y^* + E_A(y)$

Suponha que o erro final é ARREDONDADO ou TRUNCADO, mas NÃO CONSIDERAMOS esse passo nas fórmulas a seguir.

OBS.: ERROS com SINAL!

i) ADIÇÃO:

$$x + y = (x^* + \varepsilon_A(x)) + (y^* + \varepsilon_A(y)) =$$

$$= (x^* + y^*) + (\varepsilon_A(x) + \varepsilon_A(y))$$

$\Rightarrow$ ERRO ABSOLUTO: $\varepsilon_A(x+y) = \varepsilon_A(x) + \varepsilon_A(y)$

ERRO RELATIVO: $\varepsilon_R(x+y) = \dfrac{\varepsilon_A(x+y)}{x^* + y^*} = \dfrac{\varepsilon_A(x) + \varepsilon_A(y)}{x^* + y^*}$

$$= \frac{\varepsilon_A(x)}{x^*}\left(\frac{x^*}{x^* + y^*}\right) + \frac{\varepsilon_A(y)}{y^*}\left(\frac{y^*}{x^* + y^*}\right) =$$

$$= \varepsilon_R(x)\left(\frac{x^*}{x^* + y^*}\right) + \varepsilon_R(y)\left(\frac{y^*}{x^* + y^*}\right)$$

ii) SUBTRAÇÃO:

ANALOGAMENTE, OBTEMOS

ERRO ABSOLUTO: $\varepsilon_A(x-y) = \varepsilon_A(x) - \varepsilon_A(y)$

ERRO RELATIVO:

$$\varepsilon_R(x-y) = \varepsilon_R(x)\left(\frac{x^*}{x^* - y^*}\right) - \varepsilon_R(y)\left(\frac{y^*}{x^* - y^*}\right)$$

iii) MULTIPLICAÇÃO:

$$xy = (x^* + \varepsilon_A(x))(y^* + \varepsilon_A(y)) =$$

$$= x^*y^* + x^*\varepsilon_A(y) + y^*\varepsilon_A(x) + \varepsilon_A(x)\varepsilon_A(y)$$

Considerando que $\varepsilon_A(x)\varepsilon_A(y)$ seja muito pequeno, podemos desprezá-lo na expressão acima

Assim,

ERRO ABSOLUTO: $\varepsilon_A(xy) \simeq x^*\varepsilon_A(y) + y^*\varepsilon_A(x)$

ERRO RELATIVO: $\varepsilon_R(xy) \simeq \dfrac{x^*\varepsilon_A(y) + y^*\varepsilon_A(x)}{x^*y^*} =$

$$= \frac{\varepsilon_A(x)}{x^*} + \frac{\varepsilon_A(y)}{y^*} = \varepsilon_R(x) + \varepsilon_R(y)$$

iv) DIVISÃO:

$$\frac{x}{y} = \frac{x^* + \varepsilon_A(x)}{y^* + \varepsilon_A(y)} = \frac{x^* + \varepsilon_A(x)}{y^*}\left(\frac{1}{1 + \dfrac{\varepsilon_A(y)}{y^*}}\right) \simeq$$

$$\simeq \left(\frac{x^* + \varepsilon_A(x)}{y^*}\right)\left(1 - \frac{\varepsilon_A(y)}{y^*}\right) \simeq \frac{x^*}{y^*} + \frac{\varepsilon_A(x)}{y^*} - \frac{x^*\varepsilon_A(y)}{y^{*2}}$$

Assim,

ERRO ABSOLUTO: $\varepsilon_A(x/y) \simeq \dfrac{\varepsilon_A(x)}{y^*} - \dfrac{x^*\varepsilon_A(y)}{y^{*2}} = \dfrac{y^*\varepsilon_A(x) - x^*\varepsilon_A(y)}{y^{*2}}$

ERRO RELATIVO: $\varepsilon_R(x/y) \simeq \left(\dfrac{y^*\varepsilon_A(x) - x^*\varepsilon_A(y)}{y^{*2}}\right)\dfrac{y^*}{x^*} =$

$$= \frac{\varepsilon_A(x)}{x^*} - \frac{\varepsilon_A(y)}{y^*} = \varepsilon_R(x) - \varepsilon_R(y)$$

OBSERVAÇÃO: Dos resultados acima, podemos concluir que: i) A subtração de dois números aproximadamente iguais pode levar a um grande aumento no erro relativo.

ii) Multiplicação por um número de grande magnitude, ou divisão por um número de pequena magnitude, pode acarretar um grande aumento no erro absoluto.

- ARTIFÍCIOS PARA MINIMIZAR ERROS NA ARITMÉTICA COMPUTACIONAL

i) EVITAR A SUBTRAÇÃO DE VALORES PRÓXIMOS

e.g., equação do 2º grau: $ax^2+bx+c=0$

Raízes: $x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$, $x_2 = \frac{-b-\sqrt{b^2-4ac}}{2a}$

$\Rightarrow$ Quando $b^2 \gg 4ac$, x_1 envolve subtração de valores muito próximos

$\Rightarrow$ Fórmula alternativa: Racionalizando o numerador

$$x_1 = \frac{-b+\sqrt{b^2-4ac}}{2a} \times \left(\frac{-b-\sqrt{b^2-4ac}}{-b-\sqrt{b^2-4ac}}\right) = \frac{b^2-(b^2-4ac)}{2a(-b-\sqrt{b^2-4ac})}$$

$$= -2c/b+\sqrt{b^2-4ac}$$

ii) Reagrupar termos, no cálculo de polinómios

e.g., $f(x) = x^3-6.1x^2+3.2x+1.5$ (4 produtos, 3 adições)

$= [(x-6.1)x+3.2]x+1.5$ (2 produtos, 3 adições)

Menos operações $\Rightarrow$ menor propagação de erros

- PADRÃO DE PONTO FLUTUANTE BINÁRIO DO IEEE (FORMA 754-1985)

⇒ REPRESENTAÇÃO DE 64 BITS PARA NÚMERO REAL EXTENSO

1 BIT INDICADOR DE SINAL (+ OU −): s

PARTE EXPONENCIAL DE 11 BITS (CARACTERÍSTICA): c

PARTE FRACIONÁRIA DE 52 BITS (MANTISSA): f

⇒ NÚMEROS EM PONTO FLUTUANTE DA FORMA

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

e.g., NÚMERO DE MÁQUINA: 0 10000000011 1011100100 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000

s | c | f

S = 0 ⇒ NÚMERO POSITIVO

$$c = 10000000011 = 1.2^{10} + 0.2^9 + \ldots + 0.2^2 + 1.2^1 + 1.2^0$$

$$= 1027$$

⇒ PARTE EXPONENCIAL: $2^{1027-1023} = 2^4$

$$f = 1.\left(\frac{1}{2}\right)^1 + 1.\left(\frac{1}{2}\right)^3 + 1\left(\frac{1}{2}\right)^4 + 1.\left(\frac{1}{2}\right)^5 + 1.\left(\frac{1}{2}\right)^8 + 1.\left(\frac{1}{2}\right)^{12}$$

⇒ MANTISSA: 0.72230039

⇒ O NÚMERO DE MÁQUINA REPRESENTA O DECIMAL

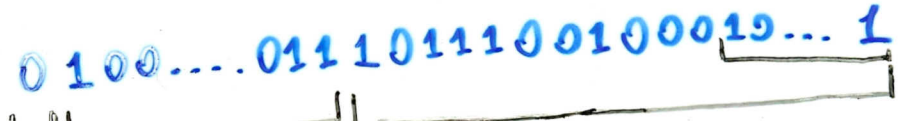
$$(-1)^s 2^{c-1023}(1+f) = (-1)^0 . 2^4 . (1.72230039)$$

$$= 27.56640625$$

NÚMERO DADO: 0 100....011 1011100100 100...0

Nº IMEDIATAMENTE INFERIOR: 0 100....011 1011100100 011...1

Nº IMEDIATAMENTE SUPERIOR: 0 100....011 1011100100 10... 1



⇒ O NÚMERO DE MÁQUINA ORIGINAL REPRESENTA METADE DOS NÚMEROS REAIS QUE ESTÃO ENTRE ELE E OS SEUS DOIS VIZINHOS MAIS PRÓXIMOS EM LINGUAGEM DE MÁQUINA

⇒ ELE REPRESENTA UM INTERVALO

- UNDERFLOW E OVERFLOW

* O MENOR NÚMERO POSITIVO QUE SE PODE REPRESENTAR NO PADRÃO IEEE TEM TODOS OS DÍGITOS ZERO, EXCETO O ÚLTIMO À DIREITA, QUE É 1

$\Rightarrow 2^{-1023} \cdot (1 + 2^{-52}) \simeq 10^{-308}$

* O MAIOR NÚMERO POSITIVO TEM UM ZERO INICIAL SEGUIDO POR 63 DÍGITOS 1

$\Rightarrow 2^{1024} \cdot (2 - 2^{-52}) \simeq 10^{308}$

Se nos cálculos computacionais aparecem números menores do que $\simeq 10^{-308}$, estes são ajustados para zero (UNDERFLOW).

Se aparecem números maiores do que $\simeq 10^{308}$, se caracteriza OVERFLOW, o que geralmente leva a falhas no processamento.

3. ALGORITMOS E CONVERGÊNCIA

* **ALGORITMO** ≡ PROCEDIMENTO QUE DESCREVE, SEM AMBIGÜIDADES, UMA SEQÜÊNCIA FINITA DE PASSOS A SEREM EXECUTADOS, NUMA ORDEM ESPECIFICADA, COM O OBJETIVO DE RESOLVER UM PROBLEMA OU OBTER UMA SOLUÇÃO APROXIMADA PARA ELE.

DESCRIÇÃO DOS ALGORITMOS: PSEUDOCÓDIGO

* **PSEUDOCÓDIGO** ⇒ ESPECIFICA A FORMA DOS DADOS DE ENTRADA E DOS RESULTADOS A SEREM PRODUZIDOS
INCORPORA UMA PRESCRIÇÃO INDEPENDENTE PARA A INTERRUPÇÃO DOS PROCEDIMENTOS

CONVENÇÕES: i) PONTO (.): INDICA O FIM DE UM PASSO

ii) PONTO-E-VÍRGULA (;): SEPARA TAREFAS EM UM MESMO PASSO

iii) PARÁGRAFO: INDICA GRUPO DE INSTRUÇÕES A SEREM TRATADAS COMO UMA ÚNICA ENTIDADE

iv) CÁLCULOS EM LOOP:
SÃO CONTROLADOS POR CONTAGEM:
e.g., PARA $i = 1, 2, \ldots, n$
FAÇA ....
OU POR CONDICIONAIS: e.g., ENQUANTO $i < n$

v) EXECUÇÕES CONDICIONAIS: SE ... ENTÃO ....
SENÃO ...

vi) COMENTÁRIOS: SÃO COLOCADOS ENTRE PARÊNTESES

EXEMPLO: CALCULAR $\sum_{i=1}^{N} x_i = x_1 + x_2 + \ldots + x_N$

QUANDO $x_1, x_2, \ldots, x_N$ SÃO DADOS

ALGORITMO:

ENTRADA $N, x_1, x_2, \ldots, x_N$

SAÍDA $SUM = \sum_{i=1}^{N} x_i$

Passo 1 Faça $SUM = 0$. (INICIALIZA O ACUMULADOR)

Passo 2 PARA $i = 1, 2, \ldots, N$

FAÇA $SUM = SUM + x_i$ (ADICIONA O TERMO SEGUINTE)

Passo 3 SAÍDA (SUM);

PARE.

- ESTABILIDADE DOS ALGORITMOS

* ALGORITMO ESTÁVEL => PEQUENAS ALTERAÇÕES NOS DADOS INICIAIS PRODUZEM PEQUENAS ALTERAÇÕES NO RESULTADO FINAL

* ALGORITMO CONDICIONALMENTE ESTÁVEL
  => ESTÁVEL APENAS PARA CERTOS VALORES DOS DADOS INICIAIS

A ESTABILIDADE DE UM ALGORITMO ESTÁ RELACIONADA AO CRESCIMENTO DO ERRO DE ARREDONDAMENTO

* DEFINIÇÃO: SEJA $E_0 > 0$ UM ERRO DE ARREDONDAMENTO INICIAL.
  SEJA E_n A MAGNITUDE DO ERRO APÓS n OPERAÇÕES SUCESSIVAS.
  SE $E_n \simeq C n . E_0$, ONDE C É UMA CONSTANTE INDEPENDENTE DE n, O CRESCIMENTO DO ERRO É DITO LINEAR.
  SE $E_n \simeq C^n E_0$ PARA ALGUM $C > 1$, O CRESCIMENTO DO ERRO É DITO EXPONENCIAL.

ALGORITMOS COM CRESCIMENTO LINEAR DO ERRO SÃO ESTÁVEIS.

ALGORITMOS COM CRESCIMENTO EXPONENCIAL DO ERRO SÃO INSTÁVEIS.

– CONVERGÊNCIA DOS ALGORITMOS

* DEFINIÇÃO: Sejam $\{\beta_n\}_{n=1}^{\infty}$ uma sequência convergente para 0.

Sejam $\{\alpha_n\}_{n=1}^{\infty}$ uma sequência convergente para o número α.

Se existe uma constante positiva K tal que

$$|\alpha_n - \alpha| \leq K|\beta_n|$$

para um n grande, então dizemos que $\{\alpha_n\}_{n=1}^{\infty}$ converge para α com taxa de convergência $O(\beta_n)$, i.e., $\alpha_n = \alpha + O(\beta_n)$

OBSERVAÇÃO: Em geral, utilizamos a sequência $\beta_n = \frac{1}{n^p}$, para um $p > 0$ qualquer, e estamos interessados no maior valor de p para o qual $\alpha_n = \alpha + O\left(\frac{1}{n^p}\right)$

e.g., se $\alpha_n = \frac{n+1}{n^2}$ e $\hat{\alpha}_n = \frac{n+3}{n^3}$

temos

$$\alpha_n = 0 + O\left(\frac{1}{n}\right)$$

$$\hat{\alpha}_n = 0 + O\left(\frac{1}{n^2}\right)$$

De maneira semelhante, podemos definir a **Taxa de Convergência de uma Função**:

* Definição: Sejam $\lim_{h \to 0} G(h) = 0$ e $\lim_{h \to 0} F(h) = L$

Se existe uma constante $K > 0$ para a qual

$$|F(h) - L| \leq K|G(h)|$$

para h suficientemente pequeno, então escrevemos

$$F(h) = L + O(G(h))$$

Observação: As funções utilizadas para comparação geralmente têm a forma $G(h) = h^p$, com $p > 0$, e estamos interessados no maior valor p para o qual $F(h) = L + O(h^p)$