NOÇÕES DE CÁLCULO NUMÉRICO

NOÇÕES DE CÁLCULO NUMÉRICO

Ana Flora P. de Castro Humes
Inês S. Homem de Melo
Luzia Kazuko Yoshida
Wagner Tunis Martins

Instituto de Matemática e Estatística — USP

McGraw-Hill
São Paulo
Rua Tabapuã, 1.105, Itaim-Bibi
CEP 04533
(011) 881-8604 e (011) 881-8528

Rio de Janeiro • Lisboa • Porto • Bogotá • Buenos Aires • Guatemala • Madrid • México • New York • Panamá • San Juan • Santiago

Auckland • Hamburg • Kuala Lumpur • London • Milan • Montreal • New Delhi • Paris • Singapore • Sydney • Tokyo • Toronto

CIP-Brasil. Catalogação-na-Publicação
Câmara Brasileira do Livro, SP

N685 Noções de cálculo numérico / Ana Flora P. de Castro Humes . . . [et al.] . — São Paulo : McGraw-Hill do Brasil , 1984.

Bibliografia.

1. Cálculo numérico 2. Cálculo numérico – Problemas , exercícios etc. I. Humes , Ana Flora Pereira de Castro.

17. CDD-517
18. -511.7
17. -517.076
18. -511.7076

83-2168

Índices para catálogo sistemático:

1. Cálculo numérico : Matemática 517 (17.) 511 (18.)
2. Exercícios : Cálculo numérico : Matemática 517.076 (17.) 511.7076 (18.)

Sumário

Prefácio

Este texto originou-se de Notas de Aulas usadas pelos Professores da disciplina Cálculo Numérico, ministrada no primeiro ano de graduação dos Cursos de Engenharia, Matemática, Química e Geologia da Universidade de São Paulo. Estas Notas de Aulas foram elaboradas pelos seguintes Professores do Departamento de Matemática Aplicada do Instituto de Matemática e Estatística da Universidade de São Paulo: Ana Flora P. de Castro Humes, Anselmo Moraes Neto, Dirceu D. Salvetti, Edna Espírito Santo, George E. Freund, Inês S. Homem de Melo, José Roberto M. Bezerra, Luzia Kazuko Yoshida, Maria Elisabete B. Vivian, Miguel Haddad, Nami Kobayashi, Nestor de Mattos Cunha Junior, Priscila Goldenberg e Wagner Tunis Martins. A nós coube o trabalho de organizar e redigir este texto de uma forma didática.

Os tópicos aqui apresentados destinam-se a um curso de um semestre de Cálculo Numérico, para alunos do primeiro ano de graduação, supondo-se conhecimentos relativos a um semestre do curso de Cálculo. Os Apêndices constituem uma complementação teórica não necessária ao entendimento do texto e podem exigir conhecimentos adicionais de Cálculo.

No final de cada capítulo existe uma série de exercícios propostos que servem para facilitar o entendimento, complementar o texto ou

simples treino da utilização dos métodos apresentados. Inseridos no texto existem alguns exercícios que servem como complementação teórica.

As seções de cada capítulo são numeradas seqüencialmente a partir de 1. Dentro de cada seção os pontos importantes recebem uma numeração composta do número da seção seguido de um número seqüencial. Ao nos referirmos a um desses pontos, (2.18) por exemplo, fica subentendido tratar-se do ponto (2.18) do capítulo em questão. No caso de uma referência a um outro capítulo, mencionaremos o número do capítulo em algarismo romano, seguido do número do ponto (IV.2.18).

A vírgula decimal será representada por (.) e a operação de multiplicação, quando necessária, será indicada por (*) como utilizado normalmente nos computadores.

Ana Flora P. de Castro Humes
Inês S. Homem de Melo
Luzia Kazuko Yoshida
Wagner Tunis Martins
IME-USP

Capítulo 1
Introdução

1 – GENERALIDADES

A maioria dos problemas matemáticos é originária da necessidade de resolver problemas da Natureza. Isto porque fenômenos da Natureza podem ser descritos através do uso de modelos matemáticos.

Esquematicamente podemos representar as etapas da solução de um problema através da Fig. 1.1.

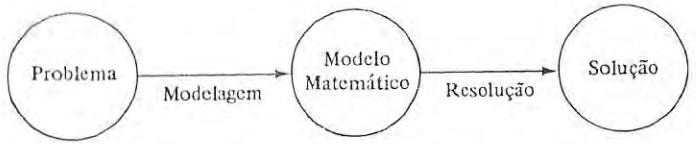


Fig. 1.1

Numa primeira etapa teríamos que obter um modelo matemático que representasse de maneira mais conveniente o problema específico que se quer estudar. Este modelo seria construído utilizando-se teorias físicas, químicas, econômicas etc. Para um mesmo problema poderíamos ter modelos matemáticos diferentes, dependendo do enfoque desejado. O modelo matemático, em geral, contém simplificações da realidade que nos levam a um problema matemático solúvel.

Construído o modelo matemático do problema, na segunda etapa procuraríamos encontrar sua solução.

Neste texto vamos estudar alguns métodos numéricos para obter a solução numérica de alguns tipos de modelos matemáticos. Por um método numérico entendemos um conjunto de regras escritas sob a forma de uma seqüência de operações elementares que levam a uma solução do problema.

Estudaremos, basicamente, métodos numéricos para resolver os seguintes problemas:

1) achar uma raiz real de uma equação;
2) resolver um sistema linear de equações;
3) aproximar uma função, dada na sua forma analítica ou tabelada, por uma outra função escolhida;
4) integrar uma função num intervalo.

A solução obtida pelo método numérico freqüentemente difere da solução do problema real. Vamos tentar enumerar algumas fontes de erro que levam a essa diferença.

Simplificações no Modelo Matemático

Como dissemos anteriormente, na construção do modelo matemático introduzimos uma série de simplificações; por exemplo, para calcular o período de um pêndulo, desprezamos sua massa.

Erro de Truncamento

Quando um modelo matemático envolve, por exemplo, a avaliação de uma série infinita, cometemos erro de truncamento ao avaliar esta série utilizando um número finito de termos. No processo de linearização de uma função, também estaremos cometendo um erro de truncamento pois a linearização corresponde a desenvolver a função em série de Taylor, tomando somente os termos lineares.

Erro de Arredondamento Cometido Durante os Cálculos

Na execução de um método numérico, são utilizados, normalmente, computadores, máquinas de calcular, réguas de cálculo etc. Todos esses instrumentos de auxílio trabalham com a representação dos números na forma decimal, com uma quantidade fixa de algarismos significativos. Entretanto, o resultado de uma operação aritmética qualquer não pode ser representado necessariamente desta forma, obrigando o seu arredondamento. O efeito do arredondamento pode ser significante quando temos um grande número de operações.

Erro nos Dados

Freqüentemente os dados necessários são obtidos através de medidas experimentais, portanto, sujeitos a imprecisões. Além disso, os erros nos dados podem ser ocasionados pela necessidade de se arredondar um dado de entrada, por exemplo, por este ser um número irracional, dízima periódica, ou mesmo não ser passível de representação no sistema adotado.

Neste capítulo estudaremos somente os erros de arredondamento. Nos demais capítulos, quando possível, estudaremos o erro de truncamento. Em geral, desprezando os erros de arredondamento, seremos capazes de delimitar o erro de truncamento cometido nos diversos métodos numéricos. Para alguns problemas forçaremos o aparecimento do erro de arredondamento para ilustrar sua propagação durante a utilização do método numérico; isto será feito utilizando uma representação dos números com poucos algarismos significativos.

2 – ERRO DE ARREDONDAMENTO

a) Aritmética de ponto flutuante

A representação usual dos números é feita utilizando um sistema de posicionamento na base 10, isto é, o número 327.302 significa

$$3 * 10^2 + 2 * 10^1 + 7 * 10^0 + 3 * 10^{-1} + 0 * 10^{-2} + 2 * 10^{-3}$$

Em geral trabalhamos na base 10, porém, qualquer número natural $B \geqslant 2$ pode ser utilizado como base. Se B é a base escolhida, o número

$$a_n a_{n-1} \ldots a_2 a_1 a_0 a_{-1} a_{-2} \ldots$$

representa, na base 10, o número

$$a_n B^n + a_{n-1}B^{n-1} + \ldots + a_2B^2 + a_1B + a_0 + a_{-1}B^{-1} + a_{-2}B^{-2} + \ldots$$

onde os coeficientes a_i são algarismos tais que $0 \leqslant a_i < B$.

Os computadores, por simplificação de funcionamento, operam normalmente na base 2, chamada base binária. Por exemplo, o número 1001.101 representa o número

$$1*2^3 + 0*2^2 + 0*2^1 + 1*2^0 + 1*2^{-1} + 0*2^{-2} + 1*2^{-3} = 9.625$$

na base decimal.

Os computadores utilizam também a seguinte normalização para representação dos números

$$\pm 0.d_1 d_2 \ldots d_t * B^e$$

onde $d_1 \neq 0$, $0 \leqslant d_i < B$, $i = 1, 2, \ldots, t$ e $m \leqslant e \leqslant M$. O número

$$0.d_1 d_2 \ldots d_t$$

é chamado de *mantissa*, B é a base, *e* o expoente, m o limite inferior do expoente, M o limite superior do expoente e t o número de algarismos significativos. Esta representação é chamada de *representação em ponto flutuante na base* B *com* t *algarismos significativos*.

Por exemplo, no computador IBM/370 temos t = 6, B = 16, m = −64 e M = 63; no computador Burroughs B6700 temos t = 13, B = 8, m = −63 e M = 63.

Vejamos dois exemplos de números binários normalizados usando t = 8:

a) $n_1 = 0.11100110 * 2^2$ cujo correspondente na base 10 é 3.59375;

b) $n_2 = 0.11100111 * 2^2$ cujo correspondente na base 10 é 3.609375.

Note que, no sistema de representação utilizado, n_1 e n_2 são dois números consecutivos, isto é, não podemos representar nenhum outro número que tenha valor intermediário. Desta forma, por exemplo, o número decimal 3.6 não tem representação exata. Este fato ilustra também o erro nos dados, devido ao arredondamento descrito na seção anterior.

Os números reais podem ser representados por uma reta contínua. Entretanto, em ponto flutuante só podemos representar pontos discretos da reta real. Por exemplo, consideremos o caso em que t = 2, B = 10, m = 0 e M = 4. Em cada intervalo

$$[10^i, 10^{i+1}), \quad -2 \leqslant i \leqslant 2,$$

podemos representar somente 90 números diferentes.

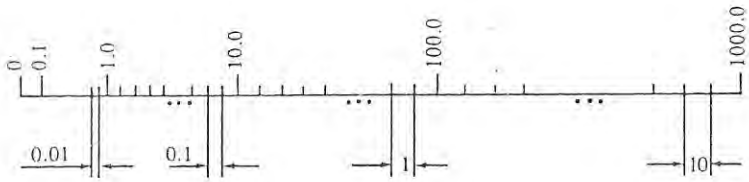


Fig. 1.2

No intervalo [0.10, 1.0) podemos representar os seguintes números:

$$\begin{array}{lll} 0.10*10^0 & 0.20*10^0 & \ldots\ 0.90*10^0 \\ 0.11*10^0 & 0.21*10^0 & \ldots\ 0.91*10^0 \\ \vdots & \vdots & \vdots \\ 0.19*10^0 & 0.29*10^0 & 0.99*10^0 \end{array}$$

b) Operações aritméticas em ponto flutuante

Ao contrário do que é válido para os números reais, as operações de adição e multiplicação em aritmética de ponto flutuante não são associativas nem distributivas. Isto se deve ao fato de, numa série de operações aritméticas, o arredondamento ser feito após cada operação. Por exemplo, para B = 10 e t = 3 temos:

b_1) $(4.26 + 9.24) + 5.04 = 13.5 + 5.04 = 18.5$

enquanto

$4.26 + (9.24 + 5.04) = 4.26 + 14.3 = 18.6$

$b_2)$ $(4210 - 4.99) - 0.02 = 4210 - 0.02 = 4210$

enquanto

$4210 - (4.99 + 0.02) = 4210 - 5.01 = 4200$

$b_3)$ $\left[\frac{0.123}{7.97}\right] * 84.9 = 0.0154 * 84.9 = 1.31$

enquanto

$$\left[\frac{0.123 * 84.9}{7.97}\right] = \frac{10.4}{7.97} = 1.30$$

$b_4)$ $15.9 * (4.99 + 0.02) = 15.9 * 5.01 = 79.7$

enquanto

$(15.9 * 4.99) + (15.9 * 0.02) = 79.3 + 0.318 = 79.6$

Assim, os erros de arredondamento introduzidos a cada operação efetuada influirão na solução obtida através do método numérico utilizado. Como conseqüência, métodos numéricos matematicamente equivalentes podem fornecer resultados diferentes.

3 – EXERCÍCIOS

1. Sabe-se que e^x pode ser escrito como

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

Num computador hipotético que usa a representação dos números em ponto flutuante com 5 algarismos significativos, na base 10, o valor de e^x é calculado através dos 4 primeiros termos da série acima. Neste caso está sendo cometido erro de arredondamento? E de truncamento? Pode existir erro nos dados?

2. Dado o número 2.47 na base 10, qual é a sua representação na base 2 usando 8 algarismos significativos? Essa representação é exata?

3. Considere a representação normalizada em ponto flutuante com 2 algarismos significativos, na base 2 e com o expoente da base variando de -2 até 2. Determine o conjunto dos números assim representáveis.

4. Represente os números que se seguem em ponto flutuante com 5 algarismos significativos usando a base 10. Se a representação não for exata, dê as duas representações:

– *truncada:* abandonando todos os significativos após o 5º algarismo.
– *arredondada:* somando 1/2 unidade ao 6º algarismo antes de truncar.

Por exemplo:

Número	*Representação truncada*	*Representação arredondada*
10/6	0.16666×10^1	0.16667×10^1

a) $\sqrt{2}$
b) 1/9
c) π
d) 1/7
e) 100/7

5. Calcule o valor das expressões que se seguem, utilizando aritmética de ponto flutuante com 3 algarismos significativos.

a) $19.3 + 1.07 - 10.3$
b) $27.2 * 1.3 - 327 * 0.00251$
c) $\frac{10.1 - 3.1 * 8.2}{14.1 + 7.09 * 3.2^2}$
d) $(365 + 0.7) + 0.5$
e) $365 + (0.7 + 0.5)$
f) $\frac{1}{3} + \frac{1}{3} + \frac{1}{3}$

6. Deseja-se calcular

$$S = \sum_{i=1}^{10} \frac{1}{2^i}$$

a) Represente cada termo da soma através da representação decimal exata e some.

b) Represente cada termo da soma em ponto flutuante com 3 algarismos significativos. Some da direita para a esquerda. Some da esquerda para a direita.

Capítulo 2
Zeros de Funções

Dada uma função real f definida e contínua num intervalo aberto I, chamaremos de zero ou raiz desta função em I a todo $\alpha \in I$ tal que $f(\alpha) = 0$.

Neste capítulo estudaremos o problema de como determinar uma raiz real de uma dada função f. Um caso particular deste problema é aquele em que a função f é um polinômio em x. Para polinômios de grau menor ou igual a 4, existem métodos diretos que fornecem todas as raízes. Foi mostrado por Galois (no século XIX) que tais métodos não podem ser obtidos para polinômios de grau maior ou igual a 5. Neste caso precisamos recorrer a processos numéricos. Estes processos numéricos podem ser utilizados na determinação de uma raiz real (se esta existir) de qualquer função f contínua dada. Vamos estudar processos numéricos iterativos. Exigiremos que a função f tenha tantas derivadas contínuas quantas forem necessárias.

1 – INTRODUÇÃO

Por processo iterativo entendemos um processo que calcula uma seqüência de aproximações $x_1, x_2, x_3, \ldots$ da solução desejada. O cálculo de uma nova aproximação é feito utilizando aproximações anteriores. Devem ser fornecidas as aproximações iniciais que o processo exigir.

Dizemos que o processo iterativo converge para $\bar{x}$, se a seqüência gerada $x_1, x_2, \ldots$ converge para $\bar{x}$*. Dizemos que o processo converge num número finito de passos se obtemos $\bar{x}$ aplicando o algoritmo um número finito de vezes.

Vamos estudar processos iterativos para calcular um valor aproximado para uma raiz real de uma função f dada. Os processos que estudaremos não fornecem raízes exatas de uma função em um número finito de passos, exceto em casos particulares. O que buscaremos então será um valor aproximado da raiz. Chamaremos este valor de $\tilde{x}$.

A diferença entre o valor exato da raiz $\bar{x}$ e o valor aproximado $\tilde{x}$ é chamada de erro. Como não podemos determinar o valor exato do erro, uma vez que a determinação depende de $\bar{x}$, o que faremos é delimitá-lo, ou seja, garantir que $|\bar{x} - \tilde{x}| \leqslant \delta$ para $\delta > 0$, previamente escolhido. Neste caso escreveremos $\bar{x} = \tilde{x} \pm \delta$ e diremos que $\tilde{x}$ tem precisão δ, onde δ é uma precisão prefixada.

Os métodos iterativos que veremos a seguir pressupõem um valor inicial dado. Desta maneira, a determinação de uma raiz real de uma função f será feita em duas etapas:

a) localização de uma raiz isolada, isto é, determinação de um intervalo que contenha exatamente uma raiz;

b) cálculo da raiz aproximada utilizando um método iterativo, com precisão prefixada ou não.

2 – LOCALIZAÇÃO DE RAÍZES ISOLADAS

Numa primeira fase a localização dos zeros será feita através de um gráfico ou de uma tabela da função f. Em qualquer caso, todas as informações sobre a função poderão ser utilizadas, como, por exemplo, pontos de máximo e de mínimo, inclinação e concavidade da curva.

Estudando o comportamento da função, teremos condições de determinar um intervalo I que contenha uma ou mais raízes da mesma. Para este intervalo I, esboçaremos o gráfico da função, determinando assim uma primeira aproximação da raiz.

* Uma seqüência $x_0, x_1, x_2 \ldots$, converge para $\bar{x}$ se dado $\varepsilon > 0$, $\exists$ N tal que qualquer que seja $n > N$, $|x_n - \bar{x}| < \varepsilon$. Neste caso temos que $\lim_{n \to \infty} x_n = \bar{x}$, o que também poderá ser indicado por $x_n \to \bar{x}$.

Na pesquisa de zeros reais através do gráfico, é muito útil o uso do Teorema de Bolzano*. Outro recurso que pode ser utilizado é transformar a equação $f(x) = 0$ numa equação equivalente da forma $g(x) = h(x)$ e buscar a intersecção do gráfico das duas funções.

Para os algoritmos que veremos a seguir, precisamos determinar uma aproximação inicial da raiz e um intervalo que contenha apenas esta raiz. Assim, se a partir do gráfico suspeitarmos da existência de duas ou mais raízes próximas ou coincidentes, é aconselhável um estudo detalhado do comportamento da função no intervalo I. Para isto devemos usar as técnicas do cálculo diferencial. Por exemplo, pode ser feita uma análise dos zeros da derivada da função, do seguinte modo. Se f' tiver um zero $\bar{z}$ nesse intervalo, não coincidente com o zero da função (caso coincida, $\bar{z}$ será um zero múltiplo), precisaremos estudar o comportamento da segunda derivada de f. Se $f(\bar{z}) * f''(\bar{z}) > 0$, não existem raízes reais nas vizinhanças de $\bar{z}$. Caso contrário existem duas raízes reais de f, próximas ao ponto $\bar{z}$ e separadas por este ponto. Nestas condições devemos restringir nosso intervalo, de maneira a não conter os dois zeros da função. Na Fig. 2.1 ilustramos as situações descritas acima.

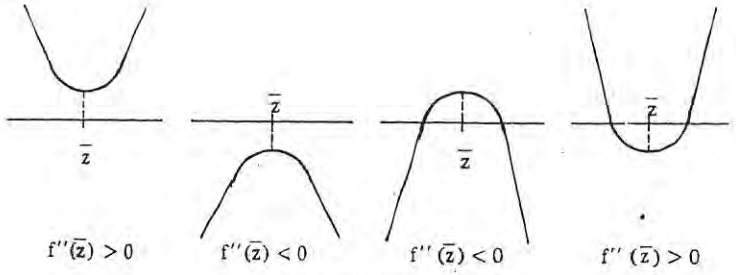


Fig. 2.1

Exemplo 2.1 Vamos pesquisar as raízes reais da função

$$f(x) = x\ell nx - 3.2$$

* Teorema de Bolzano: Se f for uma função contínua num intervalo [a, b] e trocar de sinal nos extremos desse intervalo, então existe pelo menos uma raiz real de f no intervalo [a, b].

A função $f(x) = x\ell nx - 3.2$ está definida somente para valores positivos de x. Tabelando-se f (x) nos pontos x = 1, 2, 3 e 4 escolhidos arbitrariamente, obtemos:

x	1	2	3	4
f (x)	−3.20	−1.81	0.10	2.36

Pelo Teorema de Bolzano concluímos que existe pelo menos uma raiz real no intervalo [2, 3].

Na Fig. 2.2 temos o esboço do gráfico de f (x) nesse intervalo.

x	f (x)
2.0	−1.81
2.2	−1.46
2.4	−1.10
2.6	−0.71
2.8	−0.32
3.0	0.10

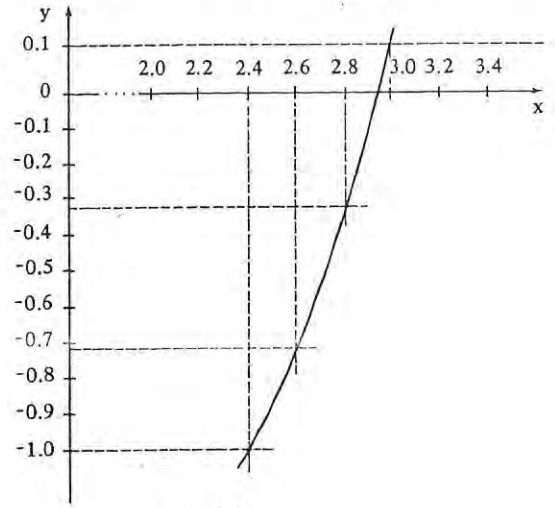


Fig. 2.2

Notamos que $f(x) < 0$, para $x \leqslant 1$ e que f (x) é monotônica estritamente crescente para $x > 1$. Desses fatos concluímos que existe uma única raiz real de f (x), isolada no intervalo [2. 8, 3. 0].

Exemplo 2.2 Vamos pesquisar agora as raízes reais da função $f(x) = 5 \log x - 2 + 0.4x$ utilizando uma técnica diferente da apresentada no exemplo anterior.

Vamos transformar a equação $5 \log x - 2 + 0.4x = 0$ na equação equivalente $5 \log x = 2 - 0.4x$. Na Fig. 2.3 construímos o gráfico das duas funções.

Propositadamente fizemos apenas um esboço grosseiro do gráfico, pois isto nos é suficiente.

Do esboço gráfico feito na Fig. 2.3 vemos que o intervalo [1, 2] contém uma raiz (no caso podemos verificar, analisando o comportamento das duas curvas, que esta raiz é única) e que essa raiz é aproximadamente 1.7.

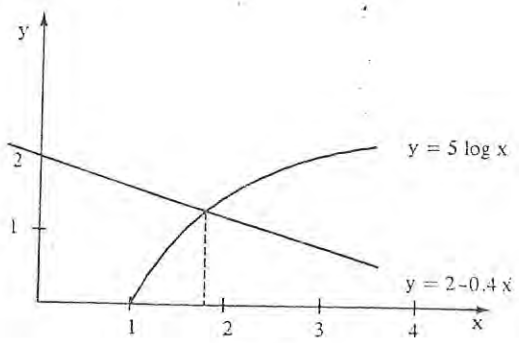


Fig. 2.3

3 – PROCESSOS ITERATIVOS

Nesta seção descreveremos processos iterativos para calcular um valor aproximado de uma raiz da função dada, supondo já encontrado um intervalo contendo somente esta raiz.

a) Método da dicotomia ou bissecção

O Teorema de Bolzano nos sugere um processo bastante simples para achar uma aproximação de uma raiz de uma função. Supondo que uma raiz da função f esteja isolada no interior do intervalo [a, b] e, portanto, f (a) * f (b) < 0, o processo consiste em dividir o intervalo dado ao meio e por aplicação do Teorema de Bolzano aos subintervalos

$$\left[a, \frac{a+b}{2}\right] \quad e \quad \left[\frac{a+b}{2}, b\right]$$

determinar qual deles contém a raiz. O processo é repetido para o novo subintervalo até que se obtenha uma precisão prefixada, isto é, o intervalo obtido seja menor ou igual a 2 vezes a precisão desejada.

Exemplo 3.1 Vamos determinar um valor aproximado da raiz quadrada de 5, com erro menor ou igual a 0.01.

Determinar $\sqrt{5}$ é equivalente a determinar o zero positivo da equação $x^2 - 5 = 0$. Sabemos que o intervalo [2, 3] contém esta raiz. Vamos aplicar o algoritmo da dicotomia. Em cada iteração i, i = 0, 1, 2, . . . , denotaremos por a_i e b_i os extremos inferior e superior, respectivamente, do intervalo que está sendo considerado, por $\tilde{x}_i$ o valor aproximado da raiz e por ε_i o erro máximo cometido na i-ésima iteração. Estes valores estão dispostos na Tabela 3.1. Inicialmente temos $f(a_0) = f(2.0) < 0$ e $f(b_0) = f(3.0) > 0$.

Tabela 3.1 Algoritmo da Dicotomia

i	a_i	b_i	$\tilde{x}_i = \frac{b_i + a_i}{2}$	$\varepsilon_i = \left\|\frac{b_i - a_i}{2}\right\|$	Sinal de $f(\tilde{x}_i)\, f(a_i)$
0	2.0	3.0	2.5	0.5	–
1	2.0	2.5	2.25	0.25	–
2	2.0	2.25	2.125	0.125	+
3	2.125	2.25	2.1875	0.0625	+
4	2.1875	2.25	2.21875	0.03125	+
5	2.21875	2.25	2.234375	0.015625	+
6	2.234375	2.25	2.2421875	0.0078125	

Portanto, $\sqrt{5} = 2.2421875 \pm 0.0078125$.

Observação: Na Tabela 3.1 os novos extremos do intervalo são a_i e $\tilde{x}_i$ se $f(\tilde{x}_i)\, f(a_i) < 0$ e $\tilde{x}_i$ e b_i, caso contrário.

É fácil mostrar que

$$\lim_{i \to \infty} \tilde{x}_i = \bar{x}$$

ou seja, que o método da bissecção converge e que, na i-ésima iteração, a resposta $\tilde{x}_i$ tem precisão ε_i. Entretanto, os erros de arredondamento podem comprometer não somente a precisão, mas também a convergência do processo para a solução exata. Para exemplificar este fato, vamos repetir o cálculo da raiz quadrada de 5 usando aritmética de ponto flutuante com 2 algarismos significativos. Obtemos os resultados da Tabela 3.2.

Tabela 3.2

i	a_i	b_i	$\tilde{x}_i = \frac{a_i + b_i}{2}$	$\varepsilon_i = \left\|\frac{b_i - a_i}{2}\right\|$	Sinal de $f(a)\, f(\tilde{x})$
0	2	3	2.5	0.5	−
1	2	2.5	2.3	0.25	−
2	2	2.3	2.2	0.15	+
3	2.2	2.3	2.3	0.05	−
4	2.2	2.3	2.3	0.05	−

Para $i = 3$ e $i = 4$ os valores da Tabela 3.2 se repetem devido aos erros de arredondamento. Assim, 2.3 é a melhor aproximação que conseguimos obter trabalhando em ponto flutuante com dois dígitos.

Exercício 3.1 Mostre que

$$\lim_{i \to \infty} |\tilde{x}_i - \bar{x}| = 0$$

sem considerar os possíveis erros de arredondamento.

b) Método das substituições ou aproximações sucessivas

Neste método a seqüência de aproximações da raiz $\bar{x}$ de uma função $f(x)$ é obtida através de uma relação de recorrência da forma

$$x_{n+1} = \phi(x_n), \quad n = 0, 1, 2, \ldots \tag{3.1}$$

onde x_0 é uma aproximação inicial de $\bar{x}$ e $\phi(x)$ é uma função que tem $\bar{x}$ como ponto fixo, isto é, $\bar{x} = \phi(\bar{x})$.

→ Mais adiante estudaremos condições que, se satisfeitas, garantem que a seqüência gerada por (3.1) converge para $\bar{x}$.

A primeira pergunta a ser respondida é: "dada uma função f com uma raiz $\bar{x}$, como determinar uma função ϕ que tenha $\bar{x}$ como ponto fixo?". Isto pode ser feito através de uma série de manipulações algébricas sobre a equação $f(x) = 0$, transformando-a numa equação equivalente da forma $x = \phi(x)$. Evidentemente nestas transformações há que se tomar os devidos cuidados para que $\phi(x)$ esteja definida em $\bar{x}$ e para que $\bar{x}$ pertença à imagem de ϕ. Como a raiz $\bar{x}$ nos é desconhecida, precisamos determinar um intervalo I que contenha $\bar{x}$ e que esteja contido tanto no domínio quanto na imagem de ϕ.

É necessário que a raiz $\bar{x}$ de $f(x)$ seja a única contida no intervalor I, caso contrário não teremos meios para discernir qual a raiz determinada.

Exemplo 3.2 A função $f(x) = x^2 + 0.96x - 2.08$ tem uma raiz positiva e outra negativa. As funções abaixo, obtidas a partir de $f(x) = 0$, têm como ponto fixo a raiz positiva de f.

$$x = \phi_1(x) = x^2 + 1.96x - 2.08$$

$$x = \phi_2(x) = \frac{2.08 - 0.96x}{x}$$

$$x = \phi_3(x) = \sqrt{2.08 - 0.96x}$$

$$x = \phi_4(x) = \frac{2.08}{x + 0.96}$$

$$x = \phi_5(x) = \frac{x^2 + 2.08}{2x + 0.96}$$

No próximo exemplo usaremos algumas dessas funções ϕ na tentativa de gerar seqüências aproximadoras de $\bar{x}$.

Exemplo 3.3 Sabendo que a função $f(x) = x^2 + 0.96x - 2.08$ tem uma raiz isolada no intervalo $I = [1, 2]$, vamos aplicar a relação de recorrência (3.1) a partir da aproximação inicial $x_0 = 1.2$, tomando $x = \phi_4(x)$. Utilizando aritmética de ponto flutuante com 4 algarismos significativos, obtemos os valores da Tabela 3.3.

Vemos que a seqüência $\{x_n\}$ converge para a raiz $\bar{x}$ ($\bar{x} = 1.04$). A repetição dos valores 1.041 e 1.039 deve-se à limitação do número de algarismos utilizados.

Tabela 3.3

i	x_n	$\phi(x_n) = \frac{2.08}{x_n + 0.96}$
0	1.2	0.9630
1	0.9630	1.082
2	1.082	1.019
3	1.019	1.051
4	1.051	1.034
5	1.034	1.043
6	1.043	1.038
7	1.038	1.041
8	1.041	1.039
9	1.039	1.041

Se tomarmos $x = \phi_1(x) = x^2 + 1.96x - 2.08$ com a mesma aproximação inicial $x_0 = 1.2$ e ainda aritmética de ponto flutuante com 4 algarismos, obteremos os valores da Tabela 3.4. Neste caso vemos claramente que não há convergência para a raiz.

Tabela 3.4

n	x_n	$\phi_1(x_n) = x_n^2 + 1.96x_n - 2.08$
0	1.2	1.712
1	1.712	4.207
2	4.207	23.87
3	23.87	614.5

O Exemplo 3.3 nos mostra que dependendo da transformação $x = \phi(x)$ escolhida a relação de recorrência (3.1) pode ou não nos fornecer uma seqüência convergente. O nosso problema agora é como determinar, *a priori*, quais transformações nos fornecerão seqüências convergentes.

O Teorema 3.1 a seguir estabelece condições suficientes para garantir a convergência do processo iterativo definido por (3.1). Vale notar que, como as condições são apenas suficientes, dada uma ϕ que não satisfaça estas condições, não podemos garantir que a seqüência $x_0, x_1, x_2, \ldots$ diverge.

Teorema 3.1 Seja $\bar{x}$ uma raiz de uma função f, isolada num intervalo $I = [a, b]$ e seja ϕ uma função tal que $\bar{x} = \phi(\bar{x})$. Se

a) ϕ e ϕ' são funções contínuas em I;

b) $k = \max_{x \in I} |\phi'(x)| < 1$;

c) $x_0 \in I$ e $x_{n+1} = \phi(x_n) \in I$ para $n = 0, 1, 2, \ldots$

então a seqüência $\{x_n\}$ converge para $\bar{x}$.

Prova: Para provar que $x_n \to \bar{x}$ basta mostrar que $|\bar{x} - x_n| \to 0$.

Vamos delimitar o erro $|\bar{x} - x_n|$. De a) e pelo Teorema do Valor Médio* do Cálculo Diferencial, temos

* Teorema do Valor Médio: Seja f uma função real definida e contínua num intervalo fechado [a, b], derivável em (a, b). Então existe $\xi \in (a, b)$ tal que $f(b) - f(a) = f'(\xi)(b - a)$.

$$\bar{x} - x_n = \phi(\bar{x}) - \phi(x_{n-1}) = \phi'(\xi_n)(\bar{x} - x_{n-1})$$

com ξ_n entre $\bar{x}$ e x_{n-1}.

Tomando o valor absoluto, vem:

$$|\bar{x} - x_n| = |\phi'(\xi_n)||\bar{x} - x_{n-1}| \tag{3.2}$$

Como $\bar{x} \in I$ e por c) $x_{n-1} \in I$, então $\xi_n \in I$ e por b)

$$|\phi'(\xi_n)| \leq k < 1$$

Substituindo em (3.2) obtemos:

$$|\bar{x} - x_n| \leq k\,|\bar{x} - x_{n-1}|$$

Como esta desigualdade vale para qualquer $n \geq 1$, temos

$$|\bar{x} - x_n| \leq k\,|\bar{x} - x_{n-1}| \leq k^2 |\bar{x} - x_{n-2}| \leq \ldots \leq k^n\,|\bar{x} - x_0|$$

Como $k < 1$, tomando o limite para $n \to \infty$, obtemos:

$$0 \leq \lim_{n\to\infty} |\bar{x} - x_n| \leq |\bar{x} - x_0| \lim_{n\to\infty} k^n = 0$$

Logo, $|\bar{x} - x_0| \to 0$, ou seja, $x_n \to \bar{x}$.

Vejamos como utilizar esse teorema para determinar uma raiz $\bar{x}$ de uma função f. Inicialmente determinamos um intervalo I, onde $\bar{x}$ esteja isolada, e uma função ϕ que tenha $\bar{x}$ como ponto fixo. Analisando ϕ e ϕ' podemos verificar se as condições a) e b) do Teorema 3.1 estão satisfeitas. Estas condições podem não estar satisfeitas pelo fato do intervalo I ter sido superdimensionado. Neste caso procuramos substituir o intervalo I por um subintervalo I′ satisfazendo as condições do teorema.

Na demonstração do Teorema 3.1, vimos que as condições a) e b) garantem que se $x_{n-1} \in I$ então $|\bar{x} - x_n| < |\bar{x} - x_{n-1}|$. Entretanto, isto não implica que $x_n \in I$ pois poderíamos ter uma situação como representada na Fig. 3.1, havendo necessidade de se verificar explicitamente a condição c).

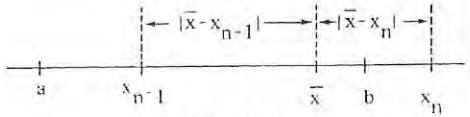


Fig. 3.1

Uma maneira simples para garantir que $x_n \in I = [a, b]$, $\forall n \geq 0$ é tomar como x_0 o extremo de I mais próximo de $\bar{x}$. Vamos mostrar que neste caso $x_1 = \phi(x_0) \in I$.

Supondo que a seja o extremo de I mais próximo de $\bar{x}$, temos:

$$|x_1 - \bar{x}| < |x_0 - \bar{x}| = |a - \bar{x}| \leq |b - \bar{x}|$$

Logo, $x_1 \in I$.

Exercício 3.2 Verifique que, se as condições a) e b) do Teorema 3.1 estão satisfeitas e se x_0 é o extremo de I mais próximo de $\bar{x}$, então $x_n \in I, \forall n$.

Exercício 3.3 Mostre que a condição c) do Teorema 3.1 pode ser substituída por

c') $\bar{x}$ é o ponto médio do intervalo I.

Na verdade, se temos um intervalo $I = [a, b]$, onde estão satisfeitas as condições a) e b) do Teorema 3.1, e se a estiver mais próximo de $\bar{x}$ do que b então, denotando $|a - \bar{x}|$ por r, temos que para qualquer $x_0 \in [a, \bar{x} + r]$ a hipótese c) do teorema é verificada. Mais ainda, para todo $I = [a, b]$ nas condições do Teorema 3.1, existe $I' \subset I$ tal que qualquer que seja $x_0 \in I'$ temos que $x_n \in I'$, $n \geq 1$.

Exercício 3.4 Com as mesmas hipóteses do Teorema 3.1, mostre que se $\phi'(x) > 0$, $x \in I$, então

$$x_1 > x_2 > x_3 > \ldots$$

ou

$$x_1 < x_2 < x_3 < \ldots$$

ou seja, a convergência da seqüência $\{x_n\}$ é monotônica. O que acontece quando $\phi'(x) < 0, \forall x \in I$?

(Sugestão: Utilize o Teorema do Valor Médio.) □

A determinação do extremo de $I = [a, b]$ mais próximo da raiz $\bar{x}$ pode ser feita da seguinte maneira. Suponhamos satisfeitas as hipóteses a) e b) do Teorema 3.1. Nestas condições, seja $\tilde{x} = (a + b)/2$ (ponto médio do intervalo I). Sabemos que $\phi(\tilde{x})$ está mais próximo de $\bar{x}$ do que $\tilde{x}$. Se $\tilde{x} < \phi(\tilde{x})$, então $\bar{x}$ está entre $\tilde{x}$ e b, ou seja, b é o

extremo de I mais próximo de $\bar{x}$ (veja Fig. 3.2). Analogamente, se $\tilde{x} > \phi(\tilde{x})$, então a é o extremo de I mais próximo de $\bar{x}$. É claro que se $\tilde{x} = \phi(\tilde{x})$ então $\tilde{x}$ é a raiz buscada.

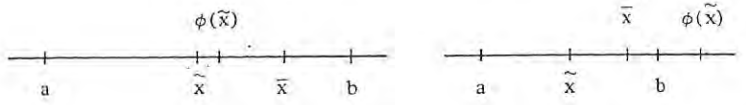


Fig. 3.2 – Casos em que b é o extremo mais próximo de $\bar{x}$.

Exercício 3.5 Utilizando o Teorema de Bolzano, determine uma outra maneira para achar o extremo de I mais próximo de $\bar{x}$.

Exemplo 3.4 Consideremos as transformações da equação $x^2 + 0.96x - 2.08 = 0$ feitas no Exemplo 3.2. Vejamos para algumas funções ϕ encontradas se estas satisfazem as hipóteses do Teorema 3.1.

Para

$$\phi_4(x) = \frac{2.08}{x + 0.96} \quad \text{e} \quad I = [1, 2],$$

temos que ϕ_4 bem como sua derivada

$$\phi_4'(x) = \frac{-2.08}{(x + 0.96)^2}$$

são contínuas em I.

Como $|\phi_4'(x)| = \dfrac{2.08}{(x + 0.96)^2}$ é estritamente decrescente, para calcularmos

$$\max_{x \in I} |\phi_4'(x)|$$

basta tomar $x = 1$, obtendo

$$k = \max_{x \in I} |\phi_4'(x)| = 0.5415 < 1$$

Para garantir a convergência precisamos tomar x_0 conveniente. Vamos tomar x_0 como o extremo de I mais próximo da raiz. Assim, como

$\tilde{x} = 1.5$ e $\phi_4(\tilde{x}) = 0.8455 < \tilde{x}$, a = 1.0 é o extremo de I mais próximo da raiz. Se tomarmos $x_0 = 1.0$, todas as hipóteses do Teorema 3.1 estarão satisfeitas e teremos garantida a convergência.

Observe que no Exemplo 3.3 tomamos $x_0 = 1.2$ e obtivemos $x_1 \notin I = [1, 2]$. Porém, como $|\phi_4'(x)| < 1$ para todo $x > 0.4822$, não houve problemas quanto à convergência.

Para $\phi_3(x) = \sqrt{2.08 - 0.96x}$ temos que

$$\phi_3'(x) = \frac{-0.96}{2\sqrt{2.08 - 0.96x}}$$

Ambas são contínuas em I = [1, 2], porém

$$\max_{x \in I} |\phi_3'(x)| > 1$$

Entretanto, se tomarmos o intervalo $I' = [1, 1.5]$, que ainda contém a raiz $\bar{x}$, teremos $\max_{x \in I'} |\phi_3'(x)| < 1$.

Para $\phi_1(x) = x^2 + 1.96x - 2.08$ que é contínua em I e $\phi_1'(x) = 2x + 1.96$ também contínua em I, temos

$$\max_{x \in I} |\phi_1'(x)| = 4.96 > 1$$

donde não podemos usar o Teorema 3.1 para tirar qualquer conclusão.

Corolário 3.1 Sejam $\phi(x)$, $\bar{x}$ e $k = \max_{x \in I} |\phi'(x)|$ satisfazendo as hipóteses do Teorema 3.1. Se $x_n = \phi(x_{n-1})$ então

$$|\bar{x} - x_n| \leqslant \frac{k}{1-k} |x_n - x_{n-1}|$$

Exercício 3.6 Prove o Corolário 3.1.

O Corolário 3.1 nos dá um limite superior para o erro de truncamento cometido na n-ésima iteração (x_n).

Exemplo 3.5 A função $f(x) = 2x - \cos x$ possui uma raiz real $\bar{x}$ isolada no intervalo $I = [0, \pi/4]$. Consideremos o processo iterativo definido por $x_{n+1} = \phi(x_n)$ com $\phi(x) = \frac{\cos x}{2}$. Seja x_0 o extremo

de I mais próximo de $\bar{x}$. Observemos que as hipóteses do Teorema 3.1 estão satisfeitas:

a) $\phi(x) = \frac{\cos x}{2}$ e $\phi'(x) = -\frac{\operatorname{sen} x}{2}$ são contínuas em I;

b) $k = \max_{x \in I} |\phi'(x)| = \frac{\operatorname{sen}(\pi/4)}{2} \leqslant 0.36 < 1.0$;

c) $x_0 \in I$ e $x_n \in I$, $\forall n$, pois x_0 é o extremo de I mais próximo de $\bar{x}$.

Pelo Teorema 3.1 concluímos que a seqüência $\{x_n\}$ convergirá para $\bar{x}$.

Vamos determinar o extremo de I mais próximo de $\bar{x}$ e calcular alguns valores fornecidos pelo processo iterativo $x_{n+1} = \phi(x_n)$, n = = 0, 1, 2, . . . Como o ponto médio de I é $\tilde{x} = \pi/8$ e

$$\phi(\tilde{x}) = \phi(\pi/8) = 0.4620 > \pi/8 = 0.3927$$

tomaremos $x_0 = \pi/4 = 0.7854$, obtendo os valores da Tabela 3.5.

Tabela 3.5

n	x_n	$\phi(x_n) = \frac{\cos x_n}{2}$
0	0.7854	0.3536
1	0.3536	0.4691
2	0.4691	0.4460
3	0.4460	0.4511
4	0.4511	0.4500
5	0.4500	0.4502
6	0.4502	0.4502

Usando o Corolário 3.1 vamos delimitar o erro de truncamento cometido na 6ª iteração.

$$|\bar{x} - x_6| \leqslant \frac{k}{1-k} |x_6 - x_5|$$

$$|\bar{x} - 0.4502| \leqslant \frac{0.36}{0.64} |0.4502 - 0.4500| \leqslant 0.0001125 < 0.0002$$

Logo, $\bar{x} = 0.4502 \pm 0.0002$.

Na Fig. 3.3 temos a interpretação geométrica do método das aproximações sucessivas para os casos em que a seqüência gerada é monotônica ou oscilante:

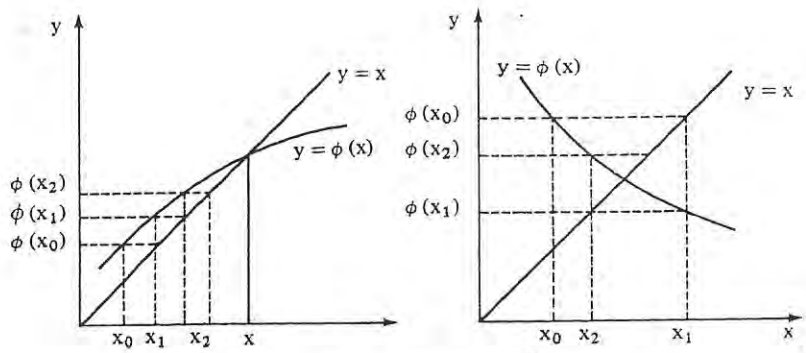


Fig. 3.3 a) seqüência monotonicamente crescente.
b) seqüência oscilante.

c) Método de Newton, Newton-Raphson ou das Tangentes

Este método é um caso particular do método das aproximações sucessivas. A idéia do método é construir uma função $\phi(x)$ para a qual exista um intervalo contendo a raiz onde $|\phi'(x)| < 1$.

Esta construção é feita impondo $\phi'(\bar{x}) = 0$. Como $\phi'(x)$ deve ser uma função contínua, existe sempre uma vizinhança I de $\bar{x}$ onde a hipótese b) do Teorema 3.1 ($k = \max_{x \in I} |\phi'(x)| < 1$) é satisfeita.

A forma mais geral de $x = \phi(x)$ equivalente a $f(x) = 0$ é dada por

$$x = x + A(x) f(x) = \phi(x) \tag{3.3}$$

onde $A(x)$ é uma função contínua qualquer, tal que $A(\bar{x}) \neq 0$. Vamos escolher $A(x)$ de forma a ter $\phi'(\bar{x}) = 0$. Derivando, temos:

$$\phi'(x) = 1 + A(x) f'(x) + A'(x) f(x)$$

Calculando no ponto $\bar{x}$ obtemos:

$$\phi'(\bar{x}) = 1 + A(\bar{x}) f'(\bar{x})$$

Supondo que $f'(\bar{x}) \neq 0$, como queremos $\phi'(\bar{x}) = 0$ devemos ter

$$A(\bar{x}) = -\frac{1}{f'(\bar{x})}$$

Uma escolha satisfatória para $A(x)$ será portanto:

$$A(x) = -\frac{1}{f'(x)}$$

De (3.3) obtemos:

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

Assim, o processo iterativo de Newton é definido por

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots * \tag{3.4}$$

Observe que como o intervalo I é determinado através da condição $|\phi'(x)| < 1$, para todo $x \in I$, em vez de determinarmos este intervalo, por razões de dificuldade de cálculo, escolheremos uma aproximação inicial x_0 suficientemente próxima da raiz. Aplicando a relação (3.4) para o x_0 escolhido, se a seqüência obtida aparentemente não estiver convergindo, abandona-se o processo e recomeça-se escolhendo outro valor inicial.

Na Fig. 3.4 temos a interpretação geométrica do método de Newton.

Da Fig. 3.4 vemos que

$$\operatorname{tg} \alpha = f'(x_n) = \frac{f(x_n)}{x_n - x_{n+1}}$$

donde

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

* Esta fórmula continua válida mesmo que $f'(\bar{x})$ seja zero, uma vez que $x_n \neq \bar{x}$.

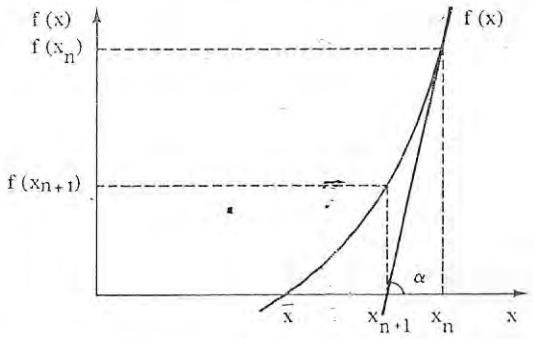


Fig. 3.4

O ponto x_{n+1} é obtido traçando-se a tangente à função f no ponto $(x_n, f(x_n))$. A intersecção da reta tangente com o eixo das abscissas fornece a nova aproximação x_{n+1}. Esta interpretação justifica o nome de método das tangentes.

Exemplo 3.6 A função $f(x) = 2x - \cos x$ possui uma raiz real $\bar{x}$ isolada no intervalo $I = [0, \pi/4]$. Vamos calcular um valor aproximado de $\bar{x}$ usando o método de Newton (os cálculos serão feitos usando aritmética de ponto flutuante com 4 algarismos significativos). Temos que

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

onde $f(x) = 2x - \cos x$ e $f'(x) = 2 + \operatorname{sen} x$. Logo

$$x_{n+1} = x_n - \frac{2x_n - \cos x_n}{2 + \operatorname{sen} x_n}, \qquad n = 0, 1, 2, \ldots$$

Tomando $x_0 = \pi/8 = 0.3927$ obtemos os valores da Tabela 3.6. Logo, $\bar{x} \simeq 0.4502$.

Dos exemplos 3.5 e 3.6 podemos ver que a convergência do método de Newton é muito mais rápida. Pode-se mostrar que a convergência do Método de Newton é quadrática, se $f'(\bar{x}) \neq 0$, e linear, caso contrário. Dizemos que a convergência é quadrática quando

$$|x_n - \bar{x}| \leqslant c\,|x_{n-1} - \bar{x}|^2$$

e é linear quando:

$$|x_n - \bar{x}| \leqslant c\,|x_{n-1} - \bar{x}|$$

onde c é uma constante.

Tabela 3.6

n	x_n	$\phi(x_n) = x_n - \dfrac{2x_n - \cos x_n}{2 + \operatorname{sen} x_n}$
0	0.3927	0.4508
1	0.4508	0.4502
2	0.4502	0.4502

Exercício 3.7 Mostre que a convergência do método de Newton é quadrática, se $f'(\bar{x}) \neq 0$, e linear, em caso contrário.

(Sugestão: Utilize o Teorema do Valor Médio e o fato que $|\phi'(\xi)| = |\phi'(\xi) - \phi'(\bar{x})|$.)

4 – CÁLCULO DE ZEROS DE UMA FUNÇÃO COM PRECISÃO PREFIXADA

No método da dicotomia sabemos, *a priori*, com que precisão é obtida a solução aproximada do zero da função.

Vejamos agora como podemos obter um zero com uma precisão δ prefixada, para o método das aproximações sucessivas. Neste caso, existe uma função $\phi(x)$ e um intervalo I onde as condições do Teorema 3.1 estão satisfeitas. Além disso, vamos supor que $\phi'(x)$ não muda de sinal no intervalo I. A seqüência de aproximações é obtida através da fórmula de recorrência:

$$x_{n+1} = \phi(x_n), \qquad n = 0, 1, 2, \ldots \tag{4.1}$$

Considerando as três primeiras aproximações podemos verificar que a convergência da seqüência obtida pode ser de dois tipos:

a) oscilante ($x_1 < x_2$ e $x_2 > x_3$ ou $x_1 > x_2$ e $x_2 < x_3$);

b) monotônica ($x_1 < x_2 < x_3$ ou $x_1 > x_2 > x_3$)*.

a) No caso da seqüência ser oscilante, sabemos que a raiz $\bar{x}$ estará sempre contida no intervalo $[x_n, x_{n+1}]$ (ou $[x_{n+1}, x_n]$). Neste caso, se tomarmos

$$\tilde{x} = \frac{x_n + x_{n+1}}{2}$$

como aproximação da raiz, o erro cometido será no máximo

$$\varepsilon_n = \frac{|x_n - x_{n-1}|}{2}$$

Portanto, deve-se repetir o processo iterativo até que $\varepsilon_n \leqslant \delta$.

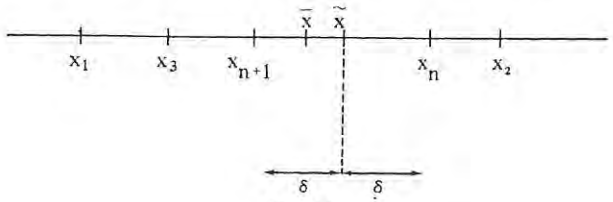


Fig. 4.1

b) No caso da seqüência ser monotônica, esta pode ser ainda decrescente ou crescente. Vamos estudar apenas o caso da seqüência ser decrescente pois o outro caso é análogo.

Como a seqüência é decrescente temos $x_n \geqslant \bar{x}$ e $x_n \geqslant \phi(x_n) \geqslant \bar{x}$.

Então, se tomarmos $x_n - 2\delta$ e calcularmos $\phi(x_n - 2\delta)$, acelerando assim o processo, se

$$x_n - 2\delta \geqslant \phi(x_n - 2\delta) \qquad (4.2)$$

sabemos que $\phi(x_n - 2\delta) \geqslant \bar{x}$. Portanto repetimos o processo iterativo com aceleração (ou seja, fazendo $x_{n+1} = \phi(x_n - 2\delta)$)até que a condição (4.2) não seja mais verificada.

* As desigualdades são estritas pois, caso contrário, não podemos classificar as seqüências obtidas baseados nas três primeiras aproximações.

Quando ocorrer:

$$x_n - 2\delta < \phi(x_n - 2\delta) \qquad (4.3)$$

então $\bar{x} \in [x_n - 2\delta, x_n]$ pois $\phi(x_n - 2\delta)$ está mais próximo de $\bar{x}$ do que $x_n - 2\delta$. Neste caso, se tomarmos $\tilde{x} = x_n - \delta$ temos que $|\tilde{x} - \bar{x}| \leqslant \delta$ (veja a Fig. 4.2).

É importante notar que não podemos garantir que tenhamos convergido para a raiz se a diferença entre dois valores consecutivos for menor ou igual a δ. Isto porque podemos ter um caso como ilustrado na Fig. 4.3 onde a convergência é muito lenta.

Tanto no caso da seqüência monotônica quanto no caso da seqüência oscilante, os erros de arredondamento poderão fazer com que a solução exata $\bar{x}$ não pertença ao intervalo $[\tilde{x} - \delta, \tilde{x} + \delta]$ obtido pelo procedimento acima. Entretanto, assumindo que somos capazes de discernir corretamente se $x_n < \phi(x_n)$ (ou $x_n > \phi(x_n)$), mesmo executando os cálculos com um número fixo de dígitos significativos em aritmética de ponto flutuante, então, fazer arredondamento a favor da segurança corrigirá os erros cometidos por arredondamento. Fazer

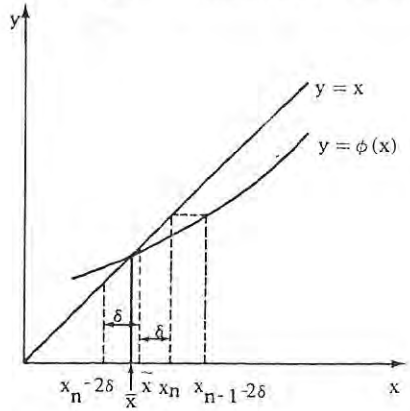


Fig. 4.2

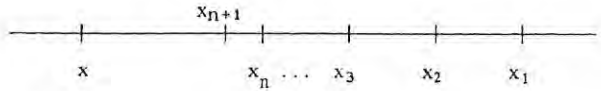


Fig. 4.3

o arredondamento a favor da segurança significa arredondar para mais se o valor de x_n obtido está à direita da raiz e truncar se o valor de x_n obtido está à esquerda da raiz.

Exemplo 4.1 Vamos calcular o zero da função $f(x) = x^4 - 2$, através do método de Newton, com precisão $\delta = 0.001$ e $x_0 = 1$.

$$\phi(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^4 - 2}{4x^3} = \frac{3x^4 + 2}{4x^3}$$

O tipo de seqüência será determinado pelas três aproximações iniciais:

$$x_1 = \phi(x_0) = 1.25$$
$$x_2 = \phi(x_1) = 1.193$$
$$x_3 = \phi(x_2) = 1.189$$

Portanto a seqüência obtida é monotônica decrescente. Como x_3 poderá vir a ser um dos extremos do intervalo que contém a raiz com precisão ± 0.001, vamos recalcular o valor de x_3, arredondando a favor da segurança

$$x_3 = \phi(x_2) = 1.190$$
$$(x_3 = 1.18934)\ .$$

Calculemos $\phi(x_3 - 2\delta) = \phi(1.190 - 0.002) = \phi(1.188) = 1.189$.

Como $\phi(1.188) > 1.188$, temos que $1.188 < \bar{x} < 1.190$.

Assim temos

$$\tilde{x} = x_3 - \delta = 1.189 \quad \text{e} \quad |\tilde{x} - \bar{x}| \leqslant 0.001$$

Observe que neste exemplo utilizamos quatro algarismos significativos.

5 – RESOLUÇÃO DE SISTEMAS NÃO LINEARES

Em problemas práticos é freqüente o caso em que temos que resolver um sistema de equações não lineares. Nestes casos, o problema pode ser apresentado na forma "determinar $x_1, x_2, \ldots, x_n$ tais que:

$$f_1(x_1, x_2, \ldots, x_n) = 0$$
$$f_2(x_1, x_2, \ldots, x_n) = 0$$
$$\vdots$$
$$f_n(x_1, x_2, \ldots, x_n) = 0$$"

Se usarmos a notação vetorial, onde $x = (x_1, x_2, \ldots, x_n)^T$ e $f = (f_1, f_2, \ldots, f_n)^T$ podemos reescrever o problema como: "determinar $x \in \mathbb{R}^n$ de modo que $f(x) = 0$".

Existem várias maneiras de se resolver este problema. Estudaremos aqui como aplicar o método de Newton a um sistema de equações não lineares.

No caso de $x \in \mathbb{R}$, podemos verificar que o método de Newton pode ser obtido da expansão linear em série de Taylor, da função $f(x)$, em torno do ponto x_n:

$$f(x_{n+1}) \simeq f(x_n) + (x_{n+1} - x_n) f'(x_n) \qquad (5.1)$$

Impondo $f(x_{n+1}) = 0$ em (5.1) temos:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Vamos fazer a mesma expansão para a função $f = (f_1, f_2, \ldots, f_n)$. Denotando a k-ésima aproximação do vetor $\bar{x} \in \mathbb{R}^n$ por $x^{(k)}$ temos:

$$f(x^{(k+1)}) \simeq f(x^{(k)}) + J(x^{(k)})(x^{(k+1)} - x^{(k)}) \qquad (5.2)$$

onde $J(x^{(k)})$ é o Jacobiano da função $f(x)$ cujo elemento $J(i,j)$ é obtido calculando-se $\partial f_i / \partial x_j$ no ponto

$$x^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})$$

Impondo $f(x^{(k+1)}) = 0$, obtemos

$$x^{(k+1)} = x^{(k)} - [J(x^{(k)})]^{-1} f(x^{(k)}) \qquad (5.3)$$

fórmula esta semelhante à fórmula (3.4) da seção 3 c.

Como a aplicação de (5.3) exige a inversão da matriz Jacobiana $n \times n$, $J(x)$, a cada iteração, podemos formular o problema de maneira equivalente: determinar $x^{(k+1)}$ resolvendo:

$$J(x^{(k)}) \cdot [x^{(k+1)} - x^{(k)}] = -f(x^{(k)}) \qquad (5.4)$$

que é um sistema linear pois $J(x^{(k)})$, $x^{(k)}$ e $f(x^{(k)})$ são conhecidos a cada iteração.

Pode-se mostrar que se $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ estiver suficientemente próximo da raiz $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$, este método convergirá.

Exemplo 5.1 Calcular (x, y) de modo que

$$x^2 + y^2 = 2 \quad \text{e} \quad x^2 - y^2 = 1,$$

tomando $x^{(0)} = y^{(0)} = 1$. Temos então

$$f_1(x, y) = x^2 + y^2 - 2 \quad \text{e} \quad f_2(x, y) = x^2 - y^2 - 1$$

$$J(x) = \begin{bmatrix} 2x & 2y \\ 2x & -2y \end{bmatrix} \qquad J^{-1}(x) = \begin{bmatrix} \frac{1}{4x} & \frac{1}{4x} \\ \frac{1}{4y} & -\frac{1}{4y} \end{bmatrix}$$

De (5.3) obtemos:

$$\begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \end{bmatrix} = \begin{bmatrix} x^{(k)} \\ y^{(k)} \end{bmatrix} - \begin{bmatrix} \frac{1}{4x^{(k)}} & \frac{1}{4x^{(k)}} \\ \frac{1}{4y^{(k)}} & -\frac{1}{4y^{(k)}} \end{bmatrix} \begin{bmatrix} [x^{(k)}]^2 + [y^{(k)}]^2 - 2 \\ [x^{(k)}]^2 + [y^{(k)}]^2 - 1 \end{bmatrix}$$

ou seja

$$\begin{bmatrix} x^{(k+1)} \\ y^{(k+1)} \end{bmatrix} = \begin{bmatrix} x^{(k)} \\ y^{(k)} \end{bmatrix} - \begin{bmatrix} \frac{[x^{(k)}]^2 + [y^{(k)}]^2 - 2}{4x^{(k)}} + \frac{[x^{(k)}]^2 - [y^{(k)}]^2 - 1}{4y^{(k)}} \\ \frac{[x^{(k)}]^2 + [y^{(k)}]^2 - 2}{4y^{(k)}} - \frac{[x^{(k)}]^2 - [y^{(k)}]^2 - 1}{4y^{(k)}} \end{bmatrix}$$

donde temos:

$$x^{(k+1)} = x^{(k)} - \frac{x^{(k)}}{2} + \frac{3}{4x^{(k)}} \quad y^{(k+1)} = y^{(k)} - \frac{y^{(k)}}{2} + \frac{1}{4x^{(k)}}$$

Para $x^{(0)} = y^{(0)} = 1$, obtemos:

$$x^{(1)} = 1.25 \qquad y^{(1)} = 0.75$$
$$x^{(2)} = 1.225 \qquad y^{(2)} = 0.7083$$
$$x^{(3)} = 1.2247 \qquad y^{(3)} = 0.7071$$

6 – EXERCÍCIOS

1. Dadas as funções

   a) $f(x) = \text{cosec}\, x - \text{tg}\, x$ c) $f(x) = x - 2.7 \ln x$

   b) $f(x) = e^{-x} - \ln x$ d) $f(x) = \ln x - \text{tg}\, hx$

   pesquisar a existência de raízes das funções acima e isolá-las em intervalos.

2. Dada a função $f(x) = \ln x - \text{tg}\, hx$, determine um intervalo que contenha uma única raiz positiva α de $f(x)$ e para este intervalo calcule o número de iterações que se executa no processo de dicotomia para obter m tal que $\alpha = m \pm 0.0001$.

3. (Método da Falsa Posição) Suponhamos que $f(x)$ seja contínua em $I = [a, b]$ e troca de sinal nos extremos de I onde um zero real α de f encontra-se isolado.

   Seja $\tilde{x}$ a intersecção do eixo 0x com a corda $\overline{PQ}$ onde $P = [a, f(a)]$ e $Q = [b, f(b)]$.

   Descreva um algoritmo para gerar uma seqüência de números que se aproxima de α baseado no seguinte procedimento:

   a) calcular $\tilde{x}$;

   b) substituir um dos extremos por $\tilde{x}$;

   c) usar como critério de parada a condição $|\tilde{x} - \tilde{x}\ (\text{anterior})| \leqslant \varepsilon$.

4. Aplique o método da Falsa Posição para calcular a raiz de $x^2 - 5 = 0$ com $\varepsilon = 0.01$.

   a) partindo do intervalo inicial [2, 2.5].

   b) partindo do intervalo inicial [2, 3].

   Podemos afirmar que a raiz exata $\bar{x} = \tilde{x} \pm \varepsilon$? Justifique.

5. Mostre que, se f for côncava ou convexa em I, o Método da Falsa Posição é um método de aproximações sucessivas. (Exiba a função $\phi(x)$.)

6. Mostre que se $\phi(x)$ é contínua com derivada contínua em I e se $\max_{x \in I} |\phi'(x)| < 1$ então $\phi(x)$ tem um único ponto fixo em I.

7. A equação $f(x) = x^2 + x - 1/4$ possui uma raiz real isolada no intervalo $[-0.5, 0.5]$. A seqüência produzida por $x_{n+1} = \phi(x_n) = -x_n^2 + 1/4$ será convergente para essa raiz?

8. Mostre que se $\phi(x)$ for contínua e se o processo $x_{n+1} = \phi(x_n)$ converge, então converge para um ponto fixo de $\phi(x)$.

9. Mostre que se $x_{n+1} = \phi(x_n) = x_n(2 - ax_n)$ converge, então o processo converge para 1/a ou 0.

   Que limites deve x_0 satisfazer para assegurar convergência? (Este é um método iterativo para fazer divisão ou determinar inversos.)

   Determine 1/7 com 4 algarismos significativos e $x_0 = 1$.

10. No método de aproximações sucessivas sejam m* o mínimo de $|\phi'(x)|$ para todo x que ocorrer nas iterações. Demonstre que se $m^* > 1$ o processo diverge.

11. Determine uma raiz de $f(x) = x^2 - 0.9$ resolvendo por $x = x^2 + x - 0.9$ pelo método de aproximações sucessivas com $x_0 = -1.0$. Poderia a outra raiz ser determinada por este método, usando a mesma função $\phi(x)$?

12. Dada a função $f(x) = x^2 - 2$ determine um intervalo I e condição inicial x_0 de modo que

$$x_{n+1} = \phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}$$

convirja para a raiz positiva de f(x).

13. Na Fig. 6.1, observe que $|\phi'(x)| < 1$ em I, $x_0 \in I$ e $\bar{x} \in I$. Entretanto o processo iterativo definido por

$$x_{n+1} = \phi(x_n)$$

não converge. Por quê?

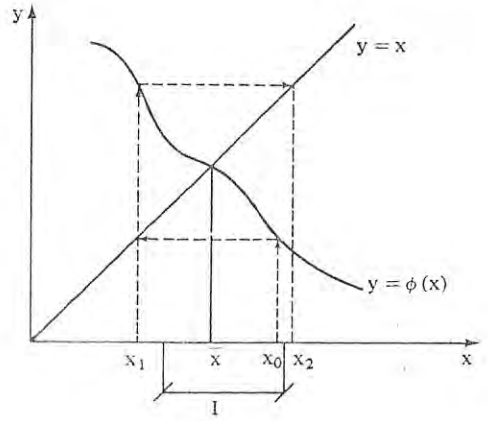


Fig. 6.1

14. Pelo Método de Newton, determinar $\bar{x}$ que minimiza a função $f(x) = 3x^4 - 8x^3 + 30x^2 - 120x + 1$ sabendo-se que este é um ponto crítico isolado no intervalo [1.5, 3.5].

15. Descreva o algoritmo correspondente ao Método de Newton Modificado para determinar zero de uma função. A construção das aproximações x_n, n = 1, 2, ... está ilustrada na Fig. 6.2. Resolva o Exercício 12 utilizando este método.

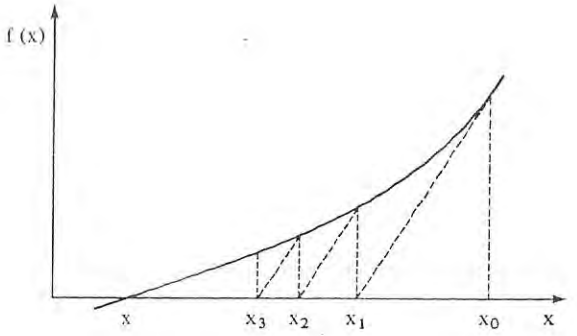


Fig. 6.2

16. Considere o seguinte problema: "dado um polinômio de grau n com coeficientes reais, $p(z)$, onde z é uma variável complexa, determinar uma raiz complexa de $p(z)$, se existir, ou seja, resolver a equação $p(z) = 0$".

Como $z = x + iy$, o polinômio $p(z)$ pode ser escrito na forma

$$p(z) = u(x, y) + iv(x, y)$$

Então, resolver a equação $p(z) = 0$ é equivalente a resolver o seguinte sistema de equações:

$$\begin{cases} u(x, y) = 0 \\ v(x, y) = 0 \end{cases}$$

Dada uma aproximação inicial $(x^{(0)}, y^{(0)})$ conveniente, podemos resolver este sistema pela extensão do Método de Newton (para sistemas não lineares).

Dado um polinômio $p(z) = u(x, y) + iv(x, y)$, tal que,

$$v(x, 0) = 0$$

podemos obter $u(x, y)$ e $v(x, y)$ expandindo $p(z)$ em série de Taylor em torno de z_0, isto é,

$$p(z) = p(z_0) + p'(z_0)(z - z_0) + p''(z_0)\frac{(z - z_0)^2}{2!} + \\ + p'''(z_0)\frac{(z - z_0)^3}{3!} + \ldots + \\ + \frac{p^{(n)}(z_0)}{n!}(z - z_0)^n$$

Tomando-se $z_0 = (x, 0)$ e $z = (x, y) = x + iy$, temos:

$$p(x + iy) = p(x) + p'(x)iy + p''(x)\frac{(iy)^2}{2!} + \\ + p'''(x)\frac{(iy)^3}{3!} + \ldots + \\ + \frac{p^{(n)}(x)}{n!}(iy)^n$$

Portanto,

$$u(x, y) = p(x) - \frac{p''(x)}{2!}y^2 + \frac{p^{(4)}(x)}{4!}y^4 + \ldots + \\ + (-1)^k \frac{p^{(2k)}(x)}{(2k)!}y^{2k}$$

e

$$v(x, y) = p'(x)y - \frac{p'''(x)}{3!}y^3 + \ldots + \\ + (-1)^k \frac{p^{(2k+1)}(x)}{(2k+1)!}y^{2k+1}$$

Observação: $p'(x)$, $p''(x)$, $\ldots$, $p^{(2k+1)}(x)$, significa calcular $p'(z)$, $p''(z)$, $\ldots$, $p^{(2k+1)}(z)$ no ponto $z_0 = (x, 0)$.

Determine uma aproximação da raiz complexa do polinômio $p(z) = z^2 - 2z + 3$, tomando como aproximação inicial $(x^{(0)}, y^{(0)}) = (1, 1)$, utilizando o método de Newton generalizado, isto é, resolvendo o sistema

$$\begin{cases} u(x, y) = 0 \\ v(x, y) = 0 \end{cases}$$

Faça 2 iterações.

Capítulo 3

Sistemas Lineares

Neste capítulo estudaremos alguns métodos para calcular a solução de sistemas de equações lineares. Apenas nos preocuparemos com sistemas quadrados, isto é, aqueles em que o número de equações é igual ao número de incógnitas.

Veremos os seguintes métodos:

a) método de eliminação de Gauss, com condensação pivotal e um processo iterativo para refinamento da solução;

b) método iterativo de Gauss-Siedel.

Noções básicas de álgebra matricial, como adição e multiplicação de matrizes, matriz inversa e identidade, determinante de uma matriz etc., serão supostas conhecidas pelo leitor.

1 – INTRODUÇÃO

Um sistema linear de ordem n

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \cdots\quad & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}$$

pode ser representado matricialmente por $Ax = b$, onde

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$ é a matriz dos coeficientes

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$ é o vetor das incógnitas e

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$ é o vetor dos termos independentes.

Em todo este texto, salvo menção em contrário, sempre indicaremos um sistema linear genérico de ordem n por $Ax = b$.

Para facilidade de notação usaremos indistintamente

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \text{ ou } x = (x_1, \ldots, x_n).$$

Um método para resolver sistemas lineares, provavelmente já conhecido do leitor, é o método de Cramer. Nele a solução do sistema $Ax = b$ é dada por

$$x_i = \frac{D_i}{D}, \quad i = 1, 2, \ldots, n$$

onde $D = \det(A)$ (determinante da matriz A);

$D_i = \det(A_i)$ e A_i é a matriz obtida de A substituindo a sua i-ésima coluna pelo vetor b dos termos independentes.

O determinante de uma matriz A de ordem n pode ser calculado através do desenvolvimento por linhas (regra de Laplace):

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

onde i é o índice de uma linha qualquer e A_{ij} é a matriz obtida de A retirando-se a i-ésima linha e a j-ésima coluna.

Observe que se $D = \det(A) \neq 0$ então o sistema $Ax = b$ tem uma única solução. Se $D = 0$ então podem ocorrer dois casos:

a) o sistema não possui solução (sistema inconsistente);

b) o sistema possui infinitas soluções (sistema indeterminado).

Por exemplo, no caso de um sistema linear de ordem 2, cada equação representa uma reta. Resolver o sistema significa determinar a intersecção das duas retas. Os três casos possíveis estão representados na Fig. 1.1.

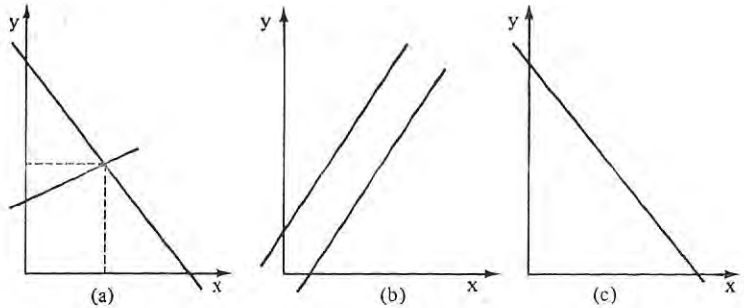


Fig. 1.1 a) retas concorrentes; b) retas paralelas e c) retas coincidentes.

Neste texto nos preocuparemos com sistemas lineares que tenham uma única solução.

A utilização do método de Cramer para sistemas lineares pode ser inviável, pois o número de operações aritméticas que devem ser efetuadas aumenta consideravelmente com pequeno aumento na ordem do sistema.

Para termos uma idéia desse número de operações, consideremos A_n e M_n, respectivamente o número de adições e o número de multiplicações necessárias para calcular o determinante de uma matriz de ordem n através do desenvolvimento por linhas. É fácil verificar que $A_1 = M_1 = 0$ e que

$$A_n = n - 1 + n * A_{n-1}$$

e

$$M_n = n + n * M_{n-1}$$

O número total de operações para calcular um determinante de ordem n é $\Delta_n = A_n + M_n$ e o número total de operações para resolver um sistema linear de ordem n pelo método de Cramer é $S_n = (n+1)\Delta_n + n$.

Na Tabela 1.1 estes números de operações são calculados para alguns valores de n.

Tabela 1.1 Número de Operações no Método de Cramer

n	M_n	A_n	Δ_n	S_n
2	2	1	3	11
3	9	5	14	59
4	40	23	63	319
5	205	119	224	1349
6	1236	719	1955	13691
7	8659	5039	13698	109591
8	62280	40319	109599	986399
9	560529	362879	923408	9234089
10	5605300	3628799	9234099	101575099
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
20	$\sim 3.758 * 10^{18}$	$\sim 2.433 * 10^{18}$	$\sim 6.191 * 10^{18}$	$\sim 1.3 * 10^{20}$

Estimando em 5 segundos o tempo necessário para realizar cada operação aritmética (com o auxílio de uma máquina de calcular), o tempo necessário para resolver um sistema de ordem 6 é

$$T = \frac{13691 * 5}{3600} \simeq 19 \text{ horas}$$

Para um sistema de ordem 8, o tempo necessário é de aproximadamente 57 dias e para um de ordem 9, de 534 dias. (Lembre-se que são 24 horas por dia, fazendo uma operação aritmética a cada 5 segundos!)

Mesmo considerando estas operações realizadas por um computador, a utilização deste método ainda é inviável. Estimando em $3.6\,\mu s = 3.6 * 10^{-6} s$ o tempo gasto pelo computador para executar cada operação aritmética*, o tempo gasto para resolver um sistema de ordem 20 é

$$T = \frac{1.3 * 10^{20} * 3.6 * 10^{-6}}{3600} = 1.3 * 10^{11} \text{horas} \simeq 1.5 * 10^{7} \text{anos}$$

Observemos que, na verdade, o tempo gasto pelo computador é bem maior, pois apenas foram consideradas as operações aritméticas.

Outro detalhe a observar é que, na solução de problemas reais, sistemas de ordem 20 e maiores ocorrem com freqüência. Veremos a seguir processos mais eficientes para resolver sistemas lineares.

2 – MÉTODO DE ELIMINAÇÃO DE GAUSS

a) Resolução de sistemas lineares triangulares

Um sistema linear da forma

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

onde $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$ é dito um *sistema triangular* (pela sua aparência).

A resolução do sistema (2.1) é feita utilizando a recorrência:

$$x_n = b_n / a_{nn}$$

$$\text{e} \quad x_i = \left[b_i - \sum_{j=i+1}^{n} a_{ij} x_j \right] \Big/ a_{ii}, \qquad i = n-1, n-2, \ldots, 1 \tag{2.2}$$

* No computador IBM370 modelo 158 os tempos das operações são: adição: 0,9 μs; multiplicação: 1,9 μs e divisão: 9,9 μs.

Veremos a seguir o método de eliminação de Gauss, que é uma maneira sistemática para transformar um sistema linear qualquer em um sistema triangular equivalente.

b) Descrição do método de eliminação de Gauss

O método de eliminação de Gauss consiste em transformar um sistema linear $Ax = b$ em um sistema triangular equivalente. Para isto utilizaremos a seguinte propriedade de Álgebra Linear.

Propriedade 2.1 A solução de um sistema linear não se altera se subtrairmos de uma equação outra equação do sistema multiplicada por uma constante.

Para maior facilidade, vamos descrever este método para um sistema de ordem 3. O mesmo processo pode ser aplicado para sistemas de qualquer ordem.

Seja o sistema

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{2.3}$$

Como vamos operar apenas sobre os coeficientes e termos independentes do sistema, este será representado através da matriz aumentada associada ao sistema:

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array}\right] \tag{2.4}$$

O nosso objetivo é obter um sistema triangular da forma (2.1) cuja matriz aumentada seja:

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a'_{22} & a'_{23} & b'_2 \\ 0 & 0 & a''_{33} & b''_3 \end{array}\right] \tag{2.5}$$

A primeira etapa do processo consiste em "zerar" os elementos da primeira coluna, abaixo da diagonal principal, supondo $a_{11} \neq 0$, obtendo a matriz aumentada

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a'_{22} & a'_{23} & b'_2 \\ 0 & a'_{32} & a'_{33} & b'_3 \end{array}\right] \tag{2.6}$$

A primeira linha de (2.4) é chamada de *linha pivô* e o elemento a_{11} é chamado *pivô* desta primeira etapa, devendo ser diferente de zero.

A transformação de (2.4) em (2.6) é feita utilizando a Propriedade 2.1 e executando os seguintes passos:

• Substituímos a segunda linha de (2.4) pelo resultado da subtração desta segunda linha pela linha pivô multiplicada por m_{21}, onde $m_{21} = a_{21}/a_{11}$. Desta maneira, obtemos como segunda linha da matriz aumentada

$$[(a_{21} - m_{21}a_{11}) \quad (a_{22} - m_{21}a_{12}) \quad (a_{23} - m_{21}a_{13}) \quad | \quad (b_2 - m_{21}b_1)]$$

Note que

$$a_{21} - m_{21}a_{11} = a_{21} - \frac{a_{21}}{a_{11}}a_{11} = 0$$

Façamos

$$a'_{22} = a_{22} - m_{21}a_{12}$$
$$a'_{23} = a_{23} - m_{21}a_{13}$$

e

$$b'_2 = b_2 - m_{21}b_1$$

Assim, a segunda linha fica sendo

$$[0 \quad a'_{22} \quad a'_{23} \quad | \quad b'_2]$$

• Analogamente, substituímos a terceira linha de (2.4) pelo resultado da subtração desta terceira linha pela linha pivô multiplicada por m_{31}, onde $m_{31} = a_{31}/a_{11}$.

Fazendo

$$a'_{32} = a_{32} - m_{31}a_{12}$$
$$a'_{33} = a_{33} - m_{31}a_{13}$$
$$b'_3 = b_3 - m_{31}b_1$$

obtemos (2.6). As constantes m_{21} e m_{31} serão chamadas *multiplicadores*.

Conforme veremos na seção 5, na resolução de um sistema linear, os multiplicadores podem ser utilizados mais de uma vez, devendo portanto ser preservados. Eles serão guardados nas posições dos correspondentes elementos zerados.

Desta forma, após a primeira etapa, obtemos

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ m_{21} & a'_{22} & a'_{23} & b'_2 \\ m_{31} & a'_{32} & a'_{33} & b'_3 \end{array}\right] \tag{2.7}$$

que corresponde à matriz aumentada dada em (2.6).

A segunda etapa consiste na repetição da primeira etapa ao sistema de ordem 2

$$\begin{bmatrix} a'_{22} & a'_{23} \\ a'_{32} & a'_{33} \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b'_2 \\ b'_3 \end{bmatrix} \tag{2.8}$$

Este sistema é obtido eliminando-se a primeira coluna e a primeira linha do sistema cuja matriz aumentada é dada por (2.7). Assim, "zerar" os elementos da 1ª coluna abaixo da diagonal principal do sistema (2.8) corresponde a "zerar" os elementos da 2ª coluna abaixo da diagonal principal em (2.7).

A linha pivô será agora a segunda linha de (2.7) e o pivô será o elemento a'_{22} que deve ser não nulo.

Calculamos o multiplicador:

$$m_{32} = a'_{32}/a'_{22}$$

Fazendo

$$a''_{33} = a'_{33} - m_{32}a'_{23}$$

e

$$b_3'' = b_3' - m_{32}b_2'$$

obtemos

$$\left[\begin{array}{lll|l} a_{11} & a_{12} & a_{13} & b_1 \\ m_{21} & a_{22}' & a_{23}' & b_2' \\ m_{31} & m_{32} & a_{33}'' & b_3'' \end{array}\right]$$

que é forma desejada.

Exemplo 2.1 Vamos resolver o sistema linear:

$$\begin{aligned} 3x_1 - 2x_2 + 5x_3 &= 20 \\ 6x_1 - 9x_2 + 12x_3 &= 51 \\ -5x_1 + 2x_3 &= 1 \end{aligned}$$

utilizando o método de eliminação de Gauss.

A matriz aumentada associada a este sistema é:

$$\left[\begin{array}{rrr|r} 3 & -2 & 5 & 20 \\ 6 & -9 & 12 & 51 \\ -5 & 0 & 2 & 1 \end{array}\right]$$

Neste e nos demais exemplos numéricos, os elementos que estão sendo calculados serão, em geral, representados por uma igualdade que tem de um lado as parcelas envolvidas no cálculo e, no outro o seu valor.

Aplicando o método de eliminação de Gauss, temos:

$$\left[\begin{array}{lll|l} 3 & -2 & 5 & 20 \\ \frac{6}{3}=2 & -9+4=-5 & 12-10=2 & 51-40=11 \\ -\frac{5}{3} & 0-\frac{10}{3}=-\frac{10}{3} & 2+\frac{25}{3}=\frac{31}{3} & 1+\frac{100}{3}=\frac{103}{3} \end{array}\right]$$

$$\left[\begin{array}{lll|l} 3 & -2 & 5 & 20 \\ 2 & -5 & 2 & 11 \\ -\frac{5}{3} & \frac{10}{(5*3)}=\frac{2}{3} & \frac{31}{3}-\frac{4}{3}=9 & \frac{103}{3}-\frac{22}{3}=27 \end{array}\right]$$

Resolvendo o sistema triangular obtido, temos:

$$\begin{aligned} x_3 &= 27/9 = 3 \\ x_2 &= (11 - 2*3)/-5 = -1 \\ x_1 &= (20 - 5*3 - 2*1)/3 = 1 \end{aligned}$$

Logo, a solução do sistema é $x = (1, -1, 3)$.

A aplicação do método de eliminação de Gauss a um sistema $Ax = b$ de ordem n é análoga. A matriz aumentada deste sistema é dada por:

$$\left[\begin{array}{lll|l} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & b_n \end{array}\right] \qquad (2.9)$$

Serão realizadas $n - 1$ etapas. Cada etapa i, $i = 1, \ldots, n-1$ consiste em "zerar" os elementos da i-ésima coluna, abaixo da diagonal principal.

Suponhamos que na aplicação do método já foram realizadas $i - 1$ etapas das $n - 1$ necessárias, obtendo-se a matriz

$$\left[\begin{array}{llllll|l} a_{11} & a_{12} & \cdots & a_{1,i-1} & a_{1i} & \cdots & a_{1n} & b_1 \\ m_{21} & a_{22} & \cdots & a_{2,i-1} & a_{2i} & \cdots & a_{2n} & b_2 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ m_{i1} & m_{i2} & \cdots & m_{i,i-1} & a_{ii} & \cdots & a_{in} & b_i \\ m_{i+1,1} & m_{i+1,2} & \cdots & m_{i+1,i-1} & a_{i+1,i} & \cdots & a_{i+1,n} & b_{i+1} \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot & \cdot \\ m_{n1} & m_{n2} & \cdots & m_{n,i-1} & a_{ni} & \cdots & a_{nn} & b_n \end{array}\right] \qquad (2.10)$$

Note que os elementos da matriz (2.10) não são necessariamente iguais aos elementos de (2.9). Usamos aqui os mesmos símbolos para facilidade de notação.

Na realização da i-ésima etapa, a linha i é a linha pivô e o elemento $a_{ii} \neq 0$ é o pivô desta etapa. Cada linha j, $j = i+1, i+2, \ldots, n$ é substituída pelo resultado da subtração desta linha pela linha pivô multiplicada por m_{ji}, onde $m_{ji} = a_{ji}/a_{ii}$, conforme Propriedade 2.1.

Após estas operações obtemos

$$\left[\begin{array}{cccccccc|c}
a_{11} & a_{12} & \cdots & a_{1,i-1} & a_{1,i} & a_{1,i+1} & \cdots & a_{1n} & b_1 \\
m_{21} & a_{22} & \cdots & a_{2,i-1} & a_{2,i} & a_{2,i+1} & \cdots & a_{2n} & b_2 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
m_{i1} & m_{i2} & \cdots & m_{i,i-1} & a_{ii} & a_{i,i+1} & \cdots & a_{in} & b_i \\
m_{i+1,1} & m_{i+1,2} & \cdots & m_{i+1,i-1} & m_{i+1,i} & a'_{i+1,i+1} & \cdots & a'_{i+1,n} & b'_{i+1} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
m_{n1} & m_{n2} & \cdots & m_{n,i-1} & m_{ni} & a'_{n,i+1} & \cdots & a'_{nn} & b'_n
\end{array}\right] \qquad (2.11)$$

onde:

$$a'_{jk} = a_{jk} - m_{ji}a_{ik} \qquad \text{e}$$

$$b'_j = b_j - m_{ji}b_i \qquad \text{para } j = i+1, i+2, \ldots, n$$

e $k = i+1, i+2, \ldots, n$.

Assim, na etapa i, repetimos a primeira etapa ao sistema de ordem $n - i + 1$ obtido eliminando-se as i primeiras linhas e i primeiras colunas da matriz (2.10).

Na descrição do método de eliminação de Gauss, sempre foi suposto que o elemento pivô fosse diferente de zero. Caso isto não ocorra, a linha pivô deve ser trocada com alguma das linhas que lhe estão abaixo, de maneira a obter um pivô não nulo. Os multiplicadores também devem ser trocados.

Note que uma troca de equações não altera a solução do sistema.

Exemplo 2.2 Vejamos a resolução do sistema linear

$$\begin{aligned}
2x_1 - x_2 + 3x_3 + 5x_4 &= -7 \\
6x_1 - 3x_2 + 12x_3 + 11x_4 &= 4 \\
4x_1 - x_2 + 10x_3 + 8x_4 &= 4 \\
-2x_2 - 8x_3 + 10x_4 &= -60
\end{aligned}$$

Para facilitar a identificação do elemento pivô, em cada etapa ele será assinalado por um círculo.

$$\left[\begin{array}{cccc|c}
\textcircled{2} & -1 & 3 & 5 & -7 \\
6 & -3 & 12 & 11 & 4 \\
4 & -1 & 10 & 8 & 4 \\
0 & -2 & -8 & 10 & -60
\end{array}\right]$$

$$\left[\begin{array}{cccc|c}
2 & -1 & 3 & 5 & -7 \\
\frac{6}{2}=3 & -3+3=\textcircled{0} & 12-9=3 & 11-15=-4 & 4+21=25 \\
\frac{4}{2}=2 & -1+2=1 & 10-6=4 & 8-10=-2 & 4+14=18 \\
0 & -2 & -8 & 10 & -60
\end{array}\right]$$

Como o pivô da segunda etapa é nulo, a segunda linha deve ser trocada com alguma linha que está abaixo dela e que tenha o pivô diferente de zero (neste caso escolhemos a terceira linha). Dessa maneira ficamos com a seguinte matriz:

$$\left[\begin{array}{cccc|c}
2 & -1 & 3 & 5 & -7 \\
2 & \textcircled{1} & 4 & -2 & 18 \\
3 & 0 & 3 & -4 & 25 \\
0 & -2 & -8 & 10 & -60
\end{array}\right]$$

Efetuando a 2ª etapa obtemos

$$\left[\begin{array}{cccc|c} 2 & -1 & 3 & 5 & -7 \\ 2 & \textcircled{1} & 4 & -2 & 18 \\ 3 & 0 & 3 & -4 & 25 \\ 0 & -\frac{2}{1}=-2 & -8+8=0 & 10-4=6 & -60+36=-24 \end{array}\right]$$

ou seja,

$$\left[\begin{array}{cccc|c} 2 & -1 & 3 & 5 & -7 \\ 2 & 1 & 4 & -2 & 18 \\ 3 & 0 & 3 & -4 & 25 \\ 0 & -2 & 0 & 6 & -24 \end{array}\right]$$

Resolvendo o sistema triangular temos:

$$x_4 = -24/6 = -4$$
$$x_3 = (25 - 16)/3 = 3$$
$$x_2 = (18 - (12 + 8))1 = 18 - 20 = -2$$
$$x_1 = (-7 - (2 + 9 - 20))/2 = (-7 + 9)/2 = 1$$

Assim, a solução do sistema é $x = (1, -2, 3, -4)$.

Verifica-se facilmente que, se durante a aplicação do método de eliminação de Gauss a um sistema $Ax = b$ ocorrer um pivô nulo e todos os elementos da coluna correspondente ao pivô, abaixo da diagonal principal, também forem nulos, o determinante da matriz A é igual a zero. Portanto, o sistema é indeterminado ou inconsistente e a aplicação do método de Gauss deve ser interrompida.

Exemplo 2.3 Vejamos o caso do seguinte sistema linear:

$$\begin{aligned} 3x_1 - 2x_2 + 5x_3 + x_4 &= 7 \\ -6x_1 + 4x_2 - 8x_3 + x_4 &= -9 \\ 9x_1 - 6x_2 + 19x_3 + x_4 &= 23 \\ 6x_1 - 4x_2 - 6x_3 + 15x_4 &= 11 \end{aligned}$$

A matriz aumentada deste sistema é dada por:

$$\left[\begin{array}{cccc|c} 3 & -2 & 5 & 1 & 7 \\ -6 & 4 & -8 & 1 & -9 \\ 9 & -6 & 19 & 1 & 23 \\ 6 & -4 & -6 & 15 & 11 \end{array}\right]$$

$$\left[\begin{array}{cccc|c} 3 & -2 & 5 & 1 & 7 \\ -\frac{6}{3}=-2 & 4-4=\textcircled{0} & -8+10=2 & 1+2=3 & -9+14=5 \\ \frac{9}{3}=3 & -6+6=0 & 19-15=4 & 1-3=-2 & 23-21=2 \\ \frac{6}{3}=2 & -4+4=0 & -6-10=-16 & 15-2=13 & 11-14=-3 \end{array}\right]$$

Como não existe possibilidade de troca de linhas para que o segundo pivô seja diferente de zero, a triangularização deve ser interrompida pois o sistema é indeterminado ou inconsistente ($\det(A) = 0$).

3 – ANÁLISE DO NÚMERO DE OPERAÇÕES DO MÉTODO DE GAUSS

Na seção 1 concluímos que resolver sistemas lineares utilizando o método de Cramer pode tornar-se inviável se os determinantes forem calculados através do desenvolvimento por linhas. Esta conclusão foi obtida a partir de uma análise do número de operações aritméticas que teriam que ser realizadas.

Vamos fazer uma análise semelhante para o método de eliminação de Gauss.

Sendo n a ordem do sistema $Ax = b$ e lembrando que a cada etapa do método de eliminação de Gauss repetimos as mesmas operações para um sistema de ordem menor, temos:

a) operações para "triangularizar" o sistema:

A_n = número de adições e subtrações ($A_n = n(n-1) + A_{n-1}$)

M_n = número de multiplicações ($M_n = n(n-1) + M_{n-1}$)

D_n = número de divisões ($D_n = (n-1) + D_{n-1}$);

b) operações necessárias para resolver o sistema triangular:

A'_n = número de adições e subtrações ($A'_n = (n-1) + A'_{n-1}$)

M'_n = número de multiplicações ($M'_n = (n-1) + M'_{n-1}$)

D'_n = número de divisões ($D'_n = n$);

c) total de operações necessárias para resolver o sistema:

$$G_n = A_n + M_n + D_n + A'_n + M'_n + D'_n = \frac{4n^3 + 9n^2 - 7n}{6}$$

É fácil verificar que os valores iniciais para essas fórmulas de recorrência são:

$$A_1 = M_1 = D_1 = 0$$
$$A'_1 = M'_1 = 0$$
$$D'_1 = 1$$

Na Tabela 3.1 temos um cálculo desses números de operações para alguns valores de n.

Tabela 3.1 Número de Operações para o Método de Eliminação de Gauss

n	$A_n = M_n$	$D_n = A'_n = M'_n$	D'_n	G_n
2	2	1	2	9
3	8	3	3	28
4	20	6	4	62
5	40	10	5	115
6	70	15	6	191
7	112	21	7	294
8	168	28	8	428
9	240	36	9	597
10	330	45	10	805
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
20	2660	190	20	5910

A viabilidade da utilização do método de Gauss fica evidente a partir destes números. Por exemplo, para resolver um sistema de ordem 20 utilizando um computador, o tempo gasto será:

$$T \simeq 5910 * 3.6 \times 10^{-6} = 0.021276\text{s}$$

4 – CONDENSAÇÃO PIVOTAL

Em todos os exemplos de aplicações do método de eliminação de Gauss vistos nas seções anteriores trabalhamos com frações. Em todos eles foram obtidas as soluções exatas, já que o método é exato e não ocorreram erros de arredondamento nas operações aritméticas.

No entanto, quando trabalhamos com calculadoras ou computadores, utilizamos representação dos números em ponto flutuante. Os inevitáveis erros de arredondamento que ocorrem nas operações com esses números podem comprometer seriamente a solução obtida.

Exemplo 4.1 O sistema

$$\begin{aligned} x_1 + 4x_2 + 52x_3 &= 57 \\ 27x_1 + 110x_2 - 3x_3 &= 134 \\ 22x_1 + 2x_2 + 14x_3 &= 38 \end{aligned}$$

tem solução exata x = (1, 1, 1). Resolvendo-o pelo método de eliminação de Gauss, trabalhando com aritmética de ponto flutuante com três algarismos significativos, temos:

$$\left[\begin{array}{ccc|c} \textcircled{1} & 4 & 52 & 57 \\ 27 & 110 & -3 & 134 \\ 22 & 2 & 14 & 38 \end{array}\right]$$

$$\left[\begin{array}{ccc|c} 1 & 4 & 52 & 57 \\ \frac{27}{1} = 27 & 110 - 108 = \textcircled{2} & -3 - 1400 = -1400 & 134 - 1540 = -1410 \\ \frac{22}{1} = 22 & 2 - 88 = -86 & 14 - 1140 = -1130 & 38 - 1250 = -1210 \end{array}\right]$$

$$\begin{bmatrix} 1 & 4 & 52 & 57 \\ 27 & 2 & -1400 & -1410 \\ 22 & -\frac{86}{2} = -43 & -1130 - 60200 = -61300 & -1210 - 60600 = -61800 \end{bmatrix}$$

$$x_3 = -61800/-61300 = 1.01$$
$$x_2 = (-1410 + 1410)/2 = 0.0$$
$$x_1 = (57 - 52.5 - 0.0)/1 = 4.5$$

A solução obtida é x = (4.5, 0.0, 1.01).

Neste exemplo a grande diferença entre a solução exata e a solução obtida não se deve ao método utilizado, mas aos erros de arredondamento que ocorrem durante o processo.

Para minimizar a influência destes erros de arredondamento, vamos utilizar a *condensação pivotal*, que consiste em escolher o elemento pivô no início de cada etapa i, i = 1, 2, . . . , n − 1 como sendo o número de maior valor absoluto dentre os elementos da i-ésima coluna que estão na diagonal principal ou abaixo dela. Se o elemento escolhido para pivô não estiver na diagonal principal, devemos trocar sua linha com a linha de índice i. Os multiplicadores que estão colocados nos lugares dos elementos zerados também devem ser trocados.

Exemplo 4.2 Vamos resolver novamente o sistema linear do Exemplo 4.1, desta vez utilizando condensação pivotal e trabalhando em aritmética de ponto flutuante com três algarismos significativos.

$$\begin{bmatrix} 1 & 4 & 52 & 57 \\ 27 & 110 & -3 & 134 \\ 22 & 2 & 14 & 38 \end{bmatrix}$$

O pivô é o número 27; portanto devemos trocar a primeira linha com a segunda.

$$\begin{bmatrix} \textcircled{27} & 110 & -3 & 134 \\ 1 & 4 & 52 & 57 \\ 22 & 2 & 14 & 38 \end{bmatrix}$$
$$p_1 = 2$$

$p_1 = 2$ indica que na 1ª etapa houve uma troca da 1ª com a 2ª linha.

$$\begin{bmatrix} 27 & 110 & -3 & 134 \\ \frac{1}{27} = 0.0370 & 4 - 4.07 = -0.07 & 52 + 0.111 = 52.1 & 57 - 4.96 = 52 \\ \frac{22}{27} = 0.815 & 2 - 89.7 = -87.7 & 14 + 2.45 = 16.5 & 38 - 109 = -71 \end{bmatrix}$$
$$p_1 = 2$$

O pivô é o número −87.7; portanto devemos trocar a segunda linha com a terceira linha.

$$\begin{bmatrix} 27 & 110 & -3 & 134 \\ 0.815 & -87.7 & 16.5 & -71 \\ 0.0370 & -0.07 & 52.1 & 52 \end{bmatrix}$$
$$p_1 = 2 \quad p_2 = 3$$

$p_2 = 3$ indica que na 2ª etapa houve uma troca da 2ª com a 3ª linha.

$$\begin{bmatrix} 27 & 110 & -3 & 134 \\ 0.815 & -87.7 & 16.5 & -71 \\ 0.0370 & \frac{-0.07}{-87.7} = 0.000798 & 52.1 - 0.0132 = 52.1 & 52 + 0.0567 = 52.1 \end{bmatrix}$$
$$p_1 = 2 \quad p_2 = 3$$

Resolvendo o sistema triangular temos:

$$x_3 = 52.1/52.1 = 1.0$$
$$x_2 = (-71 - 16.5)/-87.7 = 0.998$$
$$x_1 = (134 + 3 - 110*0.998)/27 = 1.0$$

A solução obtida é x = (1.0, 0.998, 1.0).

Sempre que utilizamos condensação pivotal, anotaremos $p_i = k$ abaixo de cada coluna i, i = 1, 2, . . . , n − 1 para indicar que na i-ésima etapa houve uma troca entre a linha i e a linha k. Esta informação será utilizada na seção seguinte.

Não é possível provar que utilizando esse processo sempre se obtém soluções mais precisas. Existem casos em que a utilização da condensação pivotal leva a piores resultados.

Exemplo 4.3 O sistema

$$\begin{bmatrix} 270 & 1100 & -30 \\ 0.22 & 0.02 & 0.14 \\ 1000 & 4000 & 52000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1340 \\ 0.38 \\ 57000 \end{bmatrix}$$

que tem solução exata x = (1, 1, 1) é equivalente ao sistema do Exemplo 4.1 com a ordem das equações trocadas e cada equação multiplicada por uma constante diferente.

Na solução deste sistema com condensação pivotal reconstituímos (exceto o fator de escala) o sistema do Exemplo 4.1 obtendo conseqüentemente a mesma solução x = (4.5, 0.0, 1.01).

Se resolvermos sem condensação pivotal, repetiremos a seqüência do Exemplo 4.2, obtendo a mesma solução x = (1.0, 0.998, 1.0).

Apesar disso a experiência mostra que, na solução da maioria dos sistemas que ocorrem em problemas reais, obtemos resultados mais precisos utilizando condensação pivotal.

5 – REFINAMENTO DA SOLUÇÃO

Já vimos que os erros de arredondamento que ocorrem quando se resolve um sistema linear pelo método de eliminação de Gauss, trabalhando com números em ponto flutuante, podem comprometer o resultado obtido. Mesmo utilizando condensação pivotal, não podemos assegurar que a solução obtida é exata. No Exemplo 4.2, utilizando condensação pivotal obtivemos

$$x = (1.0, 0.998, 1.0)$$

enquanto a solução exata do sistema é x = (1.0, 1.0, 1.0).

Seja $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ a solução de um sistema linear Ax = b, obtida pela aplicação do método de eliminação de Gauss, trabalhando com números em ponto flutuante. Devido aos erros de arredondamento ocorridos no processo de obtenção de $\tilde{x}$, esta solução pode ser diferente da solução exata do sistema, $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$.

Veremos agora um processo de refinamento da solução que tem por objetivo obter, a partir de $\tilde{x}$, uma melhor aproximação da solução exata $\bar{x}$.

Se $\tilde{x}$ for uma solução aproximada de $\bar{x}$, então existe $\bar{c} = (\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n)$, tal que $\bar{x} = \tilde{x} + \bar{c}$. Chamaremos o vetor $\bar{c}$ de vetor correção ou simplesmente correção.

Para determiná-lo, temos que:

$$A\bar{x} = b \tag{5.1}$$

então

$$A(\tilde{x} + \bar{c}) = b$$

$$A\bar{c} = b - A\tilde{x} \tag{5.2}$$

Portanto podemos obter a solução $\bar{c}$, resolvendo o sistema linear Ac = r, onde r é o resíduo definido por

$$r = b - A\tilde{x}$$

O sistema Ac = r pode ser resolvido de forma análoga ao sistema Ax = b.

Seja $\tilde{c}$ a solução obtida para o sistema Ac = r. Esta solução provavelmente não é exata, pois os erros de arredondamento que influíram no cálculo de $\tilde{x}$ também influíram no cálculo de $\tilde{c}$. No entanto, se $\tilde{x}$ é uma aproximação razoável de $\bar{x}$, é de se esperar que $\tilde{c}$ seja também uma aproximação razoável de $\bar{c}$, e que a solução refinada

$$\tilde{\tilde{x}} = \tilde{x} + \tilde{c}$$

seja uma melhor aproximação da solução exata $\bar{x}$.

Naturalmente este processo pode ser repetido, o que leva ao seguinte método iterativo de refinamento da solução:

Solução inicial: $x^{(0)} = \tilde{x}$

Primeira iteração: $r^{(0)} = b - Ax^{(0)}$

$c^{(0)}$ = solução de $Ac = r^{(0)}$

$x^{(1)} = x^{(0)} + c^{(0)}$ = primeira solução refinada

Segunda iteração: $r^{(1)} = b - Ax^{(1)}$

$c^{(1)}$ = solução de $Ac = r^{(1)}$

$x^{(2)} = x^{(1)} + c^{(1)}$ = segunda solução refinada

e assim por diante.

Como o resíduo envolve subtração de vetores cujos elementos

têm a mesma ordem de grandeza, para que ele seja significativo, os cálculos necessários para a sua obtenção devem ser feitos em dupla precisão. Para trabalharmos em dupla precisão, os números em ponto flutuante terão na mantissa o dobro de dígitos da precisão simples.

Apenas as operações envolvidas nos cálculos dos resíduos devem ser feitas em dupla precisão. Os valores obtidos para os resíduos devem ser convertidos para precisão simples.

Considerando a solução $x^{(0)} = (1.0, 0.998, 1.0)$ do sistema do Exemplo 4.2, se o resíduo $r^{(0)} = b - Ax^{(0)}$ fosse calculado em precisão simples, teríamos:

$$r^{(0)} = \begin{bmatrix} 57 \\ 134 \\ 38 \end{bmatrix} - \begin{bmatrix} 1 + 3.99 + 52 = 57 \\ 27 + 110 - 3 = 134 \\ 22 + 2 + 14 = 38 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

Com isso verificamos que não é possível refinar $x^{(0)}$ já que a solução de $Ac = r^{(0)}$ é $c^{(0)} = (0.0, 0.0, 0.0)$. Entretanto, se utilizarmos dupla precisão no cálculo de $r^{(0)}$, a solução poderá ser refinada.

Exemplo 5.1 Vamos refinar a solução $\bar{x} = (1.0, 0.998, 1.0)$ do sistema:

$$\begin{aligned} x_1 + \quad 4x_2 + 52x_3 &= 57 \\ 27x_1 + 110x_2 - \cdot\, 3x_3 &= 134 \\ 22x_1 + \quad 2x_2 + 14x_3 &= 38 \end{aligned}$$

calculada no Exemplo 4.2.

Solução inicial: $x^{(0)} = (1.0, 0.998, 1.0)$

Primeira iteração:

Cálculo do resíduo $r^{(0)}$ (com dupla precisão):

$$r^{(0)} = b - Ax^{(0)} = \begin{bmatrix} 57 \\ 134 \\ 38 \end{bmatrix} - \begin{bmatrix} 1 & 4 & 52 \\ 27 & 110 & -3 \\ 22 & 2 & 14 \end{bmatrix} \begin{bmatrix} 1.0 \\ 0.998 \\ 1.0 \end{bmatrix} =$$

$$= \begin{bmatrix} 57 \\ 134 \\ 38 \end{bmatrix} - \begin{bmatrix} 1 + 3.992 + 52 = 56.992 \\ 27 + 109.78 - 3.0 = 133.78 \\ 22 + 1.996 + 14 = 37.996 \end{bmatrix} = \begin{bmatrix} 0.008 \\ 0.22 \\ 0.004 \end{bmatrix}$$

Caso os elementos de $r^{(0)}$ não estivessem em ponto flutuante com três algarismos significativos, eles deveriam ser convertidos para esta precisão, que é a precisão simples.

Cálculo da correção $c^{(0)}$:

O sistema a ser resolvido é $Ac = r^{(0)}$.

Como a triangularização da matriz A pelo método de eliminação de Gauss com condensação pivotal independe do vetor dos termos independentes, ao aplicarmos este método ao sistema $Ac = r^{(0)}$, obteremos a mesma matriz triangular do Exemplo 4.2, quando foi calculada a solução inicial $x^{(0)}$. Com isso, para obter o sistema $Ac = r^{(0)}$ triangularizado, basta repetir para o vetor $r^{(0)}$ as operações que foram efetuadas com os elementos do vetor dos termos independentes do sistema original. Esta foi a razão que nos levou a guardar os multiplicadores e as trocas efetuadas na triangularização.

A matriz triangular obtida no Exemplo 4.2 é:

$$\begin{bmatrix} 27 & 110 & -3 \\ 0.815 & -87.7 & 16.5 \\ 0.0370 & 0.000798 & 52.1 \end{bmatrix}$$

$p_1 = 2 \qquad p_2 = 3$

Para repetir com o vetor $r^{(0)}$ as operações feitas na triangularização da matriz A, devemos inicialmente efetuar todas as trocas de linhas indicadas por p_i e em seguida fazer os cálculos usando os multiplicadores. Assim temos:

$$r^{(0)} = \begin{bmatrix} 0.008 \\ 0.22 \\ 0.008 \end{bmatrix} \xrightarrow{p_1 = 2} \begin{bmatrix} 0.22 \\ 0.008 \\ 0.004 \end{bmatrix} \xrightarrow{p_2 = 3} \begin{bmatrix} 0.22 \\ 0.004 \\ 0.008 \end{bmatrix} \Longrightarrow$$

$$\Longrightarrow \begin{bmatrix} 0.22 \\ 0.004 - 0.815 * 0.22 = -0.175 \\ 0.008 - 0.00370 * 0.22 = -0.00014 \end{bmatrix} \Longrightarrow$$

$$\Longrightarrow \begin{bmatrix} 0.22 \\ -0.175 \\ -0.00014 + 0.000798 * 0.175 = 0.0 \end{bmatrix}$$

O sistema triangular a ser resolvido é:

$$\left[\begin{array}{ccc|c} 27 & 110 & -3 & 0.22 \\ & -87.7 & 16.5 & -0.175 \\ & & 52.1 & 0.0 \end{array}\right]$$

Portanto

$$c_3 = 0.0$$
$$c_2 = -0.175/-87.7 = 0.002$$
$$c_1 = (0.22 - 110 * 0.002)/27 = 0.0$$

isto é, $c^{(0)} = (0.0, 0.002, 0.0)$.

A primeira solução refinada é:

$$x^{(1)} = x^{(0)} + c^{(0)} = \begin{bmatrix} 1.0 \\ 0.998 \\ 1.0 \end{bmatrix} + \begin{bmatrix} 0.0 \\ 0.002 \\ 0.0 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \\ 1.0 \end{bmatrix}$$

que sabemos ser a solução exata do sistema.

Neste último exemplo pudemos interromper o processo iterativo de refinamento após a primeira iteração, pois obtivemos a solução exata do sistema $x = (1.0, 1.0, 1.0)$, já conhecida. Entretanto, isto nem sempre ocorre. Além dos cálculos envolvidos no processo estarem sujeitos a erros de arredondamento, existem casos em que a solução do sistema envolve dízimas periódicas, que não podem ser

representadas de maneira exata, utilizando notação de ponto flutuante com um número fixo de dígitos.

Por essas razões precisamos estabelecer um critério de parada para o processo iterativo de refinamento. O ideal seria obter uma relação entre cada solução refinada $x^{(k)}$ e a solução exata, o que entretanto não é possível pois $\bar{x}$ não é conhecida.

O que faremos então é comparar duas soluções refinadas consecutivas, $x^{(k-1)}$ e $x^{(k)}$, usando a variação relativa:

$$\mathrm{Var}^{(k)} = \max \{v_1^{(k)}, v_2^{(k)}, \ldots, v_n^{(k)}\}$$

onde

$$v_i^{(k)} = \begin{cases} \left| \dfrac{x_i^{(k)} - x_i^{(k-1)}}{x_i^{(k)}} \right| & \text{se } x_i^{(k)} \neq 0 \\ 0 & \text{se } x_i^{(k)} = 0 = x_i^{(k-1)} \\ 1 & \text{se } x_i^{(k)} = 0 \text{ e } x_i^{(k-1)} \neq 0 \end{cases}$$

O processo iterativo será interrompido quando $\mathrm{Var}^{(k)} \leqslant \varepsilon$, para uma dada precisão $\varepsilon > 0$.

Devemos notar que apesar de

$$x_i^{(k)} = x_i^{(k-1)} + c_i^{(k-1)}$$

a diferença $x_i^{(k)} - x_i^{(k-1)}$ nem sempre é igual a $c_i^{(k-1)}$, devido aos erros de arredondamento.

Exemplo 5.2 Vamos resolver o sistema abaixo e refinar a solução obtida, trabalhando em ponto flutuante com dois algarismos significativos, até atingir a precisão $\varepsilon = 0.01$:

$$0.47 x_1 + 0.27 x_2 = 0.20$$
$$0.89 x_1 + 0.52 x_2 = 0.37$$

Cálculo da solução inicial:

$$\left[\begin{array}{cc|c} 0.89 & 0.52 & 0.37 \\ 0.53 & 0.27 - 0.28 = -0.01 & 0\,20 - 0.20 = 0.0 \end{array}\right]$$

$p_1 = 2$

Solução inicial:

$$x^{(0)} = \begin{bmatrix} 0.42 \\ 0.0 \end{bmatrix}$$

Refinamento da solução

Primeira iteração:

$$r^{(0)} = \begin{bmatrix} 0.20 \\ 0.37 \end{bmatrix} - \begin{bmatrix} 0.47 & 0.27 \\ 0.89 & 0.52 \end{bmatrix} \begin{bmatrix} 0.42 \\ 0.0 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.20 \\ 0.37 \end{bmatrix} - \begin{bmatrix} 0.1974 + 0.0 = 0.1974 \\ 0.3738 + 0.0 = 0.3738 \end{bmatrix} = \begin{bmatrix} 0.0026 \\ -0.0038 \end{bmatrix}$$

A solução do sistema

$$\begin{bmatrix} 0.47 & 0.27 \\ 0.89 & 0.52 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0.0026 \\ -0.0038 \end{bmatrix}$$

é

$$c^{(0)} = \begin{bmatrix} 0.27 \\ -0.46 \end{bmatrix}$$

Solução refinada:

$$x^{(1)} = \begin{bmatrix} 0.42 \\ 0.0 \end{bmatrix} + \begin{bmatrix} 0.27 \\ -0.46 \end{bmatrix} = \begin{bmatrix} 0.69 \\ -0.46 \end{bmatrix}$$

Cálculo da variação relativa:

$$\text{Var}^{(1)} = \max\left\{\left|\frac{0.69 - 0.42}{0.69}\right|, \left|\frac{-0.46 - 0}{-0.46}\right|\right\} = \max\{0.39, 1.0\} =$$

$$= 1.0 > \varepsilon$$

Segunda iteração:

$$r^{(1)} = \begin{bmatrix} 0.20 \\ 0.37 \end{bmatrix} - \begin{bmatrix} 0.3243 - 0.1242 = 0.2001 \\ 0.6141 - 0.2392 = 0.3749 \end{bmatrix} - \begin{bmatrix} -0.0001 \\ -0.0049 \end{bmatrix}$$

A solução do sistema $Ac = r^{(1)}$ é:

$$c^{(1)} = \begin{bmatrix} 0.15 \\ -0.25 \end{bmatrix}$$

Solução refinada:

$$x^{(2)} = \begin{bmatrix} 0.69 \\ -0.46 \end{bmatrix} + \begin{bmatrix} 0.15 \\ -0.25 \end{bmatrix} = \begin{bmatrix} 0.84 \\ -0.71 \end{bmatrix}$$

$$\text{Var}^{(2)} = \max\left\{\left|\frac{0.84 - 0.69}{0.84}\right|, \left|\frac{-0.71 + 0.46}{-0.71}\right|\right\} =$$

$$= \max\{0.18, 0.35\} = 0.35 > \varepsilon$$

Nas próximas iterações obteremos os resultados da Tabela 5.1. Os detalhes ficam a cargo do leitor.

Tabela 5.1

k	$r^{(k-1)}$	$c^{(k-1)}$	$x^{(k)}$	$\text{Var}^{(k)}$
3	(−0.0031, −0.0084)	(0.073, −0.14)	(0.91, −0.85)	0.16
4	(0.0018, 0.0021)	(0.043, −0.07)	(0.95, −0.92)	0.076
5	(0.0019, 0.0029)	(0.027, −0.04)	(0.98, −0.96)	0.042
6	(−0.0014, −0.0030)	(0.0079, −0.02)	(0.99, −0.98)	0.02
7	(−0.0007, −0.0015)	(0.0042, −0.01)	(0.99, −0.99)	$0.01 = \varepsilon$

A solução refinada do sistema é $x = (0.99, -0.99)$. Vale notar que, apesar da seqüência dos $x^{(k)}$ estar nitidamente se aproximando da solução exata do sistema $\bar{x} = (1.0, -1.0)$, os valores dos resíduos ora crescem e ora decrescem, não servindo como medida para verificar a precisão de cada aproximação.

6 – SISTEMAS MAL CONDICIONADOS

Alguns sistemas são muito sensíveis a pequenas alterações nos seus dados (coeficientes e termos independentes), isto é, uma pequena alteração nos dados pode provocar uma grande alteração na solução.

Por exemplo, o sistema

$$\begin{aligned} x_1 + 0.98\,x_2 &= 4.95 \\ x_1 + \quad x_2 &= 5.0 \end{aligned}$$

tem solução exata $\bar{x} = (2.5, 2.5)$.

Com uma alteração de 0.01 no coeficiente 0.98 obtemos o sistema:

$$\begin{aligned} x_1 + 0.99\,x_2 &= 4.95 \\ x_1 + \quad x_2 &= 5.0 \end{aligned}$$

cuja solução exata é $\bar{x} = (0.0, 5.0)$.

Assim vemos que uma alteração de aproximadamente 1% nos dados acarreta uma alteração de 100% na solução. Um sistema deste tipo é dito mal condicionado.

Quando resolvemos um sistema $Ax = b$, os erros de arredondamento fazem com que a solução obtida $\tilde{x}$ possa ser encarada como a solução exata de um sistema $\tilde{A}x = \tilde{b}$, com os dados ligeiramente modificados.

Com isso, se o sistema $Ax = b$ for mal condicionado, a solução obtida $\tilde{x} = x^{(0)}$ pode ser muito diferente da solução exata. O mesmo pode ocorrer com as correções $c^{(k)}$, caso façamos o refinamento desta solução. Neste caso, o processo iterativo de refinamento pode não convergir para a solução exata $\bar{x}$. Por essa razão, convém estipular um número máximo de iterações (ITMAX) a serem feitas.

7 – CÁLCULO DA MATRIZ INVERSA PELO MÉTODO DE ELIMINAÇÃO DE GAUSS

Dada uma matriz A, de ordem n, sua inversa A^{-1} pode ser obtida através da resolução de n sistemas lineares. Daí a possibilidade de calcularmos a inversa, utilizando o Método de Eliminação de Gauss.

Sendo

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

o que queremos é determinar

$$A^{-1} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix}$$

tal que $AA^{-1} = I$, ou seja

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Deste modo, para cada j, j = 1, 2, . . . , n, o resultado da multiplicação da matriz A pela j-ésima coluna de A^{-1} é igual à j-ésima coluna de I, isto é,

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_{1j} \\ \vdots \\ x_{jj} \\ \vdots \\ x_{nj} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Assim, a j-ésima coluna de A^{-1} pode ser obtida resolvendo o sistema:

$$\begin{aligned} a_{11}x_{1j} + a_{12}x_{2j} + \ldots + a_{1n}x_{nj} &= 0 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad &\vdots \\ a_{j1}x_{1j} + a_{j2}x_{2j} + \ldots + a_{jn}x_{nj} &= 1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad &\vdots \\ a_{n1}x_{1j} + a_{n2}x_{2j} + \ldots + a_{nn}x_{nj} &= 0 \end{aligned}$$

A solução de n sistemas deste tipo nos fornece A^{-1}.

Exemplo 7.1 Vamos determinar a inversa da matriz

$$A = \begin{bmatrix} 2 & -1 & -1 \\ 3 & 4 & -2 \\ 3 & -2 & 4 \end{bmatrix}$$

trabalhando em ponto flutuante com três algarismos significativos.

Precisamos resolver os sistemas lineares $Ax_i = e_i$, $i = 1, 2, 3$, onde e_i é a i-ésima coluna da matriz identidade e x_i é a i-ésima coluna de A^{-1}. Como os três sistemas lineares têm a mesma matriz de coeficientes A, faremos três triangularizações simultaneamente.

$$\left[\begin{array}{ccc|ccc} 2 & -1 & -1 & 1 & 0 & 0 \\ 3 & 4 & -2 & 0 & 1 & 0 \\ 3 & -2 & 4 & 0 & 0 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|ccc} 3 & 4 & -2 & 0 & 1 & 0 \\ 0.667 & -3.67 & 0.33 & 1 & -0.667 & 0 \\ 1 & -6.0 & 6.0 & 0 & -1 & 1 \end{array}\right]$$

$p_1 = 2$

$$\left[\begin{array}{ccc|ccc} 3 & 4 & -2 & 0 & 1 & 0 \\ 1 & -6 & 6 & 0 & -1 & 1 \\ 0.667 & 0.612 & -3.34 & 1 & -0.055 & -0.612 \end{array}\right]$$

$p_1 = 2 \quad p_2 = 3$

Esta última matriz representa os três sistemas triangulares a serem resolvidos. Os vetores dos termos independentes de cada sistema estão nas três últimas colunas.

Resolvendo os sistemas, temos:

a) cálculo da primeira coluna de A^{-1}

$x_{31} = 1/-3.34 = -0.299$

$x_{21} = (0 + 1.79)/-6 = -0.298$

$x_{11} = (0 - (-1.19 + 0.598))/3 = 0.592/3 = 0.197$

b) cálculo da segunda coluna de A^{-1}

$x_{32} = -0.055/-3.34 = 0.0165$

$x_{22} = (-1 - 0.099)/-6 = -1.1/-6 = 0.183$

$x_{12} = (1 - 0.732 + 0.033)/3 = 0.301/3 = 0.1$

c) cálculo da terceira coluna de A^{-1}

$x_{33} = -0.612/-3.34 = 0.183$

$x_{23} = (1 - 1.1)/-6 = -0.1/-6 = 0.0167$

$x_{13} = (0 - 0.668 + 0.366)/3 = 0.299/3 = 0.0997$

Portanto,

$$A^{-1} = \begin{bmatrix} 0.197 & 0.1 & 0.0997 \\ -0.298 & 0.183 & 0.0167 \\ -0.299 & 0.0165 & 0.183 \end{bmatrix}$$

8 – MÉTODO ITERATIVO DE GAUSS-SEIDEL

a) Descrição do método

A solução de um sistema linear $Ax = b$ também pode ser obtida utilizando um processo iterativo que consiste em calcular uma seqüên-

cia $x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(k)}, \ldots$ de aproximações da solução $\bar{x}$ do sistema, a partir de uma aproximação inicial $x^{(0)}$.

Dado um sistema linear $Ax = b$ de ordem n, existem n! maneiras de ordenar as equações deste sistema. Uma vez fixada a ordem das equações, se $a_{ii} \neq 0, i = 1, 2, \ldots, n$, podemos escrever:

$$\begin{aligned}
x_1 &= \frac{1}{a_{11}} [b_1 - a_{12}x_2 - a_{13}x_3 - \ldots - a_{1n}x_n] \\
x_2 &= \frac{1}{a_{22}} [b_2 - a_{21}x_1 - a_{23}x_3 - \ldots - a_{2n}x_n] \\
&\vdots \\
x_n &= \frac{1}{a_{nn}} [b_n - a_{n1}x_1 - a_{n2}x_2 - \ldots - a_{n,n-1}x_{n-1}]
\end{aligned} \tag{8.1}$$

O método iterativo de Gauss-Seidel consiste em, escolhida uma aproximação inicial

$$x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$$

calcular a seqüência de aproximações

$$x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots$$

utilizando as equações:

$$\begin{aligned}
x_1^{(k+1)} &= \frac{1}{a_{11}} [b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \ldots - a_{1n}x_n^{(k)}] \\
x_2^{(k+1)} &= \frac{1}{a_{22}} [b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \ldots - a_{2n}x_n^{(k)}] \\
&\vdots \\
x_n^{(k+1)} &= \frac{1}{a_{nn}} [b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \ldots - a_{n,n-1}x_{n-1}^{(k+1)}]
\end{aligned} \tag{8.2}$$

Dizemos que o processo iterativo converge se, para a seqüência de aproximações gerada, dado $\varepsilon > 0$, existir $\bar{k}$, tal que para todo $k > \bar{k}$ e $i = 1, 2, \ldots, n$, $|x_i^{(k)} - \bar{x}_i| \leqslant \varepsilon$.

Como na prática não conhecemos $\bar{x}$, torna-se necessário um critério de parada para o processo. Adotaremos o mesmo critério

usado no processo de refinamento (seção 5), ou seja, vamos comparar duas aproximações consecutivas

$$x^{(k-1)} = (x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)})$$

e

$$x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$$

usando a variação relativa

$$\mathrm{Var}^{(k)} = \max \{v_1^{(k)}, v_2^{(k)}, \ldots, v_n^{(k)}\}$$

onde

$$v_i^{(k)} = \begin{cases} \left| \dfrac{x_i^{(k)} - x_i^{(k-1)}}{x_i^{(k)}} \right| & \text{se } x_i^{(k)} \neq 0 \\ 0 & \text{se } x_i^{(k)} = 0 = x_i^{(k-1)} \\ 1 & \text{se } x_i^{(k)} = 0 \text{ e } x_i^{(k-1)} \neq 0 \end{cases}$$

Consideraremos que o processo convergiu quando $\mathrm{Var}^{(k)} \leqslant \varepsilon$, para algum k e adotaremos como solução do sistema a k-ésima aproximação obtida.

Como o processo iterativo pode não convergir, devemos estipular um número máximo de iterações (ITMAX) a serem feitas.

Exemplo 8.1 Vejamos a solução do sistema:

$$\begin{aligned}
4x_1 + x_2 + x_3 &= 5 \\
-2x_1 + 5x_2 + x_3 &= 0 \\
3x_1 + x_2 + 6x_3 &= -6.5
\end{aligned}$$

pelo método iterativo de Gauss-Seidel, trabalhando em aritmética de ponto flutuante com três algarismos significativos, considerando $x^{(0)} = (0.0, 0.0, 0.0)$, $\varepsilon = 0.01$ e ITMAX = 5.

As equações para o processo iterativo, conforme (8.2), são:

$$\begin{aligned}
x_1^{(k)} &= \frac{1}{4}[5 - x_2^{(k-1)} - x_3^{(k-1)}] \\
x_2^{(k)} &= \frac{1}{5}[0 + 2x_1^{(k)} - x_3^{(k-1)}] \\
x_3^{(k)} &= \frac{1}{6}[-6.5 - 3x_1^{(k)} - x_2^{(k)}]
\end{aligned}$$

Os resultados obtidos nas iterações são dados na Tabela 8.1.

Tabela 8.1

k = número de iteração	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0.0	0.0	0.0	–
1	1.25	0.5	−1.8	1
2	1.58	0.992	−2.03	0.496
3	1.51	1.01	−2.0	0.464
4	1.5	1.0	−2.0	$0.01 \leqslant \varepsilon$

Portanto devemos considerar x = (1.5, 1.0, −2.0) como solução do sistema. Neste caso particular, essa solução é exata. Convém lembrar que um processo iterativo nem sempre fornece a solução exata do sistema.

Exemplo 8.2 Vamos resolver, pelo Método de Gauss-Seidel, o sistema:

$$5x_1 + 3x_2 = 15$$
$$-4x_1 + 10x_2 = 19$$

trabalhando em aritmética de ponto flutuante com quatro algarismos significativos.

Consideremos $(x_1^{(0)}, x_2^{(0)}) = (0.0, 0.0)$, $\varepsilon = 0.005$ e ITMAX = = 10.

Utilizando as equações (8.2), obtemos os resultados das iterações, dispostos na Tabela 8.2.

Tabela 8.2

k = número de iteração	$x_1^{(k)}$	$x_2^{(k)}$	$Var^{(k)}$
0	0.0	0.0	–
1	3.0	3.1	1
2	1.14	2.356	1.632
3	1.586	2.534	0.2812
4	1.480	2.492	0.07162
5	1.505	2.502	0.01661
6	1.499	2.5	$0.004003 \leqslant \varepsilon$

Portanto, devemos considerar como solução do sistema

$$x = (1.499, 2.5)$$

Sendo o sistema deste exemplo um sistema linear de ordem 2, podemos fazer a interpretação geométrica das aproximações obtidas.

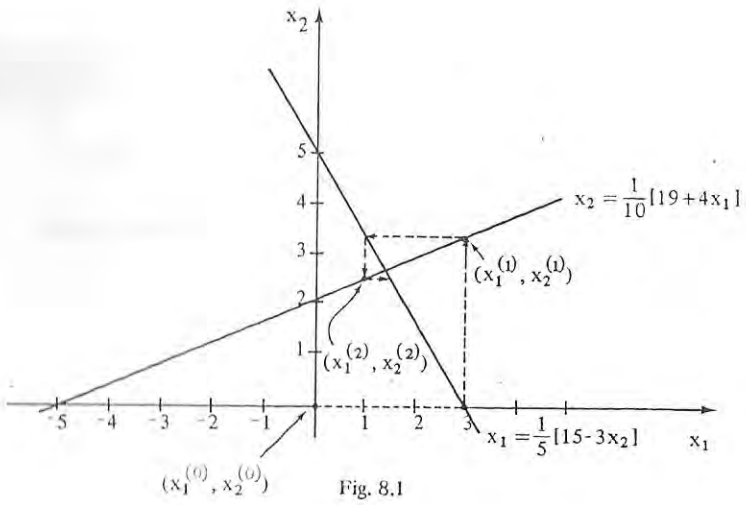


Fig. 8.1

b) Estudo da convergência do método de Gauss-Seidel

Quando utilizamos o método iterativo de Gauss-Seidel para resolver um sistema linear Ax = b, devemos nos preocupar com a convergência da seqüência de aproximações da solução. Existem condições sobre os elementos da matriz A dos coeficientes do sistema que, se satisfeitas, são suficientes para garantir a convergência do método de Gauss-Seidel. A seguir veremos condições sobre sistemas de ordem 2 que são necessárias e suficientes para a convergência deste método.

Proposição 8.1 Dado o sistema linear de ordem 2

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2 \qquad (8.3)$$

com $a_{11} \neq 0$, $a_{22} \neq 0$, o processo iterativo definido por

$$x_1^{(k)} = \frac{1}{a_{11}} [b_1 - a_{12} x_2^{(k-1)}]$$

$$k \geqslant 1 \qquad (8.4)$$

$$x_2^{(k)} = \frac{1}{a_{22}} [b_2 - a_{21} x_1^{(k)}]$$

converge se e somente se

$$\left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right| < 1$$

Demonstração: Sejam $\bar{x} = (\bar{x}_1, \bar{x}_2)$ a solução exata do sistema e $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$ a aproximação inicial escolhida.

O erro absoluto da componente i, i = 1, 2, na k-ésima iteração é dado por $\Delta x_i^{(k)} = \bar{x}_i - x_i^{(k)}$, i = 1, 2.

O processo converge se e somente se

$$\lim_{k \to \infty} |\Delta x_i^{(k)}| = 0, \quad i = 1, 2$$

De (8.3) temos:

$$x_1 = \frac{1}{a_{11}} [b_1 - a_{12} x_2]$$

$$x_2 = \frac{1}{a_{22}} [b_2 - a_{21} x_1]$$

Deste modo, para $k \geqslant 1$, temos:

$$\Delta x_1^{(k+1)} = \bar{x}_1 - x_1^{(k+1)} = \frac{1}{a_{11}} [b_1 - a_{12} \bar{x}_2] - \frac{1}{a_{11}} [b_1 - a_{12} x_2^{(k)}] =$$

$$= -\frac{a_{12}}{a_{11}} (\bar{x}_2 - x_2^{(k)}) = -\frac{a_{12}}{a_{11}} \Delta x_2^{(k)} \qquad (8.5a)$$

e

$$\Delta x_2^{(k)} = \bar{x}_2 - x_2^{(k)} = \frac{1}{a_{22}} [b_2 - a_{21} \bar{x}_1] - \frac{1}{a_{22}} [b_2 - a_{21} x_1^{(k)}] =$$

$$= -\frac{a_{21}}{a_{22}} (\bar{x}_1 - x_1^{(k)}) = -\frac{a_{21}}{a_{22}} \Delta x_1^{(k)} \qquad (8.5b)$$

Portanto, usando repetidas vezes (8.5) temos:

$$\Delta x_1^{(k+1)} = -\frac{a_{12}}{a_{11}} \Delta x_2^{(k)} = \frac{a_{12} a_{21}}{a_{11} a_{22}} \Delta x_1^{(k)} = \left[\frac{a_{12} a_{21}}{a_{11} a_{22}} \right]^2 \Delta x_1^{(k-1)} =$$

$$= \ldots = \left[\frac{a_{12} a_{21}}{a_{11} a_{22}} \right]^k \Delta x_1^{(1)}$$

e

$$\Delta x_2^{(k+1)} = -\frac{a_{21}}{a_{22}} \Delta x_1^{(k+1)} = \frac{a_{12} a_{21}}{a_{11} a_{22}} \Delta x_2^{(k)} = \left[\frac{a_{12} a_{21}}{a_{11} a_{22}} \right]^2 \Delta x_2^{(k-1)} =$$

$$= \ldots = \left[\frac{a_{12} a_{21}}{a_{11} a_{22}} \right]^k \Delta x_2^{(1)}$$

Os erros absolutos $x_i^{(1)} = \bar{x}_i - x_i^{(1)}$, i = 1, 2 são constantes pois, uma vez fixada a aproximação inicial $x^{(0)} = (x_1^{(0)}, x_2^{(0)})$, a primeira aproximação $x^{(1)} = (x_1^{(1)}, x_2^{(1)})$ é única. Portanto, para i = 1, 2 temos:

$$\lim_{k \to \infty} \left| \Delta x_i^{(k)} \right| = \lim_{k \to \infty} \left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right|^{k-1} |\Delta x_i^{(1)}| = |\Delta x_i^{(1)}| \lim_{k \to \infty} \left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right|^{k-1}$$

Logo,

$$\lim_{k \to \infty} \left| \Delta x_i^{(k)} \right| = 0 \quad \text{se e somente se} \quad \left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right| < 1 \qquad \square$$

Como a condição de convergência da Proposição 8.1 é necessária e suficiente, se um sistema linear de ordem 2 é tal que $a_{ii} \neq 0$, i = 1, 2 e

$$\left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right| > 1$$

então a seqüência gerada por (8.4) diverge.

Por exemplo, para o sistema

$$-4x_1 + 10x_2 = 19$$

$$5x_1 + 3x_2 = 15$$

temos

$$\left| \frac{a_{12} a_{21}}{a_{11} a_{22}} \right| = \frac{10 * 5}{-4 * 3} = \frac{50}{12} > 1$$

Escolhendo $x^{(0)} = (0.0, 0.0)$ como aproximação inicial, podemos observar na Fig. 8.2 que o processo diverge.

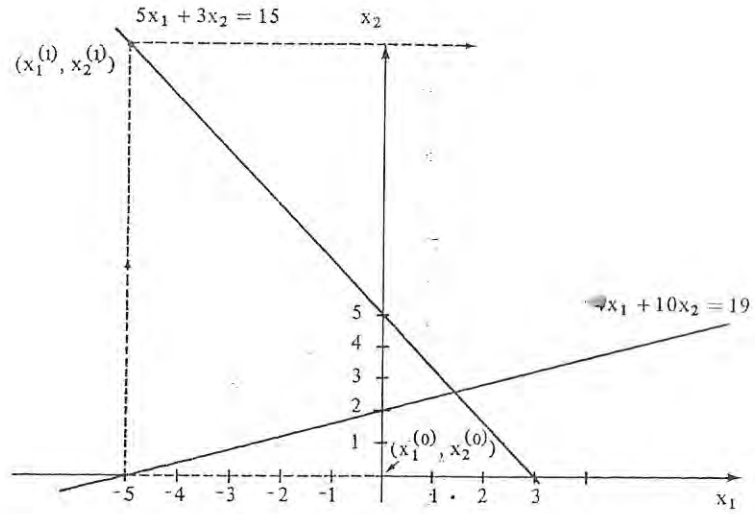


Fig. 8.2

Entretanto, se trocarmos a ordem das equações desse sistema, obteremos o mesmo do Exemplo 8.2, que converge. Note que no Exemplo 8.2, temos:

$$\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| = \left|\frac{3*(-4)}{5*10}\right| < 1$$

Corolário 8.1 Dado um sistema linear de ordem 2

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

com $a_{ii} \neq 0$, i = 1, 2, o Método de Gauss-Seidel converge em alguma ordem dessas equações se e somente se

$$\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| \neq 1$$

Exercício 8.1 Prove o Corolário 8.1.

Exercício 8.2 Dê um exemplo de um sistema linear de ordem 2 que tenha uma única solução e para o qual o Método de Gauss-Seidel não converge, qualquer que seja a ordem das equações.

Vamos estudar, agora, condições suficientes para garantir a convergência do Método de Gauss-Seidel quando aplicado a um sistema de ordem n.

Seja Ax = b um sistema linear de ordem n com

$$a_{ii} \neq 0, \qquad i = 1, 2, \ldots, n$$

A primeira condição suficiente para a convergência do Método de Gauss-Seidel que vamos estabelecer é denominada "Critério de Sassenfeld".

Para este critério de convergência precisamos dos valores $\beta_1, \beta_2, \ldots, \beta_n$, obtidos através da recorrência:

$$\beta_1 = \left|\frac{1}{a_{11}}\right| \sum_{j=2}^{n} |a_{1j}|$$

$$\beta_i = \left|\frac{1}{a_{ii}}\right| \left[\sum_{j=1}^{i-1} |a_{ij}|\beta_j + \sum_{j=i+1}^{n} |a_{ij}|\right] \qquad i = 2, 3, \ldots, n \tag{8.6}$$

Exemplo 8.3 Vamos calcular os valores $\beta_1, \beta_2, \beta_3$ e β_4 para o sistema:

$$\begin{aligned} 2x_1 + x_2 - 0.2x_3 + 0.2x_4 &= 0.4 \\ 0.6x_1 + 3x_2 - 0.6x_3 - 0.3x_4 &= -7.8 \\ -0.1x_1 - 0.2x_2 + x_3 + 0.2x_4 &= 1.0 \\ 0.4x_1 + 1.2x_2 + 0.8x_3 + 4x_4 &= -10.0 \end{aligned}$$

$$\beta_1 = \frac{1}{2}(1 + 0.2 + 0.2) = 0.7$$

$$\beta_2 = \frac{1}{3}(0.6*0.7 + 0.6 + 0.3) = 0.44$$

$$\beta_3 = \frac{1}{1}(0.1*0.7 + 0.2*0.44 + 0.2) = 0.358$$

$$\beta_4 = \frac{1}{4}(0.4*0.7 + 1.2*0.44 + 0.8*0.358) = 0.2736$$

Lema 8.1 Dado um sistema $Ax = b$, sejam β_i, $i = 1, 2, \ldots, n$, definidos em (8.6) e, para todo $k \geqslant 1$,

$$|\Delta x_i^{(k)}| = |\overline{x}_i - x_i^{(k)}|$$

Então,

$$|\Delta x_i^{(k+1)}| \leqslant \beta_i \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \qquad k \geqslant 1 \tag{8.7}$$

Demonstração: A demonstração será feita por indução finita sobre i.

Sabemos que, para todo i, $1 \leqslant i \leqslant n$.

$$\overline{x_i} = \frac{1}{a_{ii}} [b_i - a_{i1}\overline{x_1} - \ldots - a_{i,i-1}\overline{x_{i-1}} - a_{i,i+1}\overline{x_{i+1}} - \ldots - a_{in}\overline{x_n}]$$

e

$$x_i^{(k+1)} = \frac{1}{a_{ii}} [b_i - a_{i1}x_1^{(k+1)} - \ldots - a_{i,i-1}x_{i-1}^{(k+1)} - a_{i,i+1}x_{i+1}^{(k)} - \ldots - a_{in}x_n^{(k)}]$$

Com isso, temos:

- Base da indução ($i = 1$):

$$|\Delta x_1^{(k+1)}| \leqslant \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}| |\Delta x_j^{(k)}| \leqslant \left[\frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{ij}| \right] \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

portanto,

$$|\Delta x_1^{(k+1)}| \leqslant \beta_1 \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

- Hipótese de Indução: Suponhamos que, para todo ℓ, $1 \leqslant \leqslant \ell < i$, vale:

$$|\Delta x_\ell^{(k+1)}| \leqslant \beta_\ell \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

- Passo da Indução: Vamos provar que a desigualdade é válida para $\ell = i$. Como

$$|\Delta x_i^{(k+1)}| \leqslant \frac{1}{|a_{ii}|} \left[\sum_{\ell=1}^{i-1} |a_{i\ell}| |\Delta x_\ell^{(k+1)}| + \sum_{j=i+1}^{n} |a_{ij}| |\Delta x_j^{(k)}| \right],$$

aplicando a hipótese de indução temos:

$$|\Delta x_i^{(k+1)}| \leqslant \frac{1}{|a_{ii}|} \left[\left(\sum_{\ell=1}^{i-1} |a_{i\ell}| \beta_\ell \right) \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| + \sum_{j=i+1}^{n} |a_{ij}| |\Delta x_j^{(k)}| \right] \leqslant$$

$$\leqslant \frac{1}{|a_{ii}|} \left[\left(\sum_{\ell=1}^{i-1} |a_{i\ell}| \beta_\ell \right) \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| + \left(\sum_{j=i+1}^{n} |a_{ij}| \right) \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \right] = \frac{1}{|a_{ii}|} \left[\sum_{\ell=1}^{i-1} |a_{i\ell}| \beta_\ell + \sum_{j=i+1}^{n} |a_{ij}| \right] \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

Portanto,

$$|\Delta x_i^{(k+1)}| \leqslant \beta_i \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \qquad \square$$

Proposição 8.2 (Critério de Sassenfeld) Seja

$$M = \max_{1 \leqslant i \leqslant n} \beta_i$$

onde os β_i são definidos em (8.6). A condição $M < 1$ é suficiente para que as aproximações obtidas pelo método de Gauss-Seidel convirjam para a solução do sistema $Ax = b$.

Demonstração: Precisamos mostrar que

$$\lim_{k \to \infty} |\Delta x_i^{(k)}| = 0, \qquad i = 1, 2, \ldots, n$$

onde $\Delta x_i^{(k)} = \overline{x}_i - x_i^{(k)}$ é o erro absoluto da i-ésima componente, $i = 1, 2, \ldots, n$, na k-ésima iteração, $k \geqslant 1$.

Do Lema 8.1 segue que, para todo i, $1 \leqslant i \leqslant n$,

$$|\Delta x_i^{(k+1)}| \leqslant M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

o que implica que

$$\max_{1\leqslant i\leqslant n} |\Delta x_i^{(k+1)}| \leqslant M \max_{1\leqslant j\leqslant n} |\Delta x_j^{(k)}|$$

Aplicando repetidas vezes esta desigualdade, temos:

$$\max_{1\leqslant j\leqslant n} |\Delta x_j^{(k+1)}| \leqslant M \max_{1\leqslant j\leqslant n} |\Delta x_j^{(k)}| \leqslant M^2 \max_{1\leqslant j\leqslant n} |\Delta x_j^{(k-1)}| \leqslant$$

$$\leqslant \ldots \leqslant M^k \max_{1\leqslant j\leqslant n} |\Delta x_j^{(1)}|$$

Portanto, para todo i, $1 \leqslant i \leqslant n$,

$$\lim_{k\to\infty} |\Delta x_i^{(k)}| = \lim_{k\to\infty} |\Delta x_i^{(k+1)}| \leqslant \lim_{k\to\infty} \max_{1\leqslant j\leqslant n} |\Delta x_j^{(k+1)}| \leqslant$$

$$\leqslant \lim_{k\to\infty} M^k \max_{1\leqslant j\leqslant n} |\Delta x_j^{(1)}|$$

Uma vez fixada a aproximação inicial $x^{(0)}$,

$$\max_{1\leqslant j\leqslant n} |\Delta x_j^{(1)}|$$

é uma constante. Assim, temos:

$$\lim_{k\to\infty} |\Delta x_i^{(k)}| \leqslant \max_{1\leqslant j\leqslant n} |\Delta x_j^{(1)}| \lim_{k\to\infty} M^k$$

Por hipótese, $0 \leqslant M < 1$, donde concluímos que:

$$\lim_{k\to\infty} |\Delta x_i^{(k)}| = 0, \qquad i = 1, 2, \ldots, n \qquad \square$$

No caso do sistema dado no Exemplo 8.3

$$M = \max_{1\leqslant i\leqslant n} \beta_i = 0.7$$

Portanto, se aplicarmos o método de Gauss-Seidel, para resolver tal sistema, a convergência está assegurada.

É fácil ver que existem sistemas lineares que não satisfazem o critério de Sassenfeld e para os quais o método Gauss-Seidel converge.

Exemplo 8.4 O sistema linear

$$2x_1 + 4x_2 = 14$$
$$x_1 + 5x_2 = 11$$

não satisfaz o critério de Sassenfeld, pois

$$\beta_1 = \frac{4}{2} = 2 > 1$$

No entanto,

$$\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| = \frac{4}{10} < 1$$

o que garante a convergência do método de Gauss-Seidel pela Proposição 8.1.

Utilizando o Lema 8.1, podemos estabelecer um outro critério de parada para o método de Gauss-Seidel, conforme veremos a seguir.

Proposição 8.3 Seja

$$M = \max_{1\leqslant i\leqslant n} \beta_i < 1$$

onde os β_i são definidos em (8.6). Dado $\varepsilon > 0$, se

$$\max_{1\leqslant j\leqslant n} |x_j^{(k+1)} - x_j^{(k)}| \leqslant \frac{1-M}{M}\varepsilon, \qquad k \geqslant 1$$

então,

$$|\Delta x_i^{(k+1)}| \leqslant \varepsilon \text{ para todo } i = 1, 2, \ldots, n$$

Demonstração: Para todo $i = 1, 2, \ldots, n$

$$\Delta x_i^{(k)} = \bar{x}_i - x_i^{(k)} = \bar{x}_i - x_i^{(k+1)} + x_i^{(k+1)} - x_i^{(k)}$$

Portanto

$$|\Delta x_i^{(k)}| = |\bar{x}_i - x_i^{(k+1)} + x_i^{(k+1)} - x_i^{(k)}| \leqslant |\bar{x}_i - x_i^{(k+1)}| + |x_i^{(k+1)} - x_i^{(k)}| = |\Delta x_i^{(k+1)}| + |x_i^{(k+1)} - x_i^{(k)}|,$$
$$i = 1, 2, \ldots, n \qquad (8.8)$$

Pelo Lema 8.1,

$$|\Delta x_i^{(k+1)}| \leqslant M \max_{1\leqslant j\leqslant n} |\Delta x_j^{(k)}|, \qquad i = 1, 2, \ldots, n \qquad (8.9)$$

Substituindo este resultado em (8.8), temos:

$$|\Delta x_i^{(k)}| \leqslant M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| + |x_i^{(k+1)} - x_i^{(k)}|, \qquad i = 1, 2, \ldots, n$$

assim,

$$\max_{1 \leqslant i \leqslant n} |\Delta x_i^{(k)}| \leqslant \max_{1 \leqslant i \leqslant n} \left[M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| + |x_i^{(k+1)} - x_i^{(k)}| \right]$$

Como

$$M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}|$$

não depende de i, temos,

$$\max_{1 \leqslant i \leqslant n} |\Delta x_i^{(k)}| \leqslant M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| + \max_{1 \leqslant i \leqslant n} |x_i^{(k+1)} - x_i^{(k)}|.$$

Como i e j têm as mesmas variações, podemos escrever:

$$\max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| - M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \leqslant \max_{1 \leqslant j \leqslant n} |x_j^{(k+1)} - x_j^{(k)}|$$

Portanto,

$$(1 - M) \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \leqslant \max_{1 \leqslant j \leqslant n} |x_j^{(k+1)} - x_j^{(k)}|$$

Logo,

$$\max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \leqslant \frac{1}{1 - M} \max_{1 \leqslant j \leqslant n} |x_j^{(k+1)} - x_j^{(k)}|$$

Substituindo este resultado na inequação (8.9) temos, para todo $i = 1, 2, \ldots, n$,

$$|\Delta x_i^{(k+1)}| \leqslant M \max_{1 \leqslant j \leqslant n} |\Delta x_j^{(k)}| \leqslant \frac{M}{1 - M} \max_{1 \leqslant j \leqslant n} |x_j^{(k+1)} - x_j^{(k)}|$$

Portanto, dado $\varepsilon > 0$, se

$$\max_{1 \leqslant j \leqslant n} |x_j^{(k+1)} - x_j^{(k)}| \leqslant \frac{1 - M}{M} \varepsilon$$

então,

$$|\Delta x_i^{(k+1)}| \leqslant \frac{M}{1 - M} \frac{1 - M}{M} \varepsilon = \varepsilon \qquad \square$$

O resultado da Proposição 8.3 nos fornece uma delimitação para o erro absoluto

$$\Delta x_i^{(k+1)} = \bar{x} - x_i^{(k+1)},$$

através da comparação feita entre duas aproximações consecutivas $x^{(k)}$ e $x^{(k+1)}$. É evidente que assim obtemos informações mais seguras sobre a relação entre as soluções exata e aproximada, do que quando usamos a variação relativa. Entretanto, esta delimitação só pode ser utilizada se o sistema linear satisfaz o critério de Sassenfeld.

Vejamos agora um outro critério suficiente para a convergência do Método de Gauss-Seidel, que, apesar de ser mais restritivo do que o critério de Sassenfeld, é de verificação muito mais simples.

Proposição 8.4 (Critério das linhas) Uma condição suficiente para garantir a convergência do método de Gauss-Seidel, quando aplicada a um sistema $Ax = b$, com $a_{ii} \neq 0$, é

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, 2, \ldots, n$$

Demonstração: Vamos mostrar que se

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, 2, \ldots, n$$

então o sistema $Ax = b$ satisfaz o critério de Sassenfeld e, portanto, converge. Para isso vamos demonstrar que $\beta_i < 1$, $i = 1, 2, \ldots, n$, onde os β_i são definidos em (8.6). Usaremos indução finita sobre i.

- Base da indução ($i = 1$):

$$\beta_1 = \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}| < \frac{1}{|a_{11}|} |a_{11}| = 1$$

• Hipótese de indução: Seja $1 \leqslant i \leqslant n$ e suponhamos que $\beta_j < 1$ para todo $j = 1, 2, \ldots, i-1$.

• Passo da indução: Vamos mostrar que para $j = i$ temos $\beta_j < 1$. Por definição

$$\beta_i = \frac{1}{|a_{ii}|}\left[\sum_{j=1}^{i-1} |a_{ij}|\,\beta_j + \sum_{j=i+1}^{n} |a_{ij}|\right]$$

Aplicando a hipótese de indução, temos

$$\beta_i \leqslant \frac{1}{|a_{ii}|}\left[\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|\right]$$

Logo,

$$\beta_i < \frac{1}{|a_{ii}|}\,|a_{ii}| = 1$$

Portanto,

$$M = \max_{1 \leqslant i \leqslant n} \beta_i < 1 \qquad \square$$

No exemplo a seguir vemos que o critério das linhas não é equivalente ao critério de Sassenfeld.

Exemplo 8.5 O sistema linear

$$\begin{aligned} 10\,x_1 + \;\; x_2 &= 23 \\ 6\,x_1 + 2\,x_2 &= 18 \end{aligned}$$

satisfaz o critério de Sassenfeld pois $\beta_1 = \frac{1}{10}$ e $\beta_2 = \frac{3}{10}$, mas não satisfaz o critério das linhas, porque $a_{22} < a_{21}$.

Deste modo, verificamos que o conjunto dos sistemas lineares que possuem solução única obedecem a relação de inclusão exposta na Fig. 8.3, onde todas as inclusões são próprias.

Como os critérios de Sassenfeld e das linhas são apenas condições suficientes para a convergência, o método de Gauss-Seidel pode ser utilizado para resolver sistemas lineares que não satisfaçam a nenhum destes critérios. Entretanto, nestes casos deve ser feita uma análise

cuidadosa da seqüência de aproximações obtida, já que o método de Gauss-Seidel não detecta se o sistema é determinado ou não.

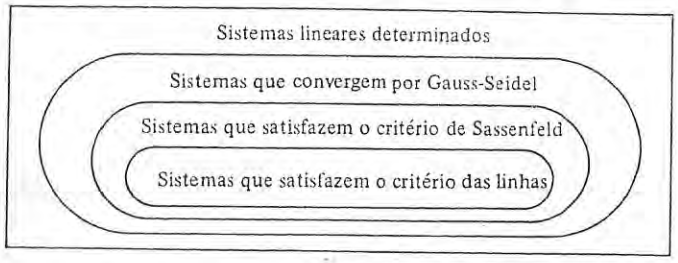


Fig. 8.3

Exemplo 8.6 Vejamos a aplicação do método de Gauss-Seidel a sistemas lineares com determinante nulo.

Dos três sistemas a seguir, os dois primeiros são indeterminados e o terceiro é inconsistente:

a) $x_1 + x_2 + x_3 = 1 \qquad x_1^{(k+1)} = 1 - x_2^{(k)} - x_3^{(k)}$

$2x_1 + 2x_2 + 2x_3 = 2 \Longrightarrow x_2^{(k+1)} = \frac{1}{2}[2 - 2x_1^{(k+1)} - 2x_3^{(k)}]$

$5x_1 + 5x_2 + 5x_3 = 5 \qquad x_3^{(k+1)} = \frac{1}{5}[5 - 5x_1^{(k+1)} - 5x_2^{(k+1)}]$

Escolhendo $x^{(0)} = (0, 0, 0)$ temos a seguinte seqüência de aproximações:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0	0	0	–
1	1	0	0	1
2	1	0	0	0

Como a variação relativa é nula, $x^{(1)} = (1, 0, 0)$ é uma solução exata do sistema. Mas, quando utilizamos um método iterativo, é altamente improvável a obtenção da solução exata na primeira iteração. Isto nos leva a "suspeitar" da solução obtida.

Escolhendo $x^{(0)} = (1, 1, 1)$ como nova aproximação inicial, temos as aproximações:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	1	1	1	–
1	−1	1	1	2
2	−1	1	1	0

e novamente obtivemos uma solução exata do sistema, $x^{(1)} = (-1, 1, 1)$. Isto se deve ao fato das equações corresponderem a três planos coincidentes.

b) $x_1 + x_2 + x_3 = 1 \qquad x_1^{(k+1)} = 1 - x_2^{(k)} - x_3^{(k)}$

$2x_1 + 2x_2 + 2x_3 = 2 \rightarrow x_2^{(k+1)} = \frac{1}{2}[2 - 2x_1^{(k+1)} - 2x_3^{(k)}]$

$2x_1 + x_2 + x_3 = 0 \qquad x_3^{(k+1)} = -2x_1^{(k+1)} - x_2^{(k+1)}$

Escolhendo $x^{(0)} = (0, 0, 0)$ como aproximação inicial, temos

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0	0	0	–
1	1	0	−2	1
2	3	0	−6	2/3
3	7	0	−14	4/7
4	15	0	−30	8/15
5	31	0	−62	16/31
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	$2^n - 1$	0	$2 - 2^{n+1}$	$2^{n-1}/(2^n - 1)$

Podemos verificar que

$$Var^{(k)} = 2^{(k-1)}/(2^k - 1)$$

logo,

$$\lim_{k \to \infty} Var^{(k)} = 1/2$$

Isto significa que se tomarmos $\varepsilon > 1/2$ e se ITMAX for razoavelmente grande, existe k tal que $Var^{(k)} < \varepsilon$. Entretanto, observando a seqüência de aproximações, é fácil constatar que não existe convergência.

Escolhendo $x^{(0)} = (0, 2, 0)$ como nova aproximação inicial, temos as aproximações:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0	2	0	–
1	−1	2	0	1
2	−1	2	0	0

Logo, $x = (-1, 2, 0)$ é solução exata do sistema.

Escolhendo $x^{(0)} = (0, 1, 1)$ como outra aproximação inicial, temos as aproximações:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0	1	1	–
1	−1	1	1	1
2	−1	1	1	0

Com isso temos uma nova solução exata $x = (-1, 1, 1)$.

Para este sistema, a intersecção dos três planos é dada pela reta que passa pelos pontos $(-1, 2, 0)$ e $(-1, 0, 2)$.

c) $x_1 + x_2 + x_3 = 1 \qquad x_1^{(k+1)} = 1 - x_2^{(k)} - x_3^{(k)}$

$2x_1 + 2x_2 + 2x_3 = 2 \rightarrow x_2^{(k+1)} = \frac{1}{2}[2 - 2x_1^{(k+1)} - 2x_3^{(k)}]$

$x_1 + x_2 + x_3 = 2 \qquad x_3^{(k+1)} = 2 - x_1^{(k+1)} - x_2^{(k+1)}$

Considerando $x^{(0)} = (0, 0, 0)$, temos:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$Var^{(k)}$
0	0	0	0	–
1	1	0	1	1
2	0	0	2	1
3	−1	0	3	1
4	−2	0	4	1/2
5	−3	0	5	1/3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	−n + 2	0	n	1/(n − 2)

Apesar da não convergência, uma vez que este sistema é inconsistente, podemos verificar que $\mathrm{Var}^{(k)} = 1/(k-2)$ para $k \geqslant 3$. Logo,

$$\lim_{k \to \infty} \mathrm{Var}^{(k)} = 0$$

existindo um n para o qual $\mathrm{Var}^{(n)} \leqslant \varepsilon$, qualquer que seja $\varepsilon > 0$ dado.

9 – COMENTÁRIOS FINAIS

Além dos dois métodos para solução de sistemas lineares vistos neste capítulo, existem muitos outros que, na maior parte dos casos, são variações dos métodos aqui apresentados.

Como cada um dos dois métodos tem suas vantagens e desvantagens, a decisão de qual deles deve ser usado depende do particular problema a ser resolvido.

A utilização do método de eliminação de Gauss, que, teoricamente, fornece a solução exata do sistema, pode ser catastrófica se o sistema for mal condicionado. Isto se deve à propagação dos erros de arredondamento, que é muito grande. Nesses casos, nem mesmo o refinamento da solução pode melhorar a solução obtida.

Com relação ao método iterativo de Gauss-Seidel, se o sistema satisfizer algum dos critérios de convergência, ele pode ser utilizado sem preocupações, uma vez que a convergência está assegurada e que a influência dos erros de arredondamento é muito pequena. Isto porque cada aproximação $x^{(k)}$ pode ser considerada como exata quando utilizada no cálculo da próxima aproximação $x^{(k+1)}$. Entretanto, nem todos os sistemas satisfazem algum dos critérios de convergência. Ao contrário, a maioria dos sistemas não os satisfaz. Para esses sistemas devemos tomar os cuidados descritos no fim da seção anterior.

Um caso particular onde a aplicação do método de Gauss-Seidel é recomendada, é a solução de sistemas lineares esparsos, isto é, aqueles que têm grande parte dos coeficientes nulos. Na prática, sistemas desse tipo ocorrem com freqüência.

O método de eliminação de Gauss teoricamente tem uma vantagem sobre o método iterativo de Gauss-Seidel: ele fornece o valor do determinante da matriz dos coeficientes do sistema. Com isso podemos saber se o sistema é determinado ou não.

10 – EXERCÍCIOS

1. Resolva o sistema abaixo pelo método de eliminação de Gauss, utilizando frações:

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ 7 \end{bmatrix}$$

2. Resolva o sistema abaixo pelo método de eliminação de Gauss com condensação pivotal, utilizando ponto flutuante com três algarismos significativos:

$$\begin{bmatrix} -1.0 & -2.3 & 4.7 & 12.0 \\ -1.1 & 2.0 & 3.1 & 3.9 \\ -2.1 & -2.2 & 3.7 & 16.0 \\ -1.2 & 2.1 & -1.1 & 4.0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4.0 \\ 3.9 \\ 12.2 \\ 6.0 \end{bmatrix}$$

3. a) Deduza a fórmula do número de operações aritméticas necessárias para "triangularizar" um sistema linear de ordem n.
   b) Deduza a fórmula do número de operações aritméticas necessárias para resolver um sistema triangular de ordem n.
   c) Deduza a fórmula do número de operações aritméticas necessárias para resolver um sistema linear de ordem n, pelo método de eliminação de Gauss.

4. Determine a inversa da matriz abaixo utilizando o método de eliminação de Gauss, utilizando frações:

$$A = \begin{bmatrix} 3 & 1 & 0 & -1 \\ 0 & -5 & 4 & 2 \\ 1 & \frac{1}{3} & 2 & -\frac{10}{3} \\ 0 & -10 & 9 & \frac{7}{2} \end{bmatrix}$$

5. Mostre que o sistema de equações abaixo é consistente se e somente se $c = 3b - 2d$:

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \\ 1 & 3 & 6 \\ 1 & 3 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

6. Resolva o sistema abaixo, utilizando o método de eliminação de Gauss com condensação pivotal e trabalhando em ponto flutuante com três algarismos significativos.

$$\begin{bmatrix} 3.2 & 1.0 & 2.0 & 0.0 \\ -1.0 & 0.0 & 1.5 & -2.4 \\ 4.1 & 2.5 & 0.0 & 1.0 \\ 3.6 & 0.0 & 0.0 & 2.8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8.2 \\ 2.84 \\ 1.0 \\ 4.72 \end{bmatrix},$$

sabendo-se que o resultado da triangularização da matriz dos coeficientes é

$$\begin{array}{cccc} 4.1 & 2.5 & 0.0 & 1.0 \\ 0.780 & -2.2 & 0.0 & 1.92 \\ 0.878 & -0.277 & 2.0 & -1.61 \\ -0.244 & 0.432 & 0.75 & -0.42 \\ p_1 = 3 & p_2 = 4 & p_3 = 3 & \end{array}$$

7. Resolva o sistema

$$\begin{bmatrix} -12 & 27 & 6 \\ 22 & -15 & 35 \\ 32 & 5 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 116 \\ -41 \\ -61.4 \end{bmatrix}$$

e efetue refinamentos sucessivos, considerando uma precisão $\varepsilon = 0.001$ e calculando no máximo 4 iterações. Trabalhe em ponto flutuante com três algarismos significativos. O resultado da triangularização deste sistema pelo método de eliminação de Gauss com condensação pivotal é:

$$\begin{array}{cccc} 32 & 5 & -4 & -61.4 \\ -0.375 & 28.9 & 4.5 & 93 \\ 0.688 & -0.637 & 40.7 & 60.4 \\ p_1 = 3 & p_2 = 3 & & \end{array}$$

8. Resolvendo o sistema

$$\begin{bmatrix} 5 & 7 \\ 7 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 12 \\ 18 \end{bmatrix}$$

pelo método de eliminação de Gauss com condensação pivotal, trabalhando em ponto flutuante com dois algarismos significativos, obteve-se a solução aproximada $x^{(0)} = (0.57, 1.3)$.

Refine esta solução considerando uma precisão $\varepsilon = 0.01$ e calculando no máximo 5 iterações.

9. Método de eliminação de Gauss como produto de matrizes: a primeira etapa do método de eliminação de Gauss corresponde à pré-multiplicação da matriz aumentada $[A|b]$ do sistema $Ax = b$ pela matriz

$$E_i = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -m_{21} & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ -m_{n1} & 0 & \dots & 1 \end{bmatrix}$$

onde

$$m_{i1} = \frac{a_{i1}}{a_{11}}, \qquad i = 2, \dots, n$$

A i-ésima etapa do método de eliminação de Gauss corresponde à pré-multiplicação da matriz aumentada obtida na etapa anterior, isto é, $E_{i-1}E_{i-2}\ldots E_2E_1[A|b]$, pela matriz

$$E_i = \text{i-ésima linha} \rightarrow \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & -m_{i+1,i} & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & -m_{n,i} & 0 & & 1 \end{bmatrix}$$

(i-ésima coluna)

(E_i = matriz elementar coluna)

onde

$$m_{k,i} = \frac{a'_{k,i}}{a'_{i,i}}, \qquad k = i+1, \ldots, n$$

e $a'_{k,i}$, $a'_{i,i}$ são elementos da matriz obtida na etapa $(i-1)$.

Desta forma, a triangularização no método de eliminação de Gauss pode ser escrita como

$$E_{n-1}E_{n-2}\ldots E_2E_1Ax = E_{n-1}E_{n-2}\ldots E_2E_1b$$

onde $E_{n-1}E_{n-2}\ldots E_2E_1A = U$, uma matriz triangular superior.

Faça a verificação da afirmação acima para um sistema genérico de ordem 3 e exiba as matrizes E_i utilizadas.

10. É dado um sistema linear $Ax = b$ onde A é uma matriz não singular.

a) Verifique que existe uma matriz U triangular superior e uma matriz L triangular inferior tais que

$$A = LU \text{ (denominada decomposição LU de A)}$$

onde

$$L = E_1^{-1}E_2^{-1}\ldots E_{n-1}^{-1}$$

(as matrizes E_i são as matrizes elementares coluna obtidas no Exercício 9).

b) Verifique que

$$L = E_1^{-1}E_2^{-1}\ldots E_{n-1}^{-1} \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ m_{21} & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ m_{n-1,1} & m_{n-1,2} & \ldots & 1 & 0 \\ m_{n,1} & m_{n,2} & \ldots & m_{n,n-1} & 1 \end{bmatrix}$$

onde m_{ik} é o multiplicador obtido pelo método de eliminação de Gauss para a i-ésima linha na k-ésima etapa.

c) Uma outra maneira de resolver um sistema linear $Ax = b$ é utilizar a decomposição LU da matriz A e resolver os dois sistemas triangulares

$$Ly = b$$

e

$$Ux = y$$

Por este método, resolva o seguinte sistema linear, utilizando frações

$$\begin{bmatrix} 2 & -1 & -1 \\ 3 & 4 & -2 \\ 3 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Observação: Note que dada a decomposição $A = LU$, podemos resolver um sistema $Ax = b$ utilizando a fórmula

$$x = U^{-1}L^{-1}b^*$$

* Ver Apêndice A.

11. O método de Doolittee para fazer a decomposição LU de uma matriz A é construir as matrizes L e U utilizando as seguintes fórmulas de recorrência:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj}, \qquad i \leqslant j$$

$$\ell_{ij} = \left[a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj} \right] \Big/ u_{jj}, \qquad i > j$$

Verifique as fórmulas acima para um sistema de ordem 4, admitindo a existência da decomposição A = LU.

12. Resolver um sistema Ax = b pelo método de eliminação de Gauss com condensação pivotal é equivalente a resolver o sistema

$$PAx = Pb$$

onde P é uma matriz de permutação* conveniente. Fazendo a decomposição LU da matriz PA obtemos

$$LUx = Pb$$

Verifique, para um sistema genérico de ordem 4, que:

a) $U = E_3(P_3E_2P_3)(P_3P_2E_1P_2P_3)(P_3P_2P_1)A$, onde P_i é a matriz de permutação da i-ésima etapa, i = 1, 2, 3, U é a matriz A "triangularizada" pelo método de eliminação de Gauss com condensação pivotal, $P = P_3P_2P_1$ é a matriz de permutação e E_i são as matrizes elementares coluna do Exercício 10.

b) $PA = (P_3P_2E_1P_2P_3)^{-1}(P_3E_2P_3)^{-1}E_3^{-1}U = LU$, utilizando a mesma notação do item a).

c) $E_1' = (P_3P_2E_1P_2P_3)$, $E_2' = (P_3E_2P_3)$ e $E_3' = E_3$ são matrizes elementares colunas.

13. Resolva o sistema:

$$\begin{bmatrix} 1 & 0.98 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4.95 \\ 5 \end{bmatrix}$$

* Ver Apêndice A.

pelo método de Gauss-Seidel. Trabalhe em ponto flutuante com três algarismos significativos e considere $x^{(0)} = (0, 0)$, uma precisão $\varepsilon = 0.01$ e o número máximo de iterações igual a 5. Faça a interpretação geométrica.

14. a) Resolva o sistema abaixo pelo método de Gauss-Seidel, trabalhando em ponto flutuante com três algarismos significativos.

$$\begin{bmatrix} 2 & 4 & 2 \\ 1 & 1 & -1 \\ -2 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ -3 \end{bmatrix}$$

Considere $x^{(0)} = (0, 0, 0)$, uma precisão $\varepsilon = 0.001$ e o número máximo de iterações igual a três.

b) Resolva o sistema acima pelo método de eliminação de Gauss com condensação pivotal, trabalhando em ponto flutuante com três algarismos significativos.

15. Seja um sistema linear de ordem 2

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

com $a_{11} \neq 0$ e $a_{22} \neq 0$.

a) Mostre que, se

$$0 < \frac{a_{12}a_{21}}{a_{11}a_{22}} < 1$$

então a convergência do método de Gauss-Seidel é monotônica, isto é,

$$x_i^{(k-1)} \leqslant x_i^{(k)} \leqslant x_i^{(k+1)}$$

ou

$$x_i^{(k-1)} \geqslant x_i^{(k)} \geqslant x_i^{(k+1)}, \qquad i = 1, 2$$

b) Mostre que, se

$$-1 < \frac{a_{12}a_{21}}{a_{11}a_{22}} < 0$$

então a convergência do método de Gauss-Seidel é oscilante, isto é.

$$x_i^{(k)} \geqslant x_i^{(k-1)} \quad e \quad x_i^{(k)} \geqslant x_i^{(k+1)}$$

ou

$$x_i^{(k)} \leqslant x_i^{(k-1)} \quad e \quad x_i^{(k)} \leqslant x_i^{(k+1)}, \qquad i = 1, 2$$

16. Dado o sistema abaixo, deseja-se resolvê-lo pelo método de Gauss-Seidel

$$\begin{cases} x_1 + 2x_2 - x_3 = 1 \\ 2x_1 - x_2 = 1 \\ - x_2 + 2x_3 - x_4 = 1 \\ - x_3 + 2x_4 = 1 \end{cases}$$

a) Este sistema satisfaz o critério das linhas? Justifique.

b) Ele satisfaz o critério de Sassenfeld? Justifique.

c) O que se pode afirmar sobre a convergência? Justifique.

d) O sistema obtido permutando-se as duas primeiras equações satisfaz o critério de Sassenfeld? Justifique.

O que se pode afirmar, então, sobre a convergência?

17. Dada a matriz

$$C = \begin{bmatrix} A & 3 & 1 \\ A & 20 & 1 \\ 1 & A & 6 \end{bmatrix}$$

considere um sistema linear que tem C como matriz dos coeficientes e sua resolução pelo método de Gauss-Seidel.

a) Em que intervalo deve estar A para que se possa afirmar que o método converge usando o critério de Sassenfeld?

b) Para quais valores de A pode-se afirmar que para todos os componentes i, i = 1, 2, 3,

$$|x_i^{(k)} - \bar{x}_i| \leqslant \frac{1}{2} |x_i^{(k-1)} - \bar{x}_i| \ ?$$

($\bar{x}$ é a solução exata e $x^{(k)}$ é a k-ésima aproximação obtida pelo método de Gauss-Seidel.)

Capítulo 4
Aproximações de Funções — Método dos Mínimos Quadrados

Neste capítulo abordaremos o problema de aproximar uma função f por uma outra função g de uma família previamente escolhida. Trataremos de dois casos em paralelo: quando a função f é tabelada – domínio discreto – e quando a função f é dada pela sua forma analítica – domínio contínuo.

Estudaremos o método dos mínimos quadrados começando pelo caso particular de ajuste de uma reta a uma tabela e depois generalizando o raciocínio para aproximar uma função f por uma g da família G das funções que são combinação linear de funções conhecidas, não nulas, g_k, k = 0, 1, . . . , m,

$$g(x) = \sum_{k=0}^{m} a_k g_k(x)$$

A seguir, daremos uma idéia de como podemos tratar do ajuste de uma função por outra não linear nos parâmetros. Trataremos também do caso particular em que as funções g_k são polinômios ortogonais entre si e do caso em que g_k são funções trigonométricas, conhecido como aproximação trigonométrica ou análise harmônica.

1 – GENERALIDADES

Quando se trata de fazer uma aproximação, surgem naturalmente algumas perguntas, como:

Por que aproximar?
Qual família de funções escolher?
Como aproximar?

Nesta seção tentaremos responder estas perguntas e justificar, assim, a escolha do método dos mínimos quadrados.

Por que aproximar?

Quando estamos fazendo um experimento, normalmente conhecemos a família da função que descreve o fenômeno envolvido. Em geral, os valores obtidos já são afetados de erros e, portanto, a função desejada não necessita fornecer exatamente os valores medidos. Basta achar, entre os diversos elementos da família, aquele que "melhor aproxima" o fenômeno medido.

Uma outra circunstância é quando se conhece a forma analítica da função que descreve um fenômeno e precisamos substituí-la por uma outra função (por exemplo, para facilitar o tratamento matemático do modelo) porém, que "se aproxime razoavelmente" da função original.

Qual família de funções escolher?

A escolha da família aproximadora G deve levar em conta os seguintes fatores:

- as características que a função aproximadora deve ter para facilitar os cálculos. Por exemplo: polinômios são facilmente integráveis; adições, subtrações, multiplicações e translações de polinômios resultam em polinômios;
- o comportamento das funções da família G deve se aproximar do comportamento da função f o mais possível, porém, com forma analítica conveniente. Por exemplo, periodicidade da função pode ser obtida com funções trigonométricas.

A escolha da família aproximadora não será tratada neste texto. O método dos mínimos quadrados assume que a família G foi escolhida a contento.

Como aproximar?

Ao aproximar uma função f por uma função g de uma família G estaremos introduzindo um erro r que será chamado de resíduo.

Assim

$$r(x) = f(x) - g\,x) \tag{1.1}$$

Aparentemente, uma "boa" aproximação seria obtida fazendo $\sum_x r(x) = 0$. Analisemos tal afirmação. Para simplificar, suponha que foi realizado um experimento em que se levantou os pontos p_1, p_2, p_3 e p_4. Sabendo-se que o fenômeno é descrito por uma reta, vamos determiná-la de modo a satisfazer $\sum_x r(x) = 0$. Pode-se observar na Fig. 1.1 que todas as retas que foram traçadas obedecem tal critério, o que mostra que $\sum_x r(x) = 0$ não é uma boa escolha.

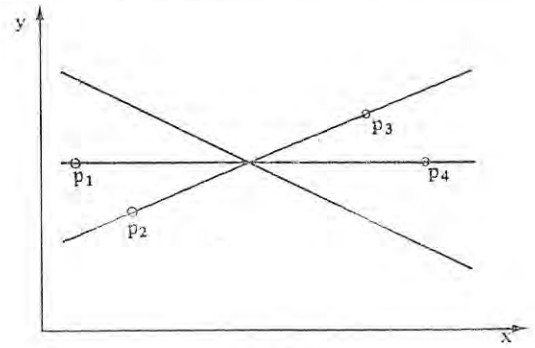


Fig. 1.1

O problema que estamos enfrentando com este critério é o fato dos erros positivos cancelarem os erros negativos. Portanto, se deixarmos de considerar o sinal dos erros, evitaremos este problema. Isso pode ser feito trabalhando com o valor absoluto dos resíduos e exigindo que $\sum_x |r(x)|$ seja mínimo. Achar o mínimo desta função nos leva a uma dificuldade matemática não desejada. Um outro critério com a mesma característica, porém com tratamento matemático mais simples, é exigir que $\sum_x r^2(x)$ seja mínimo. O método para aproximar uma função f por uma $g \in G$ utilizando esse último critério é denominado *método dos mínimos quadrados*.

Existem outros critérios para a escolha da função aproximadora g; porém, neste texto, trataremos apenas do método dos mínimos quadrados.

2 – REGRESSÃO LINEAR

Nosso objetivo agora é aproximar uma função f por uma função g da família a + bx pelo método dos mínimos quadrados.

Nesta seção vamos nos preocupar apenas em aproximar uma função f tabelada nos pontos x_i, $i = 1, 2, \ldots, n$, $n \geqslant 2$, por uma reta, utilizando o método dos mínimos quadrados. Este caso particular é conhecido como *regressão linear*.

Aproximar uma função f tabelada nos pontos x_i, $i = 1, \ldots, n$, pelo método dos mínimos quadrados, significa determinar os parâmetros a e b da reta a + bx de modo que a soma dos quadrados dos erros em cada ponto seja mínima.

O resíduo em cada ponto $(x_i, y_i) = (x_i, f(x_i))$ é dado por:

$$r(x_i) = r_i = y_i - g(x_i) \tag{2.1}$$

Portanto, queremos determinar a e b que minimizam:

$$M(a, b) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2.2}$$

Para isto é necessário* que:

$$\frac{\partial M(a, b)}{\partial a} = 0 \quad \text{e} \quad \frac{\partial M(a, b)}{\partial b} = 0 \tag{2.3}$$

* Dada uma função real contínua e diferenciável de várias variáveis, $g(a_1, a_2, \ldots, a_n)$, para que o ponto $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n) \in \mathbb{R}^n$ seja tal que $g(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n) \leqslant g(a_1, a_2, \ldots, a_n)$ (ou $g(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n) \geqslant g(a_1, a_2, \ldots, a_n)$), $\forall (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, é necessário que

$$\frac{\partial g}{\partial a_1}(\bar{a}_1, \ldots, \bar{a}_n) = \frac{\partial g}{\partial a_2}(\bar{a}_1, \ldots, \bar{a}_n) = \ldots = \frac{\partial g}{\partial a_n}(\bar{a}_1, \ldots, \bar{a}_n) = 0$$

Se $g(a_1, a_2, \ldots, a_n)$ é uma função convexa, a condição acima é necessária e suficiente para que o ponto $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n)$ seja tal que $g(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n) \leqslant g(a_1, a_2, \ldots, a_n)$, $\forall (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ (ou seja $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n)$ é um ponto de mínimo).

ou seja,

$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

e

$$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 \tag{2.4}$$

Portanto:

$$a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

e

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{2.5}$$

ou, usando a notação matricial, temos o seguinte sistema linear:

$$\begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \tag{2.6}$$

Este sistema é denominado *sistema normal*.

Veremos mais adiante que este sistema tem determinante positivo, portanto sempre tem solução.

Se considerarmos x como um vetor em $\mathbb{R}^n$ ($x \in \mathbb{R}^n$) com componentes $x_1, x_2, \ldots, x_n$; $y \in \mathbb{R}^n$, com componentes $y_1, y_2, \ldots, y_n$ e o vetor $\mathbf{1} \in \mathbb{R}^n$, com componentes 1, podemos notar que os somatórios indicados em (2.6) podem ser escritos como produtos escalares em $\mathbb{R}^n$:

$$\sum_{i=1}^{n} 1 = (\mathbf{1}|\mathbf{1}) \tag{2.7}$$

$$\sum_{i=1}^{n} x_i = (1|x) \tag{2.8}$$

$$\sum_{i=1}^{n} x_i^2 = (x|x) \tag{2.9}$$

$$\sum_{i=1}^{n} y_i = (1|y) \tag{2.10}$$

$$\sum_{i=1}^{n} x_i y_i = (x|y) \tag{2.11}$$

Reescrevendo o sistema (2.6), obtemos:

$$\begin{bmatrix} (1|1) & (1|x) \\ (1|x) & (x|x) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} (1|y) \\ (x|y) \end{bmatrix} \tag{2.12}$$

A solução é dada por:

$$\bar{a} = \frac{(1|y)(x|x) - (1|x)(x|y)}{(1|1)(x|x) - (1|x)^2} \tag{2.13}$$

$$\bar{b} = \frac{(1|1)(x|y) - (1|y)(1|x)}{(1|1)(x|x) - (1|x)^2} \tag{2.14}$$

Vamos verificar que $(\bar{a}, \bar{b})$ calculados em (2.13) e (2.14) correspondem a um ponto de mínimo da função $M(a, b) = (r|r)$. Para isto, basta verificar que*:

$$\frac{\partial^2 M}{\partial a^2}(\bar{a}, \bar{b}) > 0 \tag{2.15}$$

* Dada uma função real, contínua, diferenciável, de várias variáveis $g(a_1, a_2, \ldots, a_n)$, seja $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n) \in \mathbb{R}^n$, tal que

$$\frac{\partial g}{\partial a_1}(\bar{a}_1, \ldots, \bar{a}_n) = \ldots = \frac{\partial g}{\partial a_n}(\bar{a}_1, \ldots, \bar{a}_n) = 0$$

Então $(\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n)$ é um ponto de mínimo de $g(a_1, a_2, \ldots, a_n)$ se o seu Hessiano $\nabla^2 g(\bar{a}_1, \ldots, \bar{a}_n)$ é uma forma definida positiva.

e

$$\det \begin{bmatrix} \frac{\partial^2 M}{\partial a^2}(\bar{a}, \bar{b}) & \frac{\partial^2 M}{\partial b \partial a}(\bar{a}, \bar{b}) \\ \frac{\partial^2 M}{\partial a \partial b}(\bar{a}, \bar{b}) & \frac{\partial^2 M}{\partial b^2}(\bar{a}, \bar{b}) \end{bmatrix} > 0 \tag{2.16}$$

já que $(\bar{a}, \bar{b})$ foram determinados a partir de (2.3).

A relação (2.15) é:

$$\frac{\partial^2 M}{\partial a^2}(\bar{a}, \bar{b}) = 2(1|1) = 2n > 0 \tag{2.17}$$

e a relação (2.16) fica:

$$\det \begin{bmatrix} 2(1|1) & 2(1|x) \\ 2(1|x) & 2(x|x) \end{bmatrix}^* = 4[n(x|x) - (1|x)^2] \tag{2.18}$$

Para provar que a expressão obtida em (2.18) é positiva, vamos considerar o produto escalar $(\lambda 1 + x|\lambda 1 + x)$, $\forall \lambda \in \mathbb{R}$.

$$(\lambda 1 + x|\lambda 1 + x) = \lambda^2 (1|1) + 2\lambda (1|x) + (x|x) \tag{2.19}$$

Este produto escalar é estritamente positivo, pois só poderia ser nulo se $x = -\lambda 1$, ou seja, $x_1 = x_2 = \ldots = x_n = -\lambda$, caso que não será considerado, pois corresponde ao ajuste de uma reta a um único ponto, e estamos supondo que a função é sempre conhecida em pelo menos dois pontos. Portanto,

$$\lambda^2 n + 2\lambda \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2 > 0, \quad \forall \lambda \in \mathbb{R} \tag{2.20}$$

Como a inequação do 2º grau em λ obtida em (2.20) é verdadeira, o discriminante da equação do 2º grau correspondente é negativo, isto é,

* Note que o determinante calculado em (2.18) independe do ponto $(\bar{a}, \bar{b})$ considerado. Isto se deve ao fato da função $M(a, b) = (r|r)$ ser uma função convexa.

$$\Delta = \left[2\sum_{i=1}^{n} x_i\right]^2 - 4n\sum_{i=1}^{n} x_i^2 < 0 \qquad (2.21)$$

Portanto,

$$4n\sum_{i=1}^{n} x_i^2 > 4\left[\sum_{i=1}^{n} x_i\right]^2 \qquad (2.22)$$

o que prova que $4[n(x|x) - (1|x)^2] > 0$.

Acabamos de mostrar, também, que o sistema normal (2.12) tem uma única solução, uma vez que o determinante da matriz dos coeficientes do sistema normal (2.12) é dado por $n(x|x) - (1|x)^2$.

Exemplo 2.1 Como resultado de algum experimento, suponha que obtivemos os seguintes valores para a função f:

x	0	1	2	3	4
f(x)	0	1	1	4	4

Vamos determinar a reta que melhor se ajusta a esta função segundo o método dos mínimos quadrados.

O sistema (2.6) (ou (2.12)) correspondente é:

$$\begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 10 \\ 31 \end{bmatrix}$$

que tem solução $\bar{a} = -1/5$ e $\bar{b} = 11/10$. Portanto,

$$g(x) = \frac{11}{10}x - \frac{1}{5}$$

é a reta aproximadora obtida.

3 – MÉTODO DOS MÍNIMOS QUADRADOS – CASO GERAL

Nesta seção estudaremos a aproximação de uma função f por uma função g da família

$$\sum_{k=0}^{m} a_k g_k(x)$$

linear nos parâmetros, pelo método dos mínimos quadrados. A escolha das funções g_k deve ter sido feita, *a priori*, baseada tanto no comportamento da função f quanto nas propriedades desejadas para a função aproximadora.

Vamos considerar dois problemas distintos: o primeiro, como na seção anterior, quando a função f é tabelada, ou seja, o domínio da função que se quer aproximar é discreto, e o segundo, quando o domínio de f é contínuo.

a) Domínio discreto

Para aproximar uma função f tabelada em n pontos distintos $x_1, x_2, \ldots, x_n$ por uma função g da forma

$$\sum_{k=0}^{m} a_k g_k(x)$$

precisamos determinar $a_0, a_1, \ldots, a_m$ que minimizam a soma dos quadrados dos resíduos, $M(a_0, a_1, \ldots, a_m)$, nos pontos x_i, $i = 1, \ldots, n$. Como

$$M(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} (r(x_i))^2 = \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 = \qquad (3.1)$$

$$= \sum_{i=1}^{n} [f(x_i) - a_0 g_0(x_i) - a_1 g_1(x_i) - \ldots - a_m g_m(x_i)]^2$$

precisamos determinar $\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_m$, tais que:

$$\frac{\partial M}{\partial a_\ell}(\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_m) = 2\sum_{i=1}^{n} [f(x_i) - a_0 g_0(x_i) - a_1 g_1(x_i) - \qquad (3.2)$$

$$- \ldots - a_m g_m(x_i)](-g_\ell(x_i)) = 0, \quad 0 \leqslant \ell \leqslant m$$

ou seja,

$$\sum_{i=1}^{n} a_0 g_0(x_i)\, g_\ell(x_i) + \sum_{i=1}^{n} a_1 g_1(x_i)\, g_\ell(x_i) + \ldots + \sum_{i=1}^{n} a_m g_m(x_i)\, g_\ell(x_i) = \sum_{i=1}^{n} f(x_i)\, g_\ell(x_i) \qquad 0 \leqslant \ell \leqslant m \tag{3.3}$$

Usando a notação vetorial, $g_\ell = (g_\ell(x_1), \ldots, g_\ell(x_n)) \in \mathbb{R}^n$, para $0 \leqslant \ell \leqslant m$ e $f = (f(x_1), f(x_2), \ldots, f(x_n)) \in \mathbb{R}^n$, podemos reescrever (3.3) sob a forma de produto escalar, obtendo o sistema linear:

$$\begin{bmatrix} (g_0|g_0) & (g_0|g_1) & \ldots & (g_0|g_m) \\ (g_1|g_0) & (g_1|g_1) & \ldots & (g_1|g_m) \\ \vdots & \vdots & & \vdots \\ (g_m|g_0) & (g_m|g_1) & \ldots & (g_m|g_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} (g_0|f) \\ (g_1|f) \\ \vdots \\ (g_m|f) \end{bmatrix}$$

O sistema (3.4) denomina-se *sistema normal*. Pela propriedade do produto escalar $(g_i|g_j) = (g_j|g_i)$, vemos que o sistema normal é simétrico. Se este sistema admitir uma única solução*,**, esta nos fornece o elemento da família escolhida que melhor aproxima a função f pelo método dos mínimos quadrados.

Exemplo 3.1 Observando um sinal no osciloscópio, verifica-se que ele corresponde à superposição de dois efeitos, um oscilatório e outro crescente. Nestas condições vamos aproximá-lo por uma função g da família $ax + b\cos x = a_0 g_0(x) + a_1 g_1(x)$. Medindo alguns valores deste sinal, obtemos a tabela:

x	0	1.5	3.0	4.5	6.0
f(x)	1.00	1.57	2.00	4.30	7.00

* Observe que a função (3.1) é convexa (ver Apêndice C). Portanto a solução do sistema (3.4) nos fornece o ponto de mínimo.

** No Apêndice B são feitas algumas considerações sobre a existência de solução para o sistema normal.

Trabalhando em aritmética de ponto flutuante com 4 algarismos significativos e lembrando que x deve ser tomado em radianos, temos:

$$(g_0|g_0) = \sum_{i=1}^{5} x_i^2 = 0 + (1.5)^2 + (3)^2 + (4.5)^2 + (6)^2 = 67.5$$

$$(g_0|g_1) = (g_1|g_0) = \sum_{i=1}^{5} x_i * \cos x_i = 0\cos 0 + 1.5\cos 1.5 + 3\cos 3 + 4.5\cos 4.5 + 6\cos 6 = 1.948$$

$$(g_1|g_1) = \sum_{i=1}^{5} \cos^2 x_i = \cos^2 0 + \cos^2 1.5 + \cos^2 3 + \cos^2 4.5 + \cos^2 6 = 2.951$$

$$(g_0|f) = \sum_{i=1}^{5} x_i f(x_i) = 0\,f(0) + 1.5\,f(1.5) + 3\,f(3) + 4.5\,f(4.5) + 6\,f(6) = 69.71$$

$$(g_1|f) = \sum_{i=1}^{5} \cos x_i f(x_i) = \cos 0\, f(0) + \cos 1.5\, f(1.5) + \cos 3\, f(3) + \cos 4.5\, f(4.5) + \cos 6\, f(6) = 6.76$$

Portanto, o sistema normal fica

$$\begin{bmatrix} 67.5 & 1.948 \\ 1.948 & 2.951 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 69.71 \\ 6.76 \end{bmatrix}$$

e a sua solução

$$a_0 = 0.9855$$
$$a_1 = 1.64$$

fornece

$$g(x) = 0.9861\,x + 1.64\cos x$$

que é a função aproximadora desejada.

b) Domínio contínuo

Mesmo no caso em que a forma analítica da função f é conhecida, às vezes é de interesse aproximá-la no intervalo

$$I = [x_I, x_F]$$

por uma função g da família

$$\sum_{k=0}^{m} a_k g_k(x)$$

mais conveniente. Por exemplo, podemos ter uma função com descontinuidades mas queremos trabalhar com uma função contínua. À primeira vista, poderíamos recair no caso discreto tabelando a função dada em alguns pontos. É claro que ao fazer tal discretização estamos perdendo informações sobre o comportamento do erro. Por exemplo, na Fig. 3.1, se tabelarmos a função f apenas nos pontos assinalados, a melhor função aproximadora seria $g(x) = k$, pois teríamos $(r|r) = 0$, o que não seria uma aproximação satisfatória.

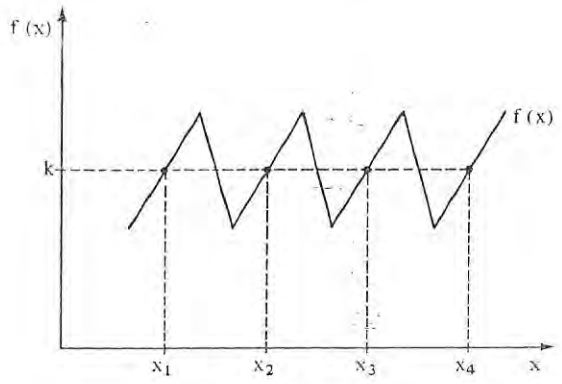


Fig. 3.1

Ao considerarmos a soma dos quadrados dos resíduos em todos os pontos do intervalo $[x_I, x_F]$, teremos, no limite, a integral do quadrado do resíduo em cada ponto do intervalo em que se quer aproximar a função dada.

Assim, teremos que determinar $a_0, a_1, \ldots, a_m$ que minimizam

$$M(a_0, a_1, \ldots, a_m) = \int_{x_I}^{x_F} r(x)^2\,dx =$$

$$= \int_{x_I}^{x_F} (f(x) - g(x))^2\,dx = \tag{3.5}$$

$$= \int_{x_I}^{x_F} (f(x) - a_0 g_0(x) - \ldots - a_m g_m(x))^2\,dx^*$$

Como nas seções anteriores, o ponto de mínimo é obtido quando:

$$\frac{\partial M}{\partial a_0} = \frac{\partial M}{\partial a_1} = \ldots = \frac{\partial M}{\partial a_m} = 0 \tag{3.6}$$

ou seja:

$$\frac{\partial M}{\partial a_\ell} = -2 \int_{x_I}^{x_F} \left(f(x) - \sum_{k=0}^{m} a_k g_k(x)\right) g_\ell(x)\,dx = 0 \tag{3.7}$$

$0 \leqslant \ell \leqslant m$.

Vamos denotar por

$$(f|g) = \int_{x_I}^{x_F} f(x)\,g(x)\,dx \tag{3.8}$$

Observe que a relação (3.8) é um produto escalar, pois obedece todas as propriedades do mesmo. Note que este produto escalar está diretamente ligado ao intervalo $[x_I, x_F]$ considerado.

Exercício 3.1 Mostre que a relação definida em (3.8) satisfaz as propriedades de produto escalar:

a) $(\alpha f|g) = \alpha\,(f|g)$, $\alpha \in \mathbb{R}$

b) $(f+g|h) = (f|h) + (g|h)$

c) $(f|g) = (g|f)$

d) $(f|f) \geqslant 0$

* Pode-se verificar que a função dada em (3.5) é convexa (ver Apêndice C).

Com a notação de produto escalar a relação (3.7) nos fornece o seguinte sistema normal:

$$\begin{bmatrix} (g_0|g_0) & (g_0|g_1) & \dots & (g_0|g_m) \\ (g_1|g_0) & (g_1|g_1) & \dots & (g_1|g_m) \\ \vdots & \vdots & & \vdots \\ (g_m|g_0) & (g_m|g_1) & \dots & (g_m|g_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} (g_0|f) \\ (g_1|f) \\ \vdots \\ (g_m|f) \end{bmatrix} \tag{3.9}$$

Se o sistema normal (3.9) admitir uma única solução* ($\overline{a}_0$, $\overline{a}_1$, ..., $\overline{a}_m$), esta determinará a melhor função aproximadora

$$g(x) = \sum_{k=0}^{m} \overline{a}_k g_k(x)$$

da família escolhida.

Note que $M(a_0, a_1, \dots, a_m)$ também pode ser escrito usando a notação de produto escalar:

$$M(a_0, a_1, \dots, a_m) = (r|r)$$

Exemplo 3.2 Vamos aproximar a função exponencial e^x no intervalo [0, 1] por uma reta, pelo método dos mínimo quadrados.

Neste caso $g(x) = a_0 + a_1 x$. Com a notação utilizada

$$g_0(x) = 1, \quad g_1(x) = x, \quad f(x) = e^x$$

e o produto escalar

$$(f|g) = \int_0^1 f(x)\, g(x)\, dx$$

Portanto, determinar a_0 e a_1 pelo método dos mínimos quadrados é determinar a solução do seguinte sistema normal:

* Ver Apêndice A.

$$\begin{bmatrix} (1|1) & (1|x) \\ (1|x) & (x|x) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} (1|e^x) \\ (x|e^x) \end{bmatrix}$$

Como

$$(1|1) = \int_0^1 1\, dx = 1$$

$$(1|x) = \int_0^1 x\, dx = \frac{1}{2}$$

$$(x|x) = \int_0^1 x^2\, dx = \frac{1}{3}$$

$$(1|e^x) = \int_0^1 e^x\, dx = e - 1$$

$$(x|e^x) = \int_0^1 xe^x\, dx = 1$$

temos

$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} e-1 \\ 1 \end{bmatrix}$$

A solução $a_0 = 4e - 10$ e $a_1 = 18 - 6e$ determina a função aproximadora $g(x) = 4e - 10 + (18 - 6e)x$.

4 – FAMÍLIAS DE FUNÇÕES NÃO LINEARES NOS PARÂMETROS

Nesta seção, preocupar-nos-emos com o ajuste de uma função f por outra função g de uma família não linear nos parâmetros. Por exemplo, funções racionais, hiperbólicas, exponenciais etc.

Ao aplicarmos o método dos mínimos quadrados para determinar os coeficientes da função aproximadora g(x), minimizamos a função $(r|r) = (f - g|f - g)$. Isto nos leva à resolução de um sistema não linear de equações, ou seja, as condições

$$\frac{\partial (r|r)}{\partial a_k} = 0, \qquad k = 0, 1, \ldots, m \tag{4.1}$$

dão origem a equações não lineares em $a_0, a_1, \ldots, a_m$.

A dificuldade, e em alguns casos até mesmo a impossibilidade de resolver (4.1), nos sugere a linearização nos parâmetros.

Exemplo 4.1 Vamos aproximar uma função f por uma função g da família ae^{bx}.

A função g não é linear nos parâmetros. Portanto, vamos linearizá-la aplicando o logaritmo natural tanto à função f quanto à função g, reduzindo o problema a: aproximar $F(x) = \ln f(x)$ por uma função G da família $c_1 + c_2 x$, onde $c_1 = \ln a$ e $c_2 = b$.

Aplicando o método dos mínimos quadrados ao novo problema obtemos c_1 e c_2. Portanto, $a = \ln^{-1} c_1$ e $b = c_2$ são os parâmetros da função aproximadora de f.

Exercício 4.1 Como fazer para aproximar f por uma função g da família ax^b?

Exemplo 4.2 Para reduzir o problema de aproximar f por uma função

$$g(x) = \frac{c_1 + c_2 x + c_3 x^2}{1 + c_4 x + c_5 x^2}$$

a um problema de aproximar uma função conhecida por uma função que seja linear nos parâmetros, vamos tomar $f(x) = g(x)$ e multiplicar os dois membros desta igualdade pelo denominador de g(x), obtendo

$$f(x) + c_4 x f(x) + c_5 x^2 f(x) = c_1 + c_2 x + c_3 x^2$$

Assim, reduzimos o problema acima ao problema de aproximar $F(x) = f(x)$ pela função

$$G(x) = c_1 + c_2 x + c_3 x^2 - c_4 x f(x) - c_5 x^2 f(x)$$

que é linear nos parâmetros.

Exercício 4.2 Determine qual o tipo de linearização que pode ser feito para aproximar uma função f qualquer por

$$g(x) = \frac{x}{a + bx}$$ ■

Existem casos em que a linearização não é possível. Por exemplo, aproximar f pela função $g(x) = c + ae^{bx}$. Uma técnica que pode ser usada, neste caso, é resolver alternadamente, até que a diferença entre dois valores consecutivos de c esteja dentro de uma tolerância aceitável, os seguintes problemas:

1) para um valor de c dado, $c = \bar{c}$, aproximar $F_1(x) = f(x) - \bar{c}$ por uma função G_1 da família ae^{bx};
2) com o valor obtido para b, digamos $b = \bar{b}$, aproximar $F_2(x) = f(x)$ por $G_2(x) = c + ae^{\bar{b}x}$ que já é linear nos parâmetros.

Convém notar que a aproximação feita usando a linearização não fornece o mínimo de $(r|r) = (f - g|f - g)$ do problema original; ela fornece o mínimo de $(F - G|F - G)$ onde F e G são as funções que substituíram f e g, respectivamente.

5 – POLINÔMIOS ORTOGONAIS

Nas seções anteriores vimos que para aproximar uma função f por uma função g da família

$$\sum_{k=0}^{m} a_k g_k(x)$$

pelo método dos mínimos quadrados é necessário resolver um sistema linear de equações denominado sistema normal. Se considerarmos um conjunto de funções $\{g_k\}$, $k = 0, 1, 2, \ldots, m$, tais que

$$(g_k|g_\ell) = 0, \qquad \forall\, k \neq \ell,\ 0 \leqslant k,\ \ell \leqslant m \tag{5.1}$$

o sistema normal se torna diagonal e os coeficientes a_k da função aproximadora são determinados por:

$$a_k = \frac{(g_k|f)}{(g_k|g_k)}, \qquad 0 \leqslant k \leqslant m \tag{5.2}$$

As funções que satisfazem a relação (5.1) são denominadas *funções ortogonais*. Nesta seção vamos nos restringir ao caso em que essas funções são polinômios. Estes polinômios são ditos *polinômios ortogonais*.

Se desejarmos aproximar f por um polinômio devemos preferir o uso de polinômios ortogonais à forma

$$g(x) = \sum_{k=0}^{m} a_k g_k(x) = \sum_{k=0}^{m} a_k x^k$$

pois nesta última forma, ao resolver o sistema normal, estaríamos trabalhando com uma matriz onde os números têm ordens de grandeza muito diferentes, o que pode acarretar muitos erros de arredondamento na resolução do sistema normal. Por exemplo, se estivéssemos aproximando uma função f por um polinômio de grau 5, teríamos na matriz números dados por

$$\sum_{i=1}^{n} 1 \quad \text{e} \quad \sum_{i=1}^{n} x_i^5 \quad \text{ou} \quad \int_{x_I}^{x_F} dx \quad \text{e} \quad \int_{x_I}^{x_F} x^5 dx$$

Um polinômio de grau k pode ser escrito na seguinte forma:

$$p_k(x) = c_k x^k + c_{k-1} x^{k-1} + \ldots + c_1 x + c_0 \qquad (5.3)$$

Os polinômios ortogonais $p_k(x)$, $k = 0, 1, 2, \ldots$, obedecem as seguintes relações:

$$(p_k|p_\ell) = 0 \qquad k \neq \ell \qquad (5.4)$$

$$(p_k|p_k) > 0 \qquad k = 0, 1, 2, \ldots \qquad (5.5)$$

Note que os polinômios de uma família de polinômios ortogonais dependem da particular definição de produto escalar escolhida.

Exemplo 5.1 Vamos construir os três primeiros polinômios ortogonais com relação ao produto escalar

$$(f|g) = \int_0^1 f(x)\, g(x)\, dx$$

e aproximar $f(x) = e^x$ no intervalo [0, 1] por um polinômio do 2º grau.

Para isto vamos utilizar a relação (5.4) e impor que o coeficiente do termo de mais alto grau de cada polinômio seja igual a um.

Vamos determinar $p_0(x)$:

$$p_0(x) = b_0 x^0 = 1\, x^0 = 1$$

O polinômio de grau 1 pode ser escrito como:

$$p_1(x) = c_1 x^1 + c_0 x^0 = x + c_0$$

onde c_0 é determinado fazendo-se $(p_1|p_0) = 0$:

$$(p_1|p_0) = \int_0^1 (x + c_0)\, 1\, dx = \left[\frac{x^2}{2} + c_0 x \right] \Bigg|_0^1 = \frac{1}{2} + c_0 = 0$$

Portanto, $p_1(x) = x - \frac{1}{2}$.

O polinômio de grau 2 pode ser escrito como:

$$p_2(x) = d_2 x^2 + d_1 x^1 + d_0 x^0 = x^2 + d_1 x + d_0$$

onde d_1 e d_0 são determinados impondo-se

$$(p_2|p_0) = 0 \quad \text{e} \quad (p_2|p_1) = 0$$

Portanto, $p_2(x) = x^2 - x + \frac{1}{6}$.

Utilizando os polinômios ortogonais calculados, vamos determinar o polinômio de 2º grau

$$g(x) = a_0 p_0(x) + a_1 p_1(x) + a_2 p_2(x)$$

que aproxima $f(x) = e^x$ no intervalo [0, 1].

De (5.2) temos:

$$a_k = \frac{(p_k|f)}{(p_k|p_k)}$$

Logo,

$$a_0 = \frac{(1|e^x)}{(1|1)} = \frac{\int_0^1 e^x dx}{\int_0^1 dx} = e - 1$$

$$a_1 = \frac{\left(x - \frac{1}{2} \middle| e^x\right)}{\left(x - \frac{1}{2} \middle| x - \frac{1}{2}\right)} = \frac{\int_0^1 e^x \left(x - \frac{1}{2}\right) dx}{\int_0^1 \left(x - \frac{1}{2}\right)^2 dx} = 6\,(3 - e)$$

$$a_2 = \frac{\left(x^2 - x + \frac{1}{6} \middle| e^x\right)}{\left(x^2 - x + \frac{1}{6} \middle| x^2 - x + \frac{1}{6}\right)} = \frac{\int_0^1 \left(x^2 - x + \frac{1}{6}\right) e^x dx}{\int_0^1 \left(x^2 - x + \frac{1}{6}\right)^2 dx} =$$

$$= 30\,(7e - 19)$$

Portanto,

$$g(x) = (e - 1) + 6(3 - e)\ \left(x - \frac{1}{2}\right) + 30\,(7e - 19)\ \left(x^2 - x + \frac{1}{6}\right)$$

Se fixássemos um valor positivo para $(p_k|p_k)$ estaríamos utilizando também a relação (5.5) para determinação dos polinômios e teríamos equações suficientes para determinar todos os coeficientes. Em particular, se tomarmos $(p_k|p_k) = 1$, obteremos uma família de polinômios ortogonais, denominados polinômios ortonormais.

Exercício 5.1 Repita o Exemplo 3.2 utilizando os polinômios ortogonais obtidos acima. Verifique que o resultado obtido é o mesmo.

Exercício 5.2 Construa os três primeiros polinômios ortogonais em relação a

$$(f|g) = \sum_{i=1}^{5} f(i)\, g(i)$$

usando $(p_k|p_\ell) = 0,\ k \neq \ell$ e $(p_k|p_k) = 1$.

Qualquer polinômio q de grau m pode ser escrito como uma combinação linear dos $(m + 1)$ primeiros polinômios ortogonais $p_0, p_1, \ldots, p_m$, ou seja

$$q(x) = \sum_{k=0}^{m} b_k p_k(x) \qquad (5.6)$$

Deste fato, decorre que, se q (x) é um polinômio qualquer de grau m,

$$(q|p_k) = b_k\,(p_k|p_k)\ \forall k,\quad k > 0 \qquad (5.7)$$

onde p_k, $k = 0, 1, 2, \ldots$, são polinômios ortogonais.

Se para os polinômios ortogonais estivermos impondo que o coeficiente do termo de mais alto grau seja unitário, podemos escrever que:

$$x p_{k-1}(x) - p_k(x) = \alpha_k p_{k-1}(x) + \beta_k p_{k-2}(x) \qquad (5.8)$$

onde α_k e β_k são constantes e p_k, p_{k-1} e p_{k-2} são polinômios ortogonais.

Exercício 5.3 Mostre que a relação (5.8) é verdadeira.

Sugestão: Escreva $x p_{k-1}(x) - p_k(x)$ como $\alpha_k p_{k-1}(x) + \beta_k p_{k-2}(x) + q(x)$, onde q (x) é um polinômio de grau $\leqslant (k - 3)$. Basta provar que $q(x) \equiv 0$, mostrando que $(q|p_j) = 0,\ 0 \leqslant j \leqslant (k - 3)$.

Da relação (5.8) podemos determinar as constantes α_k e β_k fazendo o produto escalar com $p_{k-1}(x)$ e $p_{k-2}(x)$, respectivamente:

$$(x p_{k-1} - p_k|p_{k-1}) = (\alpha_k p_{k-1} + \beta_k p_{k-2}|p_{k-1})$$

$$(x p_{k-1}|p_{k-1}) - (p_k|p_{k-1}) = \alpha_k\,(p_{k-1}|p_{k-1}) + \beta_k\,(p_{k-2}|p_{k-1})$$

Como $(p_k|p_{k-1}) = 0$ e $(p_{k-2}|p_{k-1}) = 0$, temos

$$\alpha_k = \frac{(x p_{k-1}|p_{k-1})}{(p_{k-1}|p_{k-1})} \qquad (5.9)$$

Analogamente,

$$\beta_k = \frac{(x p_{k-1}|p_{k-2})}{(p_{k-2}|p_{k-2})} \qquad (5.10)$$

Da relação (5.8), obtemos

$$p_k(x) = (x - \alpha_k)\, p_{k-1}(x) - \beta_k p_{k-2}(x)\,. \qquad (5.11)$$

As relações (5.9), (5.10) e (5.11) nos fornecem uma forma de recorrência para determinação de polinômios ortogonais com coeficiente do termo de mais alto grau igual a um. Para aplicá-las precisamos conhecer $p_1(x)$, já que $p_0(x) = 1$. O polinômio $p_1(x)$ pode ser

calculado a partir da equação $(p_0|p_1) = 0$, ou utilizando a própria fórmula de recorrência, considerando $p_{-1}(x) = 0$.

Exemplo 5.2 Obtenha os polinômios ortogonais do Exemplo 5.1, utilizando a relação de recorrência.

Sejam $p_{-1} \equiv 0$ e $p_0 \equiv 1$.

$$p_1(x) = (x - \alpha_1)\, p_0 - \beta_1 p_{-1} = x - \alpha_1$$

onde

$$\alpha_1 = \frac{(xp_0|p_0)}{(p_0|p_0)} = \frac{(x|1)}{(1|1)} = \frac{\int_0^1 x\,dx}{\int_0^1 dx} = \frac{1}{2}$$

Portanto, $p_1(x) = x - \frac{1}{2}$.

$$p_2(x) = (x - \alpha_2)\, p_1 - \beta_2 p_0 = (x - \alpha_2)\left(x - \frac{1}{2}\right) - \beta_2$$

onde

$$\alpha_2 = \frac{(xp_1|p_1)}{(p_1|p_1)} = \frac{\left(x\left(x - \frac{1}{2}\right)\middle|\left(x - \frac{1}{2}\right)\right)}{\left(x - \frac{1}{2}\middle|x - \frac{1}{2}\right)} = \frac{1}{2}$$

$$\beta_2 = \frac{(xp_1|p_0)}{(p_0|p_0)} = \frac{\left(x\left(x - \frac{1}{2}\right)\middle|1\right)}{(1|1)} = \frac{1}{12}$$

Portanto,

$$p_2(x) = \left(x - \frac{1}{2}\right)^2 - \frac{1}{12} = x^2 - x + \frac{1}{6}$$

Exercício 5.4 Construa os três primeiros polinômios ortogonais utilizando a relação de recorrência (5.9), (5.10) e (5.11) e a seguinte definição de produto escalar

$$(f|g) = \sum_{i=1}^{5} f(i)\, g(i)$$

Exercício 5.5 Dada uma família de polinômios ortogonais $p_0, p_1, p_2, \ldots, p_k, \ldots$, verifique que os polinômios da forma:

a) $q_i = \dfrac{1}{\sqrt{(p_i|p_i)}}\, p_i, \qquad i = 0, 1, 2, \ldots$

formam uma família de polinômios ortonormais;

b) $q_i = \sqrt{\dfrac{v_i}{(p_i|p_i)}}\, p_i, \qquad i = 0, 1, 2, \ldots$

formam uma família de polinômios ortogonais tais que $(q_i|q_i) = v_i$, $i = 0, 1, 2, \ldots$

Diferentes definições* de produto escalar levam a diferentes famílias de polinômios ortogonais. Um exemplo importante de uma destas famílias é a dos polinômios de Legendre que obedecem a seguinte definição:

$$(P_n|P_m) = \int_{-1}^{1} P_n(x)\, P_m(x)\, dx = \begin{cases} 0 & \text{se } m \neq n \\ \dfrac{2}{2n+1} & \text{se } m = n \end{cases} \qquad (5.12)$$

Usando esta definição podemos construir os três primeiros polinômios de Legendre

$$P_0(x) = 1, \quad P_1(x) = x \quad \text{e} \quad P_2(x) = \frac{1}{2}(3x^2 - 1) \qquad (5.13)$$

* Uma maneira mais geral de definir produto escalar é dada por:

$$(f|g) = \int_a^b w(x)\, f(x)\, g(x)\, dx \qquad w(x) > 0,\ a < x < b \quad \text{ou}$$

$$(f|g) = \sum_{i=1}^{n} w_i f(x_i)\, g(x_i) \qquad w_i > 0,\ i = 1, \ldots, n$$

A função $w(x)$ (ou w_i) é chamada função peso (ou peso). A definição que estamos utilizando neste texto é o caso particular em que a função peso é igual a 1.

Além destas, ainda existem outras famílias de polinômios ortogonais, como os polinômios de Hermite, Tchebyshev etc. Alguns destes polinômios encontram-se tabelados [Spiegel].

A seguir, vamos utilizar os polinômios ortogonais tabelados, ou previamente calculados, para ajustar uma função f por um polinômio g de grau menor ou igual a m, num intervalo [a, b].

Suponhamos que dispomos de uma tabela de polinômios ortogonais num intervalo [c, d]. Precisamos fazer uma mudança de variável $t(x) = \alpha x + \beta$* para transformar f(t), definida no intervalo [a, b], em $f(t(x)) = F(x)$ definida no intervalo [c, d]. Faremos, então, o ajuste da função F(x) por um polinômio G(x) de grau menor ou igual a m usando os polinômios tabelados. Por transformação inversa de variável $x(t) = \gamma t + \delta$ obtemos a função aproximadora $g(t) = G(x(t))$.

A transformação linear de variável conserva o mínimo do produto escalar, pois

$$(f - g|f - g) = \int_a^b (f(t) - g(t))^2 dt = \int_c^d (f(t(x)) - g(t(x)))^2 \alpha dx =$$

$$= \alpha \int_c^d (F(x) - G(x))^2 dx = \alpha (F - G|F - G)$$

Exemplo 5.3 Vamos aproximar a função $f(t) = \operatorname{sen} t$, no intervalo $0 \leqslant t \leqslant \pi$, por uma parábola, utilizando os polinômios de Legendre.

Fazendo a mudança de variável que transforma linearmente o intervalo $[0, \pi]$ em $[-1, 1]$, temos:

$$t(x) = \frac{\pi}{2}(x + 1)$$

Nestas condições

$$f(t(x)) = \operatorname{sen} t(x) = \operatorname{sen}\left(\frac{\pi}{2}(x + 1)\right) = F(x)$$

* Para transformar linearmente uma função f(t) definida num intervalo [a, b], $t \in [a, b]$ em $F(x) = f(t(x))$ definida em [c, d], $x \in [c, d]$, tomamos $t(x) = \alpha x + \beta$ que é a reta que passa pelos pontos (c, a) e (d, b).

Vamos determinar a parábola

$$G(x) = a_0 1 + a_1 x + a_2 \frac{1}{2}(3x^2 - 1) \qquad (5.14)$$

pelo método dos mínimos quadrados.

De (5.2) obtemos:

$$a_0 = \frac{(F|p_0)}{(p_0|p_0)} = \frac{2}{\pi}$$

$$a_1 = \frac{(F|p_1)}{(p_1|p_1)} = 0$$

$$a_2 = \frac{(F|p_2)}{(p_2|p_2)} = \frac{10}{\pi}\left[1 - \frac{12}{\pi^2}\right]$$

Substituindo em (5.14):

$$G(x) = \frac{2}{\pi} + \frac{10}{\pi}\left[1 - \frac{12}{\pi^2}\right]\frac{1}{2}(3x^2 - 1) \qquad x \in [-1, 1]$$

Voltando para o intervalo inicial $[0, \pi]$ através da transformação inversa, $x(t) = \frac{2}{\pi}t - 1$, obtemos:

$$g(t) = \frac{2}{\pi} + \frac{10}{\pi}\left[1 - \frac{12}{\pi^2}\right]\frac{1}{2}\left[3\left[\frac{2}{\pi}t - 1\right]^2 - 1\right]$$

que é a função aproximadora desejada.

6 – ANÁLISE HARMÔNICA

Quando a função f a ser aproximada é periódica, é conveniente utilizar como função aproximadora um elemento de uma família de funções também periódicas, com o mesmo período da que se quer aproximar. Uma particular classe é representada pelas funções:

$$1, \cos x, \cos 2x, \ldots, \cos nx, \quad \operatorname{sen} x, \operatorname{sen} 2x, \ldots, \operatorname{sen} nx \qquad (6.1)$$

Para qualquer função h (x) dada em (6.1)

$$h(x) = h(x + 2\pi)$$

Nesta seção, vamos nos preocupar em aproximar uma função f, periódica de período 2π, por uma função g da família

$$a_0 + \sum_{k=1}^{m} (a_k \cos kx + b_k \operatorname{sen} kx) \qquad (6.2)$$

pelo método dos mínimos quadrados. Se necessário, faremos uma mudança de variável para tornar o período de f igual a 2π.

Esta aproximação recebe o nome de *análise harmônica*, *aproximação trigonométrica* ou *aproximação de Fourier*.

a) Domínio contínuo

Na aplicação do método dos mínimos quadrados para determinar a função aproximadora

$$g(x) = a_0 + \sum_{k=1}^{m} (a_k \cos kx + b_k \operatorname{sen} kx)$$

basta calcular os coeficientes $a_0, a_1, b_1, a_2, b_2, \ldots, a_m, b_m$ que minimizam

$$(r|r) = \int_c^{c+2\pi} r^2(x)\,dx = \int_c^{c+2\pi} (f(x) - g(x))^2\,dx \qquad (6.3)$$

para qualquer constante real c, uma vez que por hipótese as funções f e g são periódicas de período 2π. As formas mais usuais para (6.3) são:

$$(r|r) = \int_0^{2\pi} (f(x) - g(x))^2\,dx$$

e

$$(r|r) = \int_{-\pi}^{\pi} (f(x) - g(x))^2\,dx$$

As funções dadas em (6.1) são ortogonais em relação ao produto escalar

$$(f|g) = \int_c^{c+2\pi} f(x)\,g(x)\,dx \qquad (6.4)$$

portanto o sistema normal obtido pelo método dos mínimos quadrados é um sistema diagonal e os coeficientes são dados por:

$$a_0 = \frac{(f|1)}{(1|1)}$$

$$a_k = \frac{(f|\cos kx)}{(\cos kx|\cos kx)} \qquad k = 1, 2, \ldots \qquad (6.5)$$

$$b_k = \frac{(f|\operatorname{sen} kx)}{(\operatorname{sen} kx|\operatorname{sen} kx)} \qquad k = 1, 2, \ldots$$

Usando a definição de produto escalar dada em (6.4) com c = 0, temos:

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,dx$$

$$a_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\,dx \qquad k = 1, 2, \ldots \qquad (6.6)$$

$$b_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\operatorname{sen} kx\,dx \qquad k = 1, 2, \ldots$$

Exercício 6.1 Mostre que as funções dadas em (6.1) são ortogonais em relação ao produto escalar definido em (6.4). Use as fórmulas de prostaférese.

Exercício 6.2 Mostre que $(\cos kx|\cos kx) = (\operatorname{sen} kx|\operatorname{sen} kx) = \pi$, usando o produto escalar definido em (6.4).

Uma outra forma de apresentar a análise harmônica é obtida multiplicando e dividindo a expressão de g (x) dada em (6.2) por

$$\sqrt{a_k^2 + b_k^2}$$

Assim,

$$g(x) = a_0 + \sum_{k=1}^{m} \sqrt{a_k^2 + b_k^2}\left[\frac{a_k}{\sqrt{a_k^2 + b_k^2}}\cos kx + \frac{b_k}{\sqrt{a_k^2 + b_k^2}}\operatorname{sen} kx\right] \tag{6.7}$$

Tomando-se

$$A_k = \sqrt{a_k^2 + b_k^2}, \quad \frac{a_k}{A_k} = \operatorname{sen}\phi_k \quad \text{e} \quad \frac{b_k}{A_k} = \cos\phi_k$$

temos:

$$g(x) = a_0 + \sum_{k=1}^{m} A_k(\operatorname{sen}\phi_k \cos kx + \cos\phi_k \operatorname{sen} kx) = a_0 + \sum_{k=1}^{m} A_k \operatorname{sen}(kx + \phi_k) \tag{6.8}$$

Cada termo da expressão em (6.8) é uma senóide com freqüência igual a k vezes a freqüência da função f (freqüência = 1/período). Por esta razão, o termo $A_k \operatorname{sen}(kx + \phi_k)$ é chamado de harmônico de ordem k e pode ser caracterizado somente pela sua amplitude A_k e seu ângulo de fase ϕ_k.

Exemplo 6.1 Vamos fazer a análise harmônica da função f (t) da Fig. 6.1 até o harmônico de 3ª ordem.

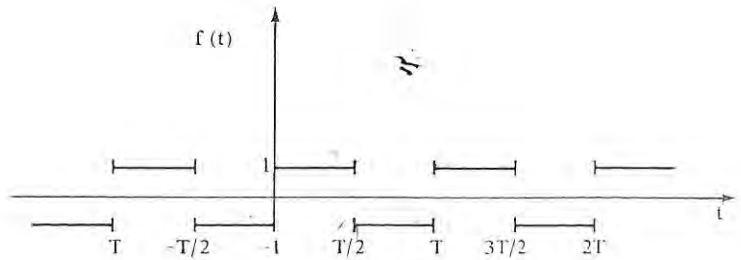


Fig. 6.1

A função f (t) é periódica com período T. Vamos fazer uma mudança de variável para obter f (t (x)) com período 2π. Esta mudança é feita através da transformação $t(x) = \frac{T}{2\pi}x$.

Calculando os coeficientes da função aproximadora

$$G(x) = a_0 + \sum_{k=1}^{3} (a_k \cos kx + b_k \operatorname{sen} kx)$$

através de (6.6), obtemos:

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right] dx = \frac{1}{2\pi}\left[\int_0^{\pi} f\left[\frac{T}{2\pi}x\right] dx + \int_{\pi}^{2\pi} f\left[\frac{T}{2\pi}x\right] dx\right] = \frac{1}{2\pi}\int_0^{\pi} 1\,dx + \frac{1}{2\pi}\int_{\pi}^{2\pi} -1\,dx = 0$$

$$a_1 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right]\cos x\,dx = \frac{1}{\pi}\int_0^{\pi}\cos x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi}\cos x\,dx = 0$$

$$a_2 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right]\cos 2x\,dx = \frac{1}{\pi}\int_0^{\pi}\cos 2x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi}\cos 2x\,dx = 0$$

$$a_3 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right]\cos 3x\,dx = \frac{1}{\pi}\int_0^{\pi}\cos 3x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi}\cos 3x\,dx = 0$$

$$b_1 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right]\operatorname{sen} x\,dx = \frac{1}{\pi}\int_0^{\pi}\operatorname{sen} x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi}\operatorname{sen} x\,dx = \frac{4}{\pi}$$

$$b_2 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right] \operatorname{sen} 2x\,dx =$$

$$= \frac{1}{\pi}\int_0^{\pi} \operatorname{sen} 2x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi} \operatorname{sen} 2x\,dx = 0$$

$$b_3 = \frac{1}{\pi}\int_0^{2\pi} f\left[\frac{T}{2\pi}x\right] \operatorname{sen} 3x\,dx =$$

$$= \frac{1}{\pi}\int_0^{\pi} \operatorname{sen} 3x\,dx - \frac{1}{\pi}\int_{\pi}^{2\pi} \operatorname{sen} 3x\,dx = \frac{4}{3\pi}$$

ou seja,

$$G(x) = \frac{4}{\pi}\operatorname{sen} x + \frac{4}{3\pi}\operatorname{sen} 3x$$

portanto,

$$g(t) = G\left[\frac{2\pi}{T}t\right] = \frac{4}{\pi}\operatorname{sen}\left[\frac{2\pi}{T}t\right] + \frac{4}{3\pi}\operatorname{sen}\left[3\frac{2\pi}{T}t\right]$$

A Fig. 6.2 mostra o esboço gráfico da função aproximadora G (x) obtida. Em a) temos o esboço das componentes de G (x), em b) o gráfico de G (x) para m = 1 e em c) o gráfico de G (x) para m = 2.

Aumentando o número de harmônicos em G (x) teríamos uma melhor aproximação* para a função f (t).

* Pode-se provar que, se f (x) é periódica com período $T = 2\pi$, contínua por trechos, com limite à esquerda e à direita nos pontos de descontinuidade, então existem $a_0, a_k, b_k, k = 1, 2, \ldots$, tal que

$$f(x) = g(x) = a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \operatorname{sen} kx)$$

Esta série, chamada série de Fourier, converge para f (x) nos pontos de continuidade de f e para

$$\frac{f(x_+) + f(x_-)}{2}$$

nos pontos de descontinuidade de f.

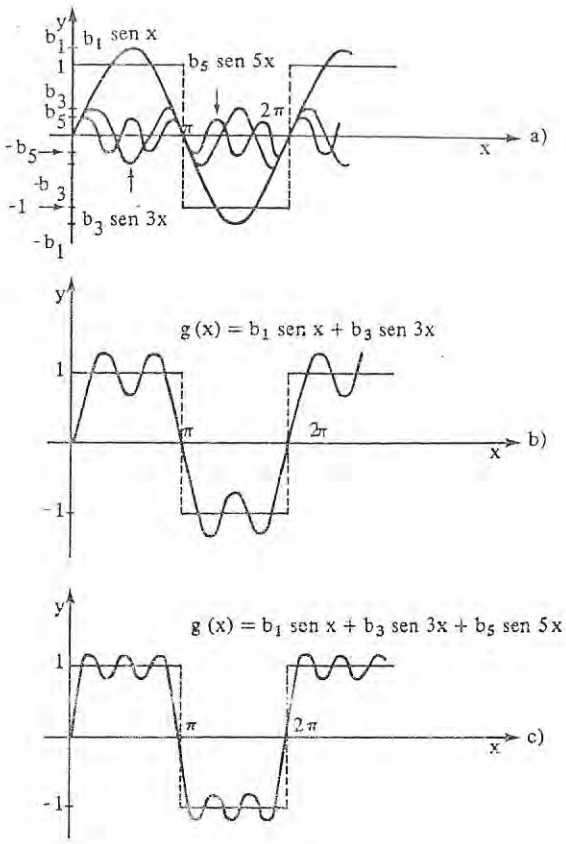


Fig. 6.2

Note que no Exemplo 6.1 a expressão obtida para G (x) coincide com a forma dada em (6.8), onde o harmônico de 1ª ordem tem amplitude $\frac{4}{\pi}$ e fase 0, o de 2ª ordem tem amplitude 0 e o de 3ª ordem tem amplitude $\frac{4}{3\pi}$ e fase 0.

Exercício 6.3 Mostre que se uma função f periódica com período 2π é par (f (x) = f (−x)), então na análise harmônica de f todos os

harmônicos terão fase $\frac{\pi}{2}$, isto é,

$$g(x) = a_0 + \sum_{k=1}^{m} a_k \cos kx$$

Exercício 6.4 Mostre que se uma função f periódica com período 2π é ímpar ($f(x) = -f(-x)$), então todos os harmônicos de f terão fase nula e o termo a_0 também será zero, ou seja,

$$g(x) = \sum_{k=1}^{m} b_k \operatorname{sen} kx$$

b) Domínio discreto

As funções

$$1, \quad \cos x, \quad \cos 2x, \ \ldots, \ \cos (N-1)x, \quad \cos Nx, \quad \operatorname{sen} x, \quad \operatorname{sen} 2x, \ \ldots, \ \operatorname{sen} (N-1)x \tag{6.9}$$

também são ortogonais em relação ao produto escalar

$$(f|g) = \sum_{j=1}^{2N} f(x_j)\, g(x_j) \tag{6.10}$$

sobre o conjunto de pontos

$$x_j = \frac{\pi}{N} j \qquad j = 1, 2, \ldots, 2N \tag{6.11}$$

Exercício 6.5 Verifique que

$$\sum_{j=1}^{2N} \operatorname{sen} \frac{\pi}{N} jk \operatorname{sen} \frac{\pi}{N} j\ell = \begin{cases} 0 & \text{se } 1 \leqslant k \neq \ell \leqslant N-1 \\ N & \text{se } 1 \leqslant k = \ell \leqslant N-1 \end{cases}$$

$$\sum_{j=1}^{2N} \operatorname{sen} \frac{\pi}{N} jk \cos \frac{\pi}{N} j\ell = 0 \qquad 1 \leqslant k \leqslant N-1 \ \text{ e } \ 1 \leqslant \ell \leqslant N$$

$$\sum_{j=1}^{2N} \cos \frac{\pi}{N} jk \cos \frac{\pi}{N} j\ell = \begin{cases} 0 & \text{se } 0 \leqslant k \neq \ell \leqslant N \\ N & \text{se } 0 < k = \ell < N \\ 2N & \text{se } k = \ell = 0 \ \text{ ou } \ k = \ell = N \end{cases}$$

Sugestão: Use as fórmulas de prostaférese e os seguintes fatos:

$$\operatorname{sen} \alpha \frac{\pi}{N} j = -\operatorname{sen} \alpha \frac{\pi}{N} (2N - j)$$

e

$$\cos \alpha \frac{\pi}{N} j = -\cos \alpha \frac{\pi}{N} (N + j) = \operatorname{sen} \left[\alpha \frac{\pi}{N} j + \frac{\pi}{2} \right]$$

Desta forma, se tivermos uma função f periódica, porém dada por meio de uma tabela com 2N pontos eqüidistantes, podemos aproximá-la por uma função

$$g(x) = a_0 + \sum_{k=1}^{m} \left[a_k \cos k \frac{\pi}{N} j + b_k \operatorname{sen} k \frac{\pi}{N} j \right]$$

com $m < N$.

Se $m = N$, o termo $b_N \operatorname{sen} \pi j$ não será calculado, pois $\operatorname{sen} \pi j = 0$

Fazendo com que as abscissas de f(x) coincidam com os pontos $x_j = \frac{\pi}{N} j$ e usando o método dos mínimos quadrados, devido à ortogonalidade das funções dadas em (6.9), os coeficientes da função aproximadora g(x) são dados por:

$$a_0 = \frac{(f|1)}{(1|1)} = \frac{1}{2N} \sum_{j=1}^{2N} f(x_j)$$

$$a_k = \frac{(f|\cos kx)}{(\cos kx|\cos kx)} = \frac{1}{N} \sum_{j=1}^{2N} f(x_j) \cos \left[k \frac{\pi}{N} j \right] \qquad 1 \leqslant k < N \tag{6.12}$$

$$b_k = \frac{(f|\operatorname{sen} kx)}{(\operatorname{sen} kx|\operatorname{sen} kx)} = \frac{1}{N} \sum_{j=1}^{2N} f(x_j) \operatorname{sen} \left[k \frac{\pi}{N} j \right] \qquad 1 \leqslant k < N$$

$$a_N = \frac{(f|\cos Nx)}{(\cos Nx|\cos Nx)} = \frac{1}{2N} \sum_{j=1}^{2N} f(x_j) \cos (\pi j)$$

De forma análoga ao que foi feito para o domínio contínuo, podemos expressar a função aproximadora g(x) para o domínio discreto, também em termos de harmônicos:

$$g(x) = a_0 + \sum_{k=1}^{m} A_k \operatorname{sen}(kx + \phi_k) \qquad (6.13)$$

onde $m \leqslant N$. No caso em que $m = N$, $A_N = a_N$ e $\phi_N = \pi/2$.

Exemplo 6.2 Vamos fazer a análise harmônica, até o 1º harmônico da função f tabelado abaixo

j	1	2	3	4
$f(x_j)$	3	5	7	6

Queremos determinar os coeficientes da função

$$g(x) = a_0 + a_1 \cos x + b_1 \operatorname{sen} x$$

pelo método dos mínimos quadrados.

Temos $2N = 4$ = número de pontos dados. Portanto, $x_j = \frac{\pi}{2} j$. Os coeficientes são calculados a partir das fórmulas (6.12):

$$a_0 = \frac{\sum_{j=1}^{4} f(x_j)}{4} = \frac{21}{4}$$

$$a_1 = \frac{1}{2}\left[3\cos\frac{\pi}{2} + 5\cos\frac{2\pi}{2} + 7\cos\frac{3\pi}{2} + 6\cos\frac{4\pi}{2}\right] = \frac{1}{2}$$

$$b_1 = \frac{1}{2}\left[3\operatorname{sen}\frac{\pi}{2} + 5\operatorname{sen}\frac{2\pi}{2} + 7\operatorname{sen}\frac{3\pi}{2} + 6\operatorname{sen}\frac{4\pi}{2}\right] = -2$$

Logo,

$$g(x) = \frac{21}{4} + \frac{1}{2}\cos x - 2\operatorname{sen} x$$

ou

$$g(x) = \frac{21}{4} + \frac{\sqrt{17}}{2}\operatorname{sen}(x - 0.245)$$

Portanto, o harmônico de 1ª ordem tem amplitude

$$\frac{\sqrt{17}}{2} = 2.06$$

e fase -0.245 radianos.

c) Funções não periódicas

Podemos estender a análise harmônica mesmo para os casos em que a função f a ser aproximada não é periódica. Seja [a, b] o intervalo em que se deseja fazer a aproximação de f. Consideramos então uma nova função h, periódica, com período $T = b - a$ e $h(t) = f(t)$, $t \in [a, b]$ e fazemos a análise harmônica desta função h.

Note que raciocínio análogo pode ser feito para uma função com domínio discreto.

7 – EXERCÍCIOS

1. Ajuste os dados da tabela

x	1	2	3
f(x)	1	2	4

por:

a) regressão linear;

b) um polinômio da forma $a + bx + cx^2$ pelo MMQ (método dos mínimos quadrados).

2. Considere a tabela

x_i	−2	−1	1	2
y_i	1	−3	1	9

a) pelo método dos mínimos quadrados, ajuste à tabela as funções $g_1(x) = ax^2 + bx$ e $g_2(x) = dx^2 + e$;

b) qual função $g_1(x)$ ou $g_2(x)$ fornece melhor ajuste segundo o critério dos mínimos quadrados? Justifique.

3. Deseja-se aproximar o polinômio $3 - x$ do intervalo $[1, 2]$ por uma função g da forma $g(x) = a_1 + a_2\frac{1}{x}$ usando o MMQ. Dê o sistema linear cuja solução fornece os valores de a_1 e a_2.

4. Deseja-se utilizar o MMQ para aproximar uma função f por uma função g que não é linear nos parâmetros. Indique, em cada caso seguinte, como transformar o problema original num problema que pode ser resolvido pelo MMQ:

   a) $g(x)$ da forma ae^{bx}

   b) $g(x)$ da forma $\frac{a + bx + cx^2}{1 + dx}$

   c) $g(x)$ da forma $\sqrt{a + bx}$

5. Considere a tabela

x	0	1	2
f(x)	1	0.5	0.7

   Ajuste os pontos tabelados por uma função da forma

$$\frac{ax}{b + x}$$

   usando o método dos mínimos quadrados.

6. Sejam $f(x)$ e $h(x)$, não identicamente nulas, funções reais distintas. O que deve acontecer ao se tentar aproximar $f(x)$ num intervalo $[a, b]$ pelo MMQ, por $g(x) = a_0 h(x) + a_1 f(x)$? Calcule a_0, a_1 e o erro médio quadrático

$$\left[\int_a^b (f(x) - g(x))^2 dx\right]$$

7. Considere os polinômios $p_0(x) = 1$, $p_1(x) = 1 + bx$, $p_2(x) = 1 + cx + ax^2$. Determine a, b, c de maneira que p_0, p_1 e p_2 sejam ortogonais relativamente à seguinte definição

$$(f|g) = \sum_{x=0}^{2} f(x)\, g(x)$$

8. Demonstre que os polinômios $1, x, x^2, x^3, \ldots$ não são ortogonais em relação a

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

   quaisquer que sejam a e b $(b > a)$.

9. Demonstre que se $p_0(x), p_1(x), \ldots$ são polinômios ortogonais, em relação a

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

   então $p_i(x)$ tem pelo menos uma raiz real em $[a, b]$, $i = 1, 2, \ldots$

10. Construa os três primeiros polinômios ortonormais em relação ao seguinte produto escalar

$$(f|g) = \int_0^1 f(x)\, g(x)\, dx$$

ou seja, construa os três primeiros polinômios ortogonais tais que

$$(p_i|p_j) = 0 \qquad i \neq j$$
$$(p_i|p_i) = 1$$

11. Aproxime a função $f(x) = \operatorname{sen} x$ no intervalo $[0, \pi/6]$ por um polinômio do 1º grau, através do MMQ, utilizando os polinômios obtidos no Exercício 10.

12. Dados os polinômios de Legendre:

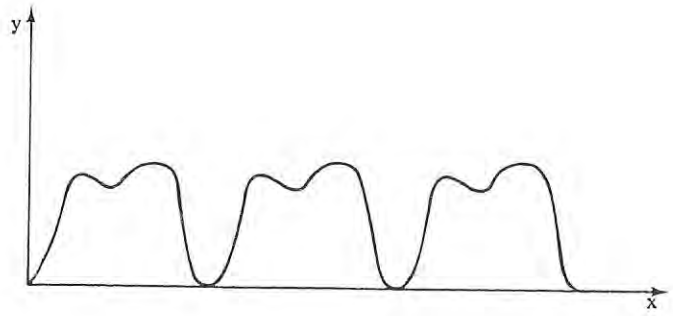
$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1)$$

que obedecem a seguinte definição

$$(P_i|P_j) = \int_{-1}^{1} P_i(x)\, P_j(x)\, dx = \begin{cases} 0 & \text{se } i \neq j \\ \dfrac{2}{2i + 1} & \text{se } i = j \end{cases}$$

aproxime a função $f(x) = \text{sen } x$ por um polinômio do 2º grau, no intervalo $[0, \pi]$, utilizando o MMQ.

13. Num osciloscópio, um certo comportamento periódico é observado na forma gráfica mostrada na figura:



Decide-se aproximar a função da figura acima por uma função da família

$$G(x) = \sum_{i=1}^{m} a_i g_i(x)$$

onde as funções $g_i(x)$ são periódicas. Fazendo as medidas obtemos a tabela

x	0	$\pi/4$	$\pi/2$	$3\pi/4$	π
y	−0.9	1.5	3.1	3.0	1.1

Ajuste estes dados por $G(x) = a_1 \text{sen } x + a_2 \cos x$, pelo MMQ.

14. Mostre que

a) na análise harmônica, a aproximação de uma função periódica $f(x)$ nunca piora com o aumento do número de harmônicos;

b) na análise harmônica de uma função *par* ($f(x) = f(-x)$) os termos em sen kx têm coeficientes nulos.

15. Faça a análise harmônica até o harmônico de ordem 2 da função

$$f(x) = \begin{cases} 1 & 0 \leqslant x < 1 \\ -1 & 1 \leqslant x < 2 \end{cases}$$

16. Provar que $f(x) = x^2$ para $-\pi \leqslant x \leqslant \pi$, f de período 2π, possui a série de Fourier

$$f(x) = \frac{\pi^2}{3} - 4\left(\cos x - \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x - \ldots\right.$$

17. Fazendo $x = \pi$ no Exercício 16, mostre que

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ldots = \frac{\pi^2}{6}$$

18. Utilizando o resultado do Exercício 16 mostre que

$$\sum_{i=1}^{\infty} \frac{(-1)^{n+1}}{n^2} = 1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \ldots = \frac{\pi^2}{12}$$

19. Desenvolver $f(x) = x^2$ $(0 \leqslant x \leqslant 2\pi)$ em série de Fourier.

20. Utilizando o resultado do Exercício 19, provar que:

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} + \ldots = \frac{\pi^2}{6}$$

Capítulo 5
Interpolação Polinomial

Neste capítulo apresentaremos os polinômios interpoladores na forma de Lagrange e na forma de Newton. O polinômio interpolador na forma de Newton será obtido de maneira construtiva. Veremos, também, sua forma para o caso particular em que os pontos dados têm abscissas eqüidistantes. Será obtida uma delimitação para o erro de truncamento na interpolação.

1 – INTRODUÇÃO

Seja uma tabela da forma

x	x_0	x_1	...	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	...	$f(x_n)$

(1.1)

correspondente aos valores de uma função f em $n+1$ pontos distintos $x_0, x_1, \ldots, x_n$ pertencentes a $\mathbb{R}$, e seja x um ponto distinto dos pontos x_i da tabela, pertencente ao intervalo que contém os pontos x_i, isto é, $x \neq x_i$, $i = 0, 1, \ldots, n$ e existem k e j, $0 \leqslant k \neq j \leqslant n$, tais que $x_k < x < x_j$.

Interpolar o ponto x à Tabela (1.1) significa calcular o valor de $f(x)$, ou seja, incluir o ponto $(x, f(x))$ à tabela.

Em problemas reais, em geral, precisamos fazer a interpolação de um ponto x em uma tabela em que não conhecemos a forma analítica da função correspondente, pois, na maioria dos casos, a Tabela (1.1) é obtida a partir de resultados experimentais. Assim, como não podemos calcular $f(x)$, pois f é desconhecida, fazemos uma *interpolação polinomial* de x, ou seja, determinamos o polinômio p de grau menor ou igual a n que passa por todos os pontos da tabela e calculamos o valor de $p(x)$. Este polinômio é chamado *polinômio interpolador* de f relativamente aos pontos $x_0, x_1, \ldots, x_n$.

A proposição a seguir nos garante a existência e unicidade do polinômio interpolador de uma tabela na forma dada em (1.1).

Proposição 1.1 Dados $n+1$ pontos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, com $x_0, x_1, \ldots, x_n$ distintos entre si, existe um único polinômio p de grau menor ou igual a n que passa por esses pontos, ou seja, $p(x_i) = y_i$, $i = 0, 1, \ldots, n$.

Demonstração: Seja $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ um polinômio de grau menor ou igual a n. Queremos determinar $a_0, a_1, \ldots, a_n$, de modo que

$$p(x_i) = y_i, \qquad i = 0, 1, 2, \ldots, n \tag{1.2}$$

ou seja, de forma que

$$\begin{aligned}
p(x_0) &= a_0 + a_1 x_0 + \ldots + a_n x_0^n = y_0 \\
p(x_1) &= a_0 + a_1 x_1 + \ldots + a_n x_1^n = y_1 \\
&\vdots \\
p(x_n) &= a_0 + a_1 x_n + \ldots + a_n x_n^n = y_n
\end{aligned}$$

Utilizando a notação matricial, temos

$$Xa = y \tag{1.3}$$

onde

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \quad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{e} \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

Precisamos provar que o sistema (1.3) tem uma única solução.

Suponhamos que $Xa = y$ não tenha solução única. Então $\det X = 0$ e o sistema homogêneo

$$Xa = 0 \tag{1.4}$$

admite soluções não triviais. Seja $\bar{a} = (\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_n)$, com pelo menos uma componente $\bar{a}_i \neq 0$, $0 \leqslant i \leqslant n$, uma solução do sistema (1.4).

Consideremos o polinômio

$$q(x) = \bar{a}_0 + \bar{a}_1 x + \ldots + \bar{a}_n x^n \tag{1.5}$$

cujos coeficientes são as componentes do vetor $\bar{a}$, solução de (1.4). Esse polinômio é não nulo e tem grau menor ou igual a n. No entanto, ele tem $n + 1$ raízes reais distintas, pois $q(x_i) = 0$, para $i = 0, 1, \ldots, n$, o que contradiz um resultado clássico de Álgebra*.

Portanto, o sistema homogêneo $Xa = 0$ não pode ter solução não trivial. Logo, $\det X \neq 0$, o que implica que o sistema $Xa = y$ tem uma única solução. ∎

Esta proposição, além de nos garantir a existência e unicidade do polinômio interpolador, nos dá um método para determiná-lo. Para isso basta resolver o sistema linear $Xa = y$, de ordem $n + 1$. Os coeficientes do polinômio interpolador serão as componentes da solução $\bar{a} = (\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_n)$ desse sistema.

Exemplo 1.1 Dada a seguinte tabela de valores da função $f(x) = e^x$, vamos interpolar o ponto $x = 1.32$:

x	1.3	1.4	1.5
e^x	3.669	4.055	4.482

Como conhecemos os valores da função em três pontos, vamos utilizar o polinômio interpolador de grau menor ou igual a dois, $p(x) = a_0 + a_1 x + a_2 x^2$.

* Seja $p(x) = a_0 + a_1 x + \ldots + a_n x^n$, $n \geqslant 1$, um polinômio não nulo, com coeficientes $a_i \in \mathbb{R}$, $i = 0, \ldots, n$. Então o número de raízes de $p(x)$ é menor ou igual a n. [Gonçalves], [Van der Waerden].

O sistema $Xa = y$ fica:

$$\begin{aligned} a_0 + a_1 * 1.3 + a_2 * 1.3^2 &= 3.669 \\ a_0 + a_1 * 1.4 + a_2 * 1.4^2 &= 4.055 \\ a_0 + a_1 * 1.5 + a_2 * 1.5^2 &= 4.482 \end{aligned}$$

Resolvendo-o, obtemos:

$$\begin{aligned} a_0 &= 2.382 \\ a_1 &= -1.675 \\ a_2 &= 2.05 \end{aligned}$$

Portanto,

$$p(x) = 2.382 - 1.675\,x + 2.05\,x^2$$

e

$$p(1.32) = 3.74292 \simeq 3.7430$$

Dada uma tabela

x	x_0	x_1	...	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	...	$f(x_n)$

(1.6)

correspondente aos valores de uma função f em $n + 1$ pontos distintos $x_0, x_1, \ldots, x_n$, sabemos que o polinômio interpolador de f

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n$$

de grau menor ou igual a n, existe e é único.

É fácil ver que a soma dos quadrados dos erros relativos aos pontos x_i, $i = 0, 1, \ldots, n$

$$\sum_{i=0}^{n} (f(x_i) - p(x_i))^2$$

é igual a zero, pois $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$.

Deste modo, se aproximarmos a função f tabelada em (1.6) por um polinômio de grau menor ou igual a n, pelo método dos mínimos quadrados, teremos como solução do sistema normal (IV-3.4) os coeficientes do polinômio interpolador.

Temos, portanto, duas maneiras distintas para calcular o polinômio interpolador de uma tabela: uma resolvendo o sistema linear $Xa = y$, conforme vimos na Proposição 1.1 e, a outra, aplicando o método dos mínimos quadrados.

Nas próximas seções veremos outros métodos para obter o polinômio interpolador.

2 – POLINÔMIO INTERPOLADOR NA FORMA DE LAGRANGE

Seja f uma função tabelada em $n + 1$ pontos distintos $x_0, x_1, \ldots, x_n$ e sejam os polinômios de grau n

$$L_i(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \quad (2.1)$$

denominados polinômios de Lagrange.

É fácil ver que para $0 \leqslant i, j \leqslant n$

$$L_i(x_j) = \begin{cases} 1 \text{ se } i = j \\ 0 \text{ se } i \neq j \end{cases} \quad (2.2)$$

Portanto, utilizando esses polinômios de Lagrange podemos determinar o polinômio interpolador de f relativamente aos pontos $x_0, x_1, \ldots, x_n$, da seguinte forma

$$p(x) = L_0(x)\,f(x_0) + L_1(x)\,f(x_1) + \ldots + L_n(x)\,f(x_n) \quad (2.3)$$

Como os polinômios $L_i(x)$ satisfazem as condições (2.2) é claro que $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, e além disso o grau de $p(x)$ é menor ou igual a n; portanto, o polinômio dado em (2.3) é o polinômio interpolador na forma de Lagrange.

Exemplo 2.1 Vamos repetir o Exemplo 1.1, utilizando agora o polinômio interpolador na forma de Lagrange. Vamos então interpolar o ponto $x = 1.32$ na seguinte tabela:

x	1.3	1.4	1.5
e^x	3.669	4.055	4.482

Da mesma forma, vamos utilizar o polinômio interpolador de grau menor ou igual a dois.

Os polinômios de Lagrange são:

$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

e

$$p(x) = L_0(x)\,f(x_0) + L_1(x)\,f(x_1) + L_2(x)\,f(x_2)$$

Portanto,

$$p(1.32) = \frac{(1.32-1.4)(1.32-1.5)}{(1.3-1.4)(1.3-1.5)} * 3.669 + \frac{(1.32-1.3)(1.32-1.5)}{(1.4-1.3)(1.4-1.5)} * 4.055 + \frac{(1.32-1.3)(1.32-1.4)}{(1.5-1.3)(1.5-1.4)} * 4.482$$

Logo,

$$p(1.32) = 3.74292 \simeq 3.7430$$

3 – POLINÔMIO INTERPOLADOR NA FORMA DE NEWTON

Sejam $x_0, x_1, \ldots, x_n$, $n + 1$ pontos distintos entre si e sejam $y_0, y_1, \ldots, y_n$ os correspondentes valores de uma função f nos pontos $x_j, j = 0, 1, \ldots, n$.

Para todo j, $j = 0, 1, \ldots, n$, consideremos $p_j(x)$ o polinômio interpolador de f relativamente aos pontos $x_0, x_1, \ldots, x_j$. É evidente que $p_0(x) = y_0$.

A construção de $p_n(x)$, polinômio interpolador de f em relação aos pontos $x_0, x_1, \ldots, x_n$, será feita de maneira recursiva, obtendo $p_j(x)$ a partir de $p_{j-1}(x)$, para $j = 1, \ldots, n$. Para isto vamos utilizar a seguinte recursão:

Como em (3.3), temos que

$$p_j^i(x) = f[x_i] + (x - x_i) f[x_i, x_{i+1}] + \ldots + (x - x_i) \ldots \\ \ldots (x - x_{j-1}) f[x_i, x_{i+1}, \ldots, x_j] \quad (3.6)$$

$0 \leqslant i \leqslant j \leqslant n$.

Observe que, para todo j, $j \leqslant n$, $p_j^0(x)$ e $p_j(x)$ são idênticos pois são polinômios interpoladores em relação aos pontos $x_0, x_1, \ldots, x_j$.

A seguir, provaremos resultados que facilitarão o cálculo de $f[x_i, x_{i+1}, \ldots, x_j]$, $0 \leqslant i < j \leqslant n$.

Proposição 3.1 Sejam y_0 e y_1 valores de uma função f calculados, respectivamente, em x_0 e x_1 distintos. Então

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

Demonstração: Pela definição 3.2

$$f[x_0, x_1] = \frac{y_1 - p_0(x_1)}{x_1 - x_0}$$

Sabemos que $y_1 = f(x_1) = f[x_1]$. Como $p_0(x)$ é o polinômio interpolador de grau zero que passa por x_0, $p_0(x_1) = y_0 = f[x_0]$. Logo,

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} \qquad \square$$

Proposição 3.2 Sejam $x_0, x_1, \ldots, x_j$, $0 \leqslant j \leqslant n$, $(j + 1)$ pontos distintos entre si, $y_0, y_1, \ldots, y_j$, os correspondentes valores de uma função f nesses pontos, $p_j^1(x)$ e $p_{j-1}(x)$ polinômios interpoladores em relação aos pontos $x_1, x_2, \ldots, x_j$ e $x_0, x_1, \ldots, x_{j-1}$, respectivamente. Então, para todo j, $2 \leqslant j \leqslant n$.

$$p_j^1(x) - p_{j-1}(x) = (x - x_1) \ldots (x - x_{j-1})(f[x_1, \ldots, x_j] - \\ - f[x_0, \ldots, x_{j-1}])$$

e

$$f[x_0, x_1, \ldots, x_j] = \frac{f[x_1, \ldots, x_j] - f[x_0, \ldots, x_{j-1}]}{x_j - x_0}$$

Demonstração: A demonstração será feita usando indução finita sobre j.

• Base da indução (j = 2): De (3.6) e (3.3) podemos escrever que:

$$p_2^1(x) - p_1(x) = f[x_1] + (x - x_1) f[x_1, x_2] - f[x_0] - \\ - (x - x_0) f[x_0, x_1] = -f[x_0] + f[x_1] - \\ - (x - x_0) f[x_0, x_1] + (x - x_1) f[x_1, x_2] \quad (3.7)$$

Da Proposição 3.1 sabemos que $-f[x_0] + f[x_1] = (x_1 - x_0) f[x_0, x_1]$. Substituindo em (3.7), temos:

$$p_2^1(x) - p_1(x) = (x_1 - x_0) f[x_0, x_1] - (x - x_0) f[x_0, x_1] + \\ + (x - x_1) f[x_1, x_2] = \\ (x_1 - x) f[x_0, x_1] + (x - x_1) f[x_1, x_2] = \\ (x - x_1)(f[x_1, x_2] - f[x_0, x_1]) \quad (3.8)$$

Pela definição 3.2, temos que

$$f[x_0, x_1, x_2] = \frac{y_2 - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

Como $p_2^1(x)$ é o polinômio interpolador relativamente aos pontos x_1, x_2, $p_2^1(x_2) = y_2$. Assim,

$$f[x_0, x_1, x_2] = \frac{p_2^1(x_2) - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)} \quad (3.9)$$

De (3.8),

$$p_2^1(x_2) - p_1(x_2) = (x_2 - x_1)(f[x_1, x_2] - f[x_0, x_1])$$

Substituindo em (3.9), temos que:

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

• Hipótese da indução: Vamos supor que para todo k, $2 \leqslant k < j$,

$$p_k^1(x) - p_{k-1}(x) = (x - x_1) \ldots (x - x_{k-1})(f[x_1, \ldots, x_k] - \\ - f[x_0, \ldots, x_{k-1}]) \quad (3.10)$$

e

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0} \quad (3.11)$$

• Passo da indução: Vamos provar que (3.10) vale para k = j.

Da construção recursiva dos polinômios $p_j^1(x)$ e $p_{j-1}(x)$, dada, respectivamente, em (3.4) e (3.1) temos que

$$p_j^1(x) = p_{j-1}^1(x) + (x - x_1) \ldots (x - x_{j-1}) \frac{y_j - p_{j-1}^1(x_j)}{(x_j - x_1) \ldots (x_j - x_{j-1})} \tag{3.12}$$

e

$$p_{j-1}(x) = p_{j-2}(x) + (x - x_0) \ldots \ldots (x - x_{j-2}) \frac{y_{j-1} - p_{j-2}(x_{j-1})}{(x_{j-1} - x_0) \ldots (x_{j-1} - x_{j-2})} \tag{3.13}$$

Usando as definições de diferenças divididas em (3.12) e (3.13), $p_j^1(x) - p_{j-1}(x)$ pode ser escrito como:

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & \{p_{j-1}^1(x) + (x - x_1) \ldots (x - x_{j-1}) f[x_1, \ldots, x_j]\} - \\ & - \{p_{j-2}(x) + (x - x_0) \ldots \\ & \ldots (x - x_{j-2}) f[x_0, \ldots, x_{j-1}]\} \end{aligned}$$

ou

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & p_{j-1}^1(x) - p_{j-2}(x) - (x - x_0) \ldots \\ & \ldots (x - x_{j-2}) f[x_0, \ldots, x_{j-1}] + \\ & + (x - x_1) \ldots (x - x_{j-1}) f[x_1, \ldots, x_j] \end{aligned}$$

Aplicando a hipótese da indução (3.10) para $p_{j-1}^1(x) - p_{j-2}(x)$, temos:

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & (x - x_1) \ldots (x - x_{j-2})(f[x_1, \ldots, x_{j-1}] - \\ & - f[x_0, \ldots, x_{j-2}]) - (x - x_0) \ldots \\ & \ldots (x - x_{j-2}) f[x_0, \ldots, x_{j-1}] + (x - x_1) \ldots \\ & \ldots (x - x_{j-1}) f[x_1, \ldots, x_j] \end{aligned}$$

Da hipótese da indução (3.11), temos que:

$$f[x_1, \ldots, x_{j-1}] - f[x_0, \ldots, x_{j-2}] = (x_{j-1} - x_0) f[x_0, x_1, \ldots, x_{j-1}]$$

Então,

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & (x - x_1) \ldots \\ & \ldots (x - x_{j-2})(x_{j-1} - x_0) f[x_0, x_1, \ldots, x_{j-1}] - \\ & - (x - x_0) \ldots (x - x_{j-2}) f[x_0, \ldots, x_{j-1}] + \\ & + (x - x_1) \ldots (x - x_{j-1}) f[x_1, \ldots, x_j] \end{aligned}$$

Colocando $(x - x_1) \ldots (x - x_{j-2}) f[x_0, \ldots, x_{j-1}]$ em evidência, temos:

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & -(x - x_{j-1})(x - x_1) \ldots \\ & \ldots (x - x_{j-2}) f[x_0, x_1, \ldots, x_{j-1}] + \\ & + (x - x_1) \ldots (x - x_{j-1}) f[x_1, \ldots, x_j] \end{aligned}$$

Colocando $(x - x_1) \ldots (x - x_{j-2})(x - x_{j-1})$ em evidência, temos:

$$\begin{aligned} p_j^1(x) - p_{j-1}(x) = {} & (x - x_1) \ldots (x - x_{j-1})(f[x_1, \ldots, x_j] - \\ & - f[x_0, \ldots, x_{j-1}]) \end{aligned} \tag{3.14}$$

Com isso provamos uma parte da proposição. Vamos provar agora que (3.11) vale para k = j.

Da definição da diferença dividida (3.2), sabemos que:

$$f[x_0, x_1, \ldots, x_j] = \frac{y_j - p_{j-1}(x_j)}{(x_j - x_0) \ldots (x_j - x_{j-1})}$$

Como $p_j^1(x)$ é o polinômio interpolador em relação aos pontos $x_1, \ldots, x_j$ e $p_j^1(x_j) = y_j$,

$$f[x_0, x_1, \ldots, x_j] = \frac{p_j^1(x_j) - p_{j-1}(x_j)}{(x_j - x_0) \ldots (x_j - x_{j-1})} \tag{3.15}$$

De (3.14), sabemos que:

$$\begin{aligned} p_j^1(x_j) - p_{j-1}(x_j) = {} & (x_j - x_1) \ldots (x_j - x_{j-1})(f[x_1, \ldots, x_j] - \\ & - f[x_0, \ldots, x_{j-1}]) \end{aligned}$$

Substituindo em (3.15), temos:

$$f[x_0, x_1, \ldots, x_j] = \frac{(x_j - x_1) \ldots (x_j - x_{j-1})(f[x_1, \ldots, x_j] - f[x_0, \ldots, x_{j-1}])}{(x_j - x_0) \ldots (x_j - x_{j-1})} =$$

$$= \frac{f[x_1, \ldots, x_j] - f[x_0, \ldots, x_{j-1}]}{(x_j - x_0)} \qquad \square$$

Proposição 3.3 Sejam x_i, $x_{i+1}, \ldots, x_j$, $(j - i + 1)$ pontos distintos entre si e y_i, $y_{i+1}, \ldots, y_j$, $0 \leqslant i < j \leqslant n$, os correspondentes valores de uma função f nestes pontos. Para todo par i, j, $0 \leqslant i < j \leqslant n$,

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{f[x_{i+1}, \ldots, x_j] - f[x_i, \ldots, x_{j-1}]}{x_j - x_i}$$

Demonstração: Vamos chamar os $(j - i + 1)$ pontos $x_i, x_{i+1}, \ldots, x_j$, respectivamente, de $x'_0, x'_1, \ldots, x'_{j-i}$. Assim,

$$f[x_i, x_{i+1}, \ldots, x_j] = f[x'_0, x'_1, \ldots, x'_{j-i}]$$

Pela Proposição 3.2,

$$f[x'_0, x'_1, \ldots, x'_{j-i}] = \frac{f[x'_1, \ldots, x'_{j-i}] - f[x'_0, x'_1, \ldots, x'_{j-i-1}]}{x'_{j-i} - x'_0}$$

Voltando à nomenclatura anterior, para i e j, $0 \leqslant i < j \leqslant n$,

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{f[x_{i+1}, \ldots, x_j] - f[x_i, \ldots, x_{j-1}]}{x_j - x_i} \qquad \square$$

Note que as Proposições 3.1 e 3.2 são casos particulares da Proposição 3.3 que foram demonstrados antes para facilidade de notação.

As diferenças divididas podem ser melhor visualizadas se colocadas numa tabela. Vamos denotar por f_i as diferenças divididas de ordem i, $0 \leqslant i \leqslant n$.

Tabela 3.1 Tabela de Diferenças Divididas

x_i	$y_i = f_0$	f_1	f_2	...	f_{n-1}	f_n
x_0	y_0					
x_1	y_1	$f[x_0, x_1]$				
x_2	y_2	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$			
x_3	y_3	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	...		
.	.	.	.		$f[x_0, x_1, \ldots, x_{n-1}]$	$f[x_0, x_1, \ldots, x_n]$
.	.	.	.		$f[x_1, x_2, \ldots, x_n]$	
x_{n-1}	y_{n-1}	.	$f[x_{n-2}, x_{n-1}, x_n]$	. . .		
		$f[x_{n-1}, x_n]$				
x_n	y_n					

Exemplo 3.1 Dada a seguinte tabela de valores:

x_i	0	1	3	4
y_i	−5	1	25	55

vamos interpolar o ponto $x = 0.5$, utilizando polinômio interpolador na forma de Newton.

Como temos 4 pontos tabelados, podemos utilizar o polinômio interpolador de grau menor ou igual a 3, dado por:

$$p_3(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + (x - x_0)(x - x_1)(x - x_2) f[x_0, x_1, x_2, x_3]$$

Vamos calcular as diferenças divididas usando a Tabela 3.1.

x_i	y_i	f_1	f_2	f_3
0	−5			
		6		
1	1		2	
		12		1
3	25		6	
		30		
4	55			

Portanto,

$$p_3(x) = -5 + 6(x - 0) + 2(x - 0)(x - 1) + 1(x - 0)(x - 1)(x - 3)$$
$$p_3(0.5) = -5 + 3 - 0.5 + 0.625 = -1.875$$

A forma explícita do polinômio interpolador da tabela dada relativamente aos pontos 0, 1, 3 e 4 é

$$p_3(x) = x^3 - 2x^2 + 7x - 5$$

Da mesma tabela de diferenças divididas podemos determinar, por exemplo, os polinômios:

$$p_3^1(x) = 1 + 12(x - 1) + 6(x - 1)(x - 3) =$$
$$= 6x^2 - 12x + 7 \quad \text{(polinômio interpolador relativamente aos pontos 1, 3 e 4)}$$

e

$$p_3^2(x) = 25 + 30(x - 3) =$$
$$= 30x - 65 \quad \text{(polinômio interpolador relativamente aos pontos 3 e 4)}$$

A seguir estudaremos o caso particular em que os valores de uma função f são tabelados em pontos eqüidistantes $x_0, x_1, \ldots, x_n$, isto é, a diferença entre dois pontos consecutivos é uma constante h, ou seja, $x_{j+1} = x_j + h$, $0 \leqslant j \leqslant n - 1$.

Para isso, vamos definir a diferença entre $f(x + h)$ e $f(x)$ por $\Delta f(x) = f(x + h) - f(x)$ denominada *diferença simples* de primeira ordem.

As diferenças simples de ordem j, $2 \leqslant j \leqslant n$, são definidas recursivamente da seguinte forma:

$$\Delta^j f(x) = \Delta^{j-1} f(x + h) - \Delta^{j-1} f(x) \tag{3.16}$$

A proposição a seguir relaciona as diferenças simples com as diferenças divididas.

Proposição 3.4 Sejam $x_0, x_1, \ldots, x_n$, $(n + 1)$ pontos distintos, tais que $x_{j+1} = x_j + h$, $0 \leqslant j \leqslant n - 1$ e $y_0, y_1, \ldots, y_n$, os correspondentes valores de uma função f nesses pontos. Para todo par i, j, $0 \leqslant i < j \leqslant n$,

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{\Delta^k f(x_i)}{k! h^k}$$

onde $k = j - i$.

Demonstração: Vamos fazer a demonstração usando indução finita sobre k.

- Base da indução: Para $k = 1$, usando a Proposição 3.3, temos:

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}$$

Como $f[x_{i+1}] = f(x_{i+1})$, $f[x_i] = f(x_i)$ e pelo fato de x_{i+1} ser $x_i + h$,

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{h}$$

Pela definição de diferença simples

$$f[x_i, x_{i+1}] = \frac{\Delta f(x_i)}{h}$$

- Hipótese da indução: Vamos supor que para todo ℓ, $1 \leqslant \ell < k$, $k = j - i$, $0 \leqslant i < j \leqslant n$,

$$f[x_i, x_{i+1}, \ldots, x_{i+\ell}] = \frac{\Delta^\ell f(x_i)}{\ell! h^\ell}$$

- Passo da indução: Pela Proposição 3.3,

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_j] - f[x_i, x_{i+1}, \ldots, x_{j-1}]}{x_j - x_i}$$

Usando a hipótese da indução e o fato de que $x_j - x_i = (j - i)h$,

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{\dfrac{\Delta^\ell f(x_{i+1})}{\ell! h^\ell} - \dfrac{\Delta^\ell f(x_i)}{\ell! h^\ell}}{(j - i)h} =$$
$$= \frac{\Delta^\ell f(x_{i+1}) - \Delta^\ell f(x_i)}{\ell!(j - i)h^\ell h}$$

Como $\ell = j - i - 1$ e $k = j - i$, usando a definição de diferença simples

$$f[x_i, x_{i+1}, \ldots, x_j] = \frac{\Delta^k f(x_i)}{k! h^k} \qquad \square$$

De (3.3) sabemos que o polinômio interpolador de f na forma de Newton em relação aos pontos $x_0, x_1, \ldots, x_n$ é:

$$p_n(x) = f[x_0] + (x - x_0) f[x_0, x_1] + \ldots + (x - x_0) \ldots$$
$$\ldots (x - x_{n-1}) f[x_0, \ldots, x_n]$$

Usando a Proposição 3.4, obtemos a fórmula do polinômio interpolador na forma de Newton com diferenças simples:

$$p_n(x) = f(x_0) + \frac{(x - x_0)}{h} \Delta f(x_0) + \ldots +$$
$$+ \frac{(x - x_0) \ldots (x - x_{n-1})}{h^n} \frac{\Delta^n f(x_0)}{n!} \tag{3.17}$$

Fazendo-se a mudança de variável $\frac{x - x_0}{h} = z$, temos:

$$p_n(x_0 + hz) = f(x_0) + z\Delta f(x_0) + z(z-1)\frac{\Delta^2 f(x_0)}{2!} + \ldots + z(z-1)(z-2)\ldots(z-n+1)\frac{\Delta^n f(x_0)}{n!} \quad (3.18)$$

Denotando $z(z-1)\ldots(z-i+1) = \binom{z}{i}$, podemos reescrever (3.18):

$$p_n(x_0 + hz) = f(x_0) + \binom{z}{1}\Delta f(x_0) + \binom{z}{2}\Delta^2 f(x_0) + \ldots + \binom{z}{n}\Delta^n f(x_0)$$

O cálculo das diferenças simples pode ser feito através da Definição 3.16, ou, de forma análoga às diferenças divididas, dispondo as diferenças simples como na Tabela 3.2. Nesta tabela denotamos as diferenças simples de ordem j simplesmente por Δ^j, $1 \leqslant j \leqslant n$.

Tabela 3.2 Tabela de diferenças simples

x_i	y_i	Δ^1	Δ^2	...	Δ^n
x_0	y_0	$\Delta f(x_0)$	$\Delta^2 f(x_0)$		
x_1	y_1	$\Delta f(x_1)$	$\Delta^2 f(x_1)$		
x_2	y_2	$\Delta f(x_2)$			
x_3	y_3			...	$\Delta^n f(x_0)$
.	.	.	.		
.	.	.	.		
.	.	.	.		
x_{n-1}	y_{n-1}	$\Delta f(x_{n-1})$	$\Delta^2 f(x_{n-2})$		
x_n	y_n				

Exemplo 3.2 Dada a tabela:

x_i	1	2	3	4
y_i	1	9	25	55

correspondente aos valores de uma função f, vamos interpolar o ponto $x = 2.5$.

Como os pontos são eqüidistantes podemos construir a tabela de diferenças simples:

x_i	y_i	Δ^1	Δ^2	Δ^3
1	1			
		8		
2	9		8	
		16		6
3	25		14	
		30		
4	55			

De (3.17),

$$p_3(x) = f(x_0) + \frac{x - x_0}{h}\Delta f(x_0) + \frac{(x-x_0)(x-x_1)}{h^2}\frac{\Delta^2 f(x_0)}{2} + \frac{(x-x_0)(x-x_1)(x-x_2)}{h^3}\frac{\Delta^3 f(x_0)}{\delta}$$

Então,

$$p_3(2.5) = 1 + (2.5-1)8 + (2.5-1)(2.5-2)4 + (2.5-1)(2.5-2)(2.5-3)1 = 15.625$$

Uma outra maneira de calcular $p_3(2.5)$ é fazer a mudança de variável

$$\frac{x - x_0}{h} = \frac{2.5 - 1}{1} = 1.5 = z$$

e usar a fórmula (3.18). Assim,

$$p_3(2.5) = f(x_0) + z\,\Delta f(x_0) + z(z-1)\frac{\Delta^2 f(x_0)}{2} + z(z-1)(z-2)\frac{\Delta^3 f(x_0)}{6} = 1 + 1.5*8 + 1.5(1.5-1)*4 + 1.5(1.5-1)(1.5-2)*1 = 15.625$$

4 – DELIMITAÇÃO DO ERRO DE TRUNCAMENTO NA INTERPOLAÇÃO POLINOMIAL

Seja f uma função contínua com (n + 1) derivadas contínuas em um intervalo I, $x_0, x_1, \ldots, x_n$, (n + 1) pontos distintos dois a dois, perten-

centes ao intervalo I e $p_n(x)$ o polinômio interpolador de f relativamente aos pontos $x_0, x_1, \ldots, x_n$.

Para cada $x \in I$, podemos definir o *erro de truncamento* na interpolação de x, usando $p_n(x)$, como sendo $E(x) = f(x) - p_n(x)$.

Para calcular E (x) vamos considerar as seguintes funções:

$$G(t) = (t - x_0)(t - x_1) \ldots (t - x_n)$$

e

$$H(t) = E(x)\,G(t) - E(t)\,G(x)$$

A função H (t) é contínua em I e derivável até ordem (n + 1), pois, por hipótese, f satisfaz essas condições e $p_n(t)$ e G (t) são polinômios.

É fácil ver que $G(x_i) = 0$ para $i = 0, 1, \ldots, n$. Como $f(x_i) = p_n(x_i)$, $i = 0, 1, \ldots, n$, $E(x_i) = 0$. Logo, $H(x_i) = 0$ para $i = 0, 1, \ldots, n$.

É fácil ver, também, que $H(x) = 0$.

Portanto, H tem (n + 2) raízes, $x, x_0, x_1, \ldots, x_n$, em I. Como H é contínua e derivável, pelo Teorema do Valor Médio*, segue que $H'(t)$ tem (n + 1) raízes, uma entre cada duas das (n + 2) raízes de H (t).

Como H é derivável até ordem (n + 1), aplicando o mesmo raciocínio, verificamos que

$H''(t)$ tem n raízes,

$H'''(t)$ tem (n − 1) raízes,

.

.

.

$H^{(n+1)}(t)$ tem uma raiz, a qual chamaremos de ξ.

Da definição de H (t) segue que a sua derivada de ordem (n + 1) em relação a t é

$$H^{(n+1)}(t) = E(x)\,G^{(n+1)}(t) - E^{(n+1)}(t)\,G(x)$$

* Teorema do Valor Médio: Seja f uma função real, definida e contínua num intervalo fechado [a, b] com primeira derivada contínua. Então $\exists \xi \in [a, b]$ tal que $f(b) - f(a) = f'(\xi)(b - a)$.

Como G (t) é um polinômio de grau (n + 1), $G^{(n+1)}(t) = (n+1)!$. Sendo $E(t) = f(t) - p_n(t)$ e $p_n(t)$ um polinômio de grau menor ou igual a n,

$$E^{(n+1)}(t) = f^{(n+1)}(t) - p_n^{(n+1)}(t) = f^{(n+1)}(t)$$

Portanto,

$$H^{(n+1)}(t) = (n+1)!\,E(x) - f^{(n+1)}(t)\,G(x)$$

Lembrando que ξ é raiz de $H^{(n+1)}(t)$, $H^{(n+1)}(\xi) = 0$. Então,

$$(n+1)!\,E(x) - f^{(n+1)}(\xi)\,G(x) = 0$$

Portanto,

$$E(x) = G(x)\,\frac{f^{(n+1)}(\xi)}{(n+1)!}$$

ou seja,

$$E(x) = (x - x_0)(x - x_1) \ldots (x - x_n)\,\frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad (4.1)$$

com $\xi \in I$.

Para utilizarmos a fórmula do erro (4.1) precisamos conhecer $f^{(n+1)}$ e o ponto $\xi \in I$. Mas, como em geral não conhecemos a forma analítica da função f não podemos determinar $f^{(n+1)}$ e conseqüentemente E (x). Mesmo se conhecermos a forma analítica de f, não saberemos o valor de ξ e, portanto, não poderemos calcular E (x) exatamente. Entretanto, podemos delimitar o erro pela desigualdade:

$$|E(x)| \leqslant \frac{|(x - x_0)(x - x_1) \ldots (x - x_n)|}{(n+1)!} \max_{z \in I} |f^{(n+1)}(z)| \qquad (4.2)$$

Exemplo 4.1 Dada a seguinte tabela para a função $f(x) = e^x$,

x	1.0	1.1	1.2
e^x	2.718	3.004	3.320

vamos calcular f (1.05) e delimitar o erro para o valor interpolado, utilizando aritmética de ponto flutuante com quatro algarismos significativos.

Como os pontos dados são eqüidistantes, podemos usar diferenças simples cuja tabela é:

x_i	$f(x_i)$	Δ^1	Δ^2
1.0	2.718		
		0.286	
1.1	3.004		0.03
		0.316	
1.2	3.320		

Como conhecemos três pontos vamos utilizar o polinômio interpolador de grau menor ou igual a dois. De (3.17), temos

$$p_2(x) = f(x_0) + \frac{(x - x_0)}{h} \Delta f(x_0) + \frac{(x - x_0)(x - x_1)}{h^2} \frac{\Delta^2 f(x_0)}{2}$$

Então

$$p_2(1.05) = 2.718 + \frac{(1.05 - 1.0)}{0.1} 0.286 + \frac{(1.05 - 1.0)(1.05 - 1.1)}{0.1^2} 0.015 = 2.857$$

A delimitação do erro de truncamento é dada por (4.2):

$$|E(x)| \leqslant \frac{|(x - x_0)(x - x_1)(x - x_2)|}{3!} \max_{z \in I} |f'''(z)|$$

Como $f(x) = e^x$, $f'''(x) = e^x$. Então,

$$\max_{z \in I} |f'''(z)| = e^{1.2} = 3.32$$

pois e^x é uma função crescente em I.

Portanto,

$$|E(x)| \leqslant \frac{|(1.05 - 1.0)(1.05 - 1.1)(1.05 - 1.2)|}{6} 3.32$$

$$|E(x)| \leqslant 0.0002075$$

5 – EXERCÍCIOS

1. Dada a tabela

x_i	5.9	6.0	6.1	6.2
y_i	34.8	36.0	37.2	38.4

a) determine o polinômio interpolador que passa pelos quatro pontos dados, usando o polinômio na forma de Lagrange;

b) determine o polinômio interpolador que passa pelos quatro pontos utilizando a forma de Newton;

c) determine o polinômio de grau menor ou igual a três que aproxima a função dada, pelo método dos mínimos quadrados. Compare-o com os obtidos nos itens anteriores;

d) calcule p(6.09).

2. Suponha que interpolamos um valor da função utilizando polinômio interpolador na forma de Newton e na forma de Lagrange, utilizando os mesmos pontos e os valores obtidos são diferentes. A que podemos atribuir essa diferença?

3. Dada a tabela da função ℓn

x_i	3.1	3.2
ℓn x_i	1.1314	1.1632

faça interpolação linear para calcular ℓn 3.16 e delimite o erro.

4. Dada a tabela de valores da função ℓn

x_i	0.40	0.50	0.70	0.80
ℓn x_i	−0.916291	−0.693147	−0.356675	−0.223144

a) estimar o valor de ℓn 0.60;

b) supondo-se que o valor "preciso" de ℓn 0.60 seja igual a −0.510826, analise o valor obtido em a) com a delimitação do erro.

5. Deseja-se construir uma tabela da função $f(x) = e^x$ no intervalo $[0, 1]$. Qual deve ser o passo da tabela (diferença entre duas abscissas consecutivas) para que o erro numa interpolação linear em qualquer parte da tabela seja menor que $\varepsilon > 0$? (Isto é, qual deve ser o passo h para que $|E(x)| \leqslant \varepsilon$, $x_k \leqslant x \leqslant x_k + h$, onde x_k e $x_k + h$ são dois pontos consecutivos quaisquer da tabela?)

6. Seja uma função f tabelada para valores x_i igualmente espaçados. Seja h o passo e suponhamos $|f''(x)| \leqslant M$ em todo o intervalo da tabela. Mostre que, ao se fazer uma interpolação linear da função f no ponto x, tomando pontos consecutivos x_i, x_{i+1} tais que $x_i \leqslant x < x_{i+1}$, o valor absoluto do erro cometido é no máximo $\frac{1}{8} \cdot M \cdot h^2$.

7. Dada a seguinte tabela de valores de uma função f

x_i	0	1	4
$f(x_i)$	1	−1	1

a) calcule o polinômio (de grau menor ou igual a dois), que passa pelos pontos tabelados, pela forma de Newton (verifique que neste caso não é possível usar diferenças simples);

b) dê o valor de p (3);

c) sabendo que

$$\max_{x \in [0,4]} |f'''(x)| = 1$$

delimite o erro de truncamento para cada um dos pontos dados. Podemos delimitar o erro de truncamento para $x = 5$? Por quê?

8. Mostre que a interpolação de um polinômio de grau n por $n + k$ pontos, $k \geqslant 1$, é exata, isto é, $E(x) = 0$, para todo x.

9. Seja Δ um *operador de diferenças* que aplicado a uma função f produz a diferença:

$$\Delta(f(x)) = f(x + h) - f(x), \quad \text{para certo } h > 0.$$

Mostre que esse operador Δ é *linear* sob as definições de adição e multiplicação por escalar, isto é,

$$\Delta(\alpha f(x) + \beta g(x)) = \alpha \Delta(f(x)) + \beta \Delta(g(x))$$

com α e β escalares quaisquer.

10. Mostre que, se f é uma função contínua com j derivadas contínuas num intervalo $(x, x + jh)$, j inteiro, $j > 0$ e $h > 0$, então existe ξ, tal que

$$\Delta^j(f(x)) = h^j f^{(j)}(\xi_j) \qquad x < \xi_j < x + jh$$

Sugestão: Demonstrar por indução sobre j. Usar o Teorema do Valor Médio. Precisamos definir, também, potências do operador de diferenças da seguinte forma:

$$\Delta^2(f(x)) = \Delta(\Delta(f(x)))$$
$$\Delta^3(f(x)) = \Delta(\Delta^2(f(x)))$$
$$\vdots \qquad\qquad \vdots$$
$$\Delta^n(f(x)) = \Delta(\Delta^{n-1}(f(x)))$$

11. Mostre que, se

$$x_{j+1} = x_j + h \qquad \forall j = 0, 1, \ldots, n-1, \; h > 0$$

então

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!} \qquad \xi \in [x_0, x_n]$$

12. Sabendo-se que

$$f[x_0, x_1, \ldots, x_n] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_n)} + \ldots + \frac{f(x_n)}{(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})}$$

a) Mostrar que

$$f[x_0, x_1, x_2] = f[x_0, x_2, x_1] = f[x_1, x_2, x_0]$$

b) Mostrar que

$$f(x) = p_2(x) + f[x_0, x_1, x_2, x](x - x_0)(x - x_1)(x - x_2)$$

onde $p_2(x)$ é o polinômio interpolador de $f(x)$ relativamente aos pontos x_0, x_1, x_2.

c) Mostrar que

$$f[x_0, x_1, \ldots x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

13. "As diferenças divididas de ordem n de um polinômio de grau n são constantes e as de ordem (n + 1) são nulas." Justifique esta afirmação.

14. Sejam a, b, c, d números reais. Mostre que existe polinômio p, de grau menor ou igual a dois, tal que:

$$p(-1) = a; \quad p(0) = b; \quad p(1) = c; \quad p(3) = d$$

se e somente se $3a - 8b + 6c - d = 0$.

15. Dada a seguinte tabela de valores

x_i	-2	-1	0	1	2	3
$p(x_i)$	-7	0	1	α	9	28

sabendo que $p(x)$ é um polinômio de grau três, determine $p(1)$:

a) usando diferenças simples;

b) usando diferenças divididas (escolha quaisquer 4 dos 5 pontos dados).

16. No Capítulo 2 estudamos alguns métodos para determinar o zero de uma função, isto é, dada uma função f, encontrar $\bar{x}$ tal que $f(\bar{x}) = 0$. Agora, supondo-se que a função f é inversível e é conhecida em n pontos x_i, $i = 1, 2, \ldots, n$, tais que $f(x_i) \neq 0$, $i = 1, 2, \ldots, n$, explique como podemos resolver o problema acima, isto é, determinar $\bar{x}$ tal que $f(\bar{x}) = 0$ usando somente os conhecimentos relativos à interpolação polinomial.

17. Dada a seguinte tabela de valores de uma função f

x	0.1	0.2	0.3	0.4	0.5
f(x)	0.70010	0.40160	0.10810	−0.17440	−0.43750

Supondo que f seja uma função inversível, calcule $\bar{x}$ tal que $f(\bar{x}) = 0$.

18. Dada a função $f(x) = \frac{3 + x}{1 + x}$ pede-se:

a) calcular um polinômio interpolador de grau dois que aproxima a função no intervalo [0, 2];

b) obter o valor aproximado da integral da função f(x) no intervalo [0, 2] através do polinômio interpolador obtido em a).

19. Seja g(x) um polinômio de grau $2n + 1$ e $P_0(x), P_1(x), P_2(x), \ldots, P_{n+1}(x)$ polinômios de grau $0, 1, 2, \ldots, n + 1$, respectivamente, que são ortogonais entre si em relação ao produto escalar

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

Sabe-se que $P_{n+1}(x)$ possui $(n + 1)$ zeros reais distintos $x_0, x_1, \ldots, x_n$ no interior do intervalo [a, b]. Seja r(x) o resto da divisão de g(x) por $P_{n+1}(x)$, isto é,

$$g(x) = P_{n+1}(x)\, q(x) + r(x)$$

onde q(x) e r(x) são polinômios em x e grau $r(x) \leq n$. Mostre que:

a) r(x) é o polinômio interpolador de g(x) em relação aos pontos $x_0, x_1, \ldots, x_n$;

b) $\int_a^b g(x)\, dx = \int_a^b r(x)\, dx.$

c) a diferença dividida de g(x) relativamente aos pontos $x_0, x_1, \ldots x_n, x$ é um polinômio de grau $\leq n$, em x.

Capítulo 6
Integração Numérica

Em muitas aplicações da Matemática é necessário efetuar o cálculo da integral de alguma função. Entretanto, muitas vezes não podemos obter uma fórmula explícita simples para a integral indefinida desejada. Em outros casos a função pode ser conhecida apenas por seus valores em alguns pontos. Por exemplo, quando medimos a velocidade de um objeto em vários instantes e queremos saber o espaço que ele percorreu neste intervalo de tempo, precisamos integrar a função velocidade.

Neste capítulo trataremos do estudo de alguns métodos numéricos para calcular a integral definida de uma função. Uma fórmula que forneça um valor numérico aproximado da integral de uma função é chamada de *quadratura numérica* ou *fórmula de integração numérica*. Serão vistos dois métodos. O primeiro se aplica a funções com valores conhecidos em pontos igualmente espaçados e são as chamadas fórmulas (fechadas) de Newton-Cotes. Dentre estas, veremos as fórmulas dos trapézios e de Simpson. O segundo método se aplica a funções com valores conhecidos em pontos pré-determinados, não necessariamente eqüidistantes e são as chamadas fórmulas de Gauss ou quadratura de Gauss.

A idéia central utilizada nesses dois métodos é a integração do polinômio interpolador.

Suporemos sempre que a função f a ser integrada é definida e contínua no intervalo fechado [a, b] e com tantas derivadas contínuas em [a, b] quantas forem necessárias.

1 – FÓRMULAS DE NEWTON-COTES

As fórmulas de Newton-Cotes são aplicadas na integração de funções conhecidas em pontos eqüidistantes. Estas fórmulas são de dois tipos: fechadas ou abertas. As fórmulas fechadas usam também os valores da função a ser integrada nos extremos do intervalo de integração; as fórmulas abertas não usam esses valores. Neste texto estudaremos apenas as fórmulas fechadas.

a) Fórmula dos trapézios

A integral de uma função f no intervalo [a, b] pode ser aproximada pela área de um trapézio, conforme a Fig. 1.1.

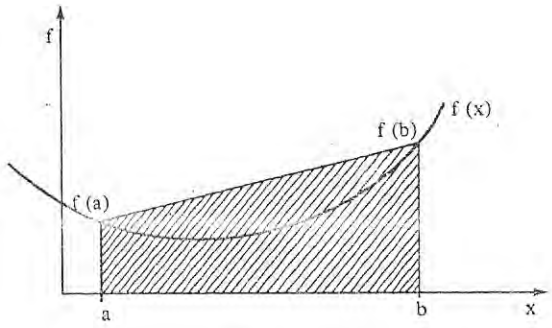


Fig. 1.1

Assim, teremos:

$$\int_a^b f(x)\,dx \simeq (f(a) + f(b))\,\frac{(b-a)}{2} \qquad (1.1)$$

O valor fornecido por (1.1) é uma aproximação do valor exato da integral. Assim, existe um erro, dado pela diferença:

$$E = \int_a^b f(x)\,dx - (f(a) + f(b))\,\frac{(b-a)}{2} \qquad (1.2)$$

Vamos calcular este erro. Para isto é importante notar que a área $(f(a) + f(b))\ \frac{(b-a)}{2}$ é a integral exata, no intervalo [a, b], da reta que passa pelos pontos (a, f(a)) e (b, f(b)). Esta reta é o polinômio interpolador da função f em relação aos pontos a e b. Sabemos do capítulo anterior que, para cada $x \in [a, b]$ temos:

$$f(x) = f[a] + (x-a) f[a, b] + R(x)$$

onde

$$R(x) = (x-a)(x-b)\frac{f''(\xi_x)}{2!} \qquad \xi_x \in [a, b]$$

Logo,

$$\int_a^b f(x)\,dx = \int_a^b (f[a] + (x-a) f[a, b])\,dx + \int_a^b R(x)\,dx$$

Como

$$\int_a^b (f[a] + (x-a) f[a, b])\,dx = \frac{(f(a) + f(b))(b-a)}{2}$$

temos de (1.2) que

$$E = \int_a^b R(x)\,dx = \frac{1}{2!}\int_a^b (x-a)(x-b) f''(\xi_x)\,dx \qquad (1.3)$$

A proposição seguinte fornece o valor deste erro.

Proposição 1.1 O erro cometido ao calcular a integral da função f no intervalo [a, b] pela fórmula dos trapézios é dado por:

$$E = \frac{-h^3}{12} f''(\varepsilon) = \frac{-(b-a)^3}{12} f''(\varepsilon) \qquad \varepsilon \in [a, b] \qquad (1.4)$$

Prova: De (1.3) temos que:

$$E = \frac{1}{2}\int_a^b (x-a)(x-b) f''(\xi_x)\,dx$$

Vamos fazer a mudança de variável $x(\alpha) = a + \alpha h$ onde $h = b - a$, $\alpha \in [0, 1]$, de modo a tornar a integral independente dos particulares valores de a e b. Assim, $dx = h d\alpha$ e existe $\xi_\alpha \in [0, 1]$ correspondente a $\xi_x \in [a, b]$.

Segue que:

$$E = \frac{1}{2}\int_0^1 (\alpha h)(\alpha h - h) f''(a + h\xi_\alpha) h\,d\alpha =$$

$$= \frac{h^3}{2}\int_0^1 \alpha(\alpha - 1) f''(a + h\xi_\alpha)\,d\alpha$$

Como $f''(a + h\xi_\alpha)$ é contínua para $\xi_\alpha \in [0, 1]$ e $\alpha(\alpha - 1)$ é contínua e não muda de sinal para $\alpha \in [0, 1]$, podemos aplicar o Teorema da Média do Cálculo Integral* obtendo:

$$E = \frac{h^3}{2} f''(a + h\xi_\alpha)\int_0^1 \alpha(\alpha - 1)\,d\alpha \qquad \xi_\alpha \in [0, 1]$$

e fazendo $\varepsilon = a + h\xi_\alpha$ vem

$$E = -\frac{h^3}{12} f''(\varepsilon) \qquad \varepsilon \in [a, b] \qquad \square$$

Como em (1.4) não conhecemos o valor de ε, podemos apenas delimitar o valor do erro.

Como

$$|E| = \frac{(b-a)^3}{12} |f''(\varepsilon)| \qquad \varepsilon \in [a, b]$$

temos

$$|E| \leqslant \frac{(b-a)^3}{12} \max_{x \in [a,b]} |f''(x)| \qquad (1.5)$$

Exemplo 1.1 Vamos calcular a integral de $f(x) = \sqrt{6x - 5}$ no intervalo [1, 9], pela fórmula dos trapézios.

* Sejam f e g funções reais, definidas e contínuas no intervalo fechado [a, b]. Se g(x) não muda de sinal no intervalo [a, b], existe $\xi \in [a, b]$ tal que

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

A demonstração do Teorema se encontra em [Apostol].

Neste caso, para $x = 1, f(x) = 1$ e para $x = 9, f(x) = 7$.

Portanto,

$$\int_1^9 \sqrt{6x-5} \simeq (f(1) + f(9)) \frac{(9-1)}{2} = 32$$

O erro cometido será, no máximo, 384, pois

$$|E| \leqslant \frac{8^3}{12} \max_{x\in[1,9]} \left| \frac{d^2\sqrt{6x-5}}{dx^2} \right| = \frac{128}{3} \max_{x\in[1,9]} \left| -9(6x-5)^{-3/2} \right| =$$

$$= \frac{128}{3} 9 = 384$$

Para reduzir o erro cometido podemos subdividir o intervalo [a, b] em n subintervalos de mesmo comprimento $h = \frac{b-a}{n}$ e considerar, como aproximação da integral, a soma das áreas dos trapézios obtidos a partir de cada um dos n subintervalos.

Uma ilustração dessa situação é mostrada na Fig. 1.2

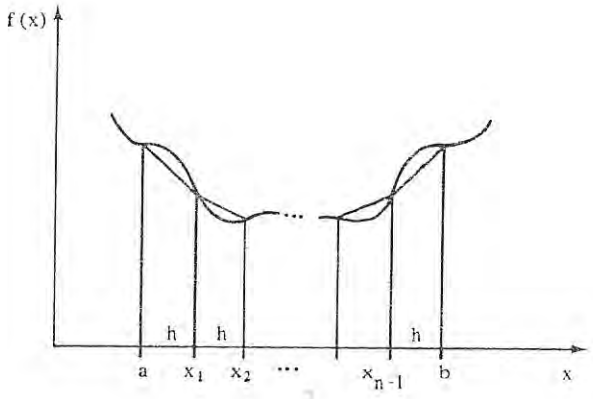


Fig. 1.2

onde $x_{i+1} = x_i + h, i = 0, 1, \ldots, n-1, x_0 = a$ e $x_n = b$.

Sabemos que:

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_n} f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)\,dx \tag{1.6}$$

Em cada subintervalo $[x_{i-1}, x_i], i = 1, 2, \ldots, n$, temos:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \frac{h}{2}(f(x_{i-1}) + f(x_i)) + E_i \tag{1.7}$$

onde

$$E_i = -\frac{h^3}{12} f''(\varepsilon_i) \qquad \varepsilon_i \in [x_{i-1}, x_i] \tag{1.8}$$

Substituindo (1.7) em (1.6), vem:

$$\int_a^b f(x)\,dx = \frac{h}{2}(f(x_0) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-1}) + f(x_n)) + \sum_{i=1}^{n} E_i \tag{1.9}$$

Portanto,

$$\int_a^b f(x)\,dx \simeq \frac{h}{2}(f(x_0) + 2f(x_1) + \ldots + 2f(x_{n-1}) + f(x_n)) \tag{1.10}$$

que é a fórmula dos trapézios repetida para n subintervalos.

Proposição 1.2 O erro cometido ao calcular a integral da função f no intervalo [a, b] pela fórmula dos trapézios repetida para n subintervalos é dado por

$$E_T = \frac{-nh^3}{12} f''(\varepsilon) = -\frac{(b-a)^3}{12n^2} f''(\varepsilon), \qquad \varepsilon \in [a, b] \tag{1.11}$$

Prova: De (1.8) temos que

$$E_i = -\frac{h^3}{12} f''(\varepsilon_i), \quad \varepsilon_i \in [x_{i-1}, x_i]$$

De (1.9)

$$E_T = \sum_{i=1}^{n} E_i = -\frac{h^3}{12} \sum_{i=1}^{n} f''(\varepsilon_i) = -\frac{(b-a)^3}{12n^3} \sum_{i=1}^{n} f''(\varepsilon_i)$$

$\varepsilon_i \in [x_{i-1}, x_i]$

Como $f''(x)$ é contínua e, portanto, assume todos os valores entre seu valor máximo e mínimo em $[a, b]$ existe algum $\varepsilon \in [a, b]$ para o qual

$$f''(\varepsilon) = \frac{\sum_{i=1}^{n} f''(\varepsilon_i)}{n}$$

Deste modo, obtemos:

$$E_T = -\frac{(b-a)^3}{12n^3} n f''(\varepsilon)$$

Portanto,

$$E_T = -\frac{(b-a)^3}{12n^2} f''(\varepsilon)$$

onde $\varepsilon \in [a, b]$. □

Analogamente ao que foi feito em (1.5), a delimitação do erro é dada por:

$$|E_T| \leqslant \frac{(b-a)^3}{12n^2} \max_{x \in [a,b]} |f''(x)| \qquad (1.12)$$

Exemplo 1.2 Vamos repetir o Exemplo 1.1, considerando $h = 1$, ou seja, vamos calcular a integral da função $f(x) = \sqrt{6x-5}$ no intervalo $[1, 9]$ repetindo a fórmula dos trapézios 8 vezes. Podemos construir a seguinte tabela:

x	1	2	3	4	5	6	7	8	9
f(x)	1	2.65	3.61	4.36	5.0	5.57	6.08	6.56	7.0

Utilizando a fórmula (1.9), temos

$$\int_1^9 \sqrt{6x-5}\,dx \simeq \frac{1}{2}(1 + 5.3 + 7.22 + 8.72 + 10 + 11.1 + 12.2 + 13.1 + 7) = \frac{1}{2}(75.6) = 37.8$$

Vamos delimitar o erro cometido nesta aproximação. Como

$$\max_{x \in [1,9]} |f''(x)| = 9$$

de (1.12), temos:

$$|E_T| \leqslant \frac{(9-1)^3}{12 * 8^2} 9 = 6$$

Observe que, neste caso particular, a função f pode ser integrada de forma exata e

$$\int_1^9 \sqrt{6x-5}\,dx = 38$$

Exemplo 1.3 Vamos calcular

$$\int_6^{10} \log x\,dx$$

repetindo a fórmula dos trapézios 8 vezes e delimitar o erro cometido. Para isto vamos construir a seguinte tabela:

x	log x
6.0	0.77815125
6.5	0.81291336
7.0	0.84509804
7.5	0.87506126
8.0	0.90308999
8.5	0.92941893
9.0	0.95424251
9.5	0.97772361
10.0	1.0

$$h = \frac{b-a}{n} = \frac{10-6}{8} = 0.5$$

$$\int_6^{10} \log x\, dx \simeq \frac{0.5}{2}(0.77815125 + 1.6258267 + 1.6901961 +$$
$$+ 1.7501225 + 1.80618 + 1.8588379 +$$
$$+ 1.908485 + 1.9554472 + 1.0) =$$
$$= 3.5933118$$

A delimitação do erro é:

$$|E_T| \leqslant \frac{(10-6)^3}{12*8^2} \max_{x\in[6,10]} \left|\frac{d^2 \log x}{dx^2}\right| =$$
$$= \frac{4^3}{12*8^2} \max_{x\in[6,10]} \left|\left[-\frac{1}{x^2}\right] \log e\right| =$$
$$= \frac{1}{12} * \frac{1}{6^2} \log e = 0.0010053113$$

O valor exato da integral é

$$\int_6^{10} \log x\, dx = 3.5939146$$

b) Fórmula de Simpson

A fórmula dos trapézios para integração de uma função f num intervalo [a, b] foi deduzida a partir da integral do polinômio interpolador de 1º grau. Neste item repetiremos o processo, utilizando o polinômio interpolador de 2º grau, $p_2(x)$, que passa pelos pontos igualmente espaçados (a, f(a)), (m, f(m)) e (b, f(b)), conforme a fig. 1.3.

Assim, como vimos no capítulo anterior, tomando $h = \frac{b-a}{2}$ temos:

$$f(x) = p_2(x) + R_2(x) \qquad \forall x \in [a, b] \qquad (1.13)$$

onde

$$p_2(x) = f(a) + (x-a)\frac{\Delta f(a)}{h} + (x-a)(x-m)\frac{\Delta^2 f(a)}{2h^2} \qquad (1.14)$$

e

$$R(x) = (x-a)(x-m)(x-b)\frac{f'''(\xi)}{3!} \quad \text{com } \xi \in [a, b] \qquad (1.15)$$

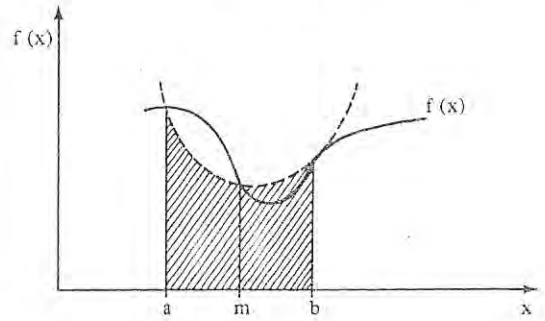


Fig. 1.3

Portanto, de (1.13) temos:

$$\int_a^b f(x)\, dx = \int_a^b p_2(x)\, dx + \int_a^b R_2(x)\, dx$$

Vamos aproximar

$$\int_a^b f(x)\, dx \quad \text{por} \quad \int_a^b p_2(x)\, dx$$

A fórmula de Simpson nos fornece o valor da integral de $p_2(x)$ no intervalo [a, b] em função dos valores de f nos pontos a, b e no ponto médio m de [a, b]. Vamos então determinar esta fórmula.

$$\int_a^b p_2(x)\, dx = \int_a^b \left[f(a) + (x-a)\frac{\Delta f(a)}{h} + \right.$$
$$\left. + (x-a)(x-m)\frac{\Delta^2 f(a)}{2h^2}\right] dx$$

Para facilitar o cálculo, vamos fazer a mudança de variável $x(\alpha) = a + \alpha h$. Assim, enquanto x percorre o intervalo [a, b], α percorre o intervalo [0, 2] e temos também que $dx = h d\alpha$.

Desta maneira:

$$\int_a^b p_2(x)\,dx = \int_0^2 \left[f(a) + \alpha\Delta f(a) + \alpha(\alpha-1)\frac{\Delta^2 f(a)}{2}\right] h\,d\alpha =$$

$$= h\left[f(a)\,\alpha\Big|_0^2 + \Delta f(a)\,\frac{\alpha^2}{2}\Big|_0^2 + \right.$$

$$\left. + \frac{\Delta^2 f(a)}{2}\left[\frac{\alpha^3}{3} - \frac{\alpha^2}{2}\right]\Big|_0^2\right] =$$

$$= h\,[2f(a) + 2(f(m) - f(a)) + \frac{1}{3}(f(b) -$$

$$- 2f(m) + f(a))]$$

Logo,

$$\int_a^b f(x)\,dx \simeq \frac{h}{3}[f(a) + 4f(m) + f(b)] \qquad (1.16)$$

Esta fórmula é denominada *fórmula de Simpson*.

A proposição a seguir fornece o valor do erro cometido ao aplicar a fórmula de Simpson para calcular a integral de f no intervalo [a, b].

Proposição 1.3. O erro cometido ao calcular a integral da função f no intervalo [a, b] pela fórmula de Simpson é dado por:

$$E = -\frac{h^5}{90}f^{IV}(\xi) = -\frac{(b-a)^5}{2880}f^{IV}(\xi) \qquad \xi \in [a, b]$$

Prova: Vamos denotar por F (x) a primitiva de f (x), ou seja,

$$F(x) = \int f(x)\,dx + c$$

onde c é uma constante.

Desta maneira temos:

$$F'(x) = f(x)$$

$$F''(x) = f'(x) \text{ etc.}$$

Além disso, fazendo $x_0 = m$, m ponto médio de [a, b], temos $a = x_0 - h$ e $b = x_0 + h$. Assim, o erro cometido é dado por:

$$E = E(h) = \int_{x_0-h}^{x_0+h} f(x)\,dx - \frac{h}{3}[f(x_0-h) + 4f(x_0) +$$

$$+ f(x_0+h)] = F(x_0+h) - F(x_0-h) - \frac{h}{3}[f(x_0-h) +$$

$$+ 4f(x_0) + f(x_0+h)]$$

Calculando as derivadas sucessivas de E (h) em relação a h, temos:

$$E'(h) = f(x_0+h) + f(x_0-h) - \frac{1}{3}[f(x_0-h) + 4f(x_0) +$$

$$+ f(x_0+h)] + \frac{h}{3}[f'(x_0-h) - f'(x_0+h)]$$

$$E''(h) = f'(x_0+h) - f'(x_0-h) - \frac{1}{3}[f'(x_0+h) - f'(x_0-h)] +$$

$$+ \frac{1}{3}[f'(x_0-h) - f'(x_0+h)] + \frac{h}{3}[-f''(x_0-h) -$$

$$- f''(x_0+h)] = f'(x_0+h) - f'(x_0-h) + \frac{2}{3}[f'(x_0-h) -$$

$$- f'(x_0+h)] - \frac{h}{3}[f''(x_0-h) + f''(x_0+h)]$$

$$E'''(h) = f''(x_0+h) + f''(x_0-h) - \frac{2}{3}[f''(x_0-h) + f''(x_0+h)] -$$

$$- \frac{1}{3}[f''(x_0-h) + f''(x_0+h)] - \frac{h}{3}[f'''(x_0+h) -$$

$$- f'''(x_0-h)] = -\frac{h}{3}[f'''(x_0+h) - f'''(x_0-h)]$$

Aplicando o Teorema da Média do Cálculo Diferencial* (ou Teorema do Valor Médio) temos:

$$E'''(h) = -\frac{h}{3}[f^{IV}(\xi)(x_0 + h - x_0 + h)]$$

onde $x_0 - h \leqslant \xi \leqslant x_0 + h$, ou seja,

$$E'''(h) = -\frac{2h^2}{3}f^{IV}(\xi)$$

* Ver (*) na página 16.

Agora vamos integrar, por três vezes consecutivas, $E'''(h)$ no intervalo $[0, h]$ e em cada integração aplicar o Teorema da Média do Cálculo Integral*. Como $E(0) = E'(0) = E''(0) = 0$, temos:

$$E''(h) = \int_0^h E'''(t)\,dt = \int_0^h -\frac{2t^2}{3} f^{IV}(\xi)\,dt =$$

$$= -f^{IV}(\xi_1) \int_0^h \frac{2t^2}{3}\,dt = -\frac{2}{9} h^3 f^{IV}(\xi_1)$$

$$E'(h) = \int_0^h E''(t)\,dt = \int_0^h -\frac{2}{9} t^3 f^{IV}(\xi_1)\,dt =$$

$$= -f^{IV}(\xi_2) \int_0^h \frac{2}{9} t^3\,dt = -\frac{h^4}{18} f^{IV}(\xi_2)$$

$$E(h) = \int_0^h E'(t)\,dt = \int_0^h -\frac{t^4}{18} f^{IV}(\xi_2)\,dt =$$

$$= -f^{IV}(\xi_3) \int_0^h \frac{t^4}{18}\,dt = -\frac{h^5}{90} f^{IV}(\xi_3)$$

onde $\xi_1, \xi_2, \xi_3 \in [a, b]$.

Portanto,

$$E = -\frac{h^5}{90} f^{IV}(\xi) = -\left[\frac{b-a}{2}\right]^5 \frac{1}{90} f^{IV}(\xi) = -\frac{(b-a)^5}{2880} f^{IV}(\xi) \tag{1.17}$$

$a \leqslant \xi \leqslant b$. □

Analogamente ao que foi feito em (1.5), a delimitação do erro é dada por

$$|E| \leqslant \frac{(b-a)^5}{2880} \max_{x \in [a,b]} |f^{IV}(x)| \tag{1.18}$$

* Ver (*) na página 161.

Exemplo 1.4 Vamos calcular

$$\int_6^{10} \log x\,dx$$

aplicando a fórmula de Simpson.

Como

$$\log 6 = 0.77815125$$
$$\log 8 = 0.90308999$$
$$\log 10 = 1.0$$

e

$$h = \frac{b-a}{2} = \frac{10-6}{2} = 2$$

então

$$\int_6^{10} \log x\,dx \simeq \frac{2}{3}(0.77815125 + 3.61236 + 1.0) =$$

$$= \frac{2}{3}(5.3905113) = 3.5936742$$

Vamos delimitar o erro cometido nesta aproximação.

$$|E| \leqslant \frac{4^5}{2880} \max_{x \in [6,10]} \left| \frac{d^4 \log x}{dx^4} \right| . =$$

$$= \frac{4^5}{2880} \max_{x \in [6,10]} \left| \left[-\frac{6}{x^4}\right] \log e \right| = \frac{4^5}{2880} \frac{\log e}{6^3} = 0.00071488783$$

Portanto,

$$|E| \leqslant 7.1488783 \times 10^{-4}$$ □

Analogamente ao que foi feito para a fórmula dos trapézios, vamos repetir a aplicação da fórmula de Simpson n vezes, dividindo o intervalo $[a, b]$ em $2n$ subintervalos igualmente espaçados. A Fig. 1.4 ilustra como faremos esta integração, sendo $x_i = x_0 + ih$, $i = 1, 2, \ldots, 2n$, $h = \frac{b-a}{2n}$ e $x_0 = a$.

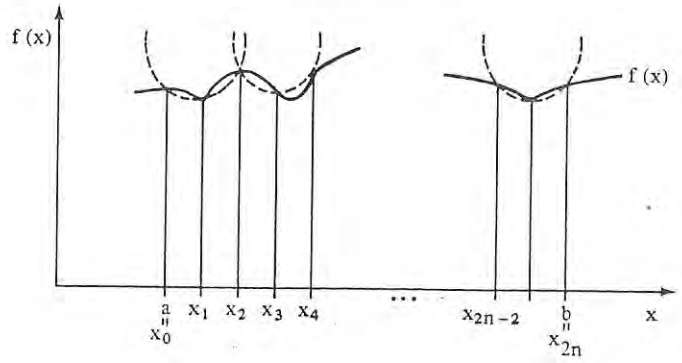


Fig. 1.4

Para cada intervalo $[x_{2i-2}, x_{2i}]$ a fórmula (1.16) nos fornece:

$$\int_{x_{2i-2}}^{x_{2i}} f(x)\,dx \simeq \frac{h}{3}[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})]$$

Com isso temos:

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_{2n}} f(x)\,dx = \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \tag{1.19}$$

$$+ \ldots + \int_{x_{2n-2}}^{x_{2n}} f(x)\,dx \simeq \frac{h}{3}[f(x_0) + 4f(x_1) +$$

$$+ 2f(x_2) + 4f(x_3) + \ldots + 2f(x_{2n-2}) +$$

$$+ 4f(x_{2n-1}) + f(x_{2n})]$$

que é a fórmula de Simpson repetida n vezes (2 n subintervalos).

Proposição 1.4 O erro cometido ao calcular a integral da função f no intervalo [a, b] pela fórmula de Simpson repetida n vezes é dado por

$$E_T = -\frac{nh^5}{90} f^{IV}(\varepsilon) = -\frac{(b-a)^5}{2880\,n^4} f^{IV}(\varepsilon)$$

onde $\varepsilon \in [a, b]$.

Prova: O erro total é calculado somando-se os erros de cada aplicação da fórmula de Simpson. Assim,

$$E_T = \sum_{i=1}^{n} E_i = \sum_{i=1}^{n} -\frac{h^5}{90} f^{IV}(\xi_i) \qquad x_{2i-2} \leqslant \xi_i \leqslant x_{2i}$$

Como $f^{IV}(x)$ é uma função contínua e portanto assume todos os valores entre seu valor máximo e mínimo em [a, b], existe algum $\varepsilon \in [a, b]$ para o qual

$$f^{IV}(\varepsilon) = \frac{\sum_{i=1}^{n} f^{IV}(\xi_i)}{n}$$

Logo,

$$E_T = -\frac{h^5}{90} n f^{IV}(\varepsilon) = -\frac{(b-a)^5}{2880\,n^5} n f^{IV}(\varepsilon) = \tag{1.20}$$

$$= -\frac{(b-a)^5}{2880\,n^4} f^{IV}(\varepsilon)$$

onde $a \leqslant \varepsilon \leqslant b$. □

Analogamente ao que foi feito em (1.5), a delimitação do erro é dada por:

$$|E_T| \leqslant \frac{(b-a)^5}{2880\,n^4} \max_{x \in [a,b]} |f^{IV}(x)| \tag{1.21}$$

Exemplo 1.5 Vamos calcular

$$\int_6^{10} \log x\,dx$$

utilizando a fórmula de Simpson repetida 4 vezes (para 8 subintervalos) e a tabela construída no Exemplo 1.3.

Neste caso

$$h = \frac{b-a}{2n} = \frac{10-6}{8} = 0.5$$

Utilizando a fórmula (1.19) temos:

$$\int_6^{10} \log x\, dx \simeq \frac{0.5}{3}\,(0.77815125 + 3.2516534 + 1.6901961 + 3.5002450 + 1.80618 + 3.176757 + 1.9084850 + 3.1908944 + 1.0) = 3.5939136$$

Vamos delimitar o erro cometido nesta aproximação. De (1.21), temos:

$$|E_T| \leqslant \frac{4^5}{2\,800 * 4^4}\,\frac{\log e}{6^3} = 0.000002792528$$

Portanto,

$$|E_T| \leqslant 2.792528 \times 10^{-6}$$

Observação final: Na prática, as fórmulas de integração mais utilizadas são as dos trapézios e de Simpson. Entretanto, o raciocínio utilizado para determinar essas fórmulas pode ser estendido para obtermos interpoladores de grau maior ou igual a 3.

2 – FÓRMULAS DE GAUSS

As fórmulas de Newton-Cotes vistas na seção anterior utilizam polinômios interpoladores para pontos eqüidistantes na integração de uma função f. Estas fórmulas são exatas para integração de polinômios de grau menor ou igual a n, onde n é o grau do polinômio interpolador.

Nesta seção, vamos desenvolver fórmulas para integração de f que serão exatas para polinômios de grau menor ou igual a $(2n + 1)$. Para estas fórmulas, chamadas fórmulas de Gauss, a determinação dos pontos em que precisamos do valor de $f(x)$ será função do grau do polinômio interpolador e da fórmula específica que vamos considerar.

Usando um polinômio interpolador de grau menor ou igual a n na forma de Lagrange (ver seção V.5), podemos escrever a integral de f no intervalo [a, b] como:

$$\int_a^b f(x)\,dx = \int_a^b \left[\sum_{i=0}^{n} L_i(x)\, f(x_i)\right] dx + \int_a^b \prod_{i=0}^{n} (x - x_i)\,\frac{f^{(n+1)}(\xi)}{(n+1)!}\,dx \qquad (2.1)$$

$\xi \in [a, b]$, onde $L_i(x)$ são os polinômios de Lagrange.

Aproximaremos esta integral por:

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)\,dx = \sum_{i=0}^{n} f(x_i)\, w_i \qquad (2.2)$$

Os coeficientes

$$w_i = \int_a^b L_i(x)\,dx$$

independem da função f e são somente função dos pontos x_i considerados. Estes pontos x_i são determinados de modo a tornar o erro nulo quando integrarmos uma função que é um polinômio de grau menor ou igual a $(2n + 1)$.

Tomemos f como sendo um polinômio de grau $(2n + 1)$. Desta forma a derivada de ordem $(n + 1)$ da função f será um polinômio de grau n. Como vimos na seção IV.5 qualquer polinômio de grau n pode ser escrito como combinação linear dos n primeiros polinômios ortogonais $p_0, p_1, \ldots, p_n$. Assim, de (IV.5.6):

$$f^{(n+1)}(x) = \sum_{j=0}^{n} b_j p_j(x) \qquad (2.3)$$

Da mesma forma:

$$\prod_{i=0}^{n} (x - x_i) = \sum_{i=0}^{n+1} c_i p_i(x) \qquad (2.4)$$

Portanto, de (2.1) e do fato de $\frac{f^{(n+1)}(\xi)}{(n+1)} = f[x_0, x_1, \ldots x_n, x]$ ser um polinômio de grau $\leqslant n$ em x, temos que:

$$\int_a^b \prod_{i=0}^{n} (x - x_i)\,\frac{f^{(n+1)}(\xi)}{(n+1)!}\,dx = \sum_{i=0}^{n+1} \sum_{j=0}^{n} b_j c_i \int_a^b p_i(x)\, p_j(x)\,dx \qquad (2.5)$$

Se os polinômios $p_0, p_1, \ldots, p_{n+1}$ são ortogonais em relação ao produto escalar

$$(f|g) = \int_a^b f(x)\, g(x)\, dx \tag{2.6}$$

a equação (2.5) se reduz a:

$$\int_a^b \prod_{i=0}^{n} (x - x_i) \frac{f^{(n+1)}(\xi)}{(n+1)!}\, dx = \sum_{i=0}^{n} b_i c_i \int_a^b p_i^2(x)\, dx \tag{2.7}$$

Lembrando que desejamos tornar esse erro nulo e que os coeficientes c_i, $i = 0, 1, \ldots, n+1$, em (2.4), são determinados em função dos pontos x_i, $i = 0, 1, \ldots, n$, podemos fazer a escolha destes pontos x_i de tal forma que $c_i = 0$, $i = 0, 1, \ldots, n$. Desta maneira (2.7) se anula. De (2.4):

$$\prod_{i=0}^{n} (x - x_i) = \sum_{i=0}^{n+1} c_i p_i(x) = c_{n+1}\, p_{n+1}(x) \tag{2.8}$$

onde $p_{n+1}(x)$ é o polinômio de grau $(n+1)$ da família de polinômios ortogonais em relação ao produto escalar definido em (2.6).

Vamos tomar

$$c_{n+1} = \frac{1}{\gamma_{n+1}}$$

onde γ_{n+1} é o coeficiente de x^{n+1} em $p_{n+1}(x)$, e os pontos x_i, $i = 0, \ldots, n$, como sendo as raízes do polinômio $p_{n+1}(x)$. O teorema a seguir garante que todos os zeros do polinômio $p_{n+1}(x)$ são reais, distintos e pertencem ao intervalo (a, b).

Teorema 2.1 Se $p_0, p_1, \ldots, p_k, \ldots$ são polinômios ortogonais de grau $0, 1, \ldots, k, \ldots$, respectivamente, em relação ao produto escalar

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

então p_k possui k zeros reais e distintos no intervalo (a, b).

Prova: Como p_0 é uma constante, sem perda de generalidade, podemos tomar $p_0 = 1$.

Sendo $(p_0|p_k) = 0$, $k \geqslant 1$, temos que

$$\int_a^b p_k(x)\, dx = 0$$

ou seja, $p_k(x)$ tem pelo menos um zero de multiplicidade ímpar no intervalo (a, b).

Sejam $x_1, x_2, \ldots, x_m$ os zeros de $p_k(x)$ em (a, b) com multiplicidade ímpar. Vamos mostrar que $m = k$, ou seja, todos os zeros de $p_k(x)$ são distintos e pertencem ao intervalo (a, b).

Tomemos o polinômio de grau $m \leqslant k$:

$$q(x) = (x - x_1)(x - x_2) \ldots (x - x_m)$$

Então $q(x) * p_k(x)$ só tem zeros de multiplicidade par em (a, b).

Como q(x) é um polinômio de grau m podemos escrever:

$$q(x) = \sum_{i=0}^{m} b_i p_i(x)$$

e

$$\int_a^b q(x) * p_k(x)\, dx = \sum_{i=0}^{m} b_i \int_a^b p_i(x)\, p_k(x)\, dx$$

Se $m < k$, essa integral é nula, o que é um absurdo pois $q(x) * p_k(x)$ só tem zeros com multiplicidade par em (a, b).

Portanto, $m = k$. □

Conforme a classe de polinômios ortogonais utilizados, teremos fórmulas de integração de Gauss diferentes. As fórmulas de Gauss-Legendre são obtidas quando tomamos os zeros dos polinômios ortogonais de Legendre. Analogamente podemos obter as fórmulas de Gauss-Tchebyshev [Carnahan], [Hildebrand].

Vamos mostrar as fórmulas de Gauss-Legendre. Como os polinômios de Legendre são ortogonais no intervalo $(-1, 1)$, vamos fazer a mudança de variável

$$z(x) = \frac{2x - (a + b)}{b - a}$$

Assim

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left[\frac{z(b-a)+(b+a)}{2}\right] dz \simeq$$

$$\simeq \sum_{i=0}^{n} w_i f\left[\frac{z_i(b-a)+(b+a)}{2}\right] = \tag{2.9}$$

$$= \sum_{i=0}^{n} w_i F(z_i)$$

onde z_i, $i = 0, \ldots, n$, são os zeros do polinômio de Legendre de grau $(n+1)$ e w_i, $i = 0, \ldots, n$, é dado por:

$$w_i = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(z-z_j)}{(z_i-z_j)}\,dz \tag{2.10}$$

Existem tabelas* dos zeros de alguns polinômios ortogonais, assim como dos correspondentes valores de w_i.

A Tabela 2.1 apresenta esses valores para os polinômios de Legendre. Em [Kopal] encontram-se as tabelas mais completas, isto é, para polinômios de grau maior e com maior precisão.

O erro cometido na integração de f usando as fórmulas de Gauss-Legendre é dado por:

$$E = \int_{-1}^{1} \prod_{i=0}^{n} (z-z_i)\,\frac{F^{(n+1)}(\xi)}{(n+1)!}\,dz =$$

$$= \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3}\,F^{(2n+2)}(\xi) \tag{2.11}$$

$\xi \in [-1, 1]$.

O cálculo desta integral não será incluído neste texto, podendo ser encontrado em [Hildebrand, pp. 314-325].

* Os zeros dos polinômios ortogonais de Laguerre e os de Tchebyshev e os correspondentes w_i podem ser encontrados em [Kopal].

Tabela 2.1

n	zeros de $p_{n+1}(z) = z_i$	coeficientes w_i	i
1	− 0.57735027	1.00000000	0
	+ 0.57735027	1.00000000	1
2	− 0.77459667	0.55555556	0
	0.00000000	0.88888889	1
	+ 0.77459667	0.55555556	2
3	− 0.86113631	0.34785485	0
	− 0.33998104	0.65214515	1
	+ 0.86113631	0.34785485	2
	+ 0.33998104	0.65214515	3
4	− 0.90617985	0.23692689	0
	− 0.53846931	0.47862867	1
	0.00000000	0.56888889	2
	+ 0.90617985	0.23692689	3
	+ 0.53846931	0.47862867	4

Exemplo 2.1 Vamos integrar $f(x) = x^4 + 1$ no intervalo $(-1, +1)$ usando as fórmulas de Gauss-Legendre, para $n = 2$.

$$I = \int_{-1}^{1} (x^4+1)\,dx = \sum_{i=0}^{2} w_i f(x_i)$$

Da Tabela 2.1, para $n = 2$, temos:

$x_0 = -0.77459667$ $\quad w_0 = 0.55555556$
$x_1 = \ \ 0.00000000$ $\quad w_1 = 0.88888889$
$x_2 = +0.77459667$ $\quad w_2 = 0.55555556$

$$I = 0.55555556 * [(-0.77459667)^4 + 1] +$$
$$+ 0.88888889 + [(0.00000000)^4 + 1] +$$
$$+ 0.55555556 * [(0.77459667)^4 + 1] =$$
$$= 2.4$$

Vamos comparar este resultado com o valor obtido para a integral de f, usando a fórmula de Simpson:

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)]$$

com $h = 1, x_0 = -1, x_1 = 0$ e $x_2 = 1$.

$$\int_{-1}^{1} f(x)\,dx = \frac{1}{3}[2 + 4 + 2] = 2.6666667$$

Podemos verificar que a fórmula de Gauss-Legendre forneceu o valor exato para a integral, enquanto a de Simpson não.

A integração usando a fórmula de Gauss-Legendre é inconveniente se os cálculos forem feitos manualmente ou mesmo com o uso de calculadoras, uma vez que os valores dos coeficientes w_i e abscissas x_i envolvidos são representados por números de muitos algarismos significativos. Entretanto, isto se torna irrelevante se os cálculos forem feitos por um computador.

3 – EXERCÍCIOS

1. Aplicando a regra de Simpson, calcular a área entre a curva que passa pelos pontos abaixo, o eixo x e as retas $x = 2$ e $x = 18$.

x	2	4	6	8	10	12	14	16	18
y	0.5	0.9	1.1	1.3	1.7	2.1	1.5	1.1	0.6

2. Calcular pela fórmula dos trapézios e pela fórmula de Simpson

$$\int_0^{0.8} \frac{dx}{x^2 - 1}$$

com passo $h = 0.2$.

3. Delimite o erro que se comete ao calcular

$$\int_0^{0.4} e^x dx$$

pela fórmula de Simpson com $h = 0.1$.

4. Deseja-se calcular

$$\log 2 = \int_1^2 \frac{dx}{x}$$

com erro inferior a $\frac{1}{2400}$ usando a fórmula dos trapézios

$$\int_{x_0}^{x_n} f(x)\,dx = h\left[\frac{f(x_0)}{2} + f(x_1) + \ldots + f(x_{n-1}) + \frac{f(x_n)}{2}\right] - \frac{nh^3}{12} f''(\varepsilon)$$

$x_0 \leqslant \varepsilon \leqslant x_n$.

Qual deve ser o passo escolhido? Repita o exercício usando a fórmula de Simpson.

5. a) Deduzir a fórmula de integração numérica que se obtém utilizando o polinômio interpolador de grau zero relativamente ao primeiro ponto de cada subintervalo e considerando o intervalo de integração [a, b] subdividido em n intervalos iguais de comprimento $h = (b - a)/n$.

   b) Qual é o erro máximo cometido em cada subintervalo?

6. Determine a fórmula de integração que se obtém ao fazer a aproximação

$$\int_{x_0}^{x_1} f(x)\,dx \simeq \int_{x_0}^{x_1} L_0(x) f(x_0)\,dx + \int_{x_0}^{x_1} L_1(x) f(x_1)\,dx$$

onde $L_0(x)$ e $L_1(x)$ são os polinômios de Lagrange de grau 1, em relação aos pontos x_0 e x_1. Delimite o erro cometido ao fazer esta aproximação.

7. Calcule

$$\int_0^1 e^x dx$$

utilizando as fórmulas de Gauss-Legendre para $n = 3$.

Capítulo 7

Introdução à Solução Numérica de Equações Diferenciais

Problemas envolvendo a taxa de variação de uma variável em relação a outra são modelados através de uma equação diferencial ou de uma equação de diferenças. Existe um número muito restrito de equações diferenciais cuja solução pode ser expressa sob uma forma analítica simples. Neste capítulo veremos alguns métodos numéricos elementares para determinar uma aproximação da solução de uma equação diferencial ordinária de primeira ordem, ou seja, uma solução da equação

$$\frac{dy}{dx} = f(x, y)$$

1 – INTRODUÇÃO

Dada uma equação diferencial de primeira ordem

$$\frac{dy}{dx} = f(x, y) \tag{1.1}$$

sujeita à condição inicial

$$y(x_0) = y_0 \tag{1.2}$$

cuja solução é dada pela função y (x), definida e contínua no intervalo $[x_0, x_F]$, gostaríamos de determinar aproximações y_ℓ, $\ell = 0, 1, \ldots,$ n dos valores da função y (x) para os pontos x_ℓ, $\ell = 0, 1, \ldots, n$.

Os pontos x_ℓ são igualmente espaçados no intervalo $[x_0, x_F]$, ou seja,

$$x_\ell = x_0 + h\ell\ , \qquad \ell = 0, 1, \ldots, n$$

onde

$$h = \frac{x_F - x_0}{n}$$

Note que não obteremos a forma analítica da função aproximadora.

Os dois métodos que estudaremos baseiam-se na expansão em série de Taylor da função y em torno do ponto x, ou seja,

$$y(x+h) = y(x) + hf(x, y(x)) + \frac{h^2}{2!} f'(x, y(x)) + \frac{h^3}{3!} f''(x, y(x)) + \ldots \tag{1.3}$$

Para estes métodos o cálculo da aproximação $y_{\ell+1}$ depende somente de y_ℓ, x_ℓ e h, e por isto são normalmente chamados de métodos de passo simples.

2 – MÉTODO DE EULER

O método de Euler utiliza os dois primeiros termos em (1.3), ou seja, a aproximação linear da função y. Assim temos

$$y_{\ell+1} = y_\ell + hf(x_\ell, y_\ell) \tag{2.1}$$

onde

$$y_0 = y(x_0),\ h = \frac{x_F - x_0}{n}\ , \qquad \ell = 0, 1, \ldots, n-1$$

Exemplo 2.1 Vamos calcular a solução aproximada de

$$\frac{dy}{dx} = f(x, y) = -xy,\ y(0) = 1$$

tomando h = 0.1, para o intervalo [0, 1]. Utilizando a fórmula (2.1) temos:

$$y_\ell = y_{\ell-1} + h(-x_{\ell-1})(y_{\ell-1})$$

e

$$x_{\ell-1} = x_0 + (\ell - 1)\,h \qquad \ell = 1, 2, \ldots, 10$$

Como $x_0 = 0$ e $y_0 = 1$, obtemos a seguinte tabela:

ℓ	x_ℓ	y_ℓ	$-x_\ell * y_\ell$	solução exata $y = e^{-x^2/2}$
0	0.	1.	0.	1.0
1	0.1	1.	−0.1	0.995012479
2	0.2	0.99	−0.198	0.980198673
3	0.3	0.9702	−0.29106	0.955997482
4	0.4	0.941094	−0.3764376	0.923116346
5	0.5	0.90345024	−0.45172512	0.882496903
6	0.6	0.858277728	−0.514966637	0.835270211
7	0.7	0.806781064	−0.564746745	0.782704538
8	0.8	0.75030639	−0.600245112	0.726149037
9	0.9	0.690281879	−0.621253691	0.666976811
10	1.0	0.62815651	−0.62815651	0.606530660

Note que o erro cometido no primeiro intervalo é carregado para o segundo intervalo. No terceiro intervalo carregamos o erro do primeiro e do segundo intervalo e assim por diante.

Se tomássemos mais termos da expansão em série de Taylor (1.3), teríamos uma melhor aproximação de $y(x)$. Entretanto seria necessário calcular derivadas da função $f(x, y(x))$ (em relação a x), o que poderia tornar muito complicado o problema.

3 – MÉTODOS DE RUNGE-KUTTA

Os métodos de Runge-Kutta de ordem m fornecem valores aproximados da solução da equação diferencial (1.1) que coincidem com os valores obtidos através da expansão em série de Taylor de y, em torno do ponto x, até o termo que inclui h^m, ou seja,

$$y(x+h) \simeq y(x) + f(x, y(x))\,h + f'(x, y(x))\,\frac{h^2}{2!} + \\ + \ldots + f^{(m-1)}(x, y(x))\,\frac{h^m}{m!} \tag{3.1}$$

Nestes métodos o cálculo do valor das derivadas da função $f(x, y(x))$ nos pontos (x_ℓ, y_ℓ) será substituído pelo cálculo da função $f(x, y)$ em pontos convenientes, produzindo resultados equivalentes.

Vamos agora desenvolver as fórmulas para um método de Runge-Kutta de segunda ordem. Assim, em vez de calcularmos $y_{\ell+1}$ utilizando (3.1) com m = 2, ou seja, fazendo

$$y_{\ell+1} = y_\ell + hf(x_\ell, y(x_\ell)) + \frac{h^2}{2!} f'(x_\ell, y(x_\ell)) \tag{3.2}$$

calcularemos $y_{\ell+1}$ através da fórmula

$$y_{\ell+1} = y_\ell + h(aK_1 + bK_2) \tag{3.3}$$

onde a e b são constantes e K_1 e K_2 são valores da função $f(x, y(x))$ calculados em pontos convenientes. Temos que determinar, então, os valores de a, b, K_1 e K_2.

De (3.2) e (3.3) temos:

$$aK_1 + bK_2 = f(x_\ell, y(x_\ell)) + \frac{h}{2} f'(x_\ell, y(x_\ell)) \tag{3.4}$$

Se tomarmos

$$K_1 = f(x_\ell, y_\ell) \quad \text{e} \\ K_2 = f(x_\ell + ph, y_\ell + qhf(x_\ell, y_\ell)) \tag{3.5}$$

o nosso problema passa a ser determinar a, b, p e q.

Usando linearização da função $f(x, y(x))$ em torno do ponto (x_ℓ, y_ℓ) através da série de Taylor*, temos:

$$K_2 = f(x_\ell + ph, y_\ell + qhf(x_\ell, y_\ell)) \simeq \\ \simeq f(x_\ell, y_\ell) + phf_x(x_\ell, y_\ell) + qhf(x_\ell, y_\ell) f_y(x_\ell, y_\ell) \tag{3.6}$$

onde

$$f_x(x_\ell, y_\ell) = \frac{\partial}{\partial x} f(x, y)\Big|_{(x_\ell, y_\ell)} \quad \text{e} \quad f_y(x_\ell, y_\ell) = \frac{\partial}{\partial y} f(x, y)\Big|_{(x_\ell, y_\ell)}$$

* Fórmula de Taylor para uma função de duas variáveis:

$$f(x, y) = f(a + h_1, b + h_2) = f(a, b) + h_1 \frac{\partial f(x, y)}{\partial x}\Big|_{(a, b)} + \\ + h_2 \frac{\partial f(x, y)}{\partial y}\Big|_{(a, b)} + E(x, y).$$

Substituindo (3.5) e (3.6) em (3.3) obtemos:

$$y_{\ell+1} \simeq y_\ell + h\{af(x_\ell, y_\ell) + b[f(x_\ell, y_\ell) + phf_x(x_\ell, y_\ell) + qhf(x_\ell, y_\ell) f_y(x_\ell, y_\ell)]\} = y_\ell + ahf(x_\ell, y_\ell) + hbf(x_\ell, y_\ell) + bh^2 pf_x(x_\ell, y_\ell) + bqh^2 f(x_\ell, y_\ell) f_y(x_\ell, y_\ell) = y_\ell + h(a+b) f(x_\ell, y_\ell) + h^2 b[pf_x(x_\ell, y_\ell) + qf(x_\ell, y_\ell) f_y(x_\ell, y_\ell)] \quad (3.7)$$

Como

$$f'(x, y(x)) = \frac{d}{dx} f(x, y(x)) = \frac{dx}{dx} f_x(x, y) + \frac{dy}{dx} f_y(x, y) = f_x(x, y) + f(x, y) f_y(x, y)$$

podemos escrever (3.2) da seguinte maneira:

$$y_{\ell+1} = y_\ell + hf(x_\ell, y_\ell) + \frac{h^2}{2!} f_x(x_\ell, y_\ell) + \frac{h^2}{2!} f(x_\ell, y_\ell) f_y(x_\ell, y_\ell) \quad (3.8)$$

Igualando (3.8) e (3.7) vem:

$$h(a+b) f(x_\ell, y_\ell) + h^2[bpf_x(x_\ell, y_\ell) + qbf(x_\ell, y_\ell) f_y(x_\ell, y_\ell)] = hf(x_\ell, y_\ell) + h^2\left[\frac{1}{2!} f_x(x_\ell, y_\ell) + \frac{1}{2!} f(x_\ell, y_\ell) f_y(x_\ell, y_\ell)\right]$$

temos então:

$$a + b = 1 \qquad bp = \frac{1}{2} \qquad bq = \frac{1}{2}$$

portanto,

$$a = 1 - b \qquad p = q = \frac{1}{2b}$$

Tomando $b = \frac{1}{2}$ obtemos

$$a = b = \frac{1}{2} \quad e \quad p = q = 1 \quad (3.9)$$

Assim, este processo de Runge-Kutta de segunda ordem fica definido pelas equações:

$$y_{\ell+1} = y_\ell + \frac{h}{2}(K_1 + K_2), \qquad \ell = 0, 1, \ldots, n-1$$
$$K_1 = f(x_\ell, y_\ell) \text{ e } K_2 = f(x_\ell + h, y_\ell + hK_1) \quad (3.10)$$

Exemplo 3.1 Vamos agora calcular a solução aproximada da equação diferencial

$$\frac{dy}{dx} = f(x, y) = -xy, \qquad y(0) = 1$$

no intervalo [0, 1], utilizando o método de Runge-Kutta de segunda ordem com $h = 0.1$.

De (3.10) temos:

$$y_{\ell+1} = y_\ell + \frac{0.1}{2}(K_1 + K_2), \qquad \ell = 0, 1, \ldots, 9$$
$$K_1 = -x_\ell y_\ell$$
$$K_2 = -(x_\ell + 0.1)(y_\ell + 0.1\,K_1)$$

Como $x_0 = 0$ e $y_0 = 1$, obtemos os seguintes resultados:

ℓ	x_ℓ	y_ℓ	K_1	K_2	$y_{\ell+1}$
0	0.	1.0	0.0	- 0.1	0.995
1	0.1	0.995	- 0.0995	- 0.19701	0.9801745
2	0.2	0.9801745	- 0.1960349	- 0.288171303	0.95596419
3	0.3	0.95596419	- 0.286789257	- 0.370914106	0.923079022
4	0.4	0.923079022	- 0.369231609	- 0.443077930	0.882463545
5	0.5	0.882463545	- 0.441231772	- 0.503004221	0.835251745
6	0.6	0.835251745	- 0.501151047	- 0.549595648	0.78271441
7	0.7	0.78271441	- 0.547900087	- 0.582339521	0.72620243
8	0.8	0.72620243	- 0.580961944	- 0.601295612	0.667089552
9	0.9	0.667089552	- 0.600380597	- 0.607051492	0.606717947

As fórmulas de Runge-Kutta de outra ordem qualquer podem ser deduzidas de forma análoga ao que foi feito aqui para segunda ordem. Vamos apresentar as fórmulas de quarta ordem, normalmente as mais utilizadas, sem deduzi-las, pois a dificuldade em obtê-las é puramente algébrica.

$$y_{\ell+1} = y_\ell + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4), \qquad \ell = 1, \ldots, n-1$$

onde

$$K_1 = f(x_\ell, y_\ell) \qquad K_2 = f\left(x_\ell + \frac{1}{2}h,\ y_\ell + \frac{1}{2}hK_1\right)$$
$$K_3 = f\left(x_\ell + \frac{1}{2}h,\ y_\ell + \frac{1}{2}hK_2\right) \qquad K_4 = f(x_\ell + h,\ y_\ell + hK_3) \tag{3.11}$$

Exemplo 3.2 Vamos recalcular a solução da equação

$$\frac{dy}{dx} = -xy\,, \qquad y(0) = 1$$

com h = 0.1, no intervalo [0, 1], pelo método de Runge-Kutta de quarta ordem.

Utilizando as fórmulas (3.11) para h = 0.1, $x_0 = 0$ e $y_0 = 1$ vem:

$$y_{\ell+1} = y_\ell + \frac{0.1}{6}(K_1 + 2K_2 + 2K_3 + K_4)\,, \qquad \ell = 0, 1, \ldots, 9$$

$$K_1 = -x_\ell y_\ell$$
$$K_2 = -(x_\ell + 0.05)(y_\ell + 0.05\,K_1)$$
$$K_3 = -(x_\ell + 0.05)(y_\ell + 0.05\,K_2)$$
$$K_4 = -(x_\ell + 0.1)(y_\ell + 0.1\,K_3)$$

Como $x_0 = 0$ e $y_0 = 1$, obtemos os seguintes resultados:

ℓ	x_ℓ	y_ℓ	K_1	K_2	K_3	K_4
0	0.	1.	0.0	−0.05	−0.049875	−0.09950125
1	0.1	0.995012479	−0.099501248	−0.14850513	−0.14813808	−0.196039734
2	0.2	0.980198673	−0.196039735	−0.242599172	−0.242017199	−0.286799087
3	0.3	0.955997481	−0.286799244	−0.329580132	−0.328831466	−0.369245734
4	0.4	0.923116345	−0.369246538	−0.407094308	−0.406242733	−0.441246036
5	0.5	0.882496901	−0.44124845	−0.473238963	−0.472359224	−0.501156587
6	0.6	0.83527021	−0.501162126	−0.526637868	−0.525809906	−0.547882454
7	0.7	0.782704542	−0.547893179	−0.566482412	−0.565785316	−0.580900808
8	0.8	0.726149051	−0.580919241	−0.592537626	−0.592043844	−0.6002502
9	0.9	0.666976845	−0.600279160	−0.605114742	−0.604885052	−0.606488339

Observação: Os resultados dos Exemplos 2.1, 3.1 e 3.2 foram obtidos com o auxílio da calculadora HP 9810 A.

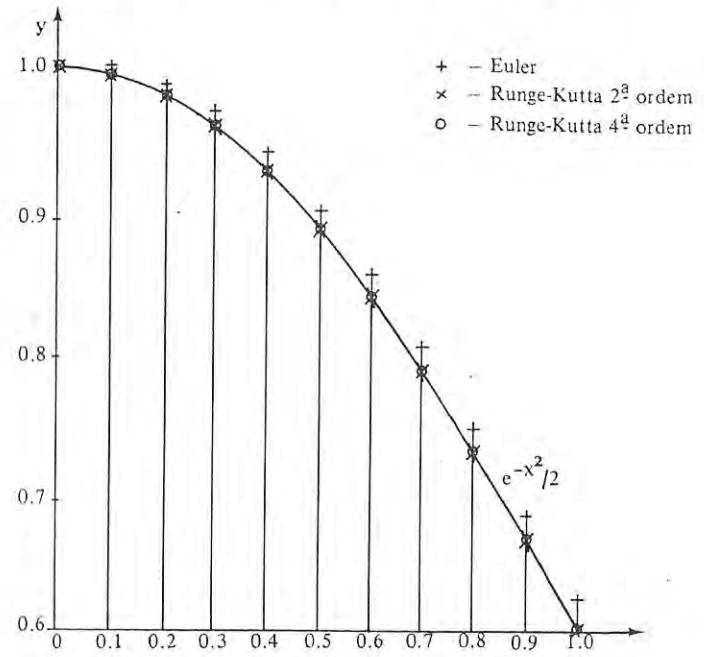


Fig. 3.1

Na Fig. 3.1 ilustramos os resultados obtidos nos Exemplos 2.1, 3.1 e 3.2, assim como a solução exata

$$y = e^{-x^2/2}$$

da equação diferencial

$$\frac{dy}{dx} = -xy$$

no intervalo [0, 1], com y (0) = 1.

4 – EXERCÍCIOS

1. Dada a equação

$$\frac{dy}{dx} = 100y - 101\,e^{-x} - 100\,, \qquad y(0) = 2$$

ache sua solução aproximada no intervalo [0, 1] utilizando o o método de Euler, com h = 0.1.

2. Resolver $\frac{dy}{dx} = y$ com condição inicial $y(0) = 1$, pelo método de Runge-Kutta de 2ª ordem, com passo h = 0.2, no intervalo [0, 1].

3. Equações diferenciais ordinárias de 2ª ordem, do tipo $\frac{d^2y}{dx^2} = y''(x) = g(x, y)$, com $y(x_0)$ e $y'(x_0)$ conhecidos, nas quais a 1ª derivada não aparece explicitamente, podem ser resolvidas segundo o seguinte algoritmo derivado do método Runge-Kutta:

$$y_{\ell+1} = y_\ell + \frac{h}{6}(K_1 + 4K_2 + K_3)$$

$$y'_{\ell+1} = h\left(y_\ell + \frac{h}{6}(K_1 + 2K_2)\right)$$

$$K_1 = g(x_\ell, y_\ell)$$

$$K_2 = g\left(x_\ell + \frac{h}{2}, y_\ell + \frac{h}{2}y'_\ell + \frac{h^2}{8}K_1\right)$$

$$K_3 = g\left(x_\ell + h, y_\ell + hy'_\ell + \frac{h^2}{2}K_2\right)$$

Utilize este algoritmo para resolver a seguinte equação:

$$\frac{d^2y}{dx^2} = -a^2y, \qquad \begin{matrix} y(0) = 0.1 \\ y'(0) = 0 \end{matrix}$$

no intervalo [0, 2] com h = 0.2.

4. Resolva a equação

$$\frac{dy}{dx} = \frac{\cos x}{(1 + 4\,\mathrm{sen}^2 x)^{1/2}}, \qquad y(0) = 2$$

no intervalo [1, 2] com h = 0.5 utilizando:

a) método de Euler;
b) método de Runge-Kutta de 2ª ordem.

5. Considere um corpo de massa m = 1 kg, sob ação de uma força que depende da velocidade v da seguinte forma:

$$F(v) = 3v^2 + v.$$

Sabendo-se que no instante t = 0 a velocidade do corpo era de 4.5 m/s, determine a velocidade do corpo no instante t = 6 s, a partir da solução da equação diferencial

$$m\frac{dv}{dt} = F(v), \qquad v(0) = 4.5 \text{ m/s}$$

utilizando o método de Runge-Kutta de 4ª ordem e h = 1 s.

6. Resolva o sistema de equações diferenciais abaixo no intervalo [0, 1] com h = 0.5 utilizando:

a) Runge-Kutta de 2ª ordem;
b) Runge-Kutta de 4ª ordem.

$$\begin{cases} \dfrac{dx}{dt} = -y \\ \dfrac{dy}{dt} = x \end{cases}, \qquad x(0) = 0,\ y(0) = 1$$

Apêndice A

Matriz Elementar Coluna e de Permutação

Definição: Uma matriz quadrada E_i de ordem n é dita *Matriz Elementar Coluna* se

$$E_i = I + ve_i^T$$

onde e_i^T é a i-ésima linha da matriz identidade de ordem n e $v^T = (v_1, v_2, \ldots, v_i, \ldots, v_n)$.

Exemplo:

$$E_3 = \begin{bmatrix} 1 & 0 & v_1 & 0 \\ 0 & 1 & v_2 & 0 \\ 0 & 0 & v_3 & 0 \\ 0 & 0 & v_4 & 1 \end{bmatrix}$$

- A inversa de uma matriz elementar coluna E_i é dada por

$$E_i^{-1} = \begin{bmatrix} 1 & 0 & \ldots & 0 & -v_1/v_i & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & -v_2/v_i & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 & -v_{i-1}/v_i & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1/v_i & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & -v_{i+1}/v_i & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & -v_n/v_i & 0 & \ldots & 1 \end{bmatrix}$$

- Uma matriz triangular inferior L de ordem n pode ser escrita como produto de n matrizes elementares coluna, também triangulares inferior, isto é,

$$L = L_1 L_2 L_3 \ldots L_n$$

onde L_k, $k = 1, 2, \ldots, n$, é uma matriz elementar coluna cuja k-ésima coluna é a k-ésima coluna da matriz L.

- Uma matriz triangular superior U de ordem n pode ser escrita como produto de n matrizes elementares coluna, isto é,

$$U = U_n U_{n-1} U_{n-2} \ldots U_2 U_1$$

onde U_k, $k = 1, 2, \ldots, n$, é a matriz elementar coluna cuja k-ésima coluna é a k-ésima coluna da matriz U.

- Dada uma matriz elementar coluna de ordem n, E_i e um vetor $z = (z_1, z_2, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n)^T$, o produto $E_i z$ é dado por

$$E_i z = \begin{bmatrix} z_1 \\ \vdots \\ z_{i-1} \\ 0 \\ z_{i+1} \\ \vdots \\ z_n \end{bmatrix} + z_i \begin{bmatrix} v_1 \\ \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \\ v_n \end{bmatrix}$$

- *Definição:* Uma matriz P quadrada de ordem n é dita de permutação se ela pode ser obtida da matriz identidade I por permutação de suas linhas.

- Dada uma matriz A qualquer, pré-multiplicá-la por uma matriz de permutação P (ou seja, efetuar o produto PA) corresponde a permutar as linhas de A. A permutação realizada é a mesma que produziu P a partir de I; pós-multiplicar uma matriz A qualquer por uma matriz de

permutação P (ou seja, efetuar o produto AP) corresponde a permutar as colunas de A.

- *Exemplo:*

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \qquad P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$PA = \begin{bmatrix} d & e & f \\ a & b & c \\ g & h & i \end{bmatrix} \qquad AP = \begin{bmatrix} b & a & c \\ e & d & f \\ h & g & i \end{bmatrix}$$

- O determinante de uma matriz de permutação P é dado por

$$\det P = (-1)^q$$

onde q é o número de permutações de linhas realizadas na matriz identidade para obter P.

- Se uma matriz de permutação P é obtida pela permutação de duas linhas quaisquer da matriz identidade, então $PP = I$.

- A inversa de uma matriz de permutação P é também uma matriz de permutação.

Apêndice B
Sistema Normal

Faremos, aqui, algumas considerações sobre o sistema normal:

$$\begin{bmatrix} (g_0|g_0) & (g_0|g_1) & \dots & (g_0|g_m) \\ (g_1|g_0) & (g_1|g_1) & \dots & (g_1|g_m) \\ \vdots & \vdots & & \vdots \\ (g_m|g_0) & (g_m|g_1) & \dots & (g_m|g_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} (g_0|f) \\ (g_1|f) \\ \vdots \\ (g_m|f) \end{bmatrix} \tag{B.1}$$

Vamos mostrar que, para o domínio discreto, se o número de pontos N em que a função está tabelada é menor do que o número de coeficientes $(m + 1)$ a serem determinados, o sistema normal é indeterminado.

Se o determinante da matriz dos coeficientes do sistema em (B.1) for nulo, existe $(a_0, a_1, \dots, a_m)$ solução não trivial do sistema homogêneo associado a (B.1). Mas,

$$(a_0, a_1, \dots, a_m) \begin{bmatrix} (g_0|g_0) & \dots & (g_0|g_m) \\ (g_1|g_0) & \dots & (g_1|g_m) \\ \vdots & & \vdots \\ (g_m|g_0) & \dots & (g_m|g_m) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = 0 \tag{B.2}$$

ou, efetuando o produto:

$$\left(\sum_{k=0}^{m} a_k g_k \,\middle|\, \sum_{k=0}^{m} a_k g_k \right) = 0 \tag{B.3}$$

Para que (B.3) seja verdadeiro é necessário que

$$\sum_{k=0}^{m} a_k g_k(x_i) = 0\,, \qquad i = 1, 2, \ldots, N \tag{B.4}$$

As equações em (B.4) fornecem um sistema linear de equações, N por (m + 1), que só admite solução não trivial se o posto da matriz G,

$$G = \begin{bmatrix} g_0(x_1) & g_1(x_1) & \cdots & g_m(x_1) \\ g_0(x_2) & g_1(x_2) & \cdots & g_m(x_2) \\ \vdots & \vdots & & \vdots \\ g_0(x_N) & g_1(x_N) & \cdots & g_m(x_N) \end{bmatrix} \tag{B.5}$$

for menor que (m + 1). Esta solução não trivial também é solução do sistema homogêneo associado a (B.1). Se $N \leqslant m$ o posto de G é logicamente menor que (m + 1), implicando a existência da solução não trivial do sistema normal e, conseqüentemente, o sistema normal é indeterminado.

Teorema B.1 O sistema normal (B.1) tem solução única se e somente se o posto da matriz G for máximo.

Prova: O posto da matriz G é máximo, portanto, (m + 1), se e somente se o posto da matriz $G^T * G$ for também (m + 1). Mas,

$$G^T * G = \begin{bmatrix} (g_0|g_0) & (g_0|g_1) & \cdots & (g_0|g_m) \\ (g_1|g_0) & (g_1|g_1) & \cdots & (g_1|g_m) \\ \vdots & \vdots & & \vdots \\ (g_m|g_0) & (g_m|g_1) & \cdots & (g_m|g_m) \end{bmatrix}$$

Portanto, o sistema (B.1) tem uma única solução se e somente se o posto da matriz G for máximo. □

É importante notar que o fato de ter $N \geqslant (m + 1)$ não implica a existência de uma única solução para o sistema normal, portanto para o problema de aproximação de uma função por uma outra, pelo

método dos mínimos quadrados. O exemplo seguinte ilustra tal fato: aproximar uma função f tabelada nos pontos 0 e $\pi/2$ pela função $g(x) = a_0 x + a_1 \operatorname{sen} x$.

No caso particular de ajuste de uma função tabelada f por uma função

$$g(x) = \sum_{k=0}^{m} a_k x^k$$

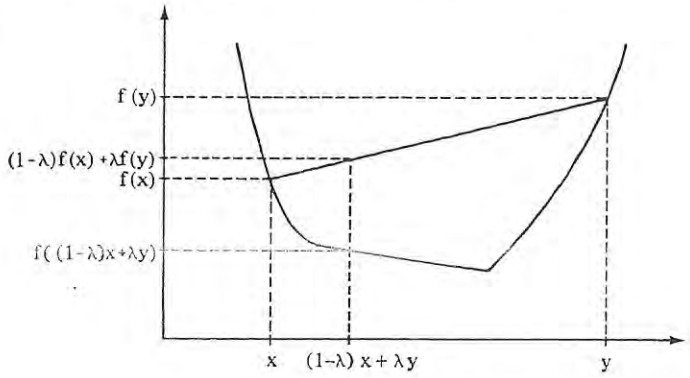
pode-se provar que se $m < N$ o sistema normal sempre tem uma única solução [Hamming].

Apêndice C
Convexidade

Definição: Diz-se que uma função $f:\mathbb{R}^n \to \mathbb{R}$ é convexa em x se para todo $y \in \mathbb{R}^n$ e $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) \leqslant (1-\lambda) f(x) + \lambda f(y)$$

Se f for convexa para todo $x \in \mathbb{R}^n$, então f é dita convexa.



Teorema C.1 (Mangasarian) Seja f uma função real duplamente diferenciável, definida num intervalo aberto $\Gamma \subset \mathbb{R}^n$. Então f é convexa se e somente se $\nabla^2 f(x)$ for semidefinida positiva em Γ, ou seja, para todo $x \in \Gamma$

$$y^T \nabla^2 f(x)\, y \geqslant 0\,, \qquad \forall\, y \in \mathbb{R}^n$$

Utilizando este resultado, vamos mostrar que

$$(r|r) = (f - g|f - g)$$

é convexa.

Tomando-se

$$g(x) = \sum_{k=0}^{m} a_k g_k(x)$$

como

$$(r|r) = (f|f) - 2(f|g) + (g|g)$$

temos

$$(r|r) = M(a_0, a_1, \ldots, a_m) = (f|f) - 2\sum_{k=0}^{m} (f|g_k)\, a_k + \sum_{k=0}^{m}\sum_{\varrho=0}^{m} (g_k|g_\varrho)\, a_k a_\varrho$$

Então:

$$\nabla (r|r) = \left[\frac{\partial (r|r)}{\partial a_k}\right]_{k=0,\ldots,m} = \left[2\sum_{\varrho=0}^{m} (g_0|g_\varrho)\, a_\varrho - 2(g_0|f),\right.$$

$$\left. 2\sum_{\varrho=0}^{m} (g_1|g_\varrho)\, a_\varrho - 2(g_1|f),\ \ldots,\ 2\sum_{\varrho=0}^{m} (g_m|g_\varrho)\, a_\varrho - 2(g_m|f)\right]$$

e o Hessiano:

$$\nabla^2 (r|r) = \begin{bmatrix} 2(g_0|g_0) & 2(g_0|g_1) & \ldots & 2(g_0|g_m) \\ 2(g_1|g_0) & 2(g_1|g_1) & \ldots & 2(g_1|g_m) \\ \vdots & \vdots & & \vdots \\ 2(g_m|g_0) & 2(g_m|g_1) & \ldots & 2(g_m|g_m) \end{bmatrix}.$$

Portanto:

$$y^T \nabla^2 (r|r)\, y = \sum_{k=0}^{m}\sum_{\varrho=0}^{m} (g_k|g_\varrho)\, y_k y_\varrho = \left[\sum_{k=0}^{m} g_k y_k \,\middle|\, \sum_{\varrho=0}^{m} g_\varrho y_\varrho\right] \geqslant 0,$$

o que mostra que $(r|r)$ é convexa.

Referências Bibliográficas

ALBRECHT, Peter. *Análise numérica: um curso moderno*. Rio de Janeiro, Livros Técnicos e Científicos, 1973, 240p.

APOSTOL, Tom M. *Mathematical analysis: a modern approach to advanced calculus*. Reading, Addison-Wesley, c1957, 559p.

BARROS, Ivan de Queiroz. *Introdução ao cálculo numérico*. São Paulo, Edgard Blücher -- EDUSP, c1972, 114p.

BARROS, Ivan de Queiroz. *Métodos numéricos*. Campinas, IMECC, 1970, v. 1.

BARROS, Ivan de Queiroz. *Notas de análise numérica*. São Paulo, s.c.p.1966, 1v (várias paginações).

CARNAHAN, Brice; LUTHER, H.A.; WILKES, James O. *Applied numerical methods*. New York, John Wiley & Sons, 1969, 604p.

DAHLQUIST, Germund & BJÖRCK, Ake. *Numerical methods*. Englewood Cliffs, Prentice-Hall, c1974, 573p (Prentice-Hall Series in Automatic Computation).

DORN, W.S. & McCRACKEN, Daniel D. *Cálculo numérico com estudos de casos em FORTRAN IV*. Rio de Janeiro, Ed. Campus, 1978, 568p.

FORSYTHE, George E. & MOLER, Cleve B. *Computer solution of linear algebraic systems*. Englewood Cliffs, Prentice-Hall, 1967, 148p.

GONÇALVES, Adilson. *Introdução à álgebra*. 11º Colóquio Brasileiro de Matemática. Rio de Janeiro, IMPA, 1977, 332p.

HAMMING, Richard W. *Numerical methods for scientists and engineers*. New York, McGraw-Hill, 1962, 411p (International Series in Pure and Applied Mathematics).

HILDEBRAND, Francis B. *Introduction to numerical analysis*. New York, McGraw-Hill, 1956 (International Series in Pure and Applied Mathematics).

KOPAL, Zdeněk. *Numerical analysis*. London, Chapman & Hall Ltd., 1961, 594p.

MANGASARIAN, Olvi L. *Nonlinear programming*. New York, McGraw-Hill, c1969, 220p.

MILNE, William E. *Cálculo numérico*. São Paulo, Editora Polígono, 1968, 346p.

RALSTON, Anthony. *A first course in numerical analysis*. New York, McGraw Hill, c1965, 578p (International Series in Pure and Applied Mathematics).

SANTOS, Vitoriano R. de B. *Curso de cálculo numérico*. Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1974, 257p (Série Ciência de Computação).

SPIEGEL, M.R. *Mathematical handbook of formulas and tables*. New York, McGraw-Hill, c1968, 271p. (Schaum's Outline Series).

VAN der Waerden, B. L. *Modern Algebra*. New York, Frederick Ungar Publishing Co., c1949, 264p, v. 1.

WILKINSON, J.H. *Rounding errors in algebraic processes*. Englewood Cliffs, Prentice-Hall, 1963, 161p.