CAPÍTULO 1

MATEMÁTICA COMPUTACIONAL

1.1. INTRODUÇÃO

O primeiro passo do aprendizado para a realização da análise numérica de qualquer problema matemático é a construção de uma base conceitual fundamental suficientemente sólida para capacitar o analista a formular, modelar e resolver numericamente um problema de Engenharia. O escopo do presente tratamento é o de apresentar métodos numéricos para solução de problemas matemáticos, com um enfoque direcionado para a aplicação em problemas práticos de engenharia, tais como a simulação e otimização de sistemas físicos e processos industriais. Dentro deste enfoque, vários assuntos foram selecionados do ensino básico de Matemática para serem resumidamente apresentados, discutidos e exemplificados dentro de possíveis aplicações de Engenharia.

O presente capítulo se inicia com alguns tópicos importantes do Cálculo Diferencial e Integral, Álgebra Linear e Análise Real que constituem a base necessária ao aluno para o entendimento dos capítulos subseqüentes. A seguir, apresenta-se de forma sumária o sistema de funcionamento de um computador típico e como as operações matemáticas são realizadas computacionalmente. A apresentação está dividida em itens que revisam a teoria de cada assunto abordado, com exemplos ilustrativos de sua possível aplicação prática no campo das Engenharias. A parte final do capítulo contém uma lista de problemas propostos e também projetos abertos propostos ao estudante visando a orientação para a formação do Engenheiro, essencialmente, um projetista.

1.2. LIMITES, CONTINUIDADE E DIFERENCIAÇÃO

Seja f uma função de uma variável real, então, o limite da função f em c, se existir, é definido como se segue:

$$\lim_{x \to c} f(x) = L \tag{1.1}$$

onde c e L são constantes. Se existir c tal que não haja nenhum número L com esta propriedade, portanto, o limite de f em c não existe. Por exemplo:

$$\lim_{x \to 1} \ln(x) = 0 \tag{1.2}$$

No entanto, o $\lim_{x \to -1} \ln(x)$ não existe, uma vez que para $x < 0$, $\ln(x)$ não é uma função real.

Continuidade é definida a partir da definição de limite enunciada anteriormente. A função f é contínua no intervalo $[a, b]$, se e somente se, para todo $c \in [a, b]$, com $a, b \in \Re$

$$\lim_{x \to c} f(x) = f(c) \tag{1.3}$$

Em conseqüência, a função $f(x) = \ln(x)$ é contínua para $x > 0$.

Por definição, uma função $f(x)$ é diferenciável em c, caso exista a derivada, definida pela equação

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c} \tag{1.4}$$

Esta definição permite também concluir que se f é diferenciável em c, então f tem que ser contínua em c.

O conjunto de funções contínuas ao longo de todo o domínio dos números reais é representado por C(R). Analogamente, o conjunto de funções em que f' é contínua ao longo de todo o domínio dos números reais é representado por $C^1(R)$. Desta maneira, a conclusão do parágrafo anterior pode ser escrita como $C^1(R) \subset C^0(R)$. O mesmo raciocínio pode ser aplicado em seqüência para derivações sucessivas de f, se existirem. Assim, define-se o conjunto genérico $C^n(R)$ para cada número natural n, representando a ordem de derivação de f, onde $f^{(n)}$ é contínua ao longo de todo o domínio dos números reais. Finalmente, define-se $C^\infty(R)$, como o conjunto de funções que têm todas as suas derivadas contínuas, i.e., $C^\infty(R) \subset \cdots \subset C^2(R) \subset C^1(R) \subset C^0(R)$. Um exemplo bastante conhecido de função desta classe é $f(x) = e^x$.

A definição acima pode ser aplicada de forma semelhante para um subconjunto dos números reais, definindo $C^n[a,b]$ como o conjunto de funções f onde exista $f^{(n)}$ e que seja contínua no intervalo $[a,b] \subset R$. O entendimento de funções deste tipo é fundamental para o entendimento de um importante teorema do Cálculo, i.e., o Teorema de Taylor, a ser estudado mais a frente neste capítulo.

1.3. NORMAS E ANÁLISE DE ERROS

A motivação principal para o estudo de Normas para o engenheiro é a necessidade de se avaliar o comportamento de sistemas que têm muitos graus de liberdade. As grandezas de interesse em um problema de engenharia (e.g., parâmetros geométricos e de operação) podem ser combinadas na forma de um ou mais vetores com muitas componentes. O comportamento de um ou mais desses vetores como um todo é um indicativo claro da resposta do sistema de engenharia em análise. Uma Norma é uma função que pode ser interpretada como o comprimento ou magnitude de cada um desses vetores.

Por exemplo, ao se buscar a solução de um problema por um método numérico, é necessário verificar-se, durante o desenvolvimento da busca da solução, se o procedimento está se aproximando ou se afastando da solução e, por fim, o momento em que a solução é encontrada. Neste ponto, diz-se que houve convergência. Esta verificação torna-se possível a partir da utilização do conceito de Norma. No problema em análise, define-se um ou mais vetores cujas componentes representem os erros ou resíduos em cada um dos graus de liberdade do problema, portanto, a cada passo nos cálculos, calcula-se o vetor erro ou residual e monitora-se o comportamento da norma desse vetor, estabelecendo-se um critério de parada para os cálculos no momento em que o valor da norma seja suficientemente pequeno de acordo com um valor pré-estabelecido $\varepsilon \cong 0$.

Uma norma pode ser definida em qualquer espaço vetorial. Uma norma é uma função $\|\cdot\|$ de $V \in R^+$ que obedeça os três postulados seguintes:

$$\|x\| > 0, \text{se } x \neq 0, \text{se } x \in \vee \tag{1.5}$$

$$\|\lambda x\| = |\lambda|\|x\|, \lambda \in R, x \in \vee \tag{1.6}$$

$$\|x + y\| \leq \|x\| + \|y\|, \text{se } x, y \in \vee \tag{1.7}$$

A função $\|x\|$ pode ser interpretada como comprimento ou magnitude de x. A norma de um vetor generaliza o conceito de valor absoluto $|r|$ para um número real ou complexo.

Um exemplo clássico de normas é representado pela classe de função conhecida como normas – p, i.e.:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i^p| \right)^{1/p} \tag{1.8}$$

onde $x = (x_1, x_2, \cdots, x_n)^T$, e $1 \leq p < \infty$.

As normas – p mais conhecidas são : i) a norma Euclidiana $(p=2)$; ii) a norma L_1 $(p=1)$, e iii) a norma L_∞ $(p \to \infty)$. No caso da norma L_∞, a Eq. (8) assume a forma:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \tag{1.9}$$

Exemplo 1.1) Utilizando a definição de normas-p, construa uma tabela comparativa dos comprimentos obtidos para os quatro vetores seguintes no R^3 a partir das normas p = 1, 2 e ∞.

$$w = (1, 2, -3)^T \qquad x = (-2, 0, 8)^T$$

$$y = (4, -4, 4)^T \qquad z = (0, 0, 7)^T$$

Solução :

	$\|\cdot\|_1$	$\|\cdot\|_2$	$\|\cdot\|_\infty$
w	6	3,74	3
x	10	8,25	8
y	12	6,93	4
z	7	7	7

Teorema 1.1

Sejam dois vetores **x** e **y**, então

$$|x \,.\, y| \leq \|x\|\|y\| \tag{1.10}$$

Prova :

a) No R^2 e R^3 :

$$|x \cdot y| = \|x\| \, \|y\| \, |\cos\theta|$$

Como $|\cos\theta| \leq 1$, então

$$|x \cdot y| \leq \|x\| \|y\|$$

b) No R^n, o resultado do Teorema 1.1 é conhecido como desigualdade de Cauchy-Schwarz:

b.1) Para $x = 0$ ou $y = 0$, $x \cdot y = \|x\| \cdot \|y\| = 0$, portanto vale a igualdade

b.2) Se $x \neq 0 \, e \, y \neq 0$

Seja $z = hx - ky, h, k \in R$

$$z \cdot z = h^2 \cdot \|x\|^2 - 2hk(x \cdot y) + k^2 \|y\|^2 \geq 0$$

para todo $h, k, \in R$, portanto, fazendo

$$h = \|y\| \, e \, k = \|x\|$$

$$2\|x\|^2 \|y\|^2 \geq 2\|x\| \|y\| (x \cdot y)$$

Como $x \neq 0 \, e \, y \neq 0, \|x\| \|y\| > 0$

Dividindo ambos os lados por $2\|x\| \|y\|$

Tem-se

$$\|x\| \|y\| \geq x \cdot y$$

Tomando os módulos, obtém-se

$$\|x\| \|y\| \geq |x \cdot y|$$

Exemplo 1.2) Verifique se a norma Euclidiana atende aos três postulados definidos pelas Eqs. (1.5) – (1.7).

Solução:

i) A partir da definição da norma Euclidiana, para p = 2 na Eq. (8), observa-se que:

$$\|x\| = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2} = (x \,.\, x)^{1/2} = 0$$

se e somente se $x = 0$.

ii) Para $\lambda \in R$, calcula-se:

$$\|\lambda x\| = \left(\lambda^2 x_1^2 + \lambda^2 x_2^2 + \cdots + \lambda^2 x_n^2\right)^{1/2}$$

$$= |\lambda| \cdot \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2}$$

$$= |\lambda| \, \|x\|$$

iii) $$\|x + y\|^2 = (x + y) \cdot (x + y)$$

$$= x \cdot x + 2x \cdot y + y \cdot y$$

$$= \|x\|^2 + 2\, x \cdot y + \|y\|^2 \qquad \text{(tomando o módulo)}$$

$$\leq \|x\|^2 + 2\, x \cdot y + \|y\|^2 \qquad \text{(Desigualdade de Cauchy-Schwarz)}$$

$$\leq \|x\|^2 + 2\|x\| \, \|y\| + \|y\|^2$$

$$= \left(\|x\| + \|y\|\right)^2$$

e considerando somente a raiz quadrada positiva nos extremos da inequação, obtém-se:

$$\|x + y\| \leq \|x\| + \|y\|$$

que é também conhecida como Desigualdade do Triângulo.

Esta denominação provém do fato de que em duas e três dimensões, a soma de dois lados de um triângulo é maior do que o terceiro lado.

Normas matriciais

A análise de algoritmos matriciais freqüentemente requer o uso de normas matriciais. Por exemplo, um algoritmo para a solução de sistemas de equações lineares (Cap. 3) pode ser de baixa qualidade se a matriz de coeficientes (A) apresenta dificuldades para ser invertida, i.e., é "praticamente singular". No entanto, para avaliar a noção do que é "praticamente

singular, é necessário que se defina uma medida de distância no espaço de matrizes. As normas matriciais fornecem essa medida.

Lembre-se que uma matriz é constituída de várias linhas e colunas, sendo que um vetor é o caso particular de uma matriz $(m\times 1)$, i.e., com m componentes (linhas) e apenas 1 coluna. Numa matriz, as várias colunas podem ser interpretadas como vetores que a constituem, os quais podem ser alinhados para armazenarem toda a informação contida na matriz em um único vetor. Assim, o espaço $R^{m\times n}$ que contém as matrizes $(m\times n)$ é isomórfico ao espaço R^{mn}, que contém os vetores $(mn\times 1)$, e a definição de uma norma matricial deve ser equivalente à definição de uma norma vetorial. Especificamente, $\|\cdot\| : R^{m\times n} \to R$ é uma norma matricial se satisfizer aos postulados definidos pelas Eqs. (1.5) – (1.7).

As normas matriciais mais usadas em Álgebra Linear numérica são:
Norma de Frobenius

$$\|A\|_F = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right\}^{1/2} \tag{1.11}$$

Normas-p

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} \tag{1.12}$$

onde sup indica *supremum*, i.e., o maior valor.

Note que as normas-p matriciais são definidas em termos das normas-p vetoriais definidas pela Eq. (1.8). A verificação de que as Eqs. (1.11) e (1.12) são de fato normas matriciais é deixada como um exercício para o leitor. Fica claro que $\|A\|_p$ é a norma-p do maior vetor obtido pela operação de A com a norma-p de um vetor unitário:

$$\|A\|_p = \sup_{x\neq 0} \left\| A\left(\frac{x}{\|x\|_p} \right) \right\| = \max_{\|x\|_p = 1} \|Ax\|_p \tag{1.13}$$

Portanto, se uma norma vetorial $\|\cdot\|$ for especificada, pode-se definir uma norma matricial subordinada a essa norma vetorial por

$$\|A\| = \sup\{\|Au\| : u \in R^n, \|u\| = 1\} \tag{1.14}$$

Uma importante conseqüência da definição apresentada pela Eq. (1.14) é que

$$\|Ax\| \le \|A\| \, \|x\| \qquad \left(x \in R^n\right) \tag{1.15}$$

A prova é simples, bastando passar $\|x\|$ para o primeiro membro e observar que o vetor $x/\|x\|$ é unitário. Assim, pela Eq. (1.14), $\|A\| \ge \|Ax\|/\|x\|$.

Além dos postulados básicos das Eqs. (1.5) – (1.7), a norma subordinada matricial tem propriedades adicionais, conforme se segue:

$$\|I\| = 1 \tag{1.16}$$

$$\|AB\| \le \|A\| \, \|B\| \tag{1.17}$$

As provas dessas propriedades são simples e deixadas como um exercício para o leitor. Prove também, usando um contra-exemplo que nem todas as normas matriciais satisfazem a propriedade estabelecida pela Eq. (1.17).

1.4. A SÉRIE DE TAYLOR

Imagine o problema de como construir uma função qualquer em um certo domínio que interessa, a partir do conhecimento de informações em apenas um ponto. Um exemplo prático típico de tal problema seria a determinação da função de variação da concentração de um determinado componente em uma mistura reativa de espécies em um reator químico ao longo do tempo, a partir de dados medidos experimentalmente.

A Figura 1.1 mostra o comportamento real de uma função hipotética em relação a uma variável independente x .

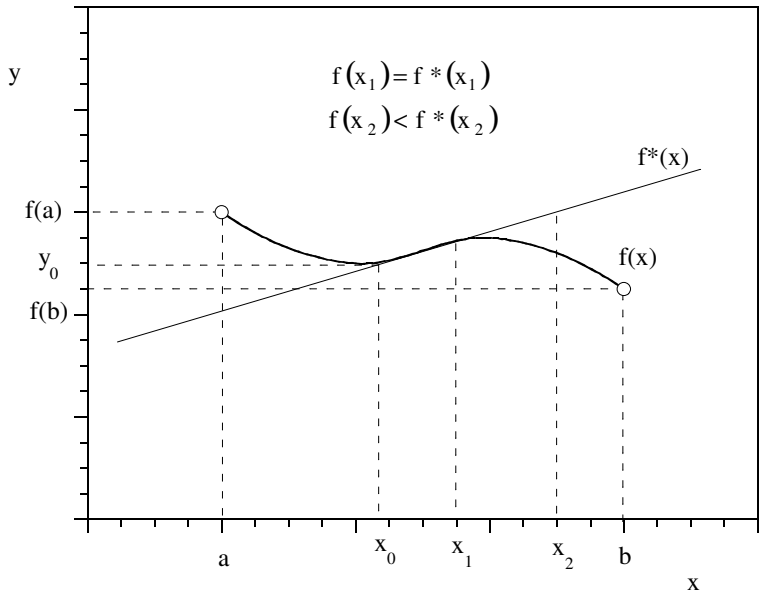

Figura 1.1 – Aproximação de f(x) por uma reta.

A princípio, assumem-se conhecidos $f(x_0)$ e $f'(x_0)$ somente. Com essas duas informações apenas, é possível aproximar-se o comportamento da função por uma reta, que é denominada $f^*(x)$ na Fig. 1.1.

A equação desta reta é determinada por

$$f^*(x) = f(x_0) + f'(x_0)(x - x_0) \tag{1.18}$$

A Figura 1.1 mostra que $f^*(x)$ não é uma boa aproximação para $f(x)$, considerando pontos distantes de $x = x_0$.

Verifica-se que, se houvesse mais informações, disponíveis em $x = x_0$, seria possível construir-se aproximações polinomiais para $f(x)$ de maior ordem. Para a construção de uma reta são necessários dois pontos ou condições; um polinômio de segundo grau, três pontos ou condições, e um polinômio de n-ésimo grau, n + 1 pontos ou condições.

Imagine um polinômio de ordem infinita como se segue:

$$f(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \cdots + b_n x^n + \cdots \tag{1.19}$$

que representa exatamente a curva desejada, pois interpola os infinitos pontos da curva real $f(x)$.

Alternativamente, a fim de se generalizar a Eq.(1.19), o polinômio pode ser construído, introduzindo-se uma posição genérica x_0:

$$f(x) = a_0 + a_1 (x - x_0) + a_2 (x - x_0)^2 + \cdots + a_n (x - x_0)^n + \cdots \tag{1.20}$$

Na equação (1.20) observa-se que $a_i \neq b_i$, uma vez que x_0 foi introduzido na formulação. O problema está em encontrar o valor dos coeficientes. Este problema é resolvido diferenciando-se sucessivamente a Eq. (1.20):

$$\begin{aligned}
f'(x) &= a_1 + 2a_2 (x - x_0) + \ldots + n a_n (x - x_0)^{n-1} + \cdots \\
f''(x) &= 2a_2 + \ldots + n(n-1) a_n (x - x_0)^{n-2} + \cdots \\
&\vdots \\
f^{(n)}(x) &= n!\, a_n + (n+1)!\, a_{n+1} (x - x_0) + \cdots
\end{aligned} \tag{1.21}$$

A seguir, na Eq. (1.21), para $x = x_0$, obtém-se :

$$f^{(n)}(x_0) = n!\, a_n$$

$$a_n = \frac{f^{(n)}(x_0)}{n!} \tag{1.22}$$

Desta maneira, pode-se representar $f(x)$, uma função qualquer, exatamente, desde que se conheçam infinitas condições em um determinado ponto x_0, i.e, $f(x_0), f'(x_0), \cdots f^{(n)}(x_0), \cdots$, através do polinômio de ordem infinita:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2!} + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!} + \cdots \tag{1.23}$$

que tem infinitos termos.

A Equação (1.23) é conhecida com Série de Taylor. Quando $x_0 = 0$, recebe a denominação de Série de Maclaurin.

Numericamente, é impossível utilizar-se uma Série de Taylor com um número infinito de termos. Por esta razão, para avaliar o valor da função numericamente, só é possível aproximar-se a função $f(x)$ por uma Série de Taylor com um certo número finito de termos. No entanto, é necessário uma avaliação da magnitude do erro cometido. Por exemplo, pode-se truncar a série no termo de ordem n na Eq. (1.23), obtendo:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2!} + \cdots + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!} + R_n \tag{1.24}$$

onde R_n é o erro cometido, igual a soma dos termos desprezados ao aproximar-se de $f(x)$ por uma série de Taylor com os $n+1$ primeiros termos.

O erro R_n pode ser estimado conforme se segue. Primeiramente, considere

$$\int_{x_0}^{x} f'(t)\,dt = [f(t)]_{x_0}^{x} = f(x) - f(x_0)$$

ou

$$f(x) = f(x_0) + \underbrace{\int_{x_0}^{x} f'(t)dt}_{R_1} \tag{1.25}$$

Integrando-se R_1 por partes, obtém-se:

$$R_1 = f'(x_0)(x - x_0) + \int_{x_0}^{x} (x - t)\, f''(t)dt$$

Portanto, reescreve-se a Eq. (1.25), i.e, a série de Taylor como:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \underbrace{\int_{x_0}^{x} (x - t)\ f''(t)dt}_{R_2} \tag{1.26}$$

Integrando-se R_2 por partes e segundo um raciocínio análogo, obtém-se

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2} + \underbrace{\int_{x_0}^{x} \frac{(x - t)^2}{2}\ f'''(t)dt}_{R_3} \tag{1.27}$$

Das equações (1.25) – (1.26) conclui-se a seguinte relação de recorrência para n qualquer:

$$R_n = \int_{x_0}^{x} \frac{(x - t)^{n-1}}{(n-1)!}\ f^{(n)}(t)dt \tag{1.28}$$

Necessita-se ainda buscar uma estimativa para o valor da integral definida pela Eq. (1.28). Para tanto utiliza-se a forma Lagrangeana, que é obtida observando-se que, na integração, t varia de x_0 a x, portanto, $f^{(n)}(t)$ possui um mínimo valor m e um máximo valor M associados a x ou a x_0 cada um, assumindo-se que $\exists\ f^{(n)}(t)$, i.e, a função é contínua. Portanto, pode-se escrever

$$\left| \int_{x_0}^{x} m \frac{(x - t)^{n-1}}{(n-1)!} \right|_{x_0}^{x} dt \le |R_n| \le \left| \int_{x_o}^{x} M \frac{(x - t)^{n-1}}{(n-1)!} \right|_{x_o}^{x} dt$$

e

$$\left| m\left[\frac{-(x - t)^n}{n!} \right]_{x_0}^{x} \right| \le |R_n| \le \left| M\left[-\frac{(x - t)^n}{n!} \right]_{x_0}^{x} \right|$$

Observa-se que

$$m = \ f^{(n)}(x_0) \text{ ou } f^{(n)}(x)$$

e

$$M = f^{(n)}(x_0) \text{ ou } f^{(n)}(x)$$

Portanto, para t entre x_0 e x, $\exists\ t = \xi$ tal que R_n é calculada exatamente por

$$|R_n| = \left| f^{(n)}(\xi) \frac{(x - x_0)}{n!} \right| \tag{1.29}$$

A partir da Eq. (1.29), outra importante forma de estimar o erro cometido ao aproximar uma função qualquer, por uma série de Taylor truncada em um termo qualquer de ordem n, pode ser estabelecida. Inicialmente, define-se $\Delta x = x - x_0$ e observa-se que $\frac{f^{(n)}}{n!}(\xi)$ pode ser entendido como uma certa constante de ordem 1, para n suficientemente grande. Assim, na realidade, o erro depende prioritariamente da proximidade da posição em que se desejar estimar $f(x)$, além do número de termos considerados na série de Taylor, i.e.:

$$R_n = O(\Delta x^n) \tag{1.30}$$

que indica que o erro é ordem Δx^n, i.e., da mesma ordem de magnitude de Δx^n.

Exemplo 1.3) Utilize uma série de Taylor construída com informações em $x_0 = 0$, para calcular o valor do cos x, onde $x = 0{,}5$ rad. Qual a ordem de n do termo em que a série deverá ser truncada para que o erro cometido seja menor do que 10^{-4}? A série truncada resultante terá $n+1$ termos.

Solução:

Deseja-se um erro menor do que 10^{-4} (0,0001), para x = 0,5 rad, portanto, para $x_0 = 0$, tem-se

$$|R_n| \le \left| M\left(\frac{x^n}{n!}\right) \right| = \left| M \frac{0{,}5^n}{n!} \right| \le 0{,}0001$$

Observa-se que M é o valor máximo de $f^{(n)}(\xi)$, para ξ entre x e x_0. A n-ésima derivada do cos x, a menos do sinal, ou será um coseno ou seno, portanto:

$$|R_n| \le \left| M \frac{0{,}5^n}{n!} \right| \le \left| \frac{0{,}5^n}{n!} \right| \le 0{,}0001$$

Desta maneira, calcula-se

$$n! \ge 0{,}5^n \times 10^4$$

$$n = 6, \text{uma vez que } 6! = 720 > 156{,}25$$

A série de Taylor para o cos x em torno de $x_0 = 0$, até o sexto termo é

$$f(x) = 1 + 0 - \frac{x^2}{2} + 0 + \frac{x^4}{24} - 0$$

Portanto:

$$f(0{,}5) = 0{,}877604166667$$

Verificação do erro:

Em uma calculadora HP 48 G+, obtém-se cos 0,5 = 0,87758256189, assim

$$|R_5| = |f(0{,}5) - \cos 0{,}5| = 2.1604777 \times 10^{-5}$$

1.5. CÁLCULO DE FUNÇÕES POR SÉRIES DE POTÊNCIAS

Um procedimento computacional comumente utilizado é a avaliação de funções matemáticas por séries de potências. A série de Taylor discutida na seção 1.4 é um exemplo de série de potências. Sempre que possível, portanto, é desejável prever-se como a série vai se comportar na avaliação de uma determinada função. Um procedimento numérico deve resultar em um valor finito, i.e., mensurável, para que tenha utilidade em uma aplicação de Engenharia, por exemplo.

Teorema 1.2 – Teorema de Leibniz

Seja uma série infinita $c_1 - c_2 + c_3 - c_4 + \cdots$

1. Estritamente alternada
2. Cada termo é menor, em módulo, do que o termo precedente
3. O limite dos termos é zero

Então a série possui soma finita.

Quando a função em questão possui um ponto singular, então existe uma região de convergência limitada.

Exemplo 1.4) As funções $\operatorname{sen} x$, $\cos x$, e^x não têm pontos singulares em $x \in R$, mas $\ln x$ tem uma singularidade em $x = 0$. Utilizando uma série de Taylor em torno de $x_0 = 1$, determine a região de convergência da série, i.e., a região em que a série pode ser utilizada para calcular $\ln x$. Além disso, $\ln x \notin R$ para $x < 0$.

Solução:

$$\ln x = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \frac{(x-1)^5}{5} - \cdots$$

Observa-se que para $x = 4$, por exemplo:

$$\ln 4 = (4-1) - \frac{(4-1)^2}{2} + \frac{(4-1)^3}{3} - \cdots$$

portanto, a soma não é finita, i.e., a série diverge e não poderá ser utilizada para avaliar o $\ln 4$.

Verifica-se que a série é estritamente alternada $\forall x \in R$, porém cada termo geral c_n somente será menor, em módulo, do que o termo precedente se e somente se $x \leq 2$. Note que, para $x = 2 + \delta$, onde $\delta > 0$, $\delta \in R$

$$\frac{|C_{n+1}|}{|C_n|} = \frac{\dfrac{(1+\delta)^{n+1}}{n+1}}{\dfrac{(1+\delta)^n}{n}} = (1+\delta)\frac{n}{n+1}$$

portanto $\lim_{n\to\infty}(1+\delta)\dfrac{n}{n+1} = 1+\delta$

Assim, o Teorema de Leibniz não é satisfeito e a série não converge para uma soma finita. Desta maneira, esta série só pode ser utilizada para calcular o $\ln x$, quando $0 < x \leq 2$, que é a região de convergência da mesma.

Observa-se que, tendo avaliado o ln 2 com esta série, pode-se utilizar uma outra série de Taylor, obtida em torno de $x_0 = 2$, para calcular o $\ln x$ para $2 < x \leq 3$, e assim sucessivamente.

1.6. TEOREMA DO VALOR MÉDIO

O Teorema do Valor Médio, TVM, pode ser interpretado como o Teorema de Taylor quando $n = 1$, sendo muito usado como argumento matemático. Seja $f \in C[a,b], \exists f''(x) \in (a,b)$, então para qualquer x e $C \in [a,b]$, verifica-se

$$f(x) = f(c) + f'(\xi)(x-c) \tag{1.31}$$

para ξ entre x e c.

Interpretação gráfica:
Tomando $x = b$ e $c = a$, escreve-se com base na Eq. (1.31)
$f(b) - f(a) = f'(\xi)(b-a)$, onde $a < \xi < b$.

Portanto, verifica-se através da Fig. 1.2 para uma função genérica $f(x)$ que

$$tgx = f'(\xi) = \frac{f(b) - f(a)}{b - a} \tag{1.32}$$

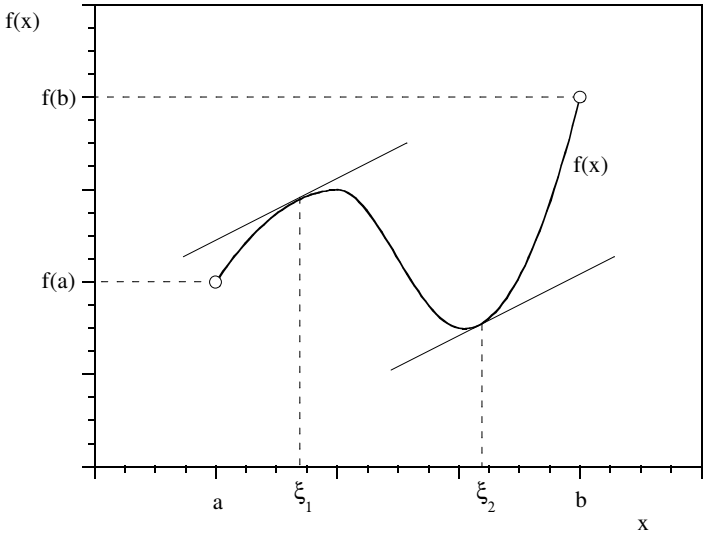

Figura 1.2 – Teorema do Valor Médio.

Assim, existirá sempre, pelo menos uma posição $x = \xi$ tal que a inclinação da reta tangente à curva $f(x)$ naquele ponto, i.e., $f'(\xi)$é calculada exatamente pela Eq.(1.32).

Caso especial:

Se $f \in C^{\circ}[a,b], \exists f' \in C^{1}(a,b)$, f(a) = f(b) = constante, tal que $f'(\xi) = 0$, para pelo menos um valor $\xi \in (a,b)$. Este caso especial é conhecido como Teorema de Rolle.

Em ambos os Teoremas, poderá haver mais de um valor de $\xi \in (a,b)$ que satisfaça as Eqs. (1.31) e (1.32).

1.7. ARREDONDAMENTO E TRUNCAMENTO

Arredondamento e truncamento são importantes conceitos para a aproximação de números reais em rotinas de cálculo. Para entender aproximadamente as diferenças entre os dois conceitos, define-se um número real $x > 0$ como

$$0,---\cdots---$$

com m dígitos à direita da vírgula.

Matematicamente, define-se arredondar x para n casas decimais (n < m) para cima ou para baixo. Se o último dígito após a vírgula $(n+1) \geq 5$, arredonda-se para cima, e caso

contrário arredonda-se para baixo. Em contrapartida, define-se truncar para n casas decimais ($n < m$), o procedimento de simplesmente desprezar todos os dígitos do número x, após o dígito n.
A Tabela 1.1 mostra alguns exemplos de arredondamento e truncamento.

Tabela 1.1 – Exemplos de arredondamento e truncamento para $n=3$.

x	Arredondamento	Truncamento
0,27448	0,274	0,274
0,38674	0,387	0,386
1,99994	2,000	1,999
3,32513	3,325	3,325
0,61208	0,612	0,612

Teorema 1.3

Se x é arredondado tal que $\tilde{x}$ é a aproximação de n-dígitos para o número, então

$$|x - \tilde{x}| \leq \frac{1}{2} \times 10^{-n} \tag{1.33}$$

Prova:

i) Se o dígito $(n+1)$ de x for 0, 1 , 2 , 3 ou 4, então $x = \tilde{x} + \varepsilon$ com $\varepsilon < \frac{1}{2} \times 10^{-n}$, portanto a Eq. (1.33) é verdadeira.

ii) Se o dígito $(n+1)$ de x for 5, 6, 6, 8 ou 9 então $\tilde{x} = \hat{x} + 10^{-n}$, onde $\hat{x}$ é um número com n dígitos e todos os outros a direita são zero (i.e., o número truncado com n dígitos.) Portanto, $x = \hat{x} + \delta \cdot 10^{-n}$, com $\delta \geq \frac{1}{2}$. Assim $\tilde{x} - x = (1-\delta) \times 10^{-n}$ como $1 - \delta \leq \frac{1}{2}$, então a Eq.(26) também é verdadeira e a demonstração está completa.

Analogamente, quando truncamos x, no dígito n, $\hat{x}$ é obtido, portanto:

$$|x - \hat{x}| < 10^{-n} \tag{1.34}$$

Prova:

Verifica-se que $x = \hat{x} + \delta \times 10^{-n}$ com $0 \leq \delta < 1$, assim $|x - \hat{x}| = |\delta| \times 10^{-n} < 10^{-n}$, então a Eq. (1.34) é verdadeira.

1.8 ORDENS DE CONVERGÊNCIA

Em um procedimento numérico executado em um computador, em muitos casos a resposta do problema não é obtida diretamente em um único cálculo. De fato, nestes casos, é necessário iniciar os cálculos a partir de uma estimativa para a resposta e, através de uma seqüência de cálculos semelhantes, a estimativa é corrigida para produzir respostas aproximadas, em um procedimento iterativo. Em cada iteração, por um critério pré-definido apropriadamente, a precisão da resposta é verificada; parando os cálculos quando a precisão da resposta aproximada for satisfatória. Neste ponto, é conveniente ressaltar que o procedimento numérico só terá utilidade prática se a cada iteração a precisão da resposta aproximada for progressivamente melhor. Esta discussão, portanto, define o conceito de **convergência** de um procedimento numérico, i.e., quando a resposta aproximada obtida satisfaz a precisão da resposta esperada para o problema, seja por um procedimento iterativo ou não. Assim, um procedimento numérico iterativo será útil para obter a resposta aproximada de um problema somente quando for gerado por uma seqüência convergente.

Seqüências convergentes

Para encontrar a resposta aproximada de um problema por um procedimento numérico, um programa de computador deve gerar uma seqüência de números $x_1, x_2, x_3 \cdots$ que se aproximem da resposta correta. Pode-se pensar, por exemplo, na raiz de uma equação não linear, no valor numérico da derivada ou integral definida de uma função complicada.

Matematicamente, deseja-se que

$$\lim_{n\to\infty} x_n = x \tag{1.35}$$

onde a seqüência real $[x_n]$ converge para o número real x, se dado $\varepsilon > 0, \exists\, n_o \in Z^+$ tal que $|x - x_n| < \varepsilon,\ n \geq n_0$.

Além de necessitarmos para um procedimento numérico uma seqüência convergente, é importante saber, se possível, *a priori* qual a rapidez com que o procedimento numérico levará a uma resposta aproximada suficientemente precisa. Desta maneira, define-se o conceito de ordem de convergência como um indicador de uma seqüência lenta ou rápida. Os exemplos simples a seguir ilustram como seqüências podem se aproximar lenta ou rapidamente da resposta de um problema.

Exemplo 1.5) A equação $e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ define o número irracional e (constante de Euler). Se for utilizada a seqüência $x_n = \left(1+\frac{1}{n}\right)^n$ em um computador, $x_1 = 2$. Prosseguindo os cálculos, verifica-se que para n = 1000, $x_{1000} = 2{,}716924$. Observa-se que , quando comparado ao valor exato $e = 2{,}7182818...$, ainda se verifica que $|e - x_{1000}| \cong 0.00136 > 10^{-3}$. Portanto, este é um exemplo de seqüência que converge lentamente para a resposta do problema.

Exemplo 1.6) Considerando a seqüência:

$$\begin{cases} x_1 = 2 \\ x_{n+1} = \dfrac{1}{2} x_n + \dfrac{1}{x_n} \qquad (n \geq 1) \end{cases}$$

Verifica-se que ao realizar os cálculos em um computador $x_1 = 2$, e $x_4 = 1{,}414216$. Portanto, a resposta aproximada calculada por esta seqüência se aproxima rapidamente do valor exato da resposta esperada, i.e., $\sqrt{2} = 1{,}41421...$

Grande "O" e pequeno "o"

Uma técnica bastante comum de verificar a rapidez com que uma seqüência converge para um determinado valor é conhecida como metodologia do grande "O" e do pequeno "o".

Trata-se de uma técnica que permite avaliar a taxa de convergência de uma seqüência qualquer a partir da comparação direta com outra sequência cuja taxa de convergência é conhecida, i.e., outra seqüência de características semelhantes à seqüência sob análise.

Sejam $[x_n]$ e $[\alpha_n]$ duas seqüências diferentes, de características matemáticas semelhantes, onde se conhece a taxa de convergência de $[\alpha_n]$. Para avaliar a taxa de convergência de $[x_n]$, consideram-se duas possibilidades:

$$\text{I) } x_n = O(\alpha_n) \tag{1.36}$$

Se houver constantes c_1 e c_2, tal que $|x_n| \leq c_1 |\alpha_n|$ quando $n \geq c_2$. Então, se $\alpha_n \neq 0$ para todo n, então $\left|\dfrac{x_n}{\alpha_n}\right|$ permanece limitada quando $n \to \infty$,

e

$$\text{II) } x_n = o(\alpha_n) \tag{1.37}$$

Significa que $\lim_{n\to\infty}\left(\dfrac{x_n}{\alpha_n}\right) = 0$

O entendimento dessas duas definições fica claro observando a situação de $x_n \to 0$ e $\alpha_n \to 0$. Observe que não há perda de generalidade nesta análise, uma vez que a situação genérica $x_n \to L_1$ e $\alpha_n \to L_2$, onde L_1 e L_2 são duas constantes quaisquer é reduzida ao caso anterior criando duas novas seqüências $y_n = x_n - L_1$ e $\beta_n = \alpha_n - L_2$. Verifica-se que $y_n \to 0$ e $\beta_n \to 0$, onde, também se conhece a taxa de convergência β_n, uma vez que, a taxa de convergência de α_n é conhecida *a priori.* Portanto, pelas Eqs. (1.36) e (1.37), pode-se afirmar que:

a) Se $x_n = O(\alpha_n) \Rightarrow x_n$ converge para zero pelo menos tão rapidamente quanto α_n, e

b) Se $x_n = o(\alpha_n) \Rightarrow x_n$ converge para zero mais rapidamente do que α_n.

Exemplo 1.7) Considere as seqüências abaixo e determine se elas têm uma convergência lenta ou rápida através da comparação com a seqüência $\frac{1}{n}$, que tem uma taxa de convergência reconhecidamente lenta.

i) $\frac{n+1}{n^2}$

Solução:

Compara-se com a seqüência $\frac{1}{n}$ fazendo:

$\lim_{n\to\infty} \frac{\frac{n+1}{n^2}}{\frac{1}{n}} = \lim_{n\to\infty} \frac{n+1}{n} = 1 \neq 0$, portanto

$\frac{n+1}{n^2} = O\left(\frac{1}{n}\right)$, i.e., lenta

ii) $\frac{1}{n \ln n}$

Solução:

$\lim_{n\to\infty} \frac{\frac{1}{n \ln n}}{\frac{1}{n}} = \lim_{n\to\infty} \frac{1}{\ln n} = 0$, portanto

$\frac{1}{n \ln n} = o\left(\frac{1}{n}\right)$, i.e., rápida

iii) $\frac{8}{n} + 4^{-n}$

$\lim_{n\to\infty} \frac{\frac{8}{n} + 4^{-n}}{\frac{1}{n}} = \lim_{n\to\infty} 8 + \lim_{n\to\infty} \frac{1}{n4^n} = 8 \neq 0$, portanto

$\frac{8}{n} + 4^{-n} = O\left(\frac{1}{n}\right)$, i.e., lenta

O conteúdo dessa seção demonstrou a necessidade de se avaliar a rapidez ou a taxa de convergência de um procedimento numérico iterativo no computador. Propõe-se, portanto, uma classificação geral para as ordens de convergência de procedimentos numéricos iterativos com base no erro absoluto da solução procurada para o problema.

Seja $[x_n]$ uma seqüência de números reais tendendo a x^* (solução exata). Apresenta-se a seguinte classificação de ordens de convergência:

i) A taxa de convergência é pelo menos **linear** se houver uma constante $c < 1$ e um inteiro N tal que

$$|x_{n+1} - x^*| \leq c|x_n - x^*| \quad (n \geq N) \tag{1.38}$$

ii) A taxa de convergência é pelo menos **superlinear** se $\exists\ \varepsilon_n \to 0$ e um inteiro N, tal que

$$|x_{n+1} - x^*| \leq \varepsilon_n |x_n - x^*| \quad (n \geq N) \tag{1.39}$$

iii) A taxa de convergência é pelo menos **quadrática** se $\exists C$ (não necessariamente menor do que 1), e um inteiro N, tal que

$$|x_{n+1} - x^*| \leq C|x_n - x^*| \quad (n \geq N) \tag{1.40}$$

iv) A taxa de convergência é pelo menos ordem α, se $\exists C, \alpha$ e um inteiro N, tal que

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^{\alpha} \quad (n \geq N) \tag{1.41}$$

1.9. OS NÚMEROS NO COMPUTADOR

O computador se comunica com o usuário no sistema decimal, mas trabalha internamente no sistema binário, através de procedimentos de conversão. O comprimento da palavra do computador impõe restrições na precisão com que os números reais podem ser representados.

Em realidade, é impossível a representação dos infinitos números da reta numérica pelo computador. Portanto, a palavra do computador determina o valor do maior e do menor número em módulo, possível de ser representado pelo mesmo. Além disso, a mesma palavra só permite ao computador representar um conjunto discreto (finito) de números, i. e., entre um número de máquina e o próximo há um "espaço vazio" que não é possível de ser preenchido devido à limitação construtiva da palavra ("hardware"). Desta maneira, o computador poderá estar "errado" até mesmo ao representar um simples.número. Este fato não torna os cálculos realizados em computador "não confiáveis", apenas demonstra a necessidade de controlar o erro cometido nas aproximações numéricas nos procedimentos computacionais. Numa rotina de cálculo numérico é, portanto, fundamental manter os erros sob controle e com baixo valor.

1.10. REPRESENTAÇÕES DE NÚMEROS

Um mesmo número pode ser expresso em diferentes bases ou sistemas numéricos, bem como em diferentes notações. Uma base numérica define o numero de caracteres básicos para a formação de qualquer número a ser expresso naquela base. As diferentes notações definem o formato de escrita de um número em qualquer base.

Se um mesmo número pode ser representado em bases diferentes, é necessário o estabelecimento de um procedimento de conversão do valor de um número de uma base para outra. O computador utiliza-se desses procedimentos para operar internamente na base 2 e se comunicar com o usuário na base 10.

Um número em qualquer base é representado por.

$$Z_\alpha = a_m a_{m-1} ... a_1 a_0 , b_1 b_2 ... b_n \tag{1.42}$$

onde $\alpha \in N$ representa a base em que o número está representado, os dígitos "a" representam a parte inteira do número e os dígitos "b" a parte fracionária.

Note que na Eq. (1.42), quando α>10 há a necessidade de utilização de caracteres adicionais para a representação dos dígitos básicos, referentes à base escolhida. Os dígitos básicos, δ, de uma determinada base formam o conjunto

$$B = \{\delta \in N \mid 0 \le \delta \le \alpha - 1\} \tag{1.43}$$

Observa-se que, para $0 \le \delta \le 9$, os dígitos básicos são os conhecidos algarismos da Aritmética. No entanto, quando δ > 9, há a necessidade de se estabelecer caracteres adicionais para representar os dígitos básicos restantes. Por exemplo, na base hexadecimal $(\alpha = 16)$, bastante utilizada computacionalmente, os dígitos são 0,1,2,3,4,5,6,7,8,9, A,B,C,D,E,F. As letras de A a F são utilizadas para representar os dígitos básicos para $10 \le \delta \le 15$, de acordo com a Eq. (1.43) para a base hexadecimal $(\delta = 16)$.

Um número na base α é convertido para a base 10 através da seguinte expressão:

$$Z_{10} = a_m \alpha^m + ... + a_0 \alpha^0 + b_1 \alpha^{-1} + ... + b_n \alpha^{-n} \tag{1.44}$$

A conversão de um número na base 10 para uma base α é feita primeiramente operando a parte inteira com a seguinte seqüência de cálculos:

$$\begin{aligned} R_0 &= (Z_{10})_{int} - q_1 \alpha \\ R_1 &= q_1 - q_2 \alpha \\ R_2 &= q_2 - q_3 \alpha \\ &. \\ &. \\ &. \\ R_p &= q_p - q_{p+1} \alpha \end{aligned} \tag{1.45}$$

onde q_1 é a parte inteira do quociente da divisão da parte inteira de $Z_{10}, (Z_{10})_{int}$, por α, q_2 a parte inteira do quociente da divisão de q_1 por α, e assim sucessivamente até que a parte

inteira do quociente $q_{p+1} < \alpha$, momento em que o procedimento se encerra. Observe que $0 \le R_i \le \alpha - 1 \ \ (0 \le i \le p)$. A parte inteira do número convertido para a base α é, portanto:

$$q_{p+1}R_pR_{p-1}...R_1R_0$$

A parte fracionária de Z_{10} é convertida para a base α realizando multiplicações sucessivas conforme se segue.

$$\begin{aligned} &0, b_1b_2...b_n \times \alpha = S_1, C_1C_2...C_n \\ &0, C_1C_2...C_n \times \alpha = S_2, d_1d_2...d_n \\ &0, d_1d_2...d_n \times \alpha = S_3, L_1L_2...L_n \\ &. \\ &. \\ &. \end{aligned} \tag{1.46}$$

As multiplicações devem prosseguir até que a parte fracionária do resultado seja 0, se o número for representável exatamente na base α. Caso contrário, o número não é exato na base α, e os cálculos devem parar quando for atingido o número de dígitos desejado para o resultado. O resultado final do número convertido na base α é

$$Z_\alpha = q_{p+1}R_pR_{p-1}...R_1R_0, S_1S_2S_3... \tag{1.47}$$

Exemplo 1.8) converta o número $Z_{10} = 9{,}8$ para a base $\alpha = 2$. Arredonde o resultado para 4 dígitos fracionários.
Solução:

i) inicialmente converte-se a parte inteira utilizando a Eq. (1.45)

$$q_1 = \frac{(Z_{10})_{int}}{\alpha} = \frac{9}{2} = 4; R_0 = (Z_{10})_{int} - q_1\alpha = 9 - 4 \times 2 = 1$$

$$q_2 = \frac{4}{2} = 2; R_1 = q_1 - q_2\alpha = 4 - 2 \times 2 = 0$$

$$q_3 = \frac{2}{2} = 1; R_2 = q_2 - q_3\alpha = 2 - 1 \times 2 = 0$$

Os cálculos param neste ponto, uma vez que $q_3 < 2$

ii) para a parte fracionária, utiliza-se a Eq.(1.46)

$$0{,}8\times 2 = 1{,}6$$
$$0{,}6\times 2 = 1{,}2$$
$$0{,}2\times 2 = 0{,}4$$
$$0{,}4\times 2 = 0{,}8$$

$$0{,}8\times 2 = 1{,}6$$
$$0{,}6\times 2 = 1{,}2$$
$$\vdots$$

iii) Combinando os itens i e ii, obtem-se

$$Z_2 = q_3 R_2 R_1 R, S_1 S_2 S_3 S_4 S_5 S_6 \ldots$$

$$Z_2 = 1001{,}110011\ldots$$

Finalmente, aproximando para 4 dígitos fracionários (arredondamento), tem-se

$$Z_2 = 1001{,}1101$$

Notação científica normalizada

Um número qualquer pode ser representado através de potências da base em que se encontra.

Em notação científica normalizada, o número é representado apenas por uma parte fracionária onde o primeiro dígito não é um zero significativo, multiplicado por uma potência de sua base.

Por exemplo, na base 10, representam-se os números:

$$592{,}4821 = 0{,}5924821\times 10^3$$

$$-0{,}006823 = -6{,}823\times 10^{-2}$$

generalizando para um número real $x \neq 0$:

$$x = \pm r\times 10^n \qquad (1.48)$$

onde $\frac{1}{10} \leq r < 1$, e n é um número inteiro (> 0, < 0 ou = 0). Note que,para $x = 0$, $r = 0$.

No sistema binário, o mesmo número $x \neq 0$ é representado por:

$$x = \pm q\times 2^m \qquad (1.49)$$

Onde $\frac{1}{2} \leq q < 1$, e m é um número inteiro, sendo que q e m recebem a denominação de mantissa e expoente, respectivamente. Note que q e m são números na base 2. A equação (1.49) recebe a denominação de forma normalizada em ponto flutuante.

Para um melhor entendimento da representação interna dos números em um computador, neste texto, define-se um computador hipotético com as seguintes características:

1. Comprimento de palavra – 32 bits
2. Sinal do nº real x – 1 bit
3. Sinal do expoente m – 1 bit
4. Expoente (nº inteiro |m|) – 7 bits
5. mantissa (nº real |q|) – 23 bits

Neste ponto, é interessante ressaltar a razão pela qual o computador utiliza internamente o sistema binário e não o decimal. Em eletrônica, para a representação dos dígitos básicos da base utilizada, faz-se uso de circuitos "flip-flop", i.e., que assumem, portanto, os valores 0 ou 1. Esse "hardware" se adequa perfeitamente ao sistema binário.

Para uma palavra definida como nos itens 1 a 5 acima, observa-se uma restrição para |m| de 7 bits. Em razão disto, nesse computador hipotético, $|\,m\,| \leq (1111111)_2$.

Note que $(1111111)_2 = 2^7 - 1 = 127$, i. e., a soma dos 7 primeiros termos de uma progressão geométrica de razão 2, cujo primeiro termo é 1. Assim, verifica-se que $2^{127} \approx 10^{38}$, de modo que a faixa de números reais representáveis nesse computador é $x \in (-10^{38}, 10^{38})$, e o menor número real possível de ser representado é da ordem de 10^{-38}.

As linguagens de programação permitem o aumento da precisão nos cálculos realizados no computador, juntando duas palavras, sendo que a mantissa e expoente passam a ter duas vezes mais bits. Este procedimento é denominado de dupla precisão.

Para uma palavra, a mantissa tem 23 bits, na realidade 24, pois o primeiro bit é assumido como tendo o valor 1. Assim, o último dígito da mantissa permite ao computador representar o menor intervalo possível entre dois números, com o valor de $2^{-24} \approx 10^{-7}$. Portanto, nesse computador hipotético, números com uma parte fracionária com mais de 7 dígitos decimais são aproximados.

Para a representação de números inteiros, o primeiro bit é reservado para o sinal. Os 31 bits restantes são utilizados para representar o número. Assim, a faixa de variação de números inteiros representáveis por esse computador vai de $-(2^{31}-1)$ até $2^{31} - 1 = 2147483647$.

A Figura 1.3 mostra graficamente a alocação de bits em uma palavra do computador hipotético em análise. Na figura 1.3, são definidas as seguintes variáveis: S-bit para o sinal de x, s-bit para o sinal de m, E é o valor do expoente (7 bits), F reserva 23 bits para a parte fracionária do número real $x(.1---...---)_2$.

Obtém-se desta maneira, o valor.

$$x = (-1)^S \times q \times 2^m \tag{1.50}$$

onde $q = (0.1F)_2$ e $m = (-1)^s \times E$, notando que o bit inicial de q é conhecido e não é explicitamente armazenado.

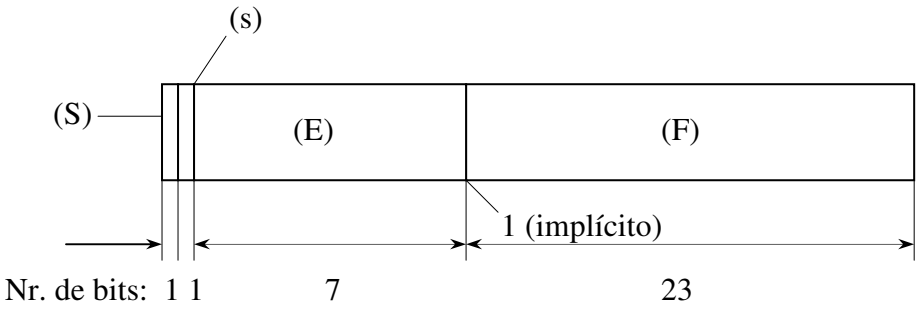

Figura 1.3 – Exemplo de palavra de computador.

Números aproximados de máquina

Continuando a utilizar o computador hipotético definido na seção anterior,verifica-se que

$$x = q \times 2^m, \quad \frac{1}{2} \le q < 1, \quad |m| \le 127 \tag{1.51}$$

A representação exata do número real x seria

$$x = (.a_1 a_2 ... a_{24} a_{25} a_{26} ...)_2 \times 2^m \tag{1.52}$$

Onde $a_1 = 1$.

O número aproximado de máquina é obtido simplesmente descartando os bits em excesso (truncamento), ou fazendo o arredondamento do número. Assim, nesse computador há duas situações possíveis

$$x' = (a_1 a_2 ... a_{24})_2 \times 2^m \tag{1.53}$$

que caracteriza o truncamento, que coincide com o arredondamento para baixo, e

$$x'' = \left((a_1 a_2 ... a_{24})_2 + 2^{-24}\right) \times 2^m \tag{1.54}$$

que caracteriza o arredondamento para cima.

A Figura 1.4 mostra graficamente as duas situações de arredondamento possíveis. O menor intervalo entre dois números nesse computador hipotético está esquematicamente representado na Fig. 1.4. No arredondamento, entre x' e x'', escolhe-se o mais próximo de x para representá-lo.

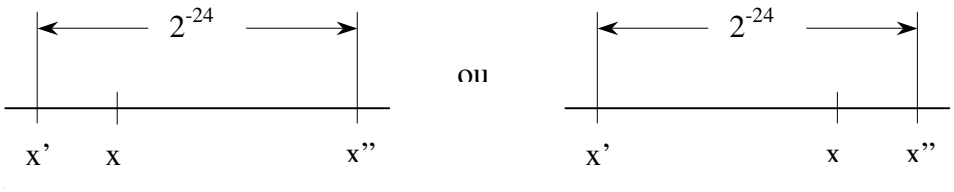

Figura 1.4 – Arredondamento de um número real x.

Se x' for número escolhido para representar x, então, o erro absoluto é dado por

$$|x - x'| \leq \frac{1}{2}|x'' - x'| = \frac{1}{2} \times 2^{-24} \times 2^{m} = 2^{m-25} \tag{1.55}$$

O erro relativo é dado por

$$\left|\frac{x - x'}{x}\right| \leq \frac{2^{m-25}}{q \times 2^{m}} = \frac{2^{-25}}{q} \leq \frac{2^{-25}}{1/2} = 2^{-24} \tag{1.56}$$

É possível, portanto que o número $x = \pm q \times 2^{m}$ esteja fora da faixa permitida pela palavra do computador. Assim, $m > 127$ ou $m < -127$, caracterizando os fenômenos de "overflow" (crítico) ou de "underflow" (alguns computadores assumem o valor zero neste caso), respectivamente. A situação de "overflow" é considerada crítica porque compromete o prosseguimento dos cálculos em uma rotina computacional, ao passo que na situação de "underflow" os cálculos podem prosseguir uma vez que seja assumido o valor nulo para o número.

Se x é um número real diferente de zero dentro da faixa de representação da máquina, então o número de máquina $\tilde{x}$ satisfaz a relação

$$\left|\frac{x - \tilde{x}}{x}\right| \leq 2^{-24} \tag{1.57}$$

Fazendo $\delta = (\tilde{x} - x)/x$, então $\tilde{x} = x(1 + \delta)$. Assim

$$\left|\frac{x - x(1+\delta)}{x}\right| \leq 2^{-24} \Rightarrow |\delta| \leq 2^{-24} \tag{1.58}$$

A equação (1.58) define o assim chamado erro unitário de arredondamento, δ. No caso do computador hipotético aqui definido, $|\delta| \leq 2^{-24}$.

A análise mostra que o número de bits alocado para a mantissa se relaciona diretamente com o erro unitário de arredondamento da máquina. Portanto, esse número de bits determina a precisão da aritmética computacional, tal que $fl(x) = \tilde{x}$, i. e., o numero de máquina de ponto flutuante $\tilde{x}$ mais próximo de x, que pode ser x' ou x''.

Em qualquer computador, verifica-se que as quatro operações aritméticas (representadas por Δ) satisfazem

$$fl(x\Delta y) = [x\Delta y](1+\delta) \tag{1.59}$$

onde $|\delta| < \varepsilon$, ε é o erro unitário de arredondamento da máquina, x e y são números de máquina.

Exemplo 1.9) Sejam x, y e z números de máquina, estime o erro cometido ao calcular $x(y+z)$ no computador hipotético definido pela Fig. 1.3.
Solução:

$$fl[x(y+z)] = [x\ \ fl(y+z)](1+\delta_1),\ \ |\delta_1| \le 2^{-24}$$

Onde δ_1 é o erro unitário de arredondamento cometido pelo computador ao realizar a operação de multiplicação entre x e $(x+y)$.

$$= [x(y+z)](1+\delta_2)(1+\delta_1), |\delta_2| \le 2^{-24}$$

Onde δ_2 é o erro unitário de arredondamento cometido pelo computador ao realizar a operação de soma entre y e z.

$$= x(y+z)(1+\delta_2+\delta_1+\delta_2\delta_1)$$
$$\cong x(y+z)(1+\delta_1+\delta_2)$$

Onde se constata que $\delta_1\delta_2 << \delta_1+\delta_2$, portanto desprezível no valor total da soma (i.e., o produto de dois infinitésimos é muito menor que o próprio infinitésimo)

$= x(y+z)(1+\delta_3),\ |\delta_3| \le 2^{-23}$, que é o limite superior do erro cometido pelo computador ao realizar a operação considerada.

Generalizando a análise apresentada nesta seção, pode-se concluir que em uma máquina operando na base β e tendo n dígitos na mantissa de seus números de ponto flutuante:

$$fl(x) = x(1+\delta),\ |\delta| \le \varepsilon \tag{1.60}$$

onde $\varepsilon = \frac{1}{2}\beta^{1-n}$ (arredondamento), $\varepsilon = \beta^{1-n}$ (truncamento), e ε é o erro de arredondamento unitário.

1.11. PERDA DE ALGARISMOS SIGNIFICATIVOS

Como um resultado de uma programação computacional “descuidada”, podem ocorrer erros nos resultados do cálculo de expressões matemáticas devido á perda de algarismos significativos. Note , por exemplo a seguinte expressão:

$$y = \sqrt{x^2 + 1} - 1 \tag{1.61}$$

Se x for um número muito pequeno, então o número de máquina $\tilde{x}$ que aproxima o número exato x, pode acarretar que a soma $\tilde{x}^2 + 1 = 1$ no computador. Neste caso, a Eq. (1.61), indicará que $y = 0$ no computador, o que não é o resultado correto.

Uma possível solução para o problema é reescrever a Eq. (1.61) como

$$y = \sqrt{x^{2+1}} - 1 \frac{\sqrt{x^2 + 1} + 1}{\sqrt{x^2 + 1} + 1} = \frac{x^2}{\sqrt{x^2 + 1} + 1} \tag{1.62}$$

A equação (1.62) evita a soma de um número muito pequeno com um número muito maior, calculando y por uma identidade matemática, evitando o erro cometido com a Eq. (1.61).

Exemplo 1.10) No caso do computador hipotético definido na seção 1.10, a mantissa da palavra é capaz de representar até 7 casas decimais após a virgula. Em um determinado ponto de um algoritmo, deseja-se realizar neste computador a soma $y = x^2 + 1$, sabendo que $x^2 = 0{,}2 \times 10^{-7}$.

Qual o valor de y calculado por esse computador?

Solução:

i) em notação cientifica normalizada, tem-se:

$$\begin{aligned} & x^2 = 0{,}2 \times 10^{-7} \\ + & \\ & 1 = 0{,}1 \times 10^{1} \end{aligned}$$

ii) Para realizar a soma, o computador representa x^2 na mesma potência de 10 que o valor 1, assim:

$$x^2 = \left(0{,}2 \times 10^{-7-1}\right) \times 10^{1}$$

$$x^2 = \left(0{,}00000002\right) \times 10^{1}$$

que o computador aproxima a mantissa para somente 7 dígitos decimais como:

$$\begin{aligned} & x^2 \cong 0 \times 10^{-1} \\ + & \\ & 1 = 0{,}1 \times 10^{1} \end{aligned}$$

Portanto:

$$y = x^2 + 1 = 0{,}1 \times 10^1$$

que demonstra a perda de algarismos significativos nessa operação matemática.

Exemplo 1.11) Calcule no computador a expressão $y = x - \text{sen } x$.
Solução:

i) sen $x \approx x$ quando x for um número pequeno, portanto isso acarretará que $y = 0$.
ii) utiliza-se, portanto, uma serie de Taylor para aproximar o sen x, como se segue.

$$y = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots \right)$$

$$y = \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

que apresentará em resultado diferente de zero, e mais próximo da solução exata, mesmo utilizando um pequeno número de termos na série de Taylor.

Duas definições importantes

1. Instabilidade numérica:

Um processo computacional é dito instável se pequenos erros cometidos durante as computações aumentam a medida que mais cálculos são feitos.

Exemplo 1.12) Verifique matematicamente se a equação

$$\begin{cases} x_0 = 1, \; x_1 = \dfrac{1}{5} \\ x_{n+1} = \dfrac{21}{5} x_n - \dfrac{4}{5} x_{n-1}, \; n \geq 1 \end{cases} \tag{1.63}$$

gera a seqüência

$$x_n = \left(\frac{1}{5} \right)^n \tag{1.64}$$

A seguir calcule no computador o valor de x_{15} com a Eq. (1.63) e com a Eq. (1.64). Comente os resultados.
Solução:

A) Prova por indução:

i) para $n = 0$ e $n = 1$, a Eq. (1.63) reproduz os valores da Eq. (1.64)

ii) assumindo que a Eq. (1.63) também reproduz os valores da Eq. (1.64) para $n \le m$. Fazendo $n = m$, pode-se obter

$$\begin{aligned} x_{m+1} &= \frac{21}{5} x_m - \frac{4}{5} x_{m-1} \\ &= \frac{21}{5}\left(\frac{1}{5}\right)^m - \frac{4}{5}\left(\frac{1}{5}\right)^{m-1} \\ &= \frac{21}{5^m} - \frac{4}{5^{m-1}} \\ &= \frac{21-20}{5^m} = \left(\frac{1}{5}\right)^m \end{aligned}$$

B) No computador, calcula-se x_{15}:

Eq. (1.63) $\rightarrow x_{15} = ...$ Eq. (1.57) $\rightarrow x_{15} = ...$

Comentário:

No cálculo realizado com a Eq. (1.63) observa-se um erro relativo maior do que...

Conclui-se, portanto, que a subtração dos dois termos da Eq. (1.63) gera erros de truncamento no computador no cálculo de cada termo, e que vão se acumulando à medida que se avança nos cálculos dos termos da seqüência. Isto caracteriza um processo computacional instável numericamente.

2. Condicionamento:

Trata-se da sensibilidade dos cálculos realizados no computador a pequenas variações em dados de entrada. Um problema é mal condicionado se pequenas variações em dados de entrada produzem grandes variações na saída.

Em alguns problemas um número de condicionamento pode ser definido. Se este número é grande, então o problema é mal condicionado.

1.12. PROBLEMAS PROPOSTOS

1.1) Considere a função definida em R pela seguinte fórmula

$$f(x) = \begin{cases} x^2 + ax + 1, \text{se } x \ge 0 \\ e^{-bx}, \text{se } x < 0 \end{cases}$$

com dois parâmetros $a, b \in R$

a) Mostre que f é contínua no seu domínio.

b) Determine os números a,b de forma que f tenha primeira derivada contínua e $f'(0) = 1$

c) Mostre que esta função têm valores positivos, é estritamente crescente e satisfaz a condição $\lim_{|x| \to +\infty} f(x) f(-x) = 0$

1.2) Calcule os limites seguintes

$$\alpha = \lim_{x \to +\infty} x \cdot \operatorname{sen}\left(\operatorname{arctg} \frac{1}{x}\right)$$

$$\beta = \lim_{x \to 0^+} \left(|\log x|\right)^x$$

1.3) Estude a função $g(x) = (1 + x^2)e^{-|x|}$, $x \in R$, considerando especialmente os aspectos seguintes: continuidade, diferenciabilidade, simetria, extremos locais e absolutos, inflexões e intervalos de monotonia. Esboce o gráfico de g.

1.4)Mostre que o erro de truncamento em $f'(x_0) \cong f\dfrac{(x_0 + h) - f(x_0)}{h}$ é $E_T = \dfrac{h}{2} f''(\xi)$, onde $x_0 \leq \xi \leq x_0 + h$

1.5) Estime a derivação para $f(x) = \operatorname{sen} x^2$ em $x = 0{,}5$ para:

a) $\dfrac{f(x+h) - f(x)}{h}$

b) $\dfrac{f(x+h) - f(x-h)}{2h}$

Qual é a aproximação mais precisa e porque?

1.6) Tendo $\lim_{x \to 0}\left(\dfrac{1 - e^x}{x}\right) = 1$

Mostre que isto é correto usando a Série de Taylor para e^x

1.7) Se a série de Taylor para o ln(x) for truncada após o termo envolvendo $(x-1)^{1000}$ e for então usada para computar o ln(2), qual o limite superior de erro? Sabe-se que para c = 1 a Série de Taylor pode ser expressa como:

$$f(x) = \sum_{K=0}^{n} \frac{1}{K!} f^{(K)}(c)(x-c)^K + E_n(x)$$

Onde o erro é dado por:

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

Note que, para c = 1, tem-se que: $1 < \xi < c$

Como $1 < \xi \Rightarrow \xi^{-(n+1)} < 1$

1.8) Para pequenos valores de x, qual a precisão da aproximação $\cos x \approx 1 - \dfrac{1}{2}x^2$? Para que faixa de valores esta aproximação dará resultados corretos até 3 casa decimais arredondadas?

1.9) Use o Teorema de Taylor com n = 2 (dois termos) para provar que a desigualdade $1 + x < e^x$ é válida para todos os números reais, exceto x = 0.

1.10) Ordens de convergência:

a) Prove que: se $x_n = o(\alpha_n)$, então $x_n = O(\alpha_n)$. Mostre que o inverso não é verdadeiro.
b) Prove que: se $x_n = o(\alpha_n)$ e $y_n = o(\alpha_n)$, então $x_n + y_n = o(\alpha_n)$.
c) Seja uma seqüência x_n definida indutivamente por $x_{n+1} = F(x_n)$. Suponha que $x_n \to x$ quando $n \to \infty$ e $F'(x) = 0$. Mostre que:

$$x_{n+2} - x_{n+1} = o(x_{n+1} - x_n)$$

1.11) Considere a seguinte função:

$$f(x) = \begin{cases} 5x^2 + x - 10, & se\ x \leq 2 \\ x^3 + 2x^2 + cx, & se\ x > 2 \end{cases}$$

(a) Determine o valor de *c* de modo que a função seja contínua em todo o domínio real.
(b) Calcule a derivada da função $f(x)$, empregando a definição de derivada, para todo o domínio real. Considere o valor de *c* calculado no item (a). A função $f'(x)$ resultante é contínua?

1.12) Mostre que as seguintes funções apresentam pelo menos uma raiz real nos intervalos dados. Apresente, graficamente, possíveis soluções (zeros) para as funções:

(a) $f(x) = x^2 + x - \cos(x), \ [0;1]$
(b) $f(x) = \ln(x) - x^2 + 3x, \ [0{,}2;0{,}8]$
(c) $f(x) = \cos(x) - \tan(x), \ [0;1]$ e $[2;3]$
(d) $f(x) = x^3 - 4x^2 + x - 4, \ [3{,}5;4{,}5]$

1.13) Obtenha o polinômio de Taylor de grau três $P_3(x)$, em torno de $x_0 = 0$, para as seguintes funções. Avalie, então, as funções, $f(x)$, e os respectivos polinômios de Taylor, $P_3(x)$, para $x = 0{,}1$ e $x = 1$:

(a) $f(x) = \sin(x)\cos(x)$
(b) $f(x) = xe^x$
(c) $f(x) = \tan(x)$

1.14) Empregando-se a aritmética de truncamento (e, em seguida, a aritmética de arredondamento), com 3 algarismos significativos, efetue os seguintes cálculos. Calcule, também, o erro absoluto e o erro relativo para cada caso:

(a) $\frac{1}{3} + \frac{1}{6}$

(b) $\frac{1}{3} \times \frac{3}{5}$

(c) $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{256} + \frac{1}{512} + \frac{1}{1024}$

(d) $\frac{1}{1024} + \frac{1}{512} + \frac{1}{256} + \frac{1}{128} + \frac{1}{64} + \frac{1}{32} + \frac{1}{16} + \frac{1}{8} + \frac{1}{4} + \frac{1}{2} + 1$

(e) $5 \times 10^4 + \sum_{i=1}^{500} 1$

(f) $\sum_{i=1}^{500} 1 + 5 \times 10^4$

1.15) Use aritmética de arredondamento, com quatro algarismos, para determinar a aproximação para as raízes das seguintes equações quadráticas. Empregue, também, formas alternativas à fórmula de Bhaskara padrão para encontrar tais raízes. Calcule os erros absolutos e relativos.

a) $1{,}002x^2 - 11{,}01x + 0{,}01265 = 0$

b) $2x^2 - 64{,}01x + 0{,}32 = 0$

1.16) A seqüência $\{F_n\}$, descrita por $F_0 = 1, F_1 = 1$ e $F_{n+2} = F_n + F_{n+1}$ se $n \geq 0$, é chamada seqüência Fibonacci. Seus termos ocorrem na natureza, como na disposição de pétalas arranjadas em espiral ou na construção da concha de caramujos. Considere a seqüência $\{x_n\}$, onde $x_n = F_{n+1}/F_n$. Supondo que $\lim_{n \to \infty} x_n = x$ exista, mostre que $x = (1 + \sqrt{5})/2$. Esse número é conhecido como razão áurea.

1.17) Determine o número de termos necessários para aproximar $\cos(x)$ até 8 algarismos significativos usando a aproximação por série de Maclaurin. Calcule a aproximação usando um valor de $x = 0{,}3\pi$. Escreva um código computacional para determinar o seu resultado.

1.18) A Lei de Stefan-Boltzmann pode ser utilizada para se fazer uma estimativa da taxa de radiação de calor q que deixa uma superfície, através da relação:

$$q = \varepsilon \sigma A T^4$$

onde q está em Watts; ε é a emissividade térmica,que caracteriza as propriedades de emissão de uma superfície (adimensional); σ é a constante de Stefan-Boltzmann (=5,67x10^{-8} W/m^{2}K^4); e T é a temperatura absoluta da superfície (em Kelvins). Determine o erro de q para uma placa de aço com $A = 0{,}20\ m^2;\ \varepsilon = 0{,}50; T = 800\ K \pm 20\ K$. Compare seus resultados com o erro exato. Repita os cálculos com $T = 800\ K \pm 40\ K$. Interprete os resultados.

1.19) Para qualquer número real $p \geq 1$, a fórmula

$$\|\vec{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

define as chamadas normas-p. Prove que para cada $\vec{x} \in \mathbb{R}^n$,

$\lim_{p\to\infty} \|\vec{x}\|_p = \|\vec{x}\|_\infty$

Sugestão: lembrar que o logaritmo do limite é o limite do logaritmo e vice-versa.

Projetos:

1. Escreva um código computacional para converter números informados em uma base binária, decimal ou octal para as outras duas bases.
2. Outro modo de se escrever um termo da seqüência de Fibonacci é empregando-se a fórmula de Binet, dada por:

$$F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$$

Implemente um código computacional empregando-se as duas fórmulas de cálculo para os termos da seqüência de Fibonacci (a primeira forma é dada no exercício 6). Compare os resultados obtidos para cada uma das duas fórmulas para a seqüência e, para o termo $n \geq 2$, calcule também a razão áurea (número Φ), dado por

$$\Phi = \frac{F_n}{F_{n-1}}$$

cujo valor, quando $n \to \infty$ é dado por $\Phi = (1+\sqrt{5})/2$. Pode-se demonstrar, matematicamente, que as séries são equivalentes. E computacionalmente, há diferenças? Se houver, justifique-as.