Cálculo Numérico A

2010

Conteúdo

1. **Nome da Disciplina** Cálculo Numérico A - MAT01032

2. **Súmula** Súmula Erros; ajustamento de equações; interpolação, derivação e integração; solução de equações lineares e não lineares; solução de sistemas de equações lineares e não lineares; noções de otimização; solução de equações diferenciais e equações diferenciais parciais; noções do método Monte Carlo em suas diferentes aplicações.

3. Regras

   (a) Alunos que não tem setenta e cinco por cento de freqüência (75%) serão reprovados com conceito FF, seguindo o que é determinado no Artigo 134 do Regimento Geral da UFRGS.

   (b) O programa da disciplina será dividido em duas áreas para fins de avaliação, as quais corresponderão duas provas escritas, cada uma com pontuação entre 0 e 10.

      i. O aluno será considerado aprovado se a média aritmética das notas das 2 provas escritas for igual ou superior a 6.0 pontos e, em cada uma delas, a nota for igual ou superior a 4.0 pontos.

      ii. Para serem aprovados, alunos que não se enquadram no critério acima, mas que têm nota iguais ou superior a 5.0 em uma das duas provas, poderão fazer uma única prova de recuperação da prova de menor nota, ao final do semestre, e serem aprovados se passarem a enquadrar-se no critério (3(b)i).

      iii. Para serem aprovados, alunos que não se enquadram nos critérios (3(b)i) e (3(b)ii), mas que possuem média aritmética das provas igual ou superior a 3.0, deverão fazer exame geral dos conteúdos da disciplina ao final do semestre.

      iv. Alunos que não se enquadrem nos critérios (3(b)i), (3(b)ii) ou (3(b)iii) estarão reprovados na disciplina.

   (c) Sendo S a média aritmética das notas das 2 provas escritas, a atribuição do conceito que será enviado ao DECORDI seguirá a regra:

      - se $6.0 \leq S < 7.5$, o conceito é C;
      - se $7.5 \leq S < 9.0$, o conceito é B;
      - se $9.0 \leq S \leq 10.0$, o conceito é A.

   (d) Alunos aprovados no exame final cumulativo receberão conceito C se obtiveram escore entre 6.0 e 8.5 (incluindo 6.0 e excluindo 8.5) e conceito B se o escore for igual ou superior a 8.5. O exame final e as provas de recuperação de área serão oferecidos em um mesmo dia e hora, para cada turma, e terão a mesma duração.

Bibliografia

1. Richard L. Burden; J. Douglas Faires - Análise Numérica - Editora CENGAGE (ISBN: 85-221-0601-0)

2. A.L. Bortoli; C. Cardoso; M.P.G. Fachin; R.D. Cunha - Introdução ao cálculo Numérico (Cadernos de Matemática e Estatística) - Editora UFRGS

3. Claudio, Dalcidio Moraes; Marins, Jussara Maria - Calculo numerico computacional :teoria e pratica - Editora Atlas (ISBN: 8522424853)

4. David Kincaid; Ward Cheney - Numerical Analysis: Mathematics of Scientific Computing - Editora American Mathematical Society (ISBN: 978-0-8218-4788-6)

5. Elementos de Análise Numérica - SD Conte

Capítulo 1

Erro de Arredondamento, Aritmética Computacional, Algoritmos e Convergência

Iniciamos o estudo de cálculo numérico fazendo, uma breve revisão de teoremas importantes do cálculo. Estes teoremas serão usados para derivar os métodos numéricos usados para aproximar funções, encontrar raízes, resolver equações diferenciais, etc... A seguir fazemos uma discussão rápida sobre a aritmética de máquina, isto é, como o arredondamento de um número real para k dígitos significativos influencia nos cálculos de forma catastrófica, provocando o chamado *erro de arredondamento*.x

1.1 Revisão do Cálculo

Definição 1 *Seja f uma função o definida em um conjunto $\mathcal{X}$ de números reais. Dizemos que f possui limite L em $x = x_0$ e escrevemos que*

$$\lim_{x \to x_0} f(x) = L$$

se

$$\forall \epsilon > 0, \exists \delta > 0, \quad 0 < |x - x_0| < \delta \Rightarrow |f(x) - L| < \epsilon.$$

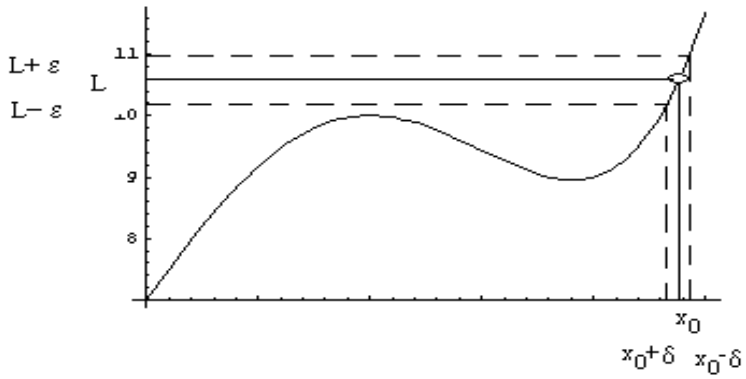


Figura 1.1: $\lim_{x\to x_0} f(x) = L$

Definição 2 *Seja f uma função definida em um conjunto $\mathcal{X}$ de números reais e $x_0 \in \mathcal{X}$. Dizemos que f é uma função contínua em $x_0 \in \mathcal{X}$ se:*

$$\lim_{x\to x_0} f(x) = f(x_0).$$

A função f(x) é contínua em $\mathcal{X}$ se ela é contínua em todo elemento $x \in \mathcal{X}$.

Definição 3 *Seja $\{x_n\}_{n=1}^{\infty}$ uma seqüência infinita de números reais (ou complexos). Dizemos que a seqüência converge para um número x se (fig. 2)*

$$\forall \epsilon > 0, \exists N(\epsilon), \quad \textit{inteiro positivo}, \quad n > N(\epsilon) \Rightarrow |x_n - x| < \epsilon,$$

isto é

$$\lim_{n\to\infty} x_n = x.$$

Os conceitos de convergência e de continuidade estão relacionados pelos teoremas:

Teorema 1 *Seja f uma função definida em um conjunto $\mathcal{X}$ de números reais, $x_0 \in \mathcal{X}$ então: f é contínua em $x_0 \Longleftrightarrow$ se $\{x_n\}_{n=1}^{\infty}$ é uma seqüência em x convergindo para x_0 então $\lim_{x\to x_0} f(x_n) = f(x_0)$.*

Definição 4 *Seja f é uma função definida num intervalo aberto contendo x_0, dizemos que f é diferenciável se:*

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

existe.

Teorema 2 *: Se a função f é diferenciável em x_0 então f é contínua em x_0.*

Observação:

- $\mathcal{C}(\mathcal{X})$: Conjunto de todas as funções contínuas em $\mathcal{X}$.
- $\mathcal{C}^n(\mathcal{X})$: Conjunto de todas as funções com n derivadas contínuas em $\mathcal{X}$.
- $\mathcal{C}^\infty(\mathcal{X})$: Conjunto de todas as funções com todas as derivadas contínuas em $\mathcal{X}$.

Teorema 3 *(**Teorema de Rolle**) $f \in \mathcal{C}[a,b]$ e diferenciável em (a,b). Se $f(a) = f(b) = 0$ então existe um número c, com $a < c < b$ de forma que $f'(c) = 0$.*

Teorema 4 *(**Teorema do Valor Médio**) Seja $f \in \mathcal{C}[a,b]$ e diferenciável em (a,b) então existe um número c, com $a < c < b$ de forma que*

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

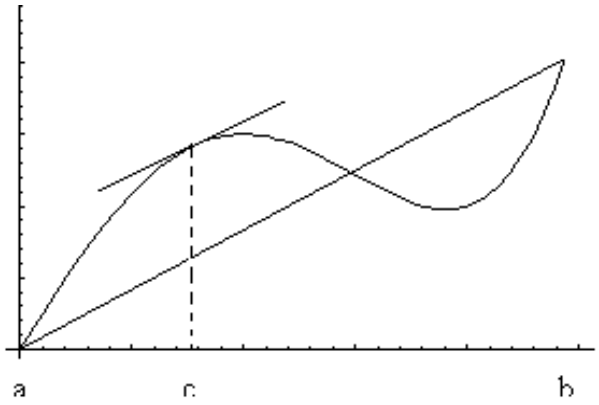


Figura 1.2: $\lim_{x \to x_0} f(x) = L$

Teorema 5 *(**Teorema do Valor Extremo**) Se $f \in \mathcal{C}[a,b]$ então existem c_1 e $c_2 \in [a,b]$, com $f(c_1) \leq f(x) \leq f(c_2)$ para cada $x \in (a,b)$. Se ainda f é diferenciável em (a,b) então os números c_1 e c_2 ocorrem nos extremos do intervalo $[a,b]$ ou onde $f'(x) = 0$.*

Definição 5 *A* **integral de Riemann** *da função f no intervalo $[a,b]$ é dada pelo seguinte limite, quando existir:*

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

onde os números $x_0, x_1, ..., x_n$ são tais que $a = x_0 \leq x_1 \leq ... \leq x_n = b$ e onde $\Delta x_i = x_i - x_{i-1}$ para cada $i = 1, 2, ..., n$ e z_i é um número de forma que $z_i \in [x_{i-1}, x_i]$

Devemos notar que toda função contínua em um intervalo fechado $[a,b]$ admite integral de Riemann neste intervalo. Geralmente escolhamos os pontos x_i igualmente espaçados em $[a,b]$ e podemos também escolher $z_i = x_i$, assim

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

onde $x_i = a + i(b-a)/n$.

Teorema 6 *(**(Teorema do valor médio com peso para integrais**) Se $f \in \mathcal{C}[a,b]$, g é integrável sobre $[a,b]$ e $g(x)$ não troca de sinal em $[a,b]$, então existe um número c, $a < c < b$, tal que*

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

Observação:

- Se $g(x) \equiv 1$ temos o valor médio de f em $[a,b]$ dado por

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

Teorema 7 *(**Teorema de Rolle Generalizado**) Seja $f \in \mathcal{C}[a,b]$, n vezes diferenciável em (a,b). Se f tende a zero nos $n+1$ números distintos $x_0, x_1, ..., x_n$ em $[a,b]$, então existe um número em (a,b) com $f^{(n)}(c) = 0$.*

Teorema 8 *(**Teorema do Valor Intermediário**) Se $f \in \mathcal{C}[a,b]$ e k é um número entre $f(a)$ e $f(b)$ então existe c em (a,b) para o qual $f(c) = k$.*

Exemplo 1: Mostre que $x^5 - 2x^3 + 3x^2 - 1 = 0$ tem zeros em $[0,1]$.
Seja $f(x) = x^5 - 2x^3 + 3x^2 - 1$ daí $f(0) = -1$ e $f(1) = 1$
Assim pelo TV Intermediário podemos afirmar que para $k = 0$, $\exists c \in [0,1]$ / $f(c) = 0$.

Teorema 9 *(**Teorema de Taylor**) Seja $f \in \mathcal{C}^n[a,b]$ e onde $f^{(n+1)}$ existe em $[a,b]$. Seja $x_0 \in [a,b]$. Para cada $x \in [a,b]$, existe $\xi(x) \in (x_0, x)$, com*

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

onde

$$\begin{aligned} P_n(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + ... + \frac{f^n(x_0)}{n!}(x-x_0)^n \\ &= \sum_{k=0}^{n} \frac{f^k(x_0)}{k!}(x-x_0)^k, \end{aligned}$$

e

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

Aqui $P_n(x)$ é chamado de polinômio de Taylor de enésimo grau para f em torno de x_0 e $R_n(x)$ é chamado de resto ou erro de truncamento associado com $P_n(x)$.
As séries infinitas obtidas quando tomamos o limite de $P_n(x)$ quando $n \to \infty$ é chamada de série de Taylor para f em torno de x_0.
No caso particular de $x_0 = 0$ chamamos de polinômio de Maclaurin e série de Maclaurin, respectivamente.

Exemplo 2: Determine os polinômios de Taylor de segundo e terceiro graus para $f(x) = \cos x$ em torno de $x_0 = 0$, e use este polinômio para calcular $\cos(0.01)$ e compare com o valor obtido por sua calculadora.
Pela calculadora, $\cos(0.01) = 0.99995000042$

Agora
$f(x) = \cos x \Rightarrow f(0) = 1$
$f'(x) = -\sin x \Rightarrow f'(0) = 0$
$f''(x) = -\cos x \Rightarrow f''(0) = -1$
$f'''(x) = \sin x \Rightarrow f'''(0) = 0$
$f^{(IV)}(x) = \cos x$

Assim para **n=2**

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \sin \xi(x),$$

com $0 < \xi(x) < 0.01$, logo

$$\cos(0.01) = 0.99995 + \frac{1}{6}(0.01)^3 \sin \xi(x)$$

onde o erro é menor que:

$$\left|\frac{1}{6}(0.01)^3 \sin \xi(x)\right| < \frac{1}{6} \times 10^{-6} \approx 1.7 \times 10^{-7}$$

e para **n=3**

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 \cos \xi(x),$$

com $0 < \xi(x) < 0.01$, logo

$$\cos(0.01) = 0.99995 + \frac{1}{24}(0.01)^4 \cos \xi(x)$$

onde o erro é menor que:

$$\left|\frac{1}{24}(0.01)^4 \cos \xi(x)\right| < \frac{1}{24} \times 10^{-8} \approx 4.2 \times 10^{-10}$$

note que esta estimativa de erro está mais próxima do erro real.

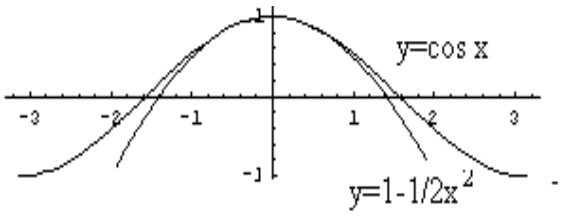


Figura 1.3: $\lim_{x\to x_0} f(x) = L$

As séries de Taylor são o fundamento dos métodos numéricos. Muitas das técnicas numéricas são derivadas diretamente das séries de Taylor, bem como as estimativas de erro na aplicação destas técnicas.

Se o valor de uma função $f(x)$ pode ser expressa em uma região de x próxima de $x = a$ pela série infinita

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \ldots \tag{1.1}$$

então $f(x)$ é dita *analítica* nesta região perto de $x = a$. A série descrita por (1.1) é única e é chamada de *expansão em série de Taylor* de $f(x)$ na vizinhança de $x = a$. Note que para que a série (1.1) exista é necessário que todas as derivadas de $f(x)$ existam em $x = a$, e neste caso podemos encontrar o valor de $f(b)$ para b próximo de a substituindo x por b na série (neste caso, dizemos que b está dentro do raio de convergência da série). Neste caso,

$$f(b) = f(a) + (b-a)f'(a) + \frac{(b-a)^2}{2!}f''(a) + \cdots + \frac{(b-a)^n}{n!}f^{(n)}(a) + \ldots \tag{1.2}$$

$f(b)$ é calculada de forma exata pela série infinita (1.2). Métodos numéricos aproximam o valor de $f(b)$ *truncando* a série (1.2) em um número finito de termos. O erro cometido na série de Taylor para $f(x)$ quando a série é truncada depois o termo contendo $(x-a)^{(n)}$ não pode ser maior que

$$\left|\frac{d^{n+1}f(\xi(x)}{dx^{n+1}}\right|_{\max} \frac{|x-a|^{n+1}}{(n+1)!}, \text{ para } \xi(x) \text{ entre } x \text{ e } a \tag{1.3}$$

onde o subscrito max denota o maior valor que a derivada da $f(x)$ pode assumir no intervalo de a até x. Na verdade este valor não é simples de ser encontrado, isto é nós não temos controle sobre o valor da derivada da função ou sobre o fatorial, mas temos controle sobre o quão próximo x está de a, isto é podemos controlar a quantidade $(x-a)^{n+1}$. Para indicar este fato dizemos que o erro é da ordem de $(x-a)^{n+1}$, e indicamos isto por $O((x-a)^{n+1})$. Por exemplo se a série (1.1) for truncada depois dos primeiros três termos, dizemos que a precisão no cálculo de $f(x)$ é $O((x-a)^3)$, isto é:

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + O((x-a)^3). \tag{1.4}$$

1.2 Erro de Arredondamento e Aritmética Computacional

Quando uma calculadora ou computador é usada para cálculos com números reais ocorre um erro inevitável, o *erro de arredondamento*. Este erro surge por causa da aritmética da máquina que possui apenas um número finito de dígitos, podendo representar assim apenas um número finitos de números reais, todos racionais. Desta forma muitos dos cálculos são feitos com representações aproximadas dos números reais. Vamos enfatizar que todos os números de máquina são todos *racionais*, *positivos* ou *negativos* e possuem apenas um *número finito de algarísmos ou dígitos*.

O sistema de numeração geralmente usado em nossas máquinas é o sistema *binário*, o qual usa os algarismos 0 e 1 para representação dos números. Estes números são armazenados como uma parte fracionária, chamada de mantissa, juntamente com uma parte exponencial, chamada de característica.

Nota 1 *Um bit é a menor unidade armazenada em um computador e é representada pelos valores 0 ou 1. Um conjunto de 8 bits é dito um byte e nele podemos armazenar uma combinação de 0 ou 1 em cada um dos seus 8 bits e desta forma 1 byte pode armazenar* 2^8 *valores inteiros distintos.*

Em geral o computador armazena palavras e o acesso de leitura (escrita) é feito usando uma ou mais palavras que são acessadas através de um endereço único. Normalmente cada palavra é composta por 32 bits ou 64 bits, dependendo do processador usado.

Por exemplo um número de ponto flutuante de precisão simples usado em alguns computadores consiste de:

- Um dígito binário (bit) que indica o sinal;
- 7-bit para o expoente com base 16;
- 24-bit para a mantissa.

24-bit correspondem entre 6 ou 7 dígitos decimais, assim vamos assumir que temos pelo menos seis dígitos de precisão para um número de ponto flutuante. O expoente de 7 dígitos binários nos dá uma variação de 0 até 127. Entretanto como desejamos representar também expoentes negativos isto é queremos representar números pequenos, subtrairmos 64 da característica e assim o campo da exponencial varia de -64 até 63.
Por exemplo seja o seguinte número de máquina:

0	1000010	10110011000001000000000

(1.5)

- O bit mais da esquerda, 0 no caso, indica que o número é positivo.

- Os próximos sete são equivalentes ao número decimal:

$$1.2^6 + 0.2^5 + 0.2^4 + 0.2^3 + 0.2^2 + 1.2^1 + 0.2^0 = 66$$

e são usados para descrever a característica, isto é o expoente de 16 (nosso computador usa base hexadecimal) será

$$66 - 64 = 2$$

.

- Os últimos 24 bits indicam que a mantissa e correspondem:

$$1.\left(\frac{1}{2}\right)^1 + 1.\left(\frac{1}{2}\right)^3 + 1.\left(\frac{1}{2}\right)^4 + 1.\left(\frac{1}{2}\right)^7 + 1.\left(\frac{1}{2}\right)^8 + 1.\left(\frac{1}{2}\right)^{14} = 0.69927978515625$$

Desta forma, o número de máquina representado por (1.5) é dado por:

$$+0.69927978515625 * 16^2 = 179.015625$$

Note que, na nossa máquina, o número imediatamente menor e o imediatamente maior que (1.5) são:

0	1000010	10110011000000111111111

$= \mathbf{179.0156}097412109375$ (1.6)

0	1000010	10110011000001000000000

$= 179.015625$ (1.7)

0	1000010	10110011000001000000001

$= \mathbf{179.0156}402587890625$ (1.8)

Assim sendo o número (1.5) é usado para representar um intervalo de números reais,

$$[179.0156173706054875, 179.01563262939453125]$$

Para assegurar unicidade de representação e obter toda a precisão disponível no sistema é imposta uma normalização, isto é ao menos um dos quatro dígitos mais a esquerda da mantissa do número de máquina é um. Assim o menor e o maior números de máquina positivos são respectivamente:

0	0000000	00010000000000000000000

$= 16^{-65} \approx 10^{-78}$ (1.9)

0	1111111	11111111111111111111111

$\approx 16^{63} \approx 10^{76}.$ (1.10)

Nos cálculos números abaixo do menor valor resultam um *underflow* e muitas vezes são vistos como zero, números acima do maior valor produzem um *overflow*, o que faz parar a computação.

Assim um número de máquina é representado como,

$$\pm 0.d_1d_2...d_k \times 10^n, \qquad 1 \leq d_1 \leq 9, \qquad 0 \leq d_i \leq 9, \qquad i = 2, ..., k \tag{1.11}$$

onde no nosso exemplo $k = 6$ e $-78 \leq n \leq 76$. Estes números são chamados de números decimais de máquina. Para representar números reais como número decimal de máquina é feito um *corte*, que despreza qualquer dígito após d_k ou o *arredondamento* que adiciona $5 \times 10^{n-(k-1)}$ ao número real antes de proceder o "corte".

Observação:

Em uma "máquina"a representação de ponto flutuante $fl(y)$ de um número y possui um erro relativo. Vamos supor que estamos usando corte com k dígitos significativos daí se o número real

$$y = 0.d_1d_2d_3...d_kd_{k+1}... \times 10^n,$$

e temos seu erro relativo como:

$$\begin{aligned}
\left|\frac{y-fl(y)}{y}\right| &= \left|\frac{0.d_1d_2d_3...d_kd_{k+1}...\times 10^n - 0.d_1d_2d_3...d_k\times 10^n}{0.d_1d_2d_3...d_k\times 10^n}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}d_{k+3}...\times 10^{n-k}}{0.d_1d_2d_3...d_k\times 10^n}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}d_{k+3}...}{0.d_1d_2d_3...d_k}\right| \times 10^{-k} \\
&\leq \frac{1}{0.1} \times 10^{-k} = 10^{-k+1}.
\end{aligned}$$

De forma similar um limite para o erro relativo quando usamos aritmética de k dígitos com arredondamento é $0.5 \times 10^{-k+1}$.

Definição 6 *Se p^* é uma aproximação de p, o erro absoluto é dado por $|p - p^*|$, e o erro relativo é dado por $|p - p^*|$ / $|p|$, providenciando que $p \neq 0$.*

O erro relativo e o erro absoluto são em geral distintos. O erro absoluto depende do problema trabalhado (este depende de unidade), se estamos por exemplo usando metros e medindo a distância entre Porto Alegre e Caxias do Sul, um erro absoluto é de $2.0 \times 10^1 m$ é desprezível. Agora se estamos medindo o comprimento de um quarto erro de dois metros é muito grande! Agora dizer que o erro relativo é 2×10^{-1} significa um erro de 20% e isto não depende de unidades.

Definição 7 *O número p^* é dito aproximar p com t algarismos significativos se t é o maior inteiro não negativo para o qual*

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

Em adição a imprecisão de representação dos números, a aritmética feita em um computador não é exata. Não entraremos a fundo neste assunto, mas vamos exemplificar o que pode ocorrer. Sejam:
• x e y números reais
• $fl(x)$ e $fl(y)$ representações em ponto-flutuante
• $\oplus$, $\ominus$, $\otimes$ e $\oslash$ representando as operações de máquina, definidas como:

$$\begin{array}{ll} x \oplus y = fl(fl(x) + fl(y)) & x \ominus y = fl(fl(x) - fl(y)) \\ x \otimes y = fl(fl(x) \times fl(y)) & x \oslash y = fl(fl(x)/fl(y)) \end{array} \tag{1.12}$$

Por exemplo seja:
Vamos considerar uma aritmética de 5 dígitos com corte.
Exemplo 1:
$x = 1/3$ e $y = 5/7$ assim $fl(x) = 0.33333 \times 10^0$ e $fl(y) = 0.71428 \times 10^0$

Operação	Resultado	Valor Real	Erro Absoluto	Erro Relativo
$x \oplus y$	0.10476×10^1	$22/21$	0.190×10^{-4}	0.182×10^{-4}
$y \ominus x$	0.38095×10^0	$8/21$	0.238×10^{-5}	0.625×10^{-5}
$x \otimes y$	0.23809×10^0	$5/21$	0.524×10^{-5}	0.220×10^{-4}
$y \oslash x$	0.21428×10^1	$15/7$	0.571×10^{-4}	0.267×10^{-4}

Exemplo 2:
$u = 0.714251$, $v = 98765.9$ e $w = 0.111111 \times 10^{-4}$ assim $fl(u) = 0.71425 \times 10^0$, $fl(v) = 0.98765 \times 10^5$ e $fl(w) = 0.11111 \times 10^{-4}$

Operação	Resultado	Valor Real	Erro Absoluto	Erro Relativo
$y \ominus u$	0.30000×10^{-4}	0.34714×10^{-4}	0.471×10^{-5}	0.136
$(y \ominus u) \oslash w$	0.27000×10^1	0.31243×10^1	0.424	0.136
$(y \ominus u) \otimes v$	0.29629×10^1	0.34285×10^1	0.465	0.136
$u \oplus v$	0.98765×10^5	0.98766×10^5	0.161×10^1	0.163×10^{-4}

Na última tabela podemos ver problemas que ocorrem quando operamos com aritmética finita.
Problemas mais comuns

- Cancelamento de dígitos significativos devido a subtração de números quase iguais
  Sejam:

$$fl(x) = 0.d_1d_2...d_p\alpha_{p+1}\alpha_{p+2}...\alpha_k \times 10^n$$
$$fl(y) = 0.d_1d_2...d_p\beta_{p+1}\beta_{p+2}...\beta_k \times 10^n$$

Note que:

$$fl(x) - fl(y) = 0.00...0\gamma_1\gamma_2...\gamma_{k-p} \times 10^n$$

onde $\gamma_i = \alpha_{p+i} - \beta_{p+i}$. Assim

$$fl(fl(x) - fl(y)) = 0.\gamma_1...\gamma_{k-p}\gamma_{k-p+1}...\gamma_k \times 10^{n-p}$$

Os números $\gamma_{k-p}...\gamma_k$ são "criados"em alguns computadores de forma aleatória e em outros são considerados nulos. Esta falta de precisão será carregada em operações posteriores.

- Divisão por número muito pequeno ou multiplicação por número muito grande
  Seja z um número real e $z + \delta$ sua representação em ponto flutuante, onde δ é o erro cometido no corte ou arredondamento. Vamos supor que $\epsilon = 10^{-n}$, assim

$$\frac{z}{\epsilon} \approx \frac{fl(z+\delta)}{fl(\epsilon)} = (z+\delta) \times 10^n$$

  Isto é o erro absoluto é de $|\delta| \times 10^n$ o que pode ser muito grande!

Observação:
Algumas vezes a perda de dígitos significativos e precisão podem ser evitadas re-ordenando as operações.

Exemplo 3:
Vamos considerar a equação quadrática:

$$x^2 + \frac{6210}{100}x + 1 = 0,$$

que possui solução exata dada por:

$$x_1 = \frac{-621 - \sqrt{385241}}{20} \approx -62.083892762591031419$$

e

$$x_2 = \frac{-621 + \sqrt{385241}}{20} \approx -0.016107237408968581$$

Vamos, agora, refazer os cálculos usando uma "máquina"que usa arredondamento com 4 algarismos significativos. Assim

$$\begin{array}{lcll} b^2 & = & 3856.41 & = & 3856. \\ 4ac & = & 4 & = & 4.000 \\ \sqrt{b^2-4ac} & = & \sqrt{3852.} & \approx & 62.06 \end{array}$$

Assim,

$$fl(x_1) = \frac{-b + \sqrt{b^2-4ac}}{2a} = \frac{-62.10 + 62.06}{2.0000} = \frac{-0.04000}{2.000} = -0.02000$$

é uma aproximação para $x_1 = -0.01611$ e o erro relativo cometido no cálculo é grande,

$$\frac{|-0.01611 + 0.02000|}{|-0.01611|} \approx 2.4 \times 10^{-1}$$

O problema desta equação é que b^2 é muito maior que $4ac$ e assim o numerador, no cálculo de x_1, envolve uma subtração de números próximos.
No cálculo de x_2 este problema não ocorre, pois envolve soma de dois números próximos, assim,

$$fl(x_2) = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{-62.10 - 62.06}{2.0000} = \frac{-124.2}{2.000} = -62.10$$

é a aproximação de $x_2 = -62.08$, e o erro relativo é

$$\frac{|-62.08 + 62.10|}{|-62.08|} \approx 3.2 \times 10^{-4}.$$

Para melhorarmos o cálculo de x_1 podemos proceder a seguinte racionalização:

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

Logo,

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}.$$

e no nosso caso teríamos

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = \frac{-2.000}{124.2} = -0.01610$$

O erro relativo agora é de 6.2×10^{-4}.
Note: Se usarmos esta fórmula no cálculo de x_2 obteríamos -50.00, isto é um erro relativo de 1.9×10^{-1}.

Perda de Precisão

Uma boa questão é saber quantos bits significativos binários são perdidos com a subtração de dois números próximos x e y? A resposta dependem dos valores de x e y, entretanto temos um teorema que nos fornece os limites em função da quantidade $\|1 = y/x\|$, que é uma medida de quanto x e y estão próximos.

Teorema 10 Perda de Precisão *Se x e y são números binários de máquina de ponto flutuante normalizados tais que $x > y$ e*

$$2^{=q} \leq 1 - \frac{y}{x} \leq 2^{-p}$$

então no máximo q e pelo menos p bits binários significantes são perdidos na subtração $x - y$.

Prova: ver Kincaid e Cheney, Numerical Analysis pg 60.

Exemplo 4:
Vamos considerar a função

$$y(x) = x - \sin(x)$$

quando x é pequeno, temos que $\sin(x) \approx x$, este cálculo irá ocasinar perda de dígitos significantes. Podemos evitar neste caso esta perda, usando série de Taylor, assim:

$$\begin{aligned} y(x) &= x - \sin(x) = x - \left(x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!} - \tfrac{x^7}{7!} + ...\right) \\ &= \tfrac{x^3}{3!} - \tfrac{x^5}{5!} + \tfrac{x^7}{7!} + ... \end{aligned}$$

Desta forma para x pequeno temos que

$$y(x) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} + ... \tag{1.13}$$

Para fazrmos um programa para calcular esta série, podemos gerar uma sequência de termos t_i como

$$\begin{cases} t_1 & = x^3/6 \\ t_{n+1} & = -t_n x^2/[(2n+2)(2n+3)], \qquad n \geq 1 \end{cases}$$

e a soma parcial desta sequência pode ser obtida por:

$$\begin{cases} s_1 & = t_1 \\ s_{n+1} & = s_n t_{n+1}\ , \qquad n \geq 1 \end{cases}$$

Vamos agora mostrar que devemos sempre calcular polinômios de forma "encaixada"para evitar erros de arredondamento. Primeiramente vamos a um exemplo:

Exemplo 5:
Vamos calcular $f(4.71)$ usando aritmética de três dígitos com arredondamento, onde f(x) é o polinômio:

$$f(x) = x^3 - 6x^2 + 3x - 0.149$$

	x	x^2	x^3	$6x^2$	$3x$
$Exato$	4.71	22.1841	104.487111	133.1046	14.13
$Três - dígitos(arred.)$	4.71	22.2	105.	133.	14.1

$$\begin{array}{lll} Exato & f(4.71) = & 104.487111 - 133.1046 + 14.13 - 0.149 = -14.636489 \\ Três - d\acute{i}gitos(arred.) & f(4.71) = & 105. - 133. + 14.1 - 0.149 = -14.1 \end{array}$$

O erro relativo cometido é:

$$\left|\frac{-14.636489 - 14.0}{-14.636489}\right| \approx 0.04.$$

Uma alternativa para este cálculo é usar a forma **encaixada**,

$$f(x) = x^3 - 6x^2 + 3x - 0.149 = ((x-6)x + 3)x - 0.149.$$

Usando esta forma de cálculos temos:

$$f(4.71) = ((4.71 - 6)4.71 + 3)4.71 - 0.149 = -14.6$$

e o erro relativo cometido baixa para 0.0025.

Obervações:

- Sempre que formos calcular o valor de um polinômio em um ponto devemos faze-lo na forma "encaixada"para evitar erros de arredondamento.
- Este texto foi adaptado da referência *Análise Numérica - Burden and Faires e Numerical Analysis - Kincaid e Cheney.*

1.3 Ordens de convergência e conceitos básicos adicionais

Em cálculos numéricos muitas vezes a resposta é obtida como uma seqüência de valores que usualmente exibe progressivamente uma maior precisão. Desta forma a convergência de seqüências é um assunto muito importante em análise numérica.

1.3.1 Convergência de Seqüências

Escrevemos

$$\lim_{n\to\infty} x_n = L$$

se existe uma correspondência para cada ϵ positivo de um número real r de forma que $|x_n - L| < \epsilon$ sempre que $n > r$.
Por exemplo:

$$\lim_{n\to\infty} \frac{n+1}{n} = 1$$

porque

$$\left|\frac{n+1}{n} - 1\right| < \epsilon$$

sempre que $n > \epsilon^{-1}$.
Como outro exemplo, lembramos do cálculo que

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$$

Se calculamos a seqüência $x_n = (1 + 1/n)^n$, temos:

n	x	erro relativo	$\lim_{n\to\infty} x_{n+1}/x_n$
1	2.00000000	0.26424112	
100	2.70481383	0.00495460	0.01875030
200	2.71151712	0.00248860	0.50227992
300	2.71376516	0.00166159	0.66768176
400	2.71489174	0.00124714	0.75057148
500	2.71556852	0.00099817	0.80036593
600	2.71602005	0.00083206	0.83358754
700	2.71634274	0.00071335	0.85732967
800	2.71658485	0.00062428	0.87514305
900	2.71677321	0.00055499	0.88900193
1000	2.71692393	0.00049954	0.90009158
1100	2.71704727	0.00045417	0.90916660
1200	2.71715008	0.00041635	0.91673027
1300	2.71723707	0.00038434	0.92313112
1400	2.71731165	0.00035691	0.92861816
1500	2.71737629	0.00033313	0.93337405
1600	2.71743285	0.00031232	0.93753578
1700	2.71748276	0.00029396	0.94120817
1800	2.71752713	0.00027764	0.94447272
1900	2.71756684	0.00026303	0.94739380
2000	2.71760257	0.00024989	0.95002291
2100	2.71763490	0.00023799	0.95240173
2200	2.71766429	0.00022718	0.95456438
2300	2.71769113	0.00021730	0.95653906
2400	2.71771574	0.00020825	0.95834924
2500	2.71773837	0.00019993	0.96001466
2600	2.71775927	0.00019224	0.96155201
2700	2.71777861	0.00018512	0.96297553
2800	2.71779658	0.00017851	0.96429740
2900	2.71781331	0.00017236	0.96552814
3000	2.71782892	0.00016662	0.96667685

Note que usando dupla precisão no FORTRAN, temos evidência que $\lim_{n\to\infty} x_{n+1}/x_n \longrightarrow 1$. Esta convergência é muito lenta, na verdade é menos que linear.

Como um exemplo de uma seqüência com convergência rápida vamos definir por recorrência a seguinte seqüência convergente para o valor de $\sqrt{2}$,

$$\left\{ \begin{array}{rcl} x_1 & = & 2 \\ x_{n+1} & = & \frac{1}{2}x_n + \frac{1}{x_n}, \quad n \geq 1 \end{array} \right.$$

Os elementos produzidos por esta seqüência são:

$$\left\{ \begin{array}{l} x_1 = 2.000000 \\ x_2 = 1.500000 \\ x_3 = 1.416667 \\ x_4 = 1.414216 \end{array} \right.$$

O limite desta seqüência é $\sqrt{2} = 1.414213562...$ e devemos notar que esta seqüência está convergindo para o limite muito rapidamente. Usando computação com dupla-precisão encontramos

a evidência que

$$\frac{|x_{n+1} - \sqrt{2}|}{|x_n - \sqrt{2}|^2} \leq 0.36$$

Esta condição corresponde a uma convergência de ordem 2, isto é quadrática.

1.3.2 Ordem de convergência

Existe uma terminologia especial para definir o quanto rápido uma seqüência converge para seu limite. Seja $\{x_n\}_{n=1}^{\infty}$ uma seqüência de números reais tendendo a um limite x^*. Dizemos que a razão de convergência desta seqüência é **linear** se existe uma constante $0 < c < 1$ e um inteiro N de forma que

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

A convergência é dita **quadrática** se existe uma constante C, não necessariamente menor que 1, e N inteiro tal que

$$|x_{n+1} - x^*| \leq C|x_n - x^*|^2 \quad (n \geq N)$$

Em geral, se existem constantes positivas C e α e um inteiro N de forma que

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

ou

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

dizemos que a taxa de convergência é de pelo menos α.

Definição 8 *Vamos supor que $\{\beta_n\}_{n=1}^{\infty}$ é uma seqüência que converge para zero, e que $\{\alpha_n\}_{n=1}^{\infty}$ é uma seqüências que converge para um número α. Se existe uma constante positiva K de forma que*

$$|\alpha_n - \alpha| \leq K|\beta_n|, \quad \textit{para um valor grande de } n,$$

então dizemos que $\{\alpha_n\}_{n=1}^{\infty}$ converge para α com uma taxa de convergência $\mathcal{O}(\beta_n)$. Indicamos a taxa de convergência da seguinte maneira:

$$\alpha_n = \alpha + \mathcal{O}(\beta_n).$$

Embora a definição (8) permita comparar duas seqüências arbitrárias, em quase todos os casos usamos

$$\beta_n = \frac{1}{n^p},$$

para um número qualquer $p > 0$. Geralmente estamos interessados no maior valor de p para o qual $\alpha_n = \alpha + \mathcal{O}(1/n^p)$.

Exemplo 1 *Para $n \geq 1$, vamos definir duas seqüências como:*

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

Podemos notar que $\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \hat{\alpha}_n = 0$, mas a seqüência $\{\hat{\alpha}_n\}$ converge muito mais rapidamente para zero que $\{\alpha_n\}$, quando utilizamos uma aproximação por arredondamento de cinco dígitos, temos:

n	1	2	3	4	5	6	7
α_n	2.00000	0.75000	0.44444	0.31250	0.24000	0.19444	0.16327
$\hat{\alpha}_n$	4.00000	0.65500	0.22222	0.10938	0.064000	0.041667	0.029155

Se fizermos $\beta_n = 1/n$ e $\hat{\beta}_n = 1/n^2$, veremos que:

$$|\alpha_n - 0| = \frac{n+1}{n^2} \leq \frac{n+n}{n^2} = 2.\frac{1}{n} = 2\beta_n$$

e

$$|\hat{\alpha}_n - 0| = \frac{n+3}{n^3} \leq \frac{n+3n}{n^3} = 4.\frac{1}{n^2} = 4\hat{\beta}_n,$$

e portanto

$$|\alpha_n| = 0 + \mathcal{O}\left(\frac{1}{n}\right) \quad e \quad \hat{\alpha}_n = 0 + \mathcal{O}\left(\frac{1}{n^2}\right).$$

1.3.3 Algoritmos

Um algoritmo é um procedimento que descreve de uma forma não ambígua uma seqüência finita de passos a serem executados em uma certa ordem. O objetivo de um algoritmo é implementar um procedimento numérico para encontrar uma solução aproximada de um problema.

Usaremos um pseudocódigo para descrever os algoritmos. Este pseudocódigo especifica a forma dos dados de entrada a serem fornecidos e a forma desejada dos resultados produzidos. Nem todos os procedimentos numéricos produzem resultados satisfatórios para dados de entrada escolhidos arbitrariamente. Desta forma uma técnica independente para interrupções dos procedimentos é incorporada para prevenir cálculos com *loops infinitos*.

Normalmente dois sinais de pontuação são usados dentro de um algoritmo: um ponto indica fim de um passo e um ponto e vírgula separa tarefas dentro de um passo. Os passos de um algoritmo geralmente seguem as regras de uma construção de programa estruturado. Ele deve ser feito de forma que não haja a mínima dificuldade para traduzi-lo em uma linguagem de programação aceitável para aplicações científicas.

Exemplo 2 *Vamos fazer um algoritmo para somar N números*

ALGORITMO 1 *Este algoritmo soma N números dados.*

INPUT	$N, x_1, x_2, ..., x_n$.
SAÍDA	$SOMA = \sum_{i=1}^{N} x_i$.
Passo1	*Faça* $SOMA = 0$.
Passo2	*Para* $i = 1, 2, ..., N$ *Faça*
	$SOMA = SOMA + x_i$.
Passo3	*SAÍDA* $SOMA$.
	PARE.

Estamos interessados em escolher métodos que produzam resultados confiáveis e precisos para um grande número de problemas.

Dizemos que um algoritmo é estável se pequenas alterações nos dados iniciais produzam pequenas alterações nos resultados.

Em caso contrário o algoritmo é dito **instável**. Devemos notar que alguns algoritmos são estáveis para alguns valores dos dados iniciais e para outros não. Estes algoritmos são ditos **condicionalmente estáveis**.

Vamos agora relacionar o crescimento do erro de arredondamento e a estabilidade de um algoritmo. Vamos supor que um erro $E_0 > 0$ seja introduzido em algum estágio do procedimento de um cálculo e que depois de n operações subseqüentes seja indicado por E_n. Aqui vamos definir os dois casos mais comuns na prática.

O crescimento do erro é dito **linear** se $E_n \approx CnE_0$, onde C é uma constante independente de n. Agora se $E_n \approx C^n E_0$ para qualquer $C > 1$ então o crescimento do erro é dito **exponencial**. Devemos notar que o crescimento linear normalmente é inevitável, agora o erro exponencial deve ser evitado. Muitas vezes um método atrativo (simples e fácil de usar) pode ser instável, isto é existem métodos que seriam ótimos em um computador perfeito, mas em computadores que usam aritimética de ponto flutuante seriam métodos imprecisos. No caso de computadores precisamos que os métodos sejam estáveis.

Exemplo 3 *O primeiro exemplo é um pouco artificial, porém é um bom exemplo. Existe um algoritmo que calcula todas as potências inteiras do número chamado na literatura de "Golden Mean", através de uma fórmula de recorrência envolvendo apenas operações simples (subtração) definida em (1.14). Este número é definido por:*

$$\phi \equiv \frac{\sqrt{5}-1}{2} = 0.61803398...$$

e podemos mostrar que para gerar a sequências das potências deste número, basta calcular a seguinte relação de recorrência que usa apenas a operação de subtração,

$$\phi^{n+1} = \phi^{n-1} - \phi^n. \tag{1.14}$$

Note que esta seqüência converge para zero e todos os seus termos são positivos! (Como exercício verifique que esta relação está correta).

Assim temos: $\phi^0 = 1$ *e* $\phi^1 = 0.6180340$, *usando um programa numérico calculamos que:*

Φ^0	1.0000000
Φ^1	0.6180340
Φ^2	0.3819660
Φ^3	0.2360680
Φ^4	0.1458980
Φ^5	0.0901699
Φ^6	0.0557282
Φ^7	0.0344417
Φ^8	0.0212864
Φ^9	0.0131553
Φ^{10}	0.0081311
Φ^{11}	0.00502422
Φ^{12}	0.00310688
Φ^{13}	0.00191734
Φ^{14}	0.00118954
Φ^{15}	0.00072780
Φ^{16}	0.00046174
Φ^{17}	0.00026606
Φ^{18}	0.00019568
Φ^{19}	0.00007038
Φ^{20}	0.00012530
Φ^{21}	-0.00005492
Φ^{22}	0.00018022
Φ^{23}	-0.00023514
Φ^{24}	0.00041536
Φ^{25}	-0.00065050
Φ^{26}	0.00106586
Φ^{27}	-0.00171636
Φ^{28}	0.00278222
Φ^{29}	-0.00449858
Φ^{30}	0.00728080

Os números negativos que aparecem na tabela anterior ocorrem porque esta fórmula envolve subtração "catastrófica", isto é subtração de números muito próximos!!!!!

Exemplo 4 *Neste segundo exemplo, vamos considerar a solução de duas equações em diferenças lineares com coeficientes constantes e ordem 2. A teoria deste tipo de equação é parecida com a de equações diferenciais ordinárias lineares com coeficientes constantes.*

1. *A primeira equação que vamos considerar é:*

$$p_n = \frac{10}{3} p_{n-1} - p_{n-2} \text{ para } n = 2, 3, ..., \tag{1.15}$$

cuja solução é dada por

$$p_n = c_1 \left(\frac{1}{3}\right)^n + c_2 3^n \tag{1.16}$$

pois, se fazemos $n = i-1$ *e* $n = i-2$ *em (1.16) temos:*

$$\begin{array}{l} p_{i-1} = c_1 \left(\frac{1}{3}\right)^{i-1} + c_2 3^{i-1} \\ p_{i-2} = c_1 \left(\frac{1}{3}\right)^{i-2} + c_2 3^{i-2} \end{array} \tag{1.17}$$

Agora, substituindo as equações (1.17) em (1.15) temos:

$$\frac{10}{3} p_{i-1} - p_{i-2} = \frac{10}{3}\left(c_1 \left(\frac{1}{3}\right)^{i-1} + c_2 3^{i-1}\right) + c_1 \left(\frac{1}{3}\right)^{i-2} + c_2 3^{i-2}$$

$$= c_1 \left(\frac{1}{3}\right)^{i-2} \left(\frac{10}{3}\,\frac{1}{3} - 1\right) + c_2 3^{i-2} \left(\frac{10}{3}\,3 - 1\right) =$$

$$= c_1 \left(\frac{1}{3}\right)^{i-2} \left(\frac{1}{3}\right)^2 + c_2 3^{i-2}(3)^2 = c_1 \left(\frac{1}{3}\right)^i + c_2 3^i = p_i$$

Desta forma acabamos de mostrar que (1.16) é solução da (1.15). Para determinarmos as constantes c_1 *e* c_2 *precisamos de 2 condições iniciais, aqui vamos considerar como condição inicial*

$$p_0 = 1 \text{ e } p_1 = 1/3.$$

Neste caso, as constantes c_1 *e* c_2 *são calculadas fazendo* $i = 0$ *e* $i = 1$ *em (1.16) e resolvendo o sistema linear resultante:*

$$\left\{ \begin{array}{l} p_0 = c_1 \left(\frac{1}{3}\right)^0 + c_2 3^0 \\ \\ p_1 = \frac{c_1}{3} + 3c_2 \end{array} \right. \rightarrow \left\{ \begin{array}{l} 1 = c_1 + c_2 \\ \\ \frac{1}{3} = \frac{c_1}{3} + 3c_2 \end{array} \right. \rightarrow c_1 = 1 \text{ e } c_2 = 0,$$

e a solução particular do nosso problema é

$$p_n = (1/3)^n$$

para todo n. Esta solução tende para zero quando n cresce.

Por outro lado, se realizamos estes cálculos em uma máquina com aritmética de cinco dígitos, a condição inicial seria

$$\hat{p}_0 = 1.0000 \text{ e } \hat{p}_1 = 0.33333.$$

Se calculamos as constantes c_1 *e* c_2 *nesta máquina com 5 dígitos encontramos,*

$$\hat{c}_1 = 1.0000 \text{ e } \hat{c}_2 = -0.12500 \times 10^{-5} \neq 0.$$

Note que o erro gerado por esta aproximação é

$$p_n - \hat{p}_n = 0.12500 \times 10^{-5} * 3^n,$$

isto é, conforme o n cresce o erro de arredondamento cresce exponencialmente, fazendo que a solução não vá para zero quando n cresce, como pode ser observado na tabela

n	$\hat{p}_n$ calculado	p_n exato	Erro Relativo
0	0.10000×10^1	0.10000×10^1	
1	0.33333×10^0	0.33333×10^0	
2	0.11110×10^0	0.11111×10^0	9×10^{-5}
3	0.37000×10^{-1}	0.37037×10^{-1}	1×10^{-3}
4	0.12230×10^{-2}	0.12346×10^{-1}	9×10^{-3}
5	0.37660×10^{-1}	0.41152×10^{-2}	8×10^{-1}
6	0.32300×10^{-3}	0.13717×10^{-2}	8×10^{-1}
7	-0.26893×10^{-2}	0.45725×10^{-3}	7×10^0
8	-0.92872×10^{-2}	0.15242×10^{-3}	6×10^1

onde podemos observar que quando n=8 o erro cometido é da ordem de 10^1*! Isto é o erro cresce exponencialmente.*

2. *Agora vamos considerar outra equação em diferenças*

$$p_n = 2p_{n-1} - p_{n-2} \text{ para } n = 2, 3, ...,$$

Sua solução é dada por

$$p_n = c_1 + c_2 n$$

Neste caso, também vamos considerar as condições iniciais,

$$p_0 = 1 \text{ e } p_1 = \frac{1}{3},$$

as constantes c_1 *e* c_2 *são dadas por:* $c_1 = 1$ *e* $c_2 = -2/3$*. Desta forma a solução geral do nosso problema é dado por:*

$$p_n = 1 - \frac{2}{3}n.$$

Agora, se consideramos uma máquina com aritmética de 5 dígitos temos:

$$\hat{p}_0 = 1.0000 \text{ e } \hat{p}_1 = 0.33333.$$

Calculando as constantes nesta máquina com 5 dígitos encontramos,

$$\hat{c}_1 = 1.0000 \text{ e } \hat{c}_2 = -0.66667.$$

Note que o erro gerado por esta aproximação é $p_n - \hat{p}_n = (0.66667 - 2/3)n$*, isto é o erro de arredondamento cresce linearmente com n, como pode ser observado na seguinte tabela:*

n	$\hat{p}_n$ calculado	p_n exato	Erro Relativo
0	0.10000×10^1	0.10000×10^1	
1	0.33333×10^0	0.33333×10^0	
2	-0.33330×10^0	-0.33333×10^0	9×10^{-5}
3	-0.10000×10^1	-0.10000×10^1	0
4	-0.16667×10^1	-0.16667×10^1	0
5	-0.23334×10^1	-0.23333×10^1	4×10^{-5}
6	-0.30000×10^1	-0.30000×10^1	0
7	-0.36667×10^1	-0.36667×10^1	0
8	-0.43334×10^1	-0.43333×10^1	2×10^{-5}

Este tipo de erro de arredondamento pode ser reduzidos se fazemos os cálculos com um número maior de dígitos significativos. Mas mesmo assim, devemos notar que este tipo de erro pode ser minorado mas não eliminado. Ele será apenas postergado quando trabalhamos com aritmética finita.

1.4 Observações Finais - Operações com Ponto Flutuante

Em 1985 o IEEE - *Institute for Eletrical And Eletronic Engineers* publicou normas para ponto flutuante binário com o artigo *Binary Floating Point Arithmetic Standard 754-1985*. Estas normas especificam formatos para precisão simples, dupla e extendida que são, em geral, seguidas pelos fabricantes. Por exemplo co-processador numérico dos PCs usa representação de 64 bits para número real,

1. 1 bit para sinal (s),
2. 11 bits para a parte exponencial (c) *característica*,
3. 52 bits para a parte fracionária (f) *mantissa* e
4. base $\beta = 2$.

- 52 bits binários significa 16 ou 17 dígitos decimais, logo um número representado neste sistema possui precisão de 16 dígitos decimais.
- 11 bits exponencial significa valores entre 0 e $2^{11} - 1 = 2047$. Usa-se somente inteiros positivos no expoente, logo para que números de pequena magnitude sejam representados o valor 1023 é subtraído da característica e assim o expoente varia entre -1023 até 1024.
- Por economia de espaço de armazenamento e com a finalidade de fornecer uma representação única para cada número de ponto flutuante, é feita uma normalização. O uso deste sistema possui a forma:
$$(-1)^s 2^{c-1023}(1+f).$$

Lembrando, **erro de arredondamento** é o erro resultante da substituição de um número real por sua notação em ponto flutuante.

Nota 2 *Existem casos onde o erro de arredondamento é muito grave. Nestes casos pode ser usada a aritmética intervalar, onde um intervalo, e não apenas um número é considerado. Aritmética intervalar é uma aritmética definida por desigualdades, isto é um "número"é definido por um intervalo no qual ele está contido. As operações são definidas de forma intervalar, por exemplo se* $a \leq x \leq b$, *e* $c \leq y \leq d$, *então* $a + c \leq x + y \leq b + d$. *As operações entre intervalos são definidas por:*

- $[a, b] + [c, d] = [a + c, b + d]$
- $[a, b] - [c, d] = [a - c, b - d]$
- $[a, b] \times [c, d] = [\min\{ac, ad, bc, bd\}, \max\{ac, ad, bc, bd\}]$
- $[a, b] \div [c, d] = [a, b] \times [d^{-1}, c^{-1}]$, *se* $0 \notin [c, d]$

Este tipo de aritmética mantém o erro de arredondamento sob controle. Para mais informações sobre o assunto ver o site http://www.cs.utep.edu/interval-comp/index.html.

Primeira Lista de Exercícios

1. Calcule o erro absoluto e erro relativo nas aproximações de p para p^*.

   (a) $p = \pi$, $p^* = 22/7$

   (b) $p = \pi$, $p^* = 3.1416$

   (c) $p = e$, $p^* = 2.718$

2. Encontre o maior intervalo em que p^* deve se encontrar para aproximar p com um erro relativo no máximo de 10^{-4}.

   (a) π

   (b) $\sqrt{2}$

   (c) 150

   (d) 1500

3. Execute os seguintes cálculos: (i) De forma exata; (ii) Usando aritmética de três dígitos e corte; (iii) Usando aritmética de três dígitos e arredondamento. (iv) Calcule os erros relativos dos ítens (ii) e (iii)

   (a) $\frac{4}{5} + \frac{1}{3}$

   (b) $\left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20}$

   (c) $\left(\frac{1}{3} - \frac{3}{11}\right) - \frac{3}{20}$

4. O número e pode ser definido por $e = \sum_{n=0}^{\infty}(1/n!)$. Calcule o erro absoluto e relativo nas seguintes aproximações de e:

   (a) $\sum_{n=0}^{5} \frac{1}{n!}$

   (b) $\sum_{n=0}^{10} \frac{1}{n!}$

5. Seja

$$f(x) = \frac{x\cos x - \sin x}{x - \sin x}.$$

   (a) Encontre o $\lim_{x\to 0} f(x)$.

   (b) Utilize arredondamento com 4 dígitos para calcular $f(0.1)$.

   (c) Substitua cada função trigonométrica por seu polinômio de McLaurin de terceiro grau e repita o item (b).

(d) O valor real é $f(0.1) = -1.99899998$. Encontre o erro relativo para os valores obtidos nos ítens (b) e (c).

6. Suponha que os pontos (x_0, y_0) e (x_1, y_1) estejam em uma linha reta com $y_0 \neq y_1$. Duas fórmulas estão disponíveis para encontrar o ponto onde esta linha corta o eixo dos x:

$$x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0} \qquad \text{e} \qquad x = x_0 - \frac{(x_1 - x_0) y_0}{y_1 - y_0}$$

   (a) Mostre que ambas as fórmulas estão algebricamente corretas

   (b) Use os dados $(x_0, y_0) = (1.31, 3.24)$ e $(x_1, y_1) = (1.93, 4, 76)$ e aritmética de arredondamento de 3 dígitos significativos para calcular o valor do ponto de intersecção com o eixo dos x por ambas as fórmulas. Qual o melhor método? explique o porque.

7. Utilize aritmética de 3 dígitos significativos, com arredondamento para computar a soma $\sum_{i=1}^{10}(1/i^3)$

   (a) Faça $\frac{1}{1} + \frac{1}{8} + ... + \frac{1}{1000}$

   (b) Faça $\frac{1}{1000} + ... + \frac{1}{8} + \frac{1}{1}$

   Existe diferença entre os dois resultados? Explique o que ocorre.

Problemas computacionais - Retirados de Numerical Analysis, Kincaid & Cheney

1. Escreva um programa para calcular

   - $$f(x) = \sqrt{x^2 + 1} - 1$$
   - e $$g(x) = \frac{x^2}{\sqrt{x^2 + 1} + 1}$$

   para uma sucessão de valores de x tais como $8^{=1}$, $8^{=2}$, $8^{=3}$,... . Apesar de $f(x) = g(x)$ o computador irá produzir diferentes resultados. Quais são os resultados confiáveis e porque?

2. Usando seu computador escreva o valor das funções

   - $f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1$

- $g(x) = (((((((x-8)x+28x)-56)x+70)x-56)x+28)x-8)x+1$
- $h(x) = (x-1)^8$

em 101 pontos igualmente espaçados do intervalo $[0.99, 1, 01]$, Calcule cada função de forma direta sem simplificação de fórmulas. Observe que as três funções são idênticas, e os resultados não, sendo alguns deles aparecem negativos apesar de pela definição de $h(x)$ sabermos que devem ser sempre positivos.

3. Um experimento numérico interessante é calcular o produto interno dos seguintes vetores:

$$x = [2.718281828; -3.141592654; 1.414213562; 0.5772156649; 0.3010299957]$$

$$y = [1486.2497; 878366.9879; -22.37492; 4773714.647; 0.000185049]$$

Calcule a soma de 4 formas diferentes:

(a) Ordem direta $\sum_{i=1}^{n} x_i y_i$,

(b) Ordem inversa $\sum_{i=n}^{1} x_i y_i$,

(c) Do maior para o menor (some primeiro os positivos do maior para o menor, depois some os negativos do menor para o maior, e então proceda a soma parcial)

(d) Menor para o maior (ordem inversa da soma no item anterior)

(e) Use precisão simples e dupla para proceder os cálculos dos ítens anteriores. Compares suas respostas com o resultado correto em sete casas decimais $-1.006571 \times 10{-}11$. Explique os resultados

4. Repita o problema anterior mas corte o final 9 em x_4 e o final 7 de x_5. Que efeito esta pequena troca provoca?

Capítulo 2

Zeros de Equações não Lineares

Neste capítulo vamos estudar alguns métodos para localizar raízes de equações. Este problema ocorre freqüentemente em computação científica. Para estimarmos intervalos onde existe uma raíz, muitas vezes usamos como auxílio os seguinte teorema:

Teorema 11 Teorema do Valor Intermediário para Funções Contínuas: *Se f é uma função contínua sobre um intervalo $[a, b]$, e se $f(a) < y < f(b)$, então $f(x) = y$ para algum $x \in (a, b)$.*

Deste teorema, podemos afirmar que se f é contínua sobre um intervalo $[a, b]$ e se $f(a)f(b) < 0$ então f deve assumir o valor 0 em algum $x \in (a, b)$, isto é que existe pelo menos um zero real neste intervalo. O primeiro e mais simples dos métodos que veremos é baseado neste princípio e é chamado de método da bisecção.

Vamos separar este capítulo em duas grandes subseções. Na primeira vamos estudar alguns métodos gerais para encontrar raízes de equações não lineares e uma segunda seção dedicada para as funções polinomiais.

2.1 Métodos Gerais

Nesta seção estudaremos métiodos para encontrar zeros de equações não lineares.

2.1.1 O Método da Bissecção

A técnica mais elementar de encontrar a raíz de $f(x) = 0$, onde f é uma função contínua, é o método da bissecção. Para darmos início primeiramente escolhemos um intervalo que contenha uma raíz do problema. Para isto escolhemos um intervalo $[a, b]$ de forma que $f(a)$ e $f(b)$ possuam sinais trocados, e assim pelo teorema do valor intermediário podemos garantir a existência de pelo menos uma raíz neste intervalo. Por simplicidade assumiremos que a raíz é única neste intervalo.

Para iniciarmos este método fazemos $a_1 = a$ e $b_1 = b$ e encontramos seu ponto médio,

$$p_1 = a_1 + \frac{b_1 - a_1}{2}.$$

Se $f(p_1) = 0$, então a raíz p procurada é dada por $p = p_1$. Se $f(p_1) \neq 0$, então o sinal de $f(p_1)$ é o mesmo de $f(a_1)$ ou de $f(b_1)$. Se $f(p_1)$ tem o sinal contrário de $f(a_1)$ então temos uma raíz do nosso problema no intervalo $[a_1, p_1]$ e neste caso fazemos

$$a_2 = a_1 \quad \text{e} \quad b_2 = p_1.$$

Caso contrário teremos que

$$a_2 = p_1 \quad \text{e} \quad b_2 = b_1.$$

O mesmo processo é agora aplicado no intervalo $[a_2, b_2]$, e encontramos assim o intervalo $[a_3, b_3]$, $[a_4, b_4]$. Devemos notar que cada intervalo $[a_i, b_i]$, $i = 1, 2, 3, ...$ contém a raíz p e que

$$|a_{i+1} - b_{i+1}| = \frac{|a_i - b_i|}{2}.$$

ALGORITMO 2 *Este algoritmo encontra solução para $f(x) = 0$, f contínua, definida em $[a, b]$ e $f(a)$ e $f(b)$ com sinais opostos*
ENTRADA a,b; Tolerância TOL; Número máximo de iterações N_0
SAIDA Solução aproximada p ou mensagem de falha

Passo 1 *Faça $i = 1$.*

Passo 2 *Enquanto $i \leq N_0$.*

Passo 3 *Faça $p = (a + b)/2$.*
Passo 4 *Se $f(p) = 0$ ou $(b - a)/2 < TOL$ então*
SAÍDA (p)
PARE
Passo 5 *$i = i + 1$*
Passo 6 *Se $f(a)f(b) > 0$*
então faça $a = p$
caso contrário $b = p$.

Passo 7 *SAÍDA ('Método falhou depois de N_0 iterações')*
STOP

OBS: Outros critérios de parada que podem ser usados no algoritmo da bissecção tais como,

$$|p_N - p_{N-1}| < \epsilon$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \epsilon$$

$$|f(p_N)| < \epsilon$$

Exemplo 5 *Vamos considerar o seguinte polinômio no intervalo I:*

- $f(x) = x^3 + 4x^2 - 10$
- *Consideremos o intervalo* $I = [1, 2]$
- $f(1) = -5$ *e* $f(2) = 14 \rightarrow$ *ao menos uma raíz no intervalo* $[1, 2]$.
- $f'(x) = 3x^2 + 8x > 0 \rightarrow$ *f(x) é crescente neste intervalo e a raíz é única.*
  - *Na primeira iteração consideramos* $a_1 = 1.0$, $b_1 = 2$ *logo a raíz é aproximada por* $p_1 = \frac{a_1+b_1}{2} = 2 + 12 = 1.5$. *O comprimento do intervalo é dado por* $\frac{b-a}{2} = 0.5$.
  - *Na segunda iteração faremos* $a_2 = 1.0$ *e* $b_2 = 1.5$ *a é raíz aproximada* $p_2 = 1.25$ *e o comprimento do intervalo é* 0.25.
  - *seguindo o mesmo raciocínio, temos os valores das 13 primeiras iterações dados na seguinte tabela:*

n	a_n	b_n	p_n	$f(p_n)$
1	1.0	2.0	1.5	2.375
2	1.0	1.5	1.25	-1.79687
3	1.25	1.5	1.375	0.16211
4	1.25	1.375	1.3125	-0.84839
5	1.3125	1.375	1.34375	-0.35098
6	1.34375	1.375	1.359375	-0.09641
7	1.359375	1.375	1.3671875	0.03236
8	1.359375	1.3671875	1.36328125	-0.03215
9	1.36328125	1.3671875	1.365324375	0.000072
10	1.36328125	1.365234375	1.364257813	-0.1605
11	1.364257813	1.362534375	1.364746094	-0.00799
12	1.364746094	1.362534375	1.364990235	-0.00396
13	1.364990235	1.365234375	1.365112305	-0.00194

  - *Depois de 13 iterações,* $p_{13} = 1.365112305$ *aproxima a raíz p com um erro*

$$|p - p_{13}| < \left|\frac{b_{13} - a_{13}}{2}\right| = |b_{14} - a_{14}| = |1.365234375 - 1.365112305| = 0.000122070,$$

como $|a_{14}| < |p|$,

$$\frac{|p - p_{13}|}{|p|} < \frac{|b_{14} - a_{14}|}{|a_{14}|} \leq 9.0 \times 10^{-5},$$

assim nossa aproximação possui pelo menos quatro dígitos significativos corretos.
OBS: O valor de p com 9 casas decimais é $p = 1.365230013$. *Note que neste caso* p_9 *está mais perto de p que a aproximação final* p_9, *mas não tem forma de saber isto sem conhecer a resposta correta!*

O método da bisecção, apesar de conceitualmente simples, possui inconvenientes significantes. Possui convergência lenta (isto é, N deve ser grande para que $|p - p_N|$ seja suficientemente pequeno), além de poder descartar uma boa aproximação intermediária. Entretanto este método possui uma propriedade importante, ele sempre converge para uma raíz e por isto é usado na prática como "chute inicial"para métodos mais eficientes.

Teorema 12 *Seja $f \in \mathcal{C}[a,b]$ e suponha que $f(a).f(b) < 0$. O método da bissecção gera uma seqüência $\{p_n\}$ aproximando p com a propriedade*

$$|p_n - p| \leq \frac{b-a}{2^n}, \quad n \geq 1. \tag{2.1}$$

Prova: Para cada $n \geq 1$, temos

$$b_n - a_n = \frac{1}{2^{n-1}}(b-a) \quad e \quad p \in (a_n, b_n).$$

Como $p_n = \frac{1}{2}(a_n + b_n)$, para todo $n \geq 1$, segue que

$$|p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}, \quad n \geq 1.$$

De acordo com definição anterior, a inequação (2.1) implica que $\{p_n\}_{n=1}^{\infty}$ converge para p e é limitada por uma seqüência que converge para zero com taxa de convergência $O(2^{-n})$.

OBS: Teoremas deste tipo apenas nos dão limites de erro na aproximação. Por exemplo no exemplo 1 o teorema assegura que:

$$|p - p_9| \leq \frac{2-1}{2^9} \approx 2 \times 10^{-3}$$

O erro real é muito menor,

$$|p - p_9| = |1.365230013 - 1.365234375| \approx 4.4 \times 10^{-6}$$

Exemplo 6 *Para determinar o número de iterções necessárias para resolver $f(x) = x^3 + 4x^2 - 10 = 0$ com precisão de $\epsilon = 10^{-5}$ usando $a_1 = 1$ e $b_1 = 2$ requer encontrar um inteiro N que satisfaça*

$$|p_n - p| \leq 2^{-N}(b-a) = 2^{-N} < 10^{-5}.$$

Para determinarmos N vamos usar logaritmo na base 10. Daí,

$$-N \log_{10} 2 < -5 \quad ou \quad N > \frac{5}{log_{10} 2} \approx 16.6.$$

Desta forma precisaríamos teoricamente de 17 iterações para termos uma aproximação para p com precisão de 10^{-5}. Devemos lembrar que estas técnicas apenas nos dão um limite do erro, em muitos casos este limite é muito maior que o erro real.

Ordem de convergência do Método da Bissecção

Seja p uma raíz de $f(x) = 0$. Neste caso, $e_i = |x_i - p|$. Agora, como no método da bissecção a aproximação de ordem $i + 1$ é dada por, $x_{i+1} = \frac{x_i + x_{i-1}}{2}$ onde $f(x_i)f(x_{i-1}) < 0$ então sabemos que o erro da iteração $i + 1$ é a metade do erro da iésima iteração, assim

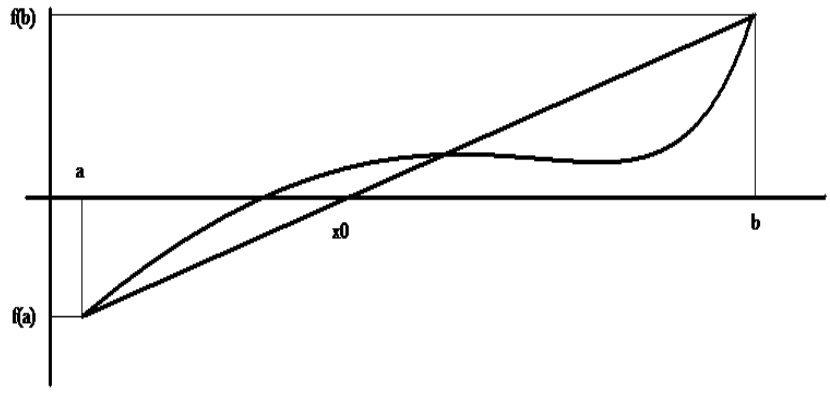
$$e_{i+1} \leq \frac{e_i}{2} \longrightarrow \frac{e_{i+1}}{e_i} \leq \frac{1}{2} \tag{2.2}$$

logo a ordem de convergência do método da bissecção é **linear**.

Observações: Uma vantagem do Método da bissecção é que sempre converge, porém sua convergência é lenta, conforme mostramos ela é linear. Podemos ainda observar que a sequência gerada por este método não decresce de forma monótona, devido a escolha da aproximação da raíz sempre como o ponto médio do intervalo considerado, não levando em conta a função considerada, mas apenas seu sinal nos extremos dos intervalos. Normalmente usamos o método da bissecção para reduzir o intervalo considerado antes de usar um outro método com maior ordem de convergência.

2.1.2 O método da Falsa-Posição

Vamos novamente considerar uma função f contínua em um intervalo [a; b] de forma que $f(a)f(b) < 0$. A idéia central deste método é usar como aproximação da raíz procurada, uma média ponderada pelos valores da função f nos extremos do intervalo a cada iteração.



Para isto, vamos considerar a sequência formada pelas abscissas dos pontos de intersecção da reta formada pelos pontos $(a_n, f(a_n))$ e $(b_n, f(b_n))$ e o eixo dos x. Com esta finalidade

primeiramente vamos encontrar a reta $y(x)$ que passa pelos pontos $(a, f(a))$ e $(b, f(b))$:

$$y(x) = \frac{x-a}{b-a}f(b) + \frac{x-b}{a-b}f(a)$$

$$y(x) = \frac{xf(b) - af(b) - xf(a) + bf(a)}{b-a} \tag{2.3}$$

$$y(x) = \frac{x(f(b) - f(a)) - (af(b) - bf(a))}{b-a}$$

Assim a abscissa x do ponto de intersecção desta reta com o eixo dos x, isto é quando $y(x) = 0$ pode ser escrita como:

$$0 = \frac{x(f(b) - f(a)) - (af(b) - bf(a))}{b-a} \tag{2.4}$$

como $a \neq b$

$$x = \frac{af(b) - bf(a)}{f(b) - f(a)} \tag{2.5}$$

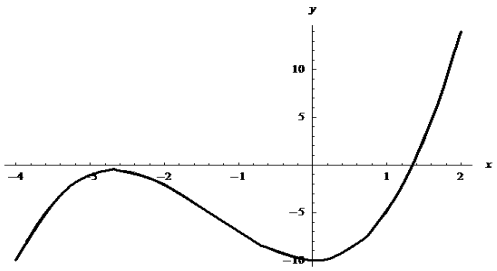
Desta forma o método da falsa posição é dado por:

- Função f definida e contínua em um intervalo $[a, b]$.
- Se $f(a)f(b) < 0$ então existe um número ξ neste intervalo de forma que $f(\xi) = 0$, logo
- Fazemos $a_0 = a$ e $b_0 = b$ e calculamos

$$p_0 = \frac{(a_0 f(b_0) - b_0 f(a_0))}{(f(b_0) - f(a_0))}$$

- Para $n = 0, 1, ...$:
- Se $f(p_n) = 0$ a raíz procurada foi encontrada, caso contrário
- Se $f(a_n)f(p_n) < 0$ fazemos: $a_{n+1} = a_n$ e $b_{n+1} = p_n$ ou se $f(a_n)f(p_n) > 0$ $a_{n+1} = p_n$ e $b_{n+1} = b_n$
- $p_{n+1} = \frac{(a_{n+1} f(b_{n+1}) - b_{n+1} f(a_{n+1}))}{(f(b_{n+1}) - f(a_{n+1}))}$
- A sequência $p_0, p_1, ...$ produzida pelo método da falsa posição converge para a raiz da função neste intervalo.

Exemplo 7 *Vamos considerar o polinômio* $p(x) = x^3 + 4x^2 - 10$. *Para esboçar o gráfico desta função notamos que:* $f'(x) = 3x^2 + 8x = (3x + 8)x$, *assim os possíveis pontos de máximo e mínimo desta função são aqueles onde* $f'(x) = 0 \to x = 0$ *ou* $x = -8/3$. *Temos ainda que* $f(0) = -10$ *e* $f[-8/3] \approx -0.5185$. *Ainda temos que* $f(x) < 0$ *se* $x \in (-8/3, 0)$ *e fora deste intervalo a função é negativa, logo seu gráfico é:*



Podemos ver que este polinômio possui uma única raíz e ela está no intervalo $[1, 2]$. *assim vamos começar fazendo* $a_1 = 1$ *e* $b_1 = 2$ *e desta forma obtemos pelo método da falsa posição:*

Se escolhemos $a_0 = 1$ *e* $b_0 = 2$ *então*

$$p_0 = \frac{1f(2) - 2f(1)}{f(2) - f(1)} = \frac{14 + 10}{-19} = -1.2631578947368$$

e desta forma:

k	a_k	b_k	p_k	$f(p_k)$
0	1.0000000000	2	1.2631578947	-1.6022743840
1	1.26315789474	2	1.338827839	-0.4303647480
2	1.3388278388	2	1.3585463418	-0.1100087885
3	1.3585463418	2	1.3648070318	-0.0277620910
4	1.3635474400	2	1.3648070318	-0.0069834154
5	1.3648070318	2	1.3651237179	-0.0017552090
6	1.3651237179	2	1.3652033037	-0.0004410630
7	1.3652033036	2	1.3652233020	-0.0001108281
8	1.3652233020	2	1.3652283270	-0.0000278479
9	1.3652283270	2	1.3652295897	-6.997390 $\times 10^{-6}$
10	1.3652295897	2	1.3652299069	-1.758239 $\times 10^{-6}$

Neste exemplo $f(p_k)$ *é sempre positivo e por este motivo os elementos* b_k, $k = 0, 1, .., 10$ *ficaram inalterados. Olhando a tabela acima, notamos que depois de* 10 *iterações o intervalo foi reduzido para*

$$[1.365229589673847, 2.000000000000].$$

Desta forma como garantimos que a raíz procurada está neste intervalo, que tem comprimento $b_{10} - a_{10} = 0.634770410$ *e é aproximada por*

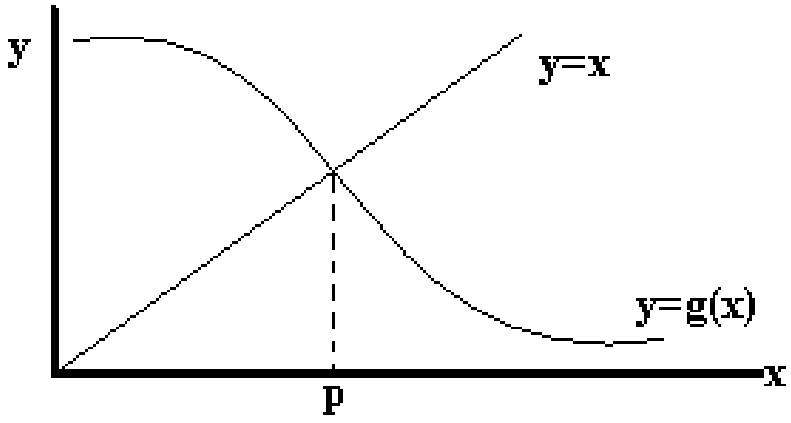
$$p_{10} = 1.365229907$$

. *Mais ainda, podemos estimar sua precisão como a sua distância até a extremidade mais próxima do intervalo* $[a_{10}, b_{10}]$*, isto é*

$$p_{10} - a_{10} = 3.172667253359407 \times 10^{-7}.$$

2.1.3 O método da Iteração Linear (Ponto Fixo)

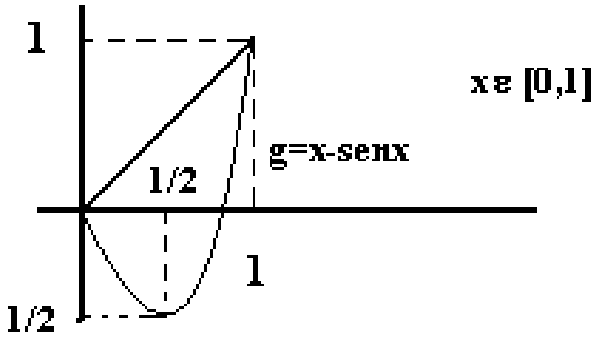
Vamos considerar uma função $f(x)$, contínua em $[a, b]$ onde existe pelo menos uma raíz de $f(x)$. O método do ponto fixo transforma um problema do tipo $f(x) = 0$ em um problema do tipo $g(x) = x$. Neste caso chamamos a função $g(x)$ de função iteração e a solução de um problema deste tipo nós chamamos de um ponto fixo para a função g.



- Quando podemos garantir a existência de ponto fixo?

  Exemplo 8

  - *Se* $g(x) = x$ *então cada* $x \in [0, 1]$ *é um ponto fixo para* g*. Neste caso temos infinitos pontos fixos.*
  - *Se* $g(x) = x - \sin \pi x$ *no intervalo* $[0, 1]$ *temos exatamente dois pontos fixos,* $x = 0$ *e* $x = 1$*.*



Condições suficientes de existência e unicidade de ponto fixo para g em um intervalo $[a, b]$ são dadas por:

Teorema 13 *Se $g(x) \in \mathcal{C}[a, b]$ e $g(x) \in [a, b]$ para todo $x \in [a, b]$, então g possui um ponto fixo em $[a, b]$.*

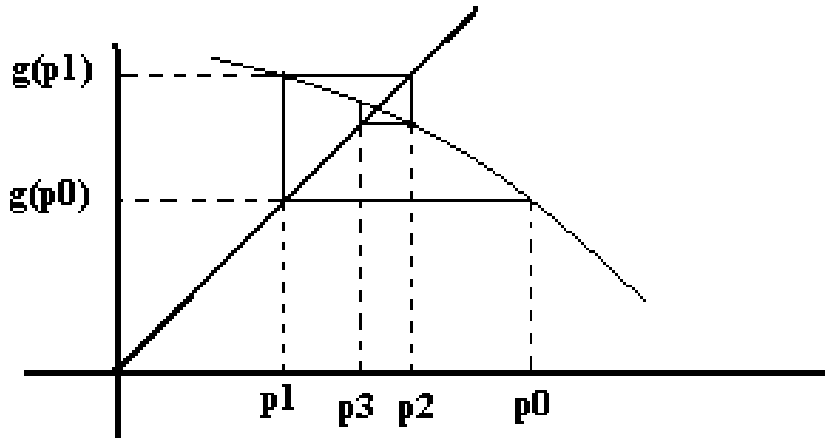
Mais, se $g'(x)$ existe em (a, b) e se existe uma constante positiva $k < 1$ de forma que

$$|g'(x)| \leq K < 1 \text{ para todo } x \in (a, b)$$

então g possui exatamente um ponto fixo em $[a, b]$

A prova deste teorema é bastante simples e pode ser encontrada em Burden and Faires "Numerical Analysis".

- Como aproximar este ponto fixo?
  - Primeiramente escolhemos um "chute inicial p_0 em $[a, b]$
  - Geramos uma seqüência $\{p_n\}_{n=0}^{\infty}$ fazendo
    - $p_1 = g(p_0)$
    - $p_2 = g(p_1)$
    - ...
    - $p_{n+1} = g(p_n)$ ...



– Se esta seqüência gerada for convergente e a função g contínua então,

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p)$$

Assim p é a solução de $x = g(x)$. Aqui devemos observar que a convergência da seqüência gerada bem como sua ordem de convergência dependem fortemente da escolha de $g(x)$.

ALGORITMO 3 *Ponto Fixo*
Encontra a solução para $p = g(p)$ com chute inicial p_0.
***Entrada**: p_0, TOL, N_0 número máximo de iterações*
***Saída**: Solução aproximada ou mensagem de erro*

1. *Faça $i = 1$*
2. *Enquanto $i \leq N_0$*
   (a) *Faça $p = g(p_0)$ (calcula p_i)*
   (b) *Se $|p - p_0| < TOL$ então*
      i. *Saída p*
      ii. ***PARE***
   (c) *Faça $i = i + 1$*
   (d) *Faça $p_0 = p$ (recobre p_0)*
3. *Saída: "Método falhou depois de N_0 iterações"*
4. ***PARE***

Vamos agora mostrar um exemplo, onde a escolha de $g(x)$ influencia na convergência e na ordem de convergência do método do ponto fixo.

n	(a)	(b)	(c)	(d)	(e)
0	1.5	1.5	1.5	1.5	1.5
1	-0.875	0.8165	1.286953768	1.348399725	1.373333333
2	6.732	2.9969	1.402540804	1.367376372	1.365262015
3	-469.7	$\sqrt{-8.65}$	1.345458374	1.364957015	1.365230014
4	1.03×10^8		1.375170253	1.365264748	1.365230013
5			1.360094193	1.365225594	
6			1.367846968	1.365230576	
7			1.363887004	1.365229942	
8			1.365916734	1.365230022	
9			1.364878217	1.365230012	
10			1.365410062	1.365230014	
15			1.365223680	1.365230013	
20			1.365230236		
25			1.365230006		
30			1.365230013		

Exemplo 9 *1. Encontre o único zero de* $f(x) = x^3 + 4x^2 - 10$ *no intervalo* $[1, 2]$.

Para usarmos o método do ponto fixo, devemos modificar o problema $f(x) = 0$ *em um problema do tipo* $g(x) = x$. *Existem muitas formas para escolher* $g(x)$, *aqui faremos cinco escolhas distintas, duas delas teremos falta de convergência e em três delas teremos oredens de convergência distintas. Nos exemplos usaremos como chute inicial* $p_0 = 1.5$.

(a) $x = x - x^3 - 4x^2 + 10 = g_1(x)$

(b) $x = \left(\frac{10-4x^2}{x}\right)^{1/2} = g_2(x)$

(c) $x = \frac{1}{2}(10 - x^3)^{1/2} = g_3(x)$

(d) $x = \left(\frac{10}{4+x}\right)^{1/2} = g_4(x)$

(e) $x = x - \frac{x^3+4x^2-10}{3x^2+8x} = g_5(x)$

Os resultados obtidos usando como função iteração as $g_i(x)$ *definidas acima são dados pela tabela 1.*

Como podemos garantir que a função escolhida $g(x)$ produza convergência para a solução de $g(x) = x$?

Esta pergunta pode ser respondida através dos seguintes teorema e corolários. Aqui apenas enunciaremos os teoremas, cujas demonstrações podem ser encontradas em livros de análise numérica, particularmente Burden and Faires, Numerical Analysis.

Teorema 14 (Ponto Fixo) *Seja* $g \in \mathcal{C}[a, b]$ *e que* $g(x) \in [a, b]$ *para todo* $x \in [a, b]$. *Suponha que* g' *existe em* (a, b) *com* $|g'(x)| \leq k < 1$ *para todo* $x \in (a, b)$.

Se p_0 é um número no intervalo $[a, b]$ então a seqüência gerado por $p_n = g(p_{n-1}$ para $n \geq 1$ converge para o único ponto fixo $p \in [a, b]$

Prova:

Pelo teorema 13 sabemos que existe um único ponto fixo $p \in [a, b]$. Uma vez que $g : [a, b] \to [a, b]$ a seqüência $\{p_n\}_{n=0}^{\infty}$ está definida para todo $n \geq 0$ e $p_n \in [a, b]$ assim,

$$|p_n - p| = |g(p_{n-1} - p| \leq k^2|P_{n-2} - p| \leq ... \leq k^n|P_0 - p|$$

como $k < 1$ temos que

$$\lim_{n\to\infty} |p_n - p| \leq lim_{n\to\infty} k^n|p_0 - p| = 0$$

Assim, $lim_{n\to\infty} p_n = p$.

Corolário 1 *Se g satisfaz as hipóteses do teorema 14 (Ponto Fixo) um limite para o erro envolvido usando-se a seqüência p_n para aproximar p é dada por:*

$$|p_n - p| \leq k^n|p_0 - p| \leq k^n \max\{p_0 - a, b - p_0\}$$

uma vez que $p \in [a, b]$.

Corolário 2 *Se g satisfaz a hipótese do teorema 14 (Ponto Fixo) então*

$$|p_n - p| \leq \frac{k^n}{1-k}|p_0 - p_1| \qquad \forall n \geq 1$$

No nosso exemplo

$$g_1(x) = x - x^3 - 4x^2 + 10$$
$$g_1'(x) = 1 - 3x^2 - 8x$$

Podemos ver que $g'(1.5) = -17.75$, assim $|g_1'(x)| > 1$ no interval e desta forma não podemos garantir a convergência (nem a não convergência) e assim esta seria uma má escolha para função iteração.

$$g_2(x) = \left(\frac{10}{x} - 4x\right)^{1/2}$$
$$g_2(1) = (10 - 4)^{1/2} = 2.44 \notin [1, 2]$$

Isto é $g_2[1, 2] \not\to [1, 2]$ e assim também não podemos garantir convergência para esta escolha de função iteração.

$$g_3(x) = \frac{1}{2}(10 - x^3)^{1/2}$$
$$g_3'(x) = -\frac{3}{4}x^2(10 - x^3)^{-1/2}$$

Neste caso vemos que $|g_3'(2)| \approx 2.12 > 1$. E assim esta também não é uma boa escolha para função iteração. (note que existe convergência mas é muito lenta).

Os casos g_4 e g_5 satisfazem o teorema.

A última escolha para $g(x)$ como pode ser mostrado é a melhor escolha possível para função iteração e esta escolha é na verdade chamada de método de Newton-Raphson que será visto na próxima seção.

2.1.4 O Método de Newton-Raphson

O método de Newton-Raphson é um dos mais poderosos para encontrar um zero de $f(x) = 0$. Sua dedução pode ser feita de forma gráfica ou como a escolha da função iteração de forma que o método do ponto fixo seja o mais rápido possível ou ainda usando-se expansão de Taylor. Aqui usaremos a expansão em Taylor.

Assim vamos supor que $f \in \mathcal{C}^2[a, b]$ e que queremos encontrar p de forma que $f(p) = 0$. Seja $\bar{x}$ uma aproximação de p de forma que $f(\bar{x}) \neq 0$ e que $|p - \bar{x}|$ seja pequeno. Assim expandindo $f(x)$ em torno de $\bar{x}$ temos que existe $\xi(x)$ entre x e $\bar{x}$ de forma que:

$$f(x) = f(\bar{x}) + (x - \bar{x})f'(\bar{x}) + \frac{(x - \bar{x})^2}{2!}f''(\xi(x)) \tag{2.6}$$

agora fazendo $x = p$ em (2.6) e lembrando que $f(p) = 0$,

$$0 = f(\bar{x}) + (p - \bar{x})f'(\bar{x}) + \frac{(p - \bar{x})^2}{2!}f''(\xi(p)) \tag{2.7}$$

como $|p - \bar{x}|$ é pequeno podemos desprezar $(p - \bar{x})^2$, assim

$$0 \approx f(\bar{x}) + (p - \bar{x})f'(\bar{x}) \tag{2.8}$$

ou

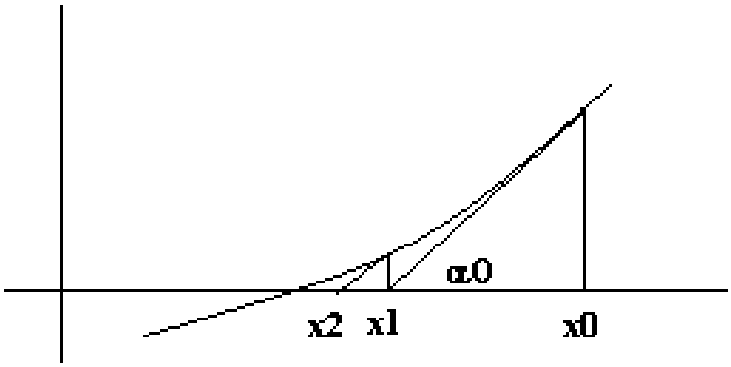
$$p = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})} \tag{2.9}$$

Da equação(2.9) parte o método de Newton-Raphson que começa com uma aproximação inicial p_0 e gera uma seqüência $\{p_n\}$, definida por:

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})} \tag{2.10}$$

Este procedimento iterativo irá convergir para o zero de muitas funções, e quando converge, será usualmente muito rápido.

Esta mesma fórmula poderia ter sido obtida de forma geométrica, vamos considerar o gráfico da figura abaixo e um chute inicial p_0. Se traçarmos a tangente a curva em p_0 sua intersecção com o eixo dos x será o segundo termo da seqüência e assim por diante. Isto é:



$$\begin{aligned}
\tan\alpha_0 =& f'(p_0) = \frac{f(p_0)}{p_0 - p_1)} \to p_1 = x_0 - \frac{f(p_0)}{f'(p_0)} \\
\tan\alpha_1 =& f'(p_1) = \frac{f(p_1)}{p_1 - p_2)} \to p_2 = x_1 - \frac{f(p_1)}{f'(p_1)} \\
&\vdots \\
\tan\alpha_{n+1} =& f'(p_0) = \frac{f(p_0)}{(p_0 - p_1)} \to p_1 = x_0 - \frac{f(p_0)}{f'(p_0)}
\end{aligned} \tag{2.11}$$

Exemplo 10 *Encontre $\sqrt{7}$ usando o método de Newton-Raphson, com 9 casas decimais. Vamos escolher $f(x) = x^2 - 7$ note que sua raíz é $\sqrt{7}$. Assim vamos endontrar a raíz de $f(x)$ pelo método pedido. Neste caso, $f'(x) = 2x$ e a fórmula de recorrência é dada por:*

$$p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} = p_n - \frac{p_n^2 - 7}{2x_n}$$

Para chute inicial vamos escolher $p_0 = 2.5$ pois $\sqrt{4} = 2$ e $\sqrt{9} = 3$,

$$\begin{aligned}
x_1 &= 2.5 - \frac{-0.75}{5} = 2.65 \\
x_2 &= 2.65 - \frac{0.0225}{5.3} = 2.645754717 \\
x_3 &= 2.645754717 - \frac{0.00001802242791}{5.291509434} = 2.645751311 \\
x_4 &= 2.645751311 - \frac{-3 \times 10^{-10}}{5.291502622} = 2.645751311
\end{aligned}$$

ALGORITMO 4 *: Newton Raphson*

Este algoritmo encontra a solução de $f(x) = 0$ *dado um "chute inicial"* p_0
ENTRADA: p_0*, Tolerância TOL, Número máximo de iterações* N_0
SAÍDA: Solução aproximada ou mensagem de erro

1. *Faça* $i = 1$
2. *Enquanto* $i \leq N_0$
   (a) $p = p_0 - f(p_0)/f'(p_0)$
   (b) *Se* $|p - p_0| <$ *TOL então*
      - *Saída* p
      - *PARE*
   (c) *Faça* $i = i + 1$
   (d) *Faça* $p_0 = p$
3. *Saída: O método falhou depois de* N_0 *iterações*
4. *Pare*

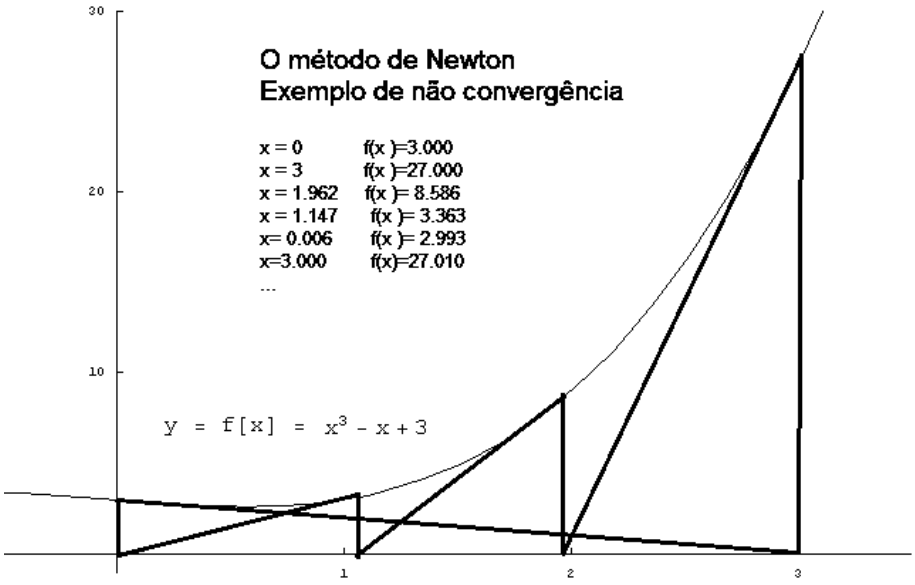
Observação: Podemos adotar outras técnicas de parada além da adotada pelo algoritmo, tais como

$$\frac{|p_n - p_{n-1}|}{|p_n|} < \epsilon \text{ se } p_n \neq 0$$

ou

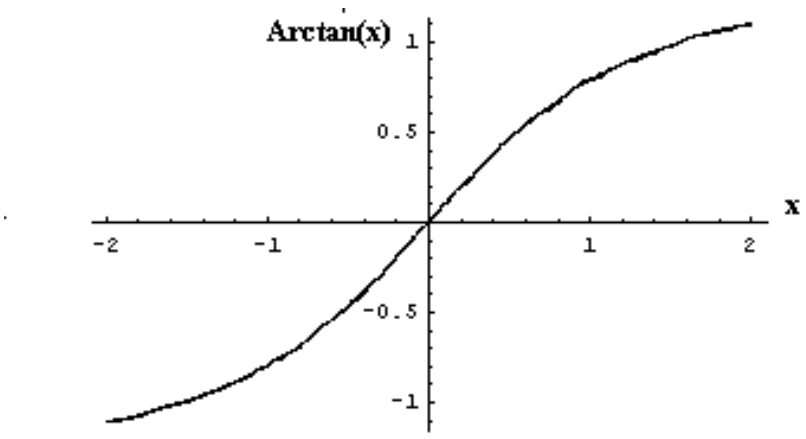
$$f(p_n) < \epsilon$$

Apesar de sua rápida convergência, o método de Newton-Raphson possui dificuldades com certos tipos de funções. Considere por exemplo uma função polinomial como a mostrada na figura abaixo, quando partimos de $x_0 = 0$, ficamos com uma não convergência oscilatória!



Ela ilustra por um caso não convergente. Outro exemplo seria um caso onde o chute é dado próximo de uma raíz mas outra raíz é aproximada e não a raíz esperada. Não existe um método simples que evite este tipo de problema em certas funções, mas um rascunho de seu gráfico ou sua tabulação em alguns pontos serão muitas vezes precisos para uma escolha certa do chute inicial. Também é problema o caso de raízes múltiplas. Neste caso o gráfico tangencia o eixo dos x e neste caso a derivada de primeira ordem na raíz é nula. Mesmo neste caso, o método de Newton pode ser convergente (dependendo do chute inicial), mas, neste caso, a convergência é muito lenta e na prática pode fazer o cálculo difícil (ela que era de ordem quadrática cai para ordem linear).

Exemplo 11 *Vamos considerar* $f(x) = \arctan x$. *Assim* $f'(x) = \frac{1}{1+x^2}$



1. *Primeiramente vamos usar* $x_0 = 1.4$ *Neste caso temos:*

$$x_1 = x_0 - \arctan(x_0)(x_0^2+1) = 1.4 - \arctan(1.4)((1.4)^2+1) = -1.41366$$
$$x_2 = x_1 - \arctan(x_1)(x_1^2+1) = 1.45013$$
$$x_3 = x_2 - \arctan(x_2)(x_2^2+1) = -1.55063$$
$$x_4 = 1.84705$$
$$x_5 = -2.89356$$
$$x_6 = 8.71033$$
$$x_7 = -103.25$$

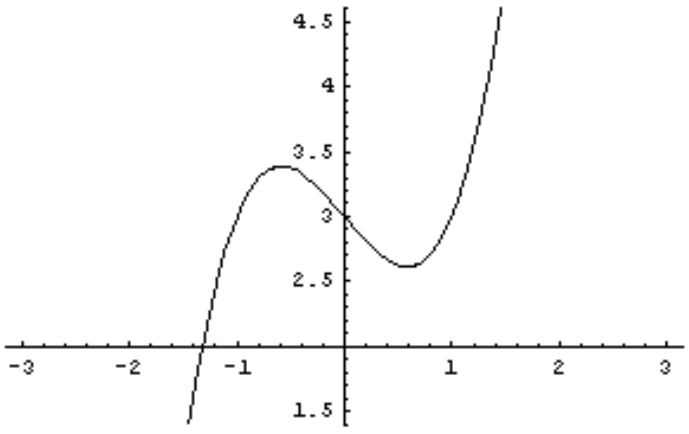
Isto é o método de Newton diverge para este valor inicial.

2. *Agora, vamos usar* $x_0 = 1.35$ *Neste caso temos:*

$$x_1 = x_0 - \arctan(x_0)(x_0^2+1) = 1.35 - \arctan(1.35)((1.35)^2+1) = -1.28409$$
$$x_2 = x_1 - \arctan(x_1)(x_1^2+1) = 1.12412$$
$$x_3 = x_2 - \arctan(x_2)(x_2^2+1) = -0.785872$$
$$x_4 = 0.291554$$
$$x_5 = -0.016251$$
$$x_6 = 2.86105 \times 10^{-6}$$
$$x_6 = -1.5 \times 10^{-17}$$

Isto é o método de Newton converge para $x = 0$ *quando o valor inicial considerado é* 1.35.

Exemplo 12 *Vamos considerar* $f(x) = x^3 - x + 3$. *Assim* $f'(x) = 3x^2 - 1$, *assim* $f'(x) = 0 \leftrightarrow x = \pm\frac{\sqrt{3}}{3} \approx \pm 0.57735$



Vamos fazer $x_0 = 0$.

$x_1 = x_0 - \frac{x_0^3 - x_0 - 3}{3x-1} = 3.000000$
$x_2 = -0.375000$
$x_3 = 1.188420$
$x_4 = -0.172081$
$x_5 = 1.91663$
$x_6 = 0.206263$
$x_7 = 7.55789$
$x_8 = -12.1509$
$x_9 = -59.647$
$x_{10} = -1238.63$

Isto é o método diverge com este "chute"inicial! Confira o que ocorre se $x_0 = -2$

Ordem de Convergência do Método de Newton

Teorema 15 *Se* $f(p) = 0$ *com* $f'(p)$ *e se* $f''(x)$ *contínuas, então existe um intervalo aberto* I *contendo* p *de forma que se* $p_0 \in I$ *então para o método de Newton-Raphson temos que*

$$\lim_{i\to\infty} p_i = p$$

Isto é o teorema garante que o método de Newton converge, dependendo do chute inicial, para a raíz simples p.
Podemos mostrar que neste caso a ordem de convergência do método é quadrática, pois:

Se p é uma raíz simples para a função $f(x)$ então fazendo $x = p$ na série de Taylor truncada no segundo termo temos

$$0 = f(x_i) + (p - x_i)f'(x_i) + \frac{1}{2}(p - x_i)^2 f''(\xi) \quad \xi \text{ entre } x_i \text{ e } p,$$

agora, dividindo a equação acima por $f'(x_i)$

$$\frac{f(x_i)}{f'(x_i)} + (p - x_i) = -\frac{1}{2}(p - x_i)^2 \frac{f''(\xi)}{f'(x_i)}$$

Reordenando a equação acima temos

$$p - \underbrace{\left(x_i - \frac{f(x_i)}{f'(x_i)}\right)}_{x_{i+1}} = -\frac{1}{2}(p - x_i)^2 \frac{f''(\xi)}{f'(x_i)},$$

aplicando módulo dos dois lados da igualdade acima e lembrando que $e_i = |x_i - p|$ podemos escrever

$$e_{i+1} = e_i^2 \left| \frac{f''(\xi)}{2f'(x_i)} \right|$$

ou

$$\frac{e_{i+1}}{e_i^2} = \frac{1}{2} \left| \frac{f''(\xi)}{f'(x_i)} \right|$$

quando $i \to \infty$ temos que $x_i \to p$ logo $\xi \to p$. Desta forma aplicando o limite na equação anterior temos

$$\lim_{i \to \infty} \frac{e_{i+1}}{e_i^2} = \frac{1}{2} \left| \frac{f''(p)}{f'(p)} \right|$$

Note que o segundo membro da equação anterior é uma constante e assim temos pela definição de ordem de convergência que a ordem de convergência do método de Newton-Raphson é quadrática (2).

OBS: O método pode ser aplicado para encontrar raízes complexas, bastando para isto termos um chute inicial complexo.

Exemplo 13 *Vamos considerar a equação polinomial*

$$f(x) = x^4 - 9x^3 + 30x^2 - 42x + 20 \tag{2.12}$$

cujas raízes são $1, 2, 3 + i, 3 - i$
usando o método de Newton Raphson com diversos chutes iniciais obtemos:

0	*0*	*3*	*1+i*	*4+2i*
1	*0.476190476*	*2.333333333*	*1.270270270+0.621621621 i*	*3.587068332+1.540044085 i*
2	*0.783857793*	*1.979591837*	*1.448937409+0.253119829 i*	*3.294613943+1.231633775 i*
3	*0.945183080*	*2.000008766*	*1.659616171+0.244371599 i*	*3.107145507+1.059849539 i*
4	*0.995239348*	*2.000000000*	*1.911995657+0.066770499 i*	*3.018207673+1.003079473 i*
5	*0.999959683*		*1.999283337+0.001360480 i*	*3.000419751+0.999731774 i*
6	*0.999999997*		*1.999999996+4.297033794* 10^{-10} *i*	*2.999999824+0.999999672 i*
7	*1.000000000*		*2.000000000+3.205324874* 10^{-24}	*3.000000000+1.000000000 i*

2.1.5 O Método de Newton Modificado

Se a raíz de $f(x)$ é simples apresenta o método de Newton apresenta uma convergência quadrática. Esta convergência pode cair para ordem linear no caso de raízes múltiplas.

Exemplo 14 *Por exemplo, vamos considerar a equação:*

$$g(x) = x^4 - 8.6x^3 - 35.51x^2 + 464.4x - 998.46 \tag{2.13}$$

cujas raízes são $x_1 = 4.3, x_2 = 4.3, x_3 = 7.34847, x_4 = -7.34847$. *Usando o método de Newton Raphson temos:*

0	*2*	*5*	*7*	*10*	*-8*
1	*3.053113553*	*4.601960784*	*7.485611511*	*8.727184466*	*-7.432656827*
2	*3.627606412*	*4.444305608*	*7.360406997*	*7.920259886*	*-7.350123736*
3	*3.944855635*	*4.370775870*	*7.348571171*	*7.497754133*	*-7.348469884*
4	*4.116183264*	*4.335071426*	*7.348469236*	*7.362461713*	*-7.348469228*
5	*4.206252614*	*4.317459659*	*7.348469228*	*7.348608992*	
6	*4.252622694*	*4.308711179*		*7.348469242*	
7	*4.276179199*	*4.304350971*			
8	*4.288055726*	*4.302174336*			
9	*4.29401929*	*4.301086882*			
10	*4.29700749*	*4.300543369*			
11	*4.29850320*	*4.300271667*			
12	*4.29925147*	*4.300135829*			
13	*4.29962570*	*4.300067913*			
14	*4.29981284*				
15	*4.29990642*				
16	*4.29995321*				
17	*4.29997660*				
18	*4.29998830*				
19	*4.29999415*				
20	*4.29999708*				
21	*4.29999854*				
22	*4.29999927*				

Note que para encontrarmos a raíz 4.3 *precisamos de muitas iterações, pois ela é dupla. A convergência do método neste caso passa a ser linear.*

Para aumentarmos a ordem de convergência, no caso de raízes múltiplas procedemos da seguinte forma. Se a função $f(x)$ possui raíz múltipla em p então

$$\mu(x) = \frac{f(x)}{f'(x)}$$

possui raíz simples em p. A idéia é usar o Método de Newton Raphson para encontrar a raíz de $\mu(x)$.

$$\frac{\mu(x)}{\mu'(x)} = \frac{f(x)/f'(x)}{[(f'(x))^2 - f(x)f''(x)]/[f'(x)]^2} \tag{2.14}$$

Assim,

$$x_{n+1} = x_n - \frac{f(x)f'(x)}{[(f'(x))^2 - f(x)f''(x)]} \tag{2.15}$$

Este método teoricamente possui convergência quadrática como o método de NRaphson para raízes simples mas pode apresentar muitos problemas de erro de arredondamento, além de exigir o conhecimento de $f''(x)$.
A dificuldade deste método é que além de precisarmos conhecer a derivada primeira da função $f(x)$, nós também precisamos conhecer $f'(x)$ e $f''(x)$. Vamos usar este método para calcular a raíz 4.3 do exemplo anterior.

Exemplo 15 *Vamos considerar a função*

$$g(x) = x^4 - 8.6x^3 - 35.51x^2 + 464.4x - 998.46$$

cuja derivada primeira é dada por,

$$g'(x) = 464.4 - 71.02x - 25.8x^2 + 4x^3$$

e a derivada segunda por,

$$g''(x) = 12x^2 - 51.6x - 71.02$$

Vamos definir

$$h(x) = x - \frac{g(x)g'(x)}{(g'(x))^2 - g(x)g''(x)}$$

os resultados obtidos são dados por:

0	*2*	*5*
1	*4.237051182*	*4.411567448*
2	*4.30451852*	*4.301674139*
3	*4.300000025*	*4.300000340*
4		*4.299999982*

Vemos que neste caso a convergência é rápida isto é volta a ser de ordem quadrática na raíz dupla.

2.1.6 O Método da Secante

O método de Newton é muito poderoso, mas algumas vezes nós não conhecemos ou é muito difícil encontrar $f'(x)$. Por exemplo se $f(x) = x^2 3^x \cos(2x)$ então sua derivada será $f'(x) = 2x3^x \cos(2x) + x^2 3x(\cos 2x) \ln 3 - 2x^2 3^x \sin 2x$, que não é simples para avaliar em cada ponto. Para evitarmos avaliar o valor da derivada da função $f(x)$ no método de Newton podemos aproximar a derivada. Para isto vamos lembrar que

$$f'(x_{n-1}) = \lim_{x \to x_{n-1}} \frac{f(x) - f(x_{n-1})}{x - x_{n-1}} \tag{2.16}$$

agora fazendo $x = x_{n-2}$ podemos escrever que

$$f'(x_{n-1}) \approx \frac{f(x_{n-2}) - f(x_{n-1})}{x_{n-2} - x_{n-1}} = \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}} \tag{2.17}$$

Usando esta aproximação para $f'(p_{n-1})$ na fórmula de Newton,

$$x_n = x_{n-1} - \frac{f(x_{n-1})(x_{n-1} - x_{n-2})}{f(x_{n-1}) - f(x_{n-2})} \tag{2.18}$$

Observe que este método necessita de dois chutes iniciais x_0 e x_1.

Exemplo 16 *Encontre um zero de* $f(x) = \cos x - x$*, usando o método da secante. Nós agora precisamos de duas condições iniciais. Esboçando os gráficos de* $\cos x$ *e de* x *vemos que estas curvas se cruzam perto de* $\frac{\pi}{4}$*. Assim vamos usar como partida* $x_0 = 0.5$ *e* $x_1 = 0.7853981635$*. A fórmula usada é:*

$$x_n = x_{n-1} - \frac{(x_{n-1} - x_{n-2})(\cos x_{n-1} - x_{n-1})}{(\cos x_{n-1} - x_{n-1}) - (\cos x_{n-2} - x_{n-2})}$$

n	p_n
0	*0.5*
1	*0.7853981635*
2	*0.7363841390*
3	*0.7390581394*
4	*0.7390851492*
5	*0.7390851334*

A ordem de convergência do método da secante é um pouco menor que do método de Newton. Neste exemplo se tivéssemos aplicado o método de Newton teríamos a mesma ordem de precisão com $n = 3$. Na verdade podemos mostrar que sua ordem d convergência é $r = (1+\sqrt{5})/2 \approx 1.62$. O método da secante e o de Newton são muitas vezes usados para refinar uma resposta obtida por outra técnica tal como o método da bissecção. Isto é feito pois estes métodos requerem uma boa primeira aproximação, e neste caso possuem convergência rápida.

2.1.7 Método de Müller

Este método foi desenvolvido por Müller (1956), ele é uma generalização do método da secante e pode ser usado para calcular uma raíz da função qualquer $f(x)$. Apesar disto, este método é particularmente usado para aproximar raízes de um polinômio. O método da secante, começa com duas aproximações iniciais x_0 e x_1 e determina a próxima aproximação x_2 como a intersecção do eixo do x com a reta que passa através dos pontos $x_0, f(x_0)$ e $x_1, f(x_1)$.
O método de Müller usa três aproximações iniciais x_0, x_1, x_2 e determina a próxima aproximação x_3 como a intersecção do eixo dos x com a parabola que passa através dos pontos $x_0, f(x_0)$, $x_1, f(x_1)$ e $x_2, f(x_2)$.
A derivação do método de Müller começa considerando três valores iniciais x_0, x_1, x_2 e o seguinte polinômio quadrático:

$$P(x) = a(x - x_2)^2 + b(x - x_2) + c \tag{2.19}$$

Como queremos que este polinômio passe pelos pontos $x_0, f(x_0)$, $x_1, f(x_1)$ e $x_2, f(x_2)$, podemos determinar as constantes a, b e c da parábola $P(x)$ como,

$$\begin{aligned} f(x_0) =& a(x_0 - x_2)^2 + b(x_0 - x_2) + c \\ f(x_1) =& a(x_1 - x_2)^2 + b(x_1 - x_2) + c \\ f(x_2) =& c \end{aligned} \tag{2.20}$$

isto é, resolvendo o sistema (2.20)

$$\begin{aligned} c =& f(x_2) \\ b =& \frac{(x_0 - x_2)^2[f(x_1) - f(x_2)] - (x_1 - x_2)^2[f(x_0) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)} \\ a =& \frac{(x_1 - x_2)[f(x_0) - f(x_2)] - (x_0 - x_2)[f(x_1) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)} \end{aligned} \tag{2.21}$$

Para determinarmos x_3, o zero da parábola $P(x)$, aplicamos a fórmula de Baskara modificada e assim evitamos problemas de erro de arredondamento, causados pela subtração de números próximos, isto é aplicamos a fórmula como,

$$(x_3 - x_2) = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}} \tag{2.22}$$

A fórmula (2.22) nos dá duas possibilidades para a escolha de x_3, dependendo do sinal que precede o termo radical. Como queremos a raíz que x_3 seja a raíz mais próxima de x_2 possível, escolhemos o sinal que precede o radical de forma que o módulo do denominador seja o maior possível.
Assim, se

$$\begin{aligned} mod1 &= |b + \sqrt{b^2 - 4ac}| \\ mod2 &= |b - \sqrt{b^2 - 4ac}| \end{aligned}$$

$$denominador = \begin{cases} b + \sqrt{b^2 - 4ac} & \text{se } mod1 \geq mod2 \\ b - \sqrt{b^2 - 4ac} & \text{se } mod1 < mod2 \end{cases}$$

E desta forma, temos

$$x_3 = x_2 - \frac{2c}{denominador} \tag{2.23}$$

onde a, b e c são dados por (2.21). Uma vez que x_3 é determinado, o procedimento descrito acima é reinicializado usando x_1, x_2 e x_3 para determinar x_4. O processo repete-se até que uma solução satisfatória seja obtida.
Nota: Note que em cada iteração o método envolve a raíz $\sqrt{b^2 - 4ac}$, logo o método irá aproximar também raízes complexas.

A seguir apresentamos o algoritmo deste método.

ALGORITMO 5 (Müller) *Encontra solução para $f(x) = 0$, dadas três aproximações x_1, x_2 e x_3;*
ENTRADA *x_1, x_2, x_3;a,b; Tol=10^{-5}; Número máximo de iterações N_0*
SAIDA *Solução aproximada p ou mensagem de falha*

Passo 1 *Faça:*

$h_1 = x_1 - x_0$;
$h_2 = x_2 - x_1$;
$\delta_1 = (f(x_1) - f(x_0))/h_1$;
$\delta_2 = (f(x_2) - f(x_1))/h_2$;
$d = (\delta_1 - \delta_2)/(h_2 + h_1)$;
$i = 1$.

Passo 2 *Enquanto $i \leq N_0$ (3-7).*

Passo 3 *Faça:*
$b = \delta_2 + h_2 d$;
$D = (b^2 - 4f(x_2)d)^{1/2}$. *(Obs: Pode ser aritmética complexa.)*

Passo 4 *Se $|b - D| < |b + D|$*
então faça $E = b + D$
caso contrário faça $E = b - D$.

Passo 5 *Faça*
$h = -2f(x_2)/E$;
$p = x_2 + h$.

Passo 6 *Se $|h| < TOL$ então*
SAÍDA (p);
PARE.

Passo 7 *Faça*

$x_0 = x_1$; *(preparando nova iteração)*
$x_1 = x_2$;
$x_2 = p$;
$h_1 = x_1 - x_0$;
$h_2 = x_2 - x_1$;
$\delta_1 = (f(x_1) - f(x_0))/h_1$;
$\delta_2 = (f(x_2) - f(x_1))/h_2$;
$d = (\delta_2 - \delta_1)/(h_2 - h_1)$;
$i = i + 1$

Passo 8 *SAÍDA ('Método falhou depois de N_0 iterações')*
STOP

Exemplo 17 *Considere o polinômio $P(x) = 16x^4 - 40x^3 + 5x^2 + 20x + 6$, considerando $TOL = 10^{-5}$. Os resultados produzem usando o o algoritmo de Müller e diferentes valores para x_0, x_1 e x_2.*

	$x_0 = 0.5$, $x_1 = -0.5$ *e* $x_2 = 0$	
i	x_i	$f(x_i)$
3	$-0.555556 + 0.598352i$	$-29.4007 - 3.89872i$
4	$-0.435450 + 0.102101i$	$+1.33223 - 1.19309$
5	$-0.390631 + 0.141852i$	$+0.375057 - 0.670164i$
6	$-0.357699 + 0.169926i$	$-0.146746 - 0.00744629i$
7	$-0.356051 + 0.162856i$	$-0.183868 \times 10^{-2} + 0.539780 \times 10^{-3}i$
8	$-0.356062 + 0.162758i$	$+0.286102 \times 10^{-5} + 0.953674 \times 10^{-6}$

	$x_0 = 0.5$, $x_1 = 1.0$ *e* $x_2 = 1.5$	
i	x_i	$f(x_i)$
3	1.28785	-1.37624
4	1.23746	$+0.126941$
5	1.24160	$+0.219440 \times 10^{-2}$
6	1.24168	$+0.257492 \times 10^{-4}$
7	1.24168	$+0.257492 \times 10^{-4}$

	$x_0 = 2.5$, $x_1 = 2.0$ *e* $x_2 = 2.25$	
i	x_i	$f(x_i)$
3	1.96059	-0.611255
4	1.97056	$+0.748825 \times 10^{-2}$
5	1.97044	-0.295639×10^{-4}
6	1.97044	-0.259639×10^{-4}

Nota: Geralmente o método de Müller converge para uma raíz da função $f(x)$ para qualquer chute inicial, problemas existem, mas quase nunca ocorrem. (ex: se $f(x_i) = f(x_{i+1}) = f(x_{i+2})$). A *ordem de convergência* do método é $\alpha = 1.84$.

2.2 Métodos para calcular raízes de Polinômios

Os métodos desenvolvidos nas seções anteriores, em particular o método de Newton, podem ser usados para encontrar raízes de polinômios. Mas para calcularmos os zeros de um polinômio devemos explorar, sempre que possível, sua estrutura especial. Além disto, freqüentemente o problema é complicado pois queremos calcular também seus zeros complexos. Desta forma encontrar zeros de polinômios merece uma atenção especial. Existem muitos métodos clássicos para encontrar raízes de polinômios tais como o método de Bairstow e o método de Laguerre (ver Numerical Analysis de Kincaid e Cheney, 1996), o método de Bernoulli, o método de Jenkins-Traub com ordem de convergência 3 (ver Jenkins e Traub, SIAMNA 7, 545-566, 1970 ou Jenkins e Traub, NM 14, 252-263, 1970) e um método baseado no método de Newton. Estes dois últimos proporcionam algoritmos estáveis para o cálculo destas raízes.

Um método numérico para determinar um zero de um polinômio geralmente toma a forma de uma receita para construir uma ou muitas seqüências z_n de números complexos supostos a convergir para um zero do polinômio. Cada algoritmos possui suas vantagens e desvantagens e desta forma a escolha para o "melhor"algoritmo para um dado problema nem sempre é fácil.

Um algoritmo razoável deve convergir, isto é, a seqüência gerada por ele deve, sob certas condições aceitáveis, convergir para um zero de um dado polinômio. Um algoritmo deve também poder produzir aproximações tanto para raíz real ou complexa de um dado polinômio. Outras propriedades que um algoritmo pode ou não ter são as seguintes:

1. *Convergência Global*: Muitos algoritmos podem ser garantidos ter convergência somente se o valor inicial z_0 está suficientemente próximo de uma raíz do polinômio. Estes são chamados métodos **localmente convergentes**. Algoritmos que não requerem isto são ditos **globalmente convergentes**

2. *Convergência Incondicional*: Alguns algoritmos só irão convergir se o dado polinômio possua algumas propriedades especiais, isto é, zeros simples, zeros não equi-modulares. Estes algoritmos são **condicionalmente convergentes**. Se um algoritmo é convergente (local ou globalmente) para todos os polinômios, ele é **incondicionalmente convergente**

3. *Estimativa a posteriori* Na prática os algoritmos são terminados artificialmente por algum critério de parada. Assim é desejável que o algoritmo possa fazer uma estimativa de erro para o erro $|z_n - \xi|$. (ξ a raíz do polinômio)

4. *Velocidade de convergência*: O conceito de *ordem* é freqüentemente usado para medir a velocidade de convergência de um algoritmo. A ordem ν é definida como o supremo dos números reais α para que

$$\lim_{n\to\infty} \sup \frac{|z_{n+1} - \xi|}{z_n - \xi|^\alpha} < \infty$$

5. *Determinação simultânea de zeros*

6. *Insensibilidade de grupos de zeros muito próximos ou iguais*

7. *Estabilidade Numérica* isto é Estabilidade significa pouca sensibilidade ao arredondamento nas operações efetuadas.

Para estudarmos métodos desenvolvidos exclusivamente para polinômios, vamos considerar o seguinte polinômio de grau n:

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + ... + a_1 z + a_0 = 0 \tag{2.24}$$

na qual os coeficientes a_k são coeficientes e z pode ser número complexo. Se $a_n \neq 0$ então p tem grau n.

Teorema 16 (Teorema Fundamental da Álgebra) *Cada polinômio não constante possui ao menos uma raíz no plano complexa.*

Teorema 17 (Número de Zeros) *: Um polinômio de grau n possui exatamente n zeros no campo complexo. (Considera-se que um zero de multiplicidade m é m vezes raíz do polinômio)*

Teorema 18 (Teorema do Resto:) *Se um polinômio P, de grau $n \geq 1$, é dividido por um fator linear $(z - c)$, então $p(z) = (z - c)q(z) + r$, onde $q(z)$ é o polinômio quociente de grau $n - 1$ e R é um número complexo chamado de resto.*

Observe no teorema acima que

1. Se $z = c$, então temos que $p(c) = r$.
2. Se c é um zero de $p(z)$ então no teorema 18 $r = 0$ e $(z - c)$ é um fator de $p(z)$.

Teorema 19 (Disco contendo todos os zeros de um polinômio) *Todos os zeros do polinômio descrito pela equação (2.24) estão dentro de um círculo aberto de centro na origem do plano complexo e cujo raio é*

$$\rho = 1 + |a_n|^{-1} \max_{0 \leq k < n} |a_k|$$

Uma outra idéia útil para a análise da localização de raízes de polinômios é que se considerarmos o polinômio $p(z)$ definido por (2.24) e definirmos o polinômio,

$$s(z) = z^n \left(a_n \left(\frac{1}{z} \right)^n + a_{n-1} \left(\frac{1}{z} \right)^{n-1} + ... + a_0 \right) = a_n + a_{n-1} z + ... + a_0 z^n,$$

temos que $s(z)$ é um polinômio de grau igual ou menor que n e que se $z_0 \neq 0$ é um número complexo de forma que $p(z_0) = 0$ é equivalente dizer que $s(\frac{1}{z_0}) = 0$. Desta observação resulta que:

Teorema 20 (Zeros Não nulos de um polinômio) *Se todos os zeros do polinômio descrito pela equação $s(z) = z^n p(1/z)$ estão em um disco centrado em zero e com raio ρ, então todos os zeros não nulos de $p(z)$ encontram-se fora do disco $|z| = \frac{1}{\rho}$.*

Uma outra idéia útil para a análise de polinômios é que: Se tomamos o polinômio p da equação (2.24), e consideramos a função $s(z) = z^n p(1/z)$. Então

$$\begin{aligned} s(z) =& z^n \left[a_n \left(\frac{1}{z}\right)^n + a_{n-1} \left(\frac{1}{z}\right)^{n-1} + ... + a_0 \right] \\ =& a_n + a_{n-1} z + a_{n-2} z^2 + ... + a_0 z^n \end{aligned} \tag{2.25}$$

Exemplo 18 *Vamos considerar $p(z) = z^4 - 4z^3 + 7z^2 - 5z - 2$. Pelos teoremas (19) e (20), concluímos que os zeros não nulos deste polinômio estão no anel $r^{-1} \leq |z| \leq R$ onde:*

$$R = 1 + |1|^{-1}.7 = 8 \quad e \quad r = 1 + |2|^{-1}.7 = \frac{9}{2}$$

isto é

$$\frac{2}{9} \leq |z| \leq 8$$

Teorema 21 *Se f é contínua em um intervalo $[a, b]$,*

1. *Se $f(a)f(b) < 0$ então temos um número ímpar de raízes neste intervalo.*
2. *Se $f(a)f(b) > 0$ então temos um número par de raízes neste intervalo.*
3. *Se f' é contínua e possui sinal constante neste intervalo,*
   (a) *$f(a)f(b) < 0$ então existe uma raíz neste intervalo*
   (b) *$f(a)f(b) > 0$ então não existe raíz neste intervalo*

1. **Regra de Descartes** O número de raízes positivas de um polinômio $P(z)$ com coeficientes reais nunca é maior que T, o número de troca de sinal na seqüência de seus coeficientes não nulos; e se é menor, então é sempre um número par.

   A mesma regra pode ser aplicada para enumerar as raízes negativas, analisando $P(-z)$.

   Exemplo 19 *$P(z) = 3z^3 + z^2 - z - 1$. A seqüência é $++--$. Logo $T = 1$ e $P(z)$ possui então uma raíz positiva. $P(-z) = -3z^3 + z^2 + z - 1$. A seqüência é $-++-$. Logo $T = 2$ assim o número de raízes negativas é 2 ou 0.*

2. **Regra de Du Gua** Seja $P(z) = a_n z^n + ... + a_1 z + a_0$ um polinômio de grau n sem raízes nulas. Se para algum k, $1 < k < n$, tem-se $a_k^2 \leq a_{k+1} a_{k-1}$, então $P(z)$ possui raízes complexas.

3. **Regra da Lacuna**

   (a) Se os coeficientes de $p(z)$ são todos reais e para algum k, $1 \leq k < n$, tem-se $a_k = 0$ e $a_{k-1} a_{k+1} > 0$, então $p(z)$ terá raízes complexas.

(b) Se os coeficientes de $p(z)$ são todos reais e existem dois ou mais coeficientes sucessivos nulos, então $p(z) = 0$ tem raízes complexas.

Exemplo 20 *Seja* $p(z) = 2z^5 + 3z^4 + z^3 + 2z^2 - 5z + 3$. *Para* $p(z)$, *o número de trocas de sinal é* $T = 2$, *logo pela regra de Descartes* $p(z)$ *possui duas ou nenhuma raíz positiva. Para* $p(-z)$ *o número de trocas é* $T = 3$ *logo* $p(z)$ *possui três ou uma raíz negativa.*

Testando $a_k^2 \leq a_{k+1}a_{k-1}$ *temos para* $k = 2$, $a_2^2 \leq a_3 a_1$ *ié* $1 \leq (3)(2) = 5$. *Assim por Du Gua, temos raízes complexas para* $p(z)$.
A regra da Lacuna nada afirma neste exemplo.

4. **Cota de Laguerre-Thibault** Dado um polinômio de coeficientes reais, procede-se a divisão do polinômio por $z - 1$, $z - 2$, ... até $z - m$, onde o quociente $Q(z)$ tenha todos os coeficientes positivos ou nulos, assim $R > 0$. Tal m é uma cota superior das raízes de P. Para determinar uma cota inferior basta aplicar o mesmo procedimento em $P(-z)$.

5. **Cota de Fujiwara** Seja $\hat{z}$ uma raíz real ou complexa de $P(z) = a_0 z^n + ... + a_{n-1}z + a_n$. Então,

$$|\hat{z}| \leq 2 \max \left\{ \left| \frac{a_1}{a_0} \right|, \left| \frac{a_2}{a_0} \right|^{1/2}, ..., \left| \frac{a_{n-1}}{a_0} \right|^{1/(n-1)}, \left| \frac{a_n}{a_0} \right|^{1/n} \right\}.$$

Exemplo 21 *Que região do plano contém as raízes de* $P(z) = z^4 - 14z^2 + 24z - 10$. *Pela cota de Fujiwara temos:*

$$|\hat{z}| \leq 2 \max\{0, 14^{1/2}, 24^{1/3}, 10^{1/4}\} \leq 2(3.74) = 7.48$$

6. **Cota de Kojima** Seja $p(z) = a_0 z^n + a_1 z^{n-1} + ... + a_{n-1}z + a_n$. Toda raíz real ou complexa $\hat{z}$ satisfaz

$$|\hat{z}| \leq q_1 + q_2$$

onde q_1 e q_2 são os maiores valores de

$$\left\{ \left| \frac{a_i}{a_0} \right|^{1/i} \right\}, \quad i = 1, 2, ..., n.$$

Exemplo 22 *Seja* $p(z) = z^5 + z^4 - 9z^3 - z^2 + 20z - 12$. *Temos que:*

$$\left| \frac{a_1}{a_0} \right|^{1/1} = 1; \ \left| \frac{a_2}{a_0} \right|^{1/2} = 3; \ \left| \frac{a_3}{a_0} \right|^{1/3} = 1; \ \left| \frac{a_4}{a_0} \right|^{1/4} = 2.114742527; \ \left| \frac{a_5}{a_0} \right|^{1/5} = 1.643751829.$$

logo

$$|\hat{z}| \leq 3 + 2.114742527 = 5.114742527$$

Entre os métodos desenvolvidos especificamente para o cálculo de raízes de polinômios vamos estudar os métodos de Newton-Horner, Bairstow e o método de Laguerre. Os dois primeiros foram escolhidos por ter entendimento mais simples e o método de Laguerre por ser simples e apresentar ordem de convergência 3 e ser usado internamente por muitos "pacotes"matemáticos. Para isto, vamos considerar o seguinte problema:

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = 0 \tag{2.26}$$

onde os coeficientes a_i são números reais e z uma variável complexa. Os métodos iterativos já vistos podem ser aplicados na resolução deste problema, no entanto encontrar as raízes de um polinômio $p(z)$, na prática, aparece com tanta freqüência que justifica o desenvolvimento de métodos particularmente adaptados a este problema que leve em conta sua forma polinomial. Além disto a necessidade de encontrar raízes complexas ou todas as raízes de um polinômio dão a este problema um enfoque especial. Muitas vezes não sabemos onde estão estas raízes e desta forma precisamos de métodos que sejam incondicionalmente convergentes principalmente em se tratando de polinômios de alto grau. Antes de vermos o método de Bairstow é importante salientar que existe uma vasta literatura sobre este assunto como por exemplo ver Wilf H.S. *The numerical solution of polynomials equations*, John Wiley & sons Inc. NY, 1960. Como este temos outros métodos clássicos para encontrar raízes de polinômios tais como o método de *Bernoulli*, método de *Laguerre* e o método de *Jekins-Traub*. Deste apenas veremos o método de Laguerre, os outros podem ser encontrados no livro "A. Ralston e P. Rabinowitz, *A first course in numerical analysis*, 1978", ou "Householder A.S., *The numerical treatment of a single nonlinear equation*, McGraw-Hill, 1970". O problema de zeros de um polinômio ainda pode ser tratado de forma matricial, calculando os autovalores de uma matriz associada ao polinômio, chamada de matriz "companheira".

Devemos ainda observar que se z_j é um zero de $p(z)$ pelo teorema do resto podemos escrever,

$$p(z) = (z - z_j)(b_{n-1} z^{n-1} + b_{n-2} z^{n-2} + \cdots + b_0) + R_j \tag{2.27}$$

se $R_j = 0$ o polinômio $q(z) = b_{n-1} z^{n-1} + b_{n-2} z^{n-2} + \cdots + b_0)$ é o polinômio deflatado de grau $n-1$ isto é o polinômio cujos zeros são idênticos aos restantes $n-1$ zeros de $p(z)$. Assim uma vez que encontramos z_j, uma raíz de $p(z)$ por algum método de encontrar raíz, através do algoritmo de Horner (divisão sintética), podemos encontrar o polinômio deflatado $q(z)$. Para encontrarmos os zeros adicionais aplicamos um método para encontrar uma raíz de $q(z)$ e encontrarmos um novo polinômio deflatado de $q(z)$ e assim sucessivamente. Este procedimento é chamado de **deflação**. Ainda notamos que para termos mais precisão nestas raízes, cada raíz de um polinômio deflatado encontrada pode servir de chute inicial para algum método iterativo, como por exemplo o de Newton-Raphson.

2.2.1 O método de Newton para polinômios

Para usarmos o procedimento de Newton Raphson para localizar zeros de um polinômio $P(x)$ é necessário avaliar seu valor em cada x_i. Para o algoritmo ser computacionalmente eficiente, podemos calcular estes valores, a cada iteração, usando o algoritmo de Horner. Por este procedimento fazemos uma decomposição do polinômio

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_0 \tag{2.28}$$

do tipo

$$P(x) = (x - x_0)Q(x) + b_0 \tag{2.29}$$

onde

$$Q(x) = b_n x^{n-1} + b_{n-1}x^{n-2} + ... + b_2 x + b_1 \tag{2.30}$$

assim se fizermos $x = x_0$ em (2.29) temos que $P(x_0) = b_0$, além disto se derivamos (2.29) temos,

$$P'(x) = Q(x) + (x - x_0)Q'(x) \tag{2.31}$$

Agora fazendo $x = x_0$ em (2.31) vemos que $P'(x_0) = Q(x_0)$. Assim aplicamos o método de Horner, agora em $Q(x)$ e obtemos o valor de $Q(x_0) = P'(x_0)$.

O algoritmo de Horner é dado por:

ALGORITMO 6 (Horner) *Para calcular o polinômio*

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$$

E e sua derivada em x_0;
ENTRADA *Grau n; Coeficiêntes $a_0, a_1, ..., a_n$; x_0.*
SAIDA *$y = P(x_0)$; $z = P'(x_0)$.*

Passo 1 *Faça*

$y = a_n$; *(calcula b_n para P)*
$z = a_n$; *(calcula b_{n-1} para Q)*

Passo 2 *Para $j = n-1, n-2, ..., 1$*

Faça *Faça* $y = x_0 y + a_j$; *(Calcula b_j para P.)*
$z = x_0 z + y$. *(Calcula b_{j-1} para Q.*

Passo 3 *Faça $y = x_0 y + a_0$. (Calcula b_0 para P.)*

Passo 4 *SAÍDA (y, z);*
STOP.

A seguir, vamos exemplificar este procedimento.

Exemplo 23 *Encontre uma aproximação de um zero de*

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

com "chute"inicial $x_0 = -2$ e usando os métodos de Newton para encontrar a raíz e o de Horner para avaliar $P(x_i)$ e $P'(x_i)$.

- *Fazendo a primeira iteração com* $x_0 = -2$,

	2	0	-3	3	-4	Coef. de $P(x)$
-2	2	-4	5	-7	10	$= P(-2)$
	2	-8	21	-49		$= Q(-2) = P'(-2)$

Note que do esquema acima temos que

$$Q(x) = 2x^3 - 4x^2 + 5x - 7 \quad \text{e} \quad P'(-2) = Q(-2)$$

assim,

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)} = -2 - \left(\frac{10}{-49}\right) \approx -1.796$$

- *Repetindo o processo, agora para* x_1 *temos,*

	2	0	-3	3	-4	Coef. de $P(x)$
-1.796	2	-3.592	3.451	-3.197	1.7242	$= P(-1.796)$
	2	-7.184	16.353	-32.565		$= P'(-1.796)$

assim,

$$x_2 = x_1 - \frac{P(x_1)}{P'(x_1)} = -1.7946 - \left(\frac{1.742}{-32.565}\right) \approx -1.7425$$

x_2, x_3, *etc... são calculados da mesma forma.*

O zero deste problema com cinco casas decimais é -1.73896.

Note que o polinômio chamado de Q depende da aproximação que está sendo usada e troca a cada iteração. Se a n-ésima iteração, x_N, no procedimento de Newton Raphson é aproximadamente um zero de P, então

$$P(x) = (x - x_N)Q(x) + b_0 = (x - x_N)Q(x) + P(x_N) \approx (x - x_N)Q(x);$$

logo $x - x_N$ é uma aproximação de um fator de P. Chamando $\hat{x}_1 = x_N$ uma aproximação de um zero de P e $Q_1(x)$ um fator aproximado, temos:

$$P(x) \approx (x - \hat{x}_1)Q_1(x),$$

agora, podemos encontrar uma aproximação de uma outra raíz de P aplicando NR sobre $Q_1(x)$.Se P tem grau n, com n zeros reais, este procedimento é aplicado repetidamente, encontrando assim $(n-2)$ aproximações de $n-2$ zeros de P e um fator quadrático Q_{n-2} que será resolvido pos Baskara. Este procedimento é chamado de deflação.

Obs: Para melhorar a deflação cada zero aproximado é usado como "chute"e o método de NR, com este chute é aplicado novamente sobre P.

Exemplo 24 *Vamos usar o método de Newton Raphson e deflação para encontrar todas as raízes do seguinte polinômio:*

$$p(x) = 10x^4 + 11x^3 + 20x^2 + 45x - 12$$

Podemos ver que as raízes deste polinômio estão localizados em um anel, com raios:

$$R = 1 + \frac{45}{10} = 5.5 \quad e \quad r = \frac{1}{1 + \frac{45}{12}} = 0.210526$$

pela regra da troca de sinal podemos afirmar que possui exatamente uma raíz real positiva e 3 ou 1 raíz real negativa. Ainda como $20^2 \leq (45)(11)$*, podemos afirmar que temos um par complexo. Assim temos 1 raíz real positiva, 1 real negativa e um par complexo conjugado.*

Para iniciarmos o método de Newton, sabemos que existe uma raíz real entre 0.2 *e* 5.5*. Como* $p(0.2) < 0$ *e* $p(1) > 0$*, vamos escolher* $x_0 = 0.5$*.*

m Repetindo o processo, agora para x_1 *temos,*

	10	*11*	*20*	*45*	*-12*	*Coef. de* $P(x)$
0.5	*10*	*16*	*28*	*59*	*17.5*	$= P(0.5)$
	10	*21*	*38.5*	*78.25*		$= P'(0.5)$

$$x_1 = 0.5 - \frac{17.5}{78.25} = 0.276357827$$

	10	*11*	*20*	*45*	*-12*	*Coef. de* $P(x)$
0.276358	*10*	*13.763578*	*23.803673*	*51.578331*	*2.254076*	$= P(0.276358)$
	10	*16.527157*	*28.371082*	*59.418902*		$= P'(0.276358)$

$$x_2 = 0.276358 - \frac{2.254076}{59.418902} = 0.238422$$

	10	*11*	*20*	*45*	*-12*	*Coef. de* $P(x)$
0.238422	*10*	*13.384225*	*23.191100*	*50.529280*	*0.047317*	$= P(0.238422)$
	10	*15.768450*	*26.950654*	*56.954922*		$= P'(0.238422)$

$$x_3 = 0.238422 - \frac{0.0473172}{56.954922} = 0.237592$$

	10	*11*	*20*	*45*	*-12*	*Coef. de* $P(x)$
0.23759172	*10*	*13.375917*	*23.178007*	*50.506902*	*0.0000215768*	$= P(0.237591)$
	10	*15.751834*	*26.920512*	*56.902993*		$= P'(0.237591)$

$$x_4 = 0.23759172 - \frac{0.0000215768}{56.90299317} = 0.23759134$$

Desta forma aceitamos a primeira raíz como 0.237591*, fatorando usando o algoritmo de Horner temos:*

	10	*11*	*20*	*45*	*-12*
0.23759134	*10*	*13.375913*	*23.178001*	*50.506892*	4.48×10^{-12}

Desta forma pelo teorema do resto temos que

$$p(x) = (x - 0.237591)(10x^3 + 13.375913x^2 + 23.178001x + 50.506892) + 4.48 \times 10^{-12}$$

Desprezando o resto obtemos uma fatoração para $p(x)$*. Sabemos que o polinômio* $p(x)$ *possui uma raíz real negativa entre* -5.5 *e* -0.2*, assim aplicando o método anterior sobre o polinômio* $q(x)$ *e notando que* $q(0) = 850.506892$ *,* $q(-1) = 30.704804$*,* $q(-2) = -22.345456$*,* $q(-1.5) = 12.0857$*, vemos que a raíz está entre* -2 *e* -1.5*. Vamos escolher* $x_0 = -1.75$*. Assim seguindo procedimento semelhante ao anterior obtemos:*

$$x_0 = -1.75$$

$$x_1 = -1.7106575$$

$$x_2 = -1.7097381$$

$$x_3 = -1.7097376$$

$$x_4 = -1.7097376$$

assumimos a raíz como x_4*, assim fatoramos novamente:*

$$p(x) = (x - 0.237591)(x + 1.7097376)g(x) + r$$

os coeficientes de $g(x)$ *são obtidos pelo algoritmo de Horner como:*

	10	*13.375913*	*23.178001*	*50.506892*
-1.7097376	*10*	*-3.721464*	*29.540727*	*-7.545* $\times10^{-25}$

Desta forma temos que o resto é -7.545×10^{-25}

$$g(x) = 10x^2 - 3.721464x + 29.540727$$

cujas raízes são $0.186073 \pm 1.708640i$*. ou seja:*

$$p(x) \approx (x - 0.237591)(x + 1.7097376)(x - 0.186073 + 1.708640i)(x - 0.186073 - 1.708640i)$$

2.2.2 O Método de Bairstow

Vamos observar que se z_j é um zero de $p(z)$ pelo teorema do resto podemos escrever,

$$p(z) = (z - z_j)(b_{n-1}z^{n-1} + b_{n-2}z^{n-2} + \cdots + b_0) + R_j \tag{2.32}$$

se $R_j = 0$ o polinômio $q(z) = b_{n-1}z^{n-1} + b_{n-2}z^{n-2} + \cdots + b_0)$ é o polinômio deflatado de grau $n-1$ isto é o polinômio cujos zeros são idênticos aos restantes $n-1$ zeros de $p(z)$. Assim uma vez que

encontramos z_j, uma raíz de $p(z)$ por algum método de encontrar raíz, através do algoritmo de Horner (divisão sintética), podemos encontrar o polinômio deflatado $q(z)$. Para encontrarmos os zeros adicionais aplicamos um método para encontrar uma raíz de $q(z)$ e encontaramos um novo polinômio deflatado de $q(z)$ e assim sucessivamente. Este procedimento é chamado de **deflação**. Devemos ainda notar que para termos mais precisão nestas raízes, cada raíz de um polinômio deflatado encontrada pode servir de chute inicial para algum método iterativo, como por exemplo o de Newton-Raphson.

Ainda temos que observar que se $p(z)$ é um polinômio real e nós precisamos encontrar raízes complexas, elas aparecem aos pares conjugados, isto é, $z_j = x_j + iy_j$ e $\bar{z}_j = x_j - iy_j$ e assim

$$(z - z_j)(z - \bar{z}_j) = z^2 + p_j z + q_j, \qquad \text{onde } p_j = 2x_j \text{ e } q_j = x_j^2 + y_j^2$$

é um fator de $p(z)$. Assim como estamos interessados em usar aritmética real, nós precisamos desenvolver um algoritmo tipo o de Horner para dividirmos por um fator quadrático, isto é:

$$\begin{aligned} p(z) =& (z^2 + p_j z + q_j)(b_{n-2}z^{n-2} + b_{n-3}z^{n-3} + \cdots + b_0) + \underbrace{b_{-1}(z + p_j) + b_{-2}}_{resto} \\ =& b_{n-2}z^n + (b_{n-3} + p_j b_{n-2})z^{n-1} + (b_{n-4} + p_j b_{n-3} + q_j b_{n-2})z^{n-2} + \ldots \\ & + (b_{-1} + p_j b_0 + q_j b_1)z + (b_{-2} + p_j b_{-1} + q_j b_0) \end{aligned} \tag{2.33}$$

Comparando as equações (2.26) e (2.26), vemos que os coeficientes b_k são dados por:

$$\begin{cases} b_{n-2} &= a_n \\ b_{n-3} &= a_{n-1} - p_j b_{n-2} \\ b_k &= a_{k+2} - p_j b_{k+1} - q_j b_{k+2} \qquad k = n-4, \ldots, 0, -1, -2 \end{cases} \tag{2.34}$$

O resto também pode ser escrito na forma $Rz + S$ onde:

$$\begin{aligned} R =& b_{-1} = a_1 - p_j b_0 - q b_1 \\ S =& b_{-2} + p_j b_{-1} = a_0 - p_j b_{-1} - q_j b_0 + p_j b_{-1} = a_0 - q_j b_0 \end{aligned} \tag{2.35}$$

Desta forma para $z^2 + p_j z + q_j$ ser um fator de $p(z)$ devemos ter $R(p_j, q_j) = S(p_j, q_j) = 0$. Neste caso o quociente polinomial é um polinômio deflatado de grau $n - 2$.

Descrição do método de Bairstow

No caso de polinômios reais, sabemos que zeros complexos ocorrem aos pares conjugados. Assim, ao invés de procurarmos um zero por vez, podemos procurar por pares de zeros os quais geram um fator real quadrático. Esta é a idéia básica da iteração de Bairstow, a qual assume um bom chute inicial. Assim queremos encontrar p e q de forma que as funções $R(p, q) = S(p, q) = 0$ definidas por (2.35). Note que b_1 e b_0 também são funções de p e q. Para resolver

estas equações usando um método iterativo,onde assumimos uma aproximação inicial p_0 e q_0, precisamos encontrar aproximações p_{i+1} e q_{i+1} de forma que $p_{i+1} = p_i + \triangle p_i$ e $q_{i+1} = q_i + \triangle q_i$ de forma que $R(p_{i+1}, q_{i+1}) = S(p_{i+1}, q_{i+1}) = 0$ ou o mais próximo de zero possível.

Agora, desejamos que as trocas $\triangle p_i$ e $\triangle q_i$ que irão resultar em p_{i+1} e q_{i+1}. Conseqüentemente devemos expandir as duas equações

$$\begin{aligned} R(p_i + \triangle p_i, q_i + \triangle q_i) &= 0 \\ S(p_i + \triangle p_i, q_i + \triangle q_i) &= 0 \end{aligned} \tag{2.36}$$

usando expansão em série de Taylor em torno de (p_i, q_i) e considerando apenas potências de $\triangle p_i$ e $\triangle q_i$ lineares, obtemos duas equações lineares aproximadas para $\triangle p_i$ e $\triangle q_i$,

$$\begin{aligned} R(p_i, q_i) + \left(\frac{\partial R}{\partial p}\right)_i \triangle p_i + \left(\frac{\partial R}{\partial q}\right)_i \triangle q_i &= 0 \\ S(p_i, q_i) + \left(\frac{\partial S}{\partial p}\right)_i \triangle p_i + \left(\frac{\partial S}{\partial q}\right)_i \triangle q_i &= 0 \end{aligned} \tag{2.37}$$

O sub-índice i denota que as derivadas parciais são calculadas no ponto p_i, q_i. Uma vez que as correções forem encontradas a iteração pode ser repetida até R e S serem suficientemente perto de zero. Este método é o método de Newton Raphson para duas variáveis. Podemos isolar p_{i+1} e q_{i+1}

$$\begin{aligned} p_{i+1} =& p_i - \frac{1}{D}\left[R\frac{\partial S}{\partial q} - S\frac{\partial R}{\partial q}\right]_{\substack{p = p_i \\ q = q_i}} \\ q_{i+1} =& q_i - \frac{1}{D}\left[S\frac{\partial R}{\partial p} - R\frac{\partial S}{\partial p}\right]_{\substack{p = p_i \\ q = q_i}} \end{aligned} \tag{2.38}$$

onde

$$D = \left|\begin{matrix} \frac{\partial R}{\partial p} & \frac{\partial S}{\partial p} \\ \frac{\partial R}{\partial q} & \frac{\partial S}{\partial q} \end{matrix}\right|_{\substack{p = p_i \\ q = q_i}}$$

Agora usando (2.35)podemos escrever

$$\begin{aligned} \frac{\partial R}{\partial p} =& -p\frac{\partial b_0}{\partial p} - q\frac{\partial b_1}{\partial p} - b_0 \\ \frac{\partial R}{\partial q} =& -p\frac{\partial b_0}{\partial q} - q\frac{\partial b_1}{\partial q} - b_1 \\ \frac{\partial S}{\partial p} =& -q\frac{\partial b_0}{\partial p} \\ \frac{\partial S}{\partial q} =& p\frac{\partial b_{-2}}{\partial q} + p\frac{\partial b_{-1}}{\partial q} \end{aligned} \tag{2.39}$$

de (2.33) temos:

$$\begin{aligned}
\frac{\partial b_k}{\partial p} &= -b_{k+1} - p\frac{\partial b_{k+1}}{\partial p} - q\frac{\partial b_{k+2}}{\partial p} \qquad k = n-3, \ldots, 0, -1 \\
\frac{\partial b_{n-2}}{\partial p} &= \frac{\partial b_{n-1}}{\partial p} = 0 \\
\\
\frac{\partial b_k}{\partial q} &= -b_{k+2} - p\frac{\partial b_{k+1}}{\partial q} - q\frac{\partial b_{k+2}}{\partial q} \qquad k = n-4, \ldots, 0, -1, -2 \\
\frac{\partial b_{n-3}}{\partial q} &= \frac{\partial b_{n-2}}{\partial q} = 0
\end{aligned} \tag{2.40}$$

Se definirmos d_k pela relação de recorrência

$$\begin{aligned}
d_k &= -b_{k+1} - pd_{k+1} - qd_{k+2} \qquad k = n-3, \ldots, 0, -1 \\
d_{n-2} &= d_{n-1} = 0
\end{aligned} \tag{2.41}$$

então segue de (2.40) que

$$\frac{\partial b_k}{\partial p} = d_k \quad e \quad \frac{\partial b_{k-1}}{\partial q} = d_k, \quad k = n-3, \ldots, 0, -1 \tag{2.42}$$

e finalmente que

$$\frac{\partial R}{\partial p} = d_{-1} \ , \ \frac{\partial R}{\partial q} = d_0, \ \frac{\partial S}{\partial p} = -qd_0 \ , \ \frac{\partial S}{\partial q} = d_{-1} + pd_0. \tag{2.43}$$

Desta forma as equações (2.38) que

$$\begin{aligned}
p_{i+1} &= p_i - \frac{1}{D}\left[b_{-1}(d_{-1} + p_i d_0) - (b_{-2} + p_i b_{-1})d_0\right] \\
q_{i+1} &= q_i - \frac{1}{D}\left[(b_{-2} + p_i b_{-1})d_{-1} + d_0 b_{-1} q_i\right] \\
&\text{onde} \\
D &= d_{-1}^2 + p_i d_0 d_{-1} + q_i d_0^2.
\end{aligned} \tag{2.44}$$

Exemplo 25 *Use a iteração de Bairstow para encontrar um fator quadrático de* $z^3 - z - 1$ *começando com* $p_0 = q_0 = 1$.

Vamos arranjar os cálculos na forma:

	(p_i, q_i)	
a_n	b_{n-2}	d_{n-3}
a_{n-1}	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	d_0
$\vdots$	b_0	d_{-1}
$\vdots$	b_{-1}	
a_0	b_{-2}	

usando (2.34) e (2.41) pra calcular as últimas duas colunas. Para este problema temos, usando as equações dadas em (2.44),

	(p_0, q_0)	$(1, 1)$
1	1	-1
0	-1	2
-1	-1	
-1	1	

	(p_1, q_1)	$\left(\frac{4}{3}, \frac{2}{3}\right)$
1	1	-1
0	$-\frac{4}{3}$	$\frac{8}{3}$
-1	$\frac{1}{9}$	
-1	$-\frac{7}{27}$	

e finalmente $p_2 = 1.32246$, $q_3 = 0.7544$, onde os valores reais são $p = 1.3247$ e $q = 0.7549$. Para encontrarmos as raízes devemos usar a fórmula de Baskara para resolver $z^2 + pz + q = 0$.

Quando o método de Bairstow converge, possui como característica convergência rápida como o método de Newton-Raphson. Porém, como usual em métodos para resolver equações não lineares simultâneas, convergência requer uma boa aproximação inicial. A vantagem é poder encontrar raízes complexas, mesmo que o valor de partida seja um número real.

2.2.3 O Método de Laguerre

O método de Laguerre é um algoritmo desenvolvido para encontrar raíz de um polinômio. Este método possui uma ótima característica, a convergência é garantida, independendo do valor inicial. Este algoritmo aproxima raízes reais ou complexas e, no caso da raíz procurada ser simples, possui ordem de convergência igual a 3. Aqui, vamos dar uma idéia deste método. Para isto vamos lembrar que pelo teorema fundamental da álgebra, todo o polinômio $p(z)$ pode ser fatorado através de suas raízes, ou

$$p(z) = a_n(z - z_1)(z - z_2)...(z - z_n). \tag{2.45}$$

Da igualdade acima, aplicando a função logaritmo dos dois lados da igualdades, temos

$$\ln|p(z)| = \ln|a_n| + \ln|(z - z_1)| + \ln|(z - z_2)| + ... + \ln|(z - z_n)|. \tag{2.46}$$

Usando a equação acima vamos definir as funções $A(x)$,

$$A(z) = \frac{d}{dz}\ln|p(z)| = \frac{1}{(z - z_1)} + \frac{1}{(z - z_2)} + ... + \frac{1}{(z - z_n)}, \tag{2.47}$$

ou de outra forma,

$$A(z) = \frac{d}{dz} \ln |p(z)| = \frac{p'(z)}{p(z)} \tag{2.48}$$

e

$$B(z) = -\frac{d^2}{dz^2} \ln |p(z)| = \frac{1}{(z - z_1)^2} + \frac{1}{(z - z_2)^2} + ... + \frac{1}{(z - z_n)^2}, \tag{2.49}$$

ou de outra forma,

$$B(z) = -\frac{d^2}{dz^2} \ln |p(z)| = -\frac{d}{dz}\left(\frac{d}{dz}\frac{p'(z)}{p(z)}\right) = -\frac{p(z)p''(z) - (p'(z))^2}{p^2(z)} = A^2(z) - \frac{p''(z)}{p(z)} \tag{2.50}$$

Agora, vamos supor estamos procurando a raíz z_1 e que a diferença entre esta raíz e a nossa estimativa z é um número complexo a, isto é $a = z - z_1$. Vamos ainda supor que todas as outras raízes se afastem de nossa estimativa z pelo menos b isto é $|b| < |z - z_i|$, $i = 2, 3, ..., n$, como $|b| > |a|$. Assim, podemos estimar o valor de a substituindo as estimativas feitas nas equações (2.47) e (2.49), ou

$$A(z) = \frac{1}{a} + \frac{n-1}{b}, \tag{2.51}$$

e

$$B(z) < \frac{1}{b^2} + \frac{n-1}{b^2}, \tag{2.52}$$

O algoritmo de Laguerre, aproxima o valor de a, resolvendo o sistema não linear, para a e b, dado pelas equações (2.51). A solução deste sistema é dada por,

$$a = \frac{n}{A(z) \pm \sqrt{(n-1)(nB(z) - A^2(z))}}, \tag{2.53}$$

e assim o próximo z é calculado como

$$z = z - a. \tag{2.54}$$

Devemos observar que a aritmética empregada neste método é complexa. A ordem de convergência deste método é 3 se z_1 é uma raíz simples. Existem modificações deste algoritmo para o caso de raíz múltipla. O algoritmo é feito usando para $A(z)$ e $B(z)$ as relações (2.48), (2.48) e a fórmula (2.54), onde a é calculado como (2.53)

Algorítmo de Laguerre

- Escolhemos um ponto de partida z_0

- Para $k = 0, 1, 2, ...$
  - Calculamos $A(z_k)$ como,
$$A(z_k) = \frac{p'(z_k)}{p(z_k)}$$
  - Calculamos $B(z_k$ como,
$$B(z_k) = A^2(z_k) - \frac{p''(z_k)}{p(z_k)}$$
  - Calcule a_k como,
$$a_k = \frac{n}{A(z_k) \pm \sqrt{(n-1)(nB(z_k) - A^2(z_k))}}$$
    onde no denominador, para evitarmos erro pela diminuição de números muito próximos e consequente erro de arredondamento, devemos escolher o sinal + ou − de forma que o denominador possua o maior valor absoluto.
  - Faça
$$x_{k+1} = z_k - a_k.$$
- Repetimos o procedimento até o erro absoluto (ou relativo) ser suficientemente pequeno, ou depois de exceder um número pré determinado de iterações.

A maior vantagem do método de Laguerre é que sua convergência é quase sempre garantida para algumas raízes do polinômio, não importando qual o valor escolhido para z_0 e, além disto, aproxima tanto valores de raízes reais quanto complexas. Ele é usado em pacotes computacionais como, *Numerical Recipes* na sub-rotina *zroots* e o método de Laguerre modificado é usado pelo pacote *NAG F77 library* na sub-rotina *C02AFF*. Sobre a família de métodos de Laguerre ver o artigo: Ljiljana D. Petkovic, Miodrag S. Petkovic e Dragan Zivkovic, "Hansen-Patrick's Family IS of Laguerre's Type", *Novi Sad J. Math*, Vol **33**1, pp 109-115, 2003.

Exemplo 26 *Vamos considerar o polinômio* $P(z) = 4x^3 + 3z^2 + 2z + 1$. *Neste caso* $n = 3$. *Vamos escolher*

1. $z_0 = -1$
   - $A(-1) = -4.0000000000$
   - $B(-1) = 7.0000000000$
   - *Calculando*
     $C_1 = A + \sqrt{2(3B - A^2)} = -0.8377223398$
     $C_2 = A - \sqrt{2(3B - A^2)} = -7.1622776607$
   - $z_1 = z_0 + n/C_2 = -0.5811388300$
2. $z_1 = -0.5811388300$

- $A(-0.5811388300) = 38.9736659610$
- $B(-0.5811388300) = 1639.6623695156$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 1639.6623695156$
  $C_2 = A - \sqrt{2(3B - A^2)} = -43.4889373220$

- $z_2 = z_1 + n/C_1 = -0.6058431463$

3. $z_2 = -0.6058431463$

- $A(-0.6058431463) = -73747.0441427632$
- $B(-0.6058431463) = 5.4383991072 \times 10^9$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 73742.4185458994$
  $C_2 = A - \sqrt{2(3B - A^2)} = -221236.5068314258$

- $z_3 = z_2 + n/C_2 = -0.6058295861$

4. $z_3 = -0.6058295861$

- $A(-0.6058295861) = 4.7182194357 \times 10^{14}$
- $B(-0.6058295861) = 2.2261594643 \times 10^{29}$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 1.4154658307 \times 10^{15}$
  $C_2 = A - \sqrt{2(3B - A^2)} = -4.7182194357 \times 10^{14}$

- $z_4 = z_3 + n/C_1 = -0.6058295861$

Agora vamos fazer $z_0 = 0$

1. $z_0 = 0$

- $A(-1) = 2.0000000000$
- $B(-1) = -2.0000000000$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 2.0000000000 + 4.4721400000i$
  $C_2 = A - \sqrt{2(3B - A^2)} = 2.0000000000 - 4.4721400000i$

- $z_1 = z_0 + n/C_1 = -0.2500000000 + 0.5590170000i$

2. $z_1 = -0.2500000000 + 0.5590170000i$

- $A(-0.2500000000 + 0.5590170000i) = -4.0000000000$
- $B(-0.2500000000 + 0.5590170000i) = 16.0000000000 - 21.4663000000i$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 6.1936800000 - 6.3175200000i$
  $C_2 = A - \sqrt{2(3B - A^2)} = -14.1937000000 + 6.3175200000i$

- $z_2 = z_1 + n/C_2 = -0.0735872000 + 0.6375370000i$

3. $z_2 = -0.0735872000 + 0.6375370000i$

- $A(-0.0735872000 + 0.6375370000i) = -520.8960000000 + 272.4890000000i$
- $B(-0.0735872000 + 0.6375370000i) = 196952.0000000000 - 286081.0000000000i$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 523.2090000000 - 277.6130000000i$
  $C_2 = A - \sqrt{2(3B - A^2)} = -1565.0000000000 + 822.5920000000i$

- $z_3 = z_2 + n/C_2 = -0.0720852000 + 0.6383270000i$

4. $z_3 = -0.0720852000 + 0.6383270000i$

- $A(-0.0720852000 + 0.6383270000i) = -6.1144273889 \times 10^8 + 1.1804919396 \times 10^9$
- $B(-0.0720852000 + 0.6383270000i) = -1.0196989997 \times 10^{18} - 1.4436064535 \times 10^{18}i$
- *Calculando*
  $C_1 = A + \sqrt{2(3B - A^2)} = 6.1144274120 \times 10^8 - 1.1804919447 \times 10^9 i$
  $C_2 = A - \sqrt{2(3B - A^2)} = -1.8343282189 \times 10^9 + 3.5414758241 \times 10^9 i$

- $z_4 = z_3 + n/C_2 = -0.0720852000 + 0.6383270000i$

2.3 Solução Numérica de Sistemas de Equações Não Lineares

A solução de sistemas não lineares de equações deve, sempre que possível ser evitado. Isto as vezes é feito através de uma linearização do problema, isto é o sistema não linear é aproximado localmente por um sistema linear. Quando esta solução aproximada não satisfaz, devemos atacar o problema "de frente". Em geral os métodos para resolução de sistemas não lineares, são adaptações dos métodos usados na resolução de uma equação não linear. Aqui iremos apresentar a extensão do método de Newton e o método da secante aplicados na resolução de sistemas não lineares.

2.3.1 O método de Newton

Vamos considerar um sistema não linear de equações descrito por:

$$\begin{array}{rcl} f_1(x_1, x_2, ..., x_n) & = & 0 \\ f_2(x_1, x_2, ..., x_n) & = & 0 \\ \vdots & & \vdots \\ f_n(x_1, x_2, ..., x_n) & = & 0 \end{array} \tag{2.55}$$

onde cada função f_i para $i = 1, 2, ..., n$ está definida de $\mathbb{R}^n \to \mathbb{R}$ e $\mathbf{X} = (x_1, x_2, ..., x_n)^T$ é um vetor definido em $\mathbb{R}^n$. Se notarmos a função $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_n(\mathbf{x}))^T$, onde $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$, podemos reescrever o sistema (2.55) como:

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}. \tag{2.56}$$

No caso unidimensional, o método de Newton escolhe uma função $\phi(x)$ de forma que o método do ponto fixo definido por

$$g(x) = x - \phi(x)f(x)$$

possua uma convergência quadrática. Neste caso a escolha é $1/f'(x)$, se $f'(x) \neq 0$. No caso matricial, devemos escolher uma matriz

$$\mathbf{A}(\mathbf{x}) = \begin{pmatrix} a_{11}(\mathbf{x}) & a_{12}(\mathbf{x}) & ... & a_{1n}(\mathbf{x}) \\ a_{21}(\mathbf{x}) & a_{22}(\mathbf{x}) & ... & a_{2n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{41}(\mathbf{x}) & a_{42}(\mathbf{x}) & ... & a_{4n}(\mathbf{x}) \end{pmatrix}$$

de forma que

$$\mathbf{G}(\mathbf{x}) = \mathbf{x} - \mathbf{A}^{-1}(\mathbf{x})\mathbf{F}(\mathbf{x})$$

possua convergência quadrática. Podemos mostrar, que esta escolha recai sobre o Jacobiano de $\mathbf{F}(\mathbf{X})$. Isto é podemos provar que:

Se $\mathbf{p}$ uma solução de $\mathbf{G}(\mathbf{x}) = \mathbf{x}$ e existe um número $\delta > 0$ de forma que:

1. Se $\frac{\partial g_i(\mathbf{x})}{\partial x_j}$, para $i = 1, 2, .., n$ e $j = 1, 2, ..., n$, é contínua uma vizinhança $N_\delta = \{\mathbf{x}/||\mathbf{x} - \mathbf{p}|| < \delta\}$;
2. Se $\frac{\partial^2 g_i(\mathbf{x})}{(\partial x_j \partial x_k)}$, é contínua e $|\partial^2 g_i(\mathbf{x})/(\partial x_j \partial x_k)| \leq M$ para algum M constante sempre que $\mathbf{x} \in N_\delta$, para cada $i = 1, 2, .., n$, $j = 1, 2, ..., n$ e $k = 1, 2, ..., n$;
3. Se $\frac{\partial g_i(\mathbf{p})}{\partial x_k} = 0$, para $i = 1, 2, .., n$, $k = 1, 2, .., n$.

Então existe um número $\hat{\delta} \leq \delta$ de forma que a seqüência gerada por $\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^{(k-1)})$ converge quadraticamente para $\mathbf{p}$ para qualquer escolha de $\mathbf{x}^{(0)}$, desde que $||\mathbf{x}^{(0)} - \mathbf{p}|| < \hat{\delta}$. Além disto,

$$||\mathbf{x}^{(k)} - \mathbf{p}|| \leq \frac{n^2 M}{2}||\mathbf{x}^{(k-1)} - \mathbf{p}||^2_\infty, \text{ para cada } k \geq 1.$$

Usando este teorema, podemos mostrar que $\mathbf{A}(\mathbf{x}) = \mathbb{J}(\mathbf{F}(\mathbf{x}))$ onde $\mathbb{J}(\mathbf{F}(\mathbf{x}))$ é o *Jacobiano* de $\mathbf{F}(\mathbf{x})$ definido como:

$$\mathbb{J}(\mathbf{F}(\mathbf{x})) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}) \end{pmatrix} \tag{2.57}$$

Desta forma, o método de Newton para sistemas é definido por:

$$\mathbf{x}^{(k)} = \mathbf{G}(\mathbf{x}^(k-1)) = \mathbf{x}^(k-1) - \mathbb{J}^{-1}(\mathbf{x}^(k-1))\mathbf{F}(\mathbf{x}^(k-1)). \tag{2.58}$$

Este método é o *Método de Newton para sistemas não lineares* e em geral possui convergência quadrática, se o valor inicial for suficientemente próximo da solução e se $\mathbb{J}^{-1}(\mathbf{p})$ existe. A grande fragilidade do método de Newton está na necessidade de calcular o Jacobiano e ainda inverte-lo. Na prática a inversa do Jacobiano é evitada separando esta operação em dois passos:

1. Encontrar um vetor $\mathbf{Y}$ que satisfaça o sistema linear $\mathbb{J}(\mathbf{x}^(k-1))\mathbf{Y} = -\mathbf{F}(\mathbf{x}^(k-1))$.
2. Calcular $\mathbf{x}^{(k)}$ como sendo $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{Y}$.

Exemplo 27 *Vamos considerar o seguinte sistema não linear de equações:*

$$\begin{cases} 3x - \cos(yz) & = \frac{1}{2} \\ x^2 - 81(y+0.1)^2 + \sin(z) & = -1.06 \\ e^{-xy} + 20z & = \frac{10\pi - 3}{3} \end{cases}$$

Para aplicarmos o método de Newton, vamos considerar a função $F(x, y, z)$,

$$F(x,y,z) = \begin{pmatrix} 3x - \cos(yz) - \dfrac{1}{2} \\ x^2 - 81(y+0.1)^2 + \sin(z) + 1.06 \\ e^{-xy} + 20z - \dfrac{10\pi - 3}{3} \end{pmatrix},$$

com Jacobiano dado por:

$$\mathcal{J}(F(x,y,z)) = \begin{pmatrix} 3 & z\sin(yz) & y\sin(yz) \\ 2x & -162(0.1+y) & \cos(z) \\ -e^{-xy}y & -e^{-xy}x & 20 \end{pmatrix}$$

Se escolhemos o vetor inicial $Y_0 = (0.1\ \ , 0.1,\ \ -0.1)^T$ *então*

$$F(Y_0) = \begin{pmatrix} -1.19995 \\ -2.26983 \\ 8.4623 \end{pmatrix}, \qquad J(F(Y_0)) = \begin{pmatrix} 3 & 0.000999983 & -0.000999983 \\ 0.2 & -32.4 & 0.995004 \\ -0.099005 & -0.099005 & 20 \end{pmatrix}$$

Sabemos que $Y_k = Y_{k-1} - (J(F(Y_{k-1})))\,F(Y_{k-1})$, *ou* $Y_k = Y_{k-1} - G$, *onde* G *é um vetor encontrado pela resolução do sistema linear* $J(Y_0)G = F(Y_0)$, *isto é:*

$$\begin{pmatrix} 3 & 0.000999983 & -0.000999983 \\ 0.2 & -32.4 & 0.995004 \\ -0.099005 & -0.099005 & 20 \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} = \begin{pmatrix} -1.19995 \\ -2.26983 \\ 8.4623 \end{pmatrix}$$

Resolvendo este sistema por Crammer temos:

$$g_1 = \frac{det1}{det} = \frac{777.229}{-1943.71} = -0.39987$$

$$g_2 = \frac{det1}{det} = \frac{-156.533}{-1943.71} = 0.0805332$$

$$g_3 = \frac{det1}{det} = \frac{-819.312}{-1943.71} = 0.42152$$

onde:

$$det1 = \begin{vmatrix} -1.19995 & 0.000999983 & -0.000999983 \\ -2.26983 & -32.4 & 0.995004 \\ 8.46203 & -0.099005 & 20 \end{vmatrix} = 777.229$$

$$det2 = \begin{vmatrix} 3 & -1.19995 & -0.000999983 \\ 0.2 & -2.26983 & 0.995004 \\ -0.099005 & 8.46203 & 20 \end{vmatrix} = -156.533$$

$$det3 = \begin{vmatrix} 3 & 0.000999983 & -1.19995 \\ 0.2 & -32.4 & -2.26983 \\ -0.099005 & -0.099005 & 8.46203 \end{vmatrix} = -819.312$$

$$det = \begin{vmatrix} 3 & 0.000999983 & -0.000999983 \\ 0.2 & -32.4 & 0.995004 \\ -0.099005 & -0.099005 & 20 \end{vmatrix} = -1943.71$$

Desta forma, temos

$$Y_1 = Y_0 - G = \begin{pmatrix} 0.49987 \\ 0.0194668 \\ -0.52152 \end{pmatrix}$$

Seguindo o mesmo procedimento temos:

$$Y_0 = \begin{pmatrix} 0.1 \\ 0.1 \\ -0.1 \end{pmatrix}, \quad Y_1 = \begin{pmatrix} 0.49987 \\ 0.0194668 \\ -0.52152 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 0.500014 \\ 0.00158859 \\ -0.523557 \end{pmatrix},$$

$$Y_3 = \begin{pmatrix} 0.5 \\ 0.0000124448 \\ -0.523598 \end{pmatrix}, \quad Y_4 = \begin{pmatrix} 0.5 \\ 7.757857 \times 10^{-10} \\ -0.523599 \end{pmatrix} ...$$

EXERCÍCIOS

1. Encontre a raíz positiva de $f(x) = x^3 - e^{-x}$.

   (a) pelo método da bissecção

   (b) pelo método de Newton

   (c) pelo método da secante

   R: Esta função possui apenas um zero positivo $x = 0.77288$.

2. Encontre o valor da raíz cúbica de 75 com precisão de seis casas decimais, usando o método de Newton.

   (R: 4.217163).

3. Encontre uma raíz real de $f(x) = x^3 - 3.23x^2 - 5.54x + 9.84$ usando o método da secante com $x_0 = 0.9$ e $x_1 = 1.0$.

   R: 1.23.

4. A função $f(x) = x^2 - 2e^{-x}x + e^{-2x}$ possui uma raíz real múltipla. Encontre esta raíz usando o método de Newton modificado com $x_0 = 1$. Tente também aplicar o método de Newton tradicional com o mesmo chute inicial e compare os o número de iterações necessárias para obter a mesma precisão.

   R: 0.567143

5. Encontre as primeiras cinco raízes positivas de $f(x) = \tan x - 2x$. Cuidado, pois a tangente tende para mais e menos infinito muitas vezes no domínio que inclui estas raízes.

   R: $1.16556, 4.60422, 7.78988, 10.94994, 14.10172$.

6. Encontre todas as raízes de $f(x) = x^4 - 7.223x^3 + 13.447x^2 - 0.672x - 10.223$.

   R: -0.713967, 1.57251, 2.17178, 4.19268.

7. Encontre aproximações com precisão de 10^{-4} de todos os zeros reais de $f(x) = x^3 - 2x^2 - 5$ usando o método de Muller.

   R: $2.69065, -0.345324 \pm 1.31873i$.

8. $P(x) = 10x^3 - 8.3x^2 + 2.295x - 0.21141 = 0$ possui uma raíz $x = 0.29$. Use o método de Newton com aproximação inicial $x_0 = 0.28$ para tentar encontrar a raíz. O que acontece?

Problemas Computacionais
Bissecção

1. Escreva um programa teste do algoritmo da bissecção e teste com:

   (a) $x^{-1} - \tan x$, $[0, \pi/2]$

(b) $x^{-1} - 2^x$, $[0, 1]$

(c) $(x^3 + 4x^2 + 3x + 5)/(2x^3 - 9x^2 + 18x - 2)$, [0,4]

2. Encontre a raíz de

$$x^8 - 36x^7 + 546x^6 - 4536x^5 + 22449x^4 - 67284x^3 + 118124x^2 - 109584x + 40320 = 0$$

no intervalo $[5.5; 6.5]$. Troque -36 por -36.001 e repita.

Newton

1. Escreva um programa para resolver a equação $x = \tan x$. Encontre as raízes próximas de 4.5 e 7.7.

2. A equação $2x^4 + 24x^3 + 61x^2 - 16x + 1 = 0$ possui dois zeros perto de $0, 1$. Encontre-os pelo método de Newton.

3. Use o método de Newton com precisão dupla, para encontrar o zero negativo de $f(x) = e^x - 1.5 - \tan^{-1} x$. $(x_0 = -7)$

4. No exercício anterior, investigue a sensibilidade da raíz para perturbações no termo constante.

5. Programe o método de Newton com aritmética complexa e teste para:

(a) $f(z) = z + \sin z - 3$, $z_0 = 2 - i$

(b) $f(z) = z^4 + z^2 + 2 + 3i$, $z_0 = 1$

secante

1. Escreva um subprograma para o método da secante, assumindo que dois valores iniciais sejam dados. Teste esta rotina em:

(a) $\sin(x/2) - 1$

(b) $e^x - \tan x$

(c) $x^3 - 12x^2 + 3x + 1$

2. Use expansões em Taylor para $f(x + h)$ e $f(x + k)$, derive a seguinte aproximação para $f'(x)$:

$$f'(x) = \frac{k^2 f(x + h) - h^2 f(x + k) + (h^2 - k^2) f(x)}{(k - h)kh}$$

3. Programe e teste um refinamento do método da secante que usa a aproximação de $f'(x)$ dada no exercício acima. Isto é use esta aproximação na fórmula de Newton. Três pontos de inicialização são necessários, os dois primeiros arbitrários e o terceiro calculado pelo método da secante.

Zeros de Polinômios

1. Escreva um programa cuja entrada é os coeficientes de um polinômio p e um ponto específico z_0, e que produza como saída os valores de $p(z_0)$, $p'(z_0)$ e $p''(z_0)$ (use Horner). Teste o algorítimo calculando $p(4)$ para $p(z) = 3z^5 - 7z^4 - 5z^3 + z^2 - 8z + 2$.

2. Escreva uma rotina computacional para o método de Newton+Horner com deflação para encontrar todas as raízes de um polinômio. Teste a rotina para

$$p(x) = x^8 - 36x^7 + 546x^6 - 4536x^5 + 22449x^4 - 67284x^3 + 118124x^2 - 109584x + 40320 = 0$$

   As raízes corretas são $1, 2, 3, 4, 5, 6, 7, 8$. A seguir resolva a mesma equação fazendo o coeficiente de x^7 como 37. Observe como uma perturbação pequena pode ocasionar uma grande troca nas raízes. Assim as raízes são funções instáveis dos coeficientes.

Capítulo 3

Sistemas Lineares

Neste capítulo, vamos estudar métodos numéricos para a solução de um sistema linear do tipo $\mathbf{Ax} = \mathbf{b}$. Para isto vamos começar com um pouco de álgebra matricial.

3.1 Algebra Matricial

- Uma matriz é um arranjo retangular de números $A = [a_{ij}]$, com $1 \leq i \leq m$ linhas e $1 \leq j \leq n$ colunas.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & ... & a_{nn} \end{pmatrix}, \qquad \mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} & ... & a_{n1} \\ a_{12} & a_{22} & a_{32} & ... & a_{n2} \\ a_{13} & a_{23} & a_{33} & ... & a_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & a_{3n} & ... & a_{nn} \end{pmatrix} \tag{3.1}$$

- Transposta: $\mathbf{A}^T = [a_{ji}]$
- Simétrica: $\mathbf{A} = \mathbf{A}^T$
- $\lambda\mathbf{A} = [\lambda a_{ij}]$
- Considere $\mathbf{A}$ e $\mathbf{B}$ de mesma ordem assim: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$
- Considere: $\mathbf{A}_{m\times p}$ e $\mathbf{B}_{p\times n}$ $\mathbf{A}.\mathbf{B} = [\sum_{k=1}^{p} a_{ik}b_{kj}]$
- Dois sistemas $\mathbf{Ax} = \mathbf{b}$ e $\mathbf{Bx} = \mathbf{d}$ são ditos equivalentes se possuem a mesma solução.

3.1.1 Operações Elementares

Em métodos numéricos transformamos um sistema, através de operações elementares em outro mais simples de ser resolvido.

1. Troca de duas equações do sistema $E_i \to E_j$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$ Troca as segunda e terceira linhas

2. Multiplicação de uma equação por um número não nulo $\lambda E_i \leftrightarrow E_i$.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & 1 \end{pmatrix}$$ multiplica a segunda linha por λ

3. Somar de uma equação um múltiplo de outra equação $E_i - \lambda E_j \to E_i$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda & 1 \end{pmatrix}$$ multiplica a segunda linha por λ e soma na terceira

Teorema 22 *Se um sistema linear de equações que é obtido de outro através de um número finito de operações elementares então estes sistemas são equivalentes.*

3.1.2 Matrizes

- $\mathbf{I}_{n\times n}$ é a matriz identidade de ordem n, logo para qualquer $\mathbf{A}_{n\times n}$, $\mathbf{AI} = \mathbf{A} = \mathbf{IA}$.
- Se $\mathbf{AB} = \mathbf{I}$ então $\mathbf{B}$ é dita inversa pela direita de $\mathbf{A}$ e $\mathbf{A}$ é dita inversa pela esquerda de $\mathbf{B}$. Note que estas inversas não são necessariamente únicas, por exemplo:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ e & f \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Teorema 23 *Uma matriz quadrada pode possuir no máximo uma inversa pela direita.*

Prova:

Vamos supor que $\mathbf{AB} = \mathbf{I}$ *e que* $\mathbf{a}^{(j)}$ *representa sua j-ésima coluna.*
Assim:

$$\sum_{j=1}^{n} b_{jk}\mathbf{a}^{(j)} = \mathbf{I}^{(k)}, \text{ para } 1 \leq k \leq n \tag{3.2}$$

onde b_{jk} *são os elementos da matriz* $\mathbf{B}$*. Assim cada coluna de* $\mathbf{I}$ *é combinação linear das colunas de* $\mathbf{A}$ *e desta forma temos que as colunas de* $\mathbf{A}$ *formam uma base para* $\mathfrak{R}^n$ *logo os coeficientes de (5.1) são unicamente determinados.*

Teorema 24 *Se* $\mathbf{A}$ *e* $\mathbf{B}$ *são matrizes quadradas e* $\mathbf{AB} = \mathbf{I}$ *então* $\mathbf{BA} = \mathbf{I}$.
Prova:

Vamos supor que $\mathbf{AB} = \mathbf{I}$.
Vamos considerar uma matriz $\mathbf{C}$ *definida como:* $\mathbf{C} = \mathbf{BA} - \mathbf{I} + \mathbf{AB}$. *Multiplicando* $\mathbf{C}$ *pela esquerda por* $\mathbf{A}$ *temos:*

$$\mathbf{AC} = (\mathbf{AB})\mathbf{A} - \mathbf{A} + (\mathbf{AB}) = \mathbf{A} - \mathbf{A} + \mathbf{I} = \mathbf{I}$$

Assim C é uma inversa pela direita de $\mathbf{A}$, *então ao pelo teorema de unicidade,* $\mathbf{C} = \mathbf{B}$ *e assim*

$$\mathbf{B} = \mathbf{BA} - \mathbf{I} + \mathbf{B}, \text{ assim } \mathbf{BA} = \mathbf{I}$$

- Se $\mathbf{A}_{n\times n}$ e se existe $\mathbf{B}$ de forma que $\mathbf{AB} = \mathbf{I}$ então $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$ e neste caso $\mathbf{B}$ é dita inversa de $\mathbf{A}$ e $\mathbf{A}$ é dita não singular ou inversível.
- Se $\mathbf{A}$ é inversível então a solução do sistema $\mathbf{Ax} = \mathbf{b}$ única.
- Se uma matriz é inversível então existe uma seqüência de operações elementares tal que

$$\mathbf{E_m E_{m-1} ... E_2 E_1 A} = \mathbf{I}$$

Exemplo 28 .

Decompor a matriz em matrizes elementares e assim encontre sua inversa,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 3 \\ 2 & 4 & 7 \end{pmatrix}$$

Podemos ver que:

$$\begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 3 \\ 2 & 4 & 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

desta forma temos que a inversa da matriz $\mathbf{A}$ *é dada por:*

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 9 & -2 & -3 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$$

- Uma matriz é dita *definida positiva* se para qualquer vetor $\mathbf{x}$, com $\mathbf{x} \neq 0$, $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$.

  Exemplo 29 .

  Por exemplo, a matriz $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ *é definida positiva, pois:*

$$(x_1 \;\; x_2)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (2x_1^2 + x2x1 + 2x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 2x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2 + x_1^2 + x_2^2 \geq 0$$

- $\mathbf{x}^T\mathbf{A}\mathbf{x}$ é chamado de *forma quadrática*
- Se uma matriz $\mathbf{A}$ é positiva definida e simétrica, então todos os seus autovalores são reais e positivos.

Teorema 25 *As seguintes propriedades são equivalentes:*

1. *A inversa de* $\mathbf{A}$ *existe (é não singular).*
2. *O determinante de* $\mathbf{A}$ *é distinto de zero.*
3. *As linhas de* $\mathbf{A}$ *formam uma base de* $\Re^n$.
4. *As colunas de* $\mathbf{A}$ *formam uma base de* $\Re^n$.
5. *Como uma função de* $\Re^n$ *em* $\Re^n$, $\mathbf{A}$ *é injetiva.*
6. *Como uma função de* $\Re^n$ *em* $\Re^n$, *A é sobrejetiva.*
7. $\mathbf{A}\mathbf{x} = \mathbf{0}$ *implica que* $\mathbf{x} = \mathbf{0}$.
8. *Para qualquer* $\mathbf{b} \in \Re^n$ *existe um único* $\mathbf{x} \in \Re^n$ *de forma que* $\mathbf{A}\mathbf{x} = \mathbf{b}$.
9. $\mathbf{A}$ *é um produto de matrizes elementares.*
10. 0 *não é um autovalor de* $\mathbf{A}$.

3.1.3 Decomposição $\mathbf{A} = \mathbf{LU}$

Uma matriz $\mathbf{A}$ pode ser decomposta na forma $\mathbf{LU}$ se a eliminação Gaussiana pode ser feita sem troca de linhas. Podemos mostrar que isto é possível quando todos os menores principais são não singulares, isto é seus determinantes não são nulos. Neste caso:

- $\mathbf{A} = \mathbf{LU}$
- $\mathbf{Ax} = \mathbf{b}$ é o mesmo que resolver

  $$\mathbf{L}\underbrace{\mathbf{Ux}} = \mathbf{b}$$

  é equivalente a resolver dois sistemas triangulares,

  $$\mathbf{Lz} = \mathbf{b}, \quad \mathbf{Ux} = \mathbf{z}$$

- Algorítmo para proceder a decomposição $\mathbf{LU}$:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & ... & a_{nn} \end{pmatrix} = \begin{pmatrix} \ell_{11} & 0 & ... & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 & ... & 0 \\ \vdots & & \vdots & ... & \vdots \\ \ell_{(n-1)1} & \ell_{(n-1)2} & \dots & \ell_{(n-1)n} & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & ... & \ell_{nn} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & ... & u_{1n} \\ 0 & u_{22} & u_{23} & ... & u_{2n} \\ 0 & \ddots & u_{33} & ... & u_{3n} \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & ... & 0 & u_{nn} \end{pmatrix}$$

- Observe que a decomposição $\mathbf{A} = \mathbf{LU}$ não é única. As decomposições mais conhecida são as de Doolitle, Crout e Cholesky (esta última só aplicável para matrizes simétricas). A seguir vamos exemplificar as 3 decomposições.
- Na decomposição de Doolitle fazemos a diagonal principal de $\mathbf{A}$ unitária, isto é, $\ell_{ii} = 1$, por exemplo:

$$\mathbf{A} = \begin{pmatrix} 60 & 30 & 20 \\ 30 & 20 & 15 \\ 20 & 15 & 12 \end{pmatrix} = \mathbf{LU} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ \ell_{21}u_{11} & \ell_{21}u_{12} + u_{22} & \ell_{21}u_{13} + u_{23} \\ \ell_{31}u_{11} & \ell_{31}u_{12} + \ell_{32}u_{22} & \ell_{31}u_{13} + \ell_{32}u_{23} + u_{33} \end{pmatrix}$$

1. Usando os as igualdades da primeira linha: $u_{11} = 60$, $u_{12} = 30$ e $u_{13} = 20$
2. Usando as igualdades da primeira coluna: $u_{11}\ell_{21} = 30 \rightarrow \ell_{21} = 1/2$ e e $u_{11}\ell_{31} = 20 \rightarrow \ell_{31} = 1/3$.
3. Usando a igualdade da segunda linha: $u_{12}\ell_{21} + u_{22} = 20 - 30(1/2) = 5$ e $u_{13}\ell_{21} + u_{23} = 15 \rightarrow u_{23} = 5$
4. Usando a igualdade da segunda coluna: $u_{12}\ell_{31} + u_{22}\ell_{32} = \frac{1}{5}(15 - \frac{30}{3}) = 1$.
5. Finalmente, usando a igualdade da terceira linha: $u_{13}\ell_{31} + u_{23}\ell_{32} + u_{33} = 12 \rightarrow u_{33} = 12 - 20\frac{1}{3} - 5(1) = \frac{1}{3}$

6. Assim:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 60 & 30 & 20 \\ 0 & 5 & 5 \\ 0 & 0 & 1/3 \end{pmatrix}, \text{ Decomposição de Doolittle}$$

7. Na decomposição de Crout a diagonal principal de $\mathbf{U}$ é unitária, ou

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix} \begin{pmatrix} 60 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 60 & 0 & 0 \\ 30 & 5 & 0 \\ 20 & 5 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \text{ Crout}$$

8. Quando a matriz $\mathbf{A}$ é definida positiva, também temos a decomposição de Cholesky, onde $\mathbf{A} = \mathbf{L}\mathbf{L}^T$:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{1/3} \end{pmatrix} \begin{pmatrix} \sqrt{60} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & \sqrt{1/3} \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sqrt{60} & 0 & 0 \\ \sqrt{60}/3 & \sqrt{5} & 0 \\ \sqrt{60}/3 & \sqrt{5} & \sqrt{1/3} \end{pmatrix} \begin{pmatrix} \sqrt{60} & \sqrt{60}/2 & \sqrt{60}/3 \\ 0 & \sqrt{5} & \sqrt{5} \\ 0 & 0 & \sqrt{3}/3 \end{pmatrix}, \text{ Cholesky}$$

Teorema 26 *Se todos os* $\mathfrak{n}$ *menores principais*

$$\mathbf{A}_k = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{pmatrix}$$

de $\mathbf{A}$ *são não singulares, então* $\mathbf{A}$ *possui decomposição* $\mathbf{LU}$.

Teorema 27 (Cholesky) *Se* $\mathbf{A}$ *é uma matriz real simétrica e definida positiva então existe uma única fatoração* $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, *onde* $\mathbf{L}$ *é uma matriz triangular inferior cujos elementos da diagonal principal são positivos.*

Teorema 28 *Se todos os elementos pivôs a cada passo da eliminação Gaussiana são não nulos então existe a fatoração* $A = LU$.

3.2 Eliminação de Gauss

Na seção anterior estudamos uma versão abstrata da eliminação de Gauss através da fatoração $\mathbf{LU}$ de uma matriz. Nesta seção descrevemos a eliminação de Gauss tradicional e esta é relacionada a forma abstrata.

3.2.1 Decomposição de Gauss Básica

Vamos considerar o seguinte sistema:

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \begin{pmatrix} 12 \\ 34 \\ 27 \\ -38 \end{pmatrix} \tag{3.3}$$

- Subtraímos 2 vezes a equação 1 da equação 2.
- Subtraímos $\frac{1}{2}$ da equação 1 da terceira.
- Subtraímos -1 vezes a equação 1 da quarta.
- os números $m_{21} = 2$, $m_{31} = \frac{1}{2}$ e $m_{41} = -1$ são chamados de multiplicadores para o primeiro passo do processo de eliminação.
- O elemento 6 usado como divisor na formação de cada um destes multiplicadores é chamado de **Pivô**.

E assim obtemos:

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \begin{pmatrix} 12 \\ 10 \\ 21 \\ -26 \end{pmatrix} \tag{3.4}$$

Neste primeiro estágio, chamamos a linha 1 de **linha do pivô**. No próximo estágio do processo a linha 2 é usada como linha do pivô.

- O elemento pivô é -4.
- Subtraímos 3 vezes a a segunda linha da terceira.
- Subtraímos $\frac{-1}{2}$ vezes a segunda linha da quarta.
- Os multiplicadores são $m_{32} = 3$ e $m_{42} = \frac{-1}{2}$.

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \begin{pmatrix} 12 \\ 10 \\ -9 \\ -21 \end{pmatrix} \tag{3.5}$$

O último estágio consiste em:

- Subtraímos 2 vezes a terceira linha da quarta.
- O pivô é 2

- O multiplicador é $m_{43} = 2$.

O sistema resultante é:

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \begin{pmatrix} 12 \\ 10 \\ -9 \\ -3 \end{pmatrix} \tag{3.6}$$

Este sistema é triangular superior e equivalente ao sistema original, isto é suas soluções são idênticas. O sistema final é resolvido por retro-substituição e a solução é:

$$\mathbf{x} = \begin{pmatrix} 1 \\ -3 \\ -2 \\ 1 \end{pmatrix}$$

Os multiplicadores usados para transformar o sistema podem ser mostrados como uma matriz triangular inferior:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{pmatrix} \tag{3.7}$$

Note que cada multiplicador está no lugar de um elemento 0 da matriz o qual ele foi o responsável por criar. A matriz coeficiente do sistema final é triangular superior $\mathbf{U}$,

$$\mathbf{U} = \begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{pmatrix} \tag{3.8}$$

Estas duas matrizes fornecem a decomposição $\mathbf{LU}$ de $\mathbf{A}$, a matrix coeficiente do sistema original. Assim,

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ \frac{1}{2} & 3 & 1 & 0 \\ -1 & -\frac{1}{2} & 2 & 1 \end{pmatrix} \begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{pmatrix} \tag{3.9}$$

Para podermos descrever formalmente o processo do algoritmo de eliminação Gaussiana, interpretamos como sendo uma sucessão de $n-1$ passos maiores que resulta em uma sucessão de matrizes que vamos chamar de

$$\mathbf{A} = \mathbf{A}^{(1)} \to \mathbf{A}^{(2)} \to ... \to \mathbf{A}^{(n)}$$

Na conclusão do k-ésimo passo construímos a matriz $\mathbf{A}^{(k)}$,

$$\left(\begin{array}{ccc|ccccc} a_{1,1}^{(k)} & \ldots & a_{1,k-1}^{(k)} & a_{1,k}^{(k)} & \ldots & a_{1,j}^{(k)} & \ldots & a_{1,n}^{(k)} \\ & \ddots & & \vdots & & \vdots & & \vdots \\ 0 & & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \ldots & a_{k-1,j}^{(k)} & \ldots & a_{k-1,n}^{(k)} \\ \hline 0 & \ldots & 0 & a_{k,k}^{(k)} & \ldots & a_{k,j}^{(k)} & \ldots & a_{k,n}^{(k)} \\ \hline 0 & \ldots & 0 & a_{k+1,k}^{(k)} & \ldots & a_{k+1,j}^{(k)} & \ldots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{i,k}^{(k)} & \ldots & a_{i,j}^{(k)} & \ldots & a_{i,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{n,k}^{(k)} & \ldots & a_{n,j}^{(k)} & \ldots & a_{n,n}^{(k)} \end{array}\right)$$

- Para obtemos $\mathbf{A}^{(k+1)}$ de $\mathbf{A}^{(k)}$: Para produzir zeros abaixo do k-ésimo pivô $a_{kk}^{(k)}$ subtraímos múltiplos da linha k das linhas abaixo dela, assim as linhas $1, 2, ..., k$ não são alteradas e:

$$a_{i,j}^{(k+1)} = \begin{cases} a_{i,j}^{(k)} & \text{se} \quad i \leq k \\ a_{i,j}^{(k)} - \left(a_{i,k}^{(k)} / a_{k,k}^{(k)}\right) a_{k,j}^{(k)} & \text{se} \quad i \geq k+1 \text{ e } j \geq k+1 \\ 0 & \text{se} \quad i \geq k+1 \text{ e } j \leq k+1 \end{cases} \tag{3.10}$$

- Depois de todos os passos feitos temos uma matriz triangular superior $\mathbf{U} = \mathbf{A}^{(n)}$.

- Com os multiplicadores definimos uma matriz triangular inferior $\mathbf{L}$, definida por:

$$\ell_{i,k} = \begin{cases} \left(a_{i,k}^{(k)} / a_{k,k}^{(k)}\right) & \text{se} \quad i \geq k+1 \\ 1 & \text{se} \quad i = k \\ 0 & \text{se} \quad i \leq k-1 \end{cases} \tag{3.11}$$

Assim $\mathbf{A} = \mathbf{LU}$ é a eliminação Gaussiana comum. Note que isto vale se não tivermos nenhum **pivô nulo**.

Teorema 29 *Se todos os elementos pivô, $a_{k,k}^{(k)}$ são não nulos no processo descrito, então $\mathbf{A} = \mathbf{LU}$.*

ALGORITMO 7 (Eliminação de Gauss com substituição retroativa:) *Resolve o sistema linear $\mathbf{Ax} = \mathbf{b}$ de ordem n.*
.ENTRADA: número de equações n; Matriz expandida $\mathbf{A} = (a_i j)$, $1 \leq i \leq n$ e $1 \leq j \leq n+1$ $(a_{i,n+1} = b_i)$.
.SAÍDA: Vetor solução $\mathbf{x} = (x_i)$ ou mensagem de erro
.Passo 1: Para $i = 1, ...n-1$ siga os passos 2-4. (Processo de eliminação.)
. Passo 2: Seja p o menor inteiro com $i \leq p \leq n$ e $a_{pi} \neq 0$.
. Se nenhum p pode ser encontrado, então:

. *SAÍDA: (A solução d problema não é única)*
. *PARE*
. *Passo 3: Se* $p \neq i$ *então execute* $(E_p) \leftrightarrow (E_i)$.
. *Passo 4: Para* $j = 1, ..., n$ *siga os passos 5 e 6.*
. *Passo 5: Faça* $m_{ji} = a_{ji}/a_{ii}$.
. *Passo 6: Execute* $(E_j - m_{ji}E_i) \leftrightarrow (E_j)$;
.*Passo 7: Se* $a_{nn} = 0$ *então:*
. *SAÍDA: (Não existe solução única);*
. *PARE*
.*Passo 8: Faça* $x_n = a_{n,n+1}/a_{nn}$. *(Início da substituição retroativa.)*
.*Passo 9: Para* $i = n-1, n-2, ..., 1$ *faça* $x_i = \left[a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j a_{ii}\right.$.
.*Passo 10: SAÍDA* $(x_1, x_2, ..., x_n)$; *(O procedimento foi completado com sucesso.)* .
. *PARE*

No algoritmo anterior podemos armazenar os multiplicadores m_{ij} nos endereços de a_{ij} que já tenham o valor zero para cada $i = 1, 2, ..., n-1$ e $j = i+1, i+2, ..., n$. Assim, $\mathbf{A}$ é substituída pelos multiplicadores, abaixo da diagonal principal, e por $\mathbf{A}^{(n)}$ acima e sobre a diagonal principal. As quantidades de tempo computacional e erros de arredondamento dependem do número de operações aritméticas de ponto flutuante na rotina. Em geral o tempo necessário aritméticas necessário para proceder a multiplicação ou uma subtração é aproximadamente o mesmo e é consideravelmente maior que o tempo necessário para executar uma adição ou uma subtração. Desta forma o custo computacional é normalmente medido em número de multiplicações e/ou divisões.

- Nenhuma operação até o passo 3.
- No passo 5, são necessárias $(n-i)$ divisões e no passo 6, $m_j i$ é multiplicado a cada termo de E_i, resultando $(n-i)(n-i+1)$ multiplicações.
  Mult/Div $(n-i) + (n-i)(n-i+1) = (n-i)(n-i+2)$. Como estas multiplicações são efetuadas para cada,
  $$\sum_{i=1}^{n-1}(n-i)(n-i+2) = (n^2+2n)\sum_{i=1}^{n-1} 1 - 2(n+1)\sum_{i=1}^{n-1} i^2 = \frac{2n^3+3n^2-5n}{6}$$
- Nos passos 8 e 9 da substituição retroativa temos uma divisão, no passo 8 temos $n-i$ multiplicações e uma divisão para cada $i = n-1, ..., 1$, isto é
  $$1 + \sum_{i=1}^{n-1}((n-i)+1) = \frac{n^2+n}{2}$$
- Total de multiplicações e/ou divisões
  $$\frac{2n^3+3n^2-5n}{6} + \frac{n^2+n}{2} = \frac{n^3}{3} + n^2 - \frac{n}{3} \approx \frac{n^3}{3}$$
  para n grande.

3.2.2 Pivotamento

O algoritmo de eliminação de Gauss descrito acima não é satisfatório pois pode falhar em sistemas muito simples de serem resolvidos.
Por exemplo:

- No sistema

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

  o algoritmo descrito acima falha, pois o primeiro pivô é nulo. A dificuldade persiste no seguinte sistema quando $\varepsilon << 1$,

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

  Pois quando aplicado o algoritmo de Gauss teremos o seguinte sistema triangular superior:

$$\begin{pmatrix} \varepsilon & 1 \\ 0 & 1-\varepsilon^{-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2-\varepsilon^{-1} \end{pmatrix}$$

  cuja solução é dada por:

$$x_2 = \frac{2-\varepsilon^{-1}}{1-\varepsilon^{-1}} \approx 1 \quad \text{e} \quad x_1 = (1-x_2)\varepsilon^{-1} \approx 0$$

  Devemos notar que a solução exata deste sistema é dada por:

$$x_2 = \frac{2-\varepsilon^{-1}}{1-\varepsilon^{-1}} \approx 1 \quad \text{e} \quad x_1 = \frac{1-x_2}{\varepsilon} = 1 - \frac{2-\varepsilon^{-1}}{(1-\varepsilon^{-1})\varepsilon} = \frac{-1}{\varepsilon-1} \approx 1$$

  Este efeito é devido a aritmética de ponto flutuante e o fato de $\varepsilon \approx 0$.

- Outro exemplo mostra que não é exatamente o fato de a_{11} ser pequeno que acarreta o erro. Mas sim o fato de a_{11} ser muito menor relativamente aos elemento de sua linha. Vamos considerar o sistema equivalente ao anterior dado por:

$$\begin{pmatrix} 1 & \varepsilon^{-1} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon^{-1} \\ 2 \end{pmatrix}$$

  Neste caso ε^{-1} é um número muito grande. A aplicação do algoritmo simples de Gauss resulta:

$$\begin{pmatrix} 1 & \varepsilon^{-1} \\ 0 & 1-\varepsilon^{-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon^{-1} \\ 2-\varepsilon^{-1} \end{pmatrix}$$

  e sua solução é dada por:

$$x_2 = \frac{2-\varepsilon^{-1}}{1-\varepsilon^{-1}} \approx 1 \quad \text{e} \quad x_1 = (\varepsilon^{-1} - \varepsilon^{-1}x_2) \approx 0$$

  Vemos assim que as dificuldades causadas neste sistema irão desaparecer se a ordem das equações for simplesmente trocada.

Resumindo, sabemos que quando o candidato à pivô é nulo devemos trocar as linhas do sistema para podermos proceder o algorítmo de Gauss. Porém, também vimos que algumas vezes a troca de linhas também é necessária mesmo que o pivô não seja nulo.

Aqui vamos considerar dois tipos de pivotamento, o chamado de pivotamento parcial que escolhe para pivô o maior elemento em módulo da coluna considerada e o pivotamento escalonado que antes de escolher o pivô, compara sua grandesa com um fator de escala.

3.2.3 Pivotamento Parcial

Se $a_{kk}^{(k)}$ é o pivô e ele é pequeno em comparação com os outros elementos abaixo dele em sua coluna $a_{jk}^{(k)}$, o multiplicador

$$m_{jk} = \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}} >> 1$$

e assim um erro de arredondamento é introduzido no cálculo e. desta forma, quando procedemos a retrosubstituição para determonarmos x_k, temos:

$$x_k = \frac{a_{j,n+1}^{(k)} - \sum_{j=k+1}^{n} a_{jk}^{(k)}}{a_{kk}^{(k)}}$$

E o erro de arredondamento do numerador será drasticamente incrementado com a divisão por um número muito pequeno, no caso $a_{kk}^{(k)}$

Exemplo 30 *Vamos considerar neste exemplo o uso de aritmética de 4 algarísmos e arredondamento*

$$\begin{cases} 0.003000x_1 + 59.14x_2 = 59.17 \\ 5.291x_1 - 6.130x_2 = 46.78 \end{cases}$$

A solução exata deste sistema é $x_1 = 10.00$ *e* $x_2 = 1.000$.

O primeiro pivô é um número muito pequeno ("perigo"). O multiplicador é

$$m_{21} = \frac{5.291}{0.003000} = 1763.666... \approx 1764.$$

Assim procedendo a operação elementar

$$L_2 - m_{21}L_1 \to L2$$

$$\begin{cases} 0.003000x_1 + 59.14x_2 = 59.17 \\ - 104300.x_2 = -104400. \end{cases}$$

e desta forma temos que

$$x_2 = 1.001$$

O que é uma boa aproximação da solução para uma máquina que opera com aritmética de 4 dígitos significativos. Agora para encontrarmos x_1 *temos:*

$$x_1 = \frac{59.17 - (59.14)(1.001)}{0.003000} = \frac{59.17 - 59.20}{0.003000} = -10.00 \neq 10.00!!!!!$$

Podemos evitar este problema procedendo ao pivotamento parcial (troca de linhas)

Exemplo 31 *Vamos considerar neste exemplo o uso de aritmética de 4 algarísmos e arredondamento e pivotamento parcial*

$$\left\{\begin{array}{rcrcl} 0.003000x_1 & + & 59.14x_2 & = & 59.17 \\ 5.291x_1 & - & 6.130x_2 & = & 46.78 \end{array}\right.$$

Novamente a solução exata deste sistema é $x_1 = 10.00$ *e* $x_2 = 1.000$.

Escolhemos para primeiro pivô $\max\{0.003000, 5.291\} = 5.291$ *assim*

$$L1 \leftrightarrow L2$$

$$\left\{\begin{array}{rcrcl} 5.291x_1 & - & 6.130x_2 & = & 46.78 \\ 0.003000x_1 & + & 59.14x_2 & = & 59.17 \end{array}\right.$$

Assim

$$m21 = \frac{0.003000}{5.291} = 0.0005670$$

e procedendo a operação elementar

$$L_2 - m_{21}L_1 \to L2$$

$$\left\{\begin{array}{rcrcl} 5.291x_1 & + & 6.130x_2 & = & 46.78 \\ & - & 59.14x_2 & = & 59.14 \end{array}\right.$$

e desta forma temos que

$$x_2 = \frac{59.14}{59.14} = 1.000$$

e

$$x_1 = \frac{46.78 + 6.130x_2}{5.291} = \frac{52.91}{5.291} = 10.00$$

Na maior parte dos casos este procedimento basta, porém existem situações onde este procedimento é inadequado. Algumas vezes iremos precisar do pivotamento escalonado. O efeito do escalonamento é garantir que o maior elemento do linha possua magnitude relativa de uma unidade antes de compararmos os elementos da coluna para escolha do pivô e possível troca de linhas. A troca apropriada de linhas é dada pela escolha (s_p é o pésimo elemento do vetor de escala $\mathbf{s}$)

$$\frac{|a_{pi}|}{s_p} = \max_{k=i,i+1,\ldots,n} \frac{|a_{ki}|}{s_k}$$

Neste caso procedemos a troca

$$L_i \leftrightarrow L_p$$

Observamos que o vetor de escala é calculado uma única vez.

Exemplo 32 *Pivotamento escalonado Vamos considerar o seguinte sistema linear (aritmética de 3 dígitos e arredondamento):*

$$\begin{cases} 2.11x_1 & - & 4.21x_2 & + & 0.921x_3 & = & 2.01 \\ 4.01x_1 & + & 10.2x_2 & - & 1.10 & = & -3.09 \\ 1.09x_1 & + & 0.987x_2 & + & 0.832x_3 & = & 4.21 \end{cases}$$

O vetor de escala é

$$\mathbf{s} = \{4.21, 10.2, 1.09\}$$

Assim para escolhermos o primeiro pivô devemos comparar as seguintes grandezas:

$$\frac{|a_{11}|}{s_1} = 0.501 \qquad \frac{|a_{21}|}{s_2} = 0.393, \qquad \frac{|a_{31}|}{s_3} = 1$$

Assim devemos fazer

$$L_1 \leftrightarrow L_3$$

$$\left.\begin{array}{ccc|c} 1.09 & 0.987 & 0.832 & 4.21 \\ 4.01 & 10.2 & -1.10 & -3.09 \\ 2.11 & -4.21 & 0.921 & 2.01 \end{array}\right| \qquad \begin{array}{l} L_2 - 3.68L1 \to L_2 \quad m_{21} = 3.68 \\ L_3 - 1.94L3 \to L_3 \quad m_{31} = 1.94 \end{array}$$

Agora temos $\mathbf{s} = \{1.09, 10.2, 4.21\}$

$$\left.\begin{array}{ccc|c} 1.09 & 0.987 & 0.832 & 4.21 \\ 0 & 6.57 & -4.18 & -18.6 \\ 0 & -6.12 & -0.689 & -6.16 \end{array}\right| \qquad \begin{array}{r} L_2 \leftrightarrow L_3 \\ L_3 - 1.94L3 \to L_3 \end{array}$$

$$\left.\begin{array}{ccc|c} 1.09 & 0.987 & 0.832 & 4.21 \\ 0 & -6.12 & -0.689 & -6.16 \\ 0 & 6.57 & -4.18 & -18.6 \end{array}\right| \qquad \begin{array}{r} m_{32} = \frac{6.57}{-6.12} = -1.07 \\ L_3 - 1.07L2 \to L_3 \end{array}$$

Assim temos:

$$\left.\begin{array}{ccc|c} 1.09 & 0.987 & 0.832 & 4.21 \\ 0 & -6.12 & -0.689 & -6.16 \\ 0 & 0.02 & -4.92 & -25.2 \end{array}\right| \qquad \text{note que p/ arit 3 dígitos } 0.02 \approx 0.00$$

e assim usando retrosubstituição:

$$x_3 = 5.12, \qquad x_2 = 0.430 \qquad x_1 = -0.431$$

3.2.4 Pivotamento Escalonado

O *pivotamento* escalonado incorpora a troca de equações no sistema sempre que necessário, porém este leva em conta a grandeza relativa de cada "candidato"de pivô com os outros elementos de sua própria linha. Assim na fase de fatoração do algoritmo de Gauss, iniciamos o processo escolhendo o fator de escala de cada linha, isto é

$$s_i = \max_{1 \le j \le n} |a_{ij}| = \max\left\{|a_{i1}|, |a_{i2}|, ..., |a_{in}|\right\}$$

Estes valores ficam acumulados em um vetor que vamos chamar de **s**. Computacionalmente os pivôs são escolhidos, mas as linhas não são trocadas fisicamente na memória do computador. Ao invés disto as trocas são armazenadas em um vetor chamado de **p**. Iniciamos o processo de fatoração selecionando a linha do pivô como aquela onde o valor $|a_{i1}/si$ é maior. O índice i será chamado de de p_1 e definirá a primeira linha da matriz de permutação. Assim $|a_{p_i1}|/spi \geq |a_{i1}|/si$ para $1 \leq i \leq n$. Uma vez que p_1 foi determinado, subtraímos apropriadamente múltiplos da linha p_1 das outras linhas em ordem de proceder zeros na primeira coluna de **A**. É claro que a linha p_1 continua intocada até o fim da fatoração. Aqui começamos a armazenar os elementos p_i no vetor **p**. Agora selecionamos um índice j para o qual $|a_{p_j1}/s_{pj}|$ é o máximo e trocamos p_1 com p_j no vetor **p**. O passo de eliminação real envolve subtrair $(a_{p_i1}/a_{p_11}$ vezes a linha p_1 da linha p_i para $2 \leq i \leq n$. O processo continua desta forma. Vamos exemplificar este processo:

Exemplo 33 Pivotamento escalonado

Vamos considerar a matriz

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ 3 & -2 & 1 \end{pmatrix}$$

Inicialmente o vetor $\mathbf{p} = (1, 2, 3)$ *e* $\mathbf{s} = 6, 8, 3$. *Para selecionar o primeiro pivô, olhamos para* $\{2/6, 1/8, 3/3\}$. *O maior destes números corresponde a terceira linha, logo o primeiro pivô será* 3. *Isto será "dito"pelo vetor* **p**, *que agora será* $\mathbf{p} = (3, 2, 1)$. *Agora múltiplos da linha 3 serão subtraídos das linhas 2 e 1 produzindo zeros na primeira coluna. O resultado é:*

$$\begin{pmatrix} [\frac{2}{3}] & \frac{13}{3} & -\frac{20}{3} \\ [\frac{1}{3}] & -\frac{16}{3} & \frac{23}{3} \\ 3 & -2 & 1 \end{pmatrix}$$

Os elementos localizados em a_{11} *e* a_{21} *são os multiplicadores. No próximo passo, selecionamos novamente a linha do pivô. Para isto comparamos os números* $|a_{p_22}|/s_{p_2}$ *e* $|a_{p_32}|/s_{p_3}$. *A primeira desta razões é* $(16/3)/8$ *e a segunda* $(13/3)/6$. *Logo* $j = 3$, *e trocamos as linhas* p_2 *com* p_3. *Assim um múltiplo da linha* p_2 *é subtraída da linha* p_3. *O resultado é* $p = (3, 1, 2)$ *e*

$$\begin{pmatrix} [\frac{2}{3}] & \frac{13}{3} & -\frac{20}{3} \\ [\frac{1}{3}] & [-\frac{16}{3}] & -\frac{7}{13} \\ 3 & -2 & 1 \end{pmatrix}$$

O último multiplicador é então guardado em a_{22}.

Se as linhas da matriz original **A** *foram trocadas de acordo ao vetor de permutação* **p**, *então teríamos a decomposição* **LU** *de* **A**. *Assim temos:*

$$\mathbf{PA} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{1}{3} & -\frac{16}{3} & 1 \end{pmatrix} \begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{13}{3} & -\frac{20}{3} \\ 0 & 0 & -\frac{7}{13} \end{pmatrix} = \begin{pmatrix} 3 & -2 & 1 \\ 2 & 3 & -6 \\ 1 & -6 & 8 \end{pmatrix}$$

onde:

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad e \quad \mathbf{A} = \begin{pmatrix} 2 & 3 & -6 \\ 1 & -6 & 8 \\ 3 & -2 & 1 \end{pmatrix}$$

A matriz permutação $\mathbf{P}$ *é obtida do vetor permutação* $\mathbf{p}$*, isto é* $\mathbf{P}$ *é obtido permutando as linhas da matriz identidade* $\mathbf{I}$ *de acordo com os elemento do vetor* $\mathbf{p}$*.*

A seguir, apresentamos um algoritmo que faz a fase de fatoração da eliminação de Gauss usando a técnica de pivotamento escalonado:

ALGORITMO 8 .
Entrada*:* $n, (a_{ij})$
Para*:* $i = 1$ **até** n **faça**

$p_i \leftarrow i$
$s_i \leftarrow \max_{i \leq j \leq n} |a_{ij}|$

Fim do para
Para $k = 1$ **até** $n - 1$ **faça***:*

Se $j \geq k$ **então selecione**

$|a_{p_j,k}|/s_{p_j} \geq |a_{p_i,k}|/s_{p_i}$ *para* $i = k, k-1, ..., n$

$p_k \leftrightarrow p_j$
Para $i = k + 1$ **até** n **faça**

$z \leftarrow a_{p_i k}/a_{p_k k}$; $a_{p_i k} \leftarrow z$
Para $j = k + 1$ **até** n **faça**

$z \leftarrow a_{p_i j}/a_{p_k k}$; $a_{p_j k} \leftarrow z$
Para $j = k + 1$ *até* n **faça**

$a_{p_i j} \leftarrow ap_i j - z a_{p_k j}$
Fim do faça
Fim do faça

Fim do faça
Saída $(a_{ij}), (p_i)$

Note que neste algorítmo os multiplicadores são armazenados na matriz $\mathbf{A}$ no lugar dos "zeros"que deveriam aparecer no processo de eliminação. Vamos usar este procedimento no próximo exemplo.

Exemplo 34 *Vamos considerar, com 3 dígitos significativos, o sistema*

$$\begin{pmatrix} 2.11 & -4.21 & 0.921 \\ 4.01 & 10.2 & -1.1 \\ 1.09 & 0.987 & 0.832 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2.01 \\ -3.09 \\ 4.21 \end{pmatrix}$$

O vetor

$$S = \{4.21, 10.2, 1.09\}$$

. Para escolhermos o pivô, analisamos que

$$|a_{11}|/s_1 = 0.501, \quad |a_{21}|/s_2 = 0.393, \quad |a_{31}|/s_3 = 1.00$$

assim fazemos a troca

$$L_1 \longleftrightarrow L_3$$

				Vetor S	
1.09	0.987	0.832	4.21	1.09	$L_2 - 3.68L_1 \to L_2$
4.01	10.2	−1.1	−3.09	10.2	$L_3 - 1.94L_1 \to L_3$
2.11	−4.21	0.921	2.01	4.21	$m_{21} = 3.68 \quad m_{31} = 1.94$
1.09	0.987	0.832	4.21	1.02	
0	6.57	−4.18	−18.6	10.2	
0	−6.12	−0.689	−6.16	4.21	

Agora para escolher o novo pivô comparando:

$$\frac{|a_{22}|}{s_2} = \frac{6.57}{10.2} = 0.644 \quad e \quad \frac{|a_{32}|}{s_3} = \frac{6.12}{4.21} = 1.45$$

Temos que trocar as linhas 2 e 3:

$$L_2 \longleftrightarrow L_3$$

assim$m_{31} = 3.68$ *e* $m_{21} = 1.94$ *e*

				Vetor S	
1.09	0.987	0.832	4.21	1.09	$L_3 - (-1.07)L_2 \to L_3$
0	−6.12	−0.689	−6.16	4.21	$m_{32} = 6.57/(-6.12) = -1.07$
0	0.02	−4.92	−25.2	10.2	

Note que o valor $0.02 \approx 0$ *aparece por estarmos usando apenas 3 dígitos significativos. Resolvendo o sistema por retro-substituição temos:*

$$x_3 = \frac{-25.2}{-4.92} = 5.12, \quad x_2 = \frac{-6.16 + 6.689x_3}{-6.12} = 0.430, \quad x_1 = \frac{4.21 - 0.987x_2 - 0.832x_3}{1.09} = -0.431$$

3.3 Técnicas Iterativas em Álgebra Matricial

Para discutirmos os erros numéricos envolvidos em uma sequência de vetores que aproxima a solução de um sistema de equações, precisamos uma medida de distância entre vetores n-dimensionais.

Definição 9 *Em um espaço vetorial* V *, uma norma é uma função* $||.||$ *definida de* V *em um conjunto de números reais não negativos que obedecem as seguintes propriedades:*

1. $||x|| > 0$, *se* $x \neq 0$, $x \in V$

2. $||\lambda x|| = |\lambda|||x||$, *se* $\lambda \in \mathbb{R}$, $x \in V$

3. $||x + y|| \leq ||x|| + ||y||$, *se* x, $y \in V$ *(Desigualdade do triângulo)*

As normas mais familiares sobre $\mathbb{R}^n$ são:

1. A norma Euclideana ℓ_2 definida por: $||\mathbf{x}||_2 = (\sum_{i=1}^{n} x_i^2)^{1/2}$, onde $\mathbf{x} = (x_1, x_2, ..., x_n)^T$

2. A norma ℓ_1 definida por:$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$, onde $\mathbf{x} = (x^1, x^2, ..., x^n)^T$

3. A norma do máximo ℓ_∞, definida por: $||\mathbf{x}||_\infty = \max_{1 \leq i \leq n} |xi|$, onde $\mathbf{x} = (x^1, x^2, ..., x^n)^T$

Norma Matricial

Podemos definir normas matriciais apenas sujeitas a satisfazerem as propriedades $(1) - (3)$, mas preferimos escolher aquelas que também estejam relacionadas com as normas matriciais descritas anteriormente. Assim se uma norma vetorial $||.||$ está determinada, a norma matricial associada a ela é definida por:

$$||\mathbf{A}|| = \sup \{||\mathbf{Au}|| \text{ tal que } \mathbf{u} \in \mathbb{R}^n, ||\mathbf{u}|| = \mathbf{1}\} \tag{3.12}$$

Podemos demonstrar que:

Teorema 30 *Se* $||.||$ *é uma norma sobre* $\mathbb{R}^n$*, então a equação (3.12) define uma norma sobre o espaço linear de todas as matrizes quadradas de ordem* $n \times n$*.*

Teorema 31 *Se a norma* $||.||$ *é a norma do máximo definida, então ela está associada a norma matricial*

$$||\mathbf{A}||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \tag{3.13}$$

Deste teorema resulta que $||\mathbf{I}|| = 1$ e $||\mathbf{AB}|| \leq ||\mathbf{A}||||\mathbf{B}||$.

Uma outra norma importante é a norma Euclideana isto é a norma ℓ_2 também chamada de norma espectral, definida por:

$$||\mathbf{A}||_2 = \sup_{||\mathbf{x}||_2 = 1} ||\mathbf{Ax}||_2$$

É possível demonstrar que

$$||\mathbf{A}||_2 = \max_{1 \leq i \leq n} |\sigma_i| = \sqrt{\rho(\mathbf{A}^T\mathbf{A})}$$

onde σ_i ao os valores singulares de $\mathbf{A}$ e $\rho(\mathbf{A}^T\mathbf{A})$ é chamado de raio espectral de $\mathbf{A}^T\mathbf{A}$ e é definido como sendo o maior autovalor da matriz $\mathbf{A}^T\mathbf{A}$.

Autovalores e Autovetores

Uma matriz quadrada associa um conjunto de vetores n-dimensionais em si mesmo. Neste caso existem certos vetores não nulos $\mathbf{x}$ que são paralelos à $\mathbf{Ax}$, isto é existe uma constante λ de forma que $\mathbf{Ax} = \lambda\mathbf{x}$. Veremos que existe uma relação muito próxima entre estes números λ e a convergência de métodos iterativos.

Definição 10 *Se* $\mathbf{A}$ *é uma matriz quadrada chamamos o determinante* $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$ *de* **polinômio característico** *de* $\mathbf{A}$.

Definição 11 *Os zeros do polinômio característico de* $\mathbf{A}$ *são chamados de* **autovalores** *de* $\mathbf{A}$.

Definição 12 *O vetor* $\mathbf{x} \neq \mathbf{0}$, *de forma que* $\mathbf{Ax} = \lambda\mathbf{x}$ *é chamado de* **autovetor** *de* $\mathbf{A}$ *correspondente ao autovalor* λ.

Definição 13 *Chamamos* $\rho(\mathbf{A})$ *de* **raio espectral** *de* $\mathbf{A}$ *e definimos por,* $\rho(\mathbf{A}) = \max|\lambda|$.

Exemplo 35 *Como exemplo, vamos considerar a matriz*

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \\ -1 & 1 & 1 \end{pmatrix},$$

$$p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) = \det\begin{pmatrix} 1-\lambda & 0 & 2 \\ 0 & 1-\lambda & -1 \\ -1 & 1 & 1-\lambda \end{pmatrix} = (1-\lambda)^3+2(1-\lambda)+(1-\lambda) = (1-\lambda)(\lambda^2-2\lambda+4).$$

Desta forma, resolvendo $p(\lambda) = 0$, *os autovalores da matriz* $\mathbf{A}$ *são* $\lambda_1 = 1$, $\lambda_2 = 1 + i\sqrt{3}$ *e* $\lambda_2 = 1 - i\sqrt{3}$.

1. *Calculando autovetor correspondente ao autovalor* λ_1 *temos:*

$$\mathbf{A} - \lambda_1\mathbf{I} = \mathbf{0} \Leftrightarrow \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Leftrightarrow x_3 = 0 \text{ e } x_1 = x_2 \Leftrightarrow V_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

2. *Calculando autovetor correspondente ao autovalor* λ_2 *temos:*

$$\mathbf{A} - \lambda_2\mathbf{I} = \mathbf{0} \Leftrightarrow \begin{pmatrix} -i\sqrt{3} & 0 & 2 \\ 0 & -i\sqrt{3} & -1 \\ -1 & 1 & -i\sqrt{3} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Leftrightarrow x_1 = -\frac{2i}{\sqrt{3}}x_3 \text{ e } x_2 = x_3\frac{i}{\sqrt{3}} \Leftrightarrow V_2 = \begin{pmatrix} -2i/\sqrt{3} \\ i/\sqrt{3} \\ 1 \end{pmatrix}$$

3. Calculando autovetor correspondente ao autovalor λ_3 temos:

$$\mathbf{A} - \lambda_2\mathbf{I} = \mathbf{0} \Leftrightarrow \begin{pmatrix} i\sqrt{3} & 0 & 2 \\ 0 & i\sqrt{3} & -1 \\ -1 & 1 & i\sqrt{3} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Leftrightarrow x_1 = \frac{2i}{\sqrt{3}}x_3 \text{ e } x_2 = -x_3\frac{i}{\sqrt{3}} \Leftrightarrow V_3 = \begin{pmatrix} 2i/\sqrt{3} \\ -i/\sqrt{3} \\ 1 \end{pmatrix}$$

Teorema 32 *Se* $\mathbf{A}$ *é uma matriz de ordem* N *então:*

1. $\|\mathbf{A}\|_2 = \left\{\rho\left(\mathbf{A}^T\mathbf{A}\right)\right\}^{1/2}$,
2. $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$, *para qualquer norma natural* $\|.\|$.

Aqui não vamos demonstrar este teorema. Vamos apenas usa-lo para calcular a norma euclideana de uma matriz.

Exemplo 36 *Considere a matrix*

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$$

assim temos que:

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{pmatrix}$$

Para encontrarmos os autovalores desta matriz devemos encontrar as raízes de:

$$0 = \det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) = \begin{pmatrix} 3-\lambda & 2 & -1 \\ 2 & 6-\lambda & 4 \\ -1 & 4 & 5-\lambda \end{pmatrix} = -\lambda(\lambda^2 - 14\lambda + 42$$

Desta forma temos que os autovalores de $\mathbf{A}$ *são* $\lambda_1 = 0$, $\lambda_2 = 7 + \sqrt{7}$ *e* $\lambda_3 = 7 - \sqrt{7}$. *Desta forma podemos encontrar que:*

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^T\mathbf{A})} = \sqrt{7 + \sqrt{7}} \approx 3.106.$$

3.3.1 Cálculo da Matriz Inversa

Se consideramos uma matriz quadrada de ordem N, $\mathbf{A}$, uma matriz $mathbfV$ é dita inversa de $\mathbf{A}$ e é notada por $\mathbf{A}^{-1}$ se

$$\mathbf{AV} = \mathbf{I}$$

Se notamos por $\mathbf{v}_i$ as colunas da matriz $\mathbf{V}$ e por $\mathbf{e}_i$ as colunas da matriz identidade $\mathbf{I}$, para encontrarmos a matriz inversa, basta resolvermos os N sistemas lineares da forma

$$\mathbf{A}\mathbf{v}_i = \mathbf{e}_i, i = 1, 2, ..., N.$$

Aqui observamos que como a matriz $\mathbf{A}$ é sempre a mesma, a melhor forma de resolver estes sistemas é fazendo a decomposição $\mathbf{LU}$ se a matriz considerada não é simética, $\mathbf{LL}^T$ se a matriz é simétrica definida positiva ou $\mathbf{LDL}^T$ se a matriz é simétrica.

3.4 Métodos Iterativos

Para resolver sistemas lineares $\mathbf{Ax} = \mathbf{b}$, temos duas grandes classes de métodos:

- *Os Métodos Diretos*: (Eliminação de Gauss e variantes) Processam um número finito de passos até chegar na solução $\mathbf{x}$ que deveria ser exata se não houvesse erros de truncamento.
- *Os Métodos Indiretos*: Produzem uma sequência de vetores que idealmente convergem para a solução do sistema. Os cálculos são encerrados quando a solução aproximada atingir uma precisão especificada. Os métodos indiretos são quase sempre iterativos.

Assim, um método iterativo resolve um sistema de equações diferenciais através de um processo que gerando uma sequência de vetores $\{\mathbf{x}^{(n)}\}_{n=0}^{\infty}$, a partir de um velor inicial $\mathbf{x}^{(0)}$ dado, que converge para a solução do sistema. Aqui iremos discutir alguns métodos iterativos estacionários mais simples. São eles os métodos iterativos de Jacobi e Gauss Seidel. Uma técnica iterativa inicia através de uma aproximação inicial $\mathbf{x}^{(0)}$ e gera uma sequência de vetores que converge para a solução do sistema. As técnicas iterativas transformam o sistema em uma forma equivalente do tipo

$$\mathbf{x} = \mathbf{Tx} + \mathbf{c} \tag{3.14}$$

para alguma matriz $\mathbf{T}$ e algum vetor $\mathbf{c}$ fixos. O método iterativo é dito estacionário quando a matriz $\mathbf{T}$ não varia no decorrer do processo iterativo.

Os métodos iterativos, em geral, são usados para resolver sistemas grandes, pois no caso de sistemas pequenos o número de operações necessárias para atingir uma determinada precisão podem exceder as requeridas pelos métodos diretos. Desta forma, os métodos iterativos são particularmente eficientes na resolução de sistemas esparsos e grandes, o que é o caso de muitos sistemas resultantes da solução numérica de equações diferenciais parciais (EDP). Além disto os métodos iterativos são usualmente menos sensíveis a erros de arredondamento que os métodos diretos.

3.4.1 Os métodos de Gauss Jacobi e Gauss Seidel

Os métodos iterativos de Jacobi e Gauss-Seidel são clássicos e datam do final do século XVIII. Uma técnica iterativa é iniciada com uma aproximação inicial $\mathbf{x}_0$, entrando em um processo que gera uma

Exemplo 37 *Considere o sistema dado por:*

$$\begin{array}{rcrcrcrcr} 10x_1 & - & x_2 & + & 2x_3 & & & = & 6 \\ -x_1 & + & 11x_2 & - & x_3 & + & 3x_4 & = & 25 \\ 2x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & -11 \\ & & 3x_2 & - & x_3 & + & 8x_4 & = & 15 \end{array}$$

Usando um método direto vemos que o sistema acima possui solução única $\mathbf{x} = (1, 2, -1, 1)^T$. *Para converter o sistema na forma* $\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{c}$, *na primeira equação isolamos* $\mathbf{x}_1$, *na segunda* $\mathbf{x}_2$ *etc..., isto é os elementos da diagonal principal de* $\mathbf{A}$ *com,*

$$\mathbf{A} = -\mathbf{L} + \mathbf{D} - \mathbf{U} = -\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 11 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} - \begin{pmatrix} 0 & -1 & 2 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

3.4.2 Critérios de Convergência para os métodos de Gauss Jacobi e Gauss Seidel

Analisando a convergência dos métodos de Jacobi e Gauss Seidel temos o seguinte teorema:

Teorema 33 *Um método iterativo descrito pela equação (3.14) converge para qualquer valor inicial* $\mathbf{x}^{(0)}$ *se e somente se* $\rho(\mathbf{M}) < 1$, *onde* $\rho(\mathbf{M})$ *o raio espectral da matriz* $\mathbf{M}$.

Aqui lembrambos que a solução de sistemas lineares através de métodos iterativos estão fundamentados em uma aproximação sucessiva de vetores $\mathbf{x}$ na direção da solução do sistema e, como já exemplificamos, uma simples troca de ordem nas linhas do sistema pode fazer a diferença entre convergência ou divergência do método aplicado. Além disto, o uso do teorema acima pode ser bastante difícil, pois a determinação do raio espectral da matriz de iteração $\mathbf{M}$ muitas vezes é muito mais trabalhoso que encontrar a solução do sistema linear, desta forma usamos os seguintes critérios de convergência.

- *Critério das Linhas*: Este é um critério de convergência para os métodos de Gauss-Jacobi e de Gauss-Seidel que estabelece uma condição suficiente (mas não necessária) para a convergência. Isto se o critério for verdadeiro, podemos garantir que os métodos citados convergem para qualquer valor inicial $\mathbf{x}^{(0)}$ considerado. Porém, devemos notar que nada podemos afirmar sobre a convergência dos métodos se este critério não for satisfeito.

O critério das linhas garante a convergência dos métodos de Gauss-Jacobi e Gauss-Seidel para qualquer valor inicial, se a matriz $\mathbf{A}$ é diagonal dominante, isto é se

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}| < |a_{ii}|$$

- *Critério de Sanssenfeld*: Este é um critério que garante a convergência do método de Gauss Seidel. Se o critério das linhas é satisfeito o critério de Sanssenfeld é automaticamente satisfeito, porém o contrário não é verdade.

  Um sistema linear converge para o método de Gauss-Seidel, independentemente do valor inicial escolhido, se

  $$\max\{\beta_k\} < 1,$$

  onde

  $$\beta_1 = \frac{1}{|a_{11}|}\sum_{j=2}^{N}|a_{ij}|, \qquad \beta_i = \frac{1}{|a_{ii}|}\left(\sum_{j=1}^{i-1}|a_{ij}|\beta_j + \sum_{j=i+1}^{N}|a_{ij}|\right), \quad i = 2, 3, ..., N$$

  Exemplo 38 *Como exemplo, vamos considerar o seguinte sistema linear:*

  $$\begin{array}{rcrcrcrcrcr} 2.0 & x_1 & + & & x_2 & - & 0.2 & x_3 & + & 0.2 & x_4 & = & 0.4 \\ 0.6 & x_1 & + & 3.0 & x_2 & - & 0.6 & x_3 & - & 0.3 & x_4 & = & -7.8 \\ -0.1 & x_1 & - & 0.2 & x_2 & + & & x_3 & + & 0.2 & x_4 & = & 1.0 \\ 0.4 & x_1 & + & 1.2 & x_2 & + & 0.8 & x_3 & + & 2.0 & x_4 & = & 20.0 \end{array}$$

  No sistema acima podemos ver que o critério das linhas não é válido. Porém pelo critério de Sansenfeld temos:

  1. $\beta_1 = \frac{1}{2}(1 + 0.2 + 0.2) = 0.7$
  2. $\beta_2 = \frac{1}{3}((0.6)(0.7) + 0.6 + 0.3) = 0.44$
  3. $\beta_3 = \frac{1}{1}((0.1)(0.7) + (0.2)(0.44) + 0.2) = 0.358$
  4. $\beta_4 = \frac{1}{2}((0.4)(0.7) + (1.2)(0.44) + (0.8)(0,358)) = 0.5472$
  5. $\max \beta_1, \beta_2, \beta_3, \beta_4 = 0.7 < 1$

  E desta forma, pelo critério de Sansenfeld podemos garantir que a solução deste sistema pelo método de Gauss-Seidel irá convergir. (para qualquer valor inicial $\mathbf{x}^{(0)}$.)

Observações

1. Aqui vamos observar que sempre que o critério das linhas for verdadeiro, o critério de Sanssenfeld também seré, porem a reciproca não é verdadeira, por exemplo, considere o sistema:

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

   Este sistema não satisfaz o critério das linhas, pois $|a_{22}| = 2 < |a_{21}| = 6$, porém satisfaz o critério de Sanssenfeld pois:

$$\beta_1 = \left|\frac{1}{10}\right| = 0.1, \qquad \beta_2 = \left|\frac{6}{2}\right| \beta_1 = 0.3$$

   assim $\max \beta_i = 0.3 < 1$.

2. A ordem comno as equações aparecem no sistema possui grande influência na convergência de um método iterativo. Por exemplo considere o sistema:

$$\begin{aligned} -4\,x_1 \;+\; 10x_2 &= 19 \\ 5\,x_1 \;+\; 3\,x_2 &= 15 \end{aligned}$$

   Nesta forma este sistema não satisfaz aos nossos critérios de convergência, porém se trocarmos as linhas, o sistema irá satisfazer este critério e sua convergência é garantida. (Verifique)

3. A quantidade de operações requerida em uma iteração é simples conhecer, porém o número de iterações requeridas não. Os métodos iterativos de Gauss-Seidel e Gauss-Jacobi realizam por iteração $2N^2 - N$ operações de aritméticas. São elas $N - 1$ multiplicações de variáveis pelos coeficientes, $N - 1$ somas e 1 divisão para cada variável do sistema, totalizando para cada variável $2N - 1$ operações para cada uma das N variáveis. Quando o valor de N é grande temos a ordem de $2N^2$ operações.

Podemos ver que o custo computacional dos métodos iterativos é menor que dos métodos diretos quando consideramos N grande. Os métodos iterativos também apresenta a vantagem de preservbar os zeros da matriz original e possuirem um erro de arredondamento menores. O problema dos métodos iterativos é de serem menos eficientes para a solução de sistemas lineares densos e de pequeno porte. Os métodos diretos, teoricamente obtém a solução de qualquer sistema não singular e os métodos iterativos convergem apenas sobre determinadas condições. Os erros de arredondamento aparecem de forma bem mais suave nos métodos iterativos que nos métodos diretos.

3.4.3 Condicionamento de uma matriz

ver em arquivo anexo

3.5 Exercícios

1. Dado o vetor $\mathbf{x} = [1, 2, 3, 4, 5]^T$, calcular as normas

   (a) $||\mathbf{x}||_1$

   (b) $||\mathbf{x}||_2$

   (c) $||\mathbf{x}||_\infty$

2. Calcular a norma da matriz $\mathbf{A} = \begin{pmatrix} 5 & 4 & -1 \\ 2 & 9 & 3 \\ 8 & -6 & 7 \end{pmatrix}$

   (a) $||\mathbf{A}||_1$

   (b) $||\mathbf{A}||_2$

   (c) $||\mathbf{A}||_\infty$

3. Resolver os sistemas abaixo pelo método de eliminação de Gauss, com estratégia indicada e usando 3 casas decimais; verificar também a unicidade e exatidão da solução obtida.

   (a) Sem pivotamento parcial

$$\begin{pmatrix} 1 & 2 & 4 \\ -3 & -1 & 4 \\ 2 & 14 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 13 \\ 8 \\ 50 \end{pmatrix}$$

   (b) Com pivotamento parcial

$$\begin{pmatrix} -2 & 3 & 1 \\ 2 & 1 & -4 \\ 4 & 10 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -5 \\ -9 \\ 2 \end{pmatrix}$$

   (c) Com pivotamento parcial

$$\begin{pmatrix} 0 & 1 & 3 & 2 & 4 \\ 8 & -2 & 9 & -1 & 2 \\ 5 & 1 & 1 & 7 & 2 \\ -2 & 4 & 5 & 1 & 0 \\ 7 & -3 & 2 & -4 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3 \\ -5 \\ 6 \\ -1 \\ 8 \end{pmatrix}$$

   (d) Com pivotamento escalonado

$$\begin{pmatrix} -2 & 3 & 1 \\ 2 & 1 & -4 \\ 4 & 10 & -6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -5 \\ -9 \\ 2 \end{pmatrix}$$

(e) Com pivotamento escalonado

$$\begin{pmatrix} 0 & 1 & 3 & 2 & 4 \\ 8 & -2 & 9 & -1 & 2 \\ 5 & 1 & 1 & 7 & 2 \\ -2 & 4 & 5 & 1 & 0 \\ 7 & -3 & 2 & -4 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 3 \\ -5 \\ 6 \\ -1 \\ 8 \end{pmatrix}$$

4. Resolver pelo método da decomposição **LU**. Verificar unicidade e exatidão da solução.

(a) Efetue os cálculos utilizando apenas aritmética de 4 dígitos.

$$\begin{pmatrix} 2 & 6 & -3 \\ 1 & 3.001 & 2 \\ 4 & -1 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 9 \\ 29 \end{pmatrix}$$

i. sem pivotamento

ii. com pivotamento

(b) Com pivotamento parcial

$$\begin{pmatrix} 1 & 2 & 3 \\ -5 & -1 & 4 \\ 2 & 4 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 17 \\ -2 \\ 24 \end{pmatrix}$$

(c) Resolva o sistema de equações:

$$\begin{cases} x_1 + 4x_2 + 2x_3 = 3.70 \\ 5x_1 + 2x_2 + x_3 = 6.89 \\ 3x_1 + x_2 + 6x_3 = 5.49 \end{cases}$$

pelo método da decomposição LU

(d) Considere o exercício 4c e a aproximação inicial $\mathbf{x}^{(0)} = [1.10 \quad 0.50 \quad 0.28]^T$.

i. Verifique o que pode ser feito para garantir convergência do método de Jacobi e calcule 3 iterações por este método.

ii. Verifique o que pode ser feito para garantir convergência do método de Gauss-Seidel e calcule 3 iterações por este método.

5. Considere o seguinte sistema de equações lineares:

$$\begin{cases} x_1 + 2x_2 - x_3 = 1 \\ 2x_1 - x_2 = 1 \\ x_2 + 2x_3 - x_4 = 1 \\ -x_3 + 2x_4 = 1 \end{cases}$$

(a) Mostre que este sistema não satisfaz o critério de linhas,

(b) Mostre que este sistema não satisfaz o critério de Sassenfeld,

(c) O que se pode afirmar sobre a convergência dos métodos de Gauss-Jacobi e Gauss-Seidel, quando aplicados a este sistema?

(d) Mostre que o sistema obtido permutando-se as duas primeiras equações satisfaz o critério de Sassenfeld,

(e) Usando o método de Gauss-Seidel, determine a solução aproximada do sistema, com a permutação sugerida no item anterior e erro

$$||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||_\infty = \max_{i=1,2,3,4} |x_i^{(k+1)} - x_i^{(k)}| \leq 10^{-3}$$

6. Considere o sistema linear abaixo:

$$\begin{cases} x_1 - 3x_2 + x_3 - x_4 = -1 \\ x_2 - 2x_3 = 3 \\ 10x_1 + x_2 + x_4 = -8 \\ x_1 - x_3 + 3x_4 = 8 \end{cases}$$

(a) Resolva o sistema pelo método de Elimina¸c˜ao de Gauss com pivoteamento.

(b) Utilizando o resultado do item anterior, escreva o sistema linear $\mathbf{LUx} = \mathbf{Pb}$, equivalente ao sistema linear dado, onde $\mathbf{P}$ é uma matriz permutação. Determine a matriz inversa.

(c) Podemos determinar a solução aproximada do sistema , usando o Método de Gauss - Seidel para qualquer aproximação inicial ? Porquê?

Capítulo 4

Interpolação e Extrapolação

Muitas vezes experimentos científicos ou computação numérica nos fornecem valores discretos de uma função. Estes valores podem ser dados de forma igualmente espaçada ou não ao longo do eixo dos x. Aqui vamos estimar o valor da função entre os valores tabulados (interpolação) ou fora do domínio destes pontos (extrapolação). Uma das funções mais conhecidas são as funções polinomiais. Podemos mostrar que esta classe de funções mapeia uniformemente funções contínuas, isto é, considerando qualquer função contínua em um intervalo limitado e fechado, sempre existirá um polinômio tão próximo quanto se queira desta função.

Teorema 34 (Aproximação de Weierstrass) *Se f é uma função definida e contínua em um intervalo fechado $[a, b]$, então para cada $\varepsilon > 0$ existe um polinômio de forma que $|f(x) - p(x)| < \varepsilon$ para todo $x \in [a, b]$.*

Teorema 35 *Teorema: Se $x_0, x_1, ..., x_n$ são números reais distintos, então para valores arbitrários $y_1, y_2, ..., y_n$, existe um único polinômio a de grau igual ou menor que n de forma que $P_n(x_i) = y_i$ $(0 \leq i \leq n)$.*

4.1 Dados não igualmente espaçados

4.1.1 Polinômios Interpoladores de Lagrange

Nesta seção, vamos estudar polinômios interpoladores que possam ser determinados simplesmente pela especificação de certos pontos do plano por onde eles devem passar. Por várias razões numéricas ou experimentais muitas vezes é inconveniente ou impossível obter dados em intervalos igualmente espaçados. Desta forma vamos adotar uma outra forma de aproximação polinomial.
Note que, uma interpolação linear para a função f que passa pelos pontos $f(x_0) = y_0$ e $f(x_1) = y_1$ é um polinômio de grau 1 que passa pelos dois pontos. Sabe-se que este interpolador é único, pois por $n + 1$ pontos passa um e somente um polinômio de grau n, e este polinômio, no caso linear pode ser escrito como:

$$P(x) = \frac{(x - x_1)}{(x_0 - x_1)} y_0 + \frac{(x - x_0)}{(x_1 - x_0)} y_1 = L_0(x) y_0 + L_1(x) y_1 \tag{4.1}$$

quando $x = x_0$, temos que, $L_0(x_0) = 1$ e $L_1(x_0) = 0$ e quando $x = x_1$, $L_0(x_1) = 0$ e $L_1(x_1) = 1$, assim:

$$P(x_0) = 1.y_0 + 0.y_1 = y_0 = f(x_0),$$

e

$$P(x_1) = 0.y_0 + 1.y_1 = y_1 = f(x_1),$$

Desta forma, (4.1) possui as propriedades requeridas para ser o polinômio interpolador de grau 1 da função f.
Esta técnica pode ser imediatamente generalizada. Assim vamos considerar a construção de um polinômio de grau máximo n que passa através de $n+1$ pontos $(x_0, f(x_0))$, $(x_1, f(x_1))$, ... , $(x_n, f(x_n))$. Assim, a gora precisamos construir para cada $k = 1, 2, ..., n$, um quociente L_k com a propriedade que $L_k(x_k) = 1$ e $L_k(x_i) = 0$ se $i \neq k$. Para $L_k(x_i) = 0$ se $i \neq k$ devemos ter no numerador de L_k o termo:

$$(x - x_0)(x - x_1)...(x - x_{k-1})(x - x_{k+1})...(x - x_n). \tag{4.2}$$

Para satisfazer $L_k(x_k) = 1$, o denominador de L_k deve ser igual à (4.2) quando $x = x_k$. Assim:

$$L_k(x) = \frac{(x - x_0)(x - x_1)...(x - x_{k-1})(x - x_{k+1})...(x - x_n)}{(x_k - x_0)(x_k - x_1)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_n)} = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)}.$$

O polinômio interpolador é facilmente descrito quando a forma de L_k é conhecida. Este polinômio é chamado de **polinômio interpolador de Lagrange** e é definido pelo teorema:

Teorema 36 *Se $x_0, x_1, ..., x_n$ são $(n+1)$ números distintos e f é uma função cujos valores são dados nestes números, então existe um único polinômio P de grau n tal que*

$$f(x_k) = P(x_k) \text{ para cada } k = 0, 1, ..., n.$$

Este polinômio é dado por

$$P(x) = f(x_0)L_0(x) + ... + f(x_n)L_n(x) = \sum_{k=0}^{n} f(x_k)L_k(x), \tag{4.3}$$

onde

$$L_k(x) = \frac{(x-x_0)(x-x_1)...(x-x_{k-1})(x-x_{k+1})...(x-x_n)}{(x_k-x_0)(x_k-x_1)...(x_k-x_{k-1})(x_k-x_{k+1})...(x_k-x_n)} = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{(x-x_i)}{(x_k-x_i)} \tag{4.4}$$

para cada $k = 0, 1, ..., n$.

Exemplo 39 *Encontre o polinômio interpolador de Lagrange de grau 2 para* $f(x) = \frac{1}{x}$*, usando os nós* $x_0 = 2$*,* $x_1 = 2.5$ *e* $x_2 = 4$.

1. *Calculando* L_0*,* L_1 *e* L_2*:*

$$\begin{aligned} L_0 &= \frac{(x-2.5)(x-4)}{(2-2.5)(2-4)} = (x-6.5)x + 10, \\ L_1 &= \frac{(x-2)(x-4)}{(2.5-2)(2.5-4)} = \frac{(-4x+24)x-32}{3}, \\ L_2 &= \frac{(x-2)(x-2.5)}{(4-2)(4-2.5)} = \frac{(x-4.5)x+5}{3}. \end{aligned}$$

2. *Como* $f(x_0) = f(2) = 0.5$*,* $f(x_1) = f(2.5) = 0.4$ *e* $f(x_2) = f(4) = 0.25$*,*

$$\begin{aligned} P(x) &= 0.5\left((x-6.5)x+10\right)+10) + 0.4\frac{(-4x+24)x-32}{3} + 0.25\frac{(x-2)(x-2.5)}{(4-2)(4-2.5)} \\ &= \frac{(x-4.5)x+5}{3} = (0.05x-0.425)x + 1.15 \end{aligned}$$

3. *Assim uma aproximação para f(3) é:*

$$f(3) \approx P(3) = 0.325$$

Qual o **limite de erro** envolvido? Conhecer uma fórmula para o limite de erro é muito importante para este método, pois os *polinômios de Lagrange* são muito usados em métodos de derivação e integração numérica.

Teorema 37 *Se* $x_0, x_1, ..., x_n$ *são números distintos em um intervalo*$[a, b]$ *e* $f \in \mathcal{C}^{n+1}[a, b]$*, então para cada* $x \in [a, b]$ *existe um número* $\xi(x) \in (a, b)$ *com*

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)...(x-x_n), \tag{4.5}$$

onde P *é o polinômio interpolador dado em (4.3).*

Note que a forma do erro para o polinômio de Lagrange é muito similar a forma de erro do polinômio de Taylor. O polinômio de Taylor de grau n em torno de x_0 concentra a informação conhecida em x_0, isto é

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

enquanto que o polinômio de Lagrange de grau n usa informação em pontos distintos $x_0, x_1, ..., x_n$, isto é

$$\frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)(x-x_1)...(x-x_n).$$

O uso específico destas fórmulas fica restrito aquelas funções com limite conhecido.

Nota: Devemos notar que este polinômios interpoladores deste tipo podem ser muito perigosos no centro de regiões onde a variável independente possui um afastamento muito grande. Apesar do polinômio "concordar"com os dados nos pontos fixados, pode mover-se sem um curso determinado entre pontos muito afastados.

Exemplo 40 *Vamos considerar a seguinte tabela:*

x	-3.0	-1.0	1.0	2.0	2.5	3.0
$f(x)$	1.0	1.5	2.0	2.0	1.5	1.0

Aproxime $f(0.3)$ usando polinômio interpolador de Lagrange, com $n = 3$ e $n = 2$

$$\begin{aligned}
&n = 3 x_1 = -1.0, x_2 = 1.0, x_3 = 2.0, x_4 = 2.5 \\
&P_3(0.3) = \frac{(0.3+1)(0.3-1)(0.3-2)}{(2.5+1)(2.5-1)(2.5-2)}(1.5) + \frac{(0.3+1)(0.3-1)(0.3-2.5)}{(2+1)(2-1)(2-2.5)}(2) \\
&\qquad + \frac{(0.3+1)(0.3-2)(0.3-2.5)}{(1+1)(1-2)(1-2.5)}(2) + \frac{(0.3-1)(0.3-2)(0.3-2.5)}{(-1-1)(-1-2)(-1-2.5)}(1.5) \\
&P_3(0.3) = (0.589333)(1.5) - (1.33467)(2) + (1.62067)(2) + (0.124667)(1.5) = 1.64300
\end{aligned}$$

$$\begin{aligned}
&n = 2 x_1 = -1.0, x_2 = 1.0, x_3 = 2.0 \\
&P_2(0.3) = 1.9008333
\end{aligned}$$

Note que sem mais informações sobre a função, não podemos saber qual dos dois resultados é o melhor. Agora, suponha que a função f descreve um certo fenômeno físico e que neste caso sabe-se que para este fenômeno as derivadas devem satisfazer $|f^{(k)}(x)| \leq 1/k$ para $k = 1, 2, ...$ Neste caso, usando (4.5),

$$|E_{T_3}| \leq \left| \frac{1}{4} \frac{(0.3+1)(0.3-1)(0.3-2)(0.3-2.5)}{4!} \right| = 0.0355 \tag{4.6}$$

$$|E_{T_2}| \leq \left| \frac{1}{3} \frac{(0.3+1)(0.3-1)(0.3-2)}{3!} \right| = 0.0860 \tag{4.7}$$

Assim neste caso a melhor aproximação para o valor de $f(0.3) \approx P_3(0.3) = 1.64300$.

Agora para ilustrar a importância de conhecer algo sobre as derivadas da função, vamos considerar

$$f(x) = \frac{1}{1+25x^2}$$

podemos neste caso calcular $f(-0.5) = 0.1379310$. Agora vamos supor que não conhecemos a função, mas sim uma tabela com os valores,

x	-1	$-.8$	$-.6$	$-.4$	$-.2$	0
f	0.3846	0.05882	0.1000	0.2000	0.5000	1.000

usando interpolação de Lagrange sobre os dados da tabela encontramos:

n	$P(-0.5)$	Erro Absoluto
1	0.15	0.01203
2	0.1250	0.01293
3	0.133824	0.00411
4	0.2008273	0.062896

Note que o crescimento do erro de truncamento com o aumento de n é devido ao fato que as derivadas desta função são da forma:

$$\begin{aligned} f'(x) &= \frac{-50x}{(1+25x^2)^2} \\ f''(x) &= \frac{-50+3750x^2}{(1+25x^2)^3} \end{aligned}$$

etc... que crescem rapidamente quando $x = -0.5$.

Finalmente, devemos observar que podemos mostrar que o erro de arredondamento para o método de Lagrange cresce com n^2 com o crescimento de n. Este crescimento não é sério para valores de n que realmente ocorrem na prática.

4.1.2 Forma de Newton do Polinômio Interpolador

Pelo teorema anterior temos que existe um único polinômio de grau n que passa por $n+1$ pontos distintos (x_0, y_0), (x_1, y_1), ... , (x_n, y_n). Se $n = 0$ temos apenas um ponto e o polinômio de grau zero

$$p_0(x) = y_0$$

passa por (x_0, y_0). Se temos dois pontos (x_0, y_0) e (x_1, y_1), podemos construir um polinômio de primeiro grau

$$p_1(x) = c_0 + c_1(x - x_0)$$

como queremos que $p_1(x_0) = y_0$ e $p_2(x_1) = y_1$, fazemos

$$c_0 = y_0 = p_0(x_1)$$

$$c_1 = \frac{y_1 - c_0}{x_1 - x_0} = \frac{y_1 - p_0(x_1)}{x_1 - x_0}$$

Se queremos que passe por 3 pontos (x_0, y_0), (x_1, y_1), (x_2, y_2), escrevemos o polinômio como:

$$p_2(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$$

no polinômio acima temos que

$$p_2(x_0) = c_0 = y_0$$

,

$$p_2(x_1) = c_0 + c_1(x - x_0) = y_0 + \frac{y_1 - p_0(x_1)}{x_1 - x_0}(x_1 - x_0)$$

mas $p_0(x_1) = y_0$, logo

$$p_2(x_1) = y_1$$

Ainda queremos que $p_2(x_2) = y_2$, assim temos

$$y_2 = c_0 + c_1(x_2 - x_0) + c_2(x_2 - x_0)(x_2 - x_1)$$

logo

$$c_2 = \frac{y_2 - (c_0 + c_1(x_2 - x_1))}{(x_2 - x_0)(x_2 - x_1)} = \frac{y_2 - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

Seguindo este raciocínio, obtemos que o polinômio que passa por $n+1$ pontos (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) pode ser escrito como:

$$p_n(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + ... + c_n(x - x_0)...(x - x_{n-1})$$

onde

$$c_k = \frac{y_k - p_{k-1}(x_k)}{(x_k - x_0)(x_k - x_1)...(x_k - x_{k-1})}$$

Computacionalmente, para calcular os $p_k(x)$, se os coeficientes c_k já são conhecidos, um método eficiente de cálculo é o algoritmo de Horner. Chamando $d_i = (x - x_i)$ escrevemos o polinômio como:

$$u(x) = c_0 + c_1 d_0 + c_1 d_0 d_1 + ... + c_n d_0 ... d_{n-1}$$

ou

$$u(x) = (...(((c_k)d_{k-1} + c_{k-1})d_{k-2} + c_{k-2})d_{k-3} + ... + c_1)d_0 + c_0$$

Para testarmos o método acima vamos escolher o polinômio $p_3(x) = 4x^3 + 35x^2 - 84x - 954$, e vamos gerar a tabela:

x	5	-7	-6	0
y	1	-23	-54	-954

temos que:

$$c_0 = 1$$

$$p_0(x) = 1$$

assim

$$c_1 = \frac{y_1 - p_0(-7)}{-7 - 5} = 2$$

$$p_1(x) = 1 + 2(x - 5)$$

novamente

$$c_2 = \frac{y_2 - p_1(-6)}{(-6 - 5)(-6 + 7)} = 3$$

$$p_2(x) = 1 + 2(x - 5) + 3(x - 5)(x + 7) = 3x^3 + 8x - 114$$

finalmente

$$c_3 = \frac{y_3 - p_1(0)}{(0 + 6)(0 + 7)(0 - 5)} = 4$$

$$p_3(x) = 1 + 2(x - 5) + 3(x - 5)(x + 7) + 4(x - 5)(x + 7)(x + 6) = 4x^3 + 35x^2 - 84x - 954$$

4.2 Interpolação por Spline Cúbico

Uma das dificuldades com a interpolação polinomial, particularmente se o polinômio é de alta ordem, é seu caráter altamente oscilatório. Uma interpolação de uma função mais suave pode ser produzida mecanicamente por uma curva Francesa ou forçando uma barra elástica flexível a passar pelos pontos desejados. O análogo matemático desta barra flexível é a função *spline*. Um spline é uma função polinomial por partes, isto é, é um polinômio de grau n entre cada dois nós da malha considerada. O spline é particularmente usado para interpolar tabela de funções com propriedades físicas livres de ruídos.
A aproximação mais comum por polinômios por partes é usando polinômios cúbicos, os *splines cúbicos*, esta aproximação interpola pontos entre cada dois pontos consecutivos da malha, que não precisa ser igualmente espaçada, usando polinômios cúbicos. Um polinômio cúbico envolve quatro constantes, nos dando assim flexibilidade para assegurar que o interpolante além de ser contínuo em cada ponto da malha seja diferenciável, e ainda podendo garantir a existência de derivadas segundas contínuas. Assim o Spline cúbico que vamos chamar de $S(x)$ é uma função formada por polinômios cúbicos $S_j(x)$ definidos entre os nós x_j e x_{j+1} da malha de forma que:

$$\begin{cases} S_j(x_{j+1}) = f(x_{j+1}) = S_{j+1}(x_{j+1}) & \text{os polinômios coincidem em } x_{j+1} \\ S'_j(x_{j+1}) = S'_{j+1}(x_{j+1}) & \text{as declividades coincidem em } x_{j+1} \\ S''_j(x_{j+1}) = S''_{j+1}(x_{j+1}) & \text{as curvaturas coincidem em } x_{j+1} \end{cases}$$

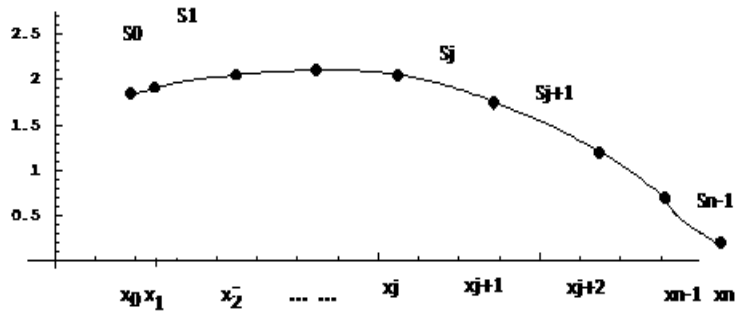


Figura 4.1: Spline Cúbico

Assim podemos definir Spline cúbico como:

Definição 14 *Dadas uma função f em $[a, b]$, um conjunto de números chamados de nós*

$$a = x_0 < x_1 < \cdots < x_n = b$$

um interpolante Spline Cúbico, S, para f é uma função que satisfaz as seguintes condições:

1. *S é um polinômio Cúbico, chamado de S_j em cada subintervalo $[x_j, x_{j+1}]$, para $j = 0, 1, \ldots, n-1$.*
2. *$S(x_j) = f(x_j)$, para $j = 0, 1, \ldots, n$.*
3. *$S_{j+1}(x_{j+1}) = S_j(x_{j+1})$, para $j = 0, 1, \ldots, n-2$.*
4. *$S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$, para $j = 0, 1, \ldots, n-2$.*
5. *$S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$, para $j = 0, 1, \ldots, n-2$.*
6. *Uma das seguintes condições de contorno é satisfeita:*
   (a) *$S''(x_0) = S''(x_n) = 0$ (coordenadas livres ou naturais).*
   (b) *$S'(x_0) = f'(x_0)$ e $S'(x_n) = f'(x_0)$ (extremos "campled").*

Devemos notar que os splines cúbicos podem ser definidos com outras condições de contorno, mas estas duas são as mais usadas.

4.2.1 Spline Cúbico Natural

Chamamos de spline cúbico natural quando ocorrem condições de contorno livres, seu gráfico aproxima a forma de de um longo cabo flexível que passa através dos pontos dados $(x_0, f(x_0))$, $(x_1, f(x_1))$ e $(x_n, f(x_n))$.

Vamos agora construir o interpolante spline cúbico para uma dada função f. Assim seja o conjunto de polinômios,

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3 \tag{4.8}$$

para cada $j = 0, 1, \ldots, n-1$. Para cada x_j temos que $S_j(x) = a_j = f(x_j)$, e se aplicamos a condição (3) temos,

$$a_{j+1} = S_{j+1}(_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3$$

para $j = 0, 1, \ldots, n-2$. Daqui em diante vamos usar a notação, $h_j = x_{j+1} - x_j$.
Se nós também definimos $\mathbf{a_n} = \mathbf{f(x_n)}$ podemos ver que a equação

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 \tag{4.9}$$

vale para $j = 0, 1, \ldots, n-1$.
De maneira similar definimos $\mathbf{b_n} = \mathbf{S'(x_n)}$ e como

$$S'_j(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

temos que $S'_j(x_j) = b_j$ para $j = 0, 1, \ldots, n-1$ e assim aplicando a condição (4),

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2 \ , \quad j = 0, 1, \ldots, n-1 \tag{4.10}$$

Podemos obter uma outra condição entre os coeficientes de S_j se definimos $\mathbf{c_n} = \frac{\mathbf{S''(x_n)}}{\mathbf{2}}$ e aplicando a condição (5). Neste caso,

$$c_{j+1} = c_j + 3d_j h_j \tag{4.11}$$

para $j = 0, 1, \ldots, n-1$.
Resolvendo para d_j na equação (4.11)

$$d_j = \frac{c_{j+1} - c_j}{3h_j} \tag{4.12}$$

e substituindo este valor na equação (4.9) e (4.10) temos as novas equações

$$a_{j+1} = a_j + b_j h_j + \frac{h_j^2}{3}(2c_j + c_{j+1}) \tag{4.13}$$

e

$$b_{j+1} = b_j + h_j(c_j + c_{j+1}) \tag{4.14}$$

para cada $j = 0, 1, 2, ..., n-1$.
A relação final envolvendo os coeficientes é obtido resolvendo a equação apropriada na forma da equação (4.13), primeiramente para b_j,

$$b_j = \frac{1}{h_j}(a_{j+1} - a_j) - \frac{h_j}{3}(2c_j + c_{j+1}), \tag{4.15}$$

e procedendo uma redução de índice da equação acima para b_{j-1},

$$b_{j-1} = \frac{1}{h_{j-1}}(a_j - a_{j-1}) - \frac{h_{j-1}}{3}(2c_{j-1} + c_j).$$

Substituindo estes valores na equação derivada de (4.14), quando o índice é reduzido de um, obtemos o sistema linear

$$h_{j-1}c_{j-1} + 2(h_{j-1} + h_j)c_j + h_j c_{j+1} = \frac{3}{h_j}(a_{j+1} - a_j) - \frac{3}{h_{j-1}}(a_j - a_{j-1}) \tag{4.16}$$

para cada $j = 1, 2, ..., n-1$. Este sistema envolve, como variáveis, somente $\{c_j\}_{j=0}^{n}$, uma vez que os valores de $\{h_j\}_{j=0}^{n-1}$ e $\{a_j\}_{j=0}^{n}$ são dados pelo espaçamento entre os nós $\{x_j\}_{j=0}^{n}$ e pelos valores $\{f(x_j)\}_{j=0}^{n}$ nos nós.
Note que uma vez que os valores de $\{c_j\}_{j=0}^{n}$ ficam conhecidos, podemos de forma simples encontrar as constantes restantes $\{b_j\}_{j=0}^{n-1}$ da equação (4.15) e $\{d_j\}_{j=0}^{n-1}$ da equação (4.11), e, assim, construir os polinômios cúbicos $\{S_j\}_{j=0}^{n-1}$. A maior questão que aparece em conexão com esta construção é quando os valores de $\{c_j\}_{j=0}^{n}$ podem ser determinados usando o sistema de equações dado por (4.16), e quando estes valores são únicos. A resposta é que com as condições de contorno naturais sempre teremos solução única para este sistema. Isto é

Teorema 38 *Se f está definida em $a = x_0 < x_1 < ... < x_n = b$, então f possui um único interpolante spline cúbico natural nos nós $x_0, x_1, ..., x_n$, isto é, um interpolador de spline cúbico que satisfaz as condições de contorno $S"(a) = S"(b) = 0$.*

A demonstração deste teorema é encontrada em Burden and Faires Numerical Analysis.
Os coeficientes de

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

podem ser calculados como:

$$\begin{aligned} &h_j = x_{j+1} - x_j \\ &a_j = f(x_j) \text{ para } j = 0, 1, ..., n \end{aligned} \tag{4.17}$$

os coeficientes c_j são calculados resolvendo o sistema linear,

$$\mathbf{Ac} = \mathbf{v} \tag{4.18}$$

onde

$$A = \begin{bmatrix} 1 & 0 & 0 & ... & ... & ... & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & ... & ... & ... & : \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & ... & ... & : \\ : & ... & ... & ... & ... & ... & : \\ : & ... & ... & ... & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & ... & ... & ... & 0 & 0 & 1 \end{bmatrix}, \tag{4.19}$$

$$\mathbf{c} = [c_0, c_1, ..., c_n]^T \tag{4.20}$$

e finalmente

$$\mathbf{v} = \begin{bmatrix} 0 \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ \vdots \\ \vdots \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 0 \end{bmatrix} \tag{4.21}$$

e finalmente,

$$\begin{aligned} b_j &= (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3 \\ d_j &= (c_{j+1} - c_j)/(3h_j) \end{aligned} \tag{4.22}$$

Obs: A matriz **A** é estritamente diagonal dominante, o que garante a unicidade de solução do sistema (4.18).

4.2.2 Spline Cúbico Restrito

Teorema 39 *Se f está definida em $a = x_0 < x_1 < ... < x_n = b$ e é diferenciável em a e b, então f possui um único interpolante spline cúbico Restrito nos nós $x_0, x_1, ..., x_n$, isto é, um interpolador de spline cúbico que satisfaz as condições de contorno $S'(a) = f'(a)$ e $S'(b) = f'(b)$.*

A demonstração deste teorema é encontrada em Burden and Faires Numerical Analysis.
Os coeficientes de

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

podem ser calculados como:

$$\begin{aligned} h_j &= x_{j+1} - x_j \\ a_j &= f(x_j) \text{ para } j = 0, 1, ..., n \end{aligned} \tag{4.23}$$

os coeficientes c_j são calculados resolvendo o sistema linear,

$$\mathbf{Ac} = \mathbf{v} \tag{4.24}$$

onde

$$A = \begin{bmatrix} 2h_0 & h_0 & 0 & ... & ... & ... & 0 \\ h_0 & 2(h_0+h_1) & h_1 & ... & ... & ... & : \\ 0 & h_1 & 2(h_1+h_2) & h_2 & ... & ... & : \\ : & ... & ... & ... & ... & ... & : \\ : & ... & ... & ... & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & ... & ... & ... & 0 & h_{n-1} & 2h_{n-1} \end{bmatrix}, \tag{4.25}$$

$$\mathbf{v} = \begin{bmatrix} \frac{3}{h_0}(a_1 - a_0) - 3f'(a) \\ \frac{3}{h_1}(a_2 - a_1) - \frac{3}{h_0}(a_1 - a_0) \\ : \\ : \\ \frac{3}{h_{n-1}}(a_n - a_{n-1}) - \frac{3}{h_{n-2}}(a_{n-1} - a_{n-2}) \\ 3f'(b) - \frac{3}{h_{n-1}}(a_n - a_{n-1}) \end{bmatrix} \tag{4.26}$$

e finalmente,

$$\mathbf{c} = [c_0, c_1, ..., c_n]^T \tag{4.27}$$

e finalmente

$$\begin{aligned} b_j &= (a_{j+1} - a_j)/h_j - h_j(c_{j+1} + 2c_j)/3 \\ d_j &= (c_{j+1} - c_j)/(3h_j) \end{aligned} \tag{4.28}$$

Obs: A matriz **A** é estritamente diagonal dominante, o que garante a unicidade de solução do sistema (4.24) e que a pode ser resolvido por eliminação Gaussiana sem troca de linhas.

Exemplo 41 *Por exemplo, vamos interpolar o seguinte conjunto de 21 pontos:*

$\{(0.9, 1.3), (1.3, 1.5), (1.9, 1.85), (2.1, 2.1), (2.6, 2.6), (3.0, 2.7), (3.9, 2.4), (4.4, 2.15), (4.7, 2.05), (5.0, 2.1),$
$(6.0, 2.25), (7.0, 2.3), (8.0, 2.25), (9.2, 1.95), (10.5, 1.4), (11.3, 0.9), (11.6, 0.7),$
$(12.0, 0.6), (12.6, 0.5), (13.0, 0.4), (13.3, 0.25)\}$

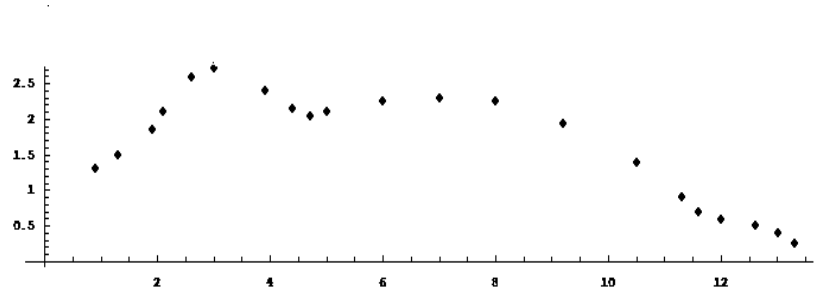


Figura 4.2: Conjunto de pontos a ser interpolado

Primeiramente vamos usar interpolador de Lagrange e desta forma obtemos um polinômio de grau 20 dado por $p(x) = -9652.785079125108 + 52462.58032870206x - 128683.40x^2 + 189994.99x^3 - 189850.97x^4 + 136777.90x^5 - 73866.57x^6 + 30677.63x^7 - 9968.98x^8 + 2564.20x^9 - 525.81x^{10} + 86.25x^{11} - 11.32x^{12} + 1.18x^{13} - 0.09769x^{14} + 0.006286x^{15} - 0.0003082x^{16} + 0.00001111x^{17} - 2.77 * 10^{-7}x^{18} + 4.28 * 10^{-9}x^{19} - 3.07 * 10^{-11}x^{20}$

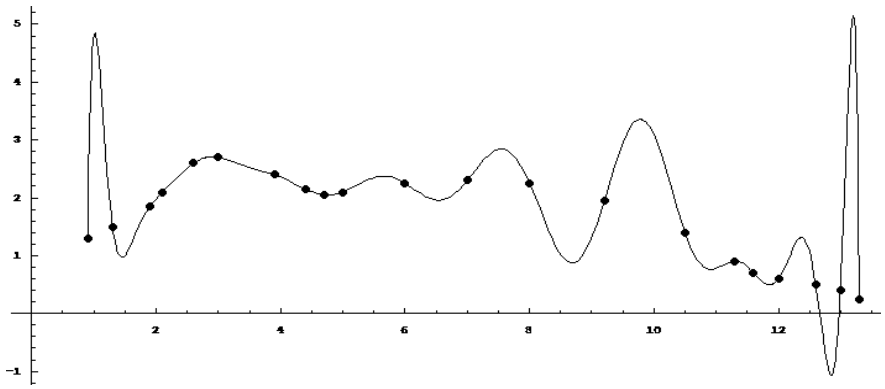


Figura 4.3: Interpolando por polinômio de grau 20

Um polinômio de grau muito alto, tende a oscilar muito, e desta forma em geral não produz o resultado desejado. Podemos neste caso fazer uma interpolação por spline-cúbico, obtendo o

seguinte spline natural,

$$\begin{cases}
p_0(x) = 1.3 + 0.54(x-0.9) + 0(x-0.9)^2 - 0.25(x-0.9)^3 & 0.9 \leq x < 1.3 \\
p_1(x) = 1.5 + 0.42(x-1.3) - 0.3(x-1.3)^2 + 0.95(x-1.3)^3 & 1.3 \leq x < 1.9 \\
p_3(x) = 1.85 + 1.09(x-1.9) + 1.41(x-1.9)^2 - 2.96(x-1.9)^3 & 1.9 \leq x < 2.1 \\
p_4(x) = 2.1 + 1.29(x-2.1) - 0.37(x-2.1)^2 - 0.45(x-2.1)^3 & 2.1 \leq x < 2.6 \\
p_5(x) = 2.6 + 0.59(x-2.6) - 1.04(x-2.6)^2 + 0.45(x-2.6)^3 & 2.6 \leq x < 3.0 \\
p_6(x) = 2.7 - 0.02(x-3.0) - 0.5(x-3.0)^2 + 0.17(x-3.0)^3 & 3.0 \leq x < 3.9 \\
p_7(x) = 2.4 - 0.50(x-3.9) - 0.03(x-3.9)^2 + 0.08(x-3.9)^3 & 3.9 \leq x < 4.4 \\
p_8(x) = 2.15 - 0.48(x-4.4) + 0.08(x-4.4)^2 + 1.31(x-4.4)^3 & 4.4 \leq x < 4.9 \\
p_9(x) = 2.05 - 0.07(x-4.7) + 1.27(x-4.7)^2 - 1.58(x-4.7)^3 & 4.9 \leq x < 5.0 \\
p_{10}(x) = 2.1 + 0.26(x-5.0) - 0.16(x-5.0)^2 + 0.04(x-5.0)^3 & 5.0 \leq x < 6.0 \\
p_{11}(x) = 2.25 + 0.08(x-6.0) - 0.03(x-6.0)^2 + 0.0(x-6.0)^3 & 6.0 \leq x < 7.0 \\
p_{12}(x) = 2.3 + 0.01(x-7.0) - 0.04(x-7.0)^2 - 0.02(x-7.0)^3 & 7.0 \leq x < 8.0 \\
p_{13}(x) = 2.25 - 0.14(x-8.0) - 0.11(x-8.0)^2 + 0.02(x-8.0)^3 & 8.0 \leq x < 9.2 \\
p_{14}(x) = 1.95 - 0.34(x-9.2) - 0.05(x-9.2)^2 - 0.01(x-9.2)^3 & 9.2 \leq x < 10.5 \\
p_{15}(x) = 1.4 - 0.53(x-10.5) - 0.1(x-10.5)^2 - 0.02(x-10.5)^3 & 10.5 \leq x < 11.3 \\
p_{16}(x) = 0.9 - 0.73(x-11.3) - 0.15(x-11.3)^2 + 1.21(x-11.3)^3 & 11.3 \leq x < 11.6 \\
p_{17}(x) = 0.7 - 0.49(x-11.6) + 0.94(x-11.6)^2 - 0.84(x-11.6)^3 & 11.6 \leq x < 12.0 \\
p_{18}(x) = 0.6 - 0.14(x-12.0) - 0.06(x-12.0)^2 + 0.04(x-12.0)^3 & 12.0 \leq x < 12.6 \\
p_{19}(x) = 0.5 - 0.18(x-12.6) + 0.0(x-12.6)^2 - 0.45(x-12.6)^3 & 12.6 \leq x < 13.0 \\
p_{20}(x) = 0.4 - 0.39(x-13.0) - 0.54(x-13.0)^2 + 0.6(x-13.0)^3 & 13.0 \leq x < 13.3
\end{cases}$$

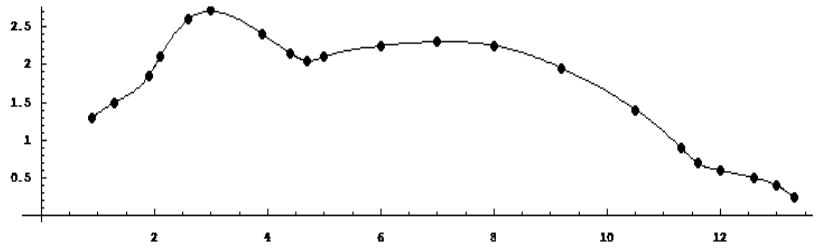


Figura 4.4: Interpolação por Spline-Cúbico Natural

Note que no exemplo acima o spline cúbico natural aproxima melhor a curva que o Restrito. A razão disto é o fato das derivadas terem sido estimadas de forma não precisa. Geralmente as condições de contorno do tipo Restrito são as preferidas quando aproximamos funções por splines cúbicos. Assim as derivadas das funções devem ser estimadas nos pontes final e inicial do intervalo. Quando os nós são igualmente espaçados, por exemplo podemos usar diferenças finitas para frente e para trás ou qualquer outro tipo de aproximação da derivada. No caso de dados não igualmente espaçados o problema fica consideravelmente mais difícil.
Devemos ainda observar que pode ser feito um estudo de limite de erro em livros de análise numérica como Schultz MH, Spline Analysis Prendice Hall 1973. Assim para o Spline cúbico Restrito se $f \in C^4[a,b]$ e com derivada quarta limitada neste intervalo por M e se chamarmos de S o único interpolante spline cúbico de f com respeito aos nós $a = x_0 < x_1 < ... < x_n = b$, então

$$\max_{a \leq x \leq b} |f(x) - S(x)| \leq \frac{5M}{384} \max_{0 \leq j \leq n-1} (x_{j+1} - x_j)^4.$$

Capítulo 5

Ajuste de Mínimos Quadrados

Neste capítulo iremos fazer uma breve introdução ao método de ajuste por mínimos quadrados.

5.1 Ajuste de Mínimos Quadrados para um conjunto discreto de pontos

Ajuste por mínimos quadrados é o procedimento matemático para encontrar a curva de melhor ajuste para um dado conjunto de pontos através da minimização da soma dos quadrados das distâncias entre os pontos e a curva considerada. A soma dos quadrados das distâncias é usada ao invés do valor absoluto, pois desta forma os resíduos poderão ser tratados como quantidades continuamente diferenciáveis. Entretanto, como quadrados de distâncias são usados, pontos muito afastados podem produzir um efeito não proporcional no ajuste. Esta propriedade pode ou não ser desejável depende muito do problema que está sendo resolvido.

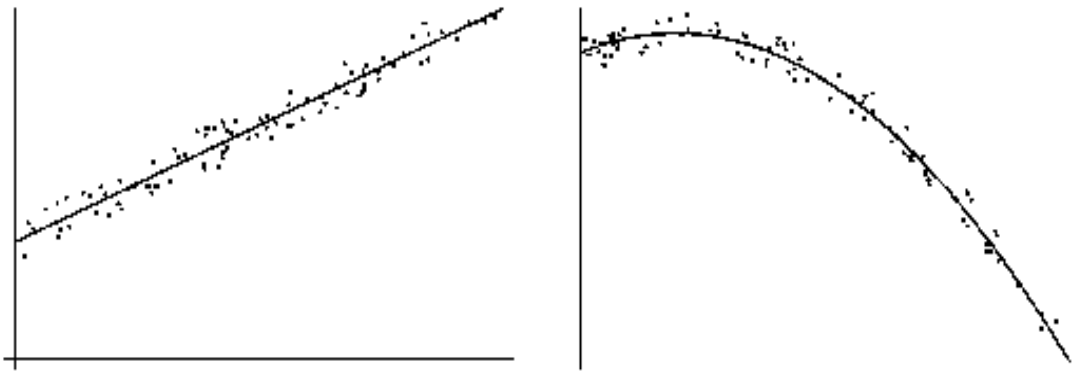


Figura 5.1: Aproximação por mínimos quadrados

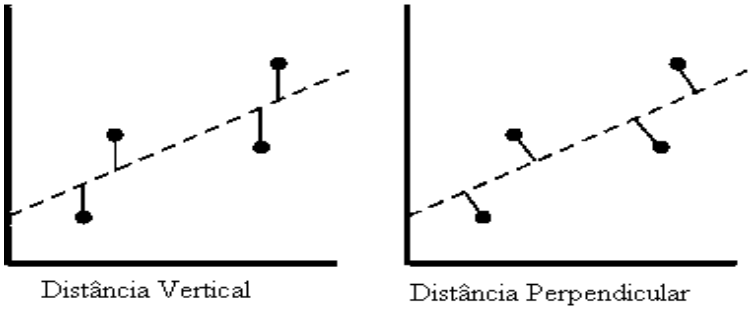


Figura 5.2: Distâncias

Na prática, as distâncias verticais da linha interpoladora são geralmente usadas para serem minimizadas ao invés das distâncias perpendiculares. Este procedimento fornece uma função de ajuste para a variável independente x para estimar o valor de y. Este procedimento permite incerteza nos pontos dos dados ao longo dos eixos x e y que serão simplesmente incorporados, e também fornece uma forma analítica muito mais simples para os parâmetros ajustados que seriam obtidos usando um ajuste baseado nas distâncias perpendiculares. Em adição, quando as distâncias verticais são consideradas, a técnica de ajuste pode ser facilmente generalizada de uma linha de melhor ajuste para um polinômio. De qualquer forma, para um número razoável de pontos de ruídos nos dados, a diferença entre a aproximação vertical e perpendicular é muito pequena.

Desta forma a abordagem de mínimos quadrados para este problema determina a determinação da melhor aproximação quando o erro envolvido é a soma dos quadrados das diferenças entre os valores de y na curva de aproximação e os valores de y dados pela tabela. Assim para caso de ajuste por uma linha reta, conhecidos os pontos $\{x_i, y_i)\}_{i=1}^m$, devemos minimizar a função:

$$E_2(a_0, a_1) = \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2,$$

em relação aos parâmetros a_0 e a_1. Para que um valor de mínimo ocorra, devemos ter que as derivadas de $E_2(a_0, a_1)$ em relação a a_0 e a a_1 sejam nulas, isto é:

$$0 = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2 = 2 \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)](-1)$$

e

$$0 = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2 = 2 \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)](-x_i)$$

Estas equações podem ser simplificadas como:

$$a_0.m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$$
$$a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i$$

ou na forma matricial:

$$\begin{pmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix} \tag{5.1}$$

Exemplo 42 *Vamos considerar os dados apresentados na tabela (42), que além disto calcula os valores necessários para aplicação da fórmula (5.1).*

x_i	y_i	x_i^2	$x_i y_i$	$P(x_i) = 1.538x_i - 0.360$
1	1.3	1	1.3	1.18
2	3.5	4	7.0	2.72
3	4.2	9	12.6	4.25
4	5.0	16	20.0	5.79
5	7.0	25	35.0	7.33
6	8.8	36	52.8	8.87
7	10.1	49	70.7	10.41
8	12.5	64	100.0	11.94
9	13.0	81	117.0	13.48
10	15.6	100	156.0	15.02
55	81.0	385	572.4	$E_2 = \sum_{i=1}^{10}(y_i - P(x_i))^2 \approx 2.34$

Resolvendo o sistema eq1 temos:

$$a_0 = \frac{385(81) - 55(572.4)}{10(385) - 55^2} = -0.360$$

e

$$a_1 = \frac{10(572.4) - 55(81)}{10(385) - 55^2} = 1.538$$

de forma que $P(x) = 1.538x - 0.360$. O gráfico deta curva e os pontos da tabela são mostrados na figura (5.3). A interpolação polinomial dos dados é mostrada na figura (5.4).

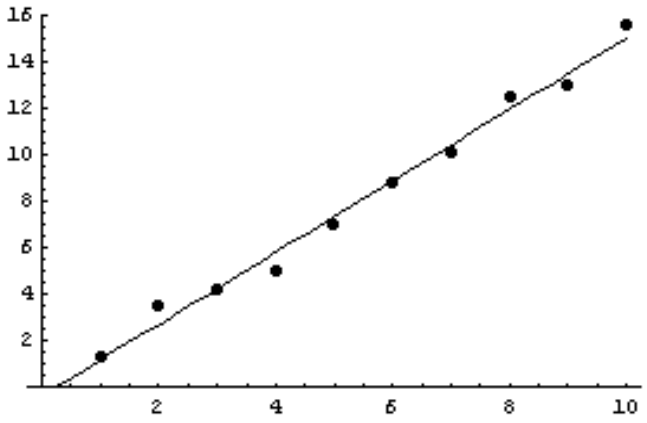


Figura 5.3: Aproximação de Mínimos Quadrados

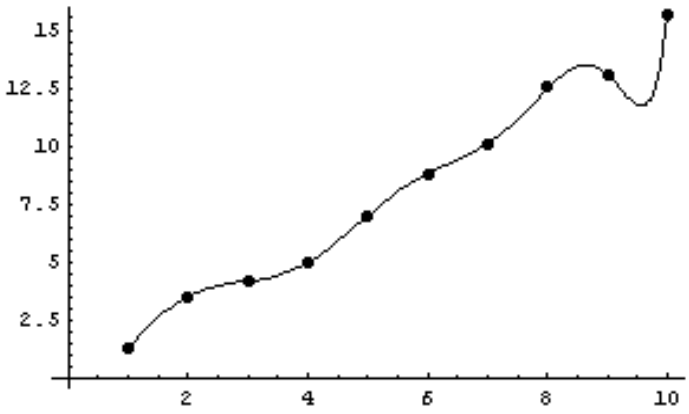


Figura 5.4: Interpolação polinomial

O problema genérico de se aproximar um conjunto de dados $\{x_i, y_i\}_{i=1,m}$ como um polinômio do tipo

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0, \tag{5.2}$$

de grau $n < m - 1$, utilizando o método dos mínimos quadrados é tratado de modo semelhante

ao caso linear. Definimos:

$$\begin{aligned} E_2 = & \sum_{i=1}^{m}(y_i - P_n(x_i))^2 \\ = & \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m} P_n(x_i)y_i + \sum_{i=1}^{m}(P_n(x_i))^2 \\ = & \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n}\left(\sum_{i=1}^{n} a_j x_1^j\right) y_i + \sum_{i=1}^{m}\left(\sum_{j=0}^{n} a_j x_i^j\right)^2 \\ = & \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n} a_j\left(\sum_{i=1}^{m} y_i x_i^j\right) + \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k\left(\sum_{i=1}^{m} x_i^{j+k}\right). \end{aligned}$$

Como no caso linear para minimizar E_2 é preciso resolver um sistema de n equações e n incógnitas definido por $\partial E_2/\partial a_j = 0$ para $j = 0, 1, ..., n$, isto é para cada j temos a equação:

$$0 = \frac{\partial E_2}{\partial a_j} = -2\sum_{i=1}^{m} y_i x_i^j + 2\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k}.$$

Manipulando as equações acima algebricamente e temos:

$$\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k} = \sum_{i=1}^{m} y_i x_i^j, \quad j = 0, 1, ..., n. \tag{5.3}$$

Que na forma matricial podem ser escritas como:

$$\begin{pmatrix} m & \sum_{i=1}^{m} x_i & \dots & \sum_{i=1}^{m} x_i^n \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 & \dots & \sum_{i=1}^{m} x_i^{n+1} \\ \vdots & \ddots & \ddots & \vdots \\ \sum_{i=1}^{m} x_i^n & \sum_{i=1}^{m} x_i^{n+1} & \dots & \sum_{i=1}^{m} x_i^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \\ \vdots \\ \sum_{i=1}^{m} x_i^n y_i \end{pmatrix} \tag{5.4}$$

Exemplo 43 *Ajuste os dados da tabela (43) através de um polinômio quadrático.*

i	x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
1	0	1.0000	0	0	0	0	0
2	0.25	1.2840	0.0625	0.01563	0.003906	0.3210	0.08025
3	0.50	1.6487	0.2500	0.1250	0.062500	0.8244	0.41227
4	0.75	2.1170	0.5625	0.4219	0.316406	1.5878	1.19081
5	1.00	2.7183	1.0000	1.0000	1.000000	2.7183	2.71830
	2.00	8.7680	1.8750	1.5625	1.382812	5.4514	4.40154

Resolvendo o sistema dado por (5.4) temos o seguinte polinômio quadrático:

$$p(x) = 1.00514 + 0.864183x + 0.843657x^2$$

E o gráfico do resultado obtido é mostrado na figura (5.5) mostrada a seguir.

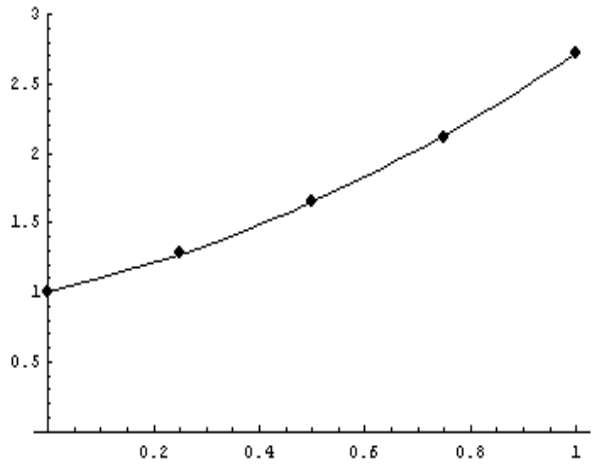


Figura 5.5: Aproximação por mínimos quadrados

O sistema definido por (5.4) pode ser resolvido por técnicas simples como eliminação Gaussiana por exemplo, mas infelizmente o conjunto de equações é muito mal condicionado. O número de equações que pode ser resolvido está severamente limitado na maior parte dos casos por causa dos erros de arredondamento. Usualmente se usarmos precisão simples o sistema terá solução sem sentido para m=7 ou m=8. Um dos problemas é a grande variação de grandezas entre os coeficientes de cada linha. Usar dupla precisão nas máquinas é altamente indicado quando trabalhamos com ajuste por mínimos quadrados.

Por sorte, na prática polinômios de baixa ordem são usados no ajuste de dados. Os polinômios de alto grau, quando usados em ajuste de curvas, tendem a reproduzir ruídos existentes nos dados. Na maior parte das vezes o ajuste de dados é feito através de linhas retas, e muitas vezes os dados são plotados em escalas diferentes (tal como a escala log-log) até os dados assumirem uma forma na qual o ajuste por uma linha reta é uma aproximação razoável.

A escolha do grau do polinômio a ser usado é muitas vezes difícil de escolher. A melhor situação é quando sabemos que os dados devem se encaixar em um polinômio de determinado grau. Julgamentos qualitativos também podem ser usados, por exemplo, se os dados parecem conter uma inflexão então um polinômio cúbico pode ser uma escolha óbvia. Outros métodos baseados na observação de $\sum_{i=0}^{m}(y_i - p(x_i))^2$ ou em análises estatísticas também podem ser usados.

5.1.1 Ajustes não polinomiais

Algumas vezes é necessário ajustar os dados através de uma função exponencial

$$y(x) = Be^{Ax} \qquad \text{ou} \qquad y(x) = Bx^A$$

Para fazermos ajuste de mínimos quadrados com estas funções precisaríamos minimizar a função erro definida por

$$E(A,B) = \sum_{i=1}^{M} (Be^{Ax} - y_i)^2 \qquad \text{ou} \qquad E(A,B) = \sum_{i=1}^{M} (Bx^A - y_i)^2$$

em ambos os casos o sistema de equações resultante desta minimização é não linear que em geral não possui solução exata. Geralmente para evitar esta dificuldade, usamos linearizar este problema e depois aplicar o método de mínimos quadrados sobre o problema linearizado.

$$\ln y = \ln B + Ax \qquad \text{ou} \qquad \ln y = \ln B + A \ln x$$

Assim em qualquer um dos casos reduzimos o problema não linear para um problema linear, onde podemos adaptar a fórmula de um polinômio linear.

Nota 3 *A aproximação obtida desta maneira, não é a aproximação de mínimos quadrados para o problema original, e em alguns casos pode ser bem distinta desta, porém em geral esta é a forma usada na prática.*

5.1.2 Alguns exercícios

1. Vamos usar mínimos quadrados (na forma linearizada) para encontrar o ajuste para o seguinte conjunto de pontos:

x_j	0.2	0.3	0.6	0.9	1.1	1.3	1.4	1.6
y_j	0.050446	0.098426	0.33277	0.72660	1.09782	1.5697	1.8487	2.5015

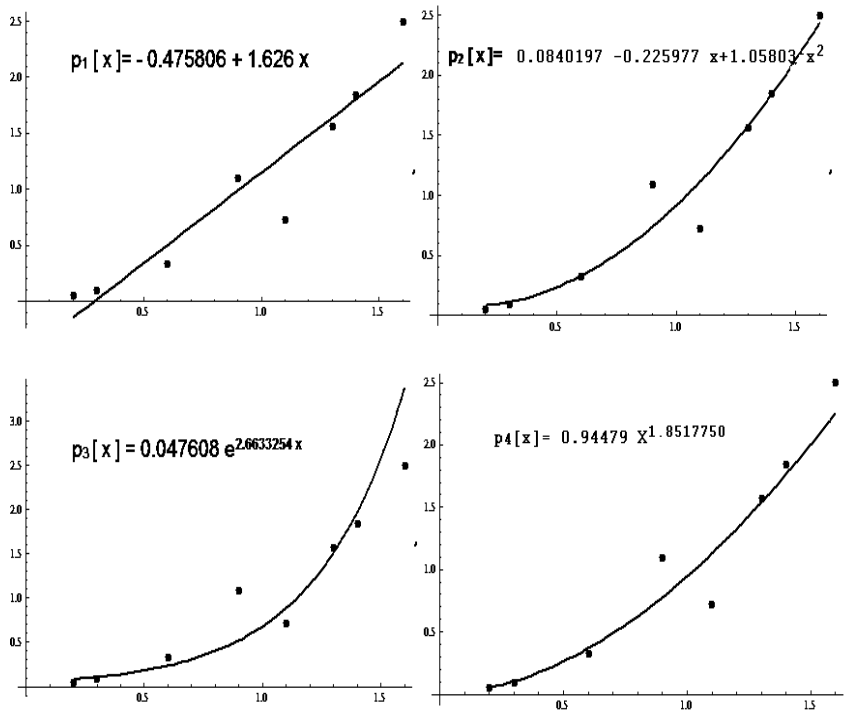
   (a) Construa o ajuste de MQ para os dados como um polinômio linear

   (b) construa o ajuste de MQ como um polinômio quadrático

   (c) construa o ajuste de MQ na forma be^{ax} (faça linearizando)

   (d) construa o ajuste de MQ na forma bx^a (faça linearizando)

   A resposta está dada na figura abaixo.



2. A lei de Hooke afirma que quando uma força é aplicada a uma mola constituída de um material uniforme o comprimento da mola é uma função linear desta força e é descrita por:

$$F(\ell) = \kappa(\ell - E)$$

onde $F(\ell)$ representa a força requerida para dsitender a mola ℓ unidades de comprimento, E é uma constante e representa o comprimento da mola quando nenhuma força é aplicada e finalmente κ é a constante da mola.

(a) Encontre uma aproximação para κ, pelo método dos mínimos quadrados, quando temos os dados da tabela:

$F(\ell)$	ℓ
2	7.0
4	9.4
6	12.3

(b) Se as seguintes medidas adicionais são feitas, qual a nova estimativa para κ usando o método dos mínimos quadrados?

$F(\ell)$	ℓ
3	8.3
5	11.3
8	14.4
10	15.9

5.2 A aproximação de funções contínuas

Outro problema de aproximação diz respeito a funções contínuas. A melhor aproximação para funções contínuas são usualmente consideradas como sendo aproximações que minimizam o erro no sentido do *minimax*. Infelizmente isto muitas vezes é difícil de encontrar para certas classes de funções e nós temos que trocar para a melhor aproximação da melhor aproximação. Por exemplo ao invés de encontramos a melhor a melhor quadrática que aproxima certa função, teremos que nos satisfazer com a quadrática que é suficientemente próxima da melhor quadrática. Boas aproximações paar funções contínuas usualmente têm um erro $d(x) = |f(x) - g(x)|$ o qual oscila em torno de zero na região de interesse na forma que os picos positivos são aproximadamente iguais aos picos negativos. Tal comportamento é muitas vezes chamado de comportamento *minimax* mesmo a aproximação não sendo a melhor em sua classe. Os métodos de aproximação considerados possuem este comportamento desejável.

A forma mais simples e mais comum de aproximação para uma função contínua é o tipo polinomial. De fato, sempre que uma representação em série de potências é usada para calcular uma função, então uma aproximação polinomial está sendo usada uma vez que a série deve ser truncada em algum ponto e uma série de potências truncada é um polinômio.

Vamos supor que $f \in \mathcal{C}[a, b]$ e que um polinômio de grau $P_n(x)$ de grau no mínimo n seja necessário para minimizar o erro;

$$\int_a^b [f(x) - P_n(x)]^2 dx.$$

Vamos definir a função erro como:

$$E = E(a_0, a_1, ..., a_n) = \int_a^b \left(f(x) - \sum_{k=0}^{n} a_k x^k \right)^2 dx.$$

O problema é encontrar os coeficientes reais $a_0, a_1, ..., a_n$ de forma que minimizem E. Uma condição necessária para que os números $a_0, a_1, ..., a_n$ minimizem E é que:

$$\frac{\partial E}{\partial a_j} = 0, \text{ para cada } j = 0, 1, ...n.$$

Como

$$E = \int_a^b (f(x))^2 dx - 2 \sum_{k=0}^{n} a_k \int_a^b x^k f(x) dx + \int_a^b \left(\sum_{k=0}^{n} a_k x^k \right)^2 dx,$$

assim temos:

$$\frac{\partial E}{\partial a_j} = -2\int_a^b x^j f(x)dx + 2\sum_{k=0}^{n}\int_a^b x^{j+k}dx.$$

Desta forma temos o sistema de equações:

$$\sum_{k=0}^{n} a_k \int_a^b x^{j+k}dx = \int_a^b x^j f(x)dx, \text{ para cada } j = 0, 1, ...n, \tag{5.5}$$

que devem ser resolvidas para as $n+1$ incógnitas a_j. Devemos observar que o sistema (5.5) sempre possuirá solução única se $f \in \mathcal{C}[a,b]$.

Exemplo 44 *Encontre o polinômio de aproximação dos mínimos quadrados de grau 2 para a função* $f(x) = \sin \pi x$ *no intervalo* $[0,1]$.
Solução: $P_2(x) = a_2x^2 + a_1x + a_0$ *e o sistema é dados por:*

$$\begin{aligned}
a_0\int_0^1 1dx + a_1\int_0^1 xdx + a_2\int_0^1 x^2dx &= \int_0^1 \sin \pi x dx,\\
a_0\int_0^1 xdx + a_1\int_0^1 x^2dx + a_2\int_0^1 x^3dx &= \int_0^1 x\sin \pi x dx,\\
a_0\int_0^1 x^2dx + a_1\int_0^1 x^3dx + a_2\int_0^1 x^4dx &= \int_0^1 x^2\sin \pi x dx.
\end{aligned}$$

Calculando as integrais temos:

$$\begin{aligned}
a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 &= \frac{2}{\pi}\\
\frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 &= \frac{1}{\pi}\\
\frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 &= \frac{\pi^2-4}{\pi^3}
\end{aligned}$$

Resolvendo o sistema temos:

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx - = 0.050465 \quad e \quad a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx - = 4.12251.$$

Assim no intervalo $[0,1]$ *a função* $f(x) = \sin \pi x$ *possui aproximação de mínimos quadrados de grau 2 dada por* $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$.

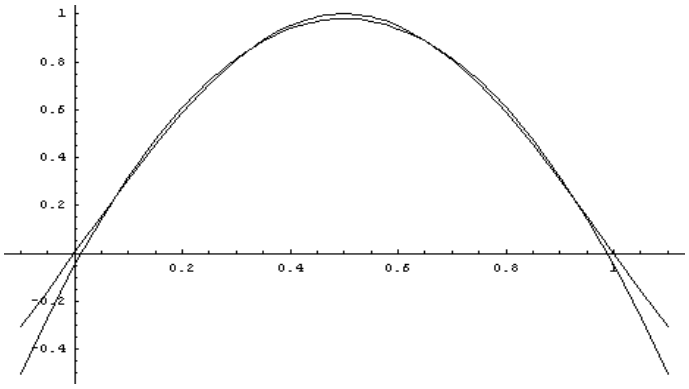


Figura 5.6: Aproximação por mínimos quadrados - exemplo

O exemplo acima mostra as dificuldades para obtermos esta aproximação de mínimos quadrados. Para a montagem do sistema de ordem $n+1$ os coeficientes para cada incógnita $a_0, ..., a_n$ é da forma

$$\int_a^b x^{j+k} dx = \frac{b^{j+k+1} - a^{j+k+1}}{j+k+1}$$

e além disto a matriz associada a este sistema é muito mal condicionada, chegando a ser um exemplo clássico e é chamada de **matriz de Hilbert**. Outra desvantagem é que para obtermos o $P_{n+1}(x)$ de nada ajudaria já conhecermos $P_n(x)$, devendo os cálculos serem feitos todos novamente. Existe uma técnica que é muito mais eficiente para fazer a aproximação de mínimos quadrados e sua principal vantagem é que conhecido $P_n(x)$ é fácil determinar $P_{n+1}(x)$. Esta nova forma envolve o conceito de **polinômios ortogonais**.

- Um conjunto de funções é dito Linearmente independente (LI) em um intervalo $[a, b]$ se sempre que

  $$c_0\varphi_0(x) + c_1\varphi_1(x) + ... + c_n\varphi_n(x) = 0,$$

  para todo $x \in [a, b]$, só existe a solução trivial, isto é $c_0 = c_1 = ... = c_n = 0$.

- Podemos mostrar que se $\varphi_j(x)$ é um polinômio de grau j, para $j = 0, 1, ..., n$, então o conjunto $\{\varphi_0, \varphi_1, ..., \varphi_n\}$ é um conjunto LI de funções em qualquer intervalo $[a, b]$.

- Vamos chamar de $\prod_n$ o conjunto de todos os polinômios de grau menor ou igual à n. Podemos mostrar que se $\mathbf{V} = \{\varphi_0, \varphi_1, ..., \varphi_n\}$ é um conjunto de polinômios LI de $\prod_n$ então qualquer polinômio de grau menor ou igual a n pode ser escrito de forma única como combinação linear dos elementos de $\mathbf{V}$.

- Dizemos que uma função integrável $\mathbf{w}$ é uma função *peso* no intervalo I se $w(x) \geq 0$, para todo x em I, sendo que $w(x) \neq 0$ em qualquer subintervalo de I. O objetivo desta função peso é atribuir graus de importância distintos às aproximações em certas

porções do intervalo. Por exemplo, no intervalo $(0,1)$ a função $w(x) = 1/\sqrt{1-x^2}$, que aparece associada aos polinômios de Chebychev, enfatiza aos pontos perto dos extremos do intervalo e menor aos pontos no interior do intervalo.

- Um conjunto de funções $\{\varphi_0, \varphi_1, ..., \varphi_n\}$ é dito ortogonal no intervalo $[a,b]$ em relação à uma função peso w se

$$\int_a^b w(x)\varphi_j(x)\varphi_k(x)dx = \begin{cases} 0, & \text{se } j \neq k \\ \alpha_k > 0 & \text{se } j = k \end{cases} .$$

  Se de forma adicional, $\alpha_k = 1$ para cada $k = 0, 1, ..., n$ então o conjunto é dito ortonormal.

- Vamos agora supor que temos um conjunto de funções ortogonais em relação a uma função peso $w(x)$ em um intervalo $[a,b]$ e seja uma função $f \in \mathcal{C}[a,b]$ se queremos aproximar

$$f(x) = \sum_{k=0}^{n} a_k \varphi_k(x)$$

  de forma que o erro quadrático

$$E(a_0, a_1, ..., a_n) = \int_a^b w(x) \left(f(x) - \sum_{k=0}^{n} a_k \varphi_k(x) \right)^2$$

  seja mínimo. Derivando as equações acima temos

$$0 = \frac{\partial E}{\partial a_j} = 2 \int_a^b w(x) \left(f(x) - \sum_{k=0}^{n} a_k \varphi_k(x) \right) \varphi_j(x) dx$$

  Assim das equações acima temos:

$$\int_a^b w(x) f(x) \varphi_j(x) dx = \sum_{k=0}^{n} a_k \int_a^b w(x) \varphi_k(x) \varphi_j(x) dx$$

  usando a ortogonalidade com relação a função peso das funções temos que:

$$a_j = \frac{\int_a^b w(x) f(x) \varphi_j(x) dx}{\int_a^b w(x) \varphi_k(x) \varphi_j(x) dx}$$

  E desta forma o problema de aproximação de uma função por mínimos quadrados fica muito simplificado.

  Devemos ainda observar que aqui apenas conjuntos de polinômios ortogonais foram considerados, e que podemos construir um conjunto de polinômios ortogonais com rel ção a uma função peso $w(x)$ em $[a,b]$ através do uso do processo de *Gram-Schmidt*.

Capítulo 6

Diferenças Finitas

No cálculo convencional a operação de diferenciação de uma função é um procedimento formal bem definido com operações altamente dependentes da forma da função envolvida. Muitos tipos de regras distintas são necessárias para funções distintas. Nos métodos numéricos em um microcomputador apenas podemos empregar as operações $+ - */$ e certas operações lógicas. Desta forma precisamos desenvolver uma técnica para diferenciar funções usando apenas estas operações. Para isto vamos estudar um pouco de cálculo de Diferenças Finitas.

6.1 Diferenças para Frente e para Trás (Forward / Backward)

Vamos considerar $f(x)$ uma função analítica num intervalo aberto contendo um ponto x, usando a série de Taylor para $f(x)$ (1.1) fazendo $x = x + h$ e $a = x$, temos que:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x)\ldots \tag{6.1}$$

isolando $f'(x)$ na equação acima temos:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}f''(x) - \frac{h^2}{6}f'''(x) - \ldots \tag{6.2}$$

e desta forma temos que,

$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h) \tag{6.3}$$

Notação: $f(x+ih) = f_{j+i}$ e $f(x) = f_j$.
Usando a notação acima:

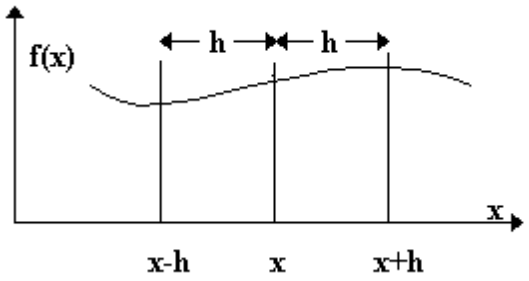


Figura 6.1: Vizinhança de x

$$f'(x) = \frac{f_{j+1} - f_j}{h} + O(h) = \frac{\triangle f_j}{h} + O(h) \tag{6.4}$$

onde $\triangle f_j = f_{j+1} - f_j$.
A expressão para $f'(x)$ pode agora ser escrita como:

$$f'(x) = \frac{\triangle f_j}{h} + O(h) \tag{6.5}$$

O termo $\triangle f_j/h$ é chamado de aproximação em diferença para frente de primeira ordem. Graficamente $(f_{j+1} - f_j)/h$ aproxima a declividade da função f em um ponto x pela declividade da linha reta que passa através de $f(x+h)$ e $f(x)$.
Para calcularmos a fórmula de diferenças para trás, fazemos um raciocínio similar expandindo $f(x-h)$. isto é:

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f'''(x) - \ldots \tag{6.6}$$

resolvendo para $f'(x)$ e truncando a série temos:

$$f'(x) = \frac{f(x) - f(x-h)}{h} + O(h) \tag{6.7}$$

ou,

$$f'(x) = \frac{f_j - f_{j-1}}{h} + O(h) = \frac{\nabla f_j}{h} + O(h) \tag{6.8}$$

onde $\nabla f_j = f_j - f_{j-1}$.
O termo $\nabla f_j/h$ é chamada de aproximação de primeira ordem em diferenças para trás com erro de ordem h para $f'(x)$. A interpretação geométrica da aproximação é a declividade da linha reta que une $f(x)$ e $f(x-h)$.

Note dos termos de erro em (6.2) e (6.7) que ambas as aproximações em diferenças para frente e para trás são exatas para linhas retas (uma vez que o termo de erro não envolve $f'(x)$ mas são somente aproximações para outras funções onde $f''(x)$ e derivadas maiores são não nulas. Agora, vamos obter aproximações de ordens maiores.

6.1.1 Cálculo de derivadas de ordem superior

Vamos expandir $f(x+2h)$ em série de Taylor, assim

$$f(x+2h) = f(x) + 2hf'(x) + \frac{(2h)^2}{2!}f''(x) + \frac{(2h)^3}{3!}f'''(x) + \dots \tag{6.9}$$

Agora, fazendo (6.9) menos duas vezes (6.1) temos,

$$f(x+2h) - 2f(x+h) = -f(x) + h^2 f''(x) + \frac{h^3}{3}f'''(x) + \dots \tag{6.10}$$

isolando $f''(x)$

$$f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} - hf'''(x) + \dots \tag{6.11}$$

$$f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} + O(h) = \frac{f_{j+2} - 2f_{j+1} + f_j}{h^2} + O(h) \tag{6.12}$$

Assim encontramos uma expressão para derivadas de f com respeito a x de segunda ordem e com erro da ordem de h. A diferença de segunda ordem para frente de f em j é definida como:

$$\triangle^2 f_j \equiv f_{j+2} - 2f_{j+1} + f_j \tag{6.13}$$

e nós podemos reescrever (6.12) para $f''(x)$ como,

$$f''(x) = \frac{f_{j+2} - 2f_{j+1} + f_j}{h^2} + O(h) = \frac{\triangle^2 f_j}{h^2} + O(h) \tag{6.14}$$

Fazendo raciocínio similar para diferenças para trás, isto é expandindo $f(x-2h)$ obtemos:

$$f''(x) = \frac{\nabla^2 f_j}{h^2} + O(h) \tag{6.15}$$

onde a diferença de segunda ordem para trás de f em j é definida como:

$$\nabla^2 f_j = f_j - 2f_{j-1} + f_{j-2} \tag{6.16}$$

Para generalizarmos o resultado para derivadas de maior ordem, definimos a fórmula de recorrência,

$$\begin{aligned} \triangle^n f_j &= \triangle(\triangle^{n-1} f_j) \\ \nabla^n f_j &= \nabla(\nabla^{n-1} f_j) \end{aligned}$$

Podemos encontrar qualquer diferença finita, tomando as diferenças da diferença finita de ordem imediatamente inferior. Por exemplo, a diferença segunda para trás de f em j pode ser encontrada como:

$$\nabla^2 f_j = \nabla(\nabla f_j) = \nabla f_j - \nabla f_{j-1} = (f_j - f_{j-1}) - (f_{j-1} - f_{j-2}) = f_j - 2f_{j-1} + f_{j-2} \tag{6.17}$$

E desta forma,

$$\begin{aligned} \left.\frac{d^n f}{dx^n}\right|_{x_j} &= \frac{\triangle^n f_j}{h^n} + O(h) \qquad dif.\ para\ \ frente \\ \left.\frac{d^n f}{dx^n}\right|_{x_j} &= \frac{\nabla^n f_j}{h^n} + O(h) \qquad dif.\ para\ \ trás \end{aligned}$$

Nota: Todas estas expressões que foram derivadas até agora são $O(h)$.
Abaixo as expressões de diferenças finitas para frente e para trás para $O(h)$ estão tabuladas até quarta ordem.

Tabelas para as fórmulas $O(h)$ de diferenças

FORWARD formulas

	f_j	f_{j+1}	f_{j+2}	f_{j+3}	f_{j+4}	
$hf'(x)$	-1	1				
$h^2 f''(x)$	1	-2	1			$+O(h)$
$h^3 f'''(x)$	-1	3	-3	1		
$h^4 f^{(4)}(x)$	1	-4	6	-4	1	

BACKWARD formulas

	f_{j-4}	f_{j-3}	f_{j-2}	f_{j-1}	f_j	
$hf'(x)$				-1	1	
$h^2 f''(x)$			1	-2	1	$+O(h)$
$h^3 f'''(x)$		-1	3	-3	1	
$h^4 f^{(4)}(x)$	1	-4	6	-4	1	

6.1.2 Expressões em diferenças com ordem de erro mais alta

As fórmulas (6.2) e (6.11) nos dão expressões em diferenças para frente de $f'(x)$ e $f''(x)$ respectivamente. Substituindo (6.11) em (6.2) temos:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}\left(\frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} - hf'''(x) + \dots\right) - \frac{h^2}{6}f'''(x) + \dots \tag{6.18}$$

logo,

$$f'(x) = \frac{-f(x+2h) + 4f(x+h) - 3f(x)}{2h} + \frac{h^2}{3}f'''(x) - \dots \tag{6.19}$$

assim,

$$f'(x) = \frac{-f_{j+2} + 4f_{j+1} - 3f_j}{2h} + O(h^2) \tag{6.20}$$

Esta é a representação da derivada primeira com uma precisão de $O(h^2)$. Note que como o erro para esta fórmula envolve a derivada terceira, ela é exata para um polinômio de segundo grau.

Fórmulas para frente e para trás, apesar de raramente usadas na prática podem ser obtidas através de substituições sucessivas de mais termos na expressão da série de Taylor por expressões de diferenças finitas de $O(h)$. Este procedimento também acarreta erros, pois cada termo substituido possui um termo de erro que automaticamente contribui para o próximo termo de erro de ordem maior.

Tabelas para as fórmulas $O(h^2)$ de diferenças

FORWARD formulas

	f_j	f_{j+1}	f_{j+2}	f_{j+3}	f_{j+4}	f_{j+5}	
$2hf'(x)$	-3	4	-1				
$h^2f''(x)$	2	-5	4	-1			$+O(h^2)$
$2h^3f'''(x)$	-5	18	-24	14	-3		
$h^4f^{(4)}(x)$	3	-14	26	-24	11	-2	

BACKWARD formulas

	f_{j-5}	f_{j-4}	f_{j-3}	f_{j-2}	f_{j-1}	f_j	
$2hf'(x)$				1	-4	3	
$h^2f''(x)$			-1	4	-5	2	$+O(h^2)$
$2h^3f'''(x)$		3	-14	24	-18	5	
$h^4f^{(4)}(x)$	-2	11	-24	26	-14	3	

6.2 Diferenças Centrais

Seja $f(x)$ uma função analítica. Vamos expandir $f(x+h)$ e $f(x-h)$ em série de Taylor, assim

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x)\ldots \tag{6.21}$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f'''(x) - \ldots \tag{6.22}$$

Diminuindo (6.22) de (6.21) temos:

$$f(x+h) - f(x-h) = 2hf'(x) + \frac{h^3}{3}f'''(x) + \ldots \tag{6.23}$$

Agora isolando $f'(x)$ temos,

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + O(h^2) = \frac{f_{j+1} - f_{j-1}}{2h} + O(h^2) \tag{6.24}$$

Note que o erro é $O(h^2)$, logo é exata para parabolas. Da mesma forma que na seção anterior podemos obter expressões para derivadas de ordem superior, e a fórmulas são:

$$f''(x) = \frac{f_{j+1} - 2f_j + f_{j-1}}{h^2} + O(h^2) \tag{6.25}$$

Generalizando temos as fórmulas:

$$\begin{cases} \frac{d^n f}{dx^n} = \frac{\nabla^n f_{j+n/2} + \triangle^n f_{j-n/2}}{2h^n} + O(h^2) & se\ \ n\ \ par \\ \frac{d^n f}{dx^n} = \frac{\nabla^n f_{j+(n-1)/2} + \triangle^n f_{j-(n-1)/2}}{2h^n} + O(h^2) & se\ \ n\ \ \acute{i}mpar \end{cases} \tag{6.26}$$

Tabelas para as fórmulas $O(h^2)$ e $O(h^4)$ de diferenças centrais

	f_{j-2}	f_{j-1}	f_j	f_{j+1}	f_{j+2}	
$2hf'(x)$		-1	0	1		
$h^2f''(x)$		1	-2	1		$+O(h^2)$
$2h^3f'''(x)$	-1	2	0	-2	1	
$h^4f^{(4)}(x)$	1	-4	6	-4	1	

	f_{j-3}	f_{j-2}	f_{j-1}	f_j	f_{j+1}	f_{j+2}	f_{j+3}	
$hf'(x)$		1	-8	0	8	-1		
$h^2f''(x)$		-1	16	-30	16	-1		$+O(h^4)$
$h^3f'''(x)$	1	-8	13	0	-13	8	-1	
$h^4f^{(4)}(x)$	-1	12	-39	56	-39	12	-1	

Expressões em diferenças de $O(h)$ podem ser obtidas empregando operações com séries de Taylor, mas não faremos aqui por ser muito extenso. Estas expressões estão tabuladas acima.

6.3 Diferenças e Polinômios

Expressões em diferenças para derivadas e polinômios possuem algumas relações diferentes que podem ser muito úteis. pois por exemplo, para diferença para frente temos:

- $f'(x) = \frac{f(x+h) - 2f(x)}{h} - \frac{h}{2}f''(x) - \frac{h^2}{6}f'''(x) - ...$
- $f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} - hf'''(x) - ...$
- ...

Assim, como podemos ver acima, o termo de erro para uma enésima diferença apenas envolverá derivadas de ordem $n+1$ ou maiores. Desta forma se consideramos um polinômio de grau n, sua enésima representação em diferença tomada em qualquer ponto deste polinômio será constante e exatamente igual a sua derivada de enésima ordem $(n!a_n)$, uma vez que os termos envolvidos no erro serão zeros.
Este conhecimento poderá ser usado para nos dar alguma idéia de o quão bem um dado polinômio pode ajustar dados obtidos em uma série de pontos igualmente espaçados em uma variável independente. Por exemplo, se a terceira diferença tomada em vários valores de uma variável independente são muito próximos e suas quarta diferenças são quase nulas, então um polinômio cúbico irá ajustar bem estes dados. Este procedimento será discutido melhor no próximo capítulo.

Exemplo 45 *Vamos considerar seguintes dados igualmente espaçados,*

x	*0*	*1*	*2*	*3*	*4*
$f(x)$	*30*	*33*	*28*	*12*	*-22*

Encontre $f'(0), f'(2), f'(4)$ *e* $f''(0)$ *usando representações* $O(h^2)$.

Solução: Em $x = 0$ *temos que usar representação em diferença para frente uma vez que não temos disponibilidade de dados para* $x < 0$. *Usando a tabela temos que,*

$$f'(x) = \frac{-3f_j + 4f_{j+1} - f_{j+2}}{2h} + O(h^2) \tag{6.27}$$

logo,

$$f'(0) = \frac{-f(2) + 4f(1) - 3f(0)}{2(1)} = \frac{-28 + 4(33) - 3(30)}{2} = -7 \tag{6.28}$$

Em $x = 2$, podemos escolher entre muitas representações. Nós vamos escolher diferença central que possui $O(h^2)$. Assim:

$$f'(x) = \frac{f_{j+1} - f_{j-1}}{2h} + O(h^2) \tag{6.29}$$

logo,

$$f'(2) = \frac{f(3) - f(1)}{2} = \frac{12 - 33}{2} = -10.5 \tag{6.30}$$

Em $x = 4$, devemos empregar representação em diferença para trás de $O(h^2)$. Assim:

$$f'(x) = \frac{f_{j-2} - 4f_{j-1} + 3f_j}{2h} + O(h^2) \tag{6.31}$$

logo,

$$f'(4) = \frac{f(2) - 4f(3) + 3f(4)}{2(1)} = \frac{28 - 4(12) + 3(-22)}{2} = -43 \tag{6.32}$$

Exemplo 46 *Os seguintes dados representam um polinômio. Qual seu grau? Qual o coeficiente de seu termo de mais alto grau?*

x	*0*	*1*	*2*	*3*	*4*	*5*
$f(x)$	*1*	*0.5*	*8.0*	*35.5*	*95.0*	*198.5*

Vamos usar diferença para frente em cada ponto (poderia ser usado para trás).

$$\begin{aligned} &\textit{Primeira diferença} \\ \triangle f_0 &= f_1 - f_0 = 0.5 - 1 = -0.5 \\ \triangle f_1 &= f_2 - f_1 = 8.0 - 0.5 = 7.5 \\ \triangle f_2 &= f_3 - f_2 = 35.5 - 8.0 = 27.5 \\ \triangle f_3 &= f_4 - f_3 = 95.0 - 35.5 \\ \triangle f_4 &= f_5 - f_4 = 198.5 - 95 = 103.5 \end{aligned} \tag{6.33}$$

$$\begin{aligned} &\textit{Segunda diferença} \\ \triangle^2 f_0 &= \triangle f_1 - \triangle f_0 = 7.5 - (-0.5) = 8.0 \\ \triangle^2 f_1 &= \triangle f_2 - \triangle f_1 = 27.5 - 7.5 = 20.0 \\ \triangle^2 f_2 &= \triangle f_3 - \triangle f_2 = 59.5 - 27.5 = 32.0 \\ \triangle^2 f_3 &= \triangle f_4 - \triangle f_3 = 103.5 - 59.5 = 44.0 \end{aligned} \tag{6.34}$$

$$\begin{array}{r} \textit{Terceira diferença} \\ \triangle^3 f_0 = \triangle^2 f_1 - \triangle^2 f_0 = 20.0 - 8.0 = 12.0 \\ \triangle^3 f_1 = \triangle^2 f_2 - \triangle^2 f_1 = 32.0 - 20.0 = 12.0 \\ \triangle^3 f_2 = \triangle^2 f_3 - \triangle^2 f_2 = 44.0 - 32.0 = 12.0 \end{array} \tag{6.35}$$

Uma vez que as terceiras diferenças são constantes o polinômio é de terceiro grau. Para diferenças para frente temos que:

$$\frac{d^3 f}{dx^3} = \frac{\triangle^3 f}{h^3} + O(h) \tag{6.36}$$

Para um polinômio de grau 3 esta expressão é exata, logo

$$\frac{d^3 f}{dx^3} = \frac{\triangle^3 f}{h^3} = \frac{12}{1} = 12 = 3!(2) \tag{6.37}$$

Finalmente, por integração vemos que $f = 2x^3 + C_1 \frac{x^2}{2} + C_2 x + C_3$.

6.4 Análise do Erro

Um assunto importante no estudo de diferenciação numérica é o efeito do erro de arredondamento. Para mostrar o que ocorre, vamos examinar a fórmula de diferença central dada por:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} - \frac{h^2}{6} f'''(\xi_1). \tag{6.38}$$

Vamos supor que no cálculo de $f(x_0 + h)$ e de $f(x_0 - h)$, encontramos erros de arredondamento $e(x_0 + h)$ e $e(x_0 - h)$; isto é por causa do arredondamento feito pela máquina valores $\bar{f}(x_0 + h)$ e $\bar{f}(x_0 - h)$ foram usados ao invés dos reais valores de $f(x_0 + h)$ e de $f(x_0 - h)$, ou

$$\begin{array}{l} f(x_0 + h) = \bar{f}(x_0 + h) + e(x_0 + h) \\ f(x_0 - h) = \bar{f}(x_0 - h) + e(x_0 - h). \end{array} \tag{6.39}$$

Neste caso o erro total na aproximação, será a soma dos erros causados, pelo arredondamento, $\frac{e(x_0+h)-e(x_0-h)}{2h}$ e pelo truncamento $\frac{h^2}{6} f'''(\xi_1)$, isto é:

$$f'(x_0) - \frac{\bar{f}(x_0 + h) - \bar{f}(x_0 - h)}{2h} = \frac{e(x_0 + h) - e(x_0 - h)}{2h} - \frac{h^2}{6} f'''(\xi_1). \tag{6.40}$$

Se nós assumirmos que os erros de arredondamento $e(x_0 \pm h)$ são limitados por um número $\epsilon > 0$, e que a terceira derivada de f é limitada por um número $M > 0$, então

$$\left| f'(x_0) - \frac{\bar{f}(x_0 + h) - \bar{f}(x_0 - h)}{2h} \right| \leq \frac{\epsilon}{h} + \frac{h^2}{6} M. \tag{6.41}$$

Assim pela fórmula (6.41) que para reduzir o erro de truncamento, $\frac{h^2}{6}M$ devemos diminuir o valor de h, mas, ao mesmo tempo, quando h é diminuído o erro de arredondamento, $\frac{\epsilon}{h}$, aumenta. Na prática raramente é vantagem fazer h muito pequeno, pois neste caso o erro de arredondamento iria dominar os cálculos.

Exemplo 47 *Considere a aproximação de* $f'(0.900)$ *para* $f(x) = \sin x$*, usando os valores dados na tabela abaixo. O valor real desta derivada é* $\cos(0.900) = 0.62161$.

x	$\sin x$	x	$\sin x$
0.800	*0.71736*	*0.901*	*0.78395*
0.850	*0.75128*	*0.902*	*0.78457*
0.880	*0.77074*	*0.905*	*0.78643*
0.890	*0.77707*	*0.910*	*0.78950*
0.895	*0.78021*	*0.920*	*0.79560*
0.898	*0.78208*	*0.950*	*0.81342*
0.899	*0.78270*	*1.000*	*0.84147*

Para fazer os cálculos amos usar a fórmula

$$f'(0.900) \approx \frac{f(0.900 + h) - f(0.900 - h)}{2h}$$

para diversos valores de h*. Os resultados obtidos são dados na tabela abaixo.*

h	*Aprox. para* $f'(0.900)$	*Erro*
0.001	*0.62500*	*0.00339*
0.002	*0.62250*	*0.00089*
0.005	*0.62200*	*0.00039*
0.010	*0.62150*	*-0.00011*
0.020	*0.62150*	*-0.00011*
0.050	*0.62140*	*-0.00021*
0.100	*0.62055*	*-0.00106*

Pelo exemplo anterior, parece que uma ótima escolha para h está entre 0.005 e 0.05. Se fizermos alguma análise sobre o termo de erro dado pela equação (6.41), isto é

$$e(h) = \frac{\epsilon}{h} + \frac{h^2}{6}M, \tag{6.42}$$

podemos usar o cálculo para verificar que o mínimo da função $e(h)$ ocorre quando sua derivada é zero, isto é:

$$e'(h) = -\frac{\epsilon}{h^2} + \frac{h}{3}M \tag{6.43}$$

assim

$$\begin{aligned} e'(h) = 0 &\iff \frac{\epsilon}{h^2} = \frac{hM}{3} \\ &\iff h = \sqrt[3]{\frac{3\epsilon}{M}} \end{aligned} \tag{6.44}$$

No nosso caso f é conhecida, assim podemos calcular o máximo valor que ela assume em um intervalo contendo o ponto de interesse, $x = 0.009$. No caso vamos considerar o interval $[0.008, 1.000]$ e neste caso,

$$M = \max_{x\in[0.008,1.000]} |f'''(x)| = \max_{x\in[0.008,1.000]} |\cos x| \approx 0.69671. \tag{6.45}$$

Uma vez que os valores de f são dados com cinco casas decimais, é razoável assumir que $\epsilon = 0.000005$. Desta forma a escolha ótima para h é aproximadamente

$$h = \sqrt[3]{\frac{3(0.000005)}{0.69671}} \approx 0.028, \tag{6.46}$$

o que é consistente com os nossos resultados.

Entretanto, na prática, não podemos calcular o valor ótimo de h para ser usado na aproximação da derivada, uma vez que não conhecemos sua derivada terceira.

Apesar de termos apenas considerado o erro de arredondamento no caso de diferença central, dificuldades similares ocorrem com todas as fórmulas de diferenciação. A razão para estes problemas é a necessidade de divisão por potências de h, e como já sabemos divisão por números muito pequenos tendem a exagerar o erro de arredondamento.

Tenha sempre em mente que como um método de aproximação a diferenciação numérica é *instável*, uma vez que os pequenos valores de h necessários para reduzir o erro de truncamento causam crescimento do erro de arredondamento. Este é o primeiro tipo de métodos instáveis que encontramos. Estas técnicas devem ser evitadas sempre que possível. Entretanto as fórmulas aqui derivadas serão necessárias na aproximação de soluções de equações diferenciais ordinárias (EDO) e parciais (EDP).

Capítulo 7

Elementos de Integração Numérica

Muitas vezes temos a necessidade de calcular a integral definida de uma função que não possui uma antiderivada explícita ou cuja antiderivada é muito difícil de ser obtida. O método básico envolvido na aproximação de uma integral é chamado de quadratura numérica e usa uma soma do tipo

$$\int_a^b f(x)dx = \sum_{i=0}^{n} a_i f(x_i) + E$$

onde E é o erro. Devemos observar que uma expressão analítica com forma fechada para integrais possui muitas vantagens sobre a integração numérica como, por exemplo, a precisão. Assim, antes de empregarmos métodos numéricos, primeiramente devemos fazer um sério esforço para proceder o cálculo analítico, incluindo para isto uma boa pesquisa em tabelas de integração. Agora, muitas vezes integração numérica é indispensável. Por exemplo, quando possuimos o valor da função apenas em pontos discretos. Tais funções muitas vezes resultam de soluções numéricas de equações diferenciais ou de dados experimentais tomados em pontos discretos. A integração numérica, em contraste com a derivação numérica é um processo estável e existem muitas fórmulas adequadas para esta finalidade.

Os métodos de quadratura são baseados em interpolação polinomial. Para tanto, consideramos uma função integrável sobre um intervalo $a \leq x \leq b$, $f(x)$. Dividimos este intervalo em n subintervalos iguais de largura $\triangle x$, onde

$$\triangle x = \frac{b-a}{n}$$

como mostrado na figura (7.1).
cada um destes subintervalos é chamado de painel (panel).

7.1 Regra Trapeizodal

Vamos agora considerar dois destes paineis consecutivos.

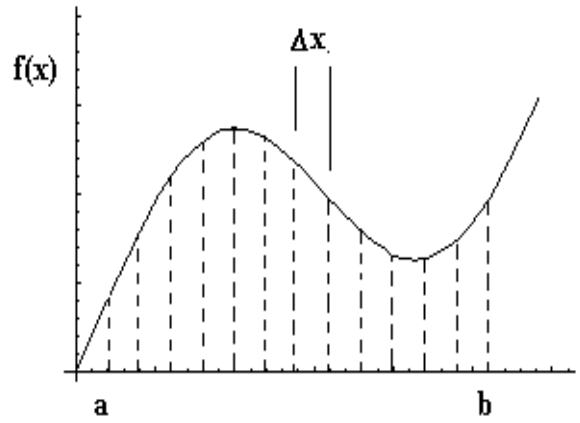


Figura 7.1: Subintervalos $\triangle x$

Nesta figura os pontos $f(x_{j-1})$, $f(x_j)$ e $f(x_{j+1})$ foram ligados por linhas retas. Estas linhas retas aproximam a função $f(x)$ e assim servem como polinômios interpoladores para $f(x)$ entre x_{j-1} e x_j, x_j e x_{j+1}, etc... Aproximando a área de cada painel sob as linha retas obtemos que

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx \frac{f_{j-1}+f_j}{2}(\triangle x) \tag{7.1}$$

e

$$\int_{x_j}^{x_{j+1}} f(x)dx \approx \frac{f_j+f_{j+1}}{2}(\triangle x) \tag{7.2}$$

Desta forma a integral sobre dois paineis é dada por:

$$\int_{x_{j-1}}^{x_{j+1}} f(x)dx = \int_{x_{j-1}}^{x_j} f(x)dx + \int_{x_j}^{x_{j+1}} f(x)dx$$

e usando (7.1) e (7.2) e assim,

$$\int_{x_{j-1}}^{x_{j+1}} f(x)dx \approx \frac{\triangle x}{2}\left(f_{j-1}+2f_j+f_{j+1}\right) \tag{7.3}$$

Como a área de cada painel foi aproximada por um trapésio, esta regra é chamada de regra do trapésio. Estendendo (7.3), a aproximação regra do trapésio sobre todo o intervalo de integração pode ser facilmente escrita como:

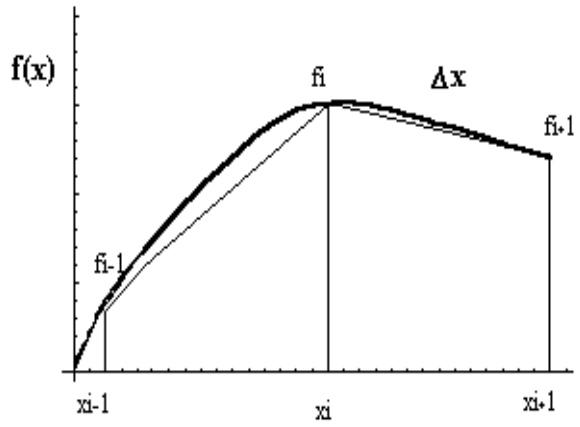


Figura 7.2: Dois paineis consecutivos

$$\int_a^b f(x)dx \approx \frac{\triangle x}{2}\left(f_0 + f_n + 2\sum_{j=1}^{n-1} f_j\right) \tag{7.4}$$

onde $f_0 = f(a)$ e $f_n = f(b)$.
Acima desenvolvemos a regra do trapésio de forma geométrica. Apesar de simples, a fórmula isto não nos fornece informações a respeito do erro cometido com esta aproximação nem sobre como incrementar sua precisão. Para obtermos informações quantitativas sobre o erro cometido vamos rederivar a fórmula (7.4) com uma base matemática mais sólida. Assim, para obtermos informações quantitativas para o erro, que mais tarde, servirá para comparação com outros métodos, vamos desenvolver a fórmula acima com auxílio de séries de Taylor.
Começamos definindo a integral indefinida

$$I(x) = \int_a^x f(x)dx \tag{7.5}$$

Se x_j está locado como mostra a figura (7.2), então assumindo que $I(x)$ é uma função analítica na região de interesse, podemos expandi-la em torno de x_j como:

$$I(x_{j+1}) = I(x_j + \triangle x) = I(x_j) + (\triangle x)I'(x_j) + \frac{(\triangle x)^2}{2}I''(x_j) + \frac{(\triangle x)^3}{3!}I'''(x_j) + O(\triangle x)^4 \tag{7.6}$$

mas, por (7.5) temos que $I'(x_j) = f(x_j)$, $I''(x_j) = f'(x_j)$, etc... Assim podemos reescrever (7.6) em termo de $f(x)$ e suas derivadas como,

$$I(x_{j+1}) = I(x_j + \triangle x) = I(x_j) + (\triangle x)f(x_j) + \frac{(\triangle x)^2}{2}f'(x_j) + \frac{(\triangle x)^3}{3!}f''(x_j) + O(\triangle x)^4 \tag{7.7}$$

Usando a fórmula de diferenças para frente na aproximação de $f'(x)$ temos,

$$I(x_{j+1}) = I(x_j)+(\triangle x)f(x_j)+\frac{(\triangle x)^2}{2}\left[\frac{f(x_{j+1})-f(x_j)}{\triangle x}-\frac{\triangle x}{2}f''(x_j)+O(\triangle x)^2\right]f'(x_j)+\frac{(\triangle x)^3}{3!}f''(x_j)+C \tag{7.8}$$

ou reagrupando os termos,

$$I(x_{j+1}) = I(x_j)+\frac{\triangle x}{2}\left[f(x_{i+1})+f(x_j)\right]-\frac{(\triangle x)^3}{12}f''(x_j)+)(\triangle x)^4 \tag{7.9}$$

Podemos ver que $I(x_{j+1}) - I(x_j)$ constitui a área de painel entre x_j e x_{j+1}. De (7.9) a área deste painel pode ser encontrada como:

$$S_{j+1} = I(x_{j+1}) - I(x_j) = \frac{\triangle x}{2}\left[f(x_{i+1})+f(x_j)\right]-\frac{(\triangle x)^3}{12}f''(x_j)+)(\triangle x)^4 \tag{7.10}$$

O termo $\frac{\triangle x}{2}\left[f(x_{i+1})+f(x_j)\right]$ é a aproximação trapeisodal para um painel simples. Os termos restantes então representam o erro. Para avaliar a integral sobre todo o intervalo, as contribuições de cada painel devem ser adicionadas. Assim

$$I = \sum_{j=1}^{n} S_j$$

ou

$$I = \frac{\triangle x}{2}\left[f(a)+f(b)+2\sum_{j=1}^{n-1}f(x_j)\right]-\frac{(\triangle x)^3}{12}\sum_{j=0}^{n-1}f''(x_j)+\text{termos de ordem mais alta} \tag{7.11}$$

O termo dominante de erro pode ser reescrito de forma fácil de entender. Primeiramente aplicamos o teorema do valor médio para a soma,

$$\sum_{j=0}^{n-1}f''(x_j) = nf''(\bar{x}), \quad \text{onde } a \le \bar{x} \le b \tag{7.12}$$

Agora,

$$nf''(\bar{x}) = \frac{b-a}{\triangle x}f''(\bar{x}) \tag{7.13}$$

assim o termo dominante de erro é escrito como

$$-\frac{(\triangle x)^3}{12}\frac{(b-a)}{\triangle x}f''(\bar{x}) = -\frac{(\triangle x)^2}{12}(b-a)f''(\bar{x}) \tag{7.14}$$

Usando (7.14), a regra do trapésio (7.11) fica

$$I = \frac{\triangle x}{2}\left[f(a)+f(b)+2\sum_{j=1}^{n-1}f(x_j)\right] - \frac{(\triangle x)^2}{12}(b-a)f''(\bar{x}) + \text{termos de ordem mais alta} \tag{7.15}$$

ou

$$I = \frac{\triangle x}{2}\left[f_0+f_n+2\sum_{j=1}^{n-1}f_j\right] + O(\triangle x)^2 \tag{7.16}$$

A regra do trapézio possui um erro de segunda ordem na integração numérica. Para funções bem comportadas é possível melhorar a aproximação estimando o termo de erro em (7.15). Usando diferença simples podemos estimar como,

$$f''(\bar{x}) \approx \frac{f'(b)-f'(a)}{b-a} \tag{7.17}$$

se esta estimativa é usada então (7.15) pode ser escrita como:

$$I = \frac{\triangle x}{2}\left[f(a)+f(b)+2\sum_{j=1}^{n-1}f(x_j)\right] - \frac{(\triangle x)^2}{12}(f'(b)-f'(a)) \tag{7.18}$$

A equação (7.18) é chamada de regra trapezoidal com correção. Com este termo de correção o método se transforma em um método de quarta ordem.

7.2 Regra de Simpson

A regra de Simpson é uma técnica de integração numérica baseada no uso de arcos parabólicos para aproximar $f(x)$ ao invés de linhas retas. Aqui vamos desenvolver esta fórmula empregando diretamente a fórmula de Taylor, pois assim teremos também uma estimativa do erro.

Vamos considerar a função $I(x) = \int_a^x f(x)dx$ e vamos assumir que a função $I(x)$ é analítica na região de interesse. Assim podemos expandir $I(x)$ em série de Taylor em torno de x_j como:

$$\begin{aligned} I(x_{j+1}) = I(x_j+\triangle x) =& I(x_j) + (\triangle x)I'(x_j) + \frac{(\triangle x)^2}{2}I''(x_j) + \frac{(\triangle x)^3}{3!}I'''(x_j) + \\ &+\frac{(\triangle x)^4}{4!}I^{iv}(x_j) + \frac{(\triangle x)^5}{5!}I^{(v)}(x_j) + \frac{(\triangle x)^6}{6!}I^{(vi)}(x_j) + O(\triangle x)^7 \end{aligned} \tag{7.19}$$

agora, notando que pela definição de $I(x)$ podemos escrever que $I'(x) = f(x)$, $I''(x) = f'(x)$, e assim por diante e substituindo estas relações em (7.19), temos que,

$$\begin{aligned} I(x_j + \triangle x) =& I(x_{j+1}) = I(x_j) + (\triangle x) f(x_j) + \frac{(\triangle x)^2}{2} f'(x_j) + \frac{(\triangle x)^3}{3!} f''(x_j) + \\ &+ \frac{(\triangle x)^4}{4!} f'''(x_j) + \frac{(\triangle x)^5}{5!} f^{(iv)}(x_j) + \frac{(\triangle x)^6}{6!} f^{(v)}(x_j) + O(\triangle x)^7 \end{aligned} \tag{7.20}$$

e

$$\begin{aligned} I(x_j - \triangle x) =& I(x_{j-1}) = I(x_j) - (\triangle x) f(x_j) + \frac{(\triangle x)^2}{2} f'(x_j) - \frac{(\triangle x)^3}{3!} f''(x_j) + \\ &+ \frac{(\triangle x)^4}{4!} f'''(x_j) - \frac{(\triangle x)^5}{5!} f^{(iv)}(x_j) + \frac{(\triangle x)^6}{6!} f^{(v)}(x_j) + O(\triangle x)^7 \end{aligned} \tag{7.21}$$

Subtraindo (7.21) de (7.20) temos

$$I(x_{j+1}) - I(x_{j-1}) = 2(\triangle x) f(x_j) + \frac{(\triangle x)^3}{3} f''(x_j) + \frac{(\triangle x)^5}{60} f^{(iv)}(x_j) + O(\triangle x)^7 \tag{7.22}$$

Agora, vamos substituir $f''(x_j)$ por sua representação em diferença central, incluindo o termo de erro, isto é

$$f''(x) = \frac{f_{j-1} - 2f_j + f_{j+1}}{h^2} - \frac{(\triangle x)^2}{12} f^{iv}(x_j) + O(\triangle x)^7 \tag{7.23}$$

e procedendo certos algebrismos obtemos,

$$I(x_{j+1}) - I(x_{j-1}) = \frac{(\triangle x)^3}{3} \left(f(x_{j-1}) + 4f(x_j) + f(x_{j+1}) \right) - \frac{(\triangle x)^5}{90} f^{(iv)}(x_j) + O(\triangle x)^7 \tag{7.24}$$

Agora, observamos que $I(x_{j+1}) - I(x_{j-1})$ constitui a área de dois painéis consecutivos entre x_{j-1} e x_{j+1}. A equação (7.24) e chamada de **regra de Simpson** para dois painéis.
para obtermos a integral sobre todo o domínio $a \leq x \leq b$, é preciso adicionar o resultado (7.24) para todos os pares de painéis. Assim, se $D_j = I(x_{j+1}) - I(x_{j-1})$ então

$$I = \sum_{\substack{j=1 \\ j\ impar}}^{n-1} D_j = D_1 + D_3 + ... + D_{n-3} + D_{n-1} \tag{7.25}$$

Note que este procedimento requer que o número de painéis seja **par**. Somando (7.24) para todos os pares de painéis teremos

$$I = \frac{\triangle x}{3}\left(f_0 + f_n + 4 \sum_{\substack{j=1 \\ j\ impar}}^{n-1} f_j + 2 \sum_{\substack{j=2 \\ j\ par}}^{n-2} f_j \right) - \frac{(\triangle x)^5}{90} \sum_{\substack{j=1 \\ j\ impar}}^{n-1} f^{(iv)}(x_j) + \frac{n}{2} O(\triangle x)^7 \tag{7.26}$$

O termo dominante de erro em (7.26) pode ser tratado da mesma forma que na regra do trapésio. Assim, pelo Teorema do valor médio temos,

$$\begin{aligned} \sum_{\substack{j=1 \\ j\ impar}}^{n-1} f^{(iv)}(x_j) =& n f^{iv}(\bar{x}),\ \text{com } a \leq \bar{x} \leq b \\ =& \frac{b-a}{\triangle x} f^{iv}(\bar{x}) \end{aligned}$$

assim temos que

$$-\frac{(\triangle x)^5}{90} \sum_{\substack{j=1 \\ j\ impar}}^{n-1} f^{(iv)}(x_j) = - - \frac{(\triangle x)^4}{180}(b-a) f^{(iv)}(\bar{x}).$$

Também devemos notar que

$$\frac{n}{2} O(\triangle x)^7 = \frac{b-a}{2(\triangle x)} O(\triangle x)^7 = O(\triangle x)^6$$

Agora, fazendo estas substituições em (7.26) temos,

$$I = \frac{\triangle x}{3}\left(f_0 + f_n + 4 \sum_{\substack{j=1 \\ j\ impar}}^{n-1} f_j + 2 \sum_{\substack{j=2 \\ j\ par}}^{n-2} f_j \right) - \frac{(\triangle x)^4}{180}(b-a) f^{(iv)}(\bar{x}) + O(\triangle x)^6 \tag{7.27}$$

A equação (7.27) é chamada de **regra de Simpson** para o intervalo completo. Ela é uma aproximação de quarta ordem. Lembramos que a interpretação geométrica deste método envolve

o uso de arcos parabólicos para aproximar $f(x)$, e que é de interesse lembrar que como todos os termos envolvendo $f'''(x)$ são cancelados em (7.27), esta técnica é exata para polinômios cúbicos.

Uma vez que o termo de erro em (7.27) envolve derivadas de ordens altas, não é prático a correção de erro através de da aproximação de seus termos. Ao invés disto, uma fórmula mais precisa envolvendo correção final pode ser encontrada se assumirmos que as derivadas de $f(x)$ são conhecidas nos pontos finais de cada painel duplo. Uma vez que $f(x)$ é conhecida em três pontos de cada painel duplo, isto é equivalente a aproximar $f(x)$por um polinômio interpolador de quarto grau sobre os dois painéis. Quando a integral é calculada pela da soma de todos os pares de painéis, as derivadas em todos os pontos interiores desaparecem e a aproximação integral fica:

$$I \approx \frac{\triangle x}{15}\left[14\left[\frac{1}{2}(f_0+f_n)+\sum_{\substack{j=2\\ j\ par}}^{n-2} f_j\right]+16\sum_{\substack{j=1\\ j\ impar}}^{n-1} f_j+\triangle x[f'(a)-f'(b)]\right] \tag{7.28}$$

A equação (7.28) é chamada de **regra de Simpson** com correção. é uma fórmula de sexta ordem e como na regra de Simpson exige que tenhamos um número **par** de painéis.

Podemos conseguir fórmulas de integração com menor ordem de erro se continuarmos substituindo as derivadas que aparecem na Série de Taylor por fórmulas em diferenças. Métodos com este procedimento são chamados de métodos de Newton-Cotes. Aqui nó estudamos os dois métodos de Newton-Cotes mais populares, isto é o Trapézio e Simpson. Os outros métodos de Newton cotes raramente são usados em micro computadores, pois suas características para o erro de arredondamento não são boas.

7.3 Integração de Romberg

O método de Romberg é baseado no fato do erro de truncamento da regra do trapézio ser proporcional a h^2. Assim, se dividimos h pela metade e reaplicamos a regra do trapézio temos uma redução do erro por um fator de $\frac{1}{4}$. Se comparamos os dois resultados chegamos a uma estimativa do erro restante. Esta estimativa pode então ser usada como uma correção. O método de *Romberg* é um refinamento sistemático desta idéia simples.
O método de integração de Romberg desta forma é um método potente baseado no uso da regra do trapézio combinado com *extrapolação de Richardson*. Para podermos aplicar a extrapolação precisamos conhecer a forma geral dos termos de erro para a regra do trapézio. Os detalhes desta derivação podem ser encontradas em livros de análise numérica. A regra do trapézio pode ser escrita como:

$$I = \frac{\triangle x}{2}\left(f(a) + f(b) + 2\sum_{j=1}^{n-1} f(a + j\triangle x)\right) + C(\triangle x)^2 + D(\triangle x)^4 + E(\triangle x)^6 + ... \quad (7.29)$$

onde C, D, E, etc. são funções de $f(x)$ e suas derivadas, mas não são funções de $\triangle x$. Os termos envolvendo as potências ímpares de $\triangle x$ desaparecem do erro.
Seja,

$$\bar{I} = \frac{\triangle x}{2}\left(f(a) + f(b) + \sum_{j=1}^{n-1} f(a + j\triangle x)\right) \quad (7.30)$$

A equação (7.29) pode ser reescrita na forma,

$$\bar{I} = I - C(\triangle x)^2 - D(\triangle x)^4 - E(\triangle x)^6 + ... \quad (7.31)$$

Vamos agora considerar que temos duas malhas de tamanhos diferentes $\triangle x_1$ e $\triangle x_2$, com $\triangle x_1 = 2\triangle x_2$. Vamos chamar de $\bar{I}_1$ e $\bar{I}_2$ os valores de $\bar{I}$ correspondentes a estas malhas. Assim, de (7.31) temos:

$$\bar{I}_1 = I_1 - C(\triangle x_1)^2 - D(\triangle x_1)^4 - E(\triangle x_1)^6 + ... \quad (32a)$$
$$\bar{I}_2 = I_2 - C(\triangle x_2)^2 - D(\triangle x_2)^4 - E(\triangle x_2)^6 + ... \quad (32b)$$

Vamos agora fazer a substituição de $\triangle x_1 = 2\triangle x_2$ na equação (32a) assim temos

$$\bar{I}_1 = I_1 - 4C(\triangle x_2)^2 - 16D(\triangle x_2)^4 - 64E(\triangle x_2)^6 + ... \quad (33)$$

Agora, multiplicando a equação (32b) por 4, subtraindo (33) e dividindo por 3:

$$\frac{4\bar{I}_2 - \bar{I}_1}{3} = I + 4D(\triangle x_2)^4 + 20E(\triangle x_2)^6 + ... \quad (34)$$

O termo de erro envolvendo $(\triangle x_2)^2$ desapareceu e (34) nos fornece, assim, uma aproximação para a integral com $O(\triangle x_2)^2$. Extrapolação deste tipo é chamada de *extrapolação de Richardson*. Agora, se calculássemos $\bar{I}_3$, onde $\triangle x_2 = 2\triangle x_3$ e se extrapolássemos $\bar{I}_2$ e $\bar{I}_3$, teríamos,

$$\frac{4\bar{I}_3 - \bar{I}_2}{3} = I + 4D(\triangle x_3)^4 + 20E(\triangle x_3)^6 + ... \quad (35)$$

Agora entre (34) e (35), o termo de erro em $(\triangle x)^4$ pode ser eliminado e desta forma fornecer uma estimativa para I de ordem $O(\triangle x)^6$. Assim a cada novo cálculo de um $\bar{I}$, um termo de erro

pode ser eliminado por extrapolação. Este procedimento sistemático é chamado de *integração de Romberg*.

Para descrevermos o algoritmo em detalhes vamos adotar outra notação. As estimativas iniciais feitas pela regra do trapézio da integral serão chamadas de $T_{1,k}$ com $k = 1, 2, 3...$ que representa a fórmula de Gauss para $n = 2^{k-1}$. Em geral podemos definir

$$T_{1,k} = \frac{\triangle x}{2}\left(f(a) + f(b) + 2\sum_{j=1}^{l} f(a + j\triangle x)\right) \tag{36}$$

onde $\triangle x = (b-a)/2^{k-1}$ e $l = 2^{k-1} - 1$. O número de painéis envolvidos em $T_{1,k}$ é 2^{k-1}. Assim, por exemplo

$$\begin{aligned}
T_{1,1} =& \frac{b-a}{2}\left(f(a) + f(b)\right) \text{ trapésio com } n = 2^0 = 1 \\
T_{1,2} =& \frac{b-a}{4}\left(f(a) + f(b) + 2f\left(a + \frac{b-a}{2}\right)\right) \text{ trapésio com } n = 2^1 = 2 \\
=& \frac{T_{1,1}}{2} + \frac{b-a}{2}\left(f\left(a + \frac{b-a}{2}\right)\right) \text{ trapésio com } n = 2^1 = 2 \\
T_{1,3} =& \frac{b-a}{8}\left(f(a) + f(b) + 2f\left(a + \frac{b-a}{4}\right) + 2f\left(a + \frac{2(b-a)}{4}\right) + 2f\left(a + \frac{3(b-a)}{4}\right)\right) \\
=& \frac{T_{1,2}}{2} + \frac{b-a}{4}\left(f\left(a + \frac{b-a}{4}\right) + f\left(a + \frac{3(b-a)}{4}\right)\right) \text{ trapésio com } n = 2^2 = 4 \\
&\text{etc...}
\end{aligned}$$

Assim vemos que cada aproximação da regra do trapézio sucessiva pode ser obtida da aproximação anterior sem ter que recalcular $f(x)$ em nenhum ponto onde ela já tinha sido calculada. A extrapolação é feita de acordo com a fórmula:

$$T_{l,k} = \frac{1}{4^{l-1} - 1}(4^{l-1}T_{l-1,k+1} - T_{l-1,k}) \tag{37}$$

Estes resultados podem ser covenientemente arranjados em uma tabela como:
Os valores extrapolados ao longo da diagonal irá convergir para a resposta correta muito mais rapidamente que a regra do trapézio cujos valores estão na primeira coluna.

Exemplo 48 *Como exemplo vamos considerar a integral*

$$I = \int_0^8 \left(\frac{5x^4}{8} - 4x^3 + 2x + 1\right) dx$$

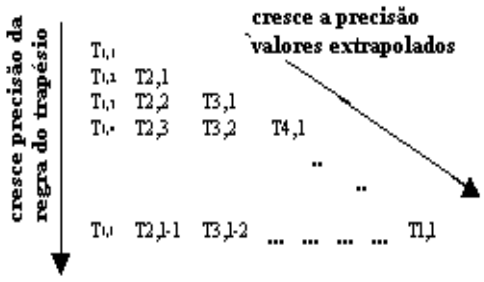


Figura 7.3: Tabela da Integração de Romberg

Esta integral pode ser calculada analiticamente e seu valor é $I = 72$. *Vamos agora montar a tabela de integração de Romberg, para encontrarmos a primeira coluna usamos a fórmula (36), assim:*

$$
\begin{aligned}
f(x) =& \frac{5x^4}{8} - 4x^3 + 2x + 1 \\
& \textit{logo} \\
f(0) =& 1 \\
f(8) =& 2560 - 2048 + 16 + 1 = 529 \\
b - a =& 8 - 0 = 8
\end{aligned}
$$

As aproximações da regra do trapézio com um e dois painéis são dadas por:

$$T_{1,1} = \frac{b-a}{2}(f(a)+f(b)) = \frac{8}{2}(1+529) = 2120$$

$$T_{1,2} = \frac{T_{1,1}}{2} + \frac{b-a}{2}\left(f\left(a+\frac{b-a}{2}\right)\right) = \frac{2120}{2} + \frac{8}{2}f(4) = 1060 + 4(160-256+8+1) = 712$$

Extrapolando estes dois valores para eliminar $O(\triangle x)^2$

$$T_{2,1} = \frac{1}{4^1-1}(4T_{1,2} - T1,1) = \frac{1}{3}(4(712)-2120) = 242\frac{2}{3}$$

A aproximação da regra do trapézio com 4 painéis

$$T_{1,3} = \frac{T_{1,2}}{2} + \frac{b-a}{4}\left(f\left(a+\frac{b-a}{4}\right) + f\left(a+\frac{3(b-a)}{4}\right)\right) = \frac{712}{2} + \frac{8}{4}(f(2)+f(6)) = 240$$

Extrapolando $T_{1,2}$ *e* $T_{1,3}$ *temos*

$$T_{2,2} = \frac{1}{3}(4(240)-712) = 82\frac{2}{3}$$

Extrapolando $T_{2,1}$ *e* $T_{2,2}$ *para eliminar* $O(\triangle x)^4$

$$T_{3,1} = \frac{1}{4^2-1}(4^2T_{2,2} - T_{2,1}) = \frac{1}{15}(16(82\frac{2}{3}) - 242\frac{2}{3}) = 72$$

E este é o valor exato. A tabela de Romberg obtida é:

$$\begin{array}{llll} 2120 & & & \\ 712 & 242\frac{2}{3} & & \\ 240 & 82\frac{2}{3} & 72 & \end{array}$$

A melhor aproximação da regra do trapézio disponível é 240 *usando 4 painéis que é ainda muito distante da resposta correta, e a grande aceleração da convergência ao longo da diagonal é aparente. É claro que na prática não conhecemos a resposta exata não é conhecida, assim uma outra linha seria calculada na tabela. Depois deste cálculo, a tabela ficaria,*

$$\begin{array}{llll} 2120 & & & \\ 712 & 242\frac{2}{3} & & \\ 240 & 82\frac{2}{3} & 72 & \\ 128\frac{1}{2} & 72\frac{2}{3} & 72 & 72 \end{array}$$

O critério de parada no procedimento de integração de Romberg pode claramente ser baseada na comparação entre valores sucessivos ao longo da diagonal da tabela. Uma vez que dois valores sucessivos concordaram exatamente na tabela acima, não existe dúvida que o método convergiu para $I = 72$. Em geral o critério de convergência pode ser da forma

$$|T_{l,1} - T_{l-1,1}| < \epsilon \tag{38}$$

Se (38) é satisfeito para algum ϵ, então o procedimento é parado e $T_{l,1}$ considerado como a resposta certa. Para uso geral, um critério de parada mais seguro é o critério de convergência absoluta,

$$\frac{|T_{l,1} - T_{l-1,1}|}{T_{l,1}} < \epsilon. \tag{39}$$

Exemplo 49 *Vamos agora aproximar a integral* $\int_0^\pi \sin(x)dx$, *com precisão de* 10^{-7}. *A montagem dos dois primeiros elementos da primeira coluna é dada pela fórmula (36)*

$$T_{1,1} = \frac{\pi}{2}(\sin(0) + \sin(\pi)) = 0$$
$$T_{1,2} = \frac{1}{2}(T_{1,1} + \pi \sin\frac{\pi}{2}) = 1.57079633$$

Assim nossa tabela é dada por: Comparando os valores obtidos na diagonal temos que $|2.09439511 - 0| > 10^{-7}$, *assim devemos continuar a tabela calculando pela fórmula (36) o valor da primeira coluna* $T_{1,3}$ *e a seguir usando a fórmula (37) para* $l = 3$ *e* $k = 1, 2$, *assim comparando os*

k	$T_{1,k}$	$T_{2,k} = \frac{1}{3}(4T_{1,k+1} - T_{1,k})$
1	0	
2	1.5707963	2.09439511

k	$T_{1,k}$	$T_{2,k} = \frac{1}{3}(4T_{1,k+1} - T_{1,k})$	$T_{3,k} = \frac{1}{15}(16T_{1,k+1} - T_{1,k})$
1	0		
2	1.5707963	2.09439511	
3	1.89611890	2.00455976	1.99857073

elementos da diagonal principal, $|1.99857070 - 2.09439511| > 10^{-7}$ *e assim devemos continuar o processo. Agora podemos parar o processo, pois* $|2.00000000 - 1.99999999| < 10^{-7}$. *Logo a*

k	$T_{1,k}$	$T_{2,k}$	$T_{3,k}$	$T_{4,k}$	$T_{5,k}$	$T_{6,k}$
1	0					
2	1.57079630	2.09439511				
3	1.89611890	2.00455976	1.99857073			
4	1.97423160	2.00269170	1.99998313	2.00000555		
5	1.99357035	2.00001659	1.99999975	2.00000001	1.99999999	
6	1.99839330	2.00000103	2.00000000	2.00000000	2.00000000	2.00000000

aproximação dada por Romberg para esta integral é 2.00000000.

7.4 Integração Adaptativa

A integração adaptativa cobre os muitos métodos que tem sido imaginados para lidar com o fato de que a maior parte das funções são mais difíceis de integrar precisamente sobre certos intervalos do que em outros. Uma seção particularmente difícil pode, por exemplo, forçar o uso de um h com valor muito pequeno para a regra de Simpson e lidar com um grande número de computação desnecessária. Métodos adaptativos usam malhas finas somente onde elas são realmente necessárias

7.5 Quadratura de Gauss

A principal idéia atrás da integração Gaussiana é escolher os nós da malha, x_i, de forma a minimizar o erro cometido na aproximação da integral pelo somatório característico das quadraturas. Todas as fórmulas que vimos nas seções anteriores assumiram espaçamento igual nos nós da malha, e este tipo de integração é principalmente usada quando os valores de $f(x_i)$ são obtidos experimentalmente. Agora, muitas integrais envolvem funções analíticas familiares as quais podem ser calculadas em qualquer argumento, apesar de possuir integral difícil de ser obtida analiticamente. Em tais casos é útil fazer uma escolha para x_i e w_i de forma a trazer o máximo de precisão na aproximação da integral. Para isto é conveniente discutir uma fórmula de quadratura mais geral como,

$$\int_a^b w(x)f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i) \tag{40}$$

na qual $w(x)$ é uma função peso que será especificada mais tarde. Note que se $w(x) = 1$ temos a fórmula de quadratura usual.

Devemos procurar que elas sejam exatas se $f(x)$ é uma das funções potências $1, x, x^2, ..., x^{2n-1}$. Isto nos fornece $2n$ condições para determinar os $2n$ números x_i e w_i necessários para a fórmula (40). De fato,

$$w_i = \int_a^b w(x)L_i(x)dx$$

onde $L_i(x)$ é a função multiplicador de Lagrange discutida anteriormente. Os argumentos $x_1, x-2, ..., x_n$ são zeros de um polinômio de grau n, $p_n(x)$, pertencentes a uma família de polinômios ortogonais, isto é possuem a propriedade de ortogonalidade

$$\int_a^b w(x)p_n(x)p_m(x)dx = 0, \qquad \text{se } m \neq n$$

Devemos notar que estes polinômios dependem da função peso $w(x)$, assim a função peso influencia ambos os valores w_i e x_i, apesar de não aparecem explicitamente na fórmula Gaussiana.

O erro de truncamento da fórmula Gaussiana é dado por,

$$\int_a^b w(x)f(x)dx - \sum_{i=1}^{n} w_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b w(x)[\pi(x)]^2 dx$$

onde $\pi(x) = (x - x_1)...(x - x_n)$. Como a fórmula de erro nos mostra, o erro cometido na aproximação é proporcional a $f^{(2n)}(\xi)$, assim esta fórmula será exata para polinômios de grau $2n - 1$ (pois sua derivada de ordem $2n$ é nula).

7.5.1 Tipos Particulares de Fórmulas Gaussianas:

Aqui vamos mostrar quatro tipos de quadraturas Gaussianas muito usadas. Todas elas são exatas para o cálculo de integrais de um polinômio de grau $2N - 1$. São elas a Quadratura de Gauss Legendre usadas para aproximar $\int_{-1}^{1} f(x)dx$, este intervalo pode ser transladado para um intervalo $[a, b]$ qualquer. Gauss Laguerre usadas para aproximar $\int_0^\infty e^{-x} f(x)dx$, Gauss Hermite usada para aproximar $\int_{-\infty}^{\infty} e^{x^2} f(x)dx$ e Gauss Tchebychev usada para aproximar $\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx$.

1. **Fórmulas de Gauss Legendre**:

   Ocorrem quando $w(x) = 1$. Normalmente normalizam o intervalo (a, b) em $(-1, 1)$. Os polinômios de Legendre são os escolhidos.

   $$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n$$

   com $P_0(x) = 1$. Os x_i são os zeros destes polinômios e os coeficiêntes são

   $$w_i = \frac{2(1 - x_i^2)}{n^2(P_{n-1}(x_i))^2}$$

   Tabelas de x_i e w_i estão disponíveis para serem substituidos diretamente na fórmula de Gauss-Legendre

   $$\int_a^b f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

   Obs: pode-se fazer uma mudança de variáveis

   $$\int_a^b f(x)dx = \int_{-1}^{1} f\left(\frac{(b-a)t + (b+a)}{2}\right)\frac{(b-a)}{2}dt \approx \frac{b-a}{2}\sum_{i=1}^{n} w_i f\left(\frac{(b-a)t_i + b + a}{2}\right)$$

   Exemplo 50 *Neste exemplo vamos usar quadratura de Gauss-Legendre com $N = 4$ para aproximar o valor da seguinte integral:*

   $$\int_0^{\pi/2} x^2 \cos x dx = \frac{\pi^2 - 8}{4} \approx 0.467401$$

Usando a tabela de Gauss-Legendre temos que para os limites de integração -1 *e* 1 *as raízes e pesos de Gauss Legendre são:* $\xi_1 = 0.8611363116$, $w_1 = 0.3478548451$, $\xi_2 = 0.3399810436$, $w_2 = 0.6521451549$, $x_3 = -x_2$, $w_3 = w_2$, $x_4 = -x_1$ *e* $w_4 = w_1$

Fazendo mudança de variáveis $\xi \in [-1, 1] \rightarrow x \in [0, \frac{\pi}{2}]$ *na equação acima temos*

- $x_4 = \dfrac{\pi/2+0}{2} + \dfrac{\pi/2-0}{2}\xi_1 = \dfrac{1.570796}{2} + \dfrac{1.5770796}{2}(-0.861136) = 0.109064$
- $x_3 = \dfrac{\pi/2+0}{2} + \dfrac{\pi/2-0}{2}\xi_1 = \dfrac{1.570796}{2} + \dfrac{1.5770796}{2}(-0.347855) = 0.518378$
- $x_2 = \dfrac{\pi/2+0}{2} + \dfrac{\pi/2-0}{2}\xi_1 = \dfrac{1.570796}{2} + \dfrac{1.5770796}{2}(0.347855) = 1.03242$
- $x_1 = \dfrac{\pi/2+0}{2} + \dfrac{\pi/2-0}{2}\xi_1 = \dfrac{1.570796}{2} + \dfrac{1.5770796}{2}(0.861136) = 1.46173$
- $f(x_4) = (0.109064)^2 \cos(0.109064) = 0.011824$
- $f(x_3) = 0.233413$
- $f(x_2) = 0.548777$
- $f(x_1) = 0.232572$

$$\begin{aligned}
\bar{I} &= \tfrac{\pi/2-0}{2}\left(w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3) + w_4 f(x_4)\right) \\
&= (0.347855)(0.232572) + (0.652145)(0.548777) + (0.652145)(0.233413) + (0.347855)(0.0118 \\
&= 0.467402
\end{aligned}$$

Exato	Gauss-Legendre $N = 4$	Trapézio $N = 4$	Simpson $N = 4$
0.467401	0.467402	0.435811	0.466890

2. **Fórmulas de Gauss Laguerre**:

   Tomam a forma

$$\int_0^\infty e^{-x} f(x) dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

   Os argumentos x_i são os zeros do polinômio de Laguerre de grau n

$$L_n(x) = e^x \frac{d^n}{dx^n}(e^{-x} x^n)$$

e os coeficiêntes w_i são

$$w_i = \frac{(n!)^2}{x_i[L'_n(x_i)]^2}$$

Os números x_i e w_i estão disponíveis em tabelas.

3. **Fórmulas de Gauss Hermite**:

   Tomam a forma

   $$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

   Os argumentos x_i são os zeros do polinômio de Hermite de grau n

   $$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

   e os coeficiêntes w_i são

   $$w_i = \frac{2^{n+1} n! \sqrt{\pi}}{[H'_n(x_i)]^2}$$

   Os números x_i e w_i estão disponíveis em tabelas.

4. **Fórmulas de Chebychev**:

   Tomam a forma

   $$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n} \sum_{i=1}^{n} f(x_i)$$

   Os argumentos x_i são os zeros do polinômio de Chebychev $T_n(x) = \cos(n \arccos x)$

7.6 Tabelas

n	x_k	w_k	**n**	x_k	w_k
2	±.57735027	1.0000000			
			12	±.98156063	.04717534
4	±.86113631	.34785485		±.90411725	.10693933
	±.33998104	.65214515		±.76990267	.16007833
				±.58731795	.30216743
6	±.93246951	.17132449		±.36783150	.23349254
	±.66120939	.36076157		±.12533341	.24914705
	±.23861919	.46791393			
			14	±.98628381	.03511946
8	±.96028986	.10122854		±.92843488	.08015809
	±.79666648	.22238103		±.82720132	.12151857
	±.52553241	.31370665		±.68729290	.15720317
	±.18343464	.36268378		±.51524864	.18553840
				±.31911237	.20519846
10	±.97390653	.06667134		±.10805495	.21526385
	±.86506337	.14945135			
	±.67940957	.21908636	5	±0.90617985	0.23692689
	±.43339539	.26926672		±0.53846931	0.47862867
	±.14887434	.29552422		0.00000000	0.56888889
3	±0.77459667	0.55555556			
	0.00000000	0.88888889			

Tabela 7.1: Tabela pesos e raízes Gauss-Legendre

n	x_k	w_k	**n**	x_k	w_k
2	±0.70710678	0.88622693	12	±0.31424038	0.57013524
				±0.94778839	0.26049231
4	±0.52464762	0.80491409		±1.59768264	0.05160799
	±1.65068012	0.08131284		±2.27950708	0.00390539
				±3.02063703	0.00008574
6	±0.43607741	0.72462960		±3.88972490	0.00000027
	±1.33584907	0.15706732			
	±2.35060497	0.00453001	14	±0.29174551	0.53640591
				±0.87871379	0.27310561
8	±0.38118699	0.66114701		±1.47668273	0.06850553
	±1.15719371	0.20780233		±2.09518326	0.00785005
	±1.98165676	0.01707798		±2.74847072	0.00035509
	±2.93063742	0.00019960		±3.46265693	0.00000472
				±4.30444857	0.00000001
10	±0.34290133	0.61086263			
	±1.03661083	0.24013861			
	±1.75668365	0.03387439			
	±2.53273167	0.00134365			
	±3.43615912	0.00000764			

Tabela 7.2: Tabela pesos e raízes Gauss-Hermite

n	x_k	w_k	**n**	x_k	w_k
2	0.58578644	0.85355339	12	0.11572212	0.26473137
	3.41421356	0.14644661		0.611775748	0.37775928
				1.51261027	0.24408201
4	0.32254769	0.60315410		2.83375134	0.09044922
	1.74576110	0.35741869		4.59922764	0.02012238
	4.53662030	0.03888791		6.84452545	0.00266397
	9.39507091	0.00053929		9.62131684	0.00020323
				13.00605499	0.00000837
6	0.22284660	0.45896467		17.11685519	0.00000017
	1.18893210	0.41700083		22.15109038	0.00000000
	2.99273633	0.11337338		28.48796725	0.00000000
	5.77514357	0.01039920		37.09912104	0.00000000
	9.83746742	0.00026102			
	15.98287398	0.00000090	14	0.09974751	0.23181558
				0.52685765	0.35378469
8	0.17027963	0.36918859		1.30062912	0.25873461
	0.90370178	0.41878678		2.43080108	0.11548289
	2.25108663	0.17579499		3.93210282	0.03319209
	4.26670017	0.03334349		5.82553622	0.00619287
	7.04590540	0.00279454		8.14024014	0.00073989
	10.75851601	0.00009077		10.91649951	0.00073989
	15.74067864	0.00000085		14.21080501	0.00000241
	22.86313174	0.00000000		18.10489222	0.00000006
				22.72338163	0.00000000
10	0.13779347	0.30844112		28.27298172	0.00000000
	0.72945455	0.40111993		35.14944366	0.00000000
	1.80834290	0.21806829		44.36608171	0.00000000
	3.40143370	0.06208746			
	5.55249614	0.00950152			
	8.33015275	0.00075301			
	11.84378584	0.00002826			
	16.27925783	0.00000042			
	21.99658581	0.00000000			
	29.92069701	0.00000000			

Tabela 7.3: Tabela pesos e raízes Gauss-Laguerre

As raízes dos polinômios de Gauss Chebychev, não necessitam de tabelas, pois são fáceis de serem encontradas. Os polinômios de Chebychev são definidos como:

$$T_n(x) = \cos\left(n \cos^{-1} x\right) \tag{41}$$

onde $T_n(x)$ é o polinômio de Chebychev de grau n. Abaixo veremos que esta definição também pode ser escrita em forma polinomial, pois:

- $n = 0 \rightarrow T_0(x) = \cos(0 \cos^{-1} x) = \cos 0 = 1$

- $T_1(x) = \cos(1\cos^{-1} x) = x$
- $T_2(x) = \cos(2\cos^{-1}(x)) = \cos y \;\; assim \;\; y = 2\cos^{-1} x$
  Desta forma:
  $\cos^{-1} x = \dfrac{y}{2}$ ou $x = \cos\left(\dfrac{y}{2}\right)$

  Da trigonometria temos:

  $\cos^2\left(\dfrac{y}{2}\right) - \sin^2\left(\dfrac{y}{2}\right) = \cos y$ e $\cos^2\left(\dfrac{y}{2}\right) + \sin^2\left(\dfrac{y}{2}\right) = 1$
  logo, $2\cos^2\left(\dfrac{y}{2}\right) = \cos y + 1$

  Além disto, como $x = \cos\left(\dfrac{y}{2}\right)$, então

  $x^2 = \cos^2\left(\dfrac{y}{2}\right)$ e portanto $2\cos^2\left(\dfrac{y}{2}\right) = 2x^2$
  logo, $2x^2 = \cos y + 1 \rightarrow \cos y = 2x^2 - 1$, isto é $T_2(x) = 2x^2 - 1$.
- Usando o mesmo raciocínio acima temos,

$$\begin{aligned} T_0(x) &= 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \\ T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1 \end{aligned}$$

  etc...

- Devemos notar que $T_n(x) \leq 1 \;\; \forall x \in [-1, 1]$ pois é um cosseno (41).
- As raízes dos polinômios de Chebychev $T_n(x)$ (41), podem ser calculadas como:

$$
\begin{aligned}
T_n(x) = 0 &\iff \cos\left[n\cos^{-1}x\right] = 0 \\
&\iff n\cos^{-1}x = \cos^{-1}0 \\
&\iff n\cos^{-1}(x) = \frac{(2m-1)\pi}{2}, \quad m = 1, 2, 3... \\
&\iff \cos^{-1}(x) = \frac{(2m-1)\pi}{2n} \\
&\iff x = \cos\left[\frac{(2m-1)\pi}{2n}\right]
\end{aligned}
$$

Assim as raízes dos polinômios de Chebychev $T_n(x)$ são:

$$x_m = \cos\left[\frac{(2m-1)\pi}{2n}\right] \qquad m = 1, 2, \ldots, n$$

Observamos que estes x_m tendem a ser mais próximos nos extremos do intervalo que do centro do intervalo.

Exemplo 51 *Para $T_6(x)$ as raízes são:*

$$
\begin{aligned}
x_1 &= \cos\left[\tfrac{(2(0)-1)\pi}{2(6)}\right] = 0.96592583 \\
x_2 &= \cos\left[\tfrac{(2(1)-1)\pi}{2(6)}\right] = 0.70710678 \\
x_3 &= 0.25881905 \\
x_4 &= -0.25881905 \\
x_5 &= -0.70710678 \\
x_6 &= -0.96592583
\end{aligned}
\tag{42}
$$

7.6.1 Exemplos

Exemplo 52 *Encontre o valor da integral:*

$$\int_0^\infty x^2 e^{-x^2} dx$$

Pela fórmula de Gauss Hermite, vemos que $f(x) = x^2$, um polinômio de grau 2, assim o valor esta integral será dado usando quadratura de Gauss Hermite escolhendo N o menor inteiro de forma que $2N - 1 \geq 2$, isto é $N \geq 3/2$. Ou seja para N maior ou igual a 2 teremos a integral calculada de forma exata por Gauss Hermite. Usando a tabele com $N = 2$ temos

$$x_1 = 0.70710678, \;\; w_1 = 0.88622693, \;\; x_2 = -0.70710678, \;\; w_2 = 0.88622693$$

Assim

$$\int_0^\infty x^2 e^{-x^2} dx = \frac{1}{2} \int_{-\infty}^\infty x^2 e^{-x^2} dx = \frac{1}{2} \sum_{i=1}^{2} f(x_i) w_i = \sum_{i=1}^{2} (x_i)^2 w_i = 0.443113$$

Exemplo 53 *Encontre o valor da integral:*

$$\int_0^6 x^5 - 7x + 2dx$$

Esta é a integral da função $f(x) = x^5 - 7x + 2$, polinômio de Grau 5. Pela fórmula de Gauss Legendre vemos que esta integral pode ser calculada de forma exata usando $2N - 1 = 5$ ou $N = 6/2 = 3$ Neste caso temos $x_1 = 0.77459667$, $x_2 = -x1$, $x_3 = 0$, $w_1 = 0.55555556$, $w_2 = w_1$ e $w[3] = 0.88888889$. Precisamos fazer uma mudança do intervalo $[-1, 1]$ para $[0, 6]$. Assim:

$$X_i = \frac{6-0}{2} x_i + \frac{a+b}{2}$$

logo

$$X_1 = 5.32379, \quad X_2 = 0.67621, \quad X_3 = 3$$

e o valor da integral é calculada por:

$$\begin{aligned} \int_0^6 x^5 - 7x + 2dx \;\; &= \tfrac{6-0}{2} \textstyle\sum_{i=1}^{3} f(X_i) w_i \\ &= 3((4241.39)w_1 + (-2.59208)w_2 + (224)w_3) \\ &3(2356.33 - 1.44005 + 199.111) = 7662. \end{aligned}$$

7.6.2 Construindo Quadraturas

Também podemos gerar quadraturas de forma que "englobem"certos tipos de singularidades. Para exemplificar este procedimento, aqui, vamos construir uma quadratura que resolve a integral

$$\int_0^1 \ln x f(x) dx$$

Para encontrar os pesos e raízes da quadratura de ordem $N = 2$, vamos supor que esta integral pode ser escrita como

$$\int_0^1 \ln x f(x) dx = w_1 f(x_1) + w_2 f(x_2)$$

Para ser uma quadratura Gaussiana, sabemos que esta integral deve ser exata se $f(x)$ é um polinômio de grau $2N-1=2(2)-1=3$, logo devemos ter que:

$$\begin{array}{rcl}\int_0^1 \ln x dx & = & w_1+w_2 \\ \int_0^1 x\ln x dx & = & w_1x_1+w_2x_2 \\ \int_0^1 x^2\ln x dx & = & w_1x_1^2+w_2x_2^2 \\ \int_0^1 x^3\ln x dx & = & w_1x_1^3+w_2x_2^3\end{array}$$

ou, resolvendo as integrais do lado esquerdo das equações acima por partes,

$$\begin{array}{rcl}-1 & = & w_1+w_2 \\ -\frac{1}{4} & = & w_1x_1+w_2x_2 \\ -\frac{1}{9} & = & w_1x_1^2+w_2x_2^2 \\ -\frac{1}{16} & = & w_1x_1^3+w_2x_2^3\end{array}$$

Resolvendo o sistema de equações não lineares acima, obtemos:

$$\begin{cases} x_1 = 0.60227691 \\ x_2 = 0.11200881 \\ w_1 = -0.28146068 \\ w_2 = -0.71853932 \end{cases}$$

Para gerar uma quadratura com $N=4$, teríamos um sistema não linear com oito equações. As oito variáveis deste sistema podem ser encontradas como:

$$w_1 = -0.383464, \quad w_2 = -0.386875, \quad w_3 = -0.190435, \quad w_4 = -0.039225$$

e

$$x_1 = 0.041448, \quad x_2 = 0.245275, \quad x_3 = 0.556165, \quad x_4 = 0.848982$$

Exemplo 54 *Para testar esta quadratura, primeiramente vamos calcular a integral de um polinômio de grau 3, cujo valor exato é dado por:*

$$\int_0^1 (x^3-5x^2+x+7)\ln x dx = -\frac{973}{144} \approx -6.75694$$

Usando a nova quadratura temos: $f(x)=x^3-5x^2+x+7$, *assim*

$$\int_0^1 (x^3-5x^2+x+7)\ln x dx \approx (x_1^3-5x_1^2+x_+7)w_1+(x_2^3-5x_2^2+x_2+7)w_2 = -6.75694$$

. Assim o erro absoluto de nossa aproximação é nulo.

Agora vamos escolher uma função não polinomial:

Exemplo 55 *Seja a integral:*

$$\int_0^1 \frac{\ln x}{x+1} dx = \frac{\pi^2}{12} \approx -0.8224670334$$

Usando quadratura $N = 2$ *temos:*

$$f(x) = \frac{1}{x+1}$$

e

$$\int_0^1 \frac{\ln x}{x+1} dx \approx f(x_1)w_1 + f(x_2)w_2 = -0.821826$$

Neste caso temos o erro já na terceira casa decimal. Se usássemos $N = 4$*,*

$$f(x) = \frac{1}{x+1}$$

e

$$\int_0^1 \frac{\ln x}{x+1} dx \approx f(x_1)w_1 + f(x_2)w_2 + f(x_3)w_3 + f(x_4)w_4 = -0.822466$$

O que nos fornece 5 casas decimais de precisão.

Se a mesma integral fosse calculada por quadratura de Gauss Legendre, teríamos:

$$f(x) = \frac{\ln x}{x+1}$$

e

Tabela 7.4: Integral $\int_0^1 \frac{\ln x}{x+1} dx$ por QG Legendre

N	Integral
2	-0.707957
12	-0.818411
22	-0.821218
32	-0.821869
42	-0.822117
492	-0.822464

Isto é, usando quadratura $N = 492$ *para GLegendre temos apenas 2 casas decimais de precisão.*

Este fato deve-se a singularidade da função $f(x) = \ln(x)/(x+1)$ considerada na quadratura de Gauss Legendre em $x = 0$. Já na quadratura construída, a função $f(x) = 1/(x+1)$ é contínua em $[0, 1]$.

7.6.3 Integrais Multiplas

Podemos avaliar integrais multiplas por uma das fórmulas de quadraturas mostradas acima. Para isto podemos fazer:

$$\int_a^b \left(\int_{c(x)}^{d(x)} f(x,y)dy \right) dx$$

Primeiramente definimos a função

$$g(x) = \int_{c(x)}^{d(x)} f(x,y)dy$$

e aplicamos uma fórmula de quadratura, isto é:

$$\int_a^b \left(\int_{c(x)}^{d(x)} f(x,y)dy \right) dx = \int_a^b g(x)dx \approx \sum_{i=1}^{N} a_i g(x_i)$$

poir sua vez, para cada i sas funções $g(x_i)$ são integrais, logo são calculadas como:

$$g(x_i) = \int_{c(x_i)}^{d(x_i)} f(x_i,y)dy \approx \sum_{j=1}^{M} b_i f(x_i,y_j)$$

ou seja

$$\int_a^b \left(\int_{c(x)}^{d(x)} f(x,y)dy \right) dx \approx \sum_{i=1}^{N} a_i \sum_{j=1}^{M} b_i f(x_i,y_j)$$

Capítulo 8

Soluções Numéricas de Equações Diferencias Ordinárias

8.1 Introdução

Existe uma enorme variedade de aproximações numéricas para a solução de ED. Não linearidades em ED e condições de contorno ou condições iniciais, muitas vezes requerem modificações nestas técnicas numéricas a serem empregadas na solução. Muitas vezes técnicas numéricas podem ser escolhidas sem realmente olharmos para a equação diferencial, o que é muito contrastante com ao problema de encontrar uma solução analítica exata para uma ED. Até problemas lineares podem algumas vezes apresentar obstáculos enormes para encontrar uma técnica aceitável para encontrar solução analítica. Ale'm disto algumas EDO e muitas EDP são virtualmente impossíveis de serem resolvidas de forma analíticas. Muitas vezes é possível aproximar soluções para estes problemas, mas a precisão da aproximação raramente podem ser propriamente avaliadas.

Técnicas numéricas para EDO são muito poderosas e podem ser aplicados em uma grande variedade de problemas. Mas é preciso ter em mente que existem dificuldades inerentes a elas mesmas, como iremos ver mais adiante. Além disto, como técnicas numéricas são estas ferramentas tão poderosas e flexíveis, é natural que sejam aplicadas na resolução de equações extremamente complexas e muitas vezes mal comportadas. Em tais circunstâncias, é inteiramente razoável esperar que dificuldades inerentes ao problema, que está sendo resolvido, sejam manifestadas de uma forma indesejável na aplicação destes métodos técnicas numéricos. Desta forma técnicas numéricas para EDO nunca devem ser consideradas como uma prova infalível, e seus resultados não devem ser aceitos como corretos sem um cuidadoso exame para cada problema importante.

Vamos agora identificar duas grandes categorias nas quais os problemas envolvendo ED devem estar envolvidos. Estas duas categorias são chamadas de *problemas de valor inicial* e *problemas de valor de contorno*. Problemas de valor inicial são aqueles para os quais condições são especificadas apenas em *um* valor da variável independente. Estas condições são chamadas de *condições iniciais*. Por simplicidade, muitas vezes dizemos que este valor é o zero. Um problema típico de condição inicial é dado por:

$$A\frac{d^2y}{dt^2} + B\frac{dy}{dt} + Cy = g(t), \qquad y(0) = y_0, \qquad \frac{dy}{dt}(0) = V_0 \tag{1}$$

Este problema poderia descrever a resposta forçada de um oscilador harmônico simples com o tempo.

Problemas de condição de contorno são aqueles para os quais as condições são especificadas em dois valores distintos da variável independente. Um problema típico de valor de contorno poderia ter a forma,

$$\frac{d^2y}{dx^2} + D\frac{dy}{dx} + Ey = h(x), \qquad y(0) = y_0, \qquad y(L) = y_l \tag{2}$$

Este problema poderia descrever a distribuição de temperatura em estado estacionário em um problema unidimensional de transferência de calor com temperatura y_0 na extremidade $x = 0$ e y_l em $x = L$.

8.2 Problemas de Valor Inicial

Qualquer problema de valor inicial pode ser representado como um conjunto de uma ou mais EDO de primeira ordem, cada uma com uma condição inicial. Por exemplo, o problema do oscilador harmônico simples descrito por,

$$A\frac{d^2y}{dt^2} + B\frac{dy}{dt} + Cy = g(t), \qquad y(0) = y_0, \qquad \frac{dy}{dt}(0) = V_0 \tag{3}$$

pode ser reescrito com a mudança de variáveis $z = \frac{dy}{dt}$, como

$$\begin{aligned} \frac{dy}{dt} &= z \\ \frac{dz}{dt} &= -\frac{B}{A}z - \frac{C}{A}y + g(t) \end{aligned} \tag{4}$$

com as condições iniciais:

$$\begin{aligned} y(0) &= y_0 \\ z(0) &= V_0 \end{aligned} \tag{5}$$

Para qualquer EDO de ordem n pode similarmente ser reduzidos a um sistema de n EDO de primeira ordem acopladas. Assim neste capítulo vamos estudar:

- Métodos para encontrar a solução aproximada $y(t)$ para problemas da forma:

$$\begin{array}{l} \frac{dy}{dt} = f(t,y) \qquad a \leq t \leq b \\ y(a) = \alpha \end{array}$$

- Extensão destes métodos para resolução aproximada do sistema de EDO de primeira ordem,

$$\begin{array}{rcl} \frac{dy_1}{dt} & = & f_1(t, y_1, y_2, ..., y_n) \\ \frac{dy_2}{dt} & = & f_2(t, y_1, y_2, ..., y_n \\ \vdots & & \vdots \\ \frac{dy_n}{dt} & = & f_n(t, y_1, y_2, ..., y_n) \end{array} \tag{6}$$

para $a \leq t \leq b$ e sujeito às condições iniciais:

$$y_1(a) = \alpha_1, \quad y_2(a) = \alpha_2, \quad ..., y_3(a) = \alpha_3 \tag{7}$$

- Relação entre um problema de valor inicial de ordem n e um sistema de EDO de primeira ordem.

8.3 Teoria Elementar de Problemas de Valor Inicial

Definição 1: Uma função $f(t,y)$ é dita de Lipschitz na variável $y \in D \subset \Re^2$, se existe uma constante, chamada de constante de Lipschitz, $L > 0$ de forma que

$$|\, f(t, y_1) - f(t, y_1)| \leq L|\, y_1 - y_2|$$

sempre que $(t, y_1),\ (t, y_2) \in D$.

Definição 2: Um conjunto $D \subset \Re^2$ é dito comvexo se sempre que $(t, y_1),\ (t, y_2) \in D$ o ponto $((1-\lambda)t_1 + \lambda t_2\,,\ (1-\lambda)y_1 + \lambda y_2)$ pertença à D para cada λ, $0 \leq \lambda \leq 1$.

Teorema 40 *Suponha que $f(t,y)$ está definida num conjunto convexo $D \subset \Re^2$. Se existe uma constante $L > 0$ de forma que,*

$$\left|\frac{\partial f}{\partial y}(t,y)\right| \leq L \qquad \textit{para todo } (t,y) \in D, \tag{8}$$

então f satisfaz a condição de Lipschitz sobre D na variável y para a constante L. (Prova: Teorema do valor médio)

Como próximo teorema iremos mostrar, que muitas vezes é de grande interesse determinar quando a função envolvida em um problema de valor inicial satisfaz as condições de Lipschitz na sua segunda variável, e a condição (8) é geralmente muito mais simples de aplicar que a definição. Deve ser notado que o teorema (40) nos dá apenas condições suficientes para uma condição valer.

O próximo teorema é uma versão do teorema fundamental de existência e unicidade para EDO de primeira ordem.

Teorema 41 *Sejam $D = \{(t, y)/a \leq t \leq b, -\infty \leq y \leq \infty\}$, e $f(t, y)$ uma função contínua sobre D. Se f satisfaz a condição de Lipschitz em D na variável y, então o problema de valor inicial*

$$\frac{dy}{dt} = f(t, y) \qquad a \leq t \leq b$$
$$y(a) = \alpha$$

tem solução única $y(t)$ para $a \leq t \leq b$. (Prova em livros de EDO).

Exemplo 56 *Seja o problema de valor inicial:*

$$y' = 1 + t\sin(ty), \qquad 0 \leq t \leq 2, \qquad y(0) = 0.$$

Deixando t constante e aplicando o TVM na função $f(t, y) = 1 + t\sin(ty)$ temos que para $y_1 < y_2$ existe um número ξ, $y_1 < \xi < y_2$ de forma que

$$t^2\cos(\xi t) = \frac{\partial f}{\partial y}(t, \xi) = \frac{f(t, y_2) - f(t, y_1)}{y_2 - y_1}$$

daí, para todo $y_1 < y_2$,

$$|f(t, y_2) - f(t, y_1)| = |y_2 - y_1| \left|t^2\cos(\xi t)\right| \leq 4\,|y_2 - y_1|$$

assim, f satisfaz a condição de Lipschtz na variável y, com $L = 4$. Além disto $f(t, y)$ é contínua quando $0 \leq t \leq 2$ e $-\infty \leq y \leq \infty$, logo, pelo teorema (41) possui solução única.

Nota 4 *Para desenvolvermos métodos numéricos de resolução de problema de valor inicial, gostaríamos de saber quando um problema tem a propriedade de que pequenas perturbações nos dados produzam pequenas mudanças na solução!*

Definição 3: O problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \qquad a \leq t \leq b, \;\; y(a) = \alpha \tag{9}$$

é dito bem-posto (well-posed) se:

1. Possui solução única $y(t)$

2. $\forall \varepsilon > 0$, existe uma constante positiva k de forma que sempre que $|\varepsilon_0| < \varepsilon$ e $|\delta(t)| < \varepsilon$, uma única solução do problema

$$\frac{dz}{dt} = f(t,y) + \delta(t) \qquad a \leq t \leq, \qquad z(a) = \alpha + \varepsilon_0 \ , \tag{10}$$

existe com, $|z(t) - y(t)| < k\varepsilon$ para todo $a \leq t \leq b$.

O problema dado pela equação (10) é muitas vezes chamado de problema perturbado associado ao problema (9) Métodos numéricos sempre estarão relacionados com a resolução de um problema perturbado, uma vez que qualquer erro introduzido na representação irá resultar em um problema deste tipo. A não ser que o problema original seja bem posto, não existe a mínima razão para esperar que a solução numérica de um problema perturbado irá se aproximar da solução do problema original. (falar de convergência e estabilidade). O próximo teorema estabelecerá condições que asseguram que um problema de valor inicial é bem posto. Sua prova pode ser encontrada em livros de EDO.

Teorema 42 *Seja $D = \{(t,y)/a \leq t \leq b, -\infty \leq y \leq \infty\}$. Se f é contínua e satisfaz a condição de Lipschtz na variável y soble D, então o problema de valor inicial*

$$\frac{dy}{dt} = f(t,y) \qquad a \leq t \leq b, \qquad y(a) = \alpha,$$

é bem posto providenciando que a função δ é contínua.

Exemplo 57 *Sejam $D = \{(t,y)/0 \leq t \leq 1, -\infty \leq y \leq \infty\}$ e o problema de valor inicial*

$$\frac{dy}{dt} = -y + t + 1 \qquad 0 \leq t \leq 1, \qquad y(0) = 1, \tag{11}$$

como $f(t,y) = -y + t + 1$ temos que $\left|\frac{\partial f}{\delta y}\right| = 1$, logo pelo teorema 1 $f(t,y)$ satisfaz a condição de Lipschtz com $L = 1$. Como f é contínua sobre D temos que o problema é bem posto. Considere agora, o problema perturbado,

$$\frac{dz}{dt} = -z + t + 1 + \delta \qquad 0 \leq t \leq 1, \qquad z(0) = 1 + \varepsilon_0, \tag{12}$$

onde δ e ε_0 são constantes. As soluções dos problemas (11) e (12) são

$$y(t) = e^{-t} + t \qquad e \qquad z(t) = (1 + \varepsilon_0 - \delta)e^{-t} + t + \delta$$

respectivamente. Também é fácil verificar que se $|\delta| < \varepsilon$ e $|\varepsilon_0| < \varepsilon$, então

$$|y(t) - z(t)| = \left|(\delta - \varepsilon_0)e^{-t} - \delta\right| \leq |\varepsilon_0| + |\delta| \left|1 - e^{-t}\right| \leq 2\varepsilon$$

para todo t, o que concorda com o teorema acima.

8.4 Soluções Numéricas: Métodos de Passo Símples

Vamos considerar o problema de valor inicial:

$$\frac{dy}{dt} = f(t,y) \qquad a \leq t \leq b \tag{13}$$
$$y(a) = \alpha \tag{14}$$

Note que muitos destes problemas podem ser resolvidos de forma analítica, mas muitos outros não. Assim necessitamos desenvolver soluções numéricas para estes problemas.

8.4.1 O Método de Euler

Euler desenvolveu este método em 1768 usando aproximação em diferenças para frente (forward). Aqui vamos desenvolve-lo pela série de Taylor. O objetivo deste método é obter uma aproximação para o PVI bem posto

$$\frac{dy}{dt} = f(t,y), \qquad a \leq t \leq b, \quad y(a) = \alpha \tag{15}$$

Na verdade uma aproximação contínua da solução $y(t)$ não será obtida; Ao invés disto, serão geradas em vários valores, chamados de *pontos da malha*, sobre o intervalo$[a,b]$. Uma vez que a solução é obtida neste ponto, ela poderá ser obtida em outros pontos por interpolação.

Primeiramente vamos dividir o intervalo $[a,b]$ em um número finito de sub-intervalos pela introdução de uma malha (mesh ou grid points) $a = t_0 < t_1 < ... < t_N = b$. Vamos considerar que os pontos desta malha são igualmente afastados entre si, apesar que se não fossem não haveria dificuldade maior. Vamos chamar este espaçamento de $h = (b-a)/N$ de tamanho do passo assim

$$t_i = a + ih, \qquad para \;\; i = 0 : N \tag{16}$$

Vamos supor que $y(t)$ seja a única solução de (15) e que possua derivada de segunda ordem contínua em $[a,b]$. Agora, pelo teorema de Taylor, temos,

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_{i+1}) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i) \tag{17}$$

para algum número ξ_i, com $t_i < \xi_i < t_{i+1}$. Agora, como $h = t_{i+1} - t_i$ temos que

$$y(t_{i+1}) = y(t_i) + h\,y'(t_{i+1}) + \frac{h^2}{2}y''(\xi_i) \tag{18}$$

e como $y(t)$ satisfaz a EDO (15) temos que

$$y(t_{i+1}) = y(t_i) + h\, f(t, y(t_i)) + \frac{h^2}{2} y''(\xi_i) \tag{19}$$

O método de Euler despreza o termo de erro da expressão acima, construindo $w_i \simeq y(t_i)$ para cada $i = 11, 2, ..., N$, simplesmente retirando o termo de erro na equação (19). Assim,

$$\begin{array}{l} w_0 = \alpha \\ w_{i+1} = w_i \quad + hf(t_i, y_i) \qquad para \;\; i = 0 : N \end{array} \tag{20}$$

A equação (20) é chamada de equação em diferença associada ao método de Euler.

- ALGORÍTMO DE EULER:
  Dados: a, b, N, $w_0 = \alpha$
  Fazer: $t_0 = a,\; h = \frac{(b-a)}{N}$
  Para: $i = 1 : N$

$$w_i = w_{i-1} + hf(t_{i-1}, w_{i-1}))$$
$$t_i = t_{i-1} + h$$

  Saída: $i = 1 : N\;,\quad (w_i, t_i)$

Para interpretar o método de Euler geometricamente, note que quando w_i é uma aproximação boa de $y(t_i)$, a afirmativa que o problema é bem posto implica que

$$f(t_i, w_i) \approx y'(t_i) = f(t_i, y(t_i))$$

O gráfico da função e da aproximação feita pelo método de Euler está na figura (8.1), feita em vários passos.

Exemplo 58 *Vamos encontrar aproximações do PVI*

$$y' = -y + t + 1, \qquad 0 \le t \le 1, \quad y(0) = 1, \tag{21}$$

faça $N = 10$.
Assim, $h = 0.1$ *e* $t_i = 0.1i$.
Usando o fato que $f(t, y) = -y + t + 1$ *temos:*

$$\begin{aligned} w_0 &= 1 \\ w_i &= w_{i-1} + h(-w_{i-1} + t_{i-1} + 1) \\ &= w_{i-1} + 0.1(-w_{i-1} + 0.1(i-1) + 1) \\ &= 0.9w_{i-1} + 0.01(i-1) + 0.1 \end{aligned}$$

Figura 8.1: geometricamente o método de Euler aproxima a solução em t_i através da tangente à curva em t_{i-1}

para $i = 1, 2, ..., 10$.

t_i	w_i	y_i	$Erro = \|w_i - y_i\|$
0.0	1.000000	1.000000	0.0
0.1	1.000000	1.004837	0.004837
0.2	1.010000	1.018731	0.008731
0.3	1.029000	1.040818	0.011818
0.4	1.056100	1.070320	0.014220
0.5	1.090490	1.106531	0.016041
0.6	1.131441	1.148812	0.017371
0.7	1.178297	1.196585	0.018288
0.8	1.230467	1.249329	0.018862
0.9	1.287420	1.306570	0.019150
1.0	1.348678	1.367879	0.019201

A solução exata de (21) é $y(t) = t + e^{-t}$. *A tabela (58) mostra uma comparação entre os valores aproximados em* t_i *e os valores reais.*

Note que o crescimento do erro cresce um pouco com o crescimento do valor de t_i. Este crescimento de erro controlado é conseqüência da estabilidade do método de Euler, que implica espera-se que o erro devido ao arredondamento não deve crescer mais que uma forma linear.

Apesar do método de Euler não ser preciso o suficiente para ser garantido seu uso na prática, ele é suficientemente elementar para nós analisarmos o erro que é produzido com sua aplicação.

Para derivarmos um limite de erro para o método de Euler, precisamos dos seguintes resultados:

Lema 1 *Para todo* $x \geq -1$ *e qualquer número positivo* m,

$$0 \leq (1 + x)^m \leq e^{mx}. \tag{22}$$

PROVA: Aplicando o teorema de Taylor com $f(x) = e^x$, *em torno de* x_0, *e* $n = 1$, *temos,*

$$e^x = 1 + x + \frac{1}{2}x^2 e^{\xi}$$

onde ξ está entre zero e x. assim,

$$0 \leq 1 + x \leq 1 + x + \frac{1}{2}x^2 e^{\xi} = e^x$$

e como $1 + x \geq 0$, $\quad 0 \leq (1 + x)^m \leq (e^x)^m = e^{mx}$.

Lema 2 *Se s e t são números reais positivos, $\{a_i\}_{i=0}^{k}$ é uma seqüência satisfazendo $a_0 \geq -t/s$, e*

$$a_{i+1} \leq (1 + s)a_i + t, \text{ para cada } i = 0, 1, 2, ..., k, \tag{23}$$

then $a_{i+1} \leq e^{(i+1)s}\left(\frac{t}{s} + a_0\right) - \frac{t}{s}$.

Teorema 43 *Suponha que f é uma função contínua e que satisfaz a condição de Lipschitz com a constante L sobre o conjunto*

$$D = \{(t, y)|a \leq t \leq b, -\infty < y < \infty\},$$

e que exista uma constante M com a propriedade qu

$$y''(t)| \leq M \text{ para todo } t \in [a, b].$$

Seja $y(t)$ a solução única do problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

e $w_0, w_1, ...w_N$ as N aproximações geradas pelo método de Euler para algum inteiro positivo N. Então

$$|y(t_i - w_i| \leq \frac{hM}{2L}[e^{L(t_i - a)} - 1] \tag{24}$$

para cada $i = 0, 1, 2, ..., N$.

8.5 Métodos de Taylor de Ordem Superior

Como o objetivo dos Métodos Numéricos é determinar aproximações suficientemente precisas de um PVI, com o mínimo de esforço, nós precisamos ter uma ferramenta para comparar a eficiência dos vários métodos de aproximação. O primeiro instrumento que iremos considerar é chamado de erro de truncamento local (ETL). O ETL em um determinado passo mede a quantidade pala qual a solução exata da Equação diferencial falha para satisfazer a equação em diferenças que está sendo usada para a aproximação.

Definição 15 *O método em diferenças*

$$\begin{aligned} w_0 =& \alpha \\ w_{i+1} =& w_i + h\phi(t_i, y_i) \quad \textit{para cada } i = 0, 1, ..., N-1 \end{aligned} \tag{25}$$

possui **erro de truncamento local** *dado por*

$$\tau_i(h) = \frac{y_i - y_{i-1}}{h} - \phi(t_{i-1}, w_{i-1}) \quad \textit{para cada } i = 1, 2, ..., N \tag{26}$$

Assim, para o método de Euler, o erro de truncamento Local no iésimo passo para o problema

$$y' = f(t, y), \quad a \leq t \leq b, \;\; y(a) = \alpha$$

é

$$\tau_i(h) = \frac{y_i - y_{i-1}}{h} - f(t_{i-1}, y_{i-1}) \quad \text{para cada } i = 1, 2, ..., N.$$

onde como usual, $y_i = y(t_i)$ denota o valor exato da solução em t_i. Este erro é dito *erro local* porque mede a precisão do método em um passo específico, assumindo que no passo anterior era exato. Pode-se notar que este erro depende da Equação Diferencial, do tamanho do passo e de um passo particular na aproximação.

Considerando que para o método de Euler temos $y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i)$, vemos que para o método de Euler

$$\tau_i(h) = \frac{h}{2}y''(\xi_i) \quad \text{para algum } \xi_i, \quad \text{onde } t_{i-1} < \xi_i < t_i, \tag{27}$$

e quando sabemos que $y''(t)$ é limitado por uma constante M em $[a, b]$, isto implica que

$$|\tau_i(h)| \leq \frac{h}{2}M.$$

Assim o ETL do método de Euler é $O(h)$. Uma forma de selecionar metodos de equações em diferenças para aproximar a solução de uma EDO é escolher de tal forma que o ETL seja

$O(h^p)$ para p tão grande quanto possível, mantendo o número e complexidade de operações dentro de um limite razoável.

Como o método de Euler foi derivado usando o teorema de Taylor com $n = 2$ para aproximar a solução da equação diferencial, nossa primeira tentativa em encontrar métodos para melhorar as propriedades de convergência dos métodos de diferenças é extender esta mesma regra para um n maior.

Suponha que a solução $y(t)$ de um PVI

$$y' = f(t, y), \quad a \leq t \leq b, \;\; y(a) = \alpha,$$

possui $(n + 1)$ derivadas contínuas. Se expandimos a solução $y(t)$ em série de Taylor em torno de t_i e calcularmos em t_{i+1} temos,

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(t_i) + ... + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i) \tag{28}$$

para algum ξ_i, $t_i < \xi_i < t_{i+1}$.

Diferenciações sucessivas da solução ,$y(t)$, nos dá,

$$\begin{aligned} y'(t) =& f(t, y(t)) \\ y''(t) =& f'(t, y(t)) \end{aligned}$$

$$y^{(k)}(t) = f^{(k-1)}(t, y(t)).$$

Substituindo estes resultados na equação (28) temos,

$$y(t_{i+1}) = y(t_i) + hf((t_i, y(t_i))) + \frac{h^2}{2}f(t_i', y(t_i))) + ... + \frac{h^n}{n!}f^{(n-1)}(t_i, y(t_i)) + \frac{h^{n+1}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)). \tag{29}$$

A equação em diferença correspondente a equação (29) é obtida não considerando o último termo envolvendo ξ_i. Este método é chamado de **Método de Taylor de ordem n**:

$$\begin{aligned} w_0 =& \alpha, \\ w_i + 1 =& w_i + hT^{(n)}(t_i, w_i) \qquad \text{para cada } i = 0, 1, ..., N - 1, \\ \text{onde } T^{(n)}(t_i, w_i) =& f(t_i, w_i) + \frac{h}{2}f(t_i, w_i) + ... + \frac{h^{n-1}}{n!}f^{(n-1)}(t_i, w_i) \end{aligned} \tag{30}$$

Note que o método de Euler é o método de Taylor de ordem um. Note também que para aplicarmos os métodos de Taylor de ordem superior, temos que derivar $(n - 1)$ vezes a função $f(t, y(t))$ em relação à t. Assim, se $f(t, y)$ não for uma função muito simples o método torna-se inviável.

8.6 Métodos de Runge Kutta

- *Métodos de Taylor*
  - Alta ordem de Erro de Truncamento Local (ETL)
  - Porém requer cálculo das derivadas de $f(t, y)$, oque consome tempo computacional e pode ser complicado
  - Quase nunca são usados na prática
- *Métodos de Runge Kutta*
  - Usa o ETL de alta ordem dos métodos de Taylor
  - Elimina o cálculo das derivadas de $f(t, y)$

8.6.1 Introdução

Fórmulas do tipo de Runge Kutta estão entre as fórmulas mais largamente usadas para a solução numérica de EDO. Suas vantagens incluem:

- São facilmente programáveis
- Possuem boas propriedades de estabilidade
- O tamanho do passo pode ser trocado livremente sem complicações
- Ele se auto inicia

Suas desventagens são em primeiro lugar que computacionalmente podem precisar de mais tempo que outros métodos para obter a mesma precisão e a estimativa do erro local não são simples de serem obtidas. Muitas vezes o método de Runge Kutta é usado para resolver problemas inteiros, mas analistas numéricos recomendam seu uso como inicializador de métodos mais eficientes de multipasso como o preditor corretor.

Preliminares Matemáticos

Teorema 44 *(Taylor em duas variáveis)*
Suponha que $f(t, y)$ e todas as suas derivadas de ordem menor ou igual que $n + 1$ são contínuas em

$$D = \{(t, y)/a \leq t \leq b, c \leq y \leq d\}$$

. *Seja* $(t_0, y_0) \in D$. *Para todo* (t, y) *em* D, *existe* ξ *entre* t *e* t_0 *e* η *entre* y *e* y_0 *com*

$$f(t, y) = P_n(t, y) + R_n(t, y),$$

onde

$$\begin{aligned} P_n(t,y) &= f(t_0,y_0) + \left[(t-t_0)\frac{\partial f}{\partial t}(t_0,y_0) + (y-y_0)\frac{\partial f}{\partial y}(t_0,y_0)\right] \\ &\quad + \left[\frac{(t-t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0,y_0) + (t-t_0)(y-y_0)\frac{\partial^2 f}{\partial t \partial y}(t_0,y_0) + \frac{(y-y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(t_0,y_0)\right] \\ &\quad + ... \\ &\quad + \left[\frac{1}{n!}\sum_{j=0}^{n}\binom{n}{j}(t-t_0)^{n-j}(y-y_0)^{n-j}\frac{\partial^n f}{\partial t^{n-j}\partial y^j}(t_0,y_0)\right] \end{aligned}$$

and

$$R_n(t,y) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1}\binom{n+1}{j}(t-t_0)^{n+1-j}(y-y_0)^j\frac{\partial^{n+1} f}{\partial t^{n+1-j}\partial y^j}(\xi,\eta)$$

A função P_n é chamada de *Polinômio de Taylor de grau n em duas variáveis* para a função f em torno de (t_0, y_0), e $R_n(t, y)$ o resto associado à $P_n(t, y)$.

Exemplo 59 *O polinômio de Taylor de segundo grau para* $f(t, y) = \sqrt{4t + 12y - t^2 - 2y^2 - 6}$ *em torno de* $(2, 3)$ *é encontrado a partir de*

$$\begin{aligned} P_2(t,y) &= f(t_0,y_0) + \left[(t-t_0)\frac{\partial f}{\partial t}(t_0,y_0) + (y-y_0)\frac{\partial f}{\partial y}(t_0,y_0)\right] \\ &\quad + \left[\frac{(t-t_0)^2}{2}\frac{\partial^2 f}{\partial t^2}(t_0,y_0) + (t-t_0)(y-y_0)\frac{\partial^2 f}{\partial t \partial y}(t_0,y_0) + \frac{(y-y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(t_0,y_0)\right] \end{aligned}$$

Calculando cada uma destas derivadas parciais em $(t_0, y_0) = (2, 3)$, $P_2(t, y)$ *reduz-se à:*

$$P_2(t,y) = 4 - \frac{1}{4}(t-2)^2 - \frac{1}{2}(y-3)^2.$$

Este polinômio fornece uma aproximação precisa de $f(t, y)$ *quando* t *está perto de* 2 *e* y *está perto de* 3, *por exemplo:*

$$P_2(2.1, 3.1) = 3.9925 \quad e \quad f(2.1, 3.1) = 3.9962.$$

Entretanto, a precisão da aproximação deteriora rapidamente quando (t, y) *se move para longe de* $(2, 3)$.

Figura 8.2: funções $f(t,y)$ e $P_2(t,y)$

O método de Runge-Kutta evita a dificuldade de derivação da função $f(t,y)$ que aparece nos métodos de Taylor de ordem superior. Para isto, assim ao invés de usar a série de Taylor este método aproxima o polinômio de Taylor por uma combinação de valores da função $f(t,y)$ em diversos pontos. Vamos ilustrar este procedimento fazendo a derivação do método de Runge Kutta de segunda ordem.

Derivação de um método de Runge Kutta de ordem 2
Para mostrar qual o procedimento seguido no desenvolvimento dos métodos de Runge-Kutta, vamos desenvolver o chamado de **Método do Ponto Médio**, que é um Método de Runge-Kutta de Segunda Ordem.

1. Primeiramente vamos supor que o polinômio de Taylor de segunda ordem possa ser aproximado por uma combinação de valores da função $f(t,y)$. Neste caso vamos supor que

   $$T^{(2)} = a_1 f(t+\alpha_1, y+\beta_1)$$

2. Para determinar os valores das constantes a_1, α_1 e $beta_1$ de forma que $a_1 f(t+\alpha_1, y+\beta_1)$ aproxime o polinômio de Taylor

   $$T^{(2)} = f(t,y) + \frac{h}{2} f'(t,y)$$

   com erro menor ou igual à $O(h^2)$, vamos primeiramente encontrar uma expressão para a derivada de $f(t,y)$ em relação a t, lembrando que $y = y(t)$,

   $$f'(t,y) = \frac{df}{dt}(t,y) = \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y).y'(t) \quad e \quad y'(t) = f(t,y).$$

3. Substituindo a expressão acima no pol. de Taylor, temos

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2}\frac{\partial f}{\partial t}(t,y) + \frac{h}{2}\frac{\partial f}{\partial y}(t,y).f(t,y). \tag{31}$$

4. Por outro lado, vamos expandindo a função $f(t+\alpha_1, y+\beta_1)$ em série de Taylor, temos que o pol. de Taylor de grau 1 em torno de (t,y) para esta expansão é dada por:

$$a_1 f(t+\alpha_1, y+\beta_1) = a_1 f(t,y) + a_1\alpha_1\frac{\partial f}{\partial t}(t,y) + a_1\beta_1\frac{\partial f}{\partial y}(t,y) + a_1.R_1(t+\alpha_1, y+\beta_1), \tag{32}$$

onde o termo de erro é dado por,

$$R_1(t+\alpha_1, y+\beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi,\eta) + \alpha_1^2\beta_1\frac{\partial^2 f}{\partial t\partial y}(\xi,\eta) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi,\eta) \tag{33}$$

5. Agora, igualando os coeficientes de f e suas derivadas nas equações (31) e (32),

$$\begin{aligned} f(t,y): &\qquad a_1 = 1; \\ \frac{\partial f}{\partial t}(t,y): &\qquad a_1\alpha_1 = \frac{h}{2} \\ \frac{\partial f}{\partial y}(t,y): &\qquad a_1\beta_1 = \frac{h}{2}f(t,y) \end{aligned}$$

Desta forma os parâmetros a_1, α_1 e β_1 são determinados de forma única como:

$$a_1 = 1, \quad \alpha_1 = \frac{h}{2} \quad e \quad \beta_1 = \frac{h}{2}f(t,y)$$

logo

$$T^{(2)}(t,y) = f\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) - R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$$

e da equação (33) obtemos a seguinte fórmula para o erro:

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi,\eta) + \frac{h^2}{4}f(t,y)\frac{\partial^2 f}{\partial t\partial y}(\xi,\eta) + \frac{h^2}{8}\frac{\partial^2 f}{\partial y^2}(\xi,\eta). \tag{34}$$

6. Assim, se todas as derivadas de segunda ordem da f são limitadas, pela fórmula do erro dada acima pela eq. (34) temos que este método é $O(h^2)$.

Assim, a equação en diferenças para o **método do Ponto Médio** é:

$$\begin{aligned} w_0 &= \alpha, \\ w_{i+1} &= w_i + hf\left(t_i + \frac{h}{2}, w_i + \frac{h}{2}f(t_i,w_i)\right) \quad para \quad i = 0,1,...,N-1. \end{aligned} \tag{35}$$

Note, uma vez que somente 3 parâmetros estão presentes na expressão $a_1 f(t+\alpha_1, y+\beta_1)$, e que todos são necessários para "casar"os termos com $T^{(2)}$, precisaremos de uma forma mais complicada para satisfazer as condições requeridas para métodos de Taylor de mais alta ordem. A forma mais apropriada com quatro constantes para aproximar,

$$T^{(3)} = f(t,y) + \frac{h}{2} f'(t,y) + \frac{h^2}{6} f''(t,y)$$

é

$$a_1 f(t,y) + a_2 f(t+\alpha_2, y+\delta_2 f(t,y)); \tag{36}$$

Fazendo os algebrismos necessários, vemos que mesmo com este número de parâmetros ainda não temos flexibilidade suficiente para "casar"o termo

$$\frac{h^2}{6}\left[\frac{\partial f}{\partial y}(t,y)\right]^2 f(t,y)$$

proveniente da expansão de $(h^2/6)f''(t,y)$. Desta forma o melhor que podemos obter com métodos vindos de (36) são métodos com ETL de $O(h^2)$. Os dois métodos mais importantes de Runge Kutta de segunda ordem são:

1. **Método de Euler Modificado**: Corresponde a escolha de $a_1 = a_2 = \frac{1}{2}$ e $\alpha_2 = \delta_2 = h$, assim sua equação em diferenças é:

$$\begin{aligned} w_0 &= \alpha, \\ w_{i+1} &= w_i + \frac{h}{2}\left(f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))\right) \quad para \quad i = 0, 1, ..., N-1 \end{aligned} \tag{37}$$

2. **Método de Heun**: Corresponde a escolha de $a_1 = \frac{1}{4}$, $a_2 = \frac{3}{4}$ e $\alpha_2 = \delta_2 = \frac{2}{3}h$, assim sua equação em diferenças é:

$$\begin{aligned} w_0 &= \alpha, \\ w_{i+1} &= w_i + \frac{h}{4}\left(f(t_i, w_i) + 3f(t_i + \frac{2}{3}h, w_i + \frac{2}{3}hf(t_i, w_i))\right) \quad para \quad i = 0, 1, ..., N \end{aligned} \tag{38}$$

Exemplo 60 *Não vamos aplicar os métodos de Runge Kutta de ordem 2 no nosso exemplo,* $y' = -y + t + 1 \quad 0 \le t \le 1, \quad y(0) = 1$ *pois por sua natureza, obtemos para todos a mesma*

equação em diferenças que o método de Taylor de ordem 2. Para comparar os vários métodos vamos aplica-lo ao PVI,

$$y' = -y + t^2 + 1 \quad 0 \leq t \leq 1, \quad y(0) = 1$$

cuja solução exata é dada por, $y(t) = -2e^{-t} + t^2 - 2t + 3$.

Considerando $N = 10$, $h = 0.1$ *e* $t_i = 0.1i$, *as equações de diferença são:*

$$\begin{array}{ll} \textit{Ponto Médio} & w_{i+1} = 0.905w_i + 0.00095i^2 + 0.001i + 0.09525 \\ \textit{Euler Modificado} & w_{i+1} = 0.905w_i + 0.00095i^2 + 0.001i + 0.0955 \\ \textit{Heun} & w_{i+1} = 0.905w_i + 0.00095i^2 + 0.001i + 0.095333333 \end{array}$$

para $i = 0, 1, ...9$, *onde* $w_0 = 1$ *em todos os casos.*

t_i	Exato	Ponto Médio	Erro	Euler Mod.	Erro	Heun	Erro
0.0	1.000000	1.000000		1.000000		1.000000	
0.1	1.000325	1.000250	7.52×10^{-5}	1.000500	1.75×10^{-4}	1.000333	8.10×10^{-6}
0.2	1.002538	1.002426	1.12×10^{-4}	1.002902	3.64×10^{-4}	1.002595	4.65×10^{-5}
0.3	1.008363	1.008246	1.18×10^{-4}	1.008927	5.63×10^{-4}	1.008473	1.09×10^{-4}
0.4	1.019360	1.019264	9.75×10^{-5}	1.020129	7.69×10^{-4}	1.019551	1.91×10^{-4}
0.5	1.036939	1.036882	5.62×10^{-5}	1.037917	9.78×10^{-4}	1.037227	2.88×10^{-4}
0.6	1.062377	1.062379	2.00×10^{-6}	1.063564	1.19×10^{-3}	1.062774	3.97×10^{-4}
0.7	1.096829	1.096903	7.33×10^{-5}	1.098226	1.40×10^{-3}	1.097344	5.14×10^{-4}
0.8	1.141342	1.141497	1.55×10^{-4}	1.142944	1.60×10^{-3}	1.141979	6.37×10^{-4}
0.9	1.196861	1.197105	2.44×10^{-4}	1.198665	1.80×10^{-3}	1.197625	7.64×10^{-4}
1.0	1.264241	1.264580	3.39×10^{-4}	1.266242	2.00×10^{-3}	1.265134	8.93×10^{-4}

Tabela 8.1: Tabela resultado do exercício (60)

Nota 5 $T^{(3)}(t, y)$ *pode ser aproximada com erro* $O(h^3)$, *mas por ser complicado o algebrismo e por quase não ser usado na prática não vamos apresenta-lo.*

O método mais comum de Runge Kutta é o **Runge Kutta de Quarta Ordem** cuja equação em diferença é dada por:

$$\begin{aligned} w_0 &= \alpha, \\ k_1 &= hf(t_i, w_i), \\ k_2 &= hf(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1), \\ k_3 &= hf(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2), \\ k_4 &= hf(t_{i+1}, w_i + k_3), \end{aligned} \tag{39}$$

$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

para $i = 0, 1, ...N-1$. O ETL deste método é $O(h^4)$, se a solução y(t) possuir cinco derivadas contínuas. O motivo da introdução de k_1, k_2, k_3 e k_4 no método é eliminar repetição de cálculos na segunda variável de $f(t, y)$.

ALGORITMO 9 *(Runge Kutta Quarta Ordem)*
Aproxima a solução do problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha,$$

em $N+1$ pontos igualmente espaçados no intervalo $[a, b]$:
ENTRADA: a, b; N; α.
SAIDA: Aproximação w de y em $N+1$ valores de t.
PASSO 1: Faça

$$\begin{aligned} h &= (b-a)/N; \\ t &= a; \\ w &= \alpha\ ; \end{aligned}$$

SAÍDA: (t, w).

PASSO2: Para $i = 1, 2, ...N$ (3-5)

PASSO 3: Faça:

$$\begin{aligned} K_1 &= hf(t, w), \\ K_2 &= hf(t + \tfrac{h}{2}, w + \tfrac{1}{2}K_1), \\ K_3 &= hf(t + \tfrac{h}{2}, w + \tfrac{1}{2}K_2), \\ K_4 &= hf(t + h, w + k), \end{aligned}$$

PASSO 4: Faça

$$\begin{aligned} w &= w + (K_1 + 2K_2 + 2K_3 + K_4)/6; \\ t &= a + ih \end{aligned}$$

PASSO 5: SAÍDA: (t, w).

Exemplo 61 *Usando o método de RK de quarta ordem obtenha uma aproximação para o PVI:*

$$y' = -y + t + 1, \quad 0 \leq t \leq 1, \quad y(0) = 1,$$

com $h = 0.1$, $N = 10$ *e* $t_i = 0.1i$.

t_i	Valor Exato	RK quarta ordem	Erro
0.0	1.0000000000	1.0000000000	
0.1	1.0048374180	1.0048375000	8.200×10^{-8}
0.2	1.0187307531	1.0187309014	1.483×10^{-7}
0.3	1.0408182207	1.0408184220	2.013×10^{-7}
0.4	1.0703200460	1.0703202889	2.429×10^{-7}
0.5	1.1065306597	1.1065309344	2.747×10^{-7}
0.6	1.1488116360	1.1488119344	2.984×10^{-7}
0.7	1.1965853038	1.1965856187	3.149×10^{-7}
0.8	1.2493289641	1.2493292897	3.256×10^{-7}
0.9	1.3065696597	1.3065699912	3.315×10^{-7}
1.0	1.3678794412	1.3678797744	3.332×10^{-7}

Tabela 8.2: Tabela dos resultados do exercício (61)

Exemplo 62 *Considere o exemplo*

$$y' = -y + 1, \quad 0 \leq t \leq 1, \quad y(0) = 0,$$

Para resolver este problema foi aplicado o método de Euler com $h = 0.025$*, o método de Euler modificado com* $h = 0.05$ *e o método de Runge-Kutta de quarta ordem com* $h = 0.1$*. Os resultados obtidos estão comparados na tabela abaixo com os valores exatos nos pontos* 0.1, 0.2, 0.3, 0.4 *e* 0.5 *da malha.*

Cada uma destas técnicas requer 20 cálculos funcionais para determinar os valores liatados na tabela. Neste exemplo o método RK de quarta ordem é claramente superior.

t	Método de Euler	Método de Euler Modificado	Runge Kutta Quarta Ordem	Valor Exato
0.1	0.096312	0.095123	0.09516250	0.095162582
0.2	0.183348	0.181198	0.18126910	0.181269247
0.3	0.262001	0.259085	0.25918158	0.259181779
0.4	0.333079	0.329563	0.32967971	0.329679954
0.5	0.397312	0.393337	0.39346906	0.393469340

Tabela 8.3: Tabela dos resultados obtidos para o exercício (62)

Exemplo 63 *1. Determine a solução numérica aproximada da seguinte Equação Diferencial Ordinária, com o passo $h = 0.2$:*

$$\begin{cases} y'(x) + 2y(x) = 0, & x \in [0,1] \\ y(0) = 1 \end{cases}$$

(a) Método de Euler

(b) Método de Euler Modificado

(c) Método de Runge-Kutta 4

(d) Sabendo-se que a solução exata da equação é $y(x) = e^{-2x}$, compare com as soluções aproximadas obtidas nos items anteriores

Solução: Primeiramente identificamos $f(x,y) = -2y$, $h = 0.2$ e $x \in [0,1]$ logo $w_0 = 1$

(a) Euler $w_{i+1} = w_i + hf(x_i, w_i)$, com $w_0 = 1$, $x_0 = 0$ e $x_i = 0 + ih$, logo temos:

$$\begin{array}{ll} i = 0 & w_1 = w_0 + hf(x_0, y_0) = 1 + 0.2(-2) = 1 - 0.4 = 0.6 \\ i = 1 & w_2 = w_1 + hf(x_1, w_1) = 0.6 + 0.2(-1.2) = 0.6 - 0.24 = 0.36 \\ i = 2 & w_3 = w_2 + hf(x_2, w_2) = 0.36 + 0.2(-0.72) = 0.36 - 0.144 = 0.216 \\ i = 3 & w_4 = w_3 + hf(x_3, w_3) = 0.216 + 0.2(-0.432) = 0.1296 \\ i = 4 & w_5 = w_4 + hf(x_4, w_4) = 0.1296 + 0.2(-0.2592) = 0.07776 \end{array}$$

(b) Euler Modificado $w_{i+1} = w_i + \frac{h}{2}\{f(x_i, w_i) + f(x_{i+1}, w_i + hf(x_i, w_i))\}$, com $w_0 = 1$, $x_0 = 0$ e $x_i = 0 + ih$, logo temos:

$$\begin{array}{ll} i = 0 & k_1 = f(x_0, w_0) = -2w_0 = -2 \\ & k_2 = f(x_1, w_0 + hf(x_0, w_0)) = f(0.2, 1 + 0.2(-2)) = f(0.2, 0.6) = -1.2 \\ & w_1 = w_0 + \frac{0.2}{2}(k_1 + k_2) = 1 + 0.1(-2 - 1.2) = 1 - 0.32 = 0.68 \\ i = 1 & k_1 = f(x_1, w_1) = -2w_1 = -1.36 \\ & k_2 = f(x_2, w_1 + hf(x_1, w_1)) = f(0.4, 0.68 + 0.2(-1.36)) = f(0.4, 0.408) = -0.816 \\ & w_2 = w_1 + \frac{0.2}{2}(k_1 + k_2) = 0.68 + 0.1(-1.36 - 0.816) = 0.4624 \\ i = 2 & k_1 = f(x_2, w_2) = -2w_2 = -0.9248 \\ & k_2 = f(x_3, w_2 + hf(x_2, w_2)) = f(0.6, 0.4624 + 0.2(-0.9248)) = f(0.6, 0.27744) = -0. \\ & w_3 = w_2 + \frac{0.2}{2}(k_1 + k_2) = 0.4624 + 0.1(-0.9248 - 0.55488) = 0.314432 \\ i = 3 & k_1 = f(x_3, w_3) = -2w_3 = -0.628864 \\ & k_2 = f(x_4, w_3 + hf(x_3, w_3)) = f(0.8, 0.314432 + 0.2(-0.1886592)) = f(0.8, 0.188659 \\ & w_4 = w_3 + \frac{0.2}{2}(k_1 + k_2) = 0.314432 + 0.1(-0.628864 - 0.3773184) = 0.21381376 \\ i = 4 & k_1 = f(x_4, w_4) = -2w_4 = -0.42762752 \\ & k_2 = f(x_5, w_4 + hf(x_4, w_4)) = f(1.0, 0.21381376 + 0.2(-0.42762752)) = f(1.0, 0.128 \\ & \qquad = -0.256576512 \\ & w_5 = w_4 + \frac{0.2}{2}(k_1 + k_2) = 0.21381376 + 0.1(-0.42762752 - 0.256576512) = \\ & \qquad = 0.21381376 + 0.1(-0.684204032) = 0.1453933 \end{array}$$

2. Deixado como exercício

3. Deixado como exercício

8.7 Problema de Valor Inicial: Sistemas de EDO e EDO de Ordem Superior

Um sistema de EDO de primeira ordem pode ser escrito como:

$$\left\{\begin{array}{rcl} \frac{d}{dt}y_1(t) & = & f_1(t, y_1, y_2, ..., y_n) \\ \frac{d}{dt}y_2(t) & = & f_2(t, y_1, y_2, ..., y_n) \\ & \vdots & \\ \frac{d}{dt}y_n(t) & = & f_n(t, y_1, y_2, ..., y_n) \end{array}\right. \tag{40}$$

O sistema acima possui n funções incógnitas, são elas $y_1(t)$, $y_2(t)$, ... e $y_n(t)$, e n EDO envolvendo estas funções incógnitas. O problema de valor inicial, associa ao sistema de equações (40) um conjunto de condições iniciais definido por:

$$\left\{\begin{array}{l} y_1(a) = \alpha_1 \\ y_2(t_0) = \alpha_2 \\ \vdots \\ y_n(t_0) = \alpha_n \end{array}\right. . \tag{41}$$

Uma notação conveniente para este sistema o sistema de equações (40) é a notação vetorial. Pra isto, chamamos do $\mathbf{Y}(t)$ ao vetor coluna definido por

$$\mathbf{Y}(t) = [y_1(t), y_2(t), ..., y_n(t)]^T.$$

Desta forma $\mathbf{Y}(t)$ é uma função de $\mathcal{R}$ (ou um intervalo de $\mathcal{R}$) em $\mathcal{R}^n$. De forma similar vamos notar o vetor $\mathbf{F} = [f_1, f_2, ..., f_n]^T$, onde cada função $f_i = f_i(t, y_1, y_2, ..., y_n)$ onde cada f_i é uma função sobre $\mathcal{R}^{n+1}$ (ou subintervalo deste. Desta forma $\mathbf{F}$ é uma função de $\mathcal{R}^{n+1}$ em $\mathcal{R}^n$. Com esta notação o sistema (40) pode ser escrito como:

$$\frac{d}{dt}\mathbf{Y} = \mathbf{F}(t, \mathbf{Y}) \tag{42}$$

e a condição inicial é dada por $\mathbf{Y}(t_0) = [\alpha_1, \alpha_2, ..., \alpha_n]^T$.

Por exemplo

Exemplo 64 *Como exemplo vamos considerar o seguinte sistema linear fazendo $x_1 = x$ e $x_2 = y$,*

$$\left\{\begin{array}{rcl} \frac{d}{dt}x(t) & = & x + 4y - e^t \\ \frac{d}{dt}y(t) & = & x + \ t + 2e^t \end{array}\right. \tag{43}$$

com condição de contorno dada por $x(0) = 4$ e $y(0) = 5/4$, possui solução única dada por:

$$\left\{\begin{array}{rcl} x(t) & = & 4e^{3t} + 2e^{-t}2e^t \\ y(t) & = & 2e^{3t} - e^{-t} + \frac{1}{4}e^t \end{array}\right. \tag{44}$$

Matricialmente temos:

$$\mathbf{Y}(t) = \left(\begin{array}{c} x(t) \\ y(t) \end{array}\right), \quad \mathbf{F}(t, \mathbf{Y}) = \left(\begin{array}{c} f_1(t, x, y) \\ f_2(t, x, t) \end{array}\right) = \left(\begin{array}{c} x + 4y - e^t \\ x + \ t + 2e^t \end{array}\right)$$

Aqui também devemos chamar a atenção que, conforme foi dito no início deste capítulo, um problema de valor inicial de ordem superior pode ser escrito como um sistema de EDO. Para isto, vamos supor que temos o seguinte PVI de ordem n

$$z^{(n)} = f(t, z, z', ..., z^{(n-1)})$$

com

$$z(t_0) = \alpha_1, \quad z'(t_0) = \alpha_2, \quad ...., \quad z^{(n-1)}(t_0) = \alpha_n.$$

Para resolvermos este sistema pelos métodos estudados anteriormente, a seguir, introduzimos a mudança de varáveis:

$$y_1 = z(t) \quad y_1 = z'(t) \quad y_3 = z''(t) \quad ... \quad y_n = z^{(n-1)}(t)$$

Estas novas variáveis satisfazem ao seguinte sistema de EDO de primeira ordem:

$$\begin{cases} \frac{d}{dt}y_1(t) &= y_2(t) \\ \frac{d}{dt}y_2(t) &= y_3(t) \\ \frac{d}{dt}y_3(t) &= y_4(t) \\ & \vdots \\ \frac{d}{dt}y_n(t) &= f(t, y_1, y_2, ..., y_n) \end{cases}$$

com a condição inicial

$$\begin{cases} y_1(t_0) &= \alpha_1 \\ y_2(t_0) &= \alpha_2 \\ y_3(t_0) &= \alpha_3 \\ & \vdots \\ y_n(t_0) &= \alpha_n) \end{cases}$$

Exemplo 65 *Vamos considerar o seguinte problema de valor inicial*

$$\begin{cases} (\sin t)z''' + \cos(ty) + \sin(t^2 + z'') + (y')^3 = \ln t \\ z(2) = 7 \\ z'(2) = 3 \\ z''(2) = -4 \end{cases}$$

Vamos converter este sistema em um PVI de primeira ordem fazendo a mudança de variáveis:

$$y_1(t) = z(t), \quad y_2(t) = z'(t), \quad \text{e } y_3(t) = z''(t)$$

o sistema é dado por:

$$\begin{cases} \dfrac{d}{dt}y_1(t) = y_2(t) \\ \\ \dfrac{d}{dt}y_2(t) = y_3(t) \\ \\ \dfrac{d}{dt}y_3(t) = \dfrac{\ln t - x_2^3 - \sin(t^2 + y_3) - \cos(ty_1)}{\sin(t)} \end{cases}$$

com condição inicial $\mathbf{Y}(2) = (7, 3, -4)^T$.

8.7.1 Resolução de Sistemas de EDO para PVI

Os sistemas de EDO podem ser resolvidos pelos métodos já discutidos nas primeiras seções deste capítulo. Assim podemos usar o métodos de Euler, Taylor de ordem superior, Runge-Kutta, os métodos de Adams Bashforth, ou preditor corretor. Por exemplo o método de Runge Kutta de quarta ordem na forma matricial é dado por:

$$\mathbf{W}(t+h) = \mathbf{W}(t) + \frac{1}{6}(\mathbf{K}_1 + 2\mathbf{K}_2 + 2\mathbf{K}_3 + \mathbf{K}_4) \tag{45}$$

onde

$$\begin{cases} \mathbf{K}_1 = h\mathbf{F}(t, \mathbf{X}) \\ \mathbf{K}_2 = h\mathbf{F}(t + \frac{1}{2}h, \mathbf{K}_1) \\ \mathbf{K}_3 = h\mathbf{F}(t + \frac{1}{2}h, \mathbf{K}_2) \\ \mathbf{K}_4 = h\mathbf{F}(t + h, \mathbf{K}_3) \end{cases}$$

Todos os métodos desenvolvidos para resolução de PVI podem ser aplicados para a solução de sistema de EDO, inclusive os métodos de Multistep. Esta extensão é feita de forma direta.

Como exemplo vamos resolver um PVI de segunda ordem por um método de Runge Kutta de ordem 2.

Exemplo 66 *Neste exemplo, vamos considerar o seguinte PVI de segunda ordem:*

$$\begin{aligned} &y'' - 2y' + 2y = e^{2t}\sin t, \quad t > 0 \\ &y(0) = -0.4 \\ &y'(0) = -0.6 \end{aligned}$$

Encontre uma aproximação para a solução deste problema em $t = 0.2$ usando o método de Heun , que é um método de Runge Kutta de segunda ordem definido por:

$$\begin{aligned} &w_0 = \alpha \\ &w_{i+1} = w_i + \tfrac{h}{4}\left(f(t_i, w_i) + 3f(t_i + \tfrac{2}{3}h\ ,\ w_i + \tfrac{2}{3}hf(t_i, w_i))\right) \quad i = 0, 1, ... \end{aligned}$$

use $h = \triangle t = 0.1$

Solução: *Primeiramente vamos proceder a seguinte mudança de variáveis:$y' = z$ e portanto $y'' = z'$, assim temos as equações:*

$$\begin{aligned} &y' = z \\ &z' = 2z - 2y + e^{2t}\sin t \end{aligned}$$

Na forma matricial temos:

$$\frac{d}{dt}\begin{pmatrix} y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} z \\ e^{2t}\sin t + 2z - 2y \end{pmatrix}, \qquad W_0 = \begin{pmatrix} y(0) \\ z(0) \end{pmatrix} = \begin{pmatrix} -0.4 \\ -0.6 \end{pmatrix}, \quad \text{onde } \mathbf{W}(t) = \begin{pmatrix} y(t) \\ z(t) \end{pmatrix}$$

Assim temos que

$$F(t, \mathbf{W}) = \begin{pmatrix} z \\ e^{2t}\sin t + 2z - 2y \end{pmatrix}$$

$h = 0.1$ *Para* $i = 0$ *temos:*

$$\mathbf{W}_1 = \mathbf{W}_0 + \frac{0.1}{4}\left\{\mathbf{F}\left(0, \begin{pmatrix}-0.4\\-0.6\end{pmatrix}\right) + 3F\left(\frac{0.2}{3}, \mathbf{W}_0 + \frac{2}{3}(0.1)\mathbf{F}\left(0, \begin{pmatrix}-0.4\\-0.6\end{pmatrix}\right)\right)\right\}$$

$$\mathbf{W}_1 = \mathbf{W}_0 + \frac{0.1}{4}\left(\begin{pmatrix}-0.6\\-0.4\end{pmatrix} + 3\mathbf{F}\left(\frac{0.2}{3}, \begin{pmatrix}-0.44\\-0.626667\end{pmatrix}\right)\right)$$

$$\mathbf{W}_1 = \begin{pmatrix}-0.4\\-0.6\end{pmatrix} + \frac{0.1}{4}\left\{\begin{pmatrix}-0.4\\-0.6\end{pmatrix} + 3\begin{pmatrix}-0.626667\\-0.297214\end{pmatrix}\right\} = \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix}$$

para $i = 1$ *temos:*

$$\mathbf{W}_2 = \mathbf{W}_1 + \frac{0.1}{4}\left\{\mathbf{F}\left(0.1, \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix}\right) + 3F\left(0.1 + \frac{0.2}{3}, \mathbf{W}_1 + \frac{2}{3}(0.1)\mathbf{F}\left(0.1, \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix}\right)\right)\right\}$$

$$\mathbf{W}_2 = \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix} + \frac{0.1}{4}\left(\begin{pmatrix}-0.632291\\-0.218645\end{pmatrix} + 3\mathbf{F}\left(0.166667, \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix} + \frac{0.2}{3}\begin{pmatrix}-0.632291\\-0.218645\end{pmatrix}\right)\right)$$

$$\mathbf{W}_2 = \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix} + \frac{0.1}{4}\left(\begin{pmatrix}-0.632291\\-0.218645\end{pmatrix} + 3\mathbf{F}\left(0.166667, \begin{pmatrix}-0.504153\\-0.646867\end{pmatrix}\right)\right)$$

$$\mathbf{W}_2 = \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix} + \frac{0.1}{4}\left(\begin{pmatrix}-0.632291\\-0.218645\end{pmatrix} + 3\begin{pmatrix}-0.646867\\-0.0539027\end{pmatrix}\right) = \begin{pmatrix}-0.462000\\-0.632291\end{pmatrix} + \frac{0.1}{4}\begin{pmatrix}-2.572890\\-0.380353\end{pmatrix}$$

$$\mathbf{W}_2 = \begin{pmatrix}-0.526322\\-0.641800\end{pmatrix}$$

E desta forma temos que $y(0.2) = -0.526322$. *Aqui notamos que por causa da mudança de variáveis feita, conhecemos também o valor de* $y'(0.2) = -0.641800$.

Capítulo 9

Problema de Valor de Contorno

Neste capítulo vamos estudar métodos numéricos para a resolução de Problemas de Valor de Contorno (PVC).

9.1 Existência de Solução

Primeiramente considerar o problema:

$$\begin{cases} y'' = f(t, y, y') \\ y(a) = \alpha \quad y(b) = \beta \end{cases} \tag{1}$$

Devemos observar que teoremas de existência para este tipo de problema tendem a ficar complicados e podem ser encontrados em textos avançados de Análise Numérica. Aqui vamos citar um resultado devido a Keller(1968):

Teorema 45 *O problema de valor de contorno*

$$\begin{cases} y'' = f(t, y) \\ y(0) = 0 \quad y(1) = 0 \end{cases}$$

possui solução única se $\partial f/\partial y$ *é contínua, não negativa e limitada na faixa infinida definida pelas desigualdades* $0 \leq t \leq 1$, $-\infty < y < \infty$.

O teorema anterior é aplicado apenas a um caso especial de PVC, mas uma simples troca de variáveis pode reduzir casos mais gerais neste caso especial. Para fazer isto primeiramente procedemos a troca no intervalo t. Vamos supor que o problema original é:

$$\begin{cases} y'' = f(t, y) \\ y(a) = \alpha \quad y(b) = \beta \end{cases} \tag{2}$$

A troca de variáveis $t = a + (b - a)s$ é feita aqui. Note que quando $s = 0$ temos $t = a$ e que quando $s = 1$ temos $t = b$. Mais ainda $x(s) = y(a+\lambda s)$ com $\lambda = b-a$. Então $x'(s) = \lambda y'(a+\lambda s)$

e $x''(s) = \lambda^2 y''(a + \lambda s)$. Da mesma forma $x(0) = y(a) = \alpha$ e $x(1) = y(b) = \beta$. Desta forma, se y é solução de (2), então x é a solução do PVC

$$\begin{cases} x'' = \lambda^2 f(a + \lambda s, x(s)) \\ x(0) = \alpha \quad x(1) = \beta \end{cases} \tag{3}$$

Na direção contrária se x é solução de (3), então a função $y(t) = x((t-a)/(b-a))$ é a solução de (2).

Teorema 46 *Considere os dois PVCs:*

$$\textbf{(i)} \begin{cases} y'' = f(t, y) \\ y(a) = \alpha \quad y(b) = \beta \end{cases} \qquad \textbf{(ii)} \begin{cases} x'' = g(t, x) \\ x(0) = \alpha \quad x(1) = \beta \end{cases}$$

no qual

$$g(p, q) = (b-a)^2 f(a + (b-a)p, q)$$

Se x é uma solução de **(ii)**, *então a função y definida por $y(t) = x((t-a)/(b-a))$ é a solução de* **(i)**. *Mais ainda, se y é uma solução de* **(i)**, *então $x(t) = y(a + (b-a)t)$ é a solução de* **(ii)**.

Ainda temos que para reduzir um PVC

$$\begin{cases} x'' = g(t, y) \\ y(0) = \alpha \quad y(1) = \beta \end{cases}$$

em um que possua condições de contorno homogêneas, simplesmente subtraímos de x uma função linear que leva os valores α e β em 0 e 1. O teorema nos diz que:

Teorema 47 *Considere os dois PVCs:*

$$\textbf{(ii)} \begin{cases} x'' = g(t, x) \\ x(0) = \alpha \quad x(1) = \beta \end{cases} \qquad \textbf{(iii)} \begin{cases} z'' = h(t, z) \\ z(0) = 0 \quad z(1) = 0 \end{cases}$$

no qual

$$h(p, q) = g(p, q + \alpha + (\beta - \alpha)p)$$

Se z é uma solução de **(iii)**, *então a função $x(t) = z(t) + \alpha + (\beta - \alpha)t$ é a solução de* **(ii)**. *Mais ainda, se x é uma solução de* **(ii)**, *então $z(t) = y(t) - [\alpha + (\beta - \alpha)t]$ é a solução de* **(iii)**.

Teorema 48 *Seja f uma função contínua de (t, s), onde $0 \leq t \leq 1$ e $-\infty < s < \infty$. Assuma que sobre este domínio*

$$|f(t, s_1) - f(t, s_2)| \leq k|s_1 - s_2| \quad (k < 8)$$

Então o PVC

$$\begin{cases} y'' = f(t, y) \\ y(0) = 0 \quad y(1) = 0 \end{cases}$$

possui solução única em $C[0, 1]$.

Exemplo

Mostre que o seguinte problema possui solução única

$$\begin{cases} y'' = 2\exp(t\cos y) \\ y(0) = 0 \quad y(1) = 0 \end{cases} \tag{4}$$

Para mostrar seja $f(t,s) = 2\exp(t\cos y)$. Pelo Teorema do valor médio temos que

$$|f(t,s_1) - f(t,s_2)| = \left|\frac{\partial f}{\partial s}\right| |s_1 - s_2|$$

onde

$$\left|\frac{\partial f}{\partial s}\right| = |2\exp(t\cos s)(-t\sin s)| \leq 2e < 5.437 < 8$$

Assim pelo teorema (48), o PVC descrito por (4) possui solução única.

9.2 Método de Diferenças Finitas Para Problemas Lineares

O método do disparo pode ser usado para resolver PVC lineares e não lineares mas, muitas vezes, ele apresenta problemas de instabilidade. Uma outra aproximação para PVC consiste de uma discretização inicial do intervalo t seguido pelo uso de fórmula de aproximação das derivadas. Os métodos de diferenças finitas substituem cada uma das derivadas que aparecem no PVC por um quociente em diferenças apropriado, escolhido de forma a obter o erro de truncamento desejado. No nosso caso estas duas fórmulas de diferenças centrais são especialmente úteis:

$$y'(t) = \frac{y(t+h) - y(t-h)}{2h} - \frac{1}{6}h^2 y'''(\xi) \tag{5}$$

$$y''(t) = \frac{y(t+h) - 2y(t) + y(t-h)}{2h} - \frac{1}{12}h^2 y^{(4)}(\tau) \tag{6}$$

EDO de segunda ordem

Vamos supor que o problema a ser resolvido é:

$$\begin{cases} y'' = f(t, y, y') \\ y(a) = \alpha \quad y(b) = \beta \end{cases} \tag{7}$$

Vamos particionar o intervalo $[a, b]$ nos pontos $a = t_0, t_1, t_2, ..., t_{n+1} = b$, os quais não necessitam ser igualmente espaçados, mas geralmente na prática são. Agora se os pontos não forem igualmente espaçados, versões mais complexas das fórmulas (5) e (6) serão necessárias. Por simplicidade, vamos assumir que:

$$t_i = a + ih \quad 0 \leq i \leq n+1 \quad h = (b-a)/(n+1) \tag{8}$$

Vamos notar um valor aproximado de $y(t_i)$ por y_i. Assim, a versão discreta de (7) é dada por:

$$\begin{cases} y_0 &= \alpha \\ h^{-2}(y_{i-1} - 2y_i + y_{i+1}) &= f(t_i, y_i, (2h)^{-1}(y_{i+1} - y_{i-1})) \qquad (1 \leq i \leq n) \\ y_{n+1} &= \beta \end{cases} \tag{9}$$

9.2.1 O caso Linear

Na equação (7) temos n incógnitas $y_1, y_2, ..., y_n$ e n equações que devem ser resolvidas. Se f envolve y_i de uma forma não linear, estas equações serão não lineares e, em geral, difíceis de serem resolvidas. Entretanto, vamos assumir que f é linear, isto é que possui a equação possui a forma:

$$\begin{aligned} y'' &= p(x)y' + q(x)y + r(x) \qquad a \leq x \leq b, \\ y(a) &= \alpha \qquad y(b) = \beta, \end{aligned} \tag{10}$$

Agora seguimos o seguinte:

- Escolhemos um inteiro N e dividimos o intervalo $[a, b]$ em $N+1$ subintervalos. Os pontos da malha dados por $x_i = a + ih$, para $i = 0, 1, ..., N + 1$ e $h = (b - a)/(N + 1)$. Por facilidade h será considerado constante.

- Em um ponto interior da malha x_i, $i = 1, ...N$, substituímos as aproximações em diferenças centrais dadas pelas fórmulas (5) e (6) para y' e y''

- Usando estas fórmulas em diferenças centrais na equação (10) resulta na seguinte equação:

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} = p(x_i)\left[\frac{y(x_{i+1}) - y(x_{i-1})}{2h}\right] + q(x_i)y(x_i) + r(x_i) - \frac{h^2}{12}\left[2p(x_i)y'''(\eta_i) - y^{(4)}(\xi_i)\right].$$

- Usando esta equação juntamente com as condições de contorno do problema $y(a) = \alpha$ e $y(b) = \beta$ obtemos um método de diferenças finitas com erro de truncamento de ordem $\mathcal{O}(h^2)$, definido por:

  $$w_0 = \alpha, \qquad w_{N+1} = \beta$$

  e

  $$\left(\frac{2w_i - w_{i+1} - w_{i-1}}{h^2}\right) + p(x_i)\left(\frac{w_{i+1} - w_{i-1}}{2h}\right) + q(x_i)w_i = -r(x_i) \tag{11}$$

  para cada $i = 1, 2, ..., N$.

- Agora rearranjando a equação (11) temos:

  $$-\left(1 + \frac{h}{2}p(x_i)\right)w_{i-1} + \left(2 + h^2q(x_i)\right)w_i - \left(1 - \frac{h}{2}p(x_i)\right)w_{i+1} = -h^2r(x_i)$$

  o que resulta em um sistema de N equações e N incógnitas,

  $$\mathbf{Aw} = \mathbf{b} \tag{12}$$

  onde,

$$\mathbf{A} = \begin{pmatrix} 2 + h^2q(x_1) & -1 + \frac{h}{2}p(x_1) & 0 & \dots & \dots & 0 \\ -1 - \frac{h}{2}p(x_2) & 2 + h^2q(x_2) & -1 + \frac{h}{2}p(x_2) & \ddots & & \vdots \\ 0 & & & \ddots & & \vdots \\ \vdots & \ddots & & & \ddots & 0 \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \dots & 0 & -1 - \frac{h}{2}p(x_N) & & 2 + h^2q(x_N) \end{pmatrix}$$

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N-1} \\ w_N \end{pmatrix} \quad \text{e} \quad \mathbf{b} = \begin{pmatrix} -h^2 r(x_1) + \left(1 + \frac{h}{2} p(x_1)\right) w_0 \\ -h^2 r(x_2) \\ \vdots \\ -h^2 r(x_{N-1}) \\ -h^2 r(x_N) + \left(1 - \frac{h}{2} p(x_N)\right) w_{N+1} \end{pmatrix}$$

O sistema acima é tridiagonal e pode ser resolvido por um algoritmo Gaussiano especial para estes casos. Note ainda que se h é pequeno e $q_i > 0$, a matriz do sistema será diagonal dominante se

$$|2 + h^2 q_i| > \left|1 + \frac{1}{2} h p_i\right| + \left|1 - \frac{1}{2} h p_i\right| = 2$$

Aqui nós assumimos que $|\frac{1}{2} h p_i| \leq 1$ porque então o termo $1 + \frac{1}{2} h p_i$ são ambos não negativos. Desta forma assumimos que $q_i > 0$ e h é pequeno o suficiente para que $|\frac{1}{2} h p_i| \leq 1$. O seguinte teorema nos apresenta condições sob as quais podemos garantir a existência de uma única solução do sistema (12).

Teorema 49 *Suponha que p, q e r são funções contínuas sobre o intervalo $[a, b]$. Se $q(x) > 0$ sobre $[a, b]$, então o sistema tridiagonal (12) possui uma única solução se $h < 2/L$, onde $L = \max_{a \leq x \leq b} |p(x)|$.*

Devemos ainda notar que a hipótese do teorema (49) garante a unicidade de solução de PVC (10), mas isto não garante que $y \in \mathcal{C}^{(4)}[a, b]$, o que é preciso para assegurar um erro de truncamento de $\mathcal{O}(h^2)$.

ALGORITMO 10

Aproxima a solução do PVC $y'' = p(x)y' + q(x)y + r(x)$, $y(a) = \alpha$, $y(b) = \beta$
ENTRADA: a, b, α, β, N.
SAÍDA: Aproximações w_i de $y(x_i)$ para cada $i = 0, 1, ..., N + 1$.

1. *Faça* $h = (b - a)/(N + 1)$

   $x = a + h$
   $a_1 = 2 + h^2 q(x)$
   $b_1 = -1 + (h/2)p(x)$
   $d_1 = -h^2 r(x) + (1 + (h/2)p(x))\alpha$

2. *Para* $= 2, \ldots, N - 1$ *faça*

   $x = a + ih$
   $a_i = 2 + h^2 q(x)$
   $b_i = -1 + (h/2)p(x)$
   $c_i = -1 - (h/2)p(x)$
   $d_i = -h^2 r(x)$

3. *Faça* $x = b - h$

$a_N = 2 + h^2 q(x)$
$c_N = -1 - (h/2)p(x)$
$d_N = -h^2 r(x) + (1 - (h/2)p(x))\beta$

4. *Faça* $l_1 = a_1$ *(4 à 10 resolve sistema tridiag.)*

   $u_1 = b_1/a_1$

5. *Para* $i = 2, ..., N$ *Faça*

   $l_i = a_i - c_i u_{i-1}$
   $u_i = b_i/l_i$

6. *Faça* $l_N = a_N - c_N u_{N-1}$
7. *Faça* $z_1 = d_1/l_1$
8. *Para* $i = 2, ..., N$ *Faça* $z_i = (d_i - c_i z_{i-1})/l_i$
9. *Faça*

   $w_0 = \alpha$
   $w_{N+1} = \beta$
   $w_N = z_N$

10. *Para* $i = N - 1, ..., 1$ *Faça* $w_i = z_i - u_i w_{i+1}$
11. *Para* $i = 0, ..., N + 1$ *Faça* $x = a + ih$

    SAÍDA: (x, w_i)

12. *PARE*

Exemplo 67 *O algoritmo (10) será usado para aproximar a solução do PVC*

$$y'' = -\frac{2}{x}y' + \frac{2}{x^2}y + \frac{\sin(\ln x)}{x^2}, \quad 1 \le x \le 2, \quad y(1) = 1, \ \ y(2) = 2,$$

neste exemplo iremos usar $N = 9$*, e assim* $h = (2 - 1)/(9 + 1) = 0.1$*. Os resultados estão listados na tabela (7). Aqui vamos comparar os resultados obtidos pelo método com a solução exata*

$$y(x) = c_1 x + \frac{c_2}{x^2} - \frac{3}{10}\sin(\ln x) - \frac{1}{10}\cos(\ln x),$$

onde $c_2 = \frac{1}{70}[8 - 12\sin(\ln 2) - 4\cos(\ln 2)] \approx -0.03920701320$ *e* $c_1 = \frac{11}{10} - c_2 \approx 1.1392070132$.

x_i	w_i	$y(x_i)$	$\lvert w_i - y(x_i)\rvert$
1.0	1.00000000	1.00000000	$--$
1.1	1.09260052	1.09262930	2.88×10^{-5}
1.2	1.18704313	1.18708484	4.17×10^{-5}
1.3	1.28333687	1.28338236	4.55×10^{-5}
1.4	1.38140205	1.38144595	4.39×10^{-5}
1.5	1.48112026	1.48115942	3.92×10^{-5}
1.6	1.58235990	1.58239246	3.26×10^{-5}
1.7	1.68498902	1.68501396	2.49×10^{-5}
1.8	1.78888175	1.78889853	1.68×10^{-5}
1.9	1.89392110	1.89392951	8.41×10^{-6}
2.0	2.00000000	2.00000000	$--$

Nota 6 *Este método é* $\mathcal{O}(h^2)$ *e por isto é menos preciso que o para resolução de PVI que é* $\mathcal{O}(h^4)$*. Para obter um método de diferenças com maior precisão, aparecem dificuldades, pois aproximações de maior ordem em diferenças finitas necessitam múltiplos valores antes e depois de* x_i*.*

Nota 7 *Pode-se usar extrapolação de Richardson para acelerar a convergência do método. Por exemplo para aproximar a solução do PVC do exemplo (66) devemos resolve-lo com* $h_1 = 0.1$*,* $h_2 = h_1/2 = 0.05$ *e* $h_3 = h_2/2 = 0.025$*. Aqui vamos notar* $w_{i,1}$*,* $w_{i,1}$ *ew*$_{i,1}$ *para as aproximações da solução* $y(x_i)$ *com* h_1*,* h_2 *e* h_3 *respectivamente. Assim usando as fórmulas, de extrapolação*

$$Ext_{1,i} = \frac{4w_{i,2} - w_{i,1}}{3}$$

$$Ext_{2,i} = \frac{4w_{i,3} - w_{i,2}}{3}$$

$$Ext_{3,i} = \frac{16w_{2,i} - w_{1,i}}{15}$$

obtemos os resultados descritos na tabela:

x_i	$w_{i,1}$	$w_{i,2})$	$w_{i,3}$	$Ext_{1,i}$	$Ext_{2,i}$	$Ext_{3,i}$
1.0	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
1.1	1.09260052	1.09262207	1.09262749	1.09262925	1.09262930	1.09262930
1.2	1.18704313	1.18707436	1.18708222	1.18708477	1.18708484	1.18708484
1.3	1.28333687	1.28337094	1.28337950	1.28338230	1.28338236	1.28338236
1.4	1.38140205	1.38143493	1.38144319	1.38144589	1.38144595	1.38144595
1.5	1.48112026	1.48114959	1.48115696	1.48115937	1.48115941	1.48115941
1.6	1.58235990	1.58238429	1.58239042	1.58239242	1.58239246	1.58239246
1.7	1.68498902	1.68500770	1.68501240	1.68501393	1.68501396	1.68501396
1.8	1.78888175	1.78889432	1.78889748	1.78889852	1.78889853	1.78889853
1.9	1.89392110	1.89392740	1.89392898	1.89392950	1.89392951	1.89392951
2.0	2.00000000	2.00000000	2.00000000	2.00000000	2.00000000	2.00000000

9.2.2 Métodos de diferenças finitas para problemas não lineares

Para o problema geral não linear de contorno

$$y'' = f(x, y, y'), \quad a \leq x \leq b, \quad y(a) = \alpha, \quad y(b) = \beta, \tag{13}$$

o método em diferença é similar ao do caso linear, porém o sistema resultante aqui é não linear e desta forma necessita de um processo iterativo para resolve-lo.

Para desenvolvermos este método vamos assumir que a função f satisfaz:

1. f e suas derivadas parciais em y e em y' são contínuas em

$$D = \{(x, y, y')/a \leq x \leq b, -\infty < y < \infty, -\infty < y' < \infty)$$

2. $f_y(x, y, y') \geq \delta > 0$ sobre D para algun $\delta > 0$

3. Existem constantes k e L ,com

$$k = \max_{(x,y,y')\in D} |f_y(x, y, y')|, \quad L = \max_{(x,y,y')\in D} |f'_y(x, y, y')|.$$

Estas três condições garantem a existência e unicidade de solução do problema (13).

Como no caso linear, dividimos o intervalo $[a, b]$ em $N + 1$ subintervalos iguais cujos pontos estão em $x_i = a + ih$, para $i = 0, 1, ..., N + 1$. Assumindo que a solução exata é limitada em sua quarta derivada nos permite substituir $y''(x_i)$ e $y'(x_i)$ em cada uma das equações

$$y''(x_i) = f(x_i, y(x_i), y'(x_i)) \tag{14}$$

pela fórmula apropriada em diferenças finitas dadas por (5) e (6). Assim obtemos para $i = 1, 2, ..., N$,

$$\frac{y(x_{i+1} - 2y(x_i) + y(x_{i-1})}{h^2} = f\left(x_i, y(x_i), \frac{y(x_{i+1}) - y(x_{i-1})}{2h} - \frac{h^2}{6}y'''(\eta_i)\right) + \frac{h^2}{12}y^{(4)}(\xi_i) \tag{15}$$

para algum ξ_i e η_i no intervalo (x_{i-1}, x_{i+1}).

Como no caso linear quando não consideramos os termos de erro e empregamos as condições de contorno,

$$\begin{gathered} w_0 = \alpha, \qquad w_{N+1} = \beta \\ -\frac{w_{i+1} - 2w_i + w_{i-1}}{h^2} = f\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right) = 0 \end{gathered} \tag{16}$$

para $i = 1, 2, ..., N$.

Desta forma obtemos um sistema não linear de N equações e N incógnitas, que possui solução única se $h < 2/L$ ver Keller[76]

$$
\begin{aligned}
2w_1 - w_2 + h^2 f\left(x_1, w_1, \frac{w_2 - \alpha}{2h}\right) - \alpha &= 0 \\
-w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3 - w_1}{2h}\right) &= 0 \\
\vdots \qquad\qquad\qquad\qquad\qquad\qquad & \vdots \\
-w_{N-1} + 2w_{N-1} - w_N + h^2 f\left(x_{N-1}, w_{N-1}, \frac{w_N - w_{N-2}}{2h}\right) &= 0 \\
-w_{N-1} + 2w_N + h^2 f\left(x_N, w_N, \frac{\beta - w_{N-1}}{2h}\right) - \beta &= 0
\end{aligned}
\tag{17}
$$

Para aproximar a solução deste sistema podemos usar o método de Newton para sistemas não lineares, conforme já disctimos. A seqüência de vetores $\mathbf{w}_k$ gerada pelas iterações converge para a solução única do sistema, se a aproximação inicial for suficientemente próxima da solução $\mathbf{w}$ e o Jacobiano da matriz seja não singular. No nosso caso o Jacobiano $\mathbf{J}(w_1, ..., w_N)$ é tridiagonal e definido por:

$$
\mathbf{J}(w_1, ..., w_N) = \begin{pmatrix}
A_{1,1} & A_{1,2} & 0 & ... & 0 \\
A_{2,1} & A_{2,2} & \ddots & & \vdots \\
0 & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & \ddots & 0 \\
 & & \ddots & & A_{N-1,N} \\
0 & \dots & 0 & A_{N,N-1} & A_{N,N}
\end{pmatrix}
$$

onde $w_0 = \alpha$, $w_{N+1} = \beta$

$$
A_{i,j} = 2 + h^2 f_y\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right), \quad \text{para } i = j, \;\; j = 1, ..., N
$$

$$
A_{i,j} = -1 + \frac{h}{2} f_{y'}\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right), \quad \text{para } i = j - 1, \;\; j = 2, ..., N
$$

$$
A_{i,j} = -1 - \frac{h}{2} f_{y'}\left(x_i, w_i, \frac{w_{i+1} - w_{i-1}}{2h}\right), \quad \text{para } i = j + 1, \;\; j = 1, ..., N - 1
$$

O método de Newton para sistemas não lineares precisa resolver em cada iteração o sistema linear

$$
\mathbf{J}(w_1, ..., w_N) \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix} = - \begin{pmatrix}
2w_1 - w_2 - \alpha + h^2 f\left(x_1, w_1, \frac{w_2 - \alpha}{2h}\right) \\
-w_1 + 2w_2 - w_3 + h^2 f\left(x_2, w_2, \frac{w_3 - w_3}{2h}\right) \\
\vdots \\
-w_{N-2} + 2w_{N-1} - w_N + h^2 f\left(x_{N-1}, w_{N-1}, \frac{w_N - w_{N-2}}{2h}\right) \\
-w_{N-1} + 2w_N + h^2 f\left(x_N, w_N, \frac{\beta - w_{N-1}}{2h}\right) - \beta
\end{pmatrix}
\tag{18}
$$

deve ser resolvido para o vetor **v**. e assim calcula-se a próxima iteração como:

$$w_i^{(k)} = w_i^{(k-1)} + v_i, \qquad \text{para cada } i = 1, 2, ..., N.$$

EXERCÍCIOS

1. Use Diferenças finitas para aproximar a solução dos seguintes PVC:

   (a) $y'' + y = 0, \quad 0 \leq 0 \leq \pi, y(0) = 1, \quad y(\pi) = -1;$. Use $h = \pi/2$.
   R: $x_1 = 1.04720, \;\; w_1 = 0.525382; \;\; x_2 = 2.09439, \;\; w_2 = -0.525382.$

   (b) $y'' + 4y = \cos x, \quad 0 \leq x \leq \pi/4, \quad y(0) = 0, \quad y(\pi/4) = 0;$. Use $h = \pi/12$.
   R: $x_1 = \frac{\pi}{12}, \;\; w_1 = -0.0877483; \;\; x_2 = \frac{\pi}{6}, \;\; w_2 = -0.0852364.$

2. Mostre que os seguintes PVC satisfazem a hipótese do teorema de existência e unicidade e aproxime sua solução.

   (a) $y'' = -3y' + 2y + 2x + 3, \;\; 0 \leq x \leq 1, \;\; y(0) = 2, \;\; y(1) = 1.$ Use $h = 0.1$ e compare os resultados com a solução exata.

   (b) $y'' = -(x+1)y' + 2y + (1 - x^2)e^{-x}, \;\; 0 \leq x \leq 1, \;\; y(0) = -1, \;\; = y(1) = 0.$ Use $h = 0.1$ e compare os resultados com a solução exata $y = (x-1)e^{-x}$.

3. Repita o exercício 2 usando extrapolação de Richardson.

4. Encontre por diferenças finitas a aproximação da solução de:

   (a) $y'' = -(y')^2 - y + \ln x, \;\; 1 \leq x \leq 2, \;\; y(1) = 0, \;\; y(2) = \ln 2.$ Use $h = 0.5$.
   R: $x_i = 1.5, \;\; w_i = 0.406800.$

   (b) $y'' = \frac{xy' - y}{x^2}, \;\; 1 \leq x \leq 1.6, \;\; y(1) = -1, \;\; y(1.6) = -0.847994.$ Use $h = 0.2$.
   R: $x = 1.2, \;\; w = -0.980942; \;\; x = 1.4, \;\; w = -0.928693.$