Notas de Aula de Cálculo Numérico

Bernardo Martins Rocha

6 de maio de 2019

Sumário

Capítulo 1

Polinômios de Taylor

Este capítulo apresenta uma das ferramentas matemáticas mais importantes dentro do cálculo numérico e análise numérica, que são os polinômios de Taylor (também conhecido como série de Taylor).

1.1 Introdução

Algumas funções matemáticas ditas *elementares* não são tão elementares assim quando tentamos avalia-las. Se p é uma função polinomial,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

então p pode ser avaliado facilmente para qualquer número x. Entretanto o mesmo não é verdadeiro para funções como e^x, $\sin(x)$, $\cos(x)$, $\log(x)$. Tente calcular essas funções sem usar a calculadora para qualquer x.

Estamos interessados em reduzir a avaliação de funções $f(x)$ por funções que sejam mais fáceis de se avaliar. Polinômios são funções fáceis de se avaliar, pois precisamos apenas de realizar operações de adição e multiplicação.

Sendo assim, deseja-se aproximar a função $f(x)$ por uma função polinomial $\hat{f}(x)$ que seja fácil de avaliar. Uma das aproximações polinomiais mais usadas são os polinômios de Taylor. Vamos estudar agora como encontrar estas funções polinomiais que aproximam $f(x)$.

A fim de encontrar um polinômio que aproxima uma função, vamos antes analisar algumas propriedades de polinômios. Considere o polinômio

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n. \tag{1.1}$$

É interessante observar que os coeficientes a_i, $i = 0, \ldots, n$, podem ser escritos em termos de valores de p e de suas várias derivadas (p', p'', ...) em $x = 0$. Para começar observe que

$$p(0) = a_0$$

Derivando $p(x)$ temos

$$p'(x) = a_1 + 2a_2 x + \ldots + n a_n x^{n-1}$$

que avaliado em $x = 0$ resulta em

$$p'(0) = a_1$$

Derivando novamente

$$p''(x) = 2a_2 + 6a_3x + \ldots + n(n-1)a_nx^{n-2}$$

e portanto

$$p''(0) = 2a_2$$

Denotando a k-ésima derivada de $p(x)$ por $p^{(k)}(x)$, de forma geral teremos a seguinte relação

$$p^{(k)}(0) = k!\, a_k$$

Lembrando que $0! = 1$ e que $p^{(0)} = p$, temos

$$a_k = \frac{p^{(k)}(0)}{k!},\ 0 \leq k \leq n$$

Se tivéssemos começado com uma função p como um polinômio em $(x-a)$,

$$p(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \ldots + a_n(x-a)^n$$

procedendo da mesma forma como anteriormente, mas agora avaliando o polinômio e suas derivadas em $x = a$, temos

$$\begin{aligned}
p(x) &= a_0 + a_1(x-a) + a_2(x-a)^2 + \ldots + a_n(x-a)^n \\
p(a) &= a_0 \\
p'(x) &= a_1 + 2a_2(x-a) + \ldots + na_n(x-a)^{n-1} \\
p'(a) &= a_1 \\
p''(x) &= 2a_2 + 6a_3(x-a) + \ldots + n(n-1)a_n(x-a)^{n-2} \\
p''(a) &= 2a_2 \\
&\ldots
\end{aligned}$$

de forma geral, temos

$$a_k = \frac{p^{(k)}(a)}{k!}$$

1.2 Polinômio de Taylor

Suponha agora que $f(x)$ seja uma função (não necessariamente um polinômio) tal que $f^{(1)}(a)$, $f^{(2)}(a)$, ..., $f^{(n)}(a)$, existam. Seja

$$\boxed{a_k = \frac{f^{(k)}(a)}{k!},\ 0 \leq k \leq n} \tag{1.2}$$

então o **polinômio de Taylor de grau n para $f(x)$ em torno de a** é definido como

$$\boxed{P_{n,a}(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \ldots + a_n(x-a)^n.} \tag{1.3}$$

Daqui em diante vamos simplificar a notação e escrever apenas $P_n(x)$, deixando o ponto a claro a partir do contexto.

O polinômio de Taylor foi definido tal que

$$P_n^{(k)}(a) = f^{(k)}(a), \quad \text{para } 0 \leq k \leq n$$

observe

$$\begin{aligned}
P_n^{(0)}(x) &= a_0 + a_1(x-a) + \ldots + a_n(x-a)^n \\
P_n^{(1)}(x) &= a_1 + 2a_2(x-a) + \ldots + na_n(x-a)^{n-1} \\
P_n^{(2)}(x) &= 2a_2 + 6a_3(x-a) + \ldots + n(n-1)a_n(x-a)^{n-2} \\
P_n^{(3)}(x) &= 6a_3 + 24a_4(x-a) + \ldots + n(n-1)(n-2)a_n(x-a)^{n-3} \\
&\ldots \\
P_n^{(n)}(x) &= n!\, a_n
\end{aligned}$$

Lembrando que

$$a_k = \frac{f^{(k)}(a)}{k!}$$

substituindo os coeficientes a_k e avaliando as expressões anteriores em $x = a$ temos

$$\begin{aligned}
P_n^{(0)}(a) &= a_0 = f^{(0)}(a) \\
P_n^{(1)}(a) &= a_1 = f^{(1)}(a) \\
P_n^{(2)}(a) &= 2a_2 = 2\tfrac{f^{(2)}(a)}{2!} = f^{(2)}(a) \\
P_n^{(3)}(a) &= 6a_3 = 6\tfrac{f^{(3)}(a)}{3!} = f^{(3)}(a) \\
&\ldots \\
P_n^{(n)}(a) &= n!\, a_n = n!\, \tfrac{f^{(n)}(a)}{n!} = f^{(n)}(a)
\end{aligned}$$

E assim confirmamos que

$$P_n^{(k)}(a) = f^{(k)}(a), \quad \text{para } 0 \leq k \leq n$$

Usando a relação

$$a_k = \frac{f^{(k)}(a)}{k!}$$

vamos escrever o polinômio de Taylor de grau n da seguinte forma

$$\boxed{P_n(x) = f(a) + f'(a)(x-a) + f''(a)\tfrac{(x-a)^2}{2} + \ldots + f^{(n)}(a)\tfrac{(x-a)^n}{n!}}$$

isto é, conhecendo-se o valor de $f(x)$ e suas derivadas em a pode-se aproximar a função $f(x)$ pelo polinômio de Taylor de grau n denotado por $P_n(x)$.

Exemplo 1: Exponencial

Encontrar o polinômio de Taylor de grau 1 (linear) que aproxima a função $f(x) = e^x$ em torno do ponto 0.
Solução: Temos

$$f(x) = e^x \quad \Rightarrow \quad f'(x) = e^x$$

portanto o polinômio de Taylor linear é dado por

$$\begin{aligned} P_1(x) &= f(a) + f'(a)(x-a) = f(0) + f'(0)(x-0) \\ &= e^0 + e^0(x-0) = 1 + x \end{aligned}$$

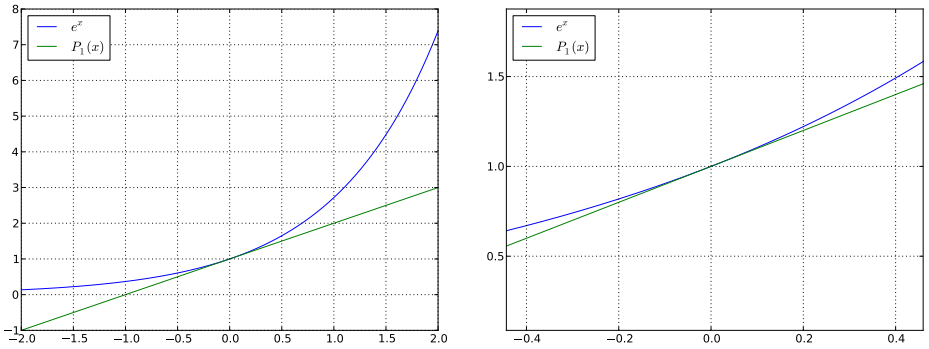


Figura 1.1: Aproximação da função exponencial.

Observação 1. Interpretação Geométrica do Exemplo 1 Equação da reta

$$y - y_0 = m(x - x_0)$$

Como o coeficiente angular da reta tangente ao gráfico de $y = f(x)$ no ponto $(x_0, f(x_0))$ é $f'(x_0) = m$, temos a seguinte eq. para a reta tangente

$$y - f(x_0) = f'(x_0)(x - x_0)$$

Comparando com nossa aproximação

$$P_1(x) - f(a) = f'(a)(x - a)$$

vemos que neste exemplo a função aproximadora é a reta tangente a curva $f(x)$ no ponto $x = a$.

Exemplo 2: Polinômio quadrático

Determinar o polinômio de Taylor de grau 2 (quadrático) para $f(x) = e^x$ em torno do ponto $a = 0$.
Solução: Lembrando que

$$f(x) = e^x \quad \Rightarrow \quad f'(x) = e^x \quad \Rightarrow \quad f''(x) = e^x$$

então

$$\begin{aligned} P_2(x) &= f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2} \\ &= f(0) + f'(0)(x-0) + f''(0)\frac{(x-0)^2}{2} \\ &= e^0 + e^0(x-0) + e^0\frac{(x-0)^2}{2} \\ &= 1 + x + \frac{x^2}{2} \end{aligned}$$

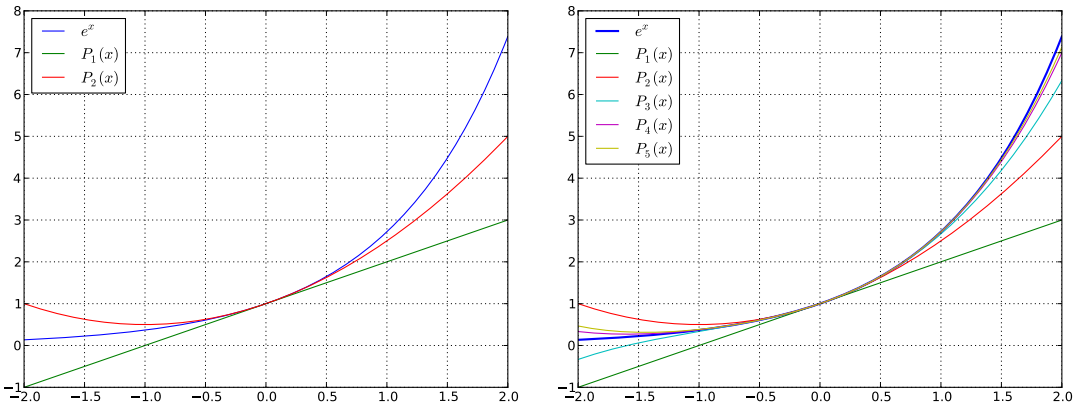


Figura 1.2: Gráficos da aproximação de $exp(x)$.

Exemplo 3: Fórmula geral para cosseno

Encontre a fórmula geral da aproximação usando polinômio de Taylor para a função $f(x) = \cos(x)$ em torno do ponto $a = 0$. Note que $\cos(x) = \cos(-x)$ é uma função par e, portanto, considere um polinômio de grau $2n$ para $n = 0, 1, 2, \ldots$.

Exemplo 4: Fórmula geral para seno

Encontre a fórmula geral da aproximação usando polinômio de Taylor para a função $f(x) = \sin(x)$ em torno do ponto $a = 0$.
Solução Note que

$$\begin{aligned} f(x) &= \sin(x) \quad &\Rightarrow \quad f(a) &= 0 \\ f'(x) &= \cos(x) \quad &\Rightarrow \quad f'(a) &= 1 \\ f''(x) &= -\sin(x) \quad &\Rightarrow \quad f''(a) &= 0 \\ f'''(x) &= -\cos(x) \quad &\Rightarrow \quad f'''(a) &= -1 \\ f^{(4)}(x) &= \sin(x) \quad &\Rightarrow \quad f^{(4)}(a) &= 0 \end{aligned}$$

A partir desse ponto as derivadas repetem em ciclo de 4. Os coeficientes do polinômio

de Taylor

$$a_k = \frac{\sin^{(k)}(0)}{k!}$$

para $k = 0, 1, 2, \ldots$ são dados por

$$0, 1, 0, -\frac{1}{3!}, 0, \frac{1}{5!}, 0, -\frac{1}{7!}, 0, \frac{1}{9!}, \ldots$$

Como seno é uma função ímpar, o polinômio de Taylor de grau $2n+1$ para $f(x) = \sin(x)$ em $a = 0$ é dado por

$$P_{2n+1}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$

Obs: note que $P_{2n+1} = P_{2n+2}$. □

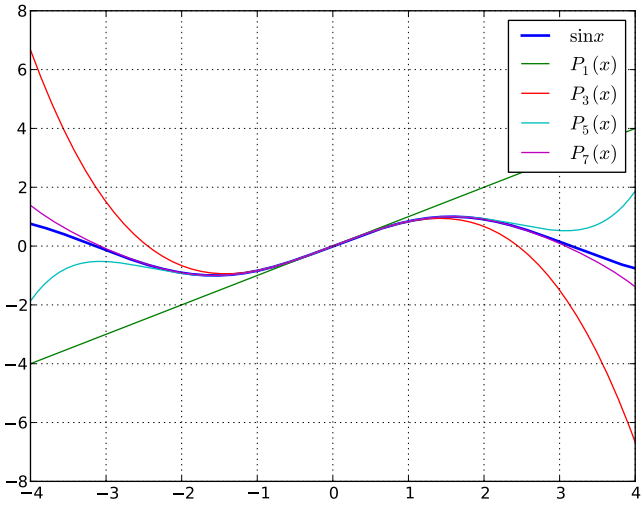


Figura 1.3: Aproximações com polinômios de Taylor para $\sin(x)$.

Exemplo 5

Exemplo 4 Encontre o valor de $f(6)$ sabendo que $f(4) = 125$, $f'(4) = 74$, $f''(4) = 30$, $f'''(4) = 6$, e que todas as outras derivadas de ordem alta são nulas.
Solução: Vamos usar uma aproximação por polinômio de Taylor de grau 3

$$P_3(x) = f(a) + f'(a)(x-a) + f''(a)\tfrac{(x-a)^2}{2!} + f'''(a)\tfrac{(x-a)^3}{3!}$$

Como temos os valores da função e suas derivadas em $x = 4$ usaremos este ponto para

aproximar $f(6)$, portanto

$$\begin{aligned} f(6) \approx P_3(6) &= f(4) + f'(4)(6-4) + f''(4)\tfrac{(6-4)^2}{2!} + f'''(4)\tfrac{(6-4)^3}{3!} \\ &= 125 + 74 \cdot 2 + 30 \cdot \tfrac{4}{2} + 6 \cdot \tfrac{8}{6} \\ &= 125 + 148 + 60 + 8 \\ &= 341 \end{aligned}$$

Exemplo 6: Raiz quadrada

Como calcular o valor de $\sqrt{13}$ numa ilha deserta, sem usar calculadora?
Solução: Podemos aproximar a função $f(x) = \sqrt{x}$ perto de um ponto a usando polinômios de Taylor. Neste caso vamos usar um polinômio linear e vamos escolher o ponto $a = 9$ (poderia ser $a = 16$). Temos

$$f(x) = \sqrt{x} = x^{1/2} \quad \Rightarrow \quad f'(x) = \frac{1}{2\sqrt{x}}$$

logo

$$P_1(x) = f(a) + f'(a)(x-a) = \sqrt{a} + \frac{1}{2\sqrt{a}}(x-a)$$

Substituindo $a = 9$ em $P_1(x)$ temos

$$P_1(x) = \sqrt{9} + \tfrac{1}{2\sqrt{9}}(x-9)$$

Sendo assim, avaliando em $x = 13$ para obter o valor de $\sqrt{13}$ obtemos

$$P_1(13) = \sqrt{9} + \frac{1}{2\sqrt{9}}(13-9) = 3 + \frac{4}{6} = 3.6666$$

O valor exato de $\sqrt{13}$ é 3.6055.
□

Exemplo 7: Raiz

Calcular o valor de $\sqrt[7]{1.1}$ ($R : 1.013708856$).
Solução: A função que queremos avaliar é $f(x) = \sqrt[7]{x}$. Vamos usar um polinômio de Taylor linear em torno de $a = 1$. Derivando

$$f(x) = x^{1/7} \quad \Rightarrow \quad f'(x) = \frac{1}{7\sqrt[7]{x^6}}$$

Assim temos a seguinte aproximação

$$\sqrt[7]{x} \approx f(a) + f'(a)(x-a) = \sqrt[7]{1} + \frac{1}{7\sqrt[7]{1^6}}(x-1)$$

e para $x = 1.1$ temos

$$\sqrt[7]{1.1} \approx 1 + \frac{1.1-1}{7} = 1.01428$$

O exemplo a seguir ilustra a importância da escolha do ponto $x = a$ em torno do qual se cria a aproximação com o polinômio de Taylor. Em alguns casos é preciso tomar cuidado com a escolha desse ponto para evitar problemas.

Exemplo 8: Logaritmo

Encontre uma aproximação para $\log(x)$.
Solução: o polinômio de Taylor para $\log(x)$ tem que ser calculado em algum ponto $a \neq 0$ já que a função não está definida neste ponto. Vamos usar então $a = 1$ para começar. Calculando as derivadas temos

$$\begin{aligned}
f'(x) &= \frac{1}{x} \quad \Rightarrow \quad f'(1) = 1 \\
f''(x) &= -\frac{1}{x^2} \quad \Rightarrow \quad f''(1) = -1 \\
f'''(x) &= \frac{2}{x^3} \quad \Rightarrow \quad f'''(1) = 2 \\
f^{(4)}(x) &= -\frac{6}{x^4} \quad \Rightarrow \quad f^{(4)}(1) = -6
\end{aligned}$$

De forma geral, para $k = 1, \ldots, n$, temos

$$f^{(k)}(x) = \frac{(-1)^{k-1}(k-1)!}{x^k} \quad \Rightarrow \quad f^{(k)}(1) = (-1)^{k-1}(k-1)!$$

Portanto temos

$$P_n(x) = f(1) + (x-1)f'(1) + \frac{(x-1)^2}{2}f''(1) + \ldots + \frac{(x-1)^n}{n!}f^{(n)}(1) \tag{1.4}$$

$$= 0 + (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{6}2 + \ldots + \frac{(x-1)^n}{n!}(-1)^{n-1}(n-1)! \tag{1.5}$$

assim

$$P_n(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} + \ldots + (-1)^{n-1}\frac{(x-1)^n}{n}$$

$$\begin{array}{cccccc} & a_3 & a_2 & a_1 & a_0 \\ + & & x \cdot b_3 & x \cdot b_2 & x \cdot b_1 \\ \hline & b_3 & b_2 & b_1 & b_0 \end{array}$$

É mais simples considerar $f(x) = \log(1+x)$ e criar o polinômio de Taylor em torno do ponto $a = 0$. Neste caso teríamos

$$P_n(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1}\frac{x^n}{n}$$

□

Através dos exemplo anteriores, foi possível observar algumas das propriedades da aproximação por polinômio de Taylor:

- quanto maior o grau do polinômio, melhor a aproximação;
- a medida que nos afastamos do ponto $x = a$, a aproximação piora;
- o polinômio de Taylor $P_n(x)$ só precisa do **valor da função e de suas derivadas** em um ponto a. Não é preciso conhecer a expressão analítica de suas derivadas.

1.3 Algoritmo de Horner

Uma tarefa computacional extremamente importante é a avaliação de polinômios, isto é: dado um ponto x qualquer, calcular o valor de $p(x)$. A forma mais direta de fazer isto é:

$$p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \tag{1.6}$$

Essa expressão pode ser reformulada como

$$p(x) = ((a_3 x + a_2)x + a_1)x + a_0 \tag{1.7}$$

Temos então o seguinte número de operações aritméticas:

- para n = 4, Eq. (1.6) faz 10 multiplicações e 4 adições
- para n = 4, Eq. (1.7) faz 4 multiplicações e 4 adições
- para n = 20, Eq. (1.6) faz 210 multiplicações e 20 adições
- para n = 20, Eq. (1.7) faz 20 multiplicações e 20 adições

Para

$$p(x) = ((a_3 x + a_2)x + a_1)x + a_0$$

podemos usar o seguinte esquema prático portanto $p(x) = b_0$. Para avaliar um polinômio de grau n

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

podemos usar a seguinte fórmula

$$\begin{aligned} b_n &= a_n \\ b_i &= a_i + x \cdot b_{i+1}, \qquad \text{para } i = n-1, \ldots, 2, 1, 0 \end{aligned}$$

Se os valores intermediários de b_i não são de interesse, podemos usar apenas uma variável b para implementar o algoritmo.

---

entrada: a_0, a_1, ..., a_n, x
saída: $p(x)$
$b = a_n$;
para $i = n - 1$ *até* 0 **faça**
  $b = a_i + x \cdot b$;
fim-para
retorne b;

1.4 Erro da aproximação

Precisamos saber qual o erro cometido ao aproximar a função $f(x)$ pelo polinômio de Taylor $P_n(x)$. Vamos representar o erro por $R_n(x)$. Se $f(x)$ é uma função para a qual $P_n(x)$ existe, definimos o erro (ou resto) $R_n(x)$ por

$$\begin{aligned} f(x) &= P_n(x) + R_n(x) \\ &= f(a) + f'(a)(x-a) + \ldots + f^{(n)}(a)\frac{(x-a)^n}{n!} + R_n(x) \end{aligned}$$

Gostaríamos de ter uma expressão para $R_n(x)$ cujo tamanho seja fácil de se estimar. Através do Teorema de Taylor iremos encontrar algumas expressões para o erro $R_n(x)$.

Teorema 1 (Teorema de Taylor)**.** Suponha que as derivadas $f^{(1)}$, ..., $f^{(n+1)}$ estejam definidas e sejam contínuas em um intervalo $[a, x]$, então temos que

$$\boxed{R_n(x) = f(x) - P_n(x) = \int_a^x f^{(n+1)}(t)\frac{(x-t)^n}{n!}\, dt}$$

onde

$$\boxed{P_n(x) = f(a) + f'(a)(x-a) + \ldots + f^{(n)}(a)\frac{(x-a)^n}{n!}}$$

Para a demonstração do teorema iremos utilizar os seguintes resultados: Integração por partes:

$$\boxed{\int_a^b u\, dv = uv\Big|_a^b - \int_a^b v\, du}$$

Teorema Fundamental do Cálculo (TFC), que diz que se $f(x)$ é definida em $[a, b]$ que admite uma anti-derivada $g(x)$ em $[a, b]$, isto é $f(x) = g'(x)$, então

$$\boxed{\int_a^b f(x)\ dx = g(b) - g(a)}$$

Prova 1. Para encontrar a expressão do erro na forma integral, vamos começar com o caso $n = 0$:

$$f(x) = f(a) + R_0(x)$$

pelo Teorema Fundamental do Cálculo, podemos escrever

$$R_0(x) = f(x) - f(a) = \int_a^x f'(t)\ dt$$

portanto

$$f(x) = f(a) + \int_a^x f'(t)\ dt$$

Vamos usar integração por partes no termo da integral

$$\begin{array}{rcl} u = f'(t) & \Rightarrow & du = f''(t)\ dt \\ dv = 1\ dt & \Leftarrow & v\ \ = t - x \end{array}$$

assim temos

$$\begin{aligned} \int_a^x f'(t)\ dt &= f'(t)(t-x)\Big|_a^x - \int_a^x f''(t)(t-x)\ dt \\ &= [f'(x)(x-x) - f'(a)(a-x)] - \int_a^x f''(t)(t-x)\ dt \\ &= -f'(a)(a-x) - \int_a^x f''(t)(t-x)\ dt \\ &= f'(a)(x-a) + \int_a^x f''(t)(x-t)\ dt \end{aligned}$$

Logo

$$f(x) = \underbrace{f(a) + f'(a)(x-a)}_{P_1(x)} + \int_a^x f''(t)(x-t)\ dt$$

Portanto

$$R_1(x) = \int_a^x f''(t)(x-t)\ dt$$

Para encontrar $R_2(x)$, usamos integração por partes novamente no termo com a integral. Para isso escolhemos

$$\begin{aligned} u = f''(t) \quad &\Rightarrow \quad du = f'''(t)\ dt \\ dv = (x-t)\ dt \quad &\Leftarrow \quad v \ = -\frac{(x-t)^2}{2} \end{aligned}$$

Assim

$$\begin{aligned} \int_a^x f''(t)(x-t)\ dt &= -f''(t)\frac{(x-t)^2}{2}\bigg|_a^x + \int_a^x f'''(t)\tfrac{(x-t)^2}{2}\ dt \\ &= f''(a)\frac{(x-a)^2}{2} + \int_a^x f'''(t)\frac{(x-t)^2}{2}\ dt \end{aligned}$$

Desta forma

$$f(x) = \underbrace{f(a) + f'(a)(x-a) + f''(a)\tfrac{(x-a)^2}{2}}_{P_2(x)} + \underbrace{\int_a^x f'''(t)\frac{(x-t)^2}{2}\ dt}_{R_2(x)}$$

ou seja

$$R_2(x) = \int_a^x f'''(t)\frac{(x-t)^2}{2}\ dt$$

Considerando que $f^{(n+1)}$ é contínua em $[a, x]$ por hipótese do teorema, podemos mostrar por indução que

$$\boxed{R_n(x) = \int_a^x f^{(n+1)}(t)\frac{(x-t)^n}{n!}\ dt} \tag{1.8}$$

□

É possível ainda obter as seguintes expressões para o erro: Forma de Cauchy

$$\boxed{R_n(x) = f^{(n+1)}(t)\frac{(x-t)^n}{n!}(x-a), \quad t \in (a,x)} \tag{1.9}$$

Forma de Lagrange

$$\boxed{R_n(x) = f^{(n+1)}(t)\frac{(x-a)^{(n+1)}}{(n+1)!}, \quad t \in (a,x)} \tag{1.10}$$

Essas expressões são muito úteis para se obter estimativas para o erro de uma aproximação usando o polinômio de Taylor. A forma do erro de Lagrange

$$R_n(x) = f^{(n+1)}(t)\frac{(x-a)^{(n+1)}}{(n+1)!} \tag{1.11}$$

é muito parecida com o próximo termo do polinômio de Taylor. A única diferença é o valor t na fórmula. t é **algum** valor entre a e x, que **não conhecemos**. Obs: t é um valor que no desenvolvimento da forma do erro de Lagrange surge da aplicação do Teorema do Valor Médio. Para **estimar** o erro, precisamos analisar os valores de $f^{(n+1)}(t)$ para todo $a < t < x$ e usar o maior deles. Ou, usar algum outro valor que com certeza é maior do que todos eles.

Exemplo 9: Erro para exponencial

Obtenha o limitante superior do erro para $e^{0.5}$ quando esta expressão é aproximada por um polinômio de Taylor de grau 4 para e^x em torno do ponto 0.
Solução: Pela fórmula de Lagrange do erro temos

$$R_4(x) = f^{(n+1)}(t)\frac{(x-0)^5}{5!} = e^t\frac{x^5}{120}, \quad \text{para algum } t \in [0, 0.5]$$

assim quando aproximamos $e^{0.5}$ o erro está limitado por

$$|R_4(x)| \leq \max\left|\frac{e^t x^5}{120}\right| \leq \left|\frac{e^{0.5}0.5^5}{120}\right| \leq 2\frac{0.5^5}{120} = 0.00052$$

Neste caso a aproximação de Taylor é

$$P_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$$

e portanto

$$e^{0.5} \approx 1 + 0.5 + \frac{0.5^2}{2} + \frac{0.5^3}{6} + \frac{0.5^4}{24} = 1.6484$$

O valor real é $e^{0.5} = 1.6487$, e vemos novamente que o erro é menor do que o estimado, pois $|1.6487 - 1.6484| = 0.0003$.

Exemplo 10: Erro para função seno

Seja $f(x) = \sin(x)$. Encontre o polinômio de Taylor cúbico em torno do ponto $a = 0$, em seguida encontre um limitante superior para este no ponto $x = \frac{\pi}{4}$ e calcule o erro.
Solução: Para $f(x) = \sin(x)$ com $a = 0$ já vimos que o polinômio cúbico de Taylor é

$$P_3(x) = x - \frac{x^3}{6}$$

Pela fórmula do erro de Lagrange, sabemos que

$$R_3(x) = f^{(4)}(t)\frac{(x-a)^4}{4!} = \sin(t)\frac{x^4}{24}$$

Portanto para o limitante superior temos

$$|R_3(x)| \leq \max \left| \frac{\sin(t)x^4}{24} \right|, \qquad \text{para } t \in [0, \pi/4]$$
$$\leq \left| \frac{\sin(\frac{\pi}{4})(\frac{\pi}{4})^4}{24} \right| \leq 0.0112$$

Avaliando $P_3(x)$ em $\frac{\pi}{4}$ temos

$$P_3(\tfrac{\pi}{4}) = \tfrac{\pi}{4} - \frac{\frac{\pi}{4}^3}{6} = 0.7046$$

O valor real é $\sin(\frac{\pi}{4}) = \frac{\sqrt{2}}{2} = 0.7071$, logo o erro cometido é $|0.7071 - 0.7046| = 0.0024$.

Exemplo 11: Grau do polinômio

Seja $f(x) = \sin(x)$ e $a = 0$. Determine n para que o erro ao se aproximar $f(x)$ por um polinômio de Taylor seja menor do que 10^{-7} para $-\frac{\pi}{4} \leq x \leq \frac{\pi}{4}$.
Solução: Pela fórmula do erro temos que

$$R_{2n+1}(x) = \sin(x) - P_{2n+1}(x) = f^{(2n+2)}(t)\frac{(x-0)^{2n+2}}{(2n+2)!}$$
$$= (-1)^{n+1}\sin(t)\frac{x^{2n+2}}{(2n+2)!}$$

Queremos saber qual o valor de n garante que

$$|R_{2n+1}(x)| \leq 10^{-7}, \quad \text{para } x \in [-\pi/4, \pi/4]$$

Então

$$|R_{2n+1}(x)| = \left| (-1)^{n+1}\sin(t)\frac{x^{2n+2}}{(2n+2)!} \right| < \frac{(\frac{\pi}{4})^{2n+2}}{(2n+2)!} < 10^{-7}$$

analisando temos

$$n = 1 \Rightarrow (2+2)! = 24 \qquad \Rightarrow \frac{(\frac{\pi}{4})^4}{4!} \approx 0.15 \times 10^{-1}$$
$$n = 2 \Rightarrow (4+2)! = 720 \qquad \Rightarrow \frac{(\frac{\pi}{4})^6}{6!} \approx 0.326 \times 10^{-3}$$
$$n = 3 \Rightarrow (6+2)! = 40320 \qquad \Rightarrow \frac{(\frac{\pi}{4})^8}{8!} \approx 3.590860 \times 10^{-6}$$
$$n = 4 \Rightarrow (8+2)! = 3628800 \qquad \Rightarrow \frac{(\frac{\pi}{4})^{10}}{10!} \approx 2.461137 \times 10^{-8}$$

ou seja, para que

$$\frac{(\frac{\pi}{4})^{2n+2}}{(2n+2)!} < 10^{-7}$$

temos que $n \geq 4$. Portanto, precisamos usar um polinômio de Taylor de grau maior ou igual a 9 para atingir a precisão desejada.

1.5 Aproximação de Derivada

Uma das principais aplicações do polinômio de Taylor é para a aproximação da derivada de uma função. Para isso, considere que uma função $f(x)$, cuja **expressão é desconhecida**, seja fornecida por meio de um conjunto de pontos $(x_0, f(x_0))$, $(x_1, f(x_1))$, ..., $(x_n, f(x_n))$. Como calcular $f'(x_i)$? Podemos usar polinômio de Taylor para aproximar as derivadas da função.

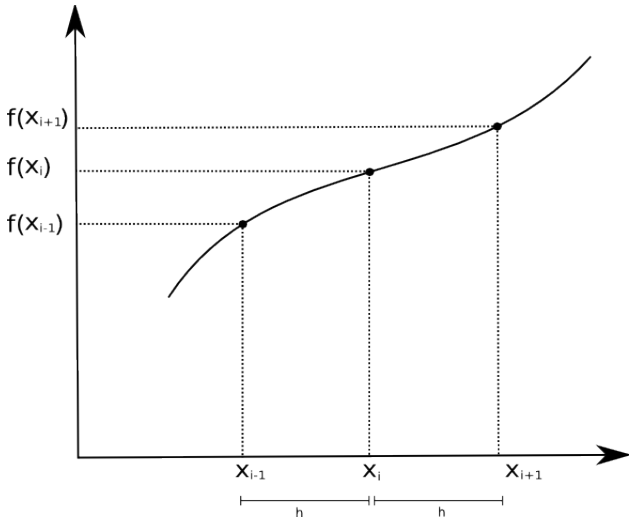


Figura 1.4: Taylor para aproximação da derivada de uma função.

Para calcular a derivada $f'(x_i)$ em cada ponto x_i, vamos usar um polinômio de Taylor linear em torno do ponto x_i. Basicamente, existem 3 formas de aproximar a primeira derivada de uma função utilizando Taylor:

- Diferença Progressiva: $x = x_{i+1}$

$$f(x_{i+1}) = f(x_i) + f'(x_i)\overbrace{(x_{i+1} - x_i)}^{h}$$

$$\boxed{f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h}}$$

- Diferença Regressiva: $x = x_{i-1}$

$$f(x_{i-1}) = f(x_i) + f'(x_i)\overbrace{(x_{i-1} - x_i)}^{-h}$$

$$\boxed{f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h}}$$

- Diferença Central: $x = x_{i+1}$ e $x = x_{i-1}$

$$f(x_{i+1}) = f(x_i) + f'(x_i)h$$
$$f(x_{i-1}) = f(x_i) - f'(x_i)h$$

subtraindo, temos

$$f(x_{i+1}) - f(x_{i-1}) = 2hf'(x_i)$$

que resulta em

$$\boxed{f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h}}$$

Observações:
- A diferença central é mais precisa para aproximar a derivada.
- As derivadas de alta ordem são calculadas de forma similar.
- Quanto mais pontos em um intervalo $[a, b]$, ou seja, quanto menor o espaçamento h entre eles, melhor a qualidade da aproximação.

Exemplo 12: Aprox. derivada

Calcule $f'(1.3)$ para $f(x) = \log(x)$ usando diferença progressiva e central para $h = 0.01$ e $h = 0.001$.
Solução: Usando $h = 0.01$, com diferença progressiva temos

$$f'(1.3) \approx \frac{\log(1.31) - \log(1.30)}{0.01} = 0.76628$$

Com diferença central temos

$$f'(1.3) \approx \frac{\log(1.31) - \log(1.29)}{2 \cdot 0.01} = 0.76924$$

Usando $h = 0.001$, com diferença progressiva temos

$$f'(1.3) \approx \frac{\log(1.301) - \log(1.300)}{0.001} = 0.76893$$

com diferença central temos

$$f'(1.3) \approx \frac{\log(1.301) - \log(1.299)}{2 \cdot 0.001} = 0.76923$$

Podemos calcular o valor real usando a derivada de $f(x)$, pois neste caso conhecemos a expressão da função. O resultado é

$$f'(x) = \frac{1}{x} \quad \Rightarrow \quad f'(1.3) = 0.76923$$

□

1.6 Exercícios

1. Verifique que o polinômio de Taylor de grau n para $f(x) = e^x$ em $a = 0$ é dado por

$$P_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{(n)!}$$

2. Verifique que o polinômio de Taylor de grau $2n$ para $f(x) = \cos(x)$ em $a = 0$ é dado por

$$P_{2n}(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots + (-1)^n \frac{x^{2n}}{(2n)!}$$

Capítulo 2

Noções de erro e sistemas de ponto flutuante

Como estamos estudando métodos numéricos para resolver problemas de ciências e engenharias, é de extrema importância entender como os números são representados no computador e como as operações aritméticas são feitas. Vamos então estudar como representar números no computador e os erros cometidos por conta da representação e durante as operações aritméticas.

Vamos começar revisando sistemas de numeração, conversão de bases, e depois estudamos como são representados os números inteiros e reais no computador. Em seguida vamos estudar as operações aritméticas em ponto flutuante e alguns efeitos numéricos como propagação do erro, cancelamento, etc.

2.1 Introdução

No dia a dia usamos números baseados no sistema decimal. O número 257, por exemplo, pode ser escrito como

$$\begin{aligned} 257 &= 2 \cdot 100 + 5 \cdot 10 + 7 \cdot 1 \\ &= 2 \cdot 10^2 + 5 \cdot 10^1 + 7 \cdot 10^0 \end{aligned}$$

Este sistema é conhecido de sistema decimal ou de sistema com base 10. Qualquer número inteiro pode ser expresso por um polinômio na base 10 com coeficientes entre 0 e 9. Usaremos a notação

$$\begin{aligned} N &= (a_n a_{n-1} \ldots a_0)_{10} \\ &= a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_0 \cdot 10^0 \end{aligned}$$

para representar qualquer inteiro no sistema decimal, onde os coeficientes são tais que $0 \leq a_i \leq 9$.

No passado outras civilizações usaram sistemas de numeração com bases diferentes como 12, 20 e 60. Nos computadores modernos, os componentes elétricos possuem apenas dois estados "*on*"e "*off*". Portanto é mais conveniente representar números no computador usando o sistema binário, cuja base é 2.

Binário	Hexadecimal	Binário	Hexadecimal
0000	0	1000	8
0001	1	1001	9
0010	2	1010	A
0011	3	1011	B
0100	4	1100	C
0101	5	1101	D
0110	6	1110	E
0111	7	1111	F

Neste sistema um número não-negativo é representado por

$$N = (a_n a_{n-1} \ldots a_0)_2 = a_n \cdot 2^n + a_{n-1} \cdot 2^{n-1} + \ldots + a_0 \cdot 2^0 \tag{2.1}$$

com a_k assumindo os valores 0 ou 1.

O sistema hexadecimal usa base 16 e portanto temos 16 dígitos diferentes. As vezes programadores usam o sistema hexadecimal pois o sistema binário pode ser cansativo e longo para representar algum valor. Exemplo:

$$(1234)_{16} = 1 \times 16^3 + 2 \times 16^2 + 3 \times 16^1 + 4 \times 16^0 = 4096 + 512 + 48 + 4$$

A representação hexadecimal usa as letras de A até F para os seis dígitos adicionais (10, 11, 12, 13, 14, 15). Outros exemplos:

$$(234)_{16} \qquad (BEEF)_{16} \qquad (DEAD)_{16} \qquad (0AFB)_{16}$$

A conversão entre os sistemas binário e hexadecimal é simples, uma vez que trabalhamos com grupos de 4 bits para cada dígito do sistema hexadecimal.

Podemos trabalhar no sistema octal, cuja base é 8, de forma análoga ao sistema hexadecimal. Quando queremos converter entre binário e octal, basta lembrar que cada grupo de 3 bits é um dígito do sistema octal.

2.1.1 Conversão binário para decimal

Pode ser obtida diretamente pelo uso da Equação (2.1).

Exemplo 1. Exemplos

$$\begin{aligned}(11)_2 &= 1 \cdot 2^1 + 1 \cdot 2^0 \\ &= 2 + 1 = 3\end{aligned}$$

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

2.1.2 Conversão decimal para binário

O procedimento é dividir o número por 2, a seguir continuar dividindo o quociente por 2, até que o quociente seja igual a 0. O número na base 2 é obtido tomando-se o resto das divisõees anteriores.

Exemplo 2. Exemplo Converter $(13)_{10}$ para binário.

$$\begin{aligned} 13 \div 2 = 6, &\quad \text{resto}=1 \\ 6 \div 2 = 3, &\quad \text{resto}=0 \\ 3 \div 2 = 1, &\quad \text{resto}=1 \\ 1 \div 2 = 0, &\quad \text{resto}=1 \quad \uparrow \end{aligned}$$

Portanto

$$(13)_{10} = (1101)_2$$

2.2 Representação de números inteiros no computador

Inteiros são armazenados usando uma palavra de 32 bits no computador.

- Se estamos interessados em números **não negativos**, a representação é simples:

$$(71)_{10} \Rightarrow \boxed{00000000}\boxed{00000000}\boxed{00000000}\boxed{01000111}$$

  Os valores que podemos representar vão de 0 a $2^{32} - 1$.

- Para números inteiros com sinal, uma possibilidade, é reservar 1 bit para indicar qual o sinal do número. Exemplo com palavra de 8 bits para simplificar.

$$(13)_{10} \Rightarrow \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ \hline \end{array}$$
$$\text{bits} \quad 7 \quad 6 \quad 5 \quad 4 \quad 3 \quad 2 \quad 1 \quad 0$$

  Bit de sinal:

  - 0 $\Rightarrow$ positivo
  - 1 $\Rightarrow$ negativo

2.2.1 Complemento a dois

Existe uma forma de representação mais esperta para números inteiros com sinal chamada de **complemento a dois**. O bit mais significativo ainda é usado para o sinal. Um número x positivo, tal que $0 \leq x \leq 2^{31} - 1$ é representado normalmente pela representação binária deste número, com bit de sinal 0. Entretanto, um número negativo $-y$, onde $1 \leq y \leq 2^{31}$ é armazenado como a representação binária do inteiro positivo:

$$2^{32} - y$$

bits	número	bits	número	$2^B - y$
0000	0	1000	-8	$2^4 - 8 = 8$
0001	1	1001	-7	$2^4 - 7 = 9$
0010	2	1010	-6	$2^4 - 6 = 10$
0011	3	1011	-5	$2^4 - 5 = 11$
0100	4	1100	-4	$2^4 - 4 = 12$
0101	5	1101	-3	$2^4 - 3 = 13$
0110	6	1110	-2	$2^4 - 2 = 14$
0111	7	1111	-1	$2^4 - 1 = 15$

Para ilustrar considere uma palavra de 4 bits. Podemos representar os seguintes números de 0 a 7 e de -8 a -1. Veja a tabela. Para negar um número em complemento a dois, use o seguinte algoritmo:

1. Inverta todos os bits do númeoro ($0 \to 1$ e $1 \to 0$).
2. Some 1 ao resultado invertido.

Exemplo 3. Exemplo Para simplificar, considere uma palavra de 4 bits. Seja $x = 2$ e $y = 2$. Qual a representação em complemento a dois de $-y$?

Solução 1. Solução do exemplo Temos que $(2)_{10} = (0010)_2$. Pelos passos do algoritmo temos:

1. Invertendo os bits $0010 \to 1101$
2. Somando 1

$$\begin{array}{r} 1101 \\ +0001 \\ \hline 1110 \end{array}$$

Ou seja, $-y = -2$ é representado por $2^4 - 2 = 16 - 2 = 14$, cuja representação binária é $(14)_{10} = (1110)_2$.

- A grande motivação para esse sistema é que não precisamos de hardware específico para a operação de subtração. O hardware de adição pode ser usado uma vez que $-y$ tenha sido representado com complemento a dois.
- Este sistema é usado em muitos dos computadores atuais.

Subtração: $x - y = 2 + (-2)$

$$\begin{array}{r} 0010 \\ +\ 1110 \\ \hline 10000 \end{array}$$

Como o bit mais à esquerda não pode ser representado neste sistema fictício cuja palavra tem 4 bits, ele é descartado e o resultado é 0, como esperado.

2.3 Representação de números fracionários

Se x é um número real positivo, sua parte integral x_i é o maior inteiro menor ou igual a x, enquanto

$$x_f = x - x_i$$

é a sua parte fracionária.

A parte fracionária sempre pode ser escrita como uma fração decimal

$$x_f = \sum_{k=1}^{\infty} b_k 10^{-k} \tag{2.2}$$

onde cada b_k é um inteiro não-negativo menor que 10. Se $b_k = 0$ para todo k maior do que algum inteiro, dizemos então que a fração **termina**. Caso contrário, dizemos que a fração **não termina**.

Exemplo 4. Exemplos

$$\frac{1}{4} = 0.25 = 2 \cdot 10^{-1} + 5 \cdot 10^{-2} \qquad \text{(termina)}$$
$$\frac{1}{3} = 0.333\ldots = 3 \cdot 10^{-1} + 3 \cdot 10^{-2} + 3 \cdot 10^{-3} \qquad \text{(não termina)}$$

□

Se a parte integral de x é um inteiro no sistema decimal da forma

$$x_i = (a_n a_{n-1} \ldots a_0)_{10}$$

enquanto a parte fracionária é dada por (2.2), é comum escrever

$$x = (a_n a_{n-1} \ldots a_0 \cdot b_1 b_2 b_3 \ldots)_{10}$$

2.3.1 Algoritmo de conversão para binário

De forma análoga para o sistema binário podemos escrever

$$x_f = \sum_{k=1}^{\infty} b_k 2^{-k}, \qquad x_i = (a_n a_{n-1} \ldots a_0)_2$$

então $x = (a_n a_{n-1} \ldots a_0 \cdot b_1 b_2 b_2 \ldots)_2$. A fração binária $(\cdot b_1 b_2 b_2 \ldots)_2$ para um dado x_f entre 0 e 1 pode ser calculada da seguinte forma: dado x entre 0 e 1, gere b_1, b_2, b_3 fazendo:

$$\begin{aligned} c_0 &= x \\ b_1 &= (2 \cdot c_0)_i, \quad c_1 = (2 \cdot c_0)_f \\ b_2 &= (2 \cdot c_1)_i, \quad c_2 = (2 \cdot c_1)_f \\ b_3 &= (2 \cdot c_2)_i, \quad c_3 = (2 \cdot c_2)_f, \quad \ldots \end{aligned}$$

onde $(.)_i$ representa a parte integral e $(.)_f$ a parte fracionária do número.

Exemplo 5. Qual a representação binária de $(0.625)_{10}$?

Solução 2.

$$\begin{aligned} 2 \cdot 0.625 = 1.25 &\quad \Rightarrow \quad b_1 = 1 \\ 2 \cdot 0.25 = 0.50 &\quad \Rightarrow \quad b_2 = 0 \\ 2 \cdot 0.5 = 1.00 &\quad \Rightarrow \quad b_3 = 1 \\ 2 \cdot 0.0 = 0.00 &\quad \Rightarrow \quad b_4 = b_5 = \ldots = 0 \end{aligned}$$

Portanto $(0.625)_{10} = (0.101)_2$.

Exemplo 6. Qual a representação binária de $(0.1)_{10}$? Temos $x = 0.1$.

$$\begin{aligned} 2 \cdot 0.1 = 0.2 &\quad \Rightarrow \quad b_1 = 0 \\ 2 \cdot 0.2 = 0.4 &\quad \Rightarrow \quad b_2 = 0 \\ 2 \cdot 0.4 = 0.8 &\quad \Rightarrow \quad b_3 = 0 \\ 2 \cdot 0.8 = 1.6 &\quad \Rightarrow \quad b_4 = 1 \\ 2 \cdot 0.6 = 1.2 &\quad \Rightarrow \quad b_5 = 1 \\ 2 \cdot 0.2 = 0.4 &\quad \Rightarrow \quad b_6 = 0 \\ 2 \cdot 0.4 = 0.8 &\quad \ldots \end{aligned}$$

Portanto $(0.1)_{10} = (0.000110011\ldots)_2 = (0.0\overline{0011})_2$.

---

entrada: x menor do que 1
saída: número binário b
$k = 1$;
faça
  se $2x >= 1$ **então**
    $b_k = 1$;
  senão
    $b_k = 0$;
  fim-se
  $x = 2x - b_k$;
  $k = k + 1$;
enquanto *(*$x \neq 0$*)*;
retorne $(b_1 b_2 \ldots b_k)$;

2.4 Mudança de base

Para transformar um número real que está representado na base 10 para a base 2, o procedimento é transformar a parte inteira e a parte fracionária usando os respectivos algoritmos já vistos.

Exemplo 7. Exemplos Converter da base 10 para a base 2.

$$(5.75)_{10} = (101.110)_2$$
$$(3.8)_{10} = (11.11001100\ldots)_2$$
$$(33.023)_{10} = (100001.00000101\ldots)_2$$

Converter da base 2 para a base 10.

$$(11.0101)_2 = (3.3125)_{10}$$

2.5 Representação de números reais

No computador números reais são armazenados usando o sistema binário e a representação destes números é finita. Por exemplo, os números

$$\pi = 3.1415\ldots$$
$$e = 2.71828\ldots$$

não podem ser representados perfeitamente no computador. O que é armazenado então é uma versão aproximada destes números. De forma geral, existem duas possibilidades para essa representação:

- Representação em Ponto Fixo
- Representação em Ponto Flutuante

2.5.1 Representação em ponto fixo

Neste esquema de representação, a palavra do computador de usualmente 32 bits é dividida em 3 campos:

- 1 bit para o sinal
- campo de bits para a parte integral
- campo de bits para a parte fracionária

Exemplo 8. Exemplo Sistema com 1 bit para sinal, 15 bits para parte integral e 16 bits para a parte fracionária. Neste sistema $(11/2)_{10} = (5.5)_{10}$ é representado da seguinte forma:

0	000000000000101	1000000000000000

O sistema de ponto fixo é severamente limitado pelo tamanho dos números que este pode armazenar. No exemplo anterior, apenas números de magnitude entre 2^{-16} (exatamente) e 2^{15} (um pouco menos) podem ser armazenados no sistema.

Em algumas aplicações essa limitação pode não ser aceitável, e portanto este formato de representação é raramente utilizado. O formato de representação mais usado para números reais é a representação de ponto flutuante, que descrevemos a seguir.

2.5.2 Representação em ponto flutuante

A representação de ponto flutuante é baseada na notação científica. Nessa notação um número real não-zero é expresso por

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

onde β é a base do sistema de numeração, d é a mantissa e e é o expoente. A mantissa é um número da forma

$$(0 \cdot d_1 d_2 d_3 \dots d_t)_\beta$$

representada por t dígitos, onde $0 \leq d_i \leq (\beta - 1)$, para $i = 1, \dots, t$ com $d_1 \neq 0$. O expoente e está no intervalo $[L, U]$. Ao exigir que $d_1 \neq 0$, dizemos que o número está **normalizado**.

Iremos denotar um sistema de ponto flutuante por

$$F(\beta, t, L, U)$$

onde

- β é a base do sistema
- t número de dígitos da mantissa
- L menor valor para o expoente
- U maior valor para o expoente

Em qualquer máquina apenas um subconjunto dos números reais é representado exatamente, e portanto nesse processo a representação de um número real será feita com **arredondamento** ou **truncamento**.

Arredondamento em ponto flutuante

Imagine que só dispomos de quatro dígitos para representar os números em uma máquina. Como seria a melhor forma de representar $\frac{15}{7} = 2.142857 = x$? Usando 2.142 ou talvez 2.143? Se calcularmos o erro vemos que

$$|2.142 - x| = 0.000857$$
$$|2.143 - x| = 0.000143$$

Como o erro é menor para 2.143 concluimos que essa é a melhor forma de representar esse número. Este número foi arredondado. O que significa arredondar um número?

Para arredondar um número na base 10, devemos apenas observar o primeiro dígito a ser descartado. Se este dígito é menor que 5 deixamos os dígitos inalterados; e se é maior ou igual a 5 devemos somar 1 ao último dígito remanescente.

Exemplo 9. Considere o seguinte sistema: $F(10, 3, -3, 3)$.
Neste sistema o número 12.5 é representado por

$$+(0.125)_{10} \times 10^2$$

O número de Euler

$$e = 2.718281\ldots$$

é representado por

$$+ (0.271)_{10} \times 10^1 \quad \text{(com truncamento)}$$
$$+ (0.272)_{10} \times 10^1 \quad \text{(com arredondamento)}$$

Exemplo 10. Considere o seguinte sistema: $F(10, 3, -5, 5)$. Os números representados neste sistema terão a forma

$$\pm(0.d_1d_2d_3) \times 10^e, \quad 0 \leq d_i \leq 9, \quad d_1 \neq 0, \quad e \in [-5, 5]$$

O menor número em valor absoluto representado nessa máquina é

$$m = 0.100 \times 10^{-5} = 10^{-1} \times 10^{-5} = 10^{-6}$$

O maior número em valor absoluto é

$$M = 0.999 \times 10^5 = 99900$$

De forma geral, o menor e o maior número são dados por

$$m = \beta^{L-1}, \qquad M = \beta^U(1 - \beta^{-t})$$

Seja um sistema $F(\beta, t, L, U)$ onde o menor e maior número são denotados por m e M, respectivamente. Dado um número real x, então temos as seguintes situações:

1. $m \leq |x| \leq M$: O número **pode** ser representado no sistema. Exemplo: $235.89 = 0.23589 \times 10^3$. No sistema $F(10, 3, -3, 3)$ temos

   $$(0.235)_{10} \times 10^3 \text{ com truncamento}$$
   $$(0.236)_{10} \times 10^3 \text{ com arredondamento}$$

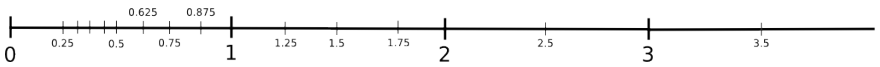
2. $|x| \leq m$: O número **não pode** ser representado no sistema. Neste caso dizemos que ocorreu **underflow**. Ex: 0.517×10^{-8}.

3. $|x| \geq M$: O número **não pode** ser representado no sistema. Neste caso dizemos que ocorreu **overflow**. Ex: 0.725×10^9.

Exemplo 11. Exemplo Considere o sistema $F(2, 3, -3, 3)$. Para simplificar, considere uma palavra com 7 bits, onde temos 1 bit para o sinal do número, 3 bits para o expoente (incluindo seu sinal) e 3 bits para a mantissa. Vamos representar $x_{min} = (0.100)_2 \times 2^{-3}$.

sinal do número	0
sinal do expoente	1
expoente	$(3)_{10} = (11)_2$
mantissa	$(0.100)_2$

Sendo assim temos a seguinte representação para $(0.0625)_{10}$ neste sistema



0	111	100

Exemplo 12. Considere o sistema $F(2, 3, -1, 2)$. Quantos e quais números podem ser representados neste sistema?

Neste sistema os números são da forma

$$\pm 0.d_1d_2d_3 \times 2^e$$

temos

- 2 possibilidades para o sinal
- $(1 \cdot 2 \cdot 2) = 4$ possibilidades para a mantissa
- 4 possibilidades para o expoente $(-1, 0, 1, 2)$

portanto temos $2 \cdot 4 \cdot 4 = 32$, e considerando que o zero também faz parte do sistema, concluimos que podemos representar 33 números distintos.

Para responder quais são os números, notemos que as possíveis formas da mantissa são 0.100, 0.101, 0.110 e 0.111 e as formas do expoente são 2^{-1}, 2^0, 2^1 e 2^2. Assim obtemos:

$(0.100)_2 \times 2^{-1}$	$= (0.25)_{10}$	$(0.101)_2 \times 2^{-1}$	$= (0.3125)_{10}$
$(0.100)_2 \times 2^0$	$= (0.5)_{10}$	$(0.101)_2 \times 2^0$	$= (0.625)_{10}$
$(0.100)_2 \times 2^1$	$= (1.0)_{10}$	$(0.101)_2 \times 2^1$	$= (1.25)_{10}$
$(0.100)_2 \times 2^2$	$= (2.0)_{10}$	$(0.101)_2 \times 2^2$	$= (2.5)_{10}$
$(0.110)_2 \times 2^{-1}$	$= (0.375)_{10}$	$(0.111)_2 \times 2^{-1}$	$= (0.4375)_{10}$
$(0.110)_2 \times 2^0$	$= (0.75)_{10}$	$(0.111)_2 \times 2^0$	$= (0.875)_{10}$
$(0.110)_2 \times 2^1$	$= (1.5)_{10}$	$(0.111)_2 \times 2^1$	$= (1.75)_{10}$
$(0.110)_2 \times 2^2$	$= (3.0)_{10}$	$(0.111)_2 \times 2^2$	$= (3.5)_{10}$

O zero é representado de uma forma especial: todos os dígitos d_i da mantissa e do expoente são nulos.

É extremamente importante observar com relação aos números de ponto flutante, que eles são **discretos** e não contínuos como um número real $x \in \mathbb{R}$, como definido usualmente na matemática.

A figura abaixo ilustra essa característica para o sistema do exemplo (por simplicidade, omitimos a exibição dos números negativos).

2.6 Formato ponto flutuante IEEE 754

Por volta dos anos 50 computação usando ponto flutuante estava em alta. Cada fabricante de computador desenvolvia o seu sistema de ponto flutuante, o que levava a muita inconsistência a como um programa funcionava em diferentes máquinas. Apesar da maioria das

máquinas usarem o sistema binário, algumas como as da série IBM 360/370 usavam o sistema hexadecimal. Em resumo, era muito difícil escrever um software **portável** que pudesse funcionar corretamente em todas as máquinas.

Em uma cooperação extraordinária entre cientistas e desenvolvedores de microprocessadores, um padrão para representação de números binários usando ponto flutuante foi desenvolvida por volta no final dos anos 70, início dos anos 80.

O padrão usa o sistema binário e dois formatos para representação de números podem ser adotados: precisão simples e precisão dupla.

Neste formato um número é representado de forma **normalizada** por

$$\pm 1.d_1 d_2 \dots d_n \times 2^e$$

Em precisão simples um número real é representado por 32 bits, sendo que:

- 1 bit para o sinal
- 8 bits para o expoente
- 23 bits para a mantissa

ou seja, tem o seguinte formato binário

$\pm$	$e_1 e_2 \dots e_8$	$d_1 d_2 \dots d_{23}$

O primeiro bit à esquerda do ponto binário, isto é $d_0 = 1$, é chamado de **bit escondido** (*hidden bit*).

Nesse formato, o **expoente** não é representado como um inteiro via complemento a dois. Os oito bits do expoente armazenam o número $s = e + 127$. Exemplos:

$$\begin{aligned} e = 1 \quad &\Rightarrow \quad s = 1 + 127 = (128)_{10} = (10000000)_2 \\ e = -3 \quad &\Rightarrow \quad s = -3 + 127 = (124)_{10} = (01111100)_2 \\ e = 52 \quad &\Rightarrow \quad s = 52 + 127 = (179)_{10} = (10110011)_2 \end{aligned}$$

Em particular as sequências de bits (00000000) e (11111111) para o expoente, são usadas para representar, respectivamente, o zero e infinito ou ocorrência de erro, que é denotado por NaN (Not a Number).

Vamos considerar agora a representação da **mantissa**. Como o sistema é normalizado temos $d_0 \neq 0$. Dado que a base é dois, a única possibilidade para o primeiro dígito será sempre igual a 1, e portanto este bit não precisa ser armazenado, por isso é chamado de **bit escondido**.

Com o uso desta normalização temos um ganho na precisão, pois a mantissa passa a ser representada com 24 bits (23 + 1 bit escondido). Exemplo: $(0.125)_{10} = (0.001)_2 = 1.0 \times 2^{-3}$ é armazenado como:

0	01111100	0000000 00000000 00000000

A mantissa só possui um dígito significativo, que é justamente o bit escondido, e portanto os demais 23 bits são representados com 0.

O **menor** número normalizado positivo representável nesse padrão é

Table 1: IEEE Single Precision

$\pm$	$a_1 a_2 a_3 \ldots a_8$	$b_1 b_2 b_3 \ldots b_{23}$

If exponent bitstring $a_1 \ldots a_8$ is	Then numerical value represented is
$(00000000)_2 = (0)_{10}$	$\pm(0.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{-126}$
$(00000001)_2 = (1)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{-126}$
$(00000010)_2 = (2)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{-125}$
$(00000011)_2 = (3)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{-124}$
$\downarrow$	$\downarrow$
$(01111111)_2 = (127)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{0}$
$(10000000)_2 = (128)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{1}$
$\downarrow$	$\downarrow$
$(11111100)_2 = (252)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{125}$
$(11111101)_2 = (253)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{126}$
$(11111110)_2 = (254)_{10}$	$\pm(1.b_1 b_2 b_3 \ldots b_{23})_2 \times 2^{127}$
$(11111111)_2 = (255)_{10}$	$\pm\infty$ if $b_1 = \ldots = b_{23} = 0$, NaN otherwise

0	00000001	0000000 00000000 00000000

nesse caso, como $s = (00000001)_2 = (1)_{10}$, o expoente é

$$\begin{aligned} s &= e + 127 \\ 1 &= e + 127 \quad \Rightarrow \quad e = -126 \end{aligned}$$

assim, o número é

$$(1.0000000\ldots)_2 \times 2^{-126} = 2^{-126} \approx 1.2 \times 10^{-38}$$

O **maior** número normalizado positivo, terá como mantissa um número com 24 bits iguais a 1 e como expoente $(11111110)_2$, isto é

0	11111110	1111111 11111111 11111111

nesse caso, como $s = (11111110)_2 = (254)_{10}$, temos

$$s = e + 127 \quad \Rightarrow \quad 254 = e + 127 \quad \Rightarrow \quad e = 127$$

e ainda

$$\begin{aligned} (1.1111\ldots1)_2 &= 1 + \frac{1}{2} + \ldots + \frac{1}{2^{23}} = \frac{2^{23} + 2^{22} + \ldots + 1}{2^{23}} \\ &= \frac{\sum_{i=0}^{23} 2^i}{2^{23}} = \frac{2^{24} - 1}{2^{23}} = 2 - \frac{1}{2^{23}} \end{aligned}$$

assim

$$(1.111111\ldots)_2 \times 2^{127} = (2 - 2^{-23}) \times 2^{127} \approx 3.4 \times 10^{38}$$

O zero é representado com zeros para expoente e mantissa

0	00000000	0000000 00000000 00000000

Os valores $+\infty$ e $-\infty$ são representados por

0	11111111	0000000 00000000 00000000
1	11111111	0000000 00000000 00000000

Se a sequência de bits para o expoente for composta por todos dígitos iguais a um e a da mantissa for não nula, isto é:

1	11111111	$xxxxxx$ $xxxxxxx$ $xxxxxxx$

temos a ocorrência de *NaN: not a number*, que representam expressões inválidas como:

$$0 * \infty, \quad 0/0, \quad \infty/\infty, \quad \infty - \infty$$

Por outro lado certas expressões que envolvam ∞ e zero possuem um resultado plausível:

$$\begin{aligned} z * 0 &= 0 \\ z/0 &= +\infty \text{ se } z > 0 \\ z/0 &= -\infty \text{ se } z < 0 \\ z * \infty &= \infty \\ \infty + \infty &= \infty \end{aligned}$$

```
#include <stdio.h>
#include <math.h>
int main() {
    float x, y;

    x = 23.0/0.0;
    printf("x = %f\n", x);

    y = sqrt(-5.0);
    printf("y = %f\n", y);

    printf("a = %f\n",       x*0);
    printf("0/0 = %f\n",     0./0.);
    printf("inf/inf = %f\n", x/x);
    printf("inf-inf = %f\n", x-x);
}
```

Saída

```
x = inf
y = -nan
a = -nan
0/0 = -nan
inf/inf = -nan
inf-inf = -nan
```

±	expoente	mantissa	valor
0	00000000	00000000000000000000000	0
1	00000000	00000000000000000000000	-0
0	11111111	00010000000000000000000	NaN
1	11111111	00010100100100100100000	NaN
0	11111111	00000000000000000000000	∞
1	11111111	00000000000000000000000	$-\infty$
0	10000001	10100000000000000000000	$1.101 \cdot 2^{129-127} = 6.5$
1	10000001	10100000000000000000000	$-1.101 \cdot 2^{129-127} = -6.5$
0	10000000	00000000000000000000000	$1.0 \cdot 2^{128-127} = 2$
0	10000010	10100000000000000000000	$1.0 \cdot 2^{130-127} = 8$

A precisão p de um sistema de ponto flutuante é definida como o número de bits da mantissa (incluindo bits escondidos). Sendo assim a precisão simples do formato IEEE 754 é $p = 24$, o que corresponde aproximadamente a 7 dígitos decimais significantes, já que

$$2^{-24} \approx 10^{-7}$$

De forma análoga, na precisão dupla com $p = 53$ temos aproximadamente 16 dígitos significantes.

Por exemplo, a representação em precisão simples de $pi = 3.141592653\ldots$ é

$$\underline{3.141592}741\ldots$$

Propriedade	Simples	Dupla	Estendida
comprimento total	32	64	80
bits na mantissa	23	52	64
bits no expoente	8	11	15
base	2	2	2
expoente máximo	127	1023	16383
expoente mínimo	-126	-1022	-16382
maior número	$\approx 3.4 \times 10^{38}$	$\approx 1.8 \times 10^{308}$	$\approx 1.2 \times 10^{4932}$
menor número	$\approx 1.2 \times 10^{-38}$	$\approx 2.2 \times 10^{-308}$	$\approx 3.4 \times 10^{-4932}$
dígitos decimais	7	16	19

Na linguagem C

- `float`: precisão simples
- `double`: precisão dupla
- `long double`: precisão estendida

Exercícios: Considere os seguintes números representados no sistema decimal:

a) 2

b) 30

c) 5.75

d) 0.1

Escreva a representação de cada um dos números no formato de ponto flutuante IEEE 754 usando precisão simples.

2.7 Operações aritméticas em ponto flutuante

As operações aritméticas em sistemas de ponto flutuante obedecem às seguintes regras:

- **Adição/subtração**: quando dois números em ponto flutuante são somados (ou subtraídos), é preciso alinhar as casas decimais do número de menor expoente até que os expoentes fiquem iguais.
- **Multiplicação/divisão**: nessa operação realizamos o produto (ou divisão) das mantissas e o expoente final da base é obtido, somando (subtraindo) os expoentes de cada parcela.
- Os resultados devem ser truncados ou arredondados.
- Truncamento ou arredondamento depende da máquina.

Exemplo 13: Adição

Adicionar 4.32 e 0.064 em uma máquina com mantissa $t = 2$ e base 10.
Solução:

$$\begin{aligned} 4.32 + 0.064 = 0.43 \times 10^1 + 0.64 \times 10^{-1} = \quad & 0.4300 \times 10^1 \\ & + 0.0064 \times 10^1 \\ & = 0.4364 \times 10^1 \end{aligned}$$

Truncamento $\rightarrow 0.43 \times 10^1$
Arredondamento $\rightarrow 0.44 \times 10^1$

Exemplo 14: Subtração

Subtrair 371 de 372 em uma máquina com mantissa $t = 2$ e base 10.
Solução:

$$\begin{aligned} 372 - 371 = 0.37 \times 10^3 + 0.37 \times 10^3 = \quad & 0.37 \times 10^3 \\ & - 0.37 \times 10^3 \\ & = 0.00 \times 10^3 \end{aligned}$$

A subtração deu 0 em vez de 1. Problema na subtração de dois números aproximadamente iguais.

Exemplo 15

Multiplicação Multiplicar 1234 por 0.016 em uma máquina com mantissa $t = 2$ e base 10.
Solução:

$$\begin{aligned} 1234 * 0.016 = 0.12 \times 10^4 * 0.16 \times 10^{-1} = \quad & 0.12 \times 10^4 \\ & * 0.16 \times 10^{-1} \\ & = 0.0192 \times 10^3 \\ & = 0.19 \times 10^2 \end{aligned}$$

Neste caso usando arredondamento ou truncamento, o resultado é 19, em vez de 19.744 que é o resultado exato.

Exemplo 16

Divisão Dividir 0.00183 por 492 em uma máquina com mantissa $t = 2$ e base 10.
Solução:

$$\begin{aligned} 0.00183 \div 492 = 0.18 \times 10^{-2} \div 0.49 \times 10^3 = \quad & 0.18 \times 10^{-2} \\ & \div 0.49 \times 10^3 \\ & = 0.3673 \times 10^{-5} \end{aligned}$$

Arredondamento $\rightarrow 0.37 \times 10^{-5}$
Truncamento $\rightarrow 0.36 \times 10^{-5}$

É importante observar que algumas propriedades aritméticas como

$$\begin{aligned} \text{associatividade: } & (a + b) + c = a + (b + c) \\ \text{distributividade: } & a(b + c) = ab + ac \end{aligned}$$

não são válidas em sistemas de ponto flutuante.

Para os exemplos a seguir, considere um sistema com base β e três dígitos na mantissa, ou seja, $F(10, 3, L, U)$. Considere ainda que o sistema trabalha com arredondamento após cada uma das operações efetuadas.

Exemplo 17: Propriedades em ponto flutuante

a) $(11.4 + 3.18) + 5.05$ e $11.4 + (3.18 + 5.05)$
b) $5.55(4.45 - 4.35)$ e $5.55 * 4.45 - 5.55 * 4.35$

Solução: a) $(11.4 + 3.18) + 5.05$

$$\begin{aligned} & 0.1140 \times 10^2 \\ + & \ 0.0318 \times 10^2 \\ = & \ 0.1458 \times 10^2 \\ = & \ 0.146 \times 10^2 \text{ (arr.)} \end{aligned}$$

$$\begin{aligned} & 0.1460 \times 10^2 \\ + & \ 0.0505 \times 10^2 \\ = & \ 0.1965 \times 10^2 \\ = & \ 0.197 \times 10^2 \text{ (arr.)} \end{aligned}$$

a) $11.4 + (3.18 + 5.05)$

$$\begin{aligned} & 0.318 \times 10^1 \\ + & \ 0.505 \times 10^1 \\ = & \ 0.823 \times 10^1 \end{aligned}$$

$$\begin{aligned} & 0.0823 \times 10^2 \\ + & \ 0.1140 \times 10^2 \\ = & \ 0.1963 \times 10^2 \\ = & \ 0.196 \times 10^2 \text{ (arr.)} \end{aligned}$$

b) $5.55(4.45 - 4.35)$

$$\begin{aligned} & 0.445 \times 10^1 \\ - & \ 0.435 \times 10^1 \\ = & \ 0.010 \times 10^1 \end{aligned}$$

$$\begin{aligned} & 0.555 \times 10^1 \\ * & \ 0.100 \times 10^0 \\ = & \ 0.055500 \times 10^1 \\ = & \ 0.555 \end{aligned}$$

b) $5.55 * 4.45 - 5.55 * 4.35$

$$
\begin{aligned}
& 0.555 \times 10^1 \\
& * 0.445 \times 10^1 \\
& = 0.246975 \times 10^2 \\
& = 0.247 \times 10^2
\end{aligned}
$$

$$
\begin{aligned}
& 0.555 \times 10^1 \\
& * 0.435 \times 10^1 \\
& = 0.241425 \times 10^2 \\
& = 0.241 \times 10^2
\end{aligned}
$$

$$
\begin{aligned}
& 0.247 \times 10^2 \\
& - 0.241 \times 10^2 \\
& = 0.006 \times 10^2 \\
& = 0.6
\end{aligned}
$$

2.8 Noções básicas sobre erros

Já vimos que introduzimos erros ao representar um número no computador. Como vamos representar *aproximadamente* um número real x por sua versão ponto flutuante no computador, precisamos definir medidas apropriadas para calcular o erro cometido nessa aproximação. Vamos usar as seguintes medidas de erro: erro absoluto e erro relativo.

2.8.1 Erro absoluto

Se $\tilde{x}$ é uma aproximação de x, então o erro absoluto é definido por

$$EA(\tilde{x}) = x - \tilde{x}$$

Exemplo 13. Exemplo Seja $x = 1428.756$. Em uma máquina com mantissa $t = 4$, usando arredondamento e truncamento, respectivamente, temos

$$
\begin{aligned}
\tilde{x_t} = 0.1428 \times 10^4 \quad &\Rightarrow \quad EA(\tilde{x_t}) = 0.756 \times 10^0 \\
\tilde{x_a} = 0.1429 \times 10^4 \quad &\Rightarrow \quad EA(\tilde{x_a}) = 0.244 \times 10^0
\end{aligned}
$$

2.8.2 Erro relativo

O erro relativo é definido por

$$ER(\tilde{x}) = \frac{x - \tilde{x}}{\tilde{x}} = \frac{EA(\tilde{x})}{\tilde{x}}$$

dado que $\tilde{x} \neq 0$.

Exemplo 14. Exemplo

$$\begin{aligned} x_1 = 1000.5, \quad \tilde{x_1} = 1000.6 \\ x_2 = 10.5, \quad \tilde{x_2} = 10.6 \end{aligned} \quad \Rightarrow \quad EA(\tilde{x_i}) = 0.1, i = 1, 2$$

$$ER(\tilde{x_1}) = \frac{0.1}{1000.6} \approx 0.00009994 = 0.9994 \times 10^{-4}$$
$$ER(\tilde{x_2}) = \frac{0.1}{10.6} \approx 0.009433 = 0.9433 \times 10^{-2}$$

Ao invés de erro relativo, as vezes o conceito de **dígitos corretos** pode ser usado. Dizemos que $\tilde{x}$ possui m dígitos (decimais) corretos com relação a x, se o erro $|x - \tilde{x}|$ possui magnitude menor ou igual a 5 no dígito de posição $(m + 1)$, contando a partir do primeiro dígito não-zero de x.

$x = 1/3,$	$\tilde{x} = 0.333,$	$EA(\tilde{x}) = 0.00033$	$\Rightarrow$	$m = 3$
$x = 23.496,$	$\tilde{x} = 23.494,$	$EA(\tilde{x}) = 0.002$	$\Rightarrow$	$m = 4$
$x = 0.02138,$	$\tilde{x} = 0.02144,$	$EA(\tilde{x}) = 0.00006$	$\Rightarrow$	$m = 2$

Apesar de termos definido erro absoluto e erro relativo para valores reais x e $\tilde{x}$, a mesma definição serve para vetores ou matrizes. Nesse caso, se denotarmos a norma de um vetor por $||\cdot||$, então teríamos:

- erro absoluto
$$||x - \tilde{x}||$$
- erro relativo
$$\frac{||x - \tilde{x}||}{||\tilde{x}||}$$

E de forma similar para matrizes. Iremos usar estes conceitos mais adiante no curso.

É claro que em geral não conhecemos o valor real que estamos interessados em aproximar, se soubessemos, não seria necessário se preocupar em aproxima-lo. O que se faz na prática é usar uma medida de erro aproximado, dada por

$$ER = \frac{x_{new} - x_{old}}{x_{new}}$$

A seguir iremos obter estimativas para o erro quando um número é aproximado em ponto flutuante em uma máquina que opera com truncamento ou arredondamento.

2.9 Erros no arredondamento e truncamento

Vamos analisar agora os erros cometidos quando a máquina opera com arredondamento ou truncamento. Para isso, vamos considerar um sistema que trabalha em aritmética de ponto flutuante com t dígitos na mantissa e base 10. Seja $x \in \mathbb{R}$. Vamos escrever x na forma

$$x = f_x \times 10^e + g_x \times 10^{e-t}$$

onde

$$\begin{aligned} e &\rightarrow \text{ expoente} \\ t &\rightarrow \text{ num. dígitos da mantissa} \\ 0.1 \leq f_x &< 1 \\ 0 \leq g_x &< 1 \end{aligned}$$

Exemplo 15. Seja $t = 4$ e $x = 234.57$, então

$$x = \underbrace{0.2345 \times 10^3}_{f_x \times 10^e} + \underbrace{0.7 \times 10^{-1}}_{g_x \times 10^{e-t}}$$

Na representação de x neste sistema, a parcela $g_x \times 10^{e-t}$ não pode ser incorporada à mantissa.

Estamos interessados em obter estimativas para o erro absoluto (ou relativo) máximo cometido quando usamos arredondamento ou truncamento.

2.9.1 Erro no truncamento

Em sistemas que usam truncamento a parcela $g_x \times 10^{e-t}$ é desprezada e portanto x é representado aproximadamente por $\tilde{x} = f_x \times 10^e$.

$$\begin{aligned} |EA(\tilde{x})| = |x - \tilde{x}| &= |f_x \times 10^e + g_x \times 10^{e-t} - f_x \times 10^e| && (2.3) \\ &= |g_x \times 10^{e-t}| && (2.4) \\ &< 10^{e-t}, \quad \text{pois } 0 \leq g_x < 1 && (2.5) \end{aligned}$$

$$\begin{aligned} |ER(\tilde{x})| = \frac{|x - \tilde{x}|}{|\tilde{x}|} &= \frac{|g_x \times 10^{e-t}|}{|f_x \times 10^e|} && (2.6) \\ &< \frac{10^{e-t}}{0.1 \times 10^e} = \frac{10^{e-t}}{10^{e-1}} = 10^{e-t}\ 10^{-(e-1)} = 10^{1-t} && (2.7) \end{aligned}$$

Em resumo

$$\boxed{|EA(\tilde{x})| < 10^{e-t}} \qquad \boxed{|ER(\tilde{x})| < 10^{1-t}}$$

2.9.2 Erro no arredondamento

Em sistemas com arredondamento, f_x é modificado de forma a levar em consideração g_x. O arredondamento é feito da seguinte forma:

$$\tilde{x} = \begin{cases} f_x, & \text{se } |g_x| < 1/2 \\ f_x \times 10^e + 1 \times 10^{e-t}, & \text{se } |g_x| \geq 1/2 \end{cases}$$

ou seja, se $|g_x| < 1/2$, a parcela g_x é desprezada, caso contrário somamos 1 ao último dígito de f_x.

Para analisar o erro máximo cometido quando representamos um número usando ponto flutuante com arredondamento, vamos considerar os dois casos

1. $|g_x| < 1/2$
2. $|g_x| \geq 1/2$

Caso 1: $|g_x| < 1/2$

$$|EA(\tilde{x})| = |x - \tilde{x}| = |g_x \times 10^{e-t}| < \tfrac{1}{2} \times 10^{e-t}$$

$$\begin{aligned} |ER(\tilde{x})| &= \frac{|x - \tilde{x}|}{|\tilde{x}|} \\ &= \frac{|g_x \times 10^{e-t}|}{|f_x \times 10^{e}|} \\ &< \frac{\frac{1}{2} \times 10^{e-t}}{0.1 \times 10^{e}} = \tfrac{1}{2} \times 10^{1-t} \end{aligned}$$

Caso 2: $|g_x| \geq 1/2$

$$\begin{aligned} |EA(\tilde{x})| = |x - \tilde{x}| &= |f_x \times 10^e + g_x \times 10^{e-t} - (f_x \times 10^e + 1 \times 10^{e-t})| \\ &= |g_x \times 10^{e-t} - 1 \times 10^{e-t}| \\ &= |(g_x - 1) \times 10^{e-t}| \\ &< \tfrac{1}{2} \times 10^{e-t} |ER(\tilde{x})| \qquad = \frac{|x - \tilde{x}|}{|\tilde{x}|} \leq \frac{\frac{1}{2} \times 10^{e-t}}{|f_x \times 10^e + 1 \times 10^{e-} \\ &< \frac{\frac{1}{2} \times 10^{e-t}}{0.1 \times 10^{e}} = \tfrac{1}{2} \times 10^{e-t} \times 10^{-(e-1)} = \tfrac{1}{2} \times 10^{1-t} \end{aligned}$$

Ou seja, em ambos os casos temos

$$\boxed{|EA(\tilde{x})| < \tfrac{1}{2} \times 10^{e-t}} \quad \boxed{|ER(\tilde{x})| < \tfrac{1}{2} \times 10^{1-t}}$$

De forma geral para uma base β qualquer, seguindo o mesmo raciocínio, podemos enunciar o seguinte teorema:

Teorema 2. Suponha uma máquina com base β e mantissa com t dígitos. Então, qualquer número real pode ser representado no intervalo de ponto flutuante da máquina com um erro relativo que não excede o epsilon de máquina ϵ_{mach} (*machine epsilon* ou *round-off unit*), o qual é definido por

$$\epsilon_{mach} = \begin{cases} \frac{1}{2}\beta^{1-t}, & \text{se arredondamento for usado} \\ \beta^{1-t}, & \text{se truncamento for usado} \end{cases}$$

2.10 Efeitos numéricos

Além dos erros causados pela representação no computador e pelas operações aritméticas, existem certos efeitos numéricos que contribuem para aumentar os erros introduzidos. A seguir iremos estudar alguns casos importantes como a adição ou subtração entre um número grande e um pequeno, subtração de dois números quase iguais, propagação do erro, etc.

2.10.1 Somar ou subtrair um número pequeno e um grande

Para exemplificar considere um sistema $F(10, 4, L, U)$. Somar 0.1 e 5000.

$$\begin{aligned} 0.1 + 5000 = &\ 0.1000 \times 10^0 + 0.5000 \times 10^4 \\ = &\ 0.00001 \times 10^4 \\ + &\ 0.50000 \times 10^4 \\ = &\ 0.50001 \times 10^4 \end{aligned}$$

Usando arredondamento (ou truncamento), obtemos 0.5×10^4.

Exemplo 18: Número pequeno e grande

Calcular

$$S = 5000 + \underbrace{0.1 + \ldots 0.1}_{10\times} = 5000$$

$$S = \underbrace{0.1 + \ldots 0.1}_{10\times} + 5000 = 5001$$

Observe que embora analiticamente o resultado das somas seja o mesmo, quando usamos uma máquina de ponto flutuante $F(10, 4, L, U)$, o resultado é diferente.

No primeiro caso, ao somar um número grande 5000 e um pequeno 0.1 cometemos um erro que se propaga em toda a soma de S. No segundo caso, evitamos este problema.

2.10.2 Cancelamento

O **cancelamento** ocorre quando subtraimos dois números quase iguais, ou quando somamos números de sinais opostos mas de magnitudes semelhantes. Como vimos, quando calculamos a diferença $x - y$, o resultado terá o mesmo expoente e. Para normalizar o resultado obtido, devemos mover os dígitos para a esquerda de tal forma que o primeiro seja diferente de zero. Desta forma, uma quantidade de dígitos iguais a zero aparece no final da mantissa do número normalizado. Estes zeros não possuem significado algum, e dizemos que ocorreu a perda de dígitos significativos.

Exemplo 19

Calcular $\sqrt{37} - \sqrt{36}$ em uma máquina $F(10, 4, L, U)$ usando arredondamento. Para efeitos de comparação, apresentamos a resposta exata aqui:

$$\sqrt{37} - \sqrt{36} = 6.08276253 - 6 = 0.08276253$$

Nessa máquina temos

$$\begin{aligned} \sqrt{37} = 6.08276253 &\quad \rightarrow \quad 0.6083 \times 10^1 \\ \sqrt{36} = 6.0 &\quad \rightarrow \quad 0.6000 \times 10^1 \end{aligned}$$

Efetuando a subtração temos

$$\begin{aligned} & 0.6083 \times 10^1 \\ - \; & 0.6000 \times 10^1 \\ = \; & 0.0083 \times 10^1 = 0.8\mathbf{300} \times 10^{-1} \end{aligned}$$

A resposta exata é $0.8277 \times 10^{-1} \rightarrow$ perda de dígitos significativos! É possível obter um resultado mais preciso? Sim, basta considerar que

$$\sqrt{x} - \sqrt{y} = \sqrt{x} - \sqrt{y}\frac{(\sqrt{x} + \sqrt{y})}{\sqrt{x} + \sqrt{y}} = \frac{x - y}{\sqrt{x} + \sqrt{y}}$$

Portanto para

$$\sqrt{37} - \sqrt{36}$$

temos

$$\begin{aligned} \frac{x - y}{\sqrt{x} + \sqrt{y}} = \frac{37 - 36}{\sqrt{37} + \sqrt{36}} &= \frac{1}{0.6083 \times 10^1 + 0.6000 \times 10^1} \\ &= \frac{1}{0.1208 \times 10^2} \\ &= 0.08278145 = 0.8278 \times 10^{-1} \end{aligned}$$

que é um resultado mais preciso que o anterior.

2.10.3 Propagação do erro

Problemas numéricos não ocorrem apenas quando dois números quase iguais são subtráidos. Também ocorrem no cálculo de uma soma, quando uma soma parcial é muito grande se comparada com o resultado final. Considere que:

$$s = \sum_{k=1}^{n} a_k$$

seja a soma a ser computada e que os a_k podem ser positivos ou negativos e de diferentes magnitudes. O cálculo pode ser feito da seguinte forma:

$$s_1 = a_k, \quad s_k = s_{k-1} + a_k, \quad k = 2, 3, \ldots, n$$

tal que $s = s_n$.

Exemplo 20

Considere uma máquina $F(10, 4, L, U)$ com truncamento. Vamos efetuar a seguinte

operação:

$$S = \sum_{i=1}^{4}(x_i + y_i) \quad \text{com} \quad x_i = 0.46709, \quad \text{e} \quad y_i = 3.5678$$

Para $i = 1$, temos: $(x_1 + y_1) = 0.4034 \times 10^1$
Erro absoluto é: $EA(\bar{S}) = |4.03569 - 4.034| = 0.00169$.

Para $i = 2$, temos: $(x_1 + y_1) + (x_2 + y_2) = 0.8068 \times 10^1$
Erro absoluto: $EA(\bar{S}) = |8.07138 - 8.068| = 0.00338$

Para $i = 3$, temos: $(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) = 0.1210 \times 10^2$
Erro absoluto: $EA(\bar{S}) = |12.10707 - 12.10| = 0.00707$
Para $i = 4$, temos

$$(x_1 + y_1) + (x_2 + y_2) + (x_3 + y_3) + (x_4 + y_4) = 0.1613 \times 10^2$$
$$EA(\bar{S}) = |16.14267 - 16.13| = 0.01276$$

De onde pode-se observar que o erro absoluto aumenta à medida em que as operações aritméticas são realizadas.

Exemplo 21

Ao aproximar a função e^x por uma série de Taylor em torno do ponto $a = 0$, temos

$$p(x) = 1 + x + \tfrac{x^2}{2!} + \tfrac{x^3}{3!} + \ldots$$

Se tentarmos avaliar $p(x)$ para valores negativos de x como por exemplo -20, -25, -30, com muitos termos para obter boa precisão o resultado obtido estará comprometido devido a propagação do erro que acontece. Para ilustrar o problema apresentamos uma pequena função implementada na linguagem Python para calcular $p(x)$. Comparamos o valor de $p(x)$ e da função $\exp(x)$ da linguagem Python para diferentes valores de x.

```
def exp_taylor(n,x):
  fat = 1.0
  term = 1.0
  soma = term
  i = 1
  while(i<=n):
     fat = fat * i
     term = term * x
     soma = soma + (term/fat)
     i = i + 1
  return soma
```

Resultados:

```
  x    exp(x)          exp_taylor(n,x)
  10   2.202647e+04    2.202646e+04
  -5   6.737947e-02    6.737947e-02
 -10   4.539993e-05    9.703415e-04
 -20   2.061153e-09    1.599694e+06
 -30   9.357622e-14    3.848426e+11
```

O problema ocorre pois, por exemplo para $x = -25$ os termos $\frac{25^{24}}{24!}$ e $-\frac{25^{25}}{25!}$ são muito próximos e portanto sofrem cancelamento. Veja:

```
 23  termo  5.057240e+09  soma -2.394348e+09
 24  termo -5.497000e+09  soma  2.662893e+09
 25  termo  5.726042e+09  soma -2.834108e+09
 26  termo -5.726042e+09  soma  2.891934e+09
 27  termo  5.505810e+09  soma -2.834108e+09
 28  termo -5.097972e+09  soma  2.671702e+09
 29  termo  4.551761e+09  soma -2.426270e+09
 30  termo -3.923932e+09  soma  2.125491e+09
```

Nesse caso, uma estratégia simples pode ser adotada para evitar o cancelamento. Para valores negativos de x, basta calcular normalmente e^x e depois retornar $\frac{1}{e^x}$ como resultado.

```
def exp_taylor(n,x):
  fat, term = 1.0, 1.0
  soma, i = term, 1
  if x<0:
    x = - x
    neg = True

  while(i<=n):
     fat = fat * i
     term = term * x
```

```
      soma = soma + (term/fat)
      i = i + 1

   if neg:
      return 1.0/soma
   else:
      return soma
```

2.11 Desastres

O uso incorreto de aritmética de ponto flutuante na implementação de programas de computador e/ou softwares científicos já foi responsável por alguns desastres.

2.11.1 Patriot missile failure - Guerra do Golfo (1991)

Uma bateria de mísseis Patriot (*Phased Array TRacking Intercept Of Target*) americano, falhou ao rastrear e interceptar um míssil Scud do Iraque. O míssil Scud acertou o acampamento americano, matou 28 soldados e feriu centenas. O relatório técnico apontou uma falha no software. A palavra do computador onde o software executava tinha 24 bits. O tempo era medido em décimos de segundo (1/10). O valor (1/10) ao ser representado em binário não termina, isso levou ao acúmulo do erro no software após longo tempo de execução do sistema e que por sua vez resultou na falha.

Figura 2.1: Patriot missile failure.

2.11.2 Ariane 5

Foguete da European Space Agency explode 40s após o lançamento. Milhões de dolares foram investidos no seu desenvolvimento e equipamento. O relatório técnico acusou um erro do software no sistema de referência inercial. Problema: número de 64 bits de ponto flutuante era convertido em um inteiro de 16 bits com sinal. Falha na conversão para números maiores que 32767, que é o maior inteiro representável com 16 bits.

Figura 2.2: Foguete Ariane 5.

2.11.3 Sleipnir offshore

Um acidente fez com que a plataforma de petróleo Sleipnir A afundasse. Após o acidente a empresa da plataforma, Statoil, uma empresa norueguesa solicita a empresa SINTEF um relatório técnico. O resultado da análise apontou uma falha em uma parede, resultando em uma rachadura e vazamento. Essa falha aconteceu como resultado de uma combinação de erros no programa de análise de elementos finitos, que subestimou a tensão na parede.

Figura 2.3: Sleipner offshore.

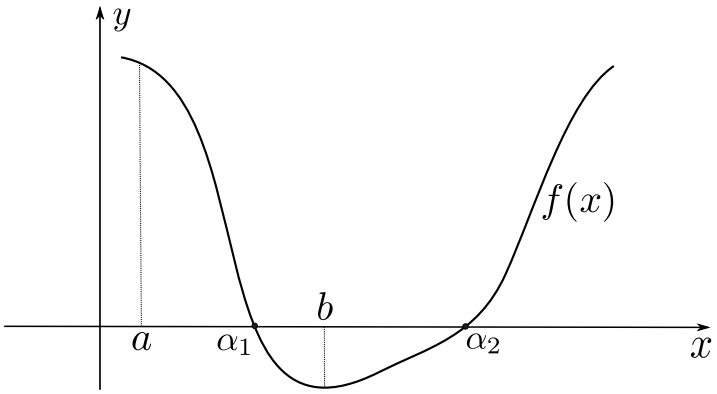
Capítulo 3

Raízes de Equações Não-Lineares

Vamos considerar agora métodos para determinar aproximações para as raízes de equações não-lineares. Esse problema pode ser escrito da seguinte forma: dada uma função não-linear $f : \mathbb{R} \to \mathbb{R}$, contínua, procuramos o valor de x para o qual

$$f(x) = 0.$$

Esse problema pode ser interpretado de forma gráfica conforme mostra a Figura **??**, onde a função apresentada possui duas raízes denotadas por α_1 e α_2.



3.1 Introdução

Para iniciar, considere a seguinte motivação e necessidade para estudar esses métodos. Sabe-se que para polinômios de grau até quatro, suas raízes podem ser calculadas através de uma expressão fechada, como por exemplo no caso de uma função quadrática

$$ax^2 + bx + c = 0 \quad \Rightarrow \quad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Entretanto, de forma geral, não podemos encontrar os zeros de uma função através de uma expressão fechada. Portanto, para encontrar os zeros de uma função temos que recorrer a *métodos numéricos* para encontrar uma *solução aproximada*.

Note que em alguns casos como $x^2 + 1 = 0$, os zeros das funções podem ser números complexos: $x = \pm\sqrt{-1} = \pm i$. Porém, nesse curso, iremos trabalhar apenas com as raízes reais.

3.1.1 Exemplos de problemas

Problema 1

Considere a seguinte função $f(x)$ não-linear

$$f(x) = x^2 - 4\sin(x) = 0,$$

isto é, o problema consiste em encontrar x que faça com que o valor da função $f(x)$ seja igual a zero.

Problema 2

Considere a seguinte equação:

$$C = \frac{M}{r}\left[1 - (1+r)^{-n}\right]$$

onde C é o capital, M é a mensalidade, r é a taxa de juros por cada período (expressa como uma fração) e n é o número de anos.

Uma pessoa pode pagar uma mensalidade de 1250 reais. Se pretende contrair um empréstimo de 10000 reais a 10 anos, qual é a taxa que poderá suportar? Para isso considere que:

$$C = 10000,\ M = 1250,\ n = 10 \quad \Rightarrow \quad 10000 = \frac{1250}{r}[1 - (1+r)^{-10}]$$
$$f(r) = 10000 - \frac{1250}{r}[1 - (1+r)^{-10}] = 0$$

Problema 3

A seguinte equação pode ser usada para calcular o nível de concentração de oxigênio c em um rio, em função da distância x, medida a partir do local de descarga de poluentes:

$$c(x) = 10 - 20(e^{-0.2x} - e^{-0.75x})$$

Calcule a distância para a qual o nível de oxigênio desce para o valor 5. Pretende-se resolver $c(x) = 5$. Podemos escrever como $c(x) - 5 = 0$, isto é:

$$10 - 20(e^{-0.2x} - e^{-0.75x}) - 5 = 0$$

O problema se resume a encontrar x tal que $f(x) = 0$, onde x nesse caso representa a distância para qual o oxigênio decai para um valor igual a 5.

Definição 1 (Zero)**.** Se $f : [a, b] \to \mathbb{R}$ é uma função dada, um ponto $\alpha \in [a, b]$ é um zero (ou raiz) de f se $f(\alpha) = 0$.

Para ilustar os zeros de um função, considere como exemplo a seguinte função $f(x) = tanh(x) - x/3$, cujos zeros são ilustrados na Figura 3.1.

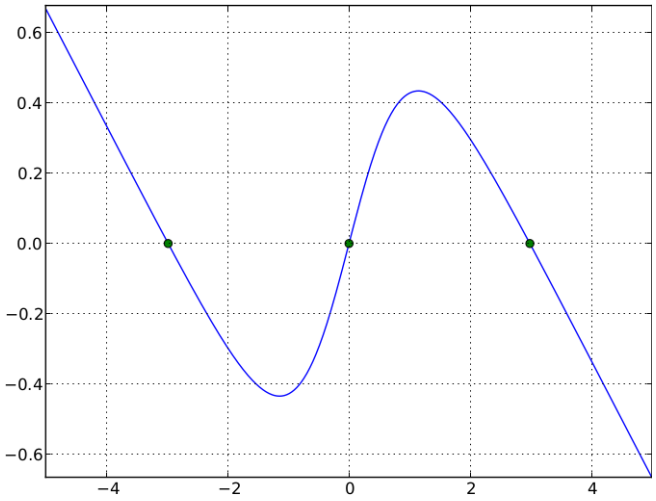


Figura 3.1: Zeros de $f(x) = tanh(x) - x/3$

Definição 2 (Multiplicidade)**.** Um ponto $\alpha \in [a, b]$ é uma raiz de multiplicidade m da equação $f(x) = 0$ se $f(\alpha) = f'(\alpha) = \ldots = f^{(m-1)}(\alpha) = 0$ e $f^{(m)}(\alpha) \neq 0$.

Seja $f(x) = x^2 + 2x + 1 = (x+1)^2$. Nesse caso temos $\alpha = -1$ com multiplicidade $m = 2$, pois $f'(x) = 2(x+1)$ e assim temos que $f(-1) = 0$ e $f'(-1) = 0$.

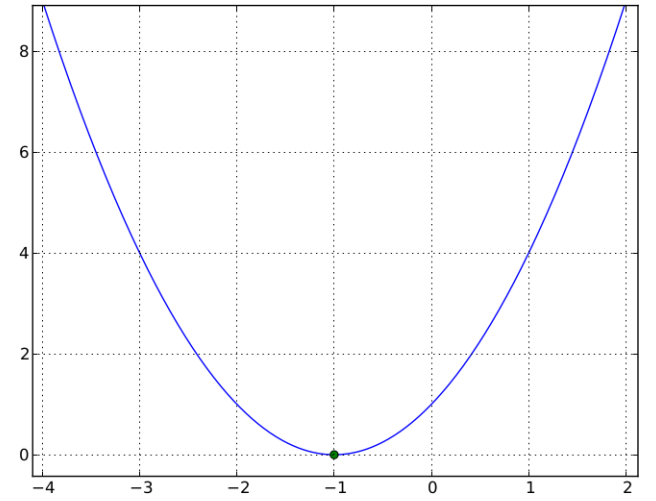


Figura 3.2: Exemplo de função com raiz de multiplicidade $m = 2$.

3.1.2 Métodos para raízes de equações

Os métodos numéricos que vamos estudar geralmente podem ser divididos em duas etapas:

1. Localização das raízes
   - Encontrar o intervalo $[a, b]$ que contenha apenas uma raiz.
2. Refinamento da aproximação

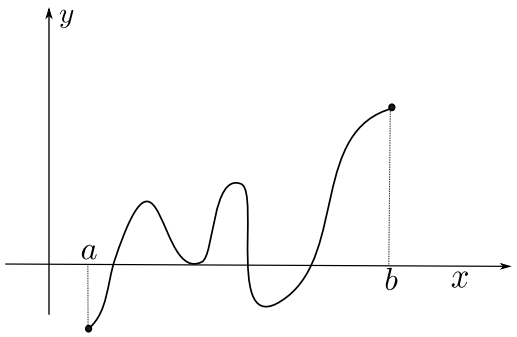
- A partir de uma aproximação inicial $x_0 \in [a, b]$, gerar uma sequência $\{x_0, x_1, x_2, \ldots\}$ que convirja para a raiz exata α de $f(x) = 0$.

Alguns métodos não precisam de um prévio isolamento de cada raiz, necessitam apenas de uma aproximação inicial x_0 (ou mais de uma, as vezes). Entretanto, boa parte deles precisa que a raiz esteja confinada em um intervalo e que ela seja única.

3.1.3 Isolamento das raízes

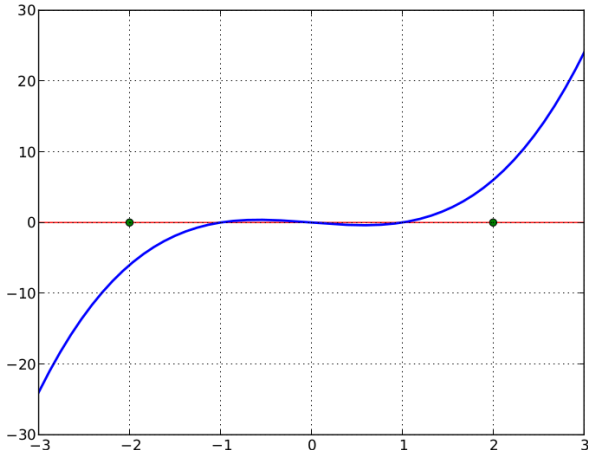
Teorema 3 (1)**.** Seja $f(x) : [a, b] \to \mathbb{R}$ uma função contínua. Se $f(a)f(b) < 0$, então existe **pelo menos** um ponto $x \in [a, b]$, tal que $f(x) = 0$.

Geometricamente, o teorema diz que qualquer gráfico de uma função contínua que começa abaixo do eixo horizontal e termina acima deste, deve cruzar este eixo em algum ponto.



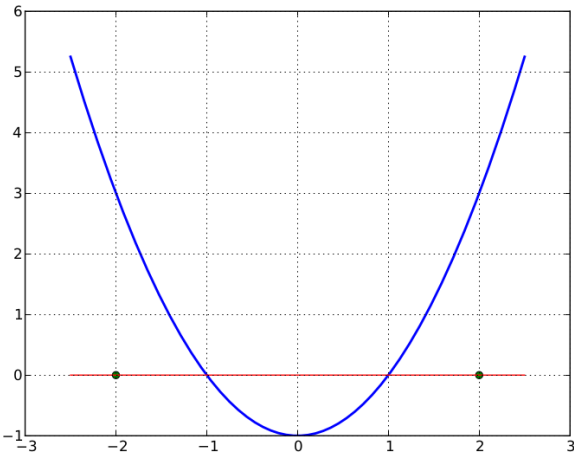
Veja alguns exemplos:

- $f(x) = x^3 - x$, $a = -2$, $b = 2$
  - f é contínua
  - $f(a) = -6$, $f(b) = 6$, sinais opostos
  - Possui 3 raízes no intervalo $[a, b]$.



- $f(x) = x^2 - 1$, $a = -2$, $b = 2$

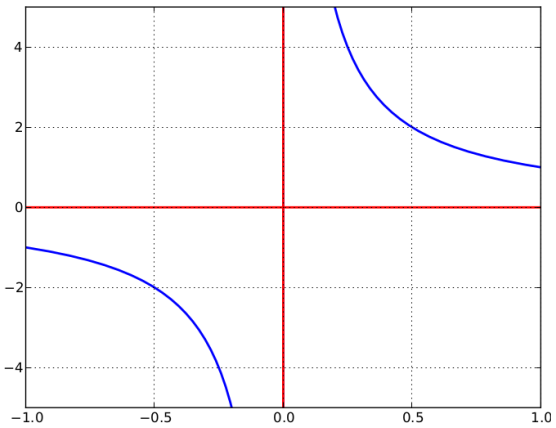
- f é contínua
- $f(a) = (b) = 2$, mesmo sinal!
- Hipótese do teorema não satisfeita! Entretanto, existem raízes.



- $a = -1$, $b = 1$

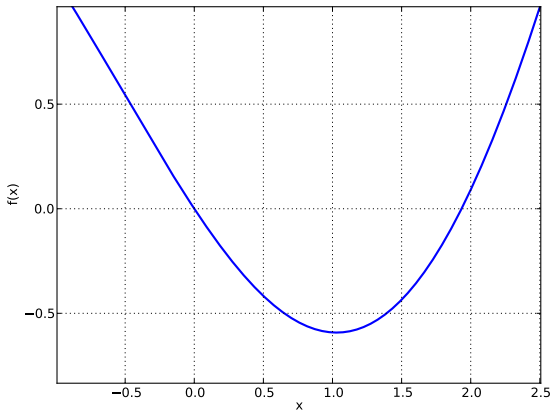
$$f(x) = \begin{cases} \frac{1}{x}, & \text{se } x \neq 0 \\ \text{indef.}, & \text{se } x = 0 \end{cases}$$

- $f(a) = -1$, $f(b) = 1$, sinais opostos
- f é discontínua!
- De fato, não existem raízes!



Exemplo 22

Como encontrar o intervalo da raiz positiva da equação $f(x) = \left(\frac{x}{2}\right)^2 - \sin(x) = 0$? Pode-se utilizar de softwares especializados e fazer um gráfico da função como mostra a figura abaixo.



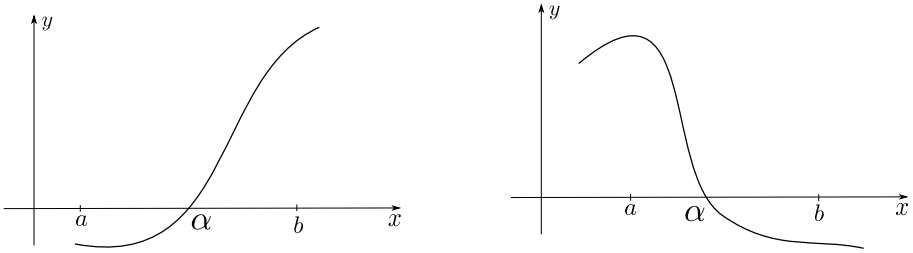
Por inspeção visual conclui-se que $\alpha \in [1.5, 2.0]$. Outra possibilidade é fazer uma tabela de valores de $f(x)$, e usar o Teorema (1).

x	$(\frac{x}{2})^2$	$\sin(x)$	$f(x)$
1.6	0.64	0.996	< 0
1.7	0.72	0.991	< 0
1.8	0.81	0.974	< 0
1.9	0.90	0.946	< 0
2.0	1.00	0.909	> 0

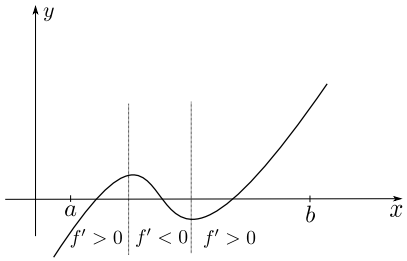
Assim fica claro que existe pelo menos uma raiz em $[1.9, 2.0]$.

Atenção! Na hora de fazer suas contas na calculadora, sempre calcule as funções trigonométricas com argumento x em radianos.

Teorema 4 (2)**.** Sob as hipóteses do Teorema 1, se $f'(x)$ existir e $f'(x)$ **preservar** o sinal em $[a, b]$ então o intervalo contém um único zero de $f(x)$.



$f'(x) > 0, \ \forall x \in [a, b]$ $\quad$ $f'(x) < 0, \ \forall x \in [a, b]$



$f'(x)$ não preserva o sinal

Figura 3.3: Exemplos de situações que a função possui um único zero no intervalo e que não possui apenas um único zero através do uso do Teorema 4.

Exemplo 23

Seja $f(x) = \sqrt{x} - 5e^{-x}$ para $x \geq 0$.
Logo, tabelando os valores da função temos

x	$\sqrt{x}$	$5e^{-x}$	$f(x)$
0.0	0.0	5.0	< 0
1.0	1.0	1.83	< 0
2.0	1.41	0.67	> 0
3.0	1.73	0.24	> 0

Logo sabemos que existe pelo menos uma raiz no intervalo $[1, 2]$. Entretanto, o Teorema 4 nos garante que existe uma única raiz pois

$$f'(x) = \frac{1}{2\sqrt{x}} + 5e^{-x} > 0, \quad \forall x > 0$$

Uma outra alternativa é rearranjar a equação $f(x)$ dada como $g(x) = h(x)$, de tal forma que os gráficos de $g(x)$ e $h(x)$ sejam mais fáceis de serem traçados do que o de f. As raízes da equação original são dadas pelos pontos onde o gráfico de $g$intercepta o gráfico de h.

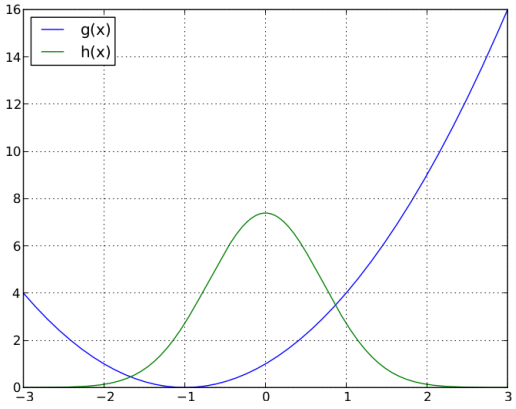
Exemplo 24

Considere a função $f(x) = (x + 1)^2 e^{(x^2-2)} - 1 = 0$.

Podemos rearranjar $f(x)$ como

$$\begin{aligned}
&\rightarrow (x+1)^2 e^{(x^2-2)} = 1 \\
&\rightarrow (x+1)^2 = \frac{1}{e^{(x^2-2)}} \\
&\rightarrow (x+1)^2 = e^{(2-x^2)} \\
&\rightarrow g(x) = h(x)
\end{aligned}$$

Os gráficos das funções (x) e $h(x)$ são apresentados abaixo.



3.1.4 Refinamento

Se o intervalo $[a, b]$ para o qual queremos procurar uma raiz de $f(x)$ já está isolado, o próximo passo consiste em gerar iterativamente uma sequência de aproximações $\{x_0, x_1, x_2, \ldots\}$ cada vez melhores que convirja para a raiz α.

Antes de estudarmos como os métodos geram as aproximações, precisamos decidir como que uma dada aproximação no passo k é suficientemente próxima da raiz exata?

Para isso precisamos definir um critério de parada que determina quando terminar o processo iterativo.

3.1.5 Critério de parada

Na prática a sequência é interrompida quando seus valores satisfizerem a pelo menos um dos seguintes critérios:

$$|x_k - x_{k-1}| \leq \epsilon \tag{3.1}$$

$$\left|\frac{x_k - x_{k-1}}{x_k}\right| \leq \epsilon \tag{3.2}$$

$$|f(x_k)| \leq \epsilon \tag{3.3}$$

onde ϵ é a precisão/tolerância fornecida como parâmetro para o processo iterativo.

As vezes não é possível atender a todos os critérios ao mesmo tempo. Observe as situações da Figura 3.4: no primeiro caso (à esquerda) note que $|f(x_k)| < \epsilon$, mas $|\alpha - x_k| > \epsilon$; no segundo caso (à direita) note que o oposto ocorre $|\alpha - x_k| < \epsilon$, mas $|f(x_k)| > \epsilon$.

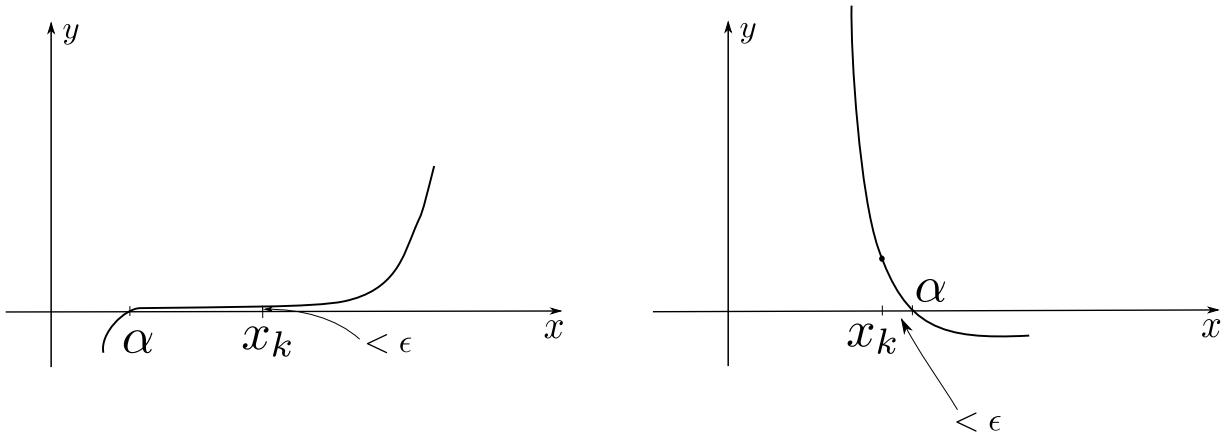


Figura 3.4: Diferentes situações e comportamentos dos critérios de convergência.

3.1.6 Ordem de convergência

É importante definir com qual rapidez a sequência de aproximações $\{x_0, x_1, \ldots\}$ converge para a raiz exata α.

Definição 3 (Ordem de convergência)**.** Uma sequência $\{x_n | n \geq 0\}$ é dita convergir com ordem $p \geq 1$ para um ponto α se

$$|\alpha - x_{n+1}| \leq c|\alpha - x_n|^p, \quad n \geq 0 \tag{3.4}$$

para uma constante $c > 0$.

Sendo $c < 1$, dizemos que :

- se $p = 1$: convergência linear
- se $1 < p < 2$: convergência super-linear
- se $p = 2$: convergência quadrática

Na prática o que isso significa? Considere a equação

$$|\alpha - x_{n+1}| \leq c|\alpha - x_n|^p$$

para os seguintes casos:

- Linear: $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, \ldots$ com $c = 10^{-1}$
- Linear: $10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, \ldots$ com $c = 10^{-2}$
- Super-linear: $10^{-2}, 10^{-3}, 10^{-5}, 10^{-8}, \ldots$
- Quadrática: $10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, \ldots$

3.1.7 Estimando a ordem de convergência

Da definição de ordem de convergência, denotando o erro por $e_k = x_k - \alpha$, temos que

$$\begin{aligned} |e_{k+1}| &= c|e_k|^p \\ |e_k| &= c|e_{k-1}|^p \quad \Rightarrow \quad \frac{e_{k+1}}{e_k} = \left(\frac{e_k}{e_{k-1}}\right)^p \end{aligned}$$

Podemos obter uma aproximação para a ordem de convergência p aplicando logaritmo em ambos os lados

$$\begin{aligned} \log\left(\frac{e_{k+1}}{e_k}\right) &= \log\left(\left[\frac{e_k}{e_{k-1}}\right]^p\right) \\ \log\left(\frac{e_{k+1}}{e_k}\right) &= p\log\left(\frac{e_k}{e_{k-1}}\right) \quad \Rightarrow \quad p = \frac{\log\left(\frac{e_{k+1}}{e_k}\right)}{\log\left(\frac{e_k}{e_{k-1}}\right)} \end{aligned}$$

Em termos práticos a ordem de convergência p pode ser determinada através de:

$$p = \frac{\log\left(\frac{x-x_0}{x_0-x_1}\right)}{\log\left(\frac{x_0-x_1}{x_1-x_2}\right)}. \tag{3.5}$$

3.2 Método da bissecção

A idéia fundamental do método da bissecção consiste em usar repetidamente o Teorema 1. O método subdivide o intervalo $[a, b]$ ao meio a cada iteração e seleciona o subintervalo que contem a raiz.

De acordo com o Teorema 1, o subintervalo que contem a raiz é aquele em que $f(x)$ tem sinais opostos nos extremos. A cada passo o intervalo é dividido ao meio:

$$m = \frac{a+b}{2} \tag{3.6}$$

então o novo intervalo será aquele que contém a raiz:

- $[a, m]$, se $f(a)f(m) < 0$
- $[m, b]$, caso contrário

A busca continua até que o **critério de parada** escolhido seja satisfeito considerando m como aproximação para a raiz. A Figura 3.5 ilustra um exemplo e a escolha dos intervalos contendo a raiz a cada passo do método.

Seja a_k e b_k os extremos do intervalo no passo k e seja ainda x_k o ponto médio e uma aproximação para a raiz. A cada iteração o método calcula o ponto médio do intervalo

$$x_k = \frac{a_k + b_k}{2} \tag{3.7}$$

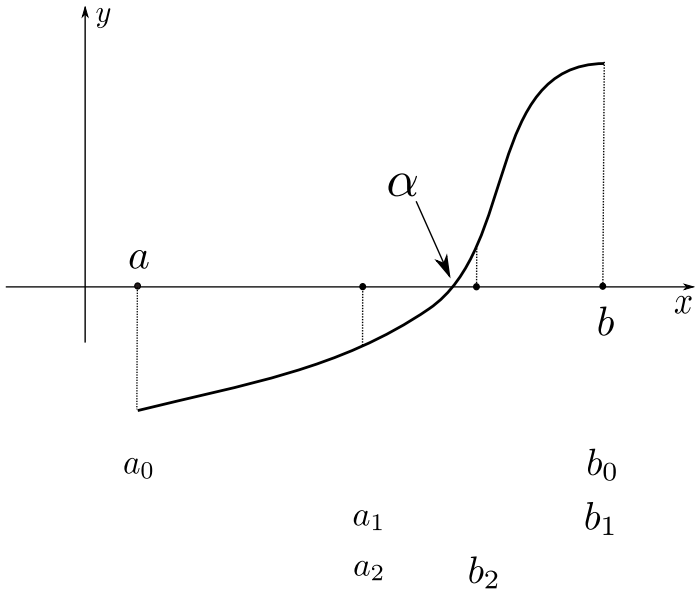


Figura 3.5: Exemplo do método da bisseção.

Podemos considerar $f(x_k) \neq 0$, caso contrário teríamos encontrado a raiz. Sendo assim o método agora calcula $f(x_k)$ e decide o novo subintervalo $[a_{k+1}, b_{k+1}]$ da seguinte forma

$$\text{se } f(a_k)f(x_k) \begin{cases} < 0, & \text{então } a_{k+1} = a_k \quad e \quad b_{k+1} = x_k \\ > 0, & \text{então } a_{k+1} = x_k \quad e \quad b_{k+1} = b_k \end{cases}$$

Exemplo 25: Exemplo

O método da bissecção aplicado à equação $f(x) = \left(\frac{x}{2}\right)^2 - \sin(x) = 0$ com intervalo inicial $[1.5, 2.0]$, gera a seguinte sequência de aproximações:

Solução: A resolução pelo método da bisseção pode ser resumida em uma tabela como segue:

k	a	b	x_k	$f(a)$	$f(b)$	$f(x_k)$
0	1.5	2.0	1.75	-0.4349	0.0907	-0.2184
1	1.75	2.0	1.875	-0.2184	0.0907	-0.0752
2	1.875	2.0	1.9375	-0.0752	0.0907	0.0050
3	1.875	1.9375	1.90625	-0.0752	0.0050	-0.0358
4	1.90625	1.9375	1.921875	-0.0358	0.0050	-0.0156
5	1.921875	1.9375	1.929688	-0.0156	0.0050	-0.0054
6	1.929688	1.9375	**1.933594**	-0.0054	0.0050	-0.0002

E assim conclui-se que a raiz pode ser aproximada por 1.933594, isto é, o ponto médio obtido na última iteração.

3.2.1 Algoritmo do método da bisseção

O método da bisseção pode ser implementado em um linguagem de programação de forma simples através do algoritmo apresentado a seguir.

Algorithm 1: Método da bisseção

entrada: função $f(x)$ contínua
função tal que $f(a)f(b) < 0$ em $[a, b]$
precisao ϵ

$k = 0$;
enquanto *critério de parada não for satisfeito* **faça**
  $x_k = \frac{(a+b)}{2}$;
  se $f(a)f(x_k) < 0$ **então**
    $b = x_k$;
  senão
    $a = x_k$;
  fim-se
fim-enquanto
retorne x_k;

3.2.2 Análise do método da bissecção

A cada iteração k a raiz α de $f(x) = 0$ está no intervalo $[a_k, b_k]$. Temos assim a seguinte relação para o erro

$$|\alpha - x_k| \leq \tfrac{1}{2}(b_k - a_k) \tag{3.8}$$

O tamanho do intervalo $(b_k - a_k)$ no passo k pode ser escrito como

$$b_k - a_k = \frac{b_{k-1} - a_{k-1}}{2} = \frac{b_{k-2} - a_{k-2}}{2^2} = \ldots = \frac{b_1 - a_1}{2^{k-1}} = \frac{b_0 - a_0}{2^k} \tag{3.9}$$

Portanto, o erro no passo k satisfaz

$$|\alpha - x_k| \leq \frac{b_0 - a_0}{2^{k+1}} \tag{3.10}$$

onde $a_0 = a$ e $b_0 = b$.

Uma propriedade interessante do método da bissecção é que a convergência é **garantida** se $f(x)$ for contínua em $[a, b]$ e se $\alpha \in [a, b]$. Também é possível determinar o número de iterações que serão necessárias para calcular a raiz com uma certa precisão ϵ, ou seja, quantas iterações são necessárias para que o erro entre a aproximação x_k da raiz α seja menor do que ϵ.

Para isso, basta encontrar o inteiro k tal que:

$$|\alpha - x_k| \leq \frac{b_0 - a_0}{2^{k+1}} \leq \epsilon \tag{3.11}$$

portanto

$$\begin{aligned}
\frac{b_0 - a_0}{2^{k+1}} &\leq \epsilon \\
\frac{b_0 - a_0}{\epsilon} &\leq 2^{k+1} \\
\log_2\left(2^{k+1}\right) &\geq \log_2\left(\frac{b_0 - a_0}{\epsilon}\right) \\
k + 1 &\geq \log_2\left(\frac{b_0 - a_0}{\epsilon}\right)
\end{aligned}$$

que resulta em

$$k \geq \frac{\ln\left(\frac{b_0 - a_0}{\epsilon}\right)}{\ln(2)} - 1 \tag{3.12}$$

Exemplo 26

Qual o número de iterações necessárias para encontrar uma aproximação para a raiz de $f(x) = \left(\frac{x}{2}\right)^2 - \sin(x)$ no intervalo $[1.5, 2.0]$ com uma precisão $\epsilon = 10^{-5}$?

Solução: Precisamos encontrar k que satisfaz

$$\begin{aligned}
k &\geq \frac{\ln\left(\frac{b_0 - a_0}{\epsilon}\right)}{\ln(2)} - 1 \\
k &\geq \frac{\ln\left(\frac{2 - 1.5}{10^{-5}}\right)}{\ln(2)} - 1 \approx 15.61 - 1 = 14.61
\end{aligned}$$

Portanto, como k deve ser inteiro, temos que depois de 15 iterações o método atinge a precisão de 10^{-5} como desejado.

3.2.3 Ordem de convergência do método da bisseção

Para o método da bissecção lembre-se que

$$|\alpha - x_k| \leq \frac{b_0 - a_0}{2^{k+1}}$$

portanto, concluimos que a ordem de convergência para o método da bissecção é linear pois $p = 1$ e que a constante é $c = \frac{1}{2}$ (veja a definição de ordem de convergência e a Equação (3.4)).

Isto é

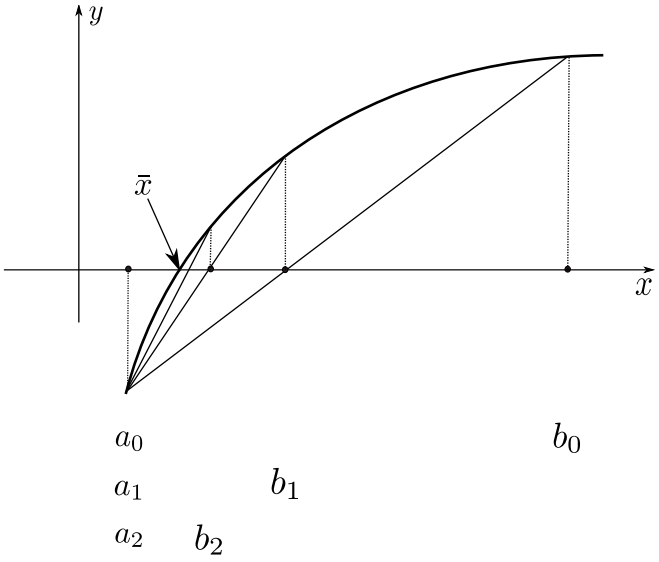
$$\frac{|\alpha - x_{k+1}|}{|\alpha - x_k|} \leq \frac{1}{2} \tag{3.13}$$

que nos diz que em média o erro cai pela metade a cada iteração do método.

3.3 Método da falsa posição

No método da bissecção a cada iteração calculamos o ponto médio do intervalo como aproximação para a raiz e então decidimos qual o próximo intervalo a continuar a busca pela raiz.

No método da falsa posição a aproximação para a raiz é dada pelo ponto x_k escolhido como sendo o zero da reta que passa pelos pontos $(a_k, f(a_k))$ e $(b_k, f(b_k))$. De forma análoga ao método da bissecção a cada iteração o método encontra um intervalo que contem a raiz e continua o processo de busca nesse intervalo. A Figura **??** ilustra a ideia básica do método.



A equação da reta que passa pelos pontos $(a_k, f(a_k))$ e $(b_k, f(b_k))$ é dada por

$$\boxed{g(x) = mx + n} \quad \Rightarrow f(a) = ma + n$$
$$\Rightarrow f(b) = mb + n$$

$$f(b) - f(a) = mb - ma \quad \Rightarrow \quad m = \frac{f(b) - f(a)}{b - a}$$

$$f(b) = \frac{f(b) - f(a)}{b - a} b + n \quad \Rightarrow \quad n = f(b) - \left[\frac{f(b) - f(a)}{b - a}\right] b$$

Assim

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

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

Queremos encontrar x tal que $g(x) = 0$, logo

$$\begin{aligned} g(x) &= 0 \\ \frac{f(b)-f(a)}{b-a}x + f(b) - \frac{f(b)-f(a)}{b-a}b &= 0 \\ \frac{f(b)-f(a)}{b-a}x + f(b) &= \frac{f(b)-f(a)}{b-a}b \\ \frac{f(b)-f(a)}{b-a}x &= \frac{f(b)-f(a)}{b-a}b - f(b) \\ x &= b - f(b)\frac{b-a}{f(b)-f(a)} \\ x &= \frac{af(b)-bf(a)}{f(b)-f(a)} \end{aligned}$$

Portanto no passo k calculamos a próxima aproximação x_k usando

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

Sendo assim, dado um intervalo $[a, b]$, o método da falsa posição pode ser descrito pelo seguinte processo:

1. calcule o ponto de interseção x_k da reta que passa por $(a_k, f(a_k))$ e $(b_k, f(b_k))$ com o eixo x usando
$$\boxed{x_k = \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)}}$$
2. selecione um novo intervalo para continuar com a busca
3. o novo intervalo será dado por
   - $[a_k, x_k]$, se $f(a_k)f(x_k) < 0$
   - $[x_k, b_k]$, caso contrário
4. o processo continua até satisfazer o critério de parada

Exemplo 27

Encontrar o zero de $f(x) = (x/2)^2 - \sin(x)$ usando o seguinte intervalo $[a, b] = [1.5, 2]$. Use $|f(x_k)| < \epsilon$ como critério de parada para $\epsilon = 0.0001$.

Solução: a aplicação do método da falsa posição pode ser resumida na tabela abaixo.

k	a	b	x	$f(a)$	$f(b)$	$f(x)$
0	1.5	2.0	1.913731	-4.349950e-01	9.070e-02	-2.618006e-02
1	1.913731	2.0	1.933054	-2.618006e-02	9.070e-02	-9.243996e-04
2	1.933054	2.0	1.933730	-9.243996e-04	9.070e-02	-3.193009e-05
3	1.933730	2.0	1.933753	-3.193009e-05	9.070e-02	-1.102069e-06
4	1.933753	2.0	1.933754	-1.102069e-06	9.070e-02	-3.903695e-08

Algorithm 2: Método da Falsa Posição

entrada: função f contínua em $[a, b]$, intervalo $[a, b]$ tal que $f(a)f(b) < 0$, precisao ϵ e número máximo de iterações *maxit*

```
xold = b;
para k de 1 até maxit faça
    x = (af(b)−bf(a))/(f(b)−f(a));
    se abs(x-xold) < ε então
        retorne x;
    fim-se
    xold = x;
    se f(a)f(x) < 0 então
        b = x;
    senão
        a = x;
    fim-se
fim-para
```

O método termina com $x = 1.933754$ como aproximação para o zero desta função.

3.3.1 Convergência do método da falsa posição

Não iremos apresentar a análise de convergência do método da falsa posição. Entretanto, cabe dizer que se as condições do método forem satisfeitas, isto é, se

- $f(x)$ for contínua no intervalo $[a, b]$ e
- $f(a)f(b) < 0$

então o método apresenta convergência de primeira ordem. Mais detalhes podem ser encontrados no livro **"Algoritmos Numéricos"** do Frederico F. Campos.

3.4 Método do ponto fixo

Para encontrar a raiz da equação

$$f(x) = 0 \tag{3.14}$$

onde f é uma função contínua no intervalo $[a, b]$ que procuramos a raiz, iremos expressar a equação (3.14) da seguinte forma:

$$x = \phi(x) \tag{3.15}$$

de forma que a solução de (3.15) também seja solução de (3.14). Para qualquer função $\phi(x)$, qualquer solução de (3.15) é chamada de **ponto fixo** de $\phi(x)$. Sendo assim temos a seguinte equivalência: problema de determinar o zero de $f(x) \leftrightarrow$ problema de determinar o ponto fixo de $\phi(x)$.

Como exemplo considere a função $f(x) = x^2 - x - 2 = 0$, a qual pode ser escrita na forma de ponto fixo como:

a) $x = x^2 - 2$

b) $x = \sqrt{2 + x}$

c) $x = 1 + \frac{2}{x}$

d) $x = \frac{x^2+2}{2x-1}$

Existem diversas formas de expressar $f(x) = 0$ como um **problema de ponto fixo** da forma $x = \phi(x)$, entretanto veremos que nem todas são satisfatórias para nossos objetivos.

Sendo, assim temos a seguinte situação para os problemas:

- Zero de função: qual o valor de x tal que $f(x) = 0$?
- Ponto fixo: qual o valor de x tal que $x = \phi(x)$?

A Figura 3.6 apresenta uma interpretação para esses problemas tomando como exemplo para a função $f(x) = \cos(x) - x$.

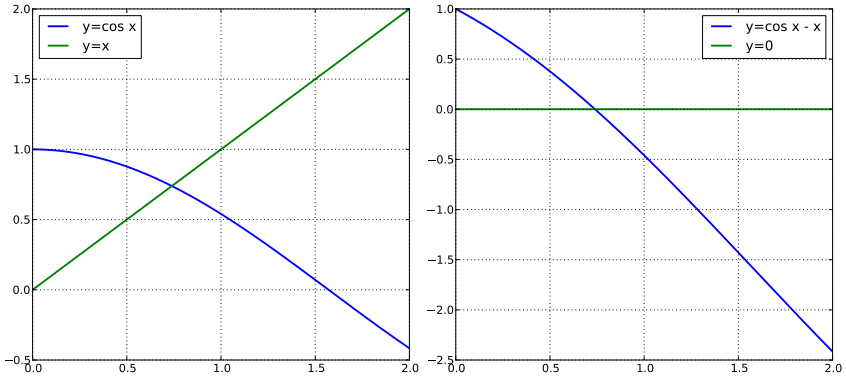


Figura 3.6: (Esquerda) Problema de ponto fixo; (Direita) Problema de zero de função.

No que segue relativo ao método do ponto fixo iremos considerar que:

- estas curvas se interceptam (existe pelo menos 1 solução)
- $\phi(x)$ e $\phi'(x)$ são contínuas no intervalo $[a, b]$

Seja x_0 uma aproximação inicial para α. O método do ponto fixo obtem aproximações sucessivas x_k para α, usando o seguinte processo iterativo

$$x_{k+1} = \phi(x_k), \quad k = 0, 1, \ldots \tag{3.16}$$

Ou seja, dado uma aproximação x_k, calculamos o valor de $\phi(x_k)$ como aproximação para a raiz. Em seguida usamos esse valor como próximo argumento para a função de iteração $\phi(x)$. Repetimos o processo até que o critério de parada estabelecido seja satisfeito.

Exemplo 28

Resolver $x^2 - x - 2 = 0$ com a função de iteração $x = \sqrt{2 + x}$ usando $x_0 = 2.5$. Econtrar a raiz $\alpha = 2$.
Solução Pelo método do ponto fixo: $x_{k+1} = \phi(x_k)$, para $k = 0, 1, \ldots$ e, portanto

$$\begin{aligned} x_1 &= \phi(x_0) = \sqrt{2 + 2.5} = \sqrt{4.5} = 2.12132 \\ x_2 &= \phi(x_1) = \sqrt{2 + 2.12132} = \sqrt{4.12132} = 2.030103 \\ x_3 &= \phi(x_2) = \sqrt{2 + 2.030103} = \sqrt{4.030103} = 2.007511, \quad \ldots \end{aligned}$$

As aproximações x_k convergem para a $\alpha = 2$. Entretanto, para certas escolhas da função de iteração $\phi(x)$ o processo iterativo diverge.

Exemplo 29

Considere o mesmo problema do exemplo anterior, entretanto agora com o seguinte esquema de ponto fixo: $x = x^2 - 2$ com $x_0 = 2.5$ como aproximação inicial.
Solução: Pelo método do ponto fixo: $x_{k+1} = \phi(x_k)$, para $k = 0, 1, \ldots$ temos

$$\begin{aligned} x_1 &= \phi(x_0) = x_0^2 - 2 = 6.25 - 2 = 4.25 \\ x_2 &= \phi(x_1) = x_1^2 - 2 = 18.0625 - 2 = 16.0625 \\ x_3 &= \phi(x_2) = x_2^2 - 2 = 258.00 - 2 = 256.00 \\ &\ldots \end{aligned}$$

que como vemos **diverge** rapidamente da raiz procurada.

Vamos analisar graficamente o que acontece com cada uma das opções, isto é, se o método converge ou diverge para cada escolha de $\phi(x)$.

No que segue

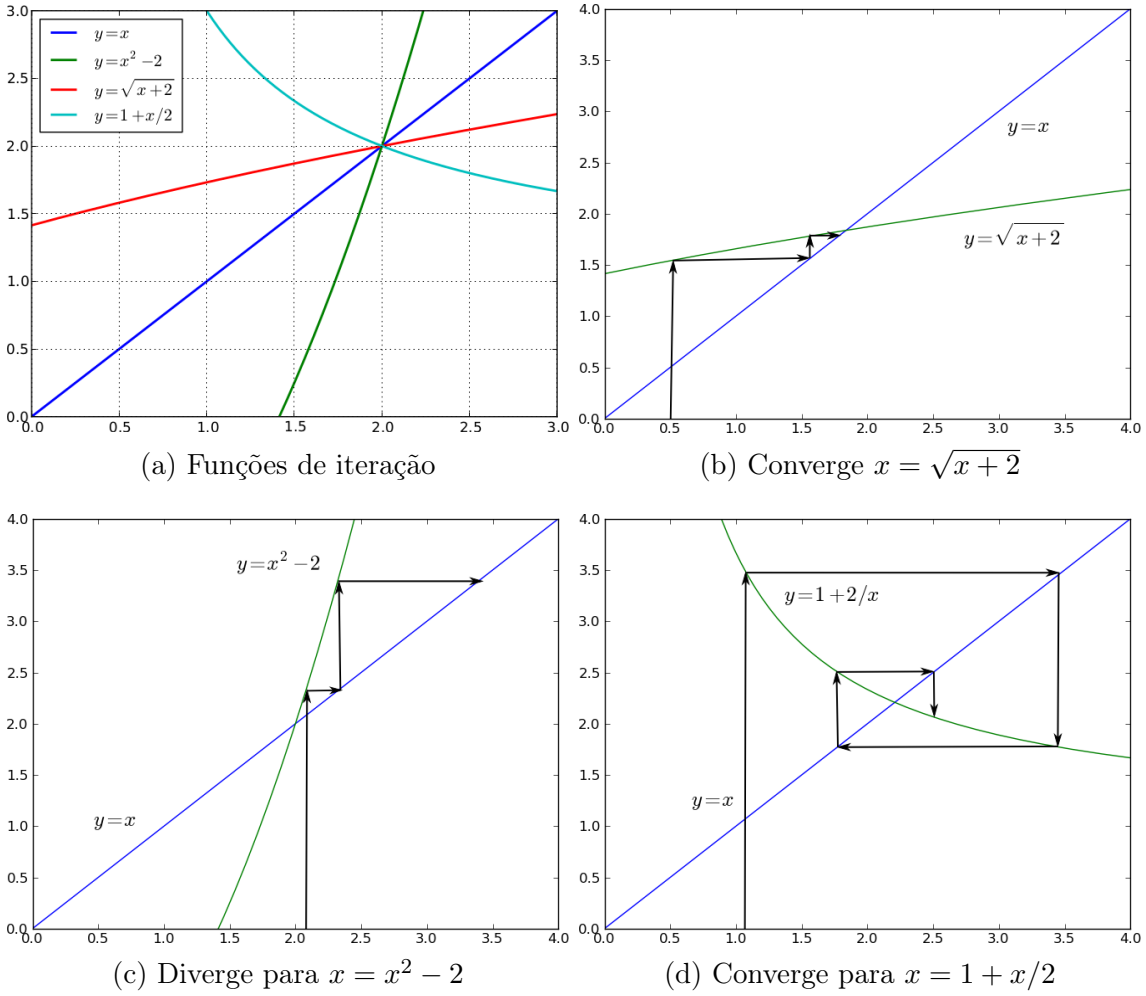
- a seta vertical corresponde à avaliação da função em um ponto
- a seta horizontal apontando para $y = x$ indica que o resultado da avaliação anterior é usado como entrada para a próxima

Para $f(x) = x^2 - x - 2$ e para as funções de iteração $\phi(x)$ listadas anteriormente, graficamente temos

Para finalizar, o Algoritmo 3 apresenta o pseudo-código que pode ser usado para implementar o método do ponto fixo, considerando uma função de iteração ϕ qualquer.

3.4.1 Convergência do método do ponto fixo

Antes de estudar a convergência do método do ponto fixo vamos recapitular alguns conceitos importantes que serão usados na demonstração do resultado que garante a convergência (ou não) do método.



(a) Funções de iteração

(b) Converge $x = \sqrt{x+2}$

(c) Diverge para $x = x^2 - 2$

(d) Converge para $x = 1 + x/2$

Figura 3.7: Comportamento do método do ponto fixo

Teorema 5 (TVM - Teorema do Valor Médio)**.** Se f é contínua em $[a, b]$ e diferenciável em (a, b), então existe pelo menos um ponto ξ entre a e b, tal que:

$$f'(\xi) = \frac{f(b) - f(a)}{b - a} \quad \Rightarrow \quad f(b) - f(a) = f'(\xi)(b - a)$$

Teorema 6. Seja f uma função real **contínua** na vizinhança de x_0. Se $f(x_0) \neq 0$, então $f(x) \neq 0$ para todo x numa vizinhança pequena de x_0.

Iremos estudar agora as condições suficientes que a função de iteração $\phi(x)$ deve satisfazer para garantir a convergência do método do ponto fixo.

Teorema 7 (Ponto fixo)**.** Seja $\phi(x)$ uma função contínua com $\phi'(x)$ contínua num intervalo fechado $I = (\alpha - h, \alpha + h)$, cujo centro α é a solução de $x = \phi(x)$. Seja $x_0 \in I$ e seja M um limitante em I para $\phi'(x)$, isto é,

$$|\phi'(x)| \leq M < 1.$$

Então:

Algorithm 3: Algoritmo do método do ponto fixo

entrada: função de iteração $\phi(x)$,
aproximação inicial x_0,
precisão ϵ
número máximo de iterações *maxit*

para k *de* 1 *até maxit* **faça**
  $x_1 = \phi(x_0)$;
  se $abs(x_1 - x_0) < \epsilon$ **então**
    retorne x_1 ;
  fim-se
  $x_0 = x_1$;
fim-para

a) a iteração $x_{k+1} = \phi(x_k)$, $k = 0, 1, \ldots$ pode ser executada indefinidamente, pois $x_k \in I, \forall k$;

b) $|x_k - \alpha| \to 0$.

Prova 2. Vamos mostrar primeiro que o item a) é válido e depois iremos mostrar que b) também é válido.
a) Para provar que $x_k \in I, \forall k$, iremos usar **indução matemática**.

i) Por hipótese $x_0 \in I$.

ii) Supomos que $x_0, x_1, \ldots, x_k \in I$.

iii) Vamos mostrar que $x_{k+1} \in I$.
Temos

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha)$$

Pelo TVM temos

$$\phi(x_k) - \phi(\alpha) = \phi'(\xi_k)(x_k - \alpha) = x_{k+1} - \alpha$$

onde ξ_k está entre x_k e α.

Tomando o módulo segue que:

$$\begin{aligned} |x_{k+1} - \alpha| &= |\phi'(\xi_k)(x_k - \alpha)| \\ &= |\phi'(\xi_k)||x_k - \alpha| \end{aligned}$$

Pela hipótese de indução: $x_k \in I \Rightarrow \xi_k \in I$.
E ainda como $|\phi'(x)| \leq M < 1$ em I temos

$$|x_{k+1} - \alpha| \leq M|x_k - \alpha|$$

Como $M < 1$, temos que $x_{k+1} \in I$. E assim concluimos que $x_k \in I$ para todo k.
b) Pelo item a), temos que

$$\begin{aligned} |x_k - \alpha| \leq M|x_{k-1} - \alpha| &\leq M^2|x_{k-2} - \alpha| \\ &\leq \ldots \leq M^k|x_0 - \alpha| \end{aligned}$$

como $M < 1$, aplicando o limite

$$\lim_{k\to\infty} M^k \to 0 \quad \Rightarrow \quad |x_k - \alpha| \to 0$$

Ou seja, $\{x_k\}$ converge para a raiz α.
□

Algumas observações sobre o resultado do Teorema do Método do Ponto Fixo:

- Se $|\phi'(\alpha)| < 1$, então existe um intervalo $I \subseteq [a, b]$ centrado em α que satisfaz as condições do Teorema do ponto fixo. Portanto, a iteração $x_{k+1} = \phi(x_k)$ irá **convergir**.
- Se $|\phi'(\alpha)| > 1$, então a iteração $x_{k+1} = \phi(x_k)$ irá **divergir**.
- Se $|\phi'(\alpha)| = 1$, nada pode ser dito a respeito da convergência do método.
- Fica a partir da demonstração do item b), que quanto menor for o valor de M, mais rápida será a convergência de $\{x_k\}$ para α.

Exemplo 30

O método do ponto fixo converge para $f(x) = x^2 - x - 2$ no intervalo $I = [1.5, 2.5]$ usando $\phi(x) = \sqrt{x+2}$?

Solução: Derivando $\phi(x)$ tem-se: $\phi'(x) = \frac{1}{2\sqrt{x+2}}$. Para mostrar que o MP converge, precisamos encontrar um limitante M para $\phi'(x)$ tal que $M < 1$, isto é

$$\max_{x\in I} |\phi'(x)| = \max_{x\in I} \left| \frac{1}{2\sqrt{x+2}} \right| = 0.267 < 1$$

Portanto, nessas condições, o Teorema do Ponto fixo garante a convergência do método.

Exemplo 31

O método do ponto fixo converge para $f(x) = x^2 - x - 2$ no intervalo $I = [1.5, 2.5]$ usando $x = x^2 - 2$?

Solução: Derivando $\phi(x)$ tem-se $\phi(x) = x^2 - 2 \quad \Rightarrow \quad \phi'(x) = 2x$. Assim temos que $\phi(x)$ e $\phi'(x)$ são contínuas. Entretanto

$$\max_{x\in I} |\phi'(x)| = \max_{x\in I} |2x| > 1$$

que nos mostra que o método do ponto fixo não converge para essa escolha da função de iteração $\phi(x)$. De fato, o método diverge (como visto anteriormente).

3.4.2 Ordem de convergência do método do ponto fixo

Do Teorema temos que

$$x_{k+1} - \alpha = \phi'(\xi_k)(x_k - \alpha)$$

para algum ξ_k entre x_k e α. Logo

$$\frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} = |\phi'(\xi_k)| \leq M$$

E portanto pela def. de ordem de convergência

$$|x_{k+1} - \alpha| \leq M|x_k - \alpha|$$

temos que $p = 1$ e dizemos então que a convergência do MPF é linear.
E ainda, o erro em qualquer iteração é proporcional ao erro da iteração anterior e a constante de proporcionalidade é dada por $\phi'(\xi_k)$.

Exemplo 32

Considere a equação $f(x) = 2x^2 - 5x + 2 = 0$, cujas raízes são $\alpha_1 = 0.5$ e $\alpha_2 = 2$. Considere os processos iterativos:

a) $x_{k+1} = \sqrt{\frac{5x_k}{2} - 1}$

b) $x_{k+1} = \frac{2x_k^2+2}{5}$

Qual dos dois processos você utilizaria para obter a raiz α_1? Porque?

Solução: Para a) temos que

$$\phi(x) = \left(\tfrac{5x}{2} - 1\right)^{1/2} \quad \Rightarrow \quad \phi'(x) = \frac{1}{2}\frac{1}{\sqrt{\frac{5x_k}{2}-1}}\frac{5}{2}$$

$$|\phi'(\alpha_1)| = \frac{5}{4\sqrt{\frac{5\cdot 0.5}{2} - 1}} = 2.5 > 1$$

Para b) temos que

$$\phi(x) = \frac{2x^2 + 2}{5} \quad \Rightarrow \quad \phi'(x) = \frac{4x}{5}$$

$$|\phi'(\alpha_1)| = \frac{4(0.5)}{5} = \frac{2}{5} = 0.4 < 1$$

Temos então que $\phi(x)$ e $\phi'(x)$ são contínuas e se x_0 for suficientemente próximo de α_1, então o processo b) irá convergir, e portanto este é mais adequado para encontrar a raiz.

Exemplo 33

Seja $f(x) = x^3 - 9x + 3$. Considere a seguinte função de iteração $x = \phi(x) = \frac{x^3+3}{9}$. Queremos encontrar a raiz de $f(x) = 0$ no intervalo $[0, 1]$. O método irá convergir?
Solução: temos que $\phi'(x) = \frac{x^2}{3}$, e portanto temos que $\phi(x)$ e $\phi'(x)$ são contínuas.

Verificamos agora que

$$|\phi'(x)| = \left|\frac{x^2}{3}\right| < 1, \forall x \in [0, 1]$$

E assim concluimos que o método irá convergir. Podemos verificar tomando $x_0 = 0.25$ e usando uma precisão $\epsilon = 0.001$.

Na primeira iteração temos

$$x_1 = \frac{0.25^3 + 3}{9} = \frac{0.015625 + 3}{9} = 0.335069$$
$$f(x_1) = x_1^3 - 9x_1 + 3 = 0.037618 - 3.015621 + 3 = 0.021997 > \epsilon$$

Mais uma iteração

$$x_2 = \frac{0.335069^3 + 3}{9} = 0.337513$$
$$f(x_2) = x_2^3 - 9x_2 + 3 = 0.038447 - 3.037617 + 3 = 0.00083 < \epsilon$$

Como o critério de parada $|f(x_2)| < \epsilon$ foi satisfeito, terminamos o processo com $x_2 = 0.337513$ como aproximação para a raiz.

Se tomarmos outra aproximação inicial $x_0 = 0.5$ mais distante da raiz temos os seguintes passos:

k	x_k	$f(x_k)$
0	0.5	-1.375
1	0.34722	-0.83137
2	0.33798	-0.0032529
3	0.33762	-0.00012219

Exemplo 34

Considere as seguintes funções:

a) $\phi_1(x) = 2x - 1$

b) $\phi_2(x) = x^2 - 2x + 2$

Qual delas você escolheria para obter a raiz 1, utilizando o processo iterativo $x_{k+1} = \phi(x_k)$? Exiba a sequência gerada com sua escolha tomando $x_0 = 1.2$.

Solução: temos que $\phi_1(x)$ e $\phi_1'(x)$ são contínuas pois

$$\phi_1(x) = 2x - 1, \quad \phi_1'(x) = 2$$

Mas $|\phi_1'(x)| = 2 > 1, \quad \forall x$ próximo de $\alpha = 1$.
Por outro lado

$$\phi_2(x) = x^2 - 2x + 2, \quad \phi_2'(x) = 2x - 2$$
$$|\phi_2'(x)| = |2x - 2| < 1$$

de onde temos

$$-1 < 2x - 2 < 1$$
$$1 < 2x < 3$$
$$\tfrac{1}{2} < x < \tfrac{3}{2}$$

Portanto $|\phi_2'(x)| < 1$ se e somente se $x \in I = [0.5, 1.5]$. Como $\phi_2(x)$ e $\phi_2'(x)$ são contínuas, tomando uma aproximação inicial $x_0 \in I$, temos a convergência do método garantida.

Exemplo 35

Vamos rever o caso $f(x) = x^2 - x - 2 = 0$. Temos os seguinte esquema, que já vimos que converge para $x = \sqrt{2 + x}$. Vamos usar $x_0 = 2.5$, então

$$x_1 = \phi(x_0) = 2.121320$$
$$x_2 = \phi(x_1) = 2.030104$$
$$x_3 = \phi(x_2) = 2.007512$$
$$x_4 = \phi(x_3) = 2.001877$$
$$x_5 = \phi(x_4) = 2.000469, \ldots$$

Vejamos agora o seguinte esquema

$$x = x - \frac{x^2 - x - 2}{2x - 1}$$

com $x_0 = 2.5$ obtemos

$$x_1 = \phi(x_0) = 2.062500$$
$$x_2 = \phi(x_1) = 2.001250$$
$$x_3 = \phi(x_2) = 2.000001, \ldots$$

de onde concluimos que o esquema converge, e ainda, de forma muito mais rápida que o esquema anterior. Porque? Que função de iteração é essa? Veremos a seguir.

3.5 Método de Newton

Na aula anterior estudamos o método do ponto fixo que expressa $f(x) = 0$ na forma

$$x = \phi(x)$$

Uma forma geral de escrever $\phi(x)$ é

$$\phi(x) = x + A(x)f(x) \tag{3.17}$$

para qualquer $A(x)$ tal que $A(\alpha) \neq 0$. Vamos estudar agora o método de Newton que é uma das técnicas mais populares para se determinar raízes de equações não lineares.

Existem diversas formas de se deduzir o método de Newton

1. método de ponto fixo
2. interpretação geométrica
3. série de Taylor

No MPF vimos que

- se $\phi(x)$ e $\phi'(x)$ forem contínuas e se $|\phi'(x)| < 1, \forall x \in I$, então o método irá convergir
- a convergência será mais rápida quanto menor for $|\phi'(\alpha)|$.

A idéia do método de Newton, quando visto como um MPF, é tentar garantir e acelerar a convergência do MPF escolhendo a função de iteração $\phi(x)$ de tal forma que $\phi'(\alpha) = 0$. Partindo da forma geral de $\phi(x)$ dada em (3.17), queremos encontrar a função $A(x)$ tal que $\phi'(\alpha) = 0$.

Derivando

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

obtemos

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

avaliando em $x = \alpha$ temos

$$\phi'(\alpha) = 1 + A'(\alpha)\underbrace{f(\alpha)}_{=0} + A(\alpha)f'(\alpha)$$

Assim

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

Como queremos que $\phi'(\alpha) = 0$, fazemos

$$1 + A(\alpha)f'(\alpha) = 0 \quad \Rightarrow \quad A(\alpha) = -\frac{1}{f'(\alpha)}, \quad \text{com } f'(\alpha) \neq 0$$

Tomando

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

temos

$$\phi(x) = x + A(x)f(x) \quad \Rightarrow \quad \boxed{\phi(x) = x - \frac{f(x)}{f'(x)}}$$

e assim o processo iterativo do método de Newton fica definido como

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots}$$

Assim dada $f(x) = 0$, obtemos a função de iteração

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

que é tal que $\phi'(\alpha) = 0$, pois

$$\begin{aligned}\phi'(x) &= 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{f'(x)^2} \\ &= \frac{f'(x)^2 - f'(x)f'(x) + f(x)f''(x)}{f'(x)^2} \\ &= \frac{f(x)f''(x)}{f'(x)^2}\end{aligned}$$

como $f(\alpha) = 0$, isto implica que $\phi'(\alpha) = 0$, desde que $f'(\alpha) \neq 0$, como queríamos.

Método de Newton

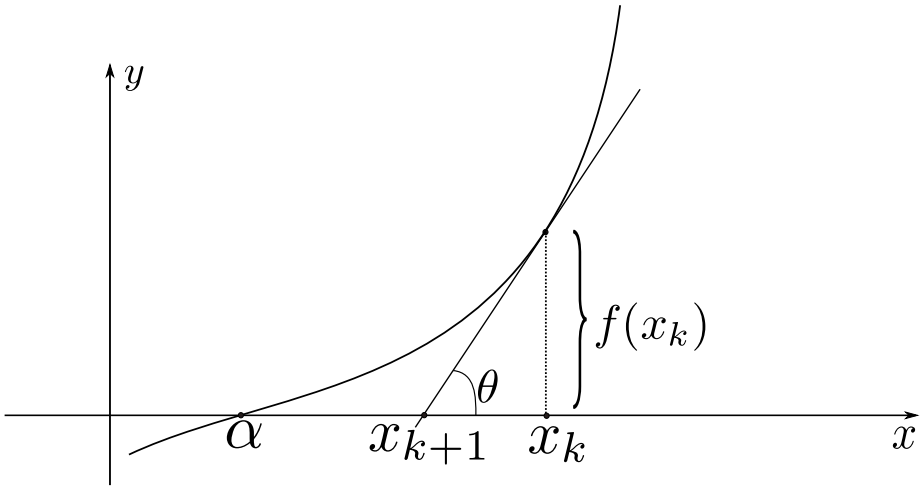
Interpretação geométrica

Antes de estudarmos exemplos, algoritmo e convergência, vejamos as outras deduções do método.

Iremos apresentar o método de Newton agora sob o ponto de vista geométrico. Seja x_k uma aproximação para a raiz α de $f(x) = 0$.

O valor de x_{k+1} é obtido graficamente traçando-se pelo ponto $(x_k, f(x_k))$ a reta tangente à curva $y = f(x)$.

O ponto de interseção da reta tangente com o eixo dos x, determina o valor de x_{k+1}.



Assim temos

$$\begin{aligned}\tan(\theta) = f'(x_k) &= \frac{f(x_k)}{x_k - x_{k+1}} \\ \Rightarrow f'(x_k)(x_k - x_{k+1}) &= f(x_k) \\ \Rightarrow x_k - x_{k+1} &= \frac{f(x_k)}{f'(x_k)} \\ \Rightarrow x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)}\end{aligned}$$

Vejamos agora a dedução através de série de Taylor.

Método de Newton

Série de Taylor Vamos deduzir o método de Newton usando série de Taylor em torno do ponto $a = x_k$.

Assim temos

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + R_1(x)$$

onde, como visto anteriormente, $R_1(x) = \frac{x-x_k}{2}f''(c_{x_k})$, com c_{x_k} entre x_k e x.

Avaliando a expressão anterior em $x = \alpha$ e desprezando o termo do erro, temos

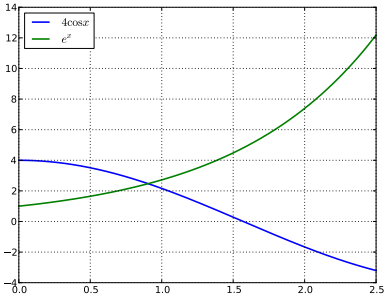
$$\begin{aligned}f(\alpha) = 0 \quad \Rightarrow \quad f(x_k) + (\alpha - x_k)f'(x_k) &\approx 0 \\ (\alpha - x_k) &\approx -\frac{f(x_k)}{f'(x_k)} \\ \alpha &\approx x_k - \frac{f(x_k)}{f'(x_k)}\end{aligned}$$

De onde definimos o método como

$$\boxed{x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}}$$

Exemplo Usando o método de Newton, determine a menor raiz positiva da equação $f(x) = 4\cos(x) - e^x = 0$ com erro inferior a $\epsilon = 10^{-2}$. Use o seguinte critério de parada $|x_{k+1} - x_k|/|x_{k+1}|$.

Solução do Exemplo Vamos fazer um gráfico para analisar e escolher um valor inicial para aproximar a raiz. Seja $y_1 = 4\cos(x)$ e $y_2 = e^x$.



Solução do Exemplo - (cont.)
Do gráfico, escolhemos $x_0 = 1$ como chute inicial. Assim

$$f(x) = 4\cos(x) - e^x \quad \Rightarrow \quad f'(x) = -4\sin(x) - e^x$$

Portanto temos

$$x_{k+1} = x_k - \frac{4\cos(x_k) - e^{x_k}}{(-4\sin(x_k) - e^{x_k})}$$

Passo 1

$$\begin{aligned} x_1 &= x_0 - \frac{4\cos(x_0) - e^{x_0}}{(-4\sin(x_0) - e^{x_0})} = 1 - \frac{4\cos(1) - e^1}{(-4\sin(1) - e^1)} \\ &= 1 - \frac{(-0.557)}{(-6.048)} = 0.908 \\ e_1 &= \frac{|x_1 - x_0|}{|x_1|} = 0.101 > \epsilon \end{aligned}$$

Solução do Exemplo - (cont.)
Passo 2

$$\begin{aligned} x_2 &= x_1 - \frac{4\cos(x_1) - e^{x_1}}{(-4\sin(x_1) - e^{x_1})} = 0.908 - \frac{4\cos(0.908) - e^{0.908}}{(-4\sin(0.908) - e^{0.908})} \\ &= 0.908 - \frac{(-0.019)}{(-5.631)} = 0.905 \\ e_2 &= \frac{|x_2 - x_1|}{|x_2|} = \frac{|0.905 - 0.908|}{|0.905|} = 0.0033 < \epsilon \end{aligned}$$

Portanto obtemos $x_2 = 0.905$ como aproximação para α com uma precisão de 10^{-2}.
□

Método de Newton

Algoritmo

entrada: $f(x)$ e sua derivada $f'(x)$, aproximação inicial x_0, precisão ϵ e número máximo de iterações *maxit*

para k *de* 1 *até maxit* **faça**
  $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$;
  se $abs(x_1 - x_0) < \epsilon$ **então**
    retorne x_1;
  fim-se
  $x_0 = x_1$;
fim-para

Exemplo Resolva $f(x) = x - x^{1/3} - 2 = 0$ usando $x_0 = 3$ como aproximação inicial.
Solução do Exemplo Temos que a derivada é

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

nesse caso a fórmula de iteração é

$$x_{k+1} = x_k - \frac{x_k - x_k^{1/3} - 2}{1 - \frac{1}{3}x_k^{-2/3}}$$

Solução do Exemplo
Aplicando o método temos

k	x_k	$f'(x_k)$	$f(x_k)$
0	3.0	0.839750	-0.442250e+00
1	3.526644	0.856130	4.506792e-03
2	3.521380	0.855986	3.771414e-07
3	**3.52137971**	0.855986	0.00000e+00

E assim ao final das iterações obtemos $\alpha \approx x_3 = 3.52137971$ como valor aproximado para a raiz.
□

3.5.1 Convergência do método de Newton

Teorema 8 (Ref: Ruggiero, página 69)**.** Sejam $f(x)$, $f'(x)$ e $f''(x)$ contínuas em um intervalo I que contém a raiz α de $f(x) = 0$. Vamos supor que $f'(\alpha) \neq 0$.
Então existe um intervalo $\bar{I} \subset I$, contendo a raiz α, tal que se $x_0 \in \bar{I}$, a sequência $\{x_k\}$ gerada pelo método de Newton $x_{k+1} = x_k - f(x_k)/f'(x_k)$ irá convergir para a raiz.

Prova 3. O método de Newton é um MPF com $\phi(x) = x - \frac{f(x)}{f'(x)}$.
Para provar a convergência do método, basta verificar, que sob as hipóteses desse Teorema, as hipóteses do Teorema do Ponto Fixo são satisfeitas para $\phi(x)$.
Precisamos provar que existe $\bar{I} \subset I$ centrado em α tal que

(i) $\phi(x)$ e $\phi'(x)$ são contínuas em $\bar{I}$

(ii) $|\phi'(x)| < 1, \forall x \in \bar{I}$

Prova 4 (cont.)**.** Sabemos que

$$\phi(x) = x - \frac{f(x)}{f'(x)}, \quad \phi'(x) = \frac{f(x)f''(x)}{f'(x)^2}$$

Pelas hipóteses temos que

- $f'(\alpha) \neq 0$
- $f'(x)$ é contínua

Então, $f'(x) \neq 0, \forall x$ na vizinhança de α.

Sendo assim é possível obter um intervalo $I_1 \subset I$ tal que $f'(x) \neq 0, \forall x \in I_1$.
Logo, em $I_1 \subset I$, temos que $f(x)$, $f'(x)$ e $f''(x)$ são contínuas e $f'(x) \neq 0$. Portanto, concluímos que $\phi(x)$ e $\phi'(x)$ são contínuas em I_1 (pela continuidade da soma, produto e divisão). Item (i) OK!

Prova 5 (cont.)**.** Como $\phi'(x)$ é contínua em I_1 e $|\phi'(\alpha)| = 0 < 1$ (por construção do método de Newton), é possível escolher $I_2 \subset I_1$ tal que $|\phi'(x)| < 1, \forall x \in I_2$. E ainda, I_2 pode ser escolhido de forma que α esteja centrado neste intervalo.
Item (ii) OK!

Sendo assim, encontramos $I_2 \subset I$, centrado em α onde $\phi(x)$ e $\phi'(x)$ são contínuas e $|\phi'(x)| < 1, \forall x \in I_2$. Ou seja, $I_2 = \bar{I}$.
□

Em resumo: se f, f' e f'' forem contínuas e $f'(\alpha) \neq 0$, o método de Newton converge, desde que a aproximação inicial x_0 seja escolhida *"suficientemente próxima"* da raiz α.

E se $f'(\alpha) = 0$? Problemas de convergência. Veremos mais detalhes adiante.

3.5.2 Ordem de convergência do método de Newton

Vamos supor que as hipóteses do Teorema estão todas satisfeitas, i.e.:

- f, f', f'' contínuas em um intervalo I com centro em α
- $f'(\alpha) \neq 0$

então subtraindo α de $x_{k+1} = \phi(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}$ temos

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha)$$

Expandindo $\phi(x_k)$ em série de Taylor em torno do ponto $a = \alpha$, resulta em

$$\phi(x_k) = \phi(\alpha) + (x_k - \alpha)\phi'(\alpha) + \frac{(x_k - \alpha)^2}{2}\phi''(\xi_k)$$

assim

$$x_{k+1} - \alpha = \phi(\alpha) + (x_k - \alpha)\underbrace{\phi'(\alpha)}_{=0} + \frac{(x_k - \alpha)^2}{2}\phi''(\xi_k) - \phi(\alpha)$$

Portanto

$$x_{k+1} - \alpha = \frac{(x_k - \alpha)^2}{2}\phi''(\xi_k)$$
$$\frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{1}{2}\phi''(\xi_k)$$

e assim obtemos

$$\frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = \left|\frac{1}{2}\phi''(\xi_k)\right| \leq c$$

isto é

$$|x_{k+1} - \alpha| \leq c|x_k - \alpha|^2$$

Pela definição de ordem de convergência, concluimos que $p = 2$ e portanto temos que o método de Newton tem convergência quadrática.

Observações

- A convergência do método de Newton é rápida
- O método requer o cálculo:
  - da derivada da função
  - da avaliação da função e da sua derivada a cada iteração
- Além disso, a função pode não ser diferenciável em alguns pontos.

Na prática, o que significa essa ordem de convergência quadrática?

Vamos supor que o erro em uma iteração k do algoritmo seja da ordem de 10^{-2}. Pela expressão anterior

$$|x_{k+1} - \alpha| \leq c|x_k - \alpha|^2$$

ou seja, o erro na próxima iteração $k + 1$ é aproximadamente 10^{-4} e assim temos

$$10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, \ldots$$

Vejamos um exemplo prático.

Problemas com o método de Newton

Em algumas situações o método de Newton pode falhar por conta de:

i) Aproximação inicial ruim.
   Em algumas situações as condições sobre a função para a convergência do método são satisfeitas, porém a escolha da aproximação inicial está fora do intervalo para o qual o método converge.
   Por exemplo, se um ponto x_0 é estacionário, i.e., $f'(x_0) = 0$. Seja $f(x) = 1 - 2x^2$, então $f'(x) = -4x$. Escolhendo $x_0 = 0$ temos que

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{1}{0}$$

Problemas com o método de Newton

i) Aproximação inicial ruim.
Outra situação que pode levar à falha do método de Newton é a escolha de um ponto inicial que faz o método entrar em loop. Exemplo com $x_0 = 0$:

$$f(x) = x^3 - 2x + 2 \quad \Rightarrow \quad f'(x) = 3x^2 - 2$$

Escolhendo $x_0 = 0$ o método entra em loop e gera uma sequência $\{1, 0, 1, 0, \ldots\}$.
Obs: nesses casos, um método diferente, como por exemplo o método da Bissecção

pode ser usado para obter uma aproximação inicial mais precisa para então ser usada no método de Newton.

Problemas com o método de Newton

ii) Problemas com a derivada

- Derivada discontínua.
- Derivada não existe na raiz.

iii) Convergência não quadrática.
Em algumas situações o método pode convergir com uma ordem não quadrática. Um exemplo é quando temos raízes com multiplicidade $m > 1$, ou seja, quando a derivada é zero na raiz. Veremos como tratar isso adiante.

3.6 Método da Secante

Como discutido uma séria desvantagem do método de Newton é a necessidade de se obter $f'(x)$ e calcular o seu valor a cada passo. Existem algumas formas de modificar o método para contornar essa desvantagem.
Uma modificação consiste em substituir $f'(x)$ pelo quociente das diferenças

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \tag{3.18}$$

onde x_k e x_{k-1} são aproximações para α.
Dessa forma temos o seguinte esquema:

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{\frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}} \\ &= x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \end{aligned}$$

Tirando o mínimo e simplificando

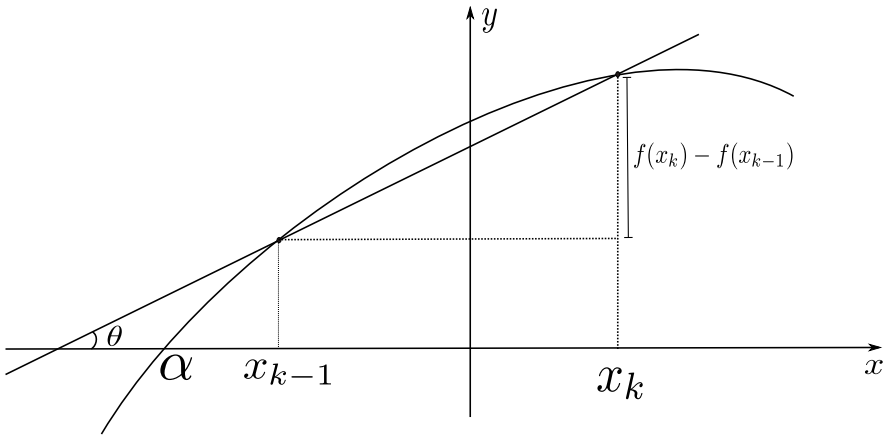
$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \tag{3.19}$$
$$= \frac{x_k f(x_k) - x_k f(x_{k-1}) - f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}$$

obtemos o seguinte processo iterativo

$$\boxed{x_{k+1} = \frac{f(x_k)x_{k-1} - f(x_{k-1})x_k}{f(x_k) - f(x_{k-1})}} \quad k = 1, 2, \ldots$$

o qual é conhecido como método da secante.
Note que para obter x_{k+1} precisamos x_k e x_{k-1}, ou seja, duas aproximações iniciais devem estar disponíveis para a equação anterior ser usada.

Interpretação Geométrica



Visto dessa forma o método consiste em tomar como aproximação a interseção da reta que passa pelos pontos $(x_k, f(x_k))$ e $(x_{k-1}, f(x_{k-1}))$ com o eixo x. Sendo assim, partindo de (3.19) temos

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} \quad \Rightarrow \quad \frac{x_{k+1} - x_k}{f(x_k)} = \frac{(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}$$
$$\Rightarrow \quad \frac{f(x_k)}{x_{k+1} - x_k} = \frac{f(x_k) - f(x_{k-1})}{(x_k - x_{k-1})}$$

$$\Rightarrow \quad \tan(\theta) = \frac{f(x_k)}{x_{k+1} - x_k} = \frac{f(x_k) - f(x_{k-1})}{(x_k - x_{k-1})}$$

Observação: note que a fórmula desse método é muito parecida com a do método da Falsa Posição, a diferença é que o método da Falsa Posição cerca a raiz α pelo intervalo $[a, b]$ e o método da Secante usa 2 aproximações sucessivas.

Exemplo

Encontre a raiz de $\sqrt{x} - 5e^{-x} = 0$, usando o método da Secante com $x_0 = 1.4$ e $x_1 = 1.5$ com uma precisão $\epsilon = 10^{-3}$.

Solução do Exemplo
Avaliando a função em x_0 e x_1 temos

$$f(x_0) = f(1.4) = \sqrt{1.4} - 5e^{-1.4} = 1.183 - 5(0.247) = -0.052$$
$$f(x_1) = f(1.5) = \sqrt{1.5} - 5e^{-1.5} = 1.225 - 5(0.223) = 0.110$$

pelo método da secante temos

$$x_2 = \frac{1.4f(1.5) - 1.5f(1.4)}{f(1.5) - f(1.4)} = \frac{1.4(0.110) - 1.5(-0.052)}{0.110 + 0.052} = 1.432$$
$$e_2 = \frac{|x_2 - x_1|}{|x_2|} = 0.047 > \epsilon \quad \Rightarrow \text{ mais iterações!}$$

Solução do Exemplo
Avaliando a função em x_2

$$f(x_2) = f(1.432) = \sqrt{1.432} - 5e^{-1.432} = 1.197 - 5(0.239) = 0.002$$

assim

$$x_3 = \frac{1.5f(1.432) - 1.432f(1.5)}{f(1.432) - f(1.5)}$$
$$= \frac{1.5(0.002) - 1.432(0.110)}{0.002 - 0.110} = 1.431$$
$$e_3 = \frac{|x_3 - x_2|}{|x_3|} = 0.0007 < \epsilon$$

Portanto, a raiz aproximada é $x_3 = 1.431$.
□

Sobre a implementação
A fórmula

$$x_{k+1} = \frac{f(x_k)x_{k-1} - f(x_{k-1})x_k}{f(x_k) - f(x_{k-1})}$$

não deve ser usada para implementação computacional do método pois problemas em sistemas de ponto flutuante como *cancelamento* podem ocorrer quando $x_k \approx x_{k-1}$ e $f(x_k)f(x_{k-1}) > 0$.
Nesse caso, para implementar o método use o seguinte esquema: dado x_0 e x_1, a sequência x_2, x_3, ... é calculada usando

$$\boxed{x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}}$$

desde que $f(x_k) \neq f(x_{k-1})$. **Algoritmo**

entrada: função $f(x)$, aproximações iniciais x0 e x1, precisão ϵ e número máximo de iterações $maxit$

para k *de* 1 *até* $maxit$ **faça**
  $f_0 = f(x_0)$;
  $f_1 = f(x_1)$;
  $x_2 = x_1 - \frac{f_1\ (x_1 - x_0)}{f_1 - f_0}$;
  se $abs(x_2 - x_1) < \epsilon$ **então**
    retorne x_2;
  fim-se
  $x_0 = x_1$;
  $x_1 = x_2$;
fim-para

3.6.1 Ordem de convergência do método da Secante

Seja

$$x_k = x_{k-1} - f(x_{k-1})\frac{x_{k-1} - x_{x-2}}{f(x_{k-1}) - f(x_{k-2})}$$

então definindo $e_k = x_k - \alpha \Rightarrow x_k = \alpha + e_k$ e assim substituindo temos

$$(\alpha + e_k) = (\alpha + e_{k-1}) - f(x_{k-1})\frac{e_{k-1} - e_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$$

$$e_k = e_{k-1} - f(x_{k-1})\frac{e_{k-1} - e_{x-2}}{f(x_{k-1}) - f(x_{k-2})}$$

$$= \frac{f(x_{k-1})e_{k-2} - f(x_{k-2})e_{x-1}}{f(x_{k-1}) - f(x_{k-2})}$$

Pelo Teorema do Valor Médio temos que existe c_{k-1} entre x_{k-1} e α tal que

$$(x_{k-1} - \alpha)f'(c_{k-1}) = f(x_{k-1}) - f(\alpha)$$

Sendo assim

$$f'(c_{k-1}) = \frac{f(x_{k-1})}{(x_{k-1} - \alpha)} = \frac{f(x_{k-1})}{e_{k-1}}$$

de onde obtemos

$$f(x_{k-1}) = f'(c_{k-1})e_{k-1}$$
$$f(x_{k-2}) = f'(c_{k-2})e_{k-2}$$

logo

$$e_k = \frac{f'(c_{k-1})e_{k-1}e_{k-2} - f'(c_{k-2})e_{k-1}e_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$$

$$= \frac{f'(c_{k-1}) - f'(c_{k-2})}{f(x_{k-1}) - f(x_{k-2})}e_{k-1}e_{k-2}$$

Vamos supor que existe $M > 0$ tal que

$$\max \left| \frac{f'(c_{k-1}) - f'(c_{k-2})}{f(x_{k-1}) - f(x_{k-2})} \right| < M$$

Então

$$|e_k| = M|e_{k-1}||e_{k-2}|$$

Suponha que para k muito grande exista algum C_k tal que

$$|e_k| = C_k|e_{k-1}|^p$$

onde p é a ordem de convergência do método.

De forma análoga temos que

$$|e_{k-1}| = C_{k-1}|e_{k-2}|^p \quad \Rightarrow \quad |e_{k-2}| = \frac{(|e_{k-1}|)^{(1/p)}}{(C_{k-1})^{(1/p)}}$$

Assim

$$|e_k| = C_k|e_{k-1}|^p$$
$$|e_k| = M|e_{k-1}||e_{k-2}| = M|e_{k-1}|\frac{(|e_{k-1}|)^{(1/p)}}{(C_{k-1})^{(1/p)}}$$

Pela igualdade temos

$$\Rightarrow \left[\frac{C_k}{M}(C_{k-1})^{(1/p)}\right]|e_{k-1}|^p = |e_{k-1}|^{1+1/p}$$

Assim temos que p deve satisfazer

$$p = 1 + \tfrac{1}{p} \quad \Rightarrow \quad p^2 - p - 1 = 0$$

de onde encontramos que

$$p = \frac{1+\sqrt{5}}{2} \approx 1.618$$

Assim concluimos que a ordem de convergência do método da secante é $p \approx 1.618$.

Observe que:

(-) ordem de convergência menor do que a do método de Newton

(+) este método não requer o conhecimento de $f'(x)$

3.7 Comparação dos métodos

- **Bisseção e Falsa Posição**: se a função $f(x)$ for contínua no intervalo $[a, b]$ e mudar de sinal nos extremos do intervalo $f(a)f(b) < 0$, então temos **garantia de convergência (!!!)**.

- **Ponto Fixo**: nem todas as escolhas da função de iteração do método do Ponto Fixo são adequadas, pois algumas divergem e outras podem convergir de forma muito lenta. O MPF irá convergir se:

  - $\phi(x)$ e $\phi'(x)$ contínuas num intervalo I centrado em α
  - $|\phi'(x)| < 1,\ \forall x \in I$

- **Newton**; possui critérios mais restritivos para convergência.

  - É preciso calcular $f(x)$ e $f'(x)$ a cada iteração
  - Convergência quadrática
  - Raiz com multiplicadade $m > 1 \Rightarrow$ convergência lenta.

- **Secante**: muito parecido com o método de Newton.

  - Precisa de duas aproximações para calcular uma nova aproximação.
  - Não é preciso conhecer a derivada.
  - O cálculo de $f'(x)$ é obtido de forma aproximada.
  - Convergência super-linear

- Existem situações que o método de Newton pode falhar:

  - má escolha para a aproximação inicial x_0
  - apresentar uma convergência não quadrática, quando temos raízes com multiplicicade $m > 1$ ou mesmo quando $f'(x_k) \approx 0$.

- De forma geral o método de Newton é o mais indicado sempre que for fácil avaliar as condições de convergência e que $f'(x)$ estiver disponível.

- Se $f'(x)$ não está disponível ou é uma função muito custosa de se avaliar, então o método da Secante é o mais indicado, uma vez que é o método que converge de forma mais rápida entre os demais.

- Podemos ainda usar um método como o da Bissecção/Falsa Posição cuja convergência é garantida para obter uma aproximação inicial mais precisa para ser usada no método de Newton, por exemplo.

Exemplo - (Ruggiero, Exemplo 18)
Considere a seguinte função $f(x) = e^{-x^2} - \cos(x)$. Para o método de Newton e do Ponto Fixo usamos:

$$f'(x) = \sin(x) - 2xe^{-x^2}$$
$$\phi(x) = \cos(x) - e^{-x^2} + x$$

Precisão $\epsilon = 10^{-8}$. Raiz encontrada $\alpha = 1.447414$.

método	iterações	dados
bissecção	24	$[1, 2]$
falsa posição	10	$[1, 2]$
ponto fixo	14	$x_0 = 1.5$
newton	4	$x_0 = 1.5$
secante	7	$x_0 = 1$, $x_1 = 2$

□

Considere a seguinte função $f(x) = x^3 - x - 1$. Para o método de Newton e do Ponto Fixo usamos:

$$f'(x) = 3x^2 - 1$$
$$\phi(x) = (x + 1)^{1/3}$$

Precisão $\epsilon = 10^{-8}$. Raiz encontrada $\alpha = 1.324718$.

método	iterações	dados
bissecção	24	$[1, 2]$
falsa posição	18	$[1, 2]$
ponto fixo	10	$x_0 = 1.0$
newton	**22**	$x_0 = 0$
secante	**27**	$x_0 = 0$, $x_1 = 0.5$

A convergência lenta do método de Newton se deve ao fato do chute inicial $x_0 = 0$ estar distante da raiz, e ainda porque x_0 gera $x_1 = 0.5$ como aproximação que está muito próximo de um zero de $f'(x) = 3x^2 - 1 = 0 \Rightarrow x = \pm\sqrt{3}/3 \approx \pm 0.57$
Idem para o método da Secante.

□

Considere a seguinte função $f(x) = 4\sin(x) - e^x$. Para o método de Newton e do Ponto Fixo usamos:

$$f'(x) = 4\cos(x) - e^x$$
$$\phi(x) = x - 2\sin(x) = 0.5e^x$$

Precisão $\epsilon = 10^{-8}$. Raiz encontrada $\alpha = 0.3705581$.

método	iterações	dados
bissecção	24	$[0, 1]$
falsa posição	9	$[0, 1]$
ponto fixo	8	$x_0 = 0.5$
newton	4	$x_0 = 0.5$
secante	8	$x_0 = 0$, $x_1 = 1$

□

Outros métodos e problemas

Outros métodos mais robustos

- Método pégaso
- Método Muller (aproximação quadrática)
- Método de van Wijngaarden-Dekker-Brent
  - Mais detalhes em [F. F. Campos, Cap. 6, Página 301]
  - Método usado na função `fzero` do MATLAB

Métodos específicos para raízes polinomiais

- Raízes complexas
- Mais detalhes em [N. B. Franco, Cap. 3, Página 92]

Capítulo 4

Resolução de Sistemas de Equações Lineares

Iremos estudar agora métodos computacionais para resolver um sistema de equações lineares da forma:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad & \qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

onde

$$a_{ij} \in \mathbb{R}, \quad b_i \in \mathbb{R}, \quad x_j \in \mathbb{R}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$$

Chamamos a_{ij} de coeficientes, b_i são constantes dadas e x_j são as variáveis ou incógnitas do problema.

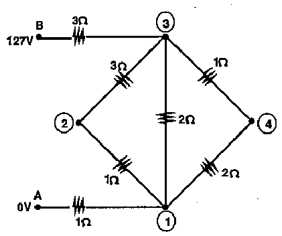
4.1 Introdução

Em uma vasta gama de problemas de ciências e engenharias a solução de um sistema de equações lineares é necessária. Podemos enumerar diversas áreas e problemas típicos, tais como:

- Solução de equações diferenciais ordinárias;
- Solução de equações diferenciais parciais através do método dos elementos finitos, diferenças finitas ou volumes finitos;
- Programação linear;
- Análise de estruturas;
- Sistemas de equações não-lineares;
- Outros métodos numéricos: interpolação, mínimos quadrados, etc.
- Circuitos elétricos

4.1.1 Exemplo em circuitos elétricos

Considere o seguinte exemplo como motivação. Calcular as tensões dos nós do circuito elétrico da figura abaixo: Para a resolução considere a seguinte modelagem do problema:



- Lei de Kirchhoff: a soma das correntes que passam em cada nó do circuito é nula.
- Lei de Ohm: a corrente do nó j para o nó k é dada pela equação

$$I_{jk} = \frac{V_j - V_k}{R_{jk}}$$

Aplicando ao circuito elétrico, para cada nó temos:

- Nó 1:

$$\begin{aligned} I_{A1} + I_{21} + I_{31} + I_{41} &= 0 \\ \tfrac{0-V_1}{1} + \tfrac{V_2-V_1}{1} + \tfrac{V_3-V_1}{2} + \tfrac{V_4-V_1}{2} &= 0 \\ -2V_1 + V_2 - V_1 + \tfrac{V_3}{2} - \tfrac{V_1}{2} + \tfrac{V_4}{2} - \tfrac{V_1}{2} &= 0 \\ -4V_1 + 2V_2 + V_3 - 2V_1 + 2V_4 &= 0 \\ \boxed{-6V_1 + 2V_2 + V_3 + V_4 = 0} & \end{aligned}$$

- Nó 2:

$$\boxed{3V_1 - 4V_2 + V_3 = 0}$$

- Nó 3:

$$\boxed{3V_1 + 2V_2 - 13V_2 + 6V_4 = -254}$$

- Nó 4:

$$\boxed{V_1 + 2V_3 - 3V_4 = 0}$$

Como as quatro equações anteriores devem ser satisfeitas simultaneamente, isto resulta no seguinte sistema de equações lineares:

$$\begin{bmatrix} -6 & 2 & 1 & 1 \\ 3 & -4 & 1 & 0 \\ 3 & 2 & -13 & 6 \\ 1 & 0 & 2 & -3 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -254 \\ 0 \end{bmatrix}$$

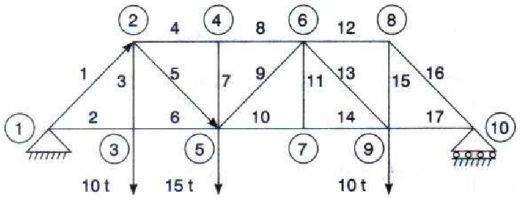
Usando algum método que iremos estudar, encontramos a solução deste sistema

$$V^* = \begin{bmatrix} 25.7 \\ 31.75 \\ 49.6 \\ 41.6 \end{bmatrix}$$

ou seja $V_1 = 25.7V$, $V_2 = 31.75V$, $V_3 = 49.6V$ e $V_4 = 41.6V$.

4.1.2 Exemplo em estruturas

Exemplo 1, Capítulo 3, Página, 105, Livro da Ruggiero.
Determinar as forças que atuam nesta treliça.



Junção 2:

$$\sum F_x = -\alpha f_1 + f_4 + \alpha f_5 = 0$$
$$\sum F_y = -\alpha f_1 - f_3 - \alpha f_5 = 0$$

Procedendo de forma análoga para todas as junções obtem-se um sistema linear de 17 equações e 17 variáveis ($f_1, \ldots, f_{17}$).

4.2 Conceitos fundamentais

Antes de estudar os métodos para solução deste tipo de problema, vamos rever alguns conceitos fundamentais de Álgebra Linear necessários para o desenvolvimento e análise dos métodos.

Definição 4. Uma matriz é um conjunto de elementos (números reais ou complexos) dispostos de forma retangular. O tamanho ou dimensão é definido pelo seu número de linhas e colunas. Uma matriz com m linhas e n colunas é dita ser $m \times n$ (m por n) e se $m = n$, então dizemos que a matriz é quadrada.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

Um elemento a_{ij} da matriz é referenciado por 2 índices: o primeiro indica a linha e o segundo a coluna.

Alguns casos particulares de matrizes:

- Matriz linha: $1 \times n$

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \end{bmatrix}$$

- Matriz coluna: $n \times 1$

$$\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}$$

- Matriz nula:

$$a_{ij} = 0, \quad \forall i, j$$

- Matriz diagonal:

$$d_{ij} = 0, \quad \forall i \neq j$$

- Matriz identidade:

$$e_{ij} = 1, \quad \forall i = j$$
$$e_{ij} = 0, \quad \forall i \neq j$$

- Matriz triangular inferior: acima da diagonal principal é nula

$$b_{ij} = 0, \quad \forall\, i < j, \qquad \text{Exemplo: } \mathbf{B} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

- Matriz triangular superior: abaixo da diagonal principal é nula

$$c_{ij} = 0, \quad \forall\, i > j, \qquad \text{Exemplo: } \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{bmatrix}$$

- Matriz simétrica:

$$m_{ij} = m_{ji}, \quad \forall\, i, j$$

Transposição A transposta de uma matriz $\mathbf{A}$, denotada por $\mathbf{A}^T$, é uma matriz obtida trocando-se as suas linhas pelas colunas. Exemplo:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}, \qquad \mathbf{A}^T = \begin{bmatrix} 1 & 5 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}$$

Adição e Subtração Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes $m \times n$. Então a matriz $\mathbf{C}$ é $m \times n$ e seus elementos são dados por

$$c_{ij} = a_{ij} + b_{ij}, \quad \forall\, i, j$$

Multiplicação por escalar Seja $\mathbf{A}$ uma matriz $m \times n$ e seja $k \in \mathbb{R}$ um escalar qualquer. Então $\mathbf{B} = k\mathbf{A}$ é tal que

$$b_{ij} = k\, a_{ij}, \quad \forall\, i, j$$

Multiplicação matriz-vetor Seja $\mathbf{A}$ uma matriz $m \times n$ e $\mathbf{x}$ um vetor $n \times 1$, então a multiplicação de $\mathbf{A}$ por $\mathbf{x}$ é

$$\mathbf{v} = \mathbf{A}\mathbf{x} \quad \Rightarrow \quad v_i = \sum_{j=1}^{n} a_{ij}\, x_j, \quad i = 1, 2, \ldots, m$$

Exemplo

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \\ 17 \end{bmatrix}$$

Multiplicação matriz-matriz Seja $\mathbf{A}$ uma matriz $m \times p$ e $\mathbf{B}$ uma matriz $p \times n$. O resultado da multiplicação $\mathbf{AB}$ é uma matriz $\mathbf{C}$ de tamanho $m \times n$.

$$c_{ij} = \sum_{k=1}^{p} a_{ik}\, b_{kj}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$$

Exemplo:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 0 \\ -1 & 6 & 4 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} -3 & 2 \\ 4 & 9 \\ 8 & -1 \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} -5 & 15 \\ 59 & 48 \end{bmatrix}$$

Produto Interno e Produto Externo O produto interno ou escalar entre dois vetores $\mathbf{x}$ e $\mathbf{y}$, ambos de tamanho n resulta em um valor escalar k dado por

$$k = \mathbf{x}^T\mathbf{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n = \sum_{i=1}^{n} x_i\, y_i$$

O produto externo entre $\mathbf{x}(m \times 1)$ e $\mathbf{y}(n \times 1)$ resulta em uma matriz $\mathbf{M}$ de tamanho $m \times n$ dada por

$$m_{ij} = x_i\, y_j, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$$

Exemplo:

$$\mathbf{x} = \begin{bmatrix} 5 \\ -1 \\ 2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{x}^T\mathbf{y} = 10, \quad \mathbf{x}\mathbf{y}^T = \mathbf{M} = \begin{bmatrix} 5 & 15 & 20 \\ -1 & -3 & -4 \\ 2 & 6 & 8 \end{bmatrix}$$

Determinante Seja $\mathbf{A}$ uma matriz quadrada de ordem n. Então $\mathbf{A}$ possui um número associado chamado de determinante, o qual pode ser calculado pela seguinte fórmula:

$$det(\mathbf{A}) = a_{11}det(\mathbf{M_{11}}) - a_{12}det(\mathbf{M_{12}}) + \ldots + (-1)^{n+1}a_{1n}det(\mathbf{M_{1n}})$$

onde $\mathbf{M_{ij}}$ é a matriz resultante da remoção da linha i e da coluna j da matriz $\mathbf{A}$. Em particular

$$\mathbf{A} = [a_{11}] \Rightarrow det(\mathbf{A}) = a_{11}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \Rightarrow det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$$

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\begin{aligned} det(\mathbf{A}) = & a_{11}(a_{22}a_{33} - a_{32}a_{23}) \\ & - a_{12}(a_{21}a_{33} - a_{31}a_{23}) \\ & + a_{13}(a_{21}a_{32} - a_{31}a_{22}) \end{aligned}$$

Definição 5 (Matriz singular)**.** Uma matriz com $det(\mathbf{A}) = 0$ é dita **singular**. Por outro lado quando $det(\mathbf{A}) \neq 0$ dizemos que a matriz é **não-singular**.

Definição 6 (Vetores Linearmente Independentes)**.** Um conjunto de vetores $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}$ é dito ser linearmente independente (LI) se

$$c_1\mathbf{x_1} + c_2\mathbf{x_2} + \ldots + c_k\mathbf{x_k} = 0$$

somente se $c_1 = c_2 = \ldots = c_k = 0$. Caso contrário, isto é, quando $c_1, c_2, \ldots, c_k$ não são todos nulos, dizemos que o conjunto de vetores é linearmente dependente (LD).

Definição 7 (Posto)**.** O posto (ou rank) de uma matriz $\mathbf{A}$ de tamanho $m \times n$ é definido como o número máximo de vetores linhas (ou de vetores colunas) linearmente independentes de $\mathbf{A}$. Escrevemos posto$(\mathbf{A}) = r$ e temos que $r \leq \min(m, n)$.

Definição 8 (Inversa)**.** A inversa de uma matriz $\mathbf{A}$ quadrada $n \times n$ é representada por $\mathbf{A^{-1}}$ e definida de tal forma que

$$\mathbf{A}\mathbf{A^{-1}} = \mathbf{A^{-1}}\mathbf{A} = \mathbf{I}$$

onde $\mathbf{I}$ é a matriz identidade de ordem n.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}, \quad \mathbf{A^{-1}} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -2 & 1 \end{bmatrix}$$

4.3 Sistemas Lineares

Um sistema de equações lineares consiste em um conjunto de m equações polinomiais com n variáveis x_i de grau um, isto é

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad & \qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

o qual pode ser escrito da seguinte forma matricial $\mathbf{Ax} = \mathbf{b}$ onde

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{4.1}$$

onde $\mathbf{A}$ é a matriz dos coeficientes, $\mathbf{b}$ é o vetor dos termos independentes e $\mathbf{x}$ é o vetor solução procurado.

4.3.1 Número de Soluções

Vamos considerar apenas sistemas cujas matrizes dos coeficientes são quadradas, isto é, onde $\mathbf{A} \in \mathbb{R}^{n \times n}$. Iremos tratar do caso onde $\mathbf{A}$ não é uma matriz quadrada e $m > n$ mais adiante, quando estudarmos mínimos quadrados.

Para o sistema $\mathbf{Ax} = \mathbf{b}$, temos as seguintes possibilidades quanto ao número de soluções:

(a) uma única solução

(b) infinitas soluções

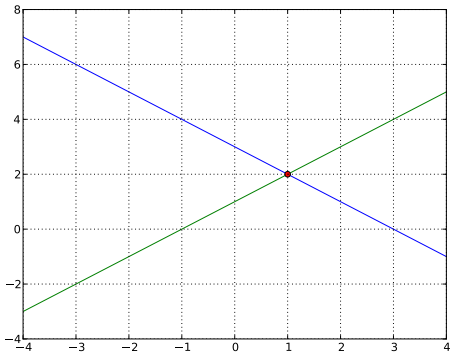
(c) sem solução

Vamos analisar cada caso em mais detalhes através de alguns exemplos de sistemas de equações lineares 2×2.

Caso (a) Única solução

$$\begin{aligned} x_1 + x_2 &= 3 \\ x_1 - x_2 &= -1 \end{aligned}$$

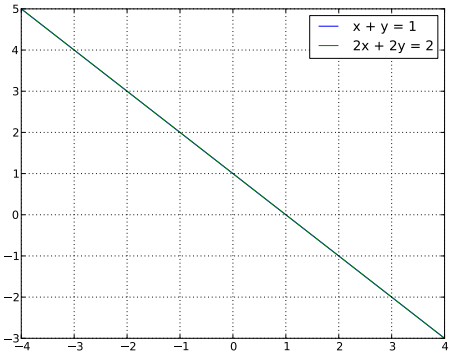
$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$



Caso (b) Infinitas Soluções

$$x_1 + x_2 = 1$$
$$2x_1 + 2x_2 = 2$$

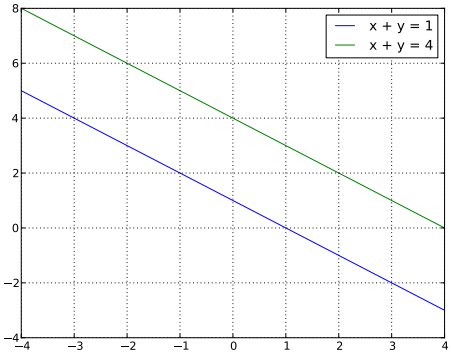
$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{x} = \begin{bmatrix} 1 - \theta \\ \theta \end{bmatrix}$$



Caso (c) Sem Solução

$$x_1 + x_2 = 1$$
$$x_1 + x_2 = 4$$

$$\Rightarrow \quad \nexists \mathbf{x} \text{ tal que } \mathbf{Ax} = \mathbf{b}$$



4.3.2 Existência e unicidade da solução

A equação $\mathbf{Ax} = \mathbf{b}$ possui uma única solução se e somente se a matriz $\mathbf{A}$ for não-singular. O Teorema a seguir, caracteriza a não-singularidade da matriz $\mathbf{A}$.

Teorema 9. Seja $\mathbf{A}$ uma matriz quadrada $n \times n$. As seguintes afirmações são equivalentes:

a) $\mathbf{A^{-1}}$ existe

b) Não existe $\mathbf{y}$ não-zero tal que $\mathbf{Ay} = \mathbf{0}$. Ou seja, a única solução do sistema homogêneo é $\mathbf{y} = \mathbf{0}$.

c) posto$(\mathbf{A}) = n$

d) det$(\mathbf{A}) \neq 0$

e) Dado qualquer vetor $\mathbf{b}$, existe exatamente um vetor $\mathbf{x}$ tal que $\mathbf{Ax} = \mathbf{b}$ (ou $\mathbf{x} = \mathbf{A^{-1}b}$).

Prova: Livro texto de Álgebra Linear.

Existência e unicidade da solução De fato, para os exemplos anteriores, temos **Caso (a)**

$$\det\left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right) = -1 - 1 = -2 \neq 0 \Rightarrow \text{ OK, solução única}$$

Caso (b)

$$\det\left(\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\right) = 2 - 2 = 0$$

Caso (c)

$$\det\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) = 1 - 1 = 0$$

4.4 Métodos Diretos

Iremos estudar agora diversos métodos numéricos para a solução de sistemas de equações lineares. Vamos considerar que $\mathbf{A}$ é quadrada e não-singular.

Os métodos de solução de sistemas lineares geralmente a conversão de um sistema quadrado em um sistema triangular que possui a mesma solução que o original.

Inicialmente, vamos estudar como resolver sistemas lineares triangulares inferiores e superiores.

4.4.1 Sistemas Triangulares

Sistema triangular inferior Considere um sistema triangular inferior de ordem n dado por

$$\begin{bmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & & & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

A solução deste sistema é feita através de um procedimento chamado de **substituição** (ou substituições sucessivas):

$$
\begin{aligned}
& l_{11}x_1 = b_1 && \Rightarrow \quad x_1 = \frac{b_1}{l_{11}} \\
& l_{21}x_1 + l_{22}x_2 = b_2 && \Rightarrow \quad x_2 = \frac{b_2 - l_{21}x_1}{l_{22}} \\
& \vdots \\
& l_{n1}x_1 + l_{n2}x_2 + \ldots + l_{nn}x_n = b_n && \Rightarrow \quad x_n = \frac{b_n - l_{n1}x_1 - l_{n2}x_2 - \ldots - l_{nn-1}x_{n-1}}{l_{nn}}
\end{aligned}
$$

De forma geral para $\mathbf{Lx} = \mathbf{b}$ temos

$$
\boxed{x_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij}\, x_j \right) \Big/ l_{ii}} \quad i = 1, \ldots, n
$$

Exemplo

$$
\begin{bmatrix} 2 & 0 & 0 & 0 \\ 3 & 5 & 0 & 0 \\ 1 & -6 & 8 & 0 \\ -1 & 4 & -3 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 48 \\ 0 \end{bmatrix}
$$

Solução

$$
\begin{aligned}
& 2x_1 = 4 && \Rightarrow \quad x_1 = 2 \\
& 3x_1 + 5x_2 = 1 && \Rightarrow \quad x_2 = \tfrac{1-6}{5} = -1 \\
& x_1 - 6x_2 + 8x_3 = 48 && \Rightarrow \quad x_3 = \tfrac{48-2-6}{8} = 5 \\
& -x_1 + 4x_2 - 3x_3 + 9x_4 = 0 && \Rightarrow \quad x_4 = \tfrac{2+4+15}{9} = \tfrac{21}{9}
\end{aligned}
$$

---

entrada: $\mathbf{L} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$
saída: $\mathbf{x} \in \mathbb{R}^n$

```
x(1) = b(1) / L(1,1);
para i=2, ..., n faça
    s = b(i);
    para j=1, ..., i-1 faça
        s = s - L(i,j) * x(j);
    fim-para
    x(i) = s/L(i,i);
fim-para
```

Sistema triangular superior O algoritmo análogo para o caso de um sistema triangular

superior $\mathbf{Ux} = \mathbf{b}$ é chamado de **retro-substituição** (ou substituições retroativas).

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

e assim temos

$$\begin{aligned} u_{nn}x_n = b_n \qquad & \Rightarrow \quad x_n = \frac{b_n}{u_{nn}} \\ u_{n-1n-1}x_{n-1} + u_{n-1n}x_n = b_{n-1} \qquad & \Rightarrow \quad x_{n-1} = \frac{b_{n-1} - u_{n-1n}x_n}{u_{n-1n-1}} \\ & \vdots \\ u_{11}x_1 + u_{12}x_2 + \ldots + u_{1n}x_n = b_1 \qquad & \Rightarrow \quad x_1 = \frac{b_n - u_{12}x_1 - u_{13}x_3 - \ldots - u_{1n}x_n}{u_{11}} \end{aligned}$$

De forma geral para $\mathbf{Ux} = \mathbf{b}$ temos

$$\boxed{x_i = \left(b_i - \sum_{j=i+1}^{n} u_{ij}\, x_j \right) \Big/ u_{ii}} \quad i = n, \ldots, 1$$

Exemplo

$$\begin{bmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}$$

Solução

$$\begin{aligned} 4x_3 = 8 \qquad & \Rightarrow \quad x_3 = 2 \\ x_2 + x_3 = 4 \qquad & \Rightarrow \quad x_2 = 2 \\ 2x_1 + 4x_2 - 2x_3 = 2 \qquad & \Rightarrow \quad x_1 = \tfrac{2-8+4}{2} = -\tfrac{2}{2} = -1 \end{aligned}$$

---

entrada: $\mathbf{U} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$
saída: $\mathbf{x} \in \mathbb{R}^n$

```
x(n) = b(n)/U(n,n);
para i=n-1, ..., 1 faça
    s = b(i);
    para j=i+1, ..., n faça
        s = s - U(i,j) * x(j);
    fim-para
    x(i) = s/U(i,i);
fim-para
```

4.4.2 Complexidade Computacional

Muitas vezes precisamos medir o custo de execução de um algoritmo. Para isso usualmente definimos uma função de complexidade que pode ser uma medida do tempo para o algoritmo resolver um problema cuja instância de entrada tem tamanho n (ou medir por exemplo o quanto de memória seria necessário para execução).

A complexidade de um algoritmo para solução de um sistema linear de ordem n é medida através do número de operações aritméticas como adição, multiplicação e divisão.

Lembrando que

$$\boxed{\sum_{i=1}^{n} i = \frac{n(n+1)}{2}}$$

Substituição:

Divisão: n

Adição: $\displaystyle\sum_{i=2}^{n}(i-1) = \sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}$

Multiplicação: $\displaystyle\sum_{i=2}^{n}(i-1) = \sum_{i=1}^{n-1} i = \frac{n(n-1)}{2}$

No total o algoritmo de substituição para sistemas triangulares inferiores realiza

$$n + \frac{n(n-1)}{2} + \frac{n(n-1)}{2} = n + n^2 - n = n^2$$

operações de ponto flutuante.

Complexidade Computacional

Retro-substituição:

Divisão: n

Adição: $\displaystyle\sum_{i=1}^{n-1}(n-i) = n(n-1) - \frac{n(n-1)}{2} = \frac{n(n-1)}{2}$

Multiplicação: $\displaystyle\sum_{i=1}^{n-1}(n-i) = n(n-1) - \frac{n(n-1)}{2} = \frac{n(n-1)}{2}$

No total o algoritmo de retro-substituição para sistemas triangulares superiores realiza

$$n + \frac{n(n-1)}{2} + \frac{n(n-1)}{2} = n + n^2 - n = n^2$$

operações de ponto flutuante.

4.4.3 Métodos para solução de sistemas lineares

Existem dois tipos de métodos para a solução de sistemas de equações lineares:

- Métodos diretos
  - Os métodos diretos são aqueles que conduzem à **solução exata** após um número finito de passos a menos de erros de arredondamento introduzidos pela máquina.
- Métodos iterativos
  - São aqueles que se baseiam na construção de **sequências de aproximações**. Em um método iterativo, a cada passo, os valores calculados anteriormente são usados para melhorar a aproximação. É claro que o método só será útil se a sequência de aproximações construídas convergir para uma solução aproximada do sistema.

4.4.4 Eliminação de Gauss

O primeiro método direto que iremos estudar é o método da eliminação de Gauss. A idéia fundamental do método é transformar a matriz **A** em uma matriz triangular superior introduzindo zeros abaixo da diagonal principal, primeiro na coluna 1, depois na coluna 2 e assim por diante.

$$\begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \end{bmatrix} \rightarrow \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix}$$

Por fim, usa-se a **retro-substituição** para obter a solução do sistema triangular superior obtido ao final dessa etapa de eliminação.

Na eliminação de Gauss, as operações efetuadas para se obter a matriz triangular superior são tais que a matriz triangular obtida possui a mesma solução que o sistema original.

Definição 9 (Sistema equivalente)**.** Dois sistemas de equações lineares são equivalentes quando possuem o mesmo vetor solução.

Um sistema pode ser transformado em um outro sistema equivalente utilizando as seguintes operações elementares:

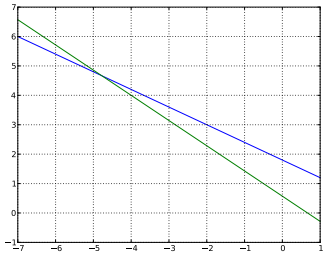
- trocar a ordem de duas equações
- multiplicar uma equação por uma constante não-nula
- somar um múltiplo de uma equação à outra

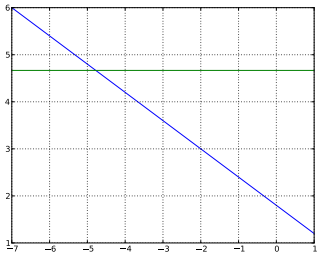
Exemplo

$$3x_1 + 5x_2 = 9$$
$$6x_1 + 7x_2 = 4$$

Podemos subtrair da linha 2 um múltiplo da linha 1, isto é

$$L_2' = L_2 - 2L_1$$





Efetuando esta operação obtemos o sistema equivalente

$$\begin{aligned} 3x_1 + 5x_2 &= 9 \\ -3x_2 &= -14 \end{aligned}$$

Vamos primeiro estudar um exemplo simples para posteriormente generalizar a idéia.
Exemplo Seja o sistema

$$\begin{aligned} x_1 + x_3 &= 0 \\ x_1 + x_2 &= 1 \\ 2x_1 + 3x_2 + x_3 &= 1 \end{aligned}$$

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Solução Como podemos eliminar os coeficientes abaixo da diagonal principal na primeira coluna?

$$\begin{aligned} L_2' &= L_2 - L_1 \\ L_3' &= L_3 - 2L_1 \end{aligned}$$

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

Precisamos agora de eliminar os coeficientes abaixo da diagonal na segunda coluna (a_{32}). Como?

$$L_3'' = L_3' - 3L_2'$$

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

Agora podemos usar a retro-substituição para encontrar facilmente a solução deste sistema:

$$\begin{aligned} 2x_3 = -2 \quad &\Rightarrow \quad x_3 = -1 \\ x_2 - x_3 = 1 \quad &\Rightarrow \quad x_2 = 1 + x_3 = 1 - 1 = 0 \\ x_1 + x_3 = 0 \quad &\Rightarrow \quad x_1 = -x_3 = 1 \end{aligned}$$

Encontramos assim a solução: $x^T = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$

4.4.5 Revisitando a Eliminação de Gauss

Considere o problema de resolver o seguinte sistema pela eliminação de Gauss

$$\begin{bmatrix} \mathbf{2} & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 9 \end{bmatrix}$$

Passo 1

$$\begin{aligned} m_{21} = \tfrac{a_{21}}{a_{11}} = 4/2 = 2 \quad &\Rightarrow \quad L_2' = L_2 - 2L_1 \\ m_{31} = \tfrac{a_{31}}{a_{11}} = -2/2 = -1 \quad &\Rightarrow \quad L_3' = L_3 + L_1 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 8 & 3 & 14 \end{array}\right]$$

Passo 2

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & \mathbf{-8} & -2 & -12 \\ 0 & 8 & 3 & 14 \end{array}\right]$$

$$m_{32} = \tfrac{a_{32}}{a_{22}} = 8/-8 = -1 \quad \Rightarrow \quad L_3'' = L_3' + L_2'$$

$$\left[\begin{array}{ccc|c} 2 & 1 & 1 & 5 \\ 0 & -8 & -2 & -12 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

Próxima etapa: resolver o sistema triangular superior obtido usando o algoritmo de **retro-substituição**.

4.4.6 Formalização da Eliminação de Gauss

Considere o seguinte sistema, de dimensão $n \times n$ escrito de forma aumentada (juntamente com o vetor $\mathbf{b}$), isto é

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right]$$

procedendo com a eliminação de Gauss temos:
Passo 1 (k=1): eliminamos os elementos abaixo da diagonal principal na primeira coluna. Suponha que $a_{11} \neq 0$. Então definimos os multiplicadores do primeiro passo:

$$\begin{aligned} m_{21} &= a_{21}/a_{11} \\ m_{31} &= a_{31}/a_{11} \\ &\vdots \\ m_{n1} &= a_{n1}/a_{11} \end{aligned}$$

ou seja para $i = 2, \dots, n$ tem-se: $m_{i1} = a_{i1}/a_{11}$. Agora, multiplicamos a 1^a equação por m_{i1} e subtraimos da i-ésima equação, isto é

$$\begin{aligned} \text{Para } i = 2:n \qquad a_{ij}^{(1)} &= a_{ij}^{(0)} - m_{i1}\ a_{1j}^{(0)} \\ b_i^{(1)} &= b_i^{(0)} - m_{i1}\ b_1^{(0)}, \quad j = 1:n \end{aligned}$$

Observe que não alteramos a primeira linha, pois $i = 2, \dots, n$, logo esta permanece inalterada:

$$a_{1j}^{(1)} = a_{1j}^{(0)} = a_{1j}, \quad b_1^{(1)} = b_1^{(0)} = b_1$$

Após essa etapa zeramos todos os elementos abaixo da diagonal principal na 1^a coluna.

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ \mathbf{0} & \mathbf{a_{22}^1} & \mathbf{a_{23}^1} & \dots & \mathbf{a_{2n}^1} & \mathbf{b_2^1} \\ \mathbf{0} & \mathbf{a_{32}^1} & \mathbf{a_{33}^1} & \dots & \mathbf{a_{3n}^1} & \mathbf{b_3^1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{a_{n2}^1} & \mathbf{a_{n3}^1} & \dots & \mathbf{a_{nn}^1} & \mathbf{b_n^1} \end{array}\right]$$

Passo 2 (k=2): consiste em introduzir zeros abaixo da diagonal principal na 2^a coluna. Suponha $a_{22} \neq 0$. Definimos então os multiplicadores

$$m_{i2} = a_{i2}/a_{22}, \quad i = 3, \dots, n$$

e assim

$$\begin{aligned} \text{para } i = 3, \dots, n \qquad a_{ij}^{(2)} &= a_{ij}^{(1)} - m_{i2}\ a_{2j}^{(1)} \\ b_i^{(2)} &= b_i^{(1)} - m_{i2}\ b_1^{(2)}, \quad j = 2, \dots, n \end{aligned}$$

o que resulta em

$$\left[\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 & b_2^1 \\ 0 & \mathbf{0} & \mathbf{a_{33}^2} & \dots & \mathbf{a_{3n}^2} & \mathbf{b_3^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \mathbf{0} & \mathbf{a_{n3}^2} & \dots & \mathbf{a_{nn}^2} & \mathbf{b_n^2} \end{array}\right]$$

O algoritmo segue com os passos 3, 4 em diante. Agora considere a descrição do algoritmo para um passo k qualquer.

Passo k: Considerando $a_{kk} \neq 0$, temos

$$m_{ik} = a_{ik}/a_{kk}, \quad i = k+1, \dots, n$$

e assim fazemos as seguintes operações

$$\begin{aligned} \text{para } i = k+1 : n \qquad a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik}\ a_{kj}^{(k-1)} \\ b_i^{(k)} &= b_i^{(k-1)} - m_{ik}\ b_k^{(k-1)}, \quad j = k, \dots, n \end{aligned}$$

Observe novamente que não alteramos as linhas de 1 até k.

No processo de eliminação os elementos $a_{11}, a_{22}^{(1)}, a_{33}^{(2)}, \dots, a_{kk}^{(k-1)}$ que aparecem na diagonal da matriz $\mathbf{A}$ são chamados de **pivôs**.

Se os pivôs não se anulam, isto é, se $a_{kk} \neq 0, k = 1 : n$, durante o processo, então a eliminação procede com sucesso e por fim chegamos ao seguinte sistema triangular superior

$$\left[\begin{array}{ccccccc} a_{11} & a_{12} & a_{13} & \dots & a_{1n-1} & a_{1n} & b_1 \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n-1}^1 & a_{2n}^1 & b_2^1 \\ 0 & 0 & a_{33}^2 & \dots & a_{3n-1}^2 & a_{3n}^2 & b_3^2 \\ \vdots & \vdots & \vdots & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{0} & \mathbf{a_{nn}^{n-1}} & \mathbf{b_n^{n-1}} \end{array}\right] \tag{4.2}$$

Em seguida resolvemos esse sistema usando **retro substituição**. O Algoritmo abaixo descreve o funcionamento do método da Eliminação de Gauss.

Algorithm 4: Eliminação de Gauss

entrada: matriz $\mathbf{A} \in \mathbb{R}^{n \times n}$, vetor $\mathbf{b} \in \mathbb{R}^n$
saída: vetor solução $\mathbf{x} \in \mathbb{R}^n$

```
para k = 1 : n − 1 faça
    para i = k + 1 : n faça
        m = A(i,k) / A(k,k);
        para j = k + 1 : n faça
            A(i,j) = A(i,j) - m * A(k,j);
        fim-para
        b(i) = b(i) - m * b(k);
    fim-para
fim-para
x = retroSubstituicao(A,b);
retorna x;
```

4.4.7 Observação importante

Mas, e se na etapa k da eliminação de Gauss, o pivô for zero? Isso significa que $a_{kk} = 0$, e assim, teríamos

$$m_{ik} = \frac{a_{ik}}{a_{kk}} \Rightarrow \quad \text{divisão por zero!}$$

Nesse caso, se um pivô for zero, o processo de eliminação tem que parar, ou temporariamente ou permanentemente.

O sistema pode ou não ser singular.

Se o sistema for singular, i.e, $\det(\mathbf{A}) = 0$, e portanto como vimos o sistema não possui uma única solução. Veremos agora um caso que a matriz não é singular e podemos resolver esse problema.

4.5 Estratégia de Pivotamento

Vamos ilustrar a idéia do pivoteamento através de um exemplo. Considere a seguinte matriz.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 6 & 8 \end{bmatrix}$$

Vamos proceder com a eliminação de Gauss:

$$\begin{aligned} m_{21} = 2, \quad & a^1_{2j} = a^0_{2j} - 2\ a^0_{1j} \\ m_{31} = 4, \quad & a^1_{3j} = a^0_{3j} - 4\ a^0_{1j}, \quad j = 1:3 \end{aligned}$$

então obtemos

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 3 \\ 0 & 2 & 4 \end{bmatrix}$$

No próximo passo, o pivô é a_{22} e usamos ele para calcular m_{32}. Entretanto

$$m_{32} = \frac{a_{32}}{a_{22}} = \frac{2}{0},$$

ocorre uma divisão por zero! E agora, o que podemos fazer?

Podemos realizar uma operação elementar de troca de linhas. Como vimos este tipo de operação quando realizado em um sistema, não altera a solução. Sendo assim, vamos trocar as linhas 2 e 3 e seguir adiante com a eliminação de Gauss.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 3 \\ 0 & 2 & 4 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix}$$

E assim chegamos a um sistema triangular superior, cuja solução pode ser obtida facilmente usando a retro-substituição.

Iremos mostrar adiante que a estratégia de **pivotamento** é importante pois:

- **evita a propagação de erros numéricos**
- nos fornece meios de evitar problemas durante a eliminação de Gauss quando o **pivô** a_{kk} no passo k é **igual a zero** e precisamos calcular o multiplicador

$$m_{ik} = \frac{a_{ik}}{a_{kk}}$$

Assim, através da troca de linhas, podemos encontrar uma linha de tal forma que o novo pivô é não-zero, permitindo que a eliminação de Gauss continue até chegar em uma matriz triangular superior. Com relação à técnica de pivotamento, existem duas possibilidades que serão discutidas a seguir:

- pivotamento parcial
- pivotamento total

4.5.1 Pivotamento Parcial

No pivotamento parcial, em cada passo k, **o pivô é escolhido como o maior elemento em módulo** abaixo de a_{kk} (inclusive), isto é

$$\text{Encontrar } r \text{ tal que: } \; |a_{rk}| = \max |a_{ik}|, \; k \leq i \leq n$$

Feita a escolha do pivô, trocamos as linhas r e k e o algoritmo procede. Isso evita a propagação de erros numéricos, pois:

- O pivoteamento parcial garante que

$$|m_{ik}| \leq 1$$

- Se a_{kk} for muito pequeno, consequentemente m_{ik} será muito grande. Dessa forma, após a multiplicação por m_{ik} podemos ampliar erros de arredondamento envolvidos no processo.
- Também evitamos o erro que pode ser causado quando somamos um número pequeno com um número grande.

Além disso, considere a seguinte situação como exemplo após a execução do passo 1 da eliminação de Gauss com pivotamento parcial:

$$\begin{bmatrix} x & x & x & x & x \\ 0 & \mathbf{0} & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \end{bmatrix}$$

neste caso, note a ocorrência de um zero na diagonal principal (da segunda coluna). Sendo assim podemos seguir para o próximo passo e completar a eliminação. Entretanto a matriz triangular superior $\mathbf{U}$ resultante do processo terá um zero na diagonal principal, o que implica que $\det \mathbf{U} = 0$. Assim, conclui-se que $\mathbf{U}$ é uma matriz singular e, portanto, a matriz $\mathbf{A}$ também é singular.

Exemplo 36

Aplique a eliminação de Gauss com pivoteamento parcial no seguinte sistema:

$$\left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 4 & 9 & -3 & 8 \\ -2 & -3 & 7 & 10 \end{array}\right]$$

A cada passo k:

- encontrar o pivô do passo k
- se necessário, trocar as linhas
- calcular multiplicador m_{ik}
- para $i = k+1 : n$, calcular

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)} \\ b_{i}^{(k)} &= b_{i}^{(k-1)} - m_{ik}\, b_{k}^{(k-1)}, \quad j = k : n \end{aligned}$$

Passo 1
Escolha do pivô: $\max\{2, 4, 2\} = 4$. Trocar as linhas 1 e 2.

$$\left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 4 & 9 & -3 & 8 \\ -2 & -3 & 7 & 10 \end{array}\right] \quad \Rightarrow \quad \left[\begin{array}{ccc|c} \mathbf{4} & 9 & -3 & 8 \\ 2 & 4 & -2 & 2 \\ -2 & -3 & 7 & 10 \end{array}\right]$$

$$\begin{aligned} m_{21} = 2/4 = 1/2 \quad &\Rightarrow \quad a_{2j}^{1} = a_{2j}^{0} - \tfrac{1}{2} a_{1j}^{0} \\ m_{31} = -2/4 = -1/2 \quad &\Rightarrow \quad a_{3j}^{1} = a_{3j}^{0} + \tfrac{1}{2} a_{1j}^{0}, \quad j = 1 : 3 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 4 & 9 & -3 & 8 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & -2 \\ 0 & \frac{3}{2} & \frac{11}{2} & 14 \end{array}\right]$$

Passo 2
Escolha do pivô: $\max\{\frac{1}{2}, \frac{3}{2}\} = \frac{3}{2}$. Trocar as linhas 2 e 3.

$$\left[\begin{array}{ccc|c} 4 & 9 & -3 & 8 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & -2 \\ 0 & \frac{3}{2} & \frac{11}{2} & 14 \end{array}\right] \quad \Rightarrow \quad \left[\begin{array}{ccc|c} 4 & 9 & -3 & 8 \\ 0 & \mathbf{\frac{3}{2}} & \frac{11}{2} & 14 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & -2 \end{array}\right]$$

$$m_{32} = -\tfrac{1}{2}\tfrac{2}{3} = -\tfrac{1}{3} \quad \Rightarrow \quad a_{3j}^{2} = a_{3j}^{1} + \tfrac{1}{3} a_{2j}^{1}, \quad j = 2 : 3$$

$$\left[\begin{array}{ccc|c} 4 & 9 & -3 & 8 \\ 0 & \frac{3}{2} & \frac{11}{2} & 14 \\ 0 & 0 & \frac{4}{3} & \frac{8}{3} \end{array}\right]$$

Retro-substituição

$$\left[\begin{array}{ccc|c} 4 & 9 & -3 & 8 \\ 0 & \frac{3}{2} & \frac{11}{2} & 14 \\ 0 & 0 & \frac{4}{3} & \frac{8}{3} \end{array}\right]$$

$$\frac{4}{3}x_3 = \frac{8}{3} \quad \Rightarrow \quad \boxed{x_3 = 2}$$
$$\frac{3}{2}x_2 + 2\frac{11}{2} = 14 \quad \Rightarrow \quad \boxed{x_2 = 2}$$
$$4x_1 + 9(2) - 3(2) = 8 \quad \Rightarrow \quad \boxed{x_1 = -1}$$

Portanto a solução é $\mathbf{x^T} = [-1, 2, 2]$.

4.5.2 Algoritmo da eliminação de Gauss com pivotamento parcial

O algoritmo a seguir descreve a eliminação de Gauss com a estratégia de pivotamento parcial.

Algorithm 5: Algoritmo Eliminação de Gauss com Pivotamento Parcial

```
para k = 1 : n − 1 faça
    w = |A(k,k)|;
    para j = k : n faça
        se |A(j,k)| > w então
            w = |A(j,k)|;
            r = j;
        fim-se
    fim-para
    trocaLinhas(k,r);
    para i = k + 1 : n faça
        m = A(i,k) / A(k,k);
        para j = k + 1 : n faça
            A(i,j) = A(i,j) - m*A(k,j) ;
        fim-para
        b(i) = b(i) - m*b(k) ;
    fim-para
fim-para
x = retroSubstituicao(A,b) ;
retorna x ;
```

4.5.3 Efeitos numéricos do pivotamento

Para ilustrar as consequências em aritmética de ponto flutuante de não se utilizar a estratégia de pivotamento considere o seguinte exemplo.

Exemplo 37

Seja o seguinte sistema

$$\begin{bmatrix} 0.0001 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Usando um sistema de ponto flutuante $F(10, 3, -10, 10)$ (sistema decimal com 3 dígitos na mantissa), com arredondamento, encontre a solução do sistema usando eliminação de Gauss sem pivoteamento.
Solução: temos que

$$m_{21} = \frac{1}{0.0001} = 10000 \quad \Rightarrow \quad L_2' = L_2 - 10000 L_1$$

assim

$$\left[\begin{array}{cc|c} 0.0001 & 1 & 1 \\ 0 & \mathbf{-10000}^* & \mathbf{-10000}^{**} \end{array}\right]$$

Note que $(*)$ foi obtido como

$$\begin{aligned} 1 - 10000 \times 1 &= 0.00001 \times 10^5 - 0.10000 \times 10^5 \\ &= 0.09999 \times 10^5 \\ &= \text{ (arredondando) } = 0.100 \times 10^5 \end{aligned}$$

e de forma análoga para $(**)$, temos

$$\begin{aligned} 2 - 10000 \times 1 &= 0.00001 \times 10^5 - 0.10000 \times 10^5 \\ &= 0.09998 \times 10^5 \\ &= \text{ (arredondando) } = 0.100 \times 10^5 \end{aligned}$$

Por fim, aplicando a retrosubstituição obtemos uma solução errada, devido aos erros de aritmética em ponto flutuante cometidos em $(*)$ e $(**)$ durante a soma/subtração de números muito pequenos com números muito grandes.

$$\text{Solução obtida} \quad \rightarrow \quad \mathbf{x}^T = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

A solução exata é dada por

$$\text{Solução exata} \quad \rightarrow \quad \mathbf{x}^T = \begin{bmatrix} 1.00010001 & 0.99989999 \end{bmatrix}$$

e o uso da eliminação de Gauss com pivotamento parcial chegaria à uma solução mais adequada nesse caso.

4.5.4 Pivoteamento Total

Na estratégia de pivoteamento total, o elemento escolhido como pivô é o maior elemento em módulo que ainda atua no processo de eliminação, isto é:

$$\text{Encontrar } r \text{ e } s \text{ tais que: } \; |a_{rs}| = \max |a_{ij}|, \; k \leq i, j \leq n$$

Feita a escolha do pivô é preciso trocar as linhas k e r e as colunas k e s.

Observe que a troca de colunas afeta a ordem das incógnitas do vetor $\mathbf{x}$.

Em geral o pivoteamento parcial é satisfatório, e o pivoteamento total não é muito usado devido ao alto esforço computacional requerido na busca pelo maior elemento em módulo no resto da matriz. A Figura 4.1 ilustra os procedimentos de pivotamento parcial e total.

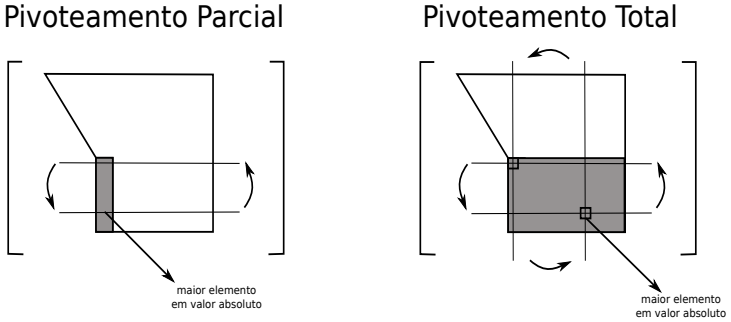


Figura 4.1: Comparação do pivotamento parcial e total.

4.6 Decomposição LU

Uma matriz quadrada pode ser escrita como o produto de duas matrizes $\mathbf{L}$ e $\mathbf{U}$, onde

- $\mathbf{L}$ é uma matriz triangular inferior unitária (elementos da diagonal principal igual a 1)
- $\mathbf{U}$ é uma matriz triangular superior

ou seja, a matriz pode ser escrita como $\mathbf{A} = \mathbf{LU}$.

Dessa forma para resolver o sistema linear $\mathbf{Ax} = \mathbf{b}$ usamos $\mathbf{A} = \mathbf{LU}$ em sua forma decomposta, isto é

$$\mathbf{LUx} = \mathbf{b}. \tag{4.3}$$

Para resolver

$$\mathbf{L} \underbrace{\mathbf{Ux}}_{\mathbf{y}} = \mathbf{b}$$

definimos o vetor $\mathbf{y}$

$$\mathbf{Ux} = \mathbf{y}$$

e resolvemos os seguintes sistemas triangulares (na sequência):

$$\mathbf{Ly} = \mathbf{b}, \tag{4.4}$$

$$\mathbf{Ux} = \mathbf{y}. \tag{4.5}$$

Portanto, para a solução de um sistema linear pela decomposição LU é preciso seguir os seguintes passos:

1. Como $\mathbf{L}$ é triangular inferior podemos resolver $\mathbf{Ly} = \mathbf{b}$ facilmente usando o algoritmo de **substituição**. Assim encontramos o vetor $\mathbf{y}$.
2. Em seguida substituimos $\mathbf{y}$ no sistema $\mathbf{Ux} = \mathbf{y}$. Como $\mathbf{U}$ é uma matriz triangular superior, podemos resolver este sistema usando o algoritmo da **retro-substituição** para encontrar a solução $\mathbf{x}$.

4.6.1 Teorema da Decomposição LU

Vamos ver agora em que condições podemos decompor uma matriz $\mathbf{A}$ na forma $\mathbf{LU}$.

Teorema 10 (LU)**.** Sejam $\mathbf{A} = (a_{ij})$ uma matriz quadrada de ordem n e $\mathbf{A_k}$ o menor principal, constituído das k primeiras linhas e k primeiras colunas de $\mathbf{A}$. Assumimos que $\det(\mathbf{A_k}) \neq 0$ para $k = 1, 2, \ldots, n-1$.

Então existe:

- uma única matriz triangular inferior $\mathbf{L} = (l_{ij})$ com $l_{ii} = 1,\ i = 1 : n$
- uma única matriz triangular superior $\mathbf{U} = (u_{ij})$

tal que $\mathbf{A} = \mathbf{LU}$. Além disso, $\det(\mathbf{A}) = u_{11}u_{22}\ldots u_{nn}$.

Prova 6. A demonstração matemática desse teorema utiliza indução matemática [**?**] e será apresentada nos passos de (i), (ii) a (iii) a seguir.
(i) Para $n = 1$ temos

$$a_{11} = 1\ a_{11} = 1\ u_{11} \quad \Rightarrow \quad u_{11} = a_{11}, l_{11} = 1$$

e ainda $\det(\mathbf{A}) = u_{11}$.
(ii) Assumimos que o teorema é verdadeiro para $n = k-1$, ou seja, que toda matriz de ordem $(k-1)$ é decomponível no produto $\mathbf{LU}$.
(iii) Vamos mostrar que podemos decompor $\mathbf{A}$ para $n = k$. Seja $\mathbf{A}$ de ordem k, escrita da forma

$$\mathbf{A} = \begin{bmatrix} \mathbf{A_{k-1}} & \mathbf{r} \\ \mathbf{s} & a_{kk} \end{bmatrix} \tag{4.6}$$

Por hipótese de indução temos que

$$\mathbf{A_{k-1}} = \mathbf{L_{k-1}U_{k-1}} \tag{4.7}$$

Usando (4.7) temos

$$\mathbf{A} = \mathbf{LU} \quad \Rightarrow \quad \mathbf{L} = \begin{bmatrix} \mathbf{L_{k-1}} & \mathbf{0} \\ \mathbf{m} & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{U_{k-1}} & \mathbf{p} \\ \mathbf{0} & u_{kk} \end{bmatrix}$$

onde $\mathbf{m}$, $\mathbf{p}$ e u_{kk} são desconhecidos. Efetuando o produto temos

$$\mathbf{LU} = \begin{bmatrix} \mathbf{L_{k-1}U_{k-1}} & \mathbf{L_{k-1}p} \\ \mathbf{mU_{k-1}} & \mathbf{mp} + u_{kk} \end{bmatrix} \tag{4.8}$$

Comparando (4.6) e (4.8)

$$\mathbf{A} = \begin{bmatrix} \mathbf{A_{k-1}} & \mathbf{r} \\ \mathbf{s} & a_{kk} \end{bmatrix} = \begin{bmatrix} \mathbf{L_{k-1}U_{k-1}} & \mathbf{L_{k-1}p} \\ \mathbf{mU_{k-1}} & \mathbf{mp} + u_{kk} \end{bmatrix}$$

Assim

$$\begin{aligned} \mathbf{A_{k-1}} &= \mathbf{L_{k-1}U_{k-1}} \\ \mathbf{r} &= \mathbf{L_{k-1}p} \\ \mathbf{s} &= \mathbf{mU_{k-1}} \\ \mathbf{mp} + u_{kk} &= a_{kk} \end{aligned}$$

Observe que

- pela hip. de indução $\mathbf{L_{k-1}}$ e $\mathbf{U_{k-1}}$ são unicamente determinadas;
- e ainda, $\mathbf{L_{k-1}}$ e $\mathbf{U_{k-1}}$ não são singulares, caso contrário $\mathbf{A_{k-1}}$ também seria, contrariando a hipótese.

Portanto

$$\begin{aligned} \mathbf{r} = \mathbf{L_{k-1}p} \quad &\Rightarrow \quad \mathbf{p} = \mathbf{L_{k-1}^{-1}r} \\ \mathbf{s} = \mathbf{mU_{k-1}} \quad &\Rightarrow \quad \mathbf{m} = \mathbf{sU_{k-1}^{-1}} \\ \mathbf{mp} + u_{kk} = a_{kk} \quad &\Rightarrow \quad u_{kk} = a_{kk} - \mathbf{mp} \end{aligned}$$

Ou seja, $\mathbf{m}$, $\mathbf{p}$ e u_{kk} são determinados unicamente nesta ordem e, portanto, $\mathbf{L}$ e $\mathbf{U}$ são determinados unicamente. Finalmente

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) = 1 \, \det(\mathbf{U}) = u_{11}u_{22}\ldots u_{nn}$$

□

4.6.2 Obtenção das matrizes L e U

Podemos obter as matrizes $\mathbf{L}$ e $\mathbf{U}$ aplicando a definição de produto e igualdade de matrizes, ou seja, impondo que $\mathbf{A}$ seja igual a $\mathbf{LU}$, onde $\mathbf{L}$ é triangular inferior unitária e $\mathbf{U}$ triangular superior. Então

$$\mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ l_{21} & 1 & 0 & \ldots & 0 \\ l_{31} & l_{32} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & 1 & 0 \\ l_{n1} & l_{n2} & l_{n3} & \ldots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ 0 & u_{22} & u_{23} & \ldots & u_{2n} \\ 0 & 0 & u_{33} & \ldots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & u_{nn} \end{bmatrix}$$

Vamos obter os elementos de **L** e **U** da seguinte forma: primeiro determina-se a 1^a linha de **U**, depois a 1^a coluna de **L**, depois a 2^a linha de **U**, depois a 2^a coluna de **L** e assim por diante, como apresentado a seguir.

1^a linha de U

$$\begin{aligned} a_{11} &= 1\ u_{11} \quad \Rightarrow \quad u_{11} = a_{11} \\ a_{12} &= 1\ u_{12} \quad \Rightarrow \quad u_{12} = a_{12} \\ &\dots \\ a_{1n} &= 1\ u_{1n} \quad \Rightarrow \quad u_{1n} = a_{1n} \end{aligned}$$

1^a coluna de L

$$\begin{aligned} a_{21} &= l_{21}\ u_{11} \quad \Rightarrow \quad l_{21} = \tfrac{a_{21}}{u_{11}} \\ a_{31} &= l_{31}\ u_{11} \quad \Rightarrow \quad l_{31} = \tfrac{a_{31}}{u_{11}} \\ &\dots \\ a_{n1} &= l_{n1}\ u_{11} \quad \Rightarrow \quad l_{n1} = \tfrac{a_{n1}}{u_{11}} \end{aligned}$$

2^a linha de U

$$\begin{aligned} a_{22} &= l_{21}u_{12} + 1\ u_{22} \quad \Rightarrow \quad u_{22} = a_{22} - l_{21}u_{12} \\ a_{23} &= l_{21}u_{13} + 1\ u_{23} \quad \Rightarrow \quad u_{23} = a_{23} - l_{21}u_{13} \\ &\dots \\ a_{2n} &= l_{21}u_{1n} + 1\ u_{2n} \quad \Rightarrow \quad u_{2n} = a_{2n} - l_{21}u_{1n} \end{aligned}$$

2^a coluna de L

$$\begin{aligned} a_{32} &= l_{31}u_{12} + l_{32}u_{22} \quad \Rightarrow \quad l_{32} = \frac{a_{32} - l_{31}u_{12}}{u_{22}} \\ a_{42} &= l_{41}u_{12} + l_{42}u_{22} \quad \Rightarrow \quad l_{42} = \frac{a_{42} - l_{41}u_1}{u_{22}} \\ &\dots \\ a_{n2} &= l_{n1}u_{12} + l_{n2}u_{22} \quad \Rightarrow \quad l_{n2} = \frac{a_{n2} - l_{n1}u_{12}}{u_{22}} \end{aligned}$$

De forma geral temos as seguintes fórmulas para determinar as componentes das matrizes **L** e **U**:

$$\boxed{u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}\ u_{kj}, \quad i \leq j} \tag{4.9}$$

$$\boxed{l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\ u_{kj} \right) \Big/ u_{jj}, \quad i > j} \tag{4.10}$$

as quais podem ser facilmente implementadas em um algoritmo.

Observação: na prática as matrizes **L** e **U** nunca são criadas e alocadas explicitamente na implementação computacional. O que fazemos é sobrescrever as entradas da matriz original **A** com as entradas de **L** e **U**.

Algorithm 6: Cálculo das matrizes L e U.

para $i = 1 : n$ **faça**
  para $j = i : n$ **faça**
    $u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}\ u_{kj}$;
  fim-para
  para $j = i + 1 : n$ **faça**
    $l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\ u_{kj}\right) \Big/ u_{jj}$;
  fim-para
fim-para

4.6.3 LU via eliminação de Gauss

O método da eliminação de Gauss pode ser interpretado como um método para obtenção das matrizes **L** e **U**. No processo da EG no passo 1, eliminamos as entradas abaixo de a_{11} na coluna 1 da matriz fazendo

$$\begin{aligned} \text{Para } i = 2 : n \quad m_{i1} &= \frac{a_{i1}}{a_{11}} \\ a_{ij}^1 &= a_{ij}^0 - m_{i1}\ a_{1j}^0 \\ b_i^1 &= b_i^0 - m_{i1}\ b_1^0, \quad j = 1 : n \end{aligned}$$

essa operação é equivalente a multiplicar $(\mathbf{A}|\mathbf{b})^0$ por uma matriz $\mathbf{M_1}$, para obter $(\mathbf{A}|\mathbf{b})^1$, onde

$$\mathbf{M_1} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -m_{21} & 1 & 0 & \dots & 0 \\ -m_{31} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ -m_{n1} & 0 & \dots & 0 & 1 \end{bmatrix}$$

Assim

$$\begin{aligned} \mathbf{M_1}(\mathbf{A}|\mathbf{b})^0 &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -m_{21} & 1 & 0 & \dots & 0 \\ -m_{31} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ -m_{n1} & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & a_{22}^1 & \dots & a_{2n}^1 & b_2^1 \\ 0 & a_{32}^1 & \dots & a_{3n}^1 & b_3^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^1 & \dots & a_{nn}^1 & b_n^1 \end{bmatrix} = (\mathbf{A}|\mathbf{b})^1 \end{aligned}$$

No passo seguinte temos

$$(\mathbf{A}|\mathbf{b})^2 = \mathbf{M_2}(\mathbf{A}|\mathbf{b})^1$$

$$\mathbf{M_2}(\mathbf{A}|\mathbf{b})^1 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -m_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & -m_{n2} & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & a_{22}^1 & \dots & a_{2n}^1 & b_2^1 \\ 0 & a_{32}^1 & \dots & a_{3n}^1 & b_3^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^1 & \dots & a_{nn}^1 & b_n^1 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & a_{22}^1 & \dots & a_{2n}^1 & b_2^1 \\ 0 & 0 & \dots & a_{3n}^2 & b_3^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & a_{nn}^2 & b_n^2 \end{bmatrix} = (\mathbf{A}|\mathbf{b})^2$$

Procedemos dessa forma, até que por fim temos

$$\begin{aligned}(\mathbf{A}|\mathbf{b})^{(n-1)} &= \mathbf{M_{n-1}}(\mathbf{A}|\mathbf{b})^{(n-2)} \\ &= \dots = \underbrace{\mathbf{M_{n-1}}\mathbf{M_{n-2}}\dots\mathbf{M_2}\mathbf{M_1}}_{\mathbf{M}}(\mathbf{A}|\mathbf{b})^{(0)}\end{aligned}$$

Deste modo temos

$$\mathbf{A^{(n-1)}} = \mathbf{MA} = \mathbf{U}$$

onde $\mathbf{U}$ é a matriz triangular superior da decomposição LU. Como $\mathbf{M}$ é um produto de matrizes não-singulares, $\mathbf{M}$ é inversível, isto é,

$$\begin{aligned}\mathbf{M} &= \mathbf{M_{n-1}}\mathbf{M_{n-2}}\dots\mathbf{M_2}\mathbf{M_1} \\ \mathbf{M^{-1}} &= \mathbf{M_1^{-1}}\mathbf{M_2^{-1}}\dots\mathbf{M_{n-2}^{-1}}\mathbf{M_{n-1}^{-1}}\end{aligned}$$

Portanto

$$\mathbf{MA} = \mathbf{U} \quad \Rightarrow \quad \mathbf{A} = \underbrace{\mathbf{M^{-1}}}_{\mathbf{L}}\mathbf{U}$$

$$\mathbf{MA} = \mathbf{U} \quad \Rightarrow \quad \mathbf{A} = \underbrace{\mathbf{M^{-1}}}_{\mathbf{L}}\mathbf{U}$$

onde

$$\mathbf{M^{-1}} = \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ m_{21} & 1 & 0 & \dots & 0 \\ m_{31} & m_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ m_{n1} & m_{n2} & m_{n3} & \dots & 1 \end{bmatrix}$$

é a matriz triangular inferior da decomposição LU.

Exemplo 1 Decomponha a matriz **A** dada abaixo nos fatores **L** e **U**, usando a eliminação de Gauss.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix}$$

Solução do Exemplo 1

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix}}_{\mathbf{U}}$$

Exemplo 2 Resolva o seguinte sistema linear:

$$\begin{bmatrix} 1 & 2 & -1 \\ 2 & 3 & -2 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$$

Solução do Exemplo 2

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{x}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Cálculo do Determinante Veremos como utilizar a decomposição $\mathbf{A} = \mathbf{LU}$ para calcular o determinante da matriz.

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U})$$

O determinante de uma matriz triangular é dado pelo produto dos elementos da diagonal principal, isto é

$$\begin{aligned} \det(\mathbf{L}) &= 1 \\ \det(\mathbf{U}) &= u_{11}u_{22}u_{33}\ldots u_{nn} \end{aligned}$$

Portanto

$$\begin{aligned} \det(\mathbf{A}) &= \det(\mathbf{L})\ \det(\mathbf{U}) \\ &= 1\ \det(\mathbf{U}) \\ &= u_{11}u_{22}u_{33}\ldots u_{nn} \end{aligned}$$

Exemplo 2 Para o exemplo anterior, temos

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Portanto o determinante é

$$\det(\mathbf{A}) = 1\ (-1)\ 2 = -2$$

□

4.7 Decomposição LU com Pivoteamento Parcial

Vamos estudar agora o uso de pivoteamento parcial para a decomposição LU. Para definir o que significa, de forma matricial, a troca de duas linhas de uma matriz, iremos apresentar o conceito de matrizes de permutação.

Uma matriz de permutação é uma matriz obtida a partir da matriz identidade através de uma reordenação de suas linhas, isto é

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Portanto se $\mathbf{P}$ é uma matriz de permutação e $\mathbf{A}$ uma matriz qualquer, então

- $\mathbf{PA}$ é uma versão da matriz $\mathbf{A}$ com as linhas permutadas
- $\mathbf{AP}$ é uma versão da matriz $\mathbf{A}$ com as colunas permutadas

Na prática (implementação) uma matriz de permutação $\mathbf{P}$ de dimensão $n \times n$ nunca é armazenada explicitamente. É muito mais eficiente representar $\mathbf{P}$ por um vetor `p` de valores inteiros de tamanho n.

Uma forma de implementar isso é fazer com que `p[k]` seja o índice da coluna que tem apenas um "1"na k-ésima linah de $\mathbf{P}$. Para o exemplo anterior

```
p = [3  2  1]
```

Para aplicar a estratégia de pivoteamento parcial nos exercícios, basta trocar efetivamente as linhas da matriz. Vejamos um exemplo.

Para resolver $\mathbf{Ax} = \mathbf{b}$, segue-se o procedimento:

- Calcular $\mathbf{P}$, $\mathbf{L}$ e $\mathbf{U}$ tal que $\mathbf{PA} = \mathbf{LU}$
- Atualizar vetor $\mathbf{b} = \mathbf{Pb}$
- Resolver $\mathbf{Ly} = \mathbf{b}$
- Resolver $\mathbf{Ux} = \mathbf{y}$

Dicas para calcular $\mathbf{L}$ e $\mathbf{U}$ via eliminação de Gauss com pivoteamento:

- se trocar linhas, atualizar o vetor $\mathbf{p}$;
- guardar os multiplicadores da eliminação de Gauss na posição que foi zerada, ao invés de colocar os zeros.

Exemplo 3 Resolver o sistema linear abaixo usando a decomposição LU com pivoteamento parcial.

$$\begin{bmatrix} 3 & -4 & 1 \\ 1 & 2 & 2 \\ 4 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 3 \\ -2 \end{bmatrix}$$

Solução do Exemplo 3

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{3}{4} & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 4 & 0 & -3 \\ 0 & -4 & \frac{13}{4} \\ 0 & 0 & \frac{35}{8} \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}$$

$$\mathbf{x^T} = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}$$

Solução do Exemplo 3 - Passo a passo Etapa 1

$$\begin{bmatrix} 3 & -4 & 1 \\ 1 & 2 & 2 \\ 4 & 0 & -3 \end{bmatrix}, \quad p = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

Troca as linhas 1 e 3 e atualiza vetor p

$$\begin{bmatrix} 4 & 0 & -3 \\ 1 & 2 & 2 \\ 3 & -4 & 1 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}$$

Elimina e guarda os multiplicadores nas suas posições (em azul):

$$\begin{bmatrix} 4 & 0 & -3 \\ 1/4 & 2 & 11/4 \\ 3/4 & -4 & 13/4 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}$$

Etapa 2

$$\begin{bmatrix} 4 & 0 & -3 \\ 1/4 & 2 & 11/4 \\ 3/4 & -4 & 13/4 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}$$

Troca as linhas 2 e 3 e atualiza vetor p

$$\begin{bmatrix} 4 & 0 & -3 \\ 3/4 & -4 & 13/4 \\ 1/4 & 2 & 11/4 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}$$

Elimina e guarda os multiplicadores nas suas posições (em azul):

$$\begin{bmatrix} 4 & 0 & -3 \\ 3/4 & -4 & 13/4 \\ 1/4 & -1/2 & 35/8 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}$$

Resultado

$$\begin{bmatrix} 4 & 0 & -3 \\ 3/4 & -4 & 13/4 \\ 1/4 & -1/2 & 35/8 \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}$$

Decomposição $\mathbf{PA} = \mathbf{LU}$:

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 3 & -4 & 1 \\ 1 & 2 & 2 \\ 4 & 0 & -3 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 3/4 & 1 & 0 \\ 1/4 & -1/2 & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 4 & 0 & -3 \\ 0 & -4 & 13/4 \\ 0 & 0 & 35/8 \end{bmatrix}$$

Para resolver $PAx = Pb \quad \Rightarrow \quad LUx = Pb$, define-se $Ux = y$ e então:

1. Resolva $Ly = Pb$
2. Resolva $Ux = y$

Procedendo desta forma, chega-se em

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{3}{4} & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 1 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 4 & 0 & -3 \\ 0 & -4 & \frac{13}{4} \\ 0 & 0 & \frac{35}{8} \end{bmatrix}, \quad p = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix}$$

$$\mathbf{x^T} = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}$$

□Fim do exemplo

4.8 Revisitando algumas definições

Definição 10 (Matriz Simétrica)**.** Uma matriz real $\mathbf{A} \in \mathbb{R}^{n \times n}$ é simétrica se possui as mesmas entradas acima e abaixo da diagonal principal, isto é, se

$$a_{ij} = a_{ji}, \quad \forall\, i, j$$

Portanto $\mathbf{A} = \mathbf{A}^T$.

Tais matrizes satisfazem a seguinte relação

$$\mathbf{x}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}\mathbf{x}, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

Definição 11 (Matriz Positiva Definida)**.** Se a matriz $\mathbf{A}$ é simétrica, então é dita ser positiva definida se

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \quad \forall \mathbf{x} \neq 0$$

Matriz Positiva Definida: de imediato verifica-se que se $\mathbf{A}$ é não singular, caso contrário haveria um $\mathbf{x}$ diferente de zero tal que $\mathbf{Ax} = \mathbf{0}$.

Além disso, escolhendo vetores escritos na forma

$$\mathbf{x^T} = \begin{bmatrix} x_1 & x_2 & \dots & x_k & 0 & 0 & \dots & 0 \end{bmatrix}$$

podemos verificar que todas as matrizes menores principais ($\mathbf{A_k}$) são positivas definidas, portanto não singular ($\det(\mathbf{A_k}) \neq 0$) e consequentemente podemos decompor $\mathbf{A}$ na forma $\mathbf{A} = \mathbf{LU}$.

Na prática muitas matrizes que surgem em aplicações de engenharias e ciências são simétricas e positiva definidas, devido a leis físicas que estão por trás da origem dessas matrizes.

Testes para matrizes positivas definidas

1. Critério de Sylvester: uma matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ é positiva definida, se e somente se

$$\det(\mathbf{A_k}) > 0, \quad k = 1, 2, \ldots, n$$

   onde $\mathbf{A_k}$ é a matriz menor principal de ordem k (a matriz $k \times k$ formada pelas k primeiras linhas e pelas k primeiras colunas).

2. Se realizarmos a eliminação de Gauss *sem troca de linha ou coluna na matriz* $\mathbf{A}$, podemos dizer que $\mathbf{A}$ é positiva definida, se e somente se, **todos os pivôs forem positivos**.

Exemplo Verifique se as seguintes matrizes são positivas definidas:

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 4 & 4 & 2 \\ 4 & 3 & 0 \\ 2 & 0 & 5 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

4.9 Decomposição de Cholesky

Quando a matriz do sistema linear é simétrica, podemos simplificar os cálculo da decomposição LU levando em conta a simetria da matriz. Essa é a idéia do método de Cholesky. Se $\mathbf{A}$ é simétrica positiva definida, pelo critério de Sylvester temos que

$$\det(\mathbf{A_k}) > 0$$

portanto, todos os menores principais são não singulares e consequentemente pelo Teorema da decomposição LU, a matriz pode ser escrita como $\mathbf{A} = \mathbf{LU}$. Se $\mathbf{A}$ é simétrica, então $\mathbf{A} = \mathbf{A}^T$. Logo

$$\mathbf{LU} = \mathbf{A} = \mathbf{A}^T = (\mathbf{LU})^T = \mathbf{U}^T\mathbf{L}^T$$

Assim

$$\begin{aligned} \mathbf{LU} &= \mathbf{U}^T\mathbf{L}^T \\ \mathbf{L}^{-1}\mathbf{LU} &= \mathbf{L}^{-1}\mathbf{U}^T\mathbf{L}^T \\ \mathbf{U} &= \mathbf{L}^{-1}\mathbf{U}^T\mathbf{L}^T \\ \mathbf{U}(\mathbf{L}^T)^{-1} &= \mathbf{L}^{-1}\mathbf{U}^T\mathbf{L}^T(\mathbf{L}^T)^{-1} \\ \mathbf{U}(\mathbf{L}^T)^{-1} &= \mathbf{L}^{-1}\mathbf{U}^T \end{aligned}$$

Temos que

$$\underbrace{\mathbf{U}(\mathbf{L}^T)^{-1}}_{\text{triangular superior}} = \underbrace{\mathbf{L}^{-1}\mathbf{U}^T}_{\text{triangular inferior}}$$

Portanto, essa igualdade só pode ser uma matriz diagonal! Vamos definir essa matriz diagonal como

$$\mathbf{D} = \mathbf{U}(\mathbf{L}^T)^{-1}. \tag{4.11}$$

Note que também poderíamos escolher $\mathbf{D} = \mathbf{L}^{-1}\mathbf{U}^T$. A matriz $\mathbf{D}$ é a mesma, só a forma calculá-la é que muda. Aqui iremos nos limitar à escolha $\mathbf{D} = \mathbf{U}(\mathbf{L}^T)^{-1}$ e analisar as suas consequências. Logo, adotando $\mathbf{D} = \mathbf{U}(\mathbf{L}^T)^{-1}$ podemos escrever $\mathbf{U}$ como

$$\mathbf{D} = \mathbf{U}(\mathbf{L}^T)^{-1} \quad \Rightarrow \quad \mathbf{D}\mathbf{L}^T = \mathbf{U}(\mathbf{L}^T)^{-1}\mathbf{L}^T \quad \Rightarrow \quad \mathbf{U} = \mathbf{D}\mathbf{L}^T \tag{4.12}$$

Sendo assim temos que

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T \tag{4.13}$$

Assim de (4.13), como todos $d_{ii} > 0$, podemos escrever

$$\mathbf{A} = \mathbf{L}\mathbf{U} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{L}(\mathbf{D})^{1/2}(\mathbf{D})^{1/2}\mathbf{L}^T = \mathbf{G}\mathbf{G}^T$$

onde

$$\mathbf{G} = \mathbf{L}(\mathbf{D})^{1/2},$$
$$\mathbf{G}^T = (\mathbf{D})^{1/2}\mathbf{L}^T.$$

A matriz diagonal $\mathbf{D}$ pode ser facilmente calculada uma vez que $\mathbf{U}$ é encontrado considerando que $\mathbf{U} = \mathbf{D}\mathbf{L}^T$ e o resultado do produto de uma matriz diagonal por uma triangular inferior unitária, isto é

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} d_{11} & d_{11}l_{21} & d_{11}l_{31} \\ 0 & d_{22} & d_{22}l_{32} \\ 0 & 0 & d_{33} \end{bmatrix} \tag{4.14}$$

de onde conclui-se que a matriz diagonal $\mathbf{D}$ é simplesmente a diagonal da matriz $\mathbf{U}$, isto é:

$$\mathbf{D} = \begin{bmatrix} u_{11} & 0 & 0 \\ 0 & u_{22} & 0 \\ 0 & 0 & u_{33} \end{bmatrix} \quad \Rightarrow \quad \mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{u_{11}} & 0 & 0 \\ 0 & \sqrt{u_{22}} & 0 \\ 0 & 0 & \sqrt{u_{33}} \end{bmatrix}. \tag{4.15}$$

4.9.1 Obtenção da matriz G da decomposição de Cholesky

A decomposição de Cholesky é um caso especial da fatoração LU aplicada para matrizes simétricas e positiva definida (SPD) e sua decomposição pode ser obtida a partir de

$$\mathbf{A} = \mathbf{G}\mathbf{G}^T$$

onde $\mathbf{G}$ é uma matriz triangular inferior tal que

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

Pelo produto e igualdade de matrizes podemos obter os elementos de $\mathbf{G}$. Elementos da diagonal principal:

$$\begin{aligned} a_{11} &= g_{11}^2 \\ a_{22} &= g_{21}^2 + g_{22}^2 \\ &\vdots \\ a_{nn} &= g_{n1}^2 + g_{n2}^2 + \ldots + g_{nn}^2 \end{aligned}$$

de forma geral

$$g_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} g_{ik}^2}\ , \quad i = 1 : n \tag{4.16}$$

Para os elementos fora da diagonal principal, temos

$$\begin{aligned} a_{21} &= g_{21} g_{11} \\ a_{31} &= g_{31} g_{11} \\ &\vdots \\ a_{n1} &= g_{n1} g_{11} \\ a_{32} &= g_{31} g_{21} + g_{32} g_{22} \\ a_{42} &= g_{41} g_{21} + g_{42} g_{22} \\ &\vdots \\ a_{n2} &= g_{n1} g_{21} + g_{n2} g_{22} \end{aligned}$$

de forma geral

$$g_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} g_{ik} g_{jk}}{g_{jj}}\ , \quad i = j+1 : n, \quad j = 1 : n \tag{4.17}$$

Observando as equações (4.16) e (4.17), vemos que podemos calcular os elementos de $\mathbf{G}$ a cada passo j do método da seguinte forma:

- calcula-se termo da diagonal principal: g_{jj}
- calcula-se termos da coluna j abaixo da diagonal principal: g_{ij} com $i = j + 1 : n$

O algoritmo abaixo descreve o a forma de se obter a matriz $\mathbf{G}$ da decomposição de Cholesky.

Algorithm 7: Decomposição de Cholesky

para $j = 1 : n$ **faça**

$$g_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2} \ ;$$

para $i = j + 1 : n$ **faça**

$$g_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} g_{ik} g_{jk} \right) \Big/ g_{jj} \ ;$$

fim-para

fim-para

4.9.2 Observações sobre a decomposição de Cholesky

Algumas observações sobre o método:

- Se $\mathbf{A}$ é SPD, então a aplicação do método de Cholesky requer menos operações de ponto flutuante do que a decomposição LU.
- Como $\mathbf{A}$ é positiva definida, isto garante que só teremos raízes quadradas de números positivos, isto é, os termos $a_{jj} - \sum_{k=1}^{j-1} g_{jk}^2$ são sempre maiores do que zero.
  - Exemplo do caso 2×2
- Caso o algoritmo falhe, podemos concluir que $\mathbf{A}$ não é simétrica e positiva definida.
- Determinante

$$\det(\mathbf{A}) = \det(\mathbf{G}\mathbf{G}^T) = \det(\mathbf{G})\det(\mathbf{G}^T) = \det(\mathbf{G})^2 = (g_{11} g_{22} \dots g_{nn})^2$$

4.9.3 Solução de sistema pela decomposição de Cholesky

Podemos usar a decomposição de Cholesky para encontrar a solução de $\mathbf{A}\mathbf{x} = \mathbf{b}$ da seguinte forma:

1. Determinar a decomposição

$$\mathbf{A} = \mathbf{G}\mathbf{G}^T$$

então

$$\mathbf{G} \underbrace{\mathbf{G}^T \mathbf{x}}_{\mathbf{y}} = \mathbf{b}$$

2. Resolver $\mathbf{G}\mathbf{y} = \mathbf{b}$, usando substituição
3. Resolver $\mathbf{G}^T\mathbf{x} = \mathbf{y}$, retro-substituição

Exemplo 38

Decomposição de Cholesky Considere a matriz

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 2 \\ -2 & 10 & -7 \\ 2 & -7 & 30 \end{bmatrix}$$

a) Verificar se $\mathbf{A}$ satisfaz as condições da decomposição de Cholesky

b) Decompor $\mathbf{A}$ em $\mathbf{G}\mathbf{G}^T$

c) Calcular o determinante

d) Resolver o sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ com $\mathbf{b} = \begin{bmatrix} 8 \\ 11 \\ -31 \end{bmatrix}$

Solução do Exemplo

a) $\mathbf{A}$ é simétrica e positiva definida

$$\det(\mathbf{A_1}) = 4, \quad \det(\mathbf{A_2}) = 36, \quad \det(\mathbf{A_3}) = 900$$

b) A decomposição é

$$\mathbf{A} = \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & -2 & 5 \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} 2 & -1 & 1 \\ 0 & 3 & -2 \\ 0 & 0 & 5 \end{bmatrix}}_{\mathbf{G}^T}$$

c) $\det(\mathbf{A}) = (2 \cdot 3 \cdot 5)^2 = 30^2 = 900$

d) $\mathbf{x} = \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}$

Podemos usar as fórmulas (4.16) e (4.17) para calcular os elementos da matriz $\mathbf{G}$ da decomposição, mas também podemos proceder de outra forma. A ideia é:

- Decompor $\mathbf{A} = \mathbf{L}\mathbf{U}$ via eliminação de Gauss
- Como $\mathbf{U} = \mathbf{D}\mathbf{L}^T$, calcular $\mathbf{D} = \text{diag}(\mathbf{U})$ (a matriz $\mathbf{D}$ é a diagonal de $\mathbf{U}$)
- Calcular $\mathbf{D}^{1/2}$
- E assim calcular $\mathbf{G} = \mathbf{L}\mathbf{D}^{1/2}$

Exemplo 39

A partir da decomposição LU da matriz $\mathbf{A}$ do exemplo anterior, obtenha $\mathbf{G}$.

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 2 \\ -2 & 10 & -7 \\ 2 & -7 & 30 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & -\frac{2}{3} & 1 \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} 4 & -2 & 2 \\ 0 & 9 & -6 \\ 0 & 0 & 25 \end{bmatrix}}_{\mathbf{U}}$$

Exercício Mostrar que, se o sistema linear $\mathbf{Ax} = \mathbf{b}$, onde $\mathbf{A}$ é não singular, é transformado no sistema linear equivalente

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$$

então esse último sistema linear pode sempre ser resolvido pelo método de Cholesky (isto é $\mathbf{B} = \mathbf{A}^T\mathbf{A}$ satisfaz as condições para a aplicação do método).

Aplicar a técnica anterior para encontrar a solução do seguinte sistema linear:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}$$

Dicas para o exercício: Mostre que $\mathbf{B}$ satisfaz as condições da decomposição de Cholesky. Será preciso usar

$$||\mathbf{x}|| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}, \quad ||\mathbf{x}||^2 = x_1^2 + x_2^2 + \ldots + x_n^2 = \mathbf{x}^T\mathbf{x}.$$

4.10 Decomposição $\mathbf{LDL}^T$

Como vimos anteriormente também podemos decompor $\mathbf{A}$ na forma $\mathbf{A} = \mathbf{LDL}^T$, onde $\mathbf{L}$ é uma matriz triangular inferior unitária e $\mathbf{D}$ é uma matriz diagonal.

De forma análoga ao que fizemos para a decomposição de Cholesky, podemos determinar os elementos da decomposição da seguinte forma:

$$d_{jj} = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_{kk} \ , \quad j = 1 : n$$

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_{kk} l_{jk}}{d_{jj}} \quad j = 1 : n-1, \quad i = j+1 : n$$

A solução do sistema linear $\mathbf{Ax} = \mathbf{b}$ é dada por

$$\mathbf{Ax} = \mathbf{b} \quad \Rightarrow \quad \mathbf{LD}\underbrace{\mathbf{L}^T\mathbf{x}}_{\mathbf{y}} = \mathbf{b}$$

$$\Rightarrow \quad \mathbf{L}\underbrace{\mathbf{Dy}}_{\mathbf{w}} = \mathbf{b}$$

e assim temos os seguintes passos para a solução do sistema:

1. $\mathbf{Lw} = \mathbf{b}$
2. $\mathbf{Dy} = \mathbf{w}$
3. $\mathbf{L}^T\mathbf{x} = \mathbf{y}$

Cálculo do determinante

$$\begin{aligned}\det(\mathbf{A}) &= \det(\mathbf{L})\det(\mathbf{D})\det(\mathbf{L}^T)\\ &= 1 \cdot \det(\mathbf{D}) \cdot 1 = d_{11}d_{22}\ldots d_{nn}\end{aligned}$$

Algorithm 8: Cálculo de L e D

para $j = 1 : n$ **faça**

$$d_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\ d_{kk}}\ ;$$

para $i = j + 1 : n$ **faça**

$$l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\ d_{kk}\ l_{jk}\right) \Big/ d_{jj}\ ;$$

fim-para

fim-para

4.11 Cálculo da Matriz Inversa

Iremos descrever como calcular a matriz inversa através da decomposição LU. Sejam $\mathbf{A}$ uma matriz de dimensão n, não singular ($\det(\mathbf{A} \neq 0)$ e $\mathbf{A}^{-1}$ a matriz inversa de $\mathbf{A}$. Vamos escrever a matriz inversa como:

$$\mathbf{A}^{-1} = \left[\begin{array}{c|c|c|c} \mathbf{v_1} & \mathbf{v_2} & \ldots & \mathbf{v_n} \end{array}\right]$$

Seja ainda $\mathbf{e_j}$ a coluna j da matriz identidade. Por exemplo, $\mathbf{e_2} = \begin{bmatrix}0 & 1 & 0 & \ldots & 0\end{bmatrix}$, $\mathbf{e_n} = \begin{bmatrix}0 & 0 & 0 & \ldots & 1\end{bmatrix}$. Resolvendo o seguinte sistema linear

$$\mathbf{Av_1} = \mathbf{e_1}$$

encontramos a primeira coluna $\mathbf{v_1}$ da matriz inversa de $\mathbf{A}$. Repetindo o procedimento para cada coluna temos

$$\mathbf{Av_j} = \mathbf{e_j}, \quad j = 1 : n \tag{4.18}$$

Agora basta usar algum dos métodos que vimos para resolver os sistemas lineares da equação (4.18). Veja algumas opções:

1. **Decomposição LU**

$$\mathbf{LUv_j} = \mathbf{e_j}, \quad j = 1 : n$$

Basta fatorar a matriz na forma LU uma única vez, e com os fatores resolver os seguintes sistemas

$$\mathbf{L}\mathbf{y_j} = \mathbf{e_j}$$
$$\mathbf{U}\mathbf{v_j} = \mathbf{y_j}$$

2. **Decomposição de Cholesky** somente se a matriz for SPD:

$$\mathbf{G}\mathbf{G}^T\mathbf{v_j} = \mathbf{e_j} \quad \Rightarrow \quad (1)\ \mathbf{G}\mathbf{y_j} = \mathbf{e_j}, \quad (2)\ \mathbf{G}^T\mathbf{v_j} = \mathbf{y_j}$$

3. **Eliminação de Gauss**.
Montar

$$\left[\ \mathbf{A}\ \middle|\ \mathbf{I}\ \right]$$

e efetuar a eliminação de Gauss de uma vez só. Assim obtemos

$$\left[\ \mathbf{U}\ \middle|\ \mathbf{T}\ \right]$$

onde $\mathbf{T}$ é uma matriz triangular inferior. Em seguida dado que temos $\mathbf{U}$ triangular superior, basta resolver a seguinte sequência de sistemas

$$\mathbf{U}\mathbf{v_j} = \mathbf{t_j}$$

onde $\mathbf{t_j}$ é a coluna j da matriz $\mathbf{T}$.

Exemplo 40

Calcular a inversa da seguinte matriz

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & -6 \\ 3 & 2 & -6 \\ 3 & 1 & -5 \end{bmatrix}$$

Assim temos

$$\left[\begin{array}{ccc|ccc} 4 & 1 & -6 & 1 & 0 & 0 \\ 3 & 2 & -6 & 0 & 1 & 0 \\ 3 & 1 & -5 & 0 & 0 & 1 \end{array}\right]$$

Efetuando a eliminação de Gauss obtemos

$$\left[\begin{array}{ccc|ccc} 4 & 1 & -6 & 1 & 0 & 0 \\ 0 & 5/4 & -3/2 & -3/4 & 1 & 0 \\ 0 & 0 & -1/5 & -3/5 & -1/5 & 1 \end{array}\right]$$

Agora basta resolver

$$\begin{bmatrix} 4 & 1 & -6 \\ 0 & 5/4 & -3/2 \\ 0 & 0 & -1/5 \end{bmatrix} \mathbf{v_1} = \begin{bmatrix} 1 \\ -3/4 \\ -3/5 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 1 & -6 \\ 0 & 5/4 & -3/2 \\ 0 & 0 & -1/5 \end{bmatrix} \mathbf{v_2} = \begin{bmatrix} 0 \\ 1 \\ -1/5 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 1 & -6 \\ 0 & 5/4 & -3/2 \\ 0 & 0 & -1/5 \end{bmatrix} \mathbf{v_3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

4.12 Métodos Iterativos

O sistema de equações lineares $\mathbf{Ax} = \mathbf{b}$ pode ser resolvido por um processo que gera a partir de um vetor inicial $\mathbf{x^{(0)}}$ uma sequência de vetores $\mathbf{x^{(1)}}$, $\mathbf{x^{(2)}}$, $\mathbf{x^{(3)}}$, ... que deve convergir para a solução.

Existem muitos métodos iterativos para a solução de sistemas lineares, entretanto só iremos estudar os chamados **métodos iterativos estacionários**.

Algumas perguntas importantes são:

- Como construir a sequência $\{\mathbf{x^{(0)}}, \mathbf{x^{(1)}}, \mathbf{x^{(2)}}, \ldots\}$?
- $\mathbf{x^{(k)}} \to \mathbf{x}^*$?
- Quais são as condições para convergência?
- Como saber se $\mathbf{x^{(k)}}$ está próximo de $\mathbf{x}^*$?
- Critério de parada?

Um método iterativo escrito na forma

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c} \tag{4.19}$$

é dito *estacionário* quando a matriz $\mathbf{B}$ for fixa durante o processo iterativo. Veremos como construir a matriz $\mathbf{B}$ para cada um dos métodos que iremos estudar: **Jacobi**, **Gauss-Seidel** e **Sobre-relaxação (SOR)**.

Antes, é preciso rever alguns conceitos como norma de vetores e matrizes, os quais serão importantes no desenvolvimento do critério de parada e na análise de convergência dos métodos.

4.12.1 Normas de Vetores e Matrizes

Para discutir o erro envolvido nas aproximações é preciso associar a cada vetor e matriz um valor escalar não negativo que de alguma forma mede sua magnitude. As normas para vetores mais comuns são:

- Norma euclideana (ou norma L_2)

$$||\mathbf{x}||_2 = (x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2}$$

- Norma infinito (ou norma do máximo)

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

Normas vetoriais devem satisfazer às seguintes propriedades:

1. $||\mathbf{x}|| > 0$ se $\mathbf{x} \neq 0$, $||\mathbf{x}|| = 0$ se $\mathbf{x} = \mathbf{0}$
2. $||\alpha\mathbf{x}|| = \alpha||\mathbf{x}||$, onde α é um escalar
3. $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$

Normas de matrizes tem que satisfazer a propridades similares:

1. $||\mathbf{A}|| > 0$ se $\mathbf{A} \neq 0$, $||\mathbf{A}|| = 0$ se $\mathbf{A} = \mathbf{0}$
2. $||\alpha\mathbf{A}|| = \alpha||\mathbf{A}||$, onde α é um escalar
3. $||\mathbf{A} + \mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$
4. $||\mathbf{AB}|| \leq ||\mathbf{A}||\ ||\mathbf{B}||$
5. $||\mathbf{Ax}|| \leq ||\mathbf{A}||\ ||\mathbf{x}||$

Iremos fazer uso em diversos momentos da seguinte norma matricial

$$||\mathbf{A}||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

Exemplo 41: Norma das linhas para matrizes

$$\mathbf{A} = \begin{bmatrix} 4 & 6 \\ -3 & 4 \end{bmatrix} \quad \Rightarrow \quad ||\mathbf{A}||_\infty = \max\{10, 7\} = 10$$

4.12.2 Critério de Parada

A distância entre dois vetores $\mathbf{x}$ e $\mathbf{y}$ pode ser calculada como

$$||\mathbf{x} - \mathbf{y}||_2 \quad \text{ou} \quad ||\mathbf{x} - \mathbf{y}||_\infty.$$

Iremos usar a norma infinito nos algoritmos que iremos descrever. Seja $\mathbf{x^{(k+1)}}$ e $\mathbf{x^{(k)}}$ duas aproximações para o vetor solução $\mathbf{x}^*$ de um sistema de equações lineares.

O seguinte critério de parada pode ser adotado

$$\frac{||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||_\infty}{||\mathbf{x}^{(k+1)}||_\infty} = \frac{\max |x_i^{(k+1)} - x_i^{(k)}|}{\max |x_i^{(k+1)}|} < \varepsilon$$

onde ε é a precisão desejada (Ex: 10^{-3}).

Na prática também adotamos um número máximo de iterações para evitar que o programa execute indefinidamente, caso o método não convirja para um determinado problema.

$$k < k_{max}$$

4.12.3 Método de Jacobi

Vamos ilustrar a idéia do método de Jacobi através de um exemplo. Seja o seguinte sistema:

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

o qual pode ser escrito como

$$\begin{aligned} x_1 &= (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 &= (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \\ x_3 &= (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \end{aligned}$$

A partir de uma aproximação inicial

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix}$$

Calculamos uma nova aproximação, denotada por $\mathbf{x}^{(1)}$, através de

$$\begin{aligned} x_1^{(1)} &= \left(b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)}\right)/a_{11} \\ x_2^{(1)} &= \left(b_2 - a_{21}x_1^{(0)} - a_{23}x_3^{(0)}\right)/a_{22} \\ x_3^{(1)} &= \left(b_3 - a_{31}x_1^{(0)} - a_{32}x_2^{(0)}\right)/a_{33} \end{aligned}$$

Após obter $\mathbf{x}^{(1)}$, calculamos $\mathbf{x}^{(2)}$ substituindo $\mathbf{x}^{(1)}$ no lugar de $\mathbf{x}^{(0)}$ na expressão anterior e assim procedemos até que o critério de parada seja satisfeito.

De forma geral o método de Jacobi para um sistema de n equações e n incógnitas consiste em, a cada passo k, calcular a seguinte aproximação da componente x_i do vetor solução da seguinte maneira:

Algorithm 9: Método de Jacobi

entrada: $\mathbf{A}$, $\mathbf{b}$, $\mathbf{x}^{(0)}$, max, ε
saída: $\mathbf{x}$

para $k = 1 : max$ **faça**
  para $i = 1 : n$ **faça**
    $x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right) \Big/ a_{ii}$;
  fim-para
  se $\max |x_i^{(k+1)} - x_i^{(k)}| < \varepsilon$ **então**
    retorna $\mathbf{x}^{(k+1)}$;
  fim-se
fim-para

Exemplo 42: Jacobi

Utilizando o método de Jacobi com $\mathbf{x}^{(0)} = \mathbf{0}$ o seguinte sistema:

$$\begin{aligned} 4x_1 + 0.24x_2 - 0.08x_3 &= 8 \\ 0.09x_1 + 3x_2 - 0.15x_3 &= 9 \\ 0.04x_1 - 0.08x_2 + 4x_3 &= 20 \end{aligned}$$

Solução do Exemplo: Temos a seguinte fórmula de iteração do método de Jacobi para este sistema linear:

$$\begin{aligned} x_1^{(k+1)} &= 2 - 0.06x_2^{(k)} + 0.02x_3^{(k)} \\ x_2^{(k+1)} &= 3 - 0.03x_1^{(k)} + 0.05x_3^{(k)} \\ x_3^{(k+1)} &= 5 - 0.01x_1^{(k)} + 0.02x_2^{(k)} \end{aligned}$$

Passo 1 $\rightarrow \mathbf{x}^{(0)} = \mathbf{0}$

$$\begin{aligned} x_1^{(1)} &= 2 - 0.06x_2^{(0)} + 0.02x_3^{(0)} = 2 \\ x_2^{(1)} &= 3 - 0.03x_1^{(0)} + 0.05x_3^{(0)} = 3 \\ x_3^{(1)} &= 5 - 0.01x_1^{(0)} + 0.02x_2^{(0)} = 5 \end{aligned}$$

Passo 2 $\rightarrow (\mathbf{x}^{(1)})^T = \begin{bmatrix} 2 & 3 & 5 \end{bmatrix}$

$$\begin{aligned} x_1^{(2)} &= 2 - 0.06(3) + 0.02(5) = 2 - 0.08 = 1.92 \\ x_2^{(2)} &= 3 - 0.03(2) + 0.05(5) = 3 + 0.19 = 3.19 \\ x_3^{(2)} &= 5 - 0.01(2) + 0.02(3) = 5 + 0.04 = 5.04 \end{aligned}$$

Passo 3 $\rightarrow (\mathbf{x}^{(2)})^T = \begin{bmatrix} 1.92 & 3.19 & 5.04 \end{bmatrix}$

$$\begin{aligned} x_1^{(3)} &= 2 - 0.06(3.19) + 0.02(5.04) = 1.91 \\ x_2^{(3)} &= 3 - 0.03(1.92) + 0.05(5.04) = 3.1944 \\ x_3^{(3)} &= 5 - 0.01(1.92) + 0.02(3.19) = 5.0446 \end{aligned}$$

Em resumo tem-se:

k	0	1	2	3
x_1	0	2	1.92	1.91
x_2	0	3	3.19	3.1944
x_3	0	5	5.04	5.0446

com um erro absoluto dado por $||\mathbf{x}^{(3)} - \mathbf{x}^{(2)}||_\infty = \max\{0.01, 0.0044, 0.0046\} = 0.01$.

4.12.4 Método de Gauss-Seidel

Observe no exemplo anterior, que o método de Jacobi, não usa os valores atualizados de $\mathbf{x}^{(k)}$ até completar por inteiro a iteração do passo k. O método de Gauss-Seidel pode ser visto como uma modificação do método de Jacobi. Nele usaremos a mesma forma de iterar que o método de Jacobi, entretanto vamos aproveitar os cálculos já atualizados, de outras componentes, para atualizar a componente que está sendo calculada. Dessa forma o valor de $x_1^{(k+1)}$ será usado para calcular $x_2^{(k+1)}$, os valores de $x_1^{(k+1)}$ e $x_2^{(k+1)}$ serão usados para calcular $x_3^{(k+1)}$, e assim por diante.

Para um sistema 3×3 temos o seguinte esquema:

$$
\begin{aligned}
x_1^{(k+1)} &= \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/a_{11} \\
x_2^{(k+1)} &= \left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right)/a_{22} \\
x_3^{(k+1)} &= \left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right)/a_{33}
\end{aligned}
$$

No caso geral temos

para $i = 1 : n$ **faça**

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \Big/ a_{ii} \ ;$$

fim-para

Note que no método de Gauss-Seidel apenas 1 aproximação para x_i precisa ser armazenada. No método de Jacobi é preciso manter 2 vetores em memória, um para $\mathbf{x}^{(k+1)}$ e outro para $\mathbf{x}^{(k)}$.

Exemplo 43

Resolva o sistema de equações do exemplo anterior usando o método de Gauss-Seidel.

Solução do Exemplo: Fórmula de iteração

$$
\begin{aligned}
x_1^{(k+1)} &= 2 - 0.06x_2^{(k)} + 0.02x_3^{(k)} \\
x_2^{(k+1)} &= 3 - 0.03x_1^{(k+1)} + 0.05x_3^{(k)} \\
x_3^{(k+1)} &= 5 - 0.01x_1^{(k+1)} + 0.02x_2^{(k+1)}
\end{aligned}
$$

Passo $1 \rightarrow \mathbf{x}^{(0)} = \mathbf{0}$

$$
\begin{aligned}
x_1^{(1)} &= 2 - 0.06(0) + 0.02(0) = 2 \\
x_2^{(1)} &= 3 - 0.03(2) + 0.05(0) = 3 - 0.06 = 2.94 \\
x_3^{(1)} &= 5 - 0.01(2) + 0.02(2.94) = 5.0388
\end{aligned}
$$

Passo 2 $\rightarrow (\mathbf{x}^{(1)})^T = \begin{bmatrix} 2 & 2.94 & 5.0388 \end{bmatrix}$

$$\begin{aligned} x_1^{(2)} &= 2 - 0.06(2.94) + 0.02(5.0388) = 1.924376 \\ x_2^{(2)} &= 3 - 0.03(1.924376) + 0.05(5.0388) = 3.194209 \\ x_3^{(2)} &= 5 - 0.01(1.924376) + 0.02(3.194209) = 5.044640 \end{aligned}$$

Passo 3 $\rightarrow (\mathbf{x}^{(2)})^T = \begin{bmatrix} 1.924376 & 3.194209 & 5.044640 \end{bmatrix}$

$$\begin{aligned} x_1^{(2)} &= 2 - 0.06(1.924376) + 0.02(5.04464) = 1.909240 \\ x_2^{(2)} &= 3 - 0.03(1.909240) + 0.05(5.04464) = 3.194955 \\ x_3^{(2)} &= 5 - 0.01(1.909240) + 0.02(3.194955) = 5.044807 \end{aligned}$$

4.12.5 Convergência dos métodos de Jacobi e Gauss-Seidel

Para estudar a convergência dos métodos, vamos primeiros escrevê-los na seguinte forma:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}$$

Para isso, vamos dividir a matriz $\mathbf{A}$ como

$$\mathbf{A} = \underbrace{\mathbf{L}}_{\text{triangular inferior}} + \underbrace{\mathbf{D}}_{\text{diagonal}} + \underbrace{\mathbf{U}}_{\text{triangular superior}}$$

isto é, para uma matriz 3×3 temos

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{bmatrix} + \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} + \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

Sendo assim o método de Jacobi pode ser escrito como:

$$\begin{aligned} \mathbf{A}\mathbf{x} = \mathbf{b} \quad &\Rightarrow \quad (\mathbf{L} + \mathbf{D} + \mathbf{U})\mathbf{x} = \mathbf{b} \\ &\Rightarrow \quad \mathbf{D}\mathbf{x} = \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x} \end{aligned}$$

e assim

$$\begin{aligned} \mathbf{D}\mathbf{x}^{(k+1)} &= \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} \\ \mathbf{x}^{(k+1)} &= -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \\ \mathbf{x}^{(k+1)} &= \mathbf{B}_J\mathbf{x}^{(k)} + \mathbf{c} \end{aligned}$$

onde para o método de Jacobi

$$\begin{aligned} \mathbf{B}_J &= -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) \\ \mathbf{c} &= \mathbf{D}^{-1}\mathbf{b} \end{aligned}$$

Para o método de Gauss-Seidel temos

$$\begin{aligned}(\mathbf{L}+\mathbf{D})\mathbf{x}^{(k+1)} &= -\mathbf{U}\mathbf{x}^{(k)} + \mathbf{b} \\ \mathbf{x}^{(k+1)} &= -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b} \\ \mathbf{x}^{(k+1)} &= \mathbf{B}_{GS}\mathbf{x}^{(k)} + \mathbf{c}\end{aligned}$$

onde para o método de Gauss-Seidel

$$\begin{aligned}\mathbf{B}_{GS} &= -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U} \\ \mathbf{c} &= (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b}\end{aligned}$$

Ou seja, ambos os métodos podem ser escritos como

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c} \tag{4.20}$$

onde $\mathbf{B}$ é chamada de matriz de iteração

$$\begin{aligned}\mathbf{B}_J &= -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) \\ \mathbf{B}_{GS} &= -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U}\end{aligned}$$

Se o método de Jacobi ou Gauss-Seidel converge ou não, depende dos autovalores da matriz de iteração $\mathbf{B}$.

Dizemos que λ_i, $i = 1:n$ é um autovalor da matriz $\mathbf{B}$ se

$$\mathbf{B}\mathbf{u} = \lambda_i \mathbf{u}$$

para algum vetor $\mathbf{u} \neq \mathbf{0}$. O seguinte teorema caracteriza a condição para convergência desses métodos.

Teorema 11. A condição necessária e suficiente para que o método iterativo descrito por $\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}$ convirja usando um vetor inicial $\mathbf{x}^{(0)}$ qualquer é

$$\rho(\mathbf{B}) = \max_{1\leq i\leq n} |\lambda_i(\mathbf{B})| < 1$$

Na prática encontrar os autovalores de $\mathbf{B}$ é tão custoso quanto resolver um sistema de equações lineares e portanto o Teorema 1 é difícil de usar. Vamos estudar outra forma de analisar a convergência para esses métodos.

Seja $\mathbf{x}^*$ a solução exata. Então $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$. Subtraindo de (4.20) temos

$$\begin{aligned}\mathbf{x}^{(k+1)} - \mathbf{x}^* &= \mathbf{B}\mathbf{x}^{(k)} - \mathbf{B}\mathbf{x}^* + \mathbf{c} - \mathbf{c} \\ &= \mathbf{B}(\mathbf{x}^{(k)} - \mathbf{x}^*)\end{aligned}$$

de forma análoga

$$\mathbf{x}^{(k)} - \mathbf{x}^* = \mathbf{B}(\mathbf{x}^{(k-1)} - \mathbf{x}^*)$$

e assim

$$\mathbf{x}^{(k+1)} - \mathbf{x}^* = \mathbf{B}^2(\mathbf{x}^{(k-1)} - \mathbf{x}^*) = \ldots = \mathbf{B}^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}^*)$$

$$\mathbf{x}^{(k+1)} - \mathbf{x}^* = \mathbf{B}^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}^*) \tag{4.21}$$

Aplicando a norma infinito em (4.21), obtemos

$$\begin{aligned} ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||_\infty &= ||\mathbf{B}^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}^*)||_\infty \\ &\leq ||\mathbf{B}^{k+1}||_\infty \; ||(\mathbf{x}^{(0)} - \mathbf{x}^*)||_\infty \\ &\leq ||\mathbf{B}||_\infty^{k+1} \; ||(\mathbf{x}^{(0)} - \mathbf{x}^*)||_\infty \end{aligned} \tag{4.22}$$

Assim de (4.22) fica claro que só haverá convergência se

$$||\mathbf{B}||_\infty < 1 \tag{4.23}$$

Vamos analisar agora critérios específicos para atender ao critério geral dado por (4.23) para o método de Jacobi e Gauss-Seidel.

Para o método de Jacobi, a matriz de iteração $\mathbf{B}_J = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ é da forma

$$\mathbf{D} = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix} \quad \Rightarrow \quad \mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & & & \\ & \frac{1}{a_{22}} & & \\ & & \ddots & \\ & & & \frac{1}{a_{nn}} \end{bmatrix}$$

portanto

$$\mathbf{B}_J = - \begin{bmatrix} 0 & \frac{a_{12}}{a_{11}} & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 0 & \frac{a_{23}}{a_{22}} & \cdots & \frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \cdots & \frac{a_{nn-1}}{a_{nn}} & 0 \end{bmatrix}$$

ou seja, seus elementos são

$$b_{ij} = -\frac{a_{ij}}{a_{ii}}$$

Para termos convergência, então precisamos que $||\mathbf{B}_J||_\infty < 1$

$$||\mathbf{B}_J||_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |b_{ij}| = \max_{1 \leq i \leq n} \sum_{j=1, i \neq j}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1$$

Teorema 12 (Critério das Linhas)**.** Seja $\mathbf{Ax} = \mathbf{b}$ e seja $\alpha_k = \displaystyle\sum_{j=1, i \neq j}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|$, para $k = 1 : n$. Se $\alpha = \max\{\alpha_k\} < 1$, então o método de Jacobi converge independentemente da aproximação inicial $\mathbf{x}^{(0)}$.

Exemplo 44

Verificar se as seguintes matrizes satisfazem o critério das linhas.

$$\begin{bmatrix} 4 & 0.24 & -0.08 \\ 0.09 & 3 & -0.15 \\ 0.04 & -0.08 & 4 \end{bmatrix}, \quad \begin{bmatrix} 1 & 3 & 1 \\ 5 & 2 & 2 \\ 0 & 6 & 8 \end{bmatrix}$$

Definição 12. Uma matriz $\mathbf{A}$ é estritamente diagonal dominante se

$$\sum_{j=1,\ j\neq i}^{n} |a_{ij}| < |a_{ii}|, \quad i = 1:n$$

Fica claro então que para matrizes estritamente diagonal dominante o critério das linhas é sempre satisfeito. Portanto, uma outra forma de verificar se o método de Jacobi converge para uma certa matriz é verificar se esta é estritamente diagonal dominante.

Exemplo 45

$$\begin{bmatrix} 10 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 3 & 10 \end{bmatrix}$$

$$\begin{aligned} |a_{12}| + |a_{13}| &= |2| + |1| < |10| = |a_{11}| \\ |a_{21}| + |a_{23}| &= |1| + |1| < |5| = |a_{22}| \\ |a_{31}| + |a_{32}| &= |2| + |3| < |10| = |a_{33}| \end{aligned}$$

Para ter convergência é preciso satisfazer pelo menos um dos critérios:

- critério das linhas, isto é

$$\boxed{\max_{1\leq i\leq n} \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1}$$

- critério de Sassenfeld

$$\boxed{\max_{1\leq i\leq n} \beta_i < 1} \tag{4.24}$$

  onde β_i são calculados como

$$\beta_i = \left(\sum_{j=1}^{i-1} |a_{ij}|\beta_j + \sum_{j=i+1}^{n} |a_{ij}| \right) \Big/ |a_{ii}|$$

É possível mostrar que para $\mathbf{B}_{GS}$ dado por

$$\mathbf{B}_{GS} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$$

temos que

$$||\mathbf{B}_{GS}||_\infty \leq \max_{1\leq i\leq n} \beta_i$$

Sendo assim, para mostrar que o método converge, basta mostrar que o critério de Sassenfeld (4.24) é satisfeito.

Para ver que o critério das linhas também é válido para o método de Gauss-Seidel, basta verificar que

$$\max_{1\leq i\leq n} \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \quad \Rightarrow \quad \beta_i < 1, \quad i = 1:n$$

Prova: Considere que: $\max \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1$ (CL $\rightarrow$ OK)

$$\beta_1 = \sum_{j=2}^{n} \frac{|a_{1j}|}{|a_{11}|} \leq \max_{1\leq i\leq n} \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

Suponha agora que $\beta_j < 1$ para $i = 1, 2, \ldots, i-1$. Então

$$\beta_i = \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|}\beta_j + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \leq \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|}$$
$$\leq \max_{1\leq i\leq n} \sum_{j=1,\ j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

Fica claro que o critério de Sassenfeld pode ser menor que o das linhas. Logo, o critério de Sassenfeld pode ser satisfeito e o critério das linhas não, e portanto o processo iterativo converge.

Algumas observações:

- Para um certo sistema de equações lineares pode acontecer do método de Jacobi convergir, enquanto o Gauss-Seidel não, ou vice-versa.
- Quanto menor o valor de $||\mathbf{B}||_\infty$, mais rápida será a convergência do método.
- Permutação de linhas ou colunas pode reduzir $||\mathbf{B}||_\infty$
- A convergência dos métodos de Jacobi e Gauss-Seidel não depende do vetor inicial $\mathbf{x}^{(0)}$.

Exemplo 46

Resolva o sistema utilizando o método de Jacobi.

$$\begin{bmatrix} 10 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -8 \\ 6 \end{bmatrix}$$

Solução do Exemplo Critério das linhas:

$$\alpha_1 = (|a_{12}| + |a_{13}|)/|10| = 0.2 + 0.1 = 0.3 < 1$$
$$\alpha_2 = (|a_{21}| + |a_{23}|)/|5| = 0.2 + 0.2 = 0.4 < 1$$
$$\alpha_3 = (|a_{31}| + |a_{32}|)/|10| = 0.2 + 0.3 = 0.5 < 1$$

Logo $\alpha = \alpha_3 = 0.5 < 1$ e portanto o método de Jacobi converge para essa matriz.

Ou então basta verificar que a matriz $\mathbf{A}$ é estritamente diagonal dominante.

Fórmula de iteração:

$$x_1^{(k+1)} = 0.7 - 0.2x_2^{(k)} - 0.1x_3^{(k)}$$
$$x_2^{(k+1)} = -1.6 - 0.2x_1^{(k)} - 0.2x_3^{(k)}$$
$$x_3^{(k+1)} = 0.6 - 0.2x_1^{(k)} - 0.3x_2^{(k)}$$

Assim temos as seguintes iterações para o vetor inicial $\mathbf{x}^{(0)} = \mathbf{0}$

k	1	2	3	4	5
x_1	0.7	0.96	0.978	0.9994	0.9979
x_2	-1.6	-1.86	-1.98	-1.9888	-1.9996
x_3	0.6	0.94	0.966	0.966	0.9968

Exemplo 47: Exemplo

Resolva o sistema utilizando o método de Gauss-Seidel.

$$\begin{bmatrix} 5 & 1 & 1 \\ 3 & 4 & 1 \\ 3 & 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ 0 \end{bmatrix}$$

Solução do Exemplo

1) A matriz não é estritamente diagonal dominante. Nada podemos afirmar sobre a convergência.

2) Critério das linhas:

$$\alpha_1 = (|a_{12}| + |a_{13}|)/|5| = 0.2 + 0.2 = 0.4 < 1$$
$$\alpha_2 = (|a_{21}| + |a_{23}|)/|4| = 0.75 + 0.25 = 1$$
$$\alpha_3 = (|a_{31}| + |a_{32}|)/|6| = 0.5 + 0.5 = 1$$

Não satisfaz o critério das linhas.

3) Critério de Sassenfeld:

$$\beta_1 = |0.2| + |0.2| = 0.4$$
$$\beta_2 = |0.75|(0.4) + |0.25| = 0.3 + 0.25 = 0.55$$
$$\beta_3 = |0.5|(0.4) + |0.5|(0.55) = 0.2 + 0.275 = 0.475$$

Assim

$$\max_{1\leq i\leq n} \beta_i = \max\{0.4, 0.55, 0.475\} = 0.55 < 1$$

Portanto, como o critério de Sassenfeld é satisfeito, podemos garantir que o processo de Gauss-Seidel converge para essa matriz.

Fórmula de iteração:

$$\begin{aligned} x_1^{(k+1)} &= 1 - 0.2x_2^{(k)} - 0.2x_3^{(k)} \\ x_2^{(k+1)} &= 1.5 - 0.75x_1^{(k+1)} - 0.25x_3^{(k)} \\ x_3^{(k+1)} &= 0 - 0.5x_1^{(k+1)} - 0.5x_2^{(k+1)} \end{aligned}$$

Usando $\mathbf{x}^{(0)} = \mathbf{0}$ como aproximação inicial, temos

$$\begin{aligned} x_1^{(1)} &= 1 - 0.2(0) - 0.2(0) = 1 \\ x_2^{(1)} &= 1.5 - 0.75(1) - 0.25(0) = 0.75 \\ x_3^{(1)} &= 0 - 0.5(1) - 0.5(0.75) \end{aligned}$$

Iterando para $k = 1, 2, \ldots$ temos

k	1	2	3	4
x_1	1.0	1.025	1.0075	1.0016
x_2	0.75	0.95	0.9913	0.9987
x_3	-0.875	-0.9875	-0.9994	-1.0002

Podemos verificar o erro

$$\begin{aligned} \frac{||\mathbf{x}^{(4)} - \mathbf{x}^{(3)}||_\infty}{||\mathbf{x}^{(4)}||_\infty} &= \frac{\max\{|1.0016-1.0075|,|0.9987-0.9913|,|-1.0002+0.9994|\}}{\max\{|1.0016|,|0.9987|,|-1.0002|\}} \\ &= \frac{0.0074}{1.0016} = 0.0074 < 10^{-2} \end{aligned}$$

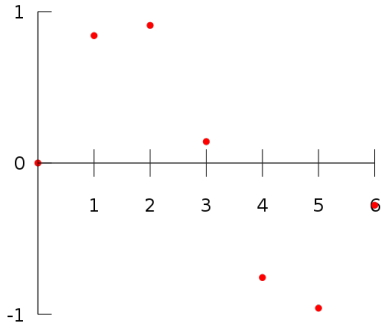
Capítulo 5

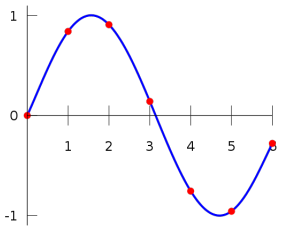
Interpolação Polinomial

5.1 Introdução

Suponha que temos um conjunto de pontos $x_0, x_1, \ldots, x_n$ e os valores de uma funçãos $f(x)$ nestes pontos $y_0 = f(x_0), \ldots, y_n = f(x_n)$. Interpolar a função $f(x)$ nos pontos $x_1, \ldots, x_n$ consiste em aproximá-la por uma função $g(x)$ tal que:

$$g(x_0) = y_0$$
$$\ldots$$
$$g(x_2) = y_2$$




1
0
-1
1
2
3
4
5
6

Capítulo 6

Método dos Mínimos Quadrados

Existem duas classes de métodos para a aproximação de dados, e a distinção entre elas está em considerarmos, ou não, a existência de erro nos dados.

No primeiro caso, consideramos que não existem erros nos dados e podemos exigir que a curva passe pelos pontos dados. Como vimos, esse problema é resolvido com **interpolação**. Nesse caso aproximamos uma função $f(x)$ por uma função polinomial $p(x)$ que passa exatamente pelos pontos dados $(x_0, y_0), \ldots, (x_n, y_n)$. Assim dado $p(x)$ é possível estimar o valor de $f(\bar{x})$ para um ponto $\bar{x}$ diferente dos pontos x_i, para $i = 0, 1, \ldots, n$.

A outra classe de métodos, que será estudada nesse capítulo, leva em consideração possíveis erros introduzidos na obtenção dos dados (limitações do instrumento, condições experimentais, etc). Nesse caso, o **método dos mínimos quadrados** tem sido amplamente utilizado.

Para ilustrar a idéia, considere agora o problema de determinar a constante de uma mola. A Lei de Hooke nos diz que o deslocamento de uma mola é proporcional à força nela aplicada, isto é $F = kx$.

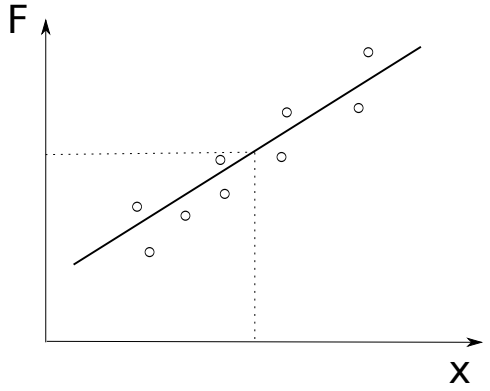
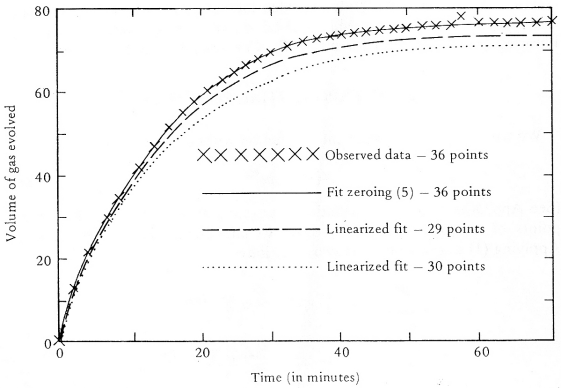


Figura 6.1: Descrever figura

A questão é como encontrar essa constante de proporcionalidade a partir de dados experimentais. Suponha que possamos realizar vários experimentos medindo forças aplicadas à mola e os seus respectivos deslocamentos. O problema consiste em encontrar uma reta que melhor aproxime esses dados. A inclinação da reta irá nos fornecer a constante k da mola.

Outro exemplo consiste da seguinte função: $f(x) = ae^{bx}$. Exemplos de dados experimentais de uma situação desse tipo podem ser vistos na Figura abaixo.



O método dos mínimos quadrados é uma das técnicas mais usadas em problemas práticos porque em geral buscamos aproximações para dados que são medidas obtidas experimentalmente e possuem um certo grau de incerteza.

Aqui compara-se o método dos mínimos quadrados com a técnica de interpolação vista no Capítulo X.

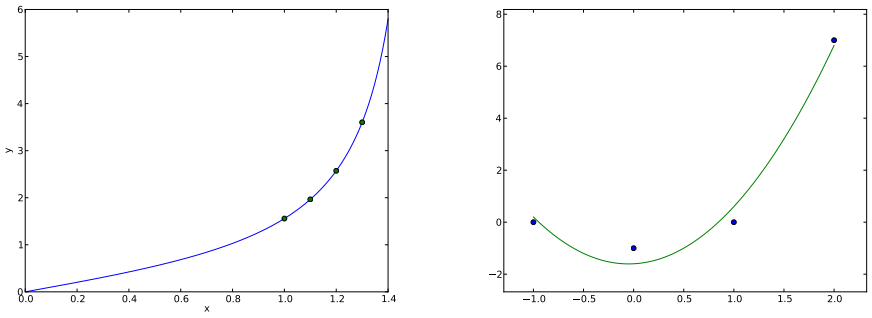


Figura 6.2: Interpolação versus Mínimos Quadrados.

6.1 Introdução

Seja $f(x)$ a função que deseja-se aproximar por uma outra função $g(x)$. Em muitas aplicações, nem sempre uma expressão analítica para $f(x)$ está disponível. No método dos mínimos quadrados (MMQ) considera-se que algumas informações sobre a forma de $g(x)$ são conhecidas. Pode-se concluir, através da observação dos dados (como no caso da Figura 6.1), por exemplo, que $g(x)$ é uma reta,

$$g(x) = c_0 + c_1 x$$

ou que é uma parábola

$$g(x) = c_0 + c_1 x + c_2 x^2$$

ou que tenha alguma outra forma específica.

De forma geral, no **caso linear**, vamos considerar que a aproximação será por uma função do tipo:

$$g(x) = c_0\phi_0(x) + c_1\phi_1(x) + \ldots + c_n\phi_n(x) \tag{6.1}$$

onde $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$ são funções pré-estabelecidas.

6.1.1 MMQ linear

: a função $g(x)$ que aproxima $f(x)$ é linear nos seus parâmetros $c_0, c_1, \ldots, c_n$.

Exemplos:

$$\phi_0(x) = 1, \quad \phi_1(x) = x, \quad \phi_2(x) = x^2, \quad \ldots$$
$$\phi_0(x) = \sin(\pi x), \quad \phi_1(x) = \sin(2\pi x), \quad \ldots$$

Veremos mais adiante como trabalhar no **caso não-linear**, como por exemplo quando $g(x) = c_0\ e^{c_1 x}$.

Para cada conjunto de coeficientes c_i, $i = 0, 1, \ldots, n$, o desvio ou resíduo da aproximação dada pela equação (6.1) no ponto x_k é dado por

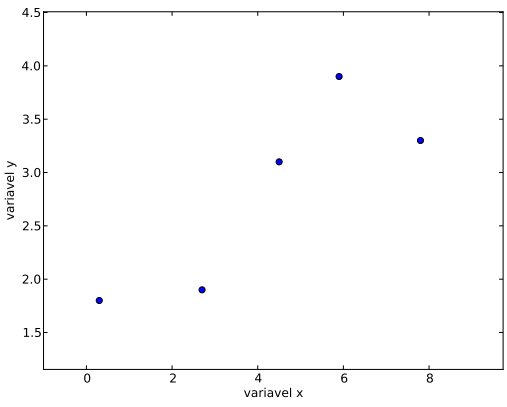
$$\begin{aligned} r(x_k) &= f(x_k) - g(x_k) \\ &= f(x_k) - [c_0\phi_0(x_k) + c_1\phi_1(x_k) + \ldots + c_n\phi_n(x_k)] \\ &= r(x_k; c_0, c_1, \ldots, c_n) \end{aligned}$$

Precisamos de estabelecer critérios de aproximação para encaminhar o problema da determinação dos parâmetros $c_0, \ldots, c_n$ que nos levarão à *melhor aproximação*.

Vamos ver um exemplo para começar a discussão. Suponha que temos a seguinte tabela de dados

x_i	0.3	2.7	4.5	5.9	7.8
y_i	1.8	1.9	3.1	3.9	3.3

Vamos analisar os dados.



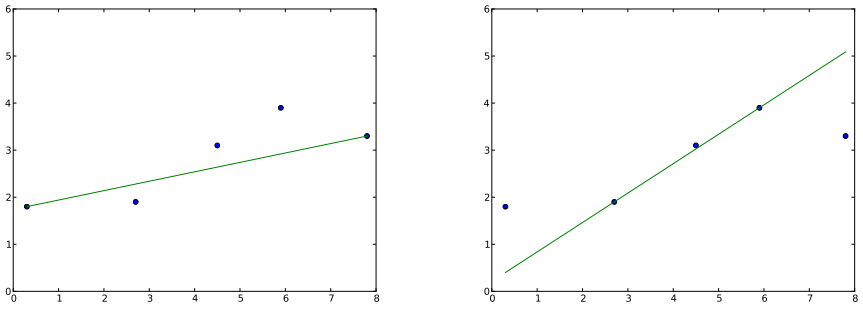


Figura 6.3: Descrever.

Podemos considerar que existe uma relação aproximadamente linear entre as variáveis. Logo, podemos desprezar alguns pontos da tabela, escolher 2 pontos e usar polinômios interpoladores lineares para aproximação. Ou seja $g(x) = c_0 + c_1 x$.

- Para $(0.3, 1.8)$ e $(7.8, 3.3)$ temos $g_1(x) = 1.74 + 0.2x$.

- Para $(2.7, 1.9)$ e $(5.9, 3.9)$ temos $g_2(x) = 0.2125 + 0.625x$.

Qual aproximação é melhor? $g_1(x)$ ou $g_2(x)$? Como verificar a qualidade da aproximação?

Para obter as aproximações g_1 e g_2 desprezamos vários dados da tabela para fazer a interpolação, o que não é muito conveniente de se fazer. Um modo de se verificar a qualidade da aproximação é calculando a soma de todas as distâncias verticais de $f(x_i)$ e $g(x_i)$ ao quadrado, isto é

$$\begin{aligned} E^2 = \sum_{i=1}^{m} r(x_i)^2 &= \sum_{i=1}^{m} [f(x_i) - g(x_i)]^2 \\ &= \sum_{i=1}^{m} [f(x_i) - c_0 - c_1 x_i]^2 \end{aligned}$$

A Figura 6.4 mostra, para o exemplo em questão, o erro em cada ponto x_i entre $f(x)$ e uma aproximação dada por $g(x)$. Note que, como nesse caso a função $g_1(x)$ que interpola os pontos (x_0, y_0) e (x_4, y_4) foi usada, o erro nesse pontos é nulo.

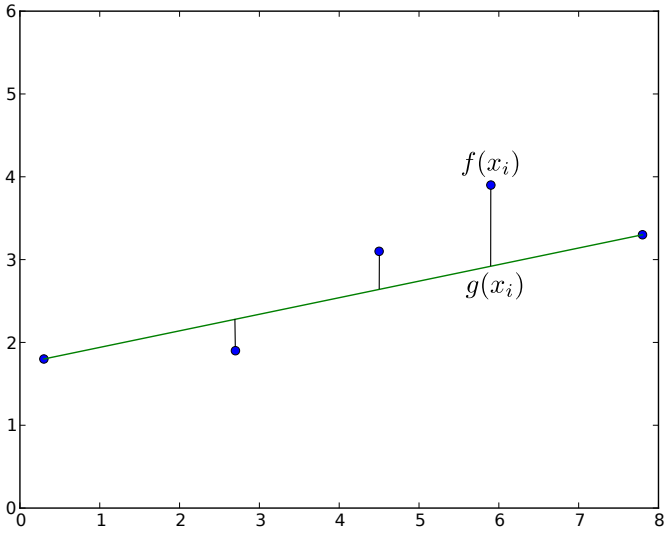


Figura 6.4: Descrever.

Para $g_1(x) = 1.74 + 0.2x$

Tabela 6.1: Descrever

i	x_i	$f(x_i)$	$g(x_i)$	r_i^2
1	0.3	1.8	1.80	0
2	2.7	1.9	2.78	0.144
3	4.5	3.1	2.64	0.2116
4	5.9	3.9	2.94	0.9604
5	7.8	3.3	3.30	0

ou seja, tem-se que $E^2 = 0^2 + 0.144^2 + 0.2116^2 + 0.9604^2 + 0^2 = 1.316$.
Para $g_2(x) = 0.2125 + 0.625x$

Tabela 6.2: Descrever

i	x_i	$f(x_i)$	$g(x_i)$	r_i^2
1	0.3	1.8	0.400	1.96
2	2.7	1.9	1.900	0
3	4.5	3.1	3.025	0.0056
4	5.9	3.9	3.900	0
5	7.8	3.3	2.275	1.051

e assim, $E^2 = 3.0166$. Ou seja, a soma dos erros (ou resíduos) ao quadrado para $g_1(x)$ é menor do que para a função $g_2(x)$ e, portanto, $g_1(x)$ é mais adequado de acordo com esse critério (soma dos erros ao quadrado).

A questão que surge então é como escolher $g(x)$ (nesse exemplo, isto se resume em encontrar c_0 e c_1) de forma que esse erro E^2 seja o menor possível? A resposta está no método dos mínimos quadrados.

6.2 Mínimos quadrados

O método dos mínimos quadrados consiste na procura de parâmetros $(c_0, c_1, \ldots, c_n)$ que minimizem a soma dos quadrados dos resíduos no caso discreto. O caso contínuo será discutido adiante.

Desta forma, introduzindo a notação do produto escalar temos

$$< r, r >= \sum_{i=1}^{m} [r(x_i)]^2 = \sum_{i=1}^{m} [f(x_i) - g(x_i)]^2$$

Como vimos $< r, r >$ é função dos parâmetros $c_0, c_1, \ldots, c_n$, isto é, $< r, r >=< r, r > (c_0, c_1, \ldots, c_n)$.

Antes de estudar o método de forma mais geral, em sua versão discreta, vamos ver um exemplo simples: **regressão linear**. Na regressão linear a função aproximadora é linear e tem a seguinte forma: $g(x) = c_0 + c_1 x$.

6.2.1 Regressão Linear

Dada uma tabela de dados $(x_i, f(x_i))$, $i = 1, \ldots, m$, queremos encontrar a reta que melhor se ajusta aos dados no critério dos mínimos quadrados. Como o ajuste será feito por uma reta temos que $\phi_0(x) = 1$ e $\phi_2(x) = x$, ou seja

$$f(x) \approx g(x) = c_0 + c_1 x$$

o resíduo é dado por

$$r = f(x) - c_0 - c_1 x$$

Pelo método dos mínimos quadrados devemos encontrar c_0 e c_1 que minimizem a função

$$< r, r > =< f(x) - c_0 - c_1 x, f(x) - c_0 - c_1 x > \tag{6.2}$$

$$= \sum_{i=1}^{m} [f(x_i) - c_0 - c_1 x_i]^2 \tag{6.3}$$

Sabe-se do Cálculo que a condição necessária de ponto crítico é que as derivadas nele sejam nulas, isto é

$$\frac{\partial}{\partial c_0} < r, r >= 0, \quad \frac{\partial}{\partial c_1} < r, r >= 0$$

Derivando a Eq. (6.2) com relação a c_0 temos

$$\begin{aligned}\frac{\partial}{\partial c_0} < r, r > &= \frac{\partial}{\partial c_0}\left\{ \sum_{i=1}^{m}[f(x_i) - c_0 - c_1 x_i]^2 \right\} \\ &= -2\sum_{i=1}^{m}[f(x_i) - c_0 - c_1 x_i]\end{aligned}$$

De forma similar, derivando a equação (6.3) com relação a c_1 temos

$$\begin{aligned}\frac{\partial}{\partial c_1} < r, r > &= \frac{\partial}{\partial c_1}\left\{ \sum_{i=1}^{m}[f(x_i) - c_0 - c_1 x_i]^2 \right\} \\ &= -2\sum_{i=1}^{m}[f(x_i) - c_0 - c_1 x_i](x_i)\end{aligned}$$

Assim, a condição de ponto crítico é que:

$$\begin{aligned}\frac{\partial}{\partial c_0} < r, r >= 0 \quad &\Rightarrow \quad \sum_{i=1}^{m}[f(x_i) - c_0 - c_1 x_i] = 0 \\ \frac{\partial}{\partial c_1} < r, r >= 0 \quad &\Rightarrow \quad \sum_{i=1}^{m}(x_i)[f(x_i) - c_0 - c_1 x_i] = 0\end{aligned}$$

Manipulando as expressões

$$\begin{aligned}\sum_{i=1}^{m} f(x_i) &= \sum_{i=1}^{m} c_0 + c_1 x_i = c_0 m + c_1 \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i f(x_i) &= \sum_{i=1}^{m} x_i(c_0 + c_1 x_i) = c_0 \sum_{i=1}^{m} x_i + c_1 \sum_{i=1}^{m} x_i^2\end{aligned}$$

e, portanto, conclui-se que

$$\begin{aligned} c_0 m + c_1 \sum_{i=1}^{m} x_i &= \sum_{i=1}^{m} f(x_i) \\ c_0 \sum_{i=1}^{m} x_i + c_1 \sum_{i=1}^{m} x_i^2 &= \sum_{i=1}^{m} x_i \ f(x_i)\end{aligned}$$

$$\begin{aligned} c_0 m + c_1 \sum_{i=1}^{m} x_i &= \sum_{i=1}^{m} f(x_i) \\ c_0 \sum_{i=1}^{m} x_i + c_1 \sum_{i=1}^{m} x_i^2 &= \sum_{i=1}^{m} x_i \ f(x_i)\end{aligned}$$

pode ser escrito na forma:

$$\begin{bmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} f(x_i) \\ \sum_{i=1}^{m} x_i\ f(x_i) \end{bmatrix} \tag{6.4}$$

Temos que, resolvendo o sistema de equações lineares encontramos c_0 e c_1, e assim determinamos $g(x) = c_0 + c_1 x$. Vejamos um exemplo numérico para ilustrar o procedimento

Exemplo 48

Encontre um polinômio linear $g(x) = c_0 + c_1 x$ que melhor se ajusta aos dados da tabela abaixo:

x_i	0	0.25	0.5	0.75	1
$f(x_i)$	1	1.2840	1.6487	2.1170	2.7183

Solução: Precisamos montar o sistema de equações lineares dado por (6.4) para encontrar c_0 e c_1 e assim determinar $g(x)$. Logo, calcula-se os coeficientes do sistema, dados por

$$\sum_{i=1}^{5} x_i = 2.5 \qquad \sum_{i=1}^{5} x_i^2 = 1.875$$
$$\sum_{i=1}^{5} f(x_i) = 8.768 \qquad \sum_{i=1}^{5} x_i f(x_i) = 5.4514$$

Logo temos o seguinte sistema

$$5c_0 + 2.5c_1 = 8.768$$
$$2.5c_0 + 1.875c_1 = 5.4514$$

Resolvendo encontramos

$$c_0 = 0.89968$$
$$c_1 = 1.70784 \quad \Rightarrow \quad g(x) = 0.89968 + 1.70784x$$

A Figura 6.5 mostra a aproximação obtida nesse exemplo pelo método dos mínimos quadrados. Note que...

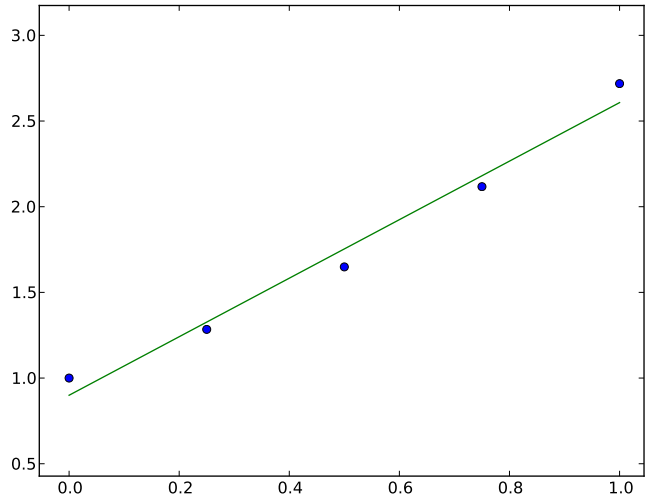


Figura 6.5: Descrever.

De forma geral, queremos aproximar $f(x)$ por

$$g(x) = c_0\phi_0(x) + c_1\phi_1(x) + \ldots + c_n\phi_n(x)$$

onde $\phi_j(x)$ são funções conhecidas. Assim para encontrar os parâmetros $c_0, c_1, \ldots, c_n$ é preciso minimizar a função:

$$\begin{aligned} < r, r > &=< f - c_0\phi_0 - \ldots - cn\phi_n, f - c_0\phi_0 - \ldots - cn\phi_n > \\ &= \sum_{i=1}^{m} \left[f(x_i) - \sum_{j=0}^{n} c_j\phi_j(x_i) \right]^2 \end{aligned}$$

Derivando com relação a cada um dos parâmetros c_k e igualando a zero temos

$$\frac{\partial}{\partial c_k} < r, r >= -2 \sum_{i=1}^{m} \left[f(x_i) - \sum_{j=0}^{n} c_j\phi_j(x_i) \right] \phi_k(x_i) = 0$$

Assim

$$\begin{aligned} & -2 \sum_{i=1}^{m} \left[f(x_i) - \sum_{j=0}^{n} c_j\phi_j(x_i) \right] \phi_k(x_i) = 0 \\ \Rightarrow \quad & \sum_{i=1}^{m} \left\{ \left[f(x_i) - \sum_{j=0}^{n} c_j\phi_j(x_i) \right] \phi_k(x_i) \right\} = 0 \\ \Rightarrow \quad & \sum_{i=1}^{m} \left[\sum_{j=0}^{n} c_j\phi_j(x_i)\phi_k(x_i) \right] = \sum_{i=1}^{m} f(x_i)\phi_k(x_i) \\ \Rightarrow \quad & \sum_{j=0}^{n} c_j \left(\sum_{i=1}^{m} \phi_j(x_i)\phi_k(x_i) \right) = \sum_{i=1}^{m} f(x_i)\phi_k(x_i) \end{aligned}$$

Lembrando que

$$< f, g >= \sum_{i=1}^{m} f(x_i)g(x_i)$$

Usando a notação de produto escalar podemos escrever

$$\Rightarrow \quad \sum_{j=0}^{n} c_j \underbrace{\left(\sum_{i=1}^{m} \phi_j(x_i)\phi_k(x_i) \right)}_{<\phi_j,\phi_k>=<\phi_k,\phi_j>} = \underbrace{\sum_{i=1}^{m} f(x_i)\phi_k(x_i)}_{<f,\phi_k>=<\phi_k,f>}$$

$$\Rightarrow \quad \boxed{\sum_{j=0}^{n} c_j < \phi_k, \phi_j >=< f, \phi_k >}, \quad \text{para } k = 0, 1, \ldots, n$$

que é um sistema de equações lineares $(n+1) \times (n+1)$.

Exemplo para $n = 2$:

$$\begin{aligned}
&< \phi_0, \phi_0 > c_0 + < \phi_0, \phi_1 > c_1 + < \phi_0, \phi_2 > c_2 =< \phi_0, f > \\
&< \phi_1, \phi_0 > c_0 + < \phi_1, \phi_1 > c_1 + < \phi_1, \phi_2 > c_2 =< \phi_1, f > \\
&< \phi_2, \phi_0 > c_0 + < \phi_2, \phi_1 > c_1 + < \phi_2, \phi_2 > c_2 =< \phi_2, f >
\end{aligned}$$

Escrevendo de forma matricial temos

$$\begin{bmatrix} < \phi_0, \phi_0 > & < \phi_0, \phi_1 > & < \phi_0, \phi_2 > \\ < \phi_1, \phi_0 > & < \phi_1, \phi_1 > & < \phi_1, \phi_2 > \\ < \phi_2, \phi_0 > & < \phi_2, \phi_1 > & < \phi_2, \phi_2 > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} < \phi_0, f > \\ < \phi_1, f > \\ < \phi_2, f > \end{bmatrix}$$

De forma geral o sistema é dado por

$$\begin{bmatrix} < \phi_0, \phi_0 > & < \phi_0, \phi_1 > & \ldots & < \phi_0, \phi_n > \\ < \phi_1, \phi_0 > & < \phi_1, \phi_1 > & \ldots & < \phi_1, \phi_n > \\ \vdots & \vdots & \ddots & \vdots \\ < \phi_n, \phi_0 > & < \phi_n, \phi_1 > & \ldots & < \phi_n, \phi_n > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} < \phi_0, f > \\ < \phi_1, f > \\ \vdots \\ < \phi_n, f > \end{bmatrix}$$

o qual é chamado de **sistema normal** ou **equações normais**. Observe que a matriz é simétrica, pois $< f, g >=< g, f >$.

PODE-se mostrar também que a matriz é PD

Exemplo 49

Dada a seguinte tabela

x	-1	0	1	2
y	0	-1	0	7

aproximar $f(x)$ por um polinômio quadrático usando o MMQ.

Solução: Deseja-se encontrar $g(x) = c_0 + c_1x + c_2x^2$ onde

$$\phi_0(x) = 1, \quad \phi_1(x) = x, \quad \phi_2(x) = x^2.$$

Sendo assim, o primeiro passo é montar os vetores que correspondem a avaliação de cada $\phi_j(x)$ nos pontos x_i dados na tabela. Isto é

$$\phi_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \phi_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}, \quad \phi_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 4 \end{bmatrix}, \quad f = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 7 \end{bmatrix}$$

Em seguida, calcula-se os produtos escalares entre estes vetores

$$\begin{aligned} &< \phi_0, \phi_0 > = 4 \\ &< \phi_0, \phi_1 > = 2 \\ &< \phi_0, \phi_2 > = 6 \\ &< \phi_1, \phi_1 > = 6 \\ &< \phi_1, \phi_2 > = 8 \\ &< \phi_2, \phi_2 > = 18 \end{aligned}$$

e também calcula-se o produto escalar de $f(x)$ com ϕ_j, isto é

$$\begin{aligned} &< f, \phi_0 > = 6 \\ &< f, \phi_1 > = 14 \\ &< f, \phi_2 > = 28. \end{aligned}$$

Assim, chega-se no seguinte sistema de equações lineares (equações normais)

$$\begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \\ 28 \end{bmatrix}$$

Resolvendo o sistema encontramos a seguinte solução

$$c_0 = -\tfrac{8}{5}, \quad c_1 = \tfrac{1}{5}, \quad c_2 = 2$$

logo a aproximação pelo MMQ é dada por

$$g(x) = -\tfrac{8}{5} + \tfrac{1}{5}x + 2x^2.$$

A Figura 6.6 apresenta uma comparação da função $f(x)$ com a $g(x)$ que é um polinômio de grau 2 obtido pelo método dos mínimos quadrados.

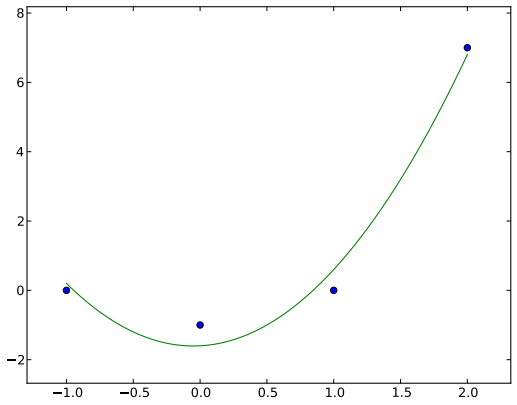


Figura 6.6: Dados do problema e $g(x) = -\frac{8}{5} + \frac{1}{5}x + 2x^2$.

6.2.2 Aproximação para funções não-lineares

Veremos a seguir como lidar com o caso em que $g(x)$ é uma função não-linear dos parâmetros $c_0, c_1, \ldots, c_n$.

Considere agora que estamos interessados em ajustar uma curva $g(x)$ que é uma função não-linear dos parâmetros. Alguns exemplos de funções $g(x)$ que são não-lineares nos parâmetros são dadas abaixo:

$$g(x) = c_0\ e^{c_1 x}$$
$$g(x) = c_0\ x^{c_1}$$
$$g(x) = \frac{1}{c_0 + c_1 x}$$
$$g(x) = c_0\ c_1^x$$

Uma das formas de tratar essa situação é através da **transformação** ou **linearização** do modelo não-linear em um modelo linear. Vamos ilustrar o procedimento através de um exemplo.

Exemplo 50

Seja $f(x) = 20e^{-15x}$. Considere a seguinte tabela de dados

x_i	0.0	0.1	0.2	0.3	0.4
y_i	20.5	4.60	1.00	0.15	0.05

Aproxime $f(x)$ por uma função da seguinte forma $g(x) = c_0\ e^{c_1 x}$.

Solução: É preciso fazer uma transformação e linearizar o modelo, isto é

$$\begin{aligned}\ln f &\approx \ln\left(c_0\ e^{c_1 x}\right)\\ &\approx \ln\left(c_0\right) + \ln\left(e^{c_1 x}\right)\\ &\approx \ln\left(c_0\right) + c_1 x\end{aligned}$$

Assim temos

$$F(x) = a_0\phi_0(x) + a_1\phi_1(x)$$

onde

$$\begin{aligned}\phi_0(x) &= 1\\ \phi_1(x) &= x\end{aligned}$$

com a seguinte mudança

$$\begin{aligned}a_0 &= \ln\left(c_0\right)\\ a_1 &= c_1\\ F &= \ln f(x)\end{aligned}$$

Construimos os vetores

$$\phi_0 = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \phi_1 = \begin{bmatrix}0.0\\0.1\\0.2\\0.3\\0.4\end{bmatrix}, F = \begin{bmatrix}\ln f(x_1)\\ \ln f(x_2)\\ \ln f(x_3)\\ \ln f(x_4)\\ \ln f(x_5)\end{bmatrix} = \begin{bmatrix}3.02\\1.52\\0\\-1.89\\-2.30\end{bmatrix}$$

para montar o sistema precisamos calcular

$$\begin{aligned}&< \phi_0, \phi_0 >, < \phi_0, \phi_1 >, < \phi_1, \phi_1 >\\ &< \phi_0, F >, < \phi_1, F >\end{aligned}$$

assim podemos montar o sistema

$$\begin{bmatrix}5 & 1.0\\ 1.0 & 0.3\end{bmatrix}\begin{bmatrix}a_0\\a_1\end{bmatrix} = \begin{bmatrix}0.35\\-1.335\end{bmatrix}$$

cuja solução é dada por $a_0 = 2.88$ e $a_1 = -14.05$. E assim, como

$$a_0 = \ln c_0 \quad \Rightarrow \quad e^{a_0} = c_0 \quad \Rightarrow \quad c_0 = e^{2.88} \quad \Rightarrow \quad \boxed{c_0 = 17.81}$$

$$a_1 = c_1 \quad \Rightarrow \quad \boxed{c_1 = -14.05}$$

x_i	-8	-6	-4	-2	0	2	4
$f(x_i)$	30	10	9	6	5	4	4

logo a função que melhor aproxima os dados do exemplo é dada por

$$\boxed{g(x) = c_0 e^{c_1 x} = 17.81\ e^{-14.05x}}$$

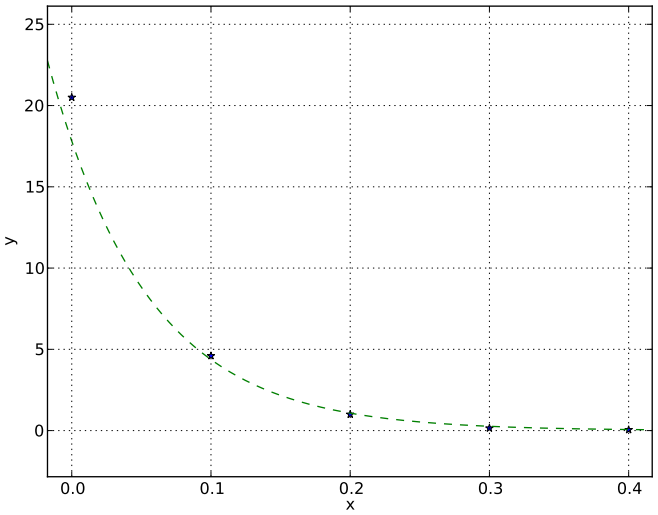


Figura 6.7: Descrever.

6.2.3 Teste de Alinhamento

Em situações que só conhecemos a função através dos dados experimentais tabelados, surge a questão: *qual família de funções melhor aproxima os dados?* Qual função $g(x)$ devemos usar: $g(x) = \frac{1}{c_0 + c_1 x}$ ou $g(x) = c_0\ c_1^x$?

A ideia é aplicar o chamado **teste de alinhamento**:

1. Transformar $y = f(x)$ em

$$F_1(x) = a_0\phi_0(x) + a_1\phi_1(x)$$
$$F_2(x) = b_0\bar{\phi}_0(x) + b_1\bar{\phi}_1(x)$$

2. Plotar x vs $F_1(x)$ e x vs $F_2(x)$

3. Escolher aquela que o gráfico estiver *"mais linear"*

EXEMPLO: Considere a seguinte tabela Qual é a melhor opção ?

a) $g(x) = \frac{1}{c_0 + c_1 x}$

b) $g(x) = c_0\ c_1^x$

Vamos linearizar as funções
a)

$$f(x) \approx g(x) = \frac{1}{c_0 + c_1 x}$$

obtemos

$$\begin{aligned} \frac{1}{f(x)} &\approx c_0 + c_1 x \\ F_1(x) = \frac{1}{f(x)} &= a_0 + a_1 x \end{aligned}$$

onde $a_0 = c_0$, $a_1 = c_1$, $\phi_0 = 1$ e $\phi_1 = x$.
b)

$$f(x) \approx g(x) = c_0\ c_1^x$$

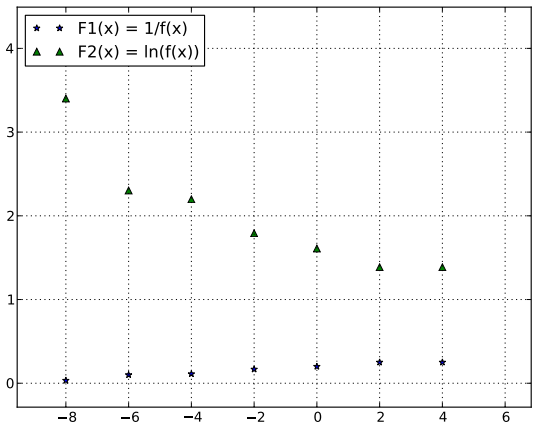
obtemos

$$\begin{aligned} \ln f &\approx \ln c_0 + x \ln c_1 \\ F_2(x) &= a_0 + a_1 x \end{aligned}$$

onde $a_0 = \ln c_0$, $a_1 = \ln c_1$, $\phi_0 = 1$ e $\phi_1 = x$.
Temos que

$$F_1(x) = \frac{1}{f(x)}, \qquad F_2(x) = \ln f(x)$$

Vamos fazer os gráficos de x contra $F_1(x)$ e x contra $F_2(x)$ para o teste de alinhamento.



Logo vemos que $F_1(x)$ é mais adequada para o ajuste de curva.

6.3 Caso Contínuo

Até então falamos em aproximação de dados discretos, isto é, aproximar $f(x)$ sabendo os valores de f em um conj. de pontos x_i. Podemos trabalhar com o caso de aproximar $f(x)$, uma função conhecida para todo x, por uma função mais simples como $g(x)$. No caso contínuo do método dos mínimos quadrados temos que minimizar

$$\int_a^b r(x)^2 \, dx = \int_a^b (f(x) - g(x))^2 \, dx$$

Vamos introduzir formalmente a definição de produto escalar entre funções para que possamos medir o resíduo:

$$< f, g >= \begin{cases} \sum_{i=1}^{m} f(x_i)g(x_i), & \text{caso discreto} \\ \\ \int_a^b f(x)g(x) \, dx, & \text{caso contínuo} \end{cases}$$

Lembrando que o produto escalar entre duas funções contínuas no intervalo $[a, b]$ é definido por

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

podemos então medir a distância de $f(x)$ a $g(x)$ fazendo

$$< f - g, f - g >= \int_a^b [f(x) - g(x)]^2 \, dx$$

no caso discreto tinhamos

$$< f - g, f - g >= \sum_{i=1}^{m} [f(x_i) - g(x_i)]^2$$

Considerando que $r = f - g$, temos

$$\begin{aligned} < r, r > &= \int_a^b [f(x) - g(x)]^2 \, dx = \int_a^b [f - c_0\phi_0 - c_1\phi_1]^2 \, dx \\ &= \int_a^b [f^2 - 2c_0 f\phi_0 - 2c_1 f\phi_1 + c_0^2\phi_0^2 + 2c_0c_1\phi_0\phi_1 + c_1^2\phi_1^2] \, dx \\ &= \int_a^b f^2 \, dx - 2c_0 \int_a^b f\phi_0 \, dx - 2c_1 \int_a^b f\phi_1 \, dx \\ &+ c_0^2 \int_a^b \phi_0^2 \, dx + 2c_0c_1 \int_a^b \phi_0\phi_1 \, dx + c_1^2 \int_a^b \phi_1^2 \, dx \end{aligned}$$

No método dos mínimos quadrados procuramos o ponto de mínimo da função $< r, r >=< r, r > (c_0, c_1)$. Precisamos calcular as derivadas parciais de $< r, r >$ com relação a c_0 e c_1.

Calculando as derivadas parciais e igualando a zero temos

$$\frac{\partial < r, r >}{\partial c_0} = -2\int_a^b f\phi_0 \; dx + 2c_0\int_a^b \phi_0\phi_0 \; dx + 2c_1\int_a^b \phi_0\phi_1 \; dx = 0$$
$$\frac{\partial < r, r >}{\partial c_1} = -2\int_a^b f\phi_1 \; dx + 2c_1\int_a^b \phi_1\phi_1 \; dx + 2c_0\int_a^b \phi_0\phi_1 \; dx = 0$$

organizando os termos, temos

$$c_0\int_a^b \phi_0\phi_0 \; dx + c_1\int_a^b \phi_0\phi_1 \; dx = \int_a^b f\phi_0 \; dx$$
$$c_1\int_a^b \phi_1\phi_1 \; dx + c_0\int_a^b \phi_0\phi_1 \; dx = \int_a^b f\phi_1 \; dx$$

usando a notação $< \phi_i, \phi_j >= \int_a^b \phi_i(x)\phi_j(x) \; dx$, podemos escrever

$$c_0 < \phi_0, \phi_0 > + \; c_1 < \phi_0, \phi_1 > = < f, \phi_0 >$$
$$c_1 < \phi_1, \phi_0 > + \; c_0 < \phi_1, \phi_1 > = < f, \phi_1 >$$

que é um sistema de equações lineares (*equações normais*), o qual pode ser escrito de forma matricial como

$$\begin{bmatrix} < \phi_0, \phi_0 > & < \phi_0, \phi_1 > \\ < \phi_1, \phi_0 > & < \phi_1, \phi_1 > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} < f, \phi_0 > \\ < f, \phi_1 > \end{bmatrix}$$

Obs: o procedimento é similar ao caso discreto, a diferença é na forma de se calcular os produtos escalares, pois no caso contínuo temos que calcular as integrais

$$< \phi_0(x), \phi_1(x) >= \int_a^b \phi_0(x)\phi_1(x) \; dx$$
$$< f(x), \phi_0(x) >= \int_a^b f(x)\phi_0(x) \; dx, \; \ldots$$

EXEMPLO: Seja $f(x) = x^4 - 5x$, $x \in [-1, 1]$. Vamos aproximar $f(x)$ por um polinômio de segundo grau usando o MMQ.

Solução: $f(x)$ é contínua no intervalo $[-1, 1]$. Queremos

$$f(x) \approx g(x) = c_0 + c_1 x + c_2 x^2$$

onde usamos as funções $\phi_0(x) = 1$, $\phi_1(x) = x$, $\phi_2(x) = x^2$. Precisamos resolver o sistema de equações normais

$$\begin{bmatrix} < \phi_0, \phi_0 > & < \phi_0, \phi_1 > & < \phi_0, \phi_2 > \\ < \phi_1, \phi_0 > & < \phi_1, \phi_1 > & < \phi_1, \phi_2 > \\ < \phi_2, \phi_0 > & < \phi_2, \phi_1 > & < \phi_2, \phi_2 > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} < f, \phi_0 > \\ < f, \phi_1 > \\ < f, \phi_2 > \end{bmatrix}$$

Substituindo $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2$, temos

$$\begin{bmatrix} < 1, 1 > & < 1, x > & < 1, x^2 > \\ < x, 1 > & < x, x > & < x, x^2 > \\ < x^2, 1 > & < x^2, x > & < x^2, x^2 > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} < f, \phi_0 > \\ < f, \phi_1 > \\ < f, x^2 > \end{bmatrix}$$

usando o produto escalar usual em $[-1,1]$ definido por

$$< f,g >= \int_a^b f(x)g(x)\ dx \quad \Rightarrow \quad \int_{-1}^1 f(x)g(x)\ dx$$

temos

$$\begin{aligned}
< 1,1 > &= \int_{-1}^1 dx = x]_{-1}^1 = 2 \\
< 1,x > &= \int_{-1}^1 x\ dx = \tfrac{x^2}{2}]_{-1}^1 = 0 =< x,1 > \\
< 1,x^2 > &= \int_{-1}^1 x^2\ dx = \tfrac{x^3}{3}]_{-1}^1 = \tfrac{2}{3} < x^2,1 >
\end{aligned}$$

$$\begin{aligned}
< x,x > &= \int_{-1}^1 x^2\ dx = \tfrac{x^3}{3}]_{-1}^1 = \tfrac{2}{3} \\
< x,x^2 > &= \int_{-1}^1 x^3\ dx = \tfrac{x^4}{4}]_{-1}^1 = 0 =< x^2,x > \\
< x^2,x^2 > &= \int_{-1}^1 x^4\ dx = \tfrac{x^5}{5}]_{-1}^1 = \tfrac{2}{5}
\end{aligned}$$

e ainda

$$\begin{aligned}
< f,1 > &= \int_{-1}^1 x^4 - 5x\ dx = \tfrac{x^5}{2} - \tfrac{5x^2}{5}\Big]_{-1}^1 = \left(\tfrac{1}{5} - \tfrac{5}{2}\right) - \left(-\tfrac{1}{5} - \tfrac{5}{2}\right) = \tfrac{2}{5} \\
< f,x > &= \int_{-1}^1 x^5 - 5x^2\ dx = \tfrac{x^6}{6} - \tfrac{5x^3}{3}\Big]_{-1}^1 = -\tfrac{10}{3} \\
< f,x^2 > &= \int_{-1}^1 x^6 - 5x^3\ dx = \tfrac{x^7}{7} - \tfrac{5x^4}{4}\Big]_{-1}^1 = \tfrac{2}{7}
\end{aligned}$$

Assim o sistema de equações normais é dado por

$$\begin{bmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \frac{2}{5} \\ -\frac{10}{3} \\ \frac{2}{7} \end{bmatrix}$$

resolvendo com o método de Cholesky, por exemplo, obtemos a seguinte solução

$$c_0 = -\tfrac{3}{35}, \quad c_1 = -5, \quad c_2 = \tfrac{6}{7}$$

e assim

$$\boxed{g(x) = -\tfrac{3}{35} - 5x + \tfrac{6}{7}x^2}$$

Figura

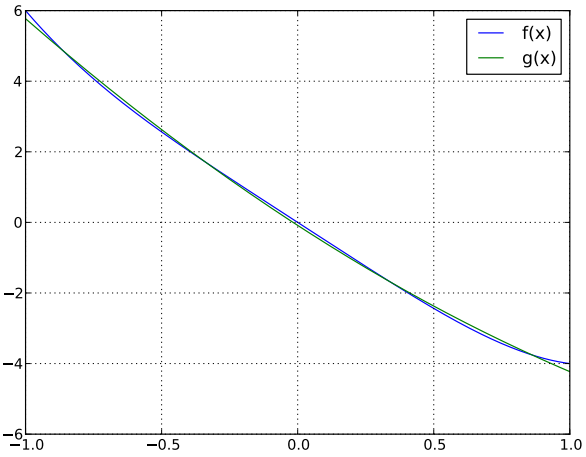


Figura 6.8: Exemplo MMQ caso contínuo.

6.4 Polinômios Ortogonais

Quando o produto escalar, no caso discreto ou no contínuo, é igual a zero, dizemos que os vetores (caso discreto) ou funções (caso contínuo) são ortogonais. Discreto

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{2} x_i y_i = x_1 y_1 + x_2 y_2 = 1(0) + 0(1) = 0$$

Contínuo

$$f(x) = \sin(x), \quad g(x) = \cos(x)$$

$$< f, g >= \int_{-\pi}^{\pi} \sin(x)\cos(x)\ dx = 0$$

Podemos também usar o produto escalar com peso, dado por uma função $w(x)$. Nesse caso o produto escalar é dado por:

- caso contínuo

$$< f, g >= \int_{a}^{b} w(x) f(x) g(x)\ dx$$

- caso discreto

$$< f, g >= \sum_{i=1}^{m} w(x_i) f(x_i) g(x_i)$$

e temos também a mesma noção de vetores ou funções ortogonais.

Note que até agora trabalhamos com o seguinte conjunto de funções

$$\phi_0(x) = 1, \quad \phi_1(x) = x, \quad \ldots, \quad \phi_n(x) = x^n$$

o qual constitui uma base para o espaço vetorial dos polinômios de grau menor ou igual a n.

Observe que se trabalharmos com um conjunto de funções ϕ_j que são ortogonais entre si, isto é, que satisfazem

$$< \phi_i, \phi_j >= 0, \quad \forall i \neq j \tag{6.5}$$

então o sistema normal tem a seguinte forma (diagonal)

$$\begin{bmatrix} < \phi_0, \phi_0 > & 0 & \ldots & 0 \\ 0 & < \phi_1, \phi_1 > & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & < \phi_n, \phi_n > \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} < f, \phi_0 > \\ < f, \phi_1 > \\ \vdots \\ < f, \phi_n > \end{bmatrix}$$

Neste caso a solução é facilmente calculada tomando

$$\begin{aligned} < \phi_0, \phi_0 > c_0 =< f, \phi_0 > \quad &\Rightarrow \quad c_0 = \frac{< f, \phi_0 >}{< \phi_0, \phi_0 >} \\ < \phi_1, \phi_1 > c_1 =< f, \phi_1 > \quad &\Rightarrow \quad c_1 = \frac{< f, \phi_1 >}{< \phi_1, \phi_1 >} \\ &\vdots \\ < \phi_n, \phi_n > c_n =< f, \phi_n > \quad &\Rightarrow \quad c_n = \frac{< f, \phi_n >}{< \phi_n, \phi_n >} \end{aligned}$$

isto é, para $i = 0, 1, \ldots, n$

$$\boxed{c_i = \frac{< f, \phi_i >}{< \phi_i, \phi_i >}} \tag{6.6}$$

Existem várias *famílias* de polinômios ortogonais, e estes podem ser construídos com *diferentes produtos escalares*, dependendo do intervalo, da função peso, ou de ser contínua ou discreta a aplicação.

A seguir veremos 2 exemplos de famílias de polinômios ortogonais:

- Polinômios de Legendre
- Polinômios de Chebyshev

6.4.1 Polinômios Ortogonais de Legendre

Os polinômios de Legendre são ortogonais no intervalo $[-1, 1]$ com a função peso $w(x) = 1$. Os primeiros polinômios são dados por

$$\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x \\
P_2(x) &= \frac{1}{2}(3x^2 - 1) \\
P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\
P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3), \ldots
\end{aligned}$$

De forma geral eles podem ser escritos como

$$P_k(x) = \frac{1}{(2^k)k!}\frac{d^k}{dx^k}[(x^2 - 1)]^k, \quad k = 1, \ldots$$

Vamos verificar para alguns casos, que de fato os polinômios de Legendre são ortogonais.

$$\begin{aligned}
< P_1, P_0 > &= \int_{-1}^{1} x \; dx = \left[\frac{x^2}{2}\right]_{-1}^{1} = 0 \\
< P_1, P_1 > &= \int_{-1}^{1} x^2 \; dx = \left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{2}{3} \\
< P_1, P_2 > &= \int_{-1}^{1} x\frac{1}{2}(3x^2 - 1) \; dx = \frac{1}{2}\int_{-1}^{1} 3x^3 - x \; dx \\
&= \frac{1}{2}\left[\frac{3}{4}x^4 - \frac{x^2}{2}\right]_{-1}^{1} = \frac{1}{2}\left[\left(\frac{3}{4} - \frac{1}{2}\right) - \left(\frac{3}{4} - \frac{1}{2}\right)\right] = 0 \\
< P_1, P_3 > &= \int_{-1}^{1} x\frac{1}{2}\left(5x^3 - 3x\right) \; dx = \ldots = 0
\end{aligned}$$

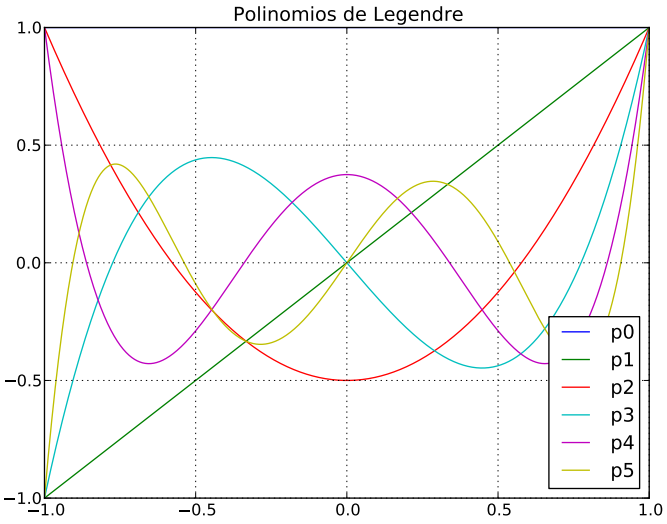


Figura 6.9: Primeiros polinômios ortogonais de Legendre.

Exemplo 51

Use o MMQ com os polinômios de Legendre para encontrar a função $g(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x)$ considerando que $\phi_0(x) = 1$, $\phi_1(x) = x$ e $\phi_2(x) = \frac{1}{2}(3x^2 - 1)$, que melhor se ajusta à $f(x) = e^x$ no intervalo $[-1, 1]$.

Solução: Como estamos trabalhando com polinômios ortogonais, podemos encontrar os parâmetros c_0, c_1 e c_2 usando

$$c_i = \frac{< f, \phi_i >}{< \phi_i, \phi_i >}$$

Para isso temos que calcular

- $< f, \phi_0 >, < f, \phi_1 >, < f, \phi_2 >$
- $< \phi_0, \phi_0 >, < \phi_1, \phi_1 >, < \phi_2, \phi_2 >$

Calculando

$$< \phi_0, \phi_0 > = \int_{-1}^{1} dx = x]_{-1}^{1} = 1 - (-1) = 2$$

$$< \phi_1, \phi_1 > = \int_{-1}^{1} x^2 \, dx = \frac{x^3}{3}\Big]_{-1}^{1} = \frac{1}{3} - \left(-\frac{1}{3}\right) = \frac{2}{3}$$

$$\begin{aligned} < \phi_2, \phi_2 > &= \int_{-1}^{1} \frac{1}{2}(3x^2 - 1)\frac{1}{2}(3x^2 - 1) \, dx \\ &= \frac{1}{4}\int_{-1}^{1} (9x^4 - 6x^2 + 1) \, dx = \frac{1}{4}\left[\frac{9x^5}{5} - \frac{6x^3}{3} + x\right]_{-1}^{1} \\ &= \frac{1}{4}\left[\left(\frac{9}{5} - 2 + 1\right) - \left(-\frac{9}{5} + 2 - 1\right)\right] \\ &= \ldots = \frac{2}{5} \end{aligned}$$

$$< f, \phi_0 > = \int_{-1}^{1} e^x \, dx = e^x]_{-1}^{1} = e^1 - e^{-1} = 2.35$$

$$\begin{aligned} < f, \phi_1 > &= \int_{-1}^{1} x \, e^x \, dx \quad \Rightarrow \quad \text{IPP : } \boxed{\int udv = uv| - \int v \, du} \\ &= xe^x\Big|_{-1}^{1} - \int_{-1}^{1} e^x \, dx = xe^x\Big|_{-1}^{1} - e^x\Big|_{-1}^{1} \\ &= e^x(x-1)\Big|_{-1}^{1} = \left[e^1(1-1) - e^{-1}(-2)\right] \\ &= 2e^{-1} = 0.736 \end{aligned}$$

$$\begin{aligned} < f, \phi_2 > &= \int_{-1}^{1} e^x \tfrac{1}{2}(3x^2 - 1) \, dx \quad \text{IPP: } \boxed{u = (3x^2 - 1), \; dv = e^x} \\ &= \frac{1}{2}\left[(3x^2 - 1)e^x\Big|_{-1}^{1} - 6\int_{-1}^{1} e^x x \, dx\right] \\ &= \frac{1}{2}\left[(2e^1 - 2e^{-1}) - 6\ (0.736)\right] = \frac{1}{2}\left[2\ (2.35) - 4.414\right] \\ &= 0.143 \end{aligned}$$

Assim temos que os coeficientes são

$$c_0 = \frac{< f, \phi_0 >}{< \phi_0, \phi_0 >} = \frac{2.35}{2} = 1.175$$
$$c_1 = \frac{< f, \phi_1 >}{< \phi_1, \phi_1 >} = \frac{0.736}{2/3} = 1.104$$
$$c_2 = \frac{< f, \phi_2 >}{< \phi_2, \phi_2 >} = \frac{0.143}{2/5} = 0.358$$

e portanto

$$\begin{aligned} g(x) &= c_0\phi_0 + c_1\phi_1 + c_2\phi_2 \\ &= c_0 + c_1 x + c_2[0.5\ (3x^2 - 1)] \\ &= 1.175 + 1.104\ x + 0.358\ [0.5(3x^2 - 1)] \end{aligned}$$

é a função que melhor aproxima $f(x) = e^x$ em $[-1, 1]$ segundo o critério dos mínimos quadrados.

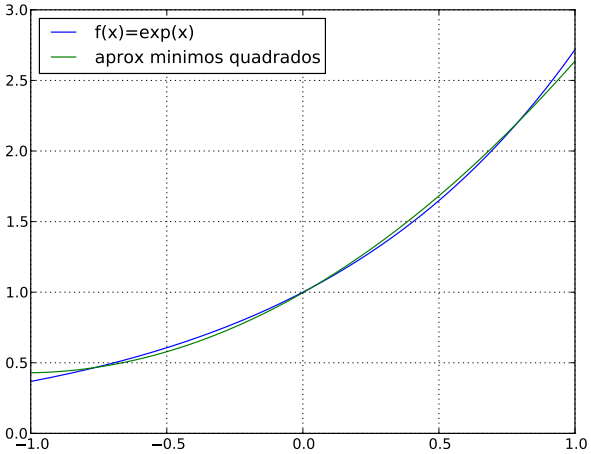


Figura 6.10: MMQ com polinômios ortogonais de Legendre.

6.4.2 Polinômios Ortogonais de Chebyshev

Os polinômios de Chebyshev são ortogonais no intervalo $[-1, 1]$ com a seguinte função peso: $w(x) = \frac{1}{\sqrt{1-x^2}}$. Os primeiros polinômios são dados por

$$P_0(x) = 1 \tag{6.7}$$
$$P_1(x) = x \tag{6.8}$$
$$P_2(x) = 2x^2 - 1 \tag{6.9}$$
$$P_3(x) = 4x^3 - 3x \tag{6.10}$$
$$P_4(x) = 8x^4 - 8x^2 + 1 \tag{6.11}$$

As Figuras 6.11 e 6.12 mostram a função peso $w(x)$ usada no produto escalar e os primeiros polinômios ortogonais de Chebyshev.

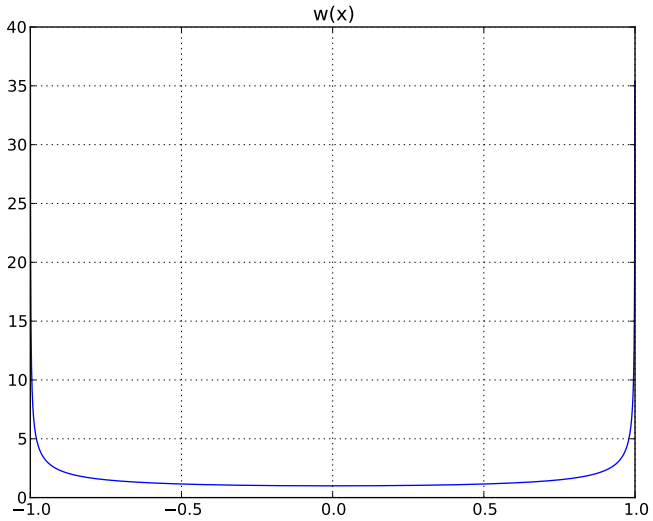


Figura 6.11: Polinômios de Chebyshev: função peso $w(x)$.

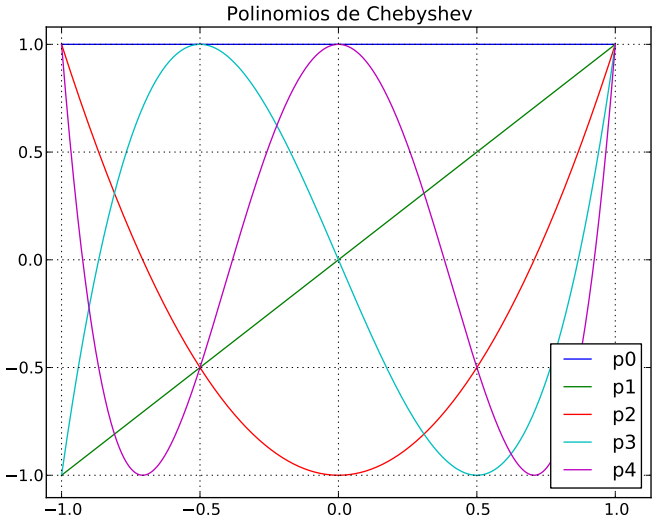


Figura 6.12: Polinômios de Chebyshev.

Capítulo 7

Integração Numérica

Neste capítulo estamos interessados em estudar métodos numéricos para calcular de forma aproximada a integral de uma função com uma variável real em um intervalo $[a, b]$. O problema consiste em: *encontrar* $I \in \mathbb{R}$ *tal que*

$$I = I(f) = \int_a^b f(x)\ dx$$

onde $f(x)$ *é uma função contínua com derivadas contínuas no intervalo* $[a, b]$.

Seja $F(x)$ a função primitiva de $f(x)$, tal que $F'(x) = f(x)$. Pelo Teorema Fundamental do Cálculo sabemos que o valor da integral é dado por

$$I = \int_a^b f(x)\ dx = F(b) - F(a).$$

Para ilustrar, considere como exemplo o problema de calcular $\int_0^2 x^4\ dx$. Como $F(x) = \frac{x^5}{5}$ satisfaz $F'(x) = x^4 = f(x)$, pelo TFC, temos

$$I = \int_0^2 x^4\ dx = \tfrac{2^5}{5} - \tfrac{0^5}{5} = \tfrac{32}{5} = 6.4$$

Porém, é preciso ressaltar as seguintes situações:

- nem sempre conseguimos determinar a primitiva $F(x)$; considere o exemplo $\int_a^b e^{x^2}\ dx$;
- em algumas situações a manipulação de $F(x)$ pode ser complexa e trabalhosa
- em outros casos, podemos não conhecer de forma analítica a função $f(x)$ que se deseja integrar e só temos os valores de $f(x)$ em pontos x_i do intervalo (situação comum em experimentos).

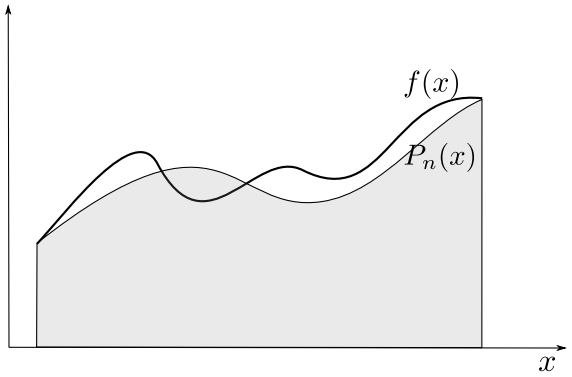
Em geral, nessas situações faz-se necessário o uso de métodos numéricos para calcular de forma aproximada o valor da integral. Neste capítulo serão apresentados três metodologias para o cálculo de integrais: fórmulas de Newton-Cotes, método dos coeficientes indeterminados e a quadratura de Gauss.

7.1 Fórmulas de Newton-Cotes

De forma geral integração numérica pelas fórmulas de Newton-Cotes consiste em integrar o polinômio interpolador $P_n(x)$ da função $f(x)$ definido em um conjunto de pontos $x_0, x_1, \ldots, x_n$ do intervalo $[a, b]$. Isto é

$$I = \int_a^b f(x)\ dx \approx \int_{x_0}^{x_n} P_n(x)\ dx$$

o que pode ser representado graficamente como apresentado pela Figura **??**.



Sendo assim, as fórmulas (ou regras) de Newton-Cotes seguem o seguinte esquema: obter o polinômio interpolador $P_n(x)$ em pontos equidistantes; aproximar o valor da integral de $f(x)$ usando $P_n(x)$. Aqui, vamos considerar apenas as fórmulas de Newton-Cotes do tipo **fechada**, isto é, quando $x_0 = a$ e $x_n = b$.

Como vimos iremos usar o polinômio interpolador $P_n(x)$ de grau n para aproximar $f(x)$. Como iremos usar pontos igualmente espaçados para desenvolver as fórmulas de Newton-Cotes, iremos usar $P_n(x)$ na forma de Newton-Gregory

$$\begin{aligned} P_n(x) = f(x_0) + (x - x_0)\frac{\Delta f(x_0)}{1!h} + (x - x_0)(x - x_1)\frac{\Delta^2 f(x_0)}{2!h} \\ + \ldots + (x - x_0)\ldots(x - x_{n-1})\frac{\Delta^n f(x_0)}{n!h} \end{aligned} \tag{7.1}$$

onde $\Delta^i f(x_0)$ é o operador de diferenças ordinárias de ordem i. Como vimos, podemos fazer uma mudança de variável na Equação 7.1 da seguinte forma

$$u = \frac{x - x_0}{h} \quad \text{ou} \quad x = x_0 + uh \tag{7.2}$$

Nesse caso, os pontos de interpolação são sempre dados por $0, 1, 2, \ldots, n$, ao invés de $x_0, x_1, \ldots, x_n$.

$$\begin{aligned} x_0 \quad &\Rightarrow \quad u = \frac{x_0 - x_0}{h} = 0 \\ x_1 \quad &\Rightarrow \quad u = \frac{x_1 - x_0}{h} = 1 \\ x_2 \quad &\Rightarrow \quad u = \frac{x_2 - x_0}{h} = \frac{2h}{h} = 2, \ldots \end{aligned}$$

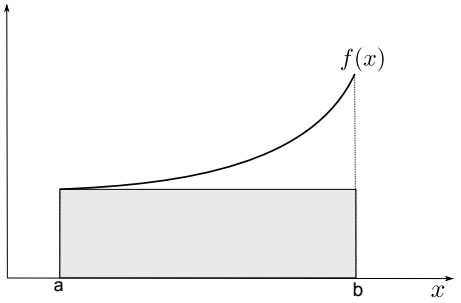
Assim podemos escrever

$$P_n(x) = f(x_0) + u\Delta f(x_0) + u(u-1)\frac{\Delta^2 f(x_0)}{2!} + \ldots + u(u-1)\ldots(u-n+1)\frac{\Delta^n f(x_0)}{n!} \tag{7.3}$$

7.1.1 Regra do Retângulo

O polinômio mais simples é uma constante. Na regra do retângulo, $f(x)$ é aproximada pelo seu valor em $x_0 = a$ (ou em $x_1 = b$), de tal forma que

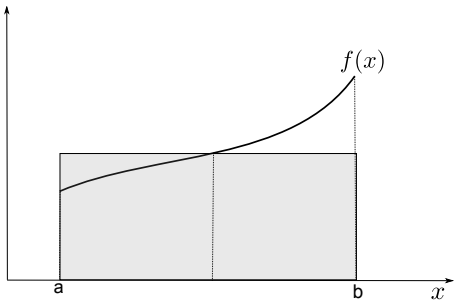
$$\int_a^b f(x)\ dx \approx \int_a^b P_0(x)\ dx = \int_a^b f(a)dx = xf(a)\Big|_a^b = (b-a)f(a) = hf(a) = I_R$$



7.1.2 Regra do Ponto Médio

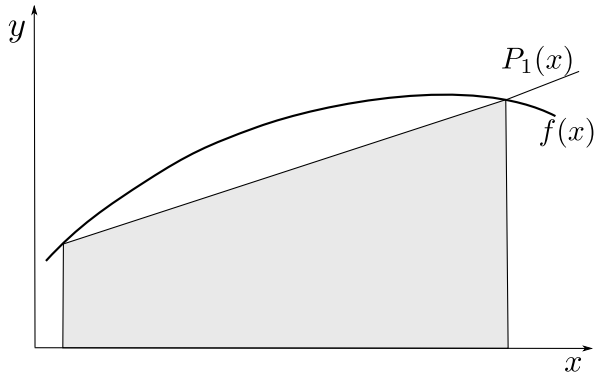
Também podemos aproximar $f(x)$ por uma outra constante tomada ao avaliar $f(x)$ em algum outro ponto do intervalor $[a, b]$; a escolha mais comum é $(a + b)/2$, o centro do intervalo. Assim temos

$$\int_a^b f(x)\ dx \approx (b-a)f\left(\frac{a+b}{2}\right) = \boxed{hf\left(\frac{a+b}{2}\right) = I_M}$$



7.1.3 Regra do Trapézio

Seja $P_1(x)$ o polinômio interpolador de $f(x)$ que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$ com $x_0 = a$ e $x_1 = b$. Para calcular a integral vamos substituir o polinômio linear $P_1(x)$ e obter o valor aproximado da integral.



Substituindo $P_1(x)$ temos

$$\int_a^b f(x)\ dx \approx \int_{a=x_0}^{b=x_1} P_1(x)\ dx = \int_{x_0}^{x_1} \left[f(x_0) + (x - x_0)\frac{\Delta f(x_0)}{h} \right]\ dx$$

Fazendo a mudança de variável de x para u, temos

$$\begin{aligned} u = \frac{x - x_0}{h} \quad &\Rightarrow \quad \frac{du}{dx} = \frac{1}{h} \quad \Rightarrow \quad h du = dx \\ x = x_0 = a \quad &\Rightarrow \quad u = \frac{x_0 - x_0}{h} = 0 \\ x = x_1 = b \quad &\Rightarrow \quad u = \frac{x_1 - x_0}{h} = 1 \end{aligned}$$

então

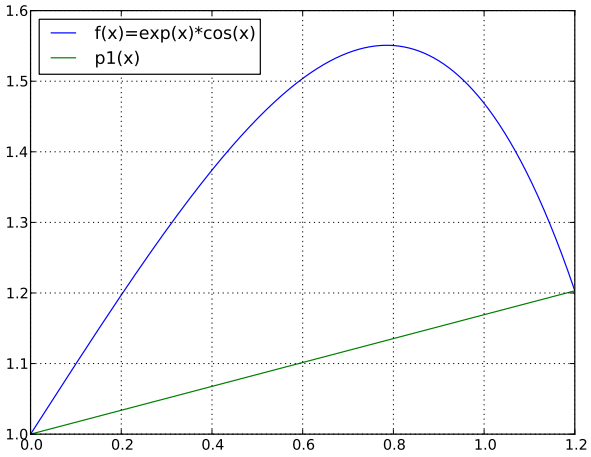
$$\int_a^b f(x)\ dx \approx= \int_0^1 \left[f(x_0) + u\frac{\Delta f(x_0)}{h} \right] h\ du$$

Integrando em u, de forma analitica, temos

$$\begin{aligned} \int_a^b f(x)\ dx \approx = \int_0^1 \left[f(x_0) + u\frac{\Delta f(x_0)}{h} \right] h\ du \\ &= h\left[uf(x_0) + \tfrac{u^2}{2}\Delta f(x_0) \right]\Big|_0^1 \\ &= h\left[f(x_0) + \tfrac{1}{2}(f(x_1) - f(x_0)) \right] \end{aligned}$$

de onde obtemos a regra do Trapézio

$$\boxed{I_T = \frac{h}{2}\left[f(x_0) + f(x_1) \right]} \tag{7.4}$$



Exemplo Calcule de forma aproximada o valor da seguinte integral $\int_0^{1.2} e^x \cos(x)\ dx$ usando a **regra do trapézio**. **Solução do Exemplo** Temos $a = x_0 = 0$ e $b = x_1 = 1.2$, logo $h = x_1 - x_0 = 1.2$. Calculando os valores da função em x_0 e x_1 temos

$$f(0) = e^0 \cos(0) = 1 \qquad\qquad f(1.2) = e^{1.2} \cos(1.2) = 1.20$$

logo

$$I = \tfrac{1.2}{2}[f(0) + f(1.2)] = 0.6[1 + 1.20] = 1.32$$

O valor exato da integral é

$$\begin{aligned}\int_0^{1.2} e^x \cos x\ dx &= \left.\tfrac{e^x(\sin x + \cos x)}{2}\right|_0^{1.2} \\ &= \tfrac{e^{1.2}(\sin 1.2 + \cos 1.2)}{2} - \frac{1}{2} \\ &= 1.648774427\end{aligned}$$

□**Obs**: os exercícios são para calcular as integrais de forma aproximada usando as fórmulas de integração numérica!

7.1.4 Regra 1/3 de Simpson

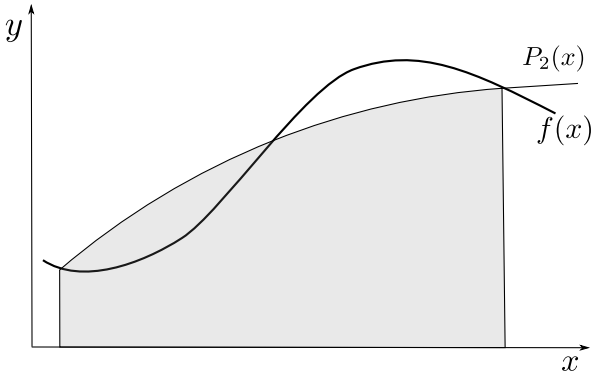
Vamos aproximar $f(x)$ por um polinômio interpolador $P_2(x)$ de grau 2. Assim temos

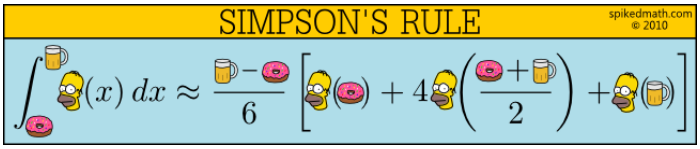
$$I = \int_a^b f(x)\ dx \approx \int_{a=x_0}^{b=x_2} P_2(x)\ dx$$

Graficamente

Novamente fazendo a mudança de variável de x para u, temos $hdu = dx$ e

$$\begin{aligned} x = x_0 = a \quad &\Rightarrow \quad u = \frac{x_0 - x_0}{h} = 0 \\ x = x_2 = b \quad &\Rightarrow \quad u = \frac{x_2 - x_0}{h} = \tfrac{2h}{h} = 2 \end{aligned}$$





então

$$\begin{aligned}\int_{x_0}^{x_2} P_2(x)\,dx &= \int_0^2 \left[f(x_0) + u\Delta f(x_0) + u(u-1)\frac{\Delta^2 f(x_0)}{2}\right] h\,du \\ &= h\left[uf(x_0) + \frac{u^2}{2}\Delta f(x_0) + \left(\tfrac{u^3}{6} - \tfrac{u^2}{4}\right)\Delta^2 f(x_0)\right]\Bigg|_0^2 \\ &= h\left[2f(x_0) + 2(f(x_1) - f(x_0)) + \tfrac{1}{3}(f(x_2) - 2f(x_1) + f(x_0)\right]\end{aligned}$$

de onde obtemos a regra de 1/3 de Simpson

$$\boxed{I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right]} \tag{7.5}$$

Regra 1/3 de Simpson

$$\boxed{I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right]}$$

Para lembrar

Exemplo Vamos calcular o valor da integral $\int_0^{1.2} e^x \cos x\,dx$. Temos que

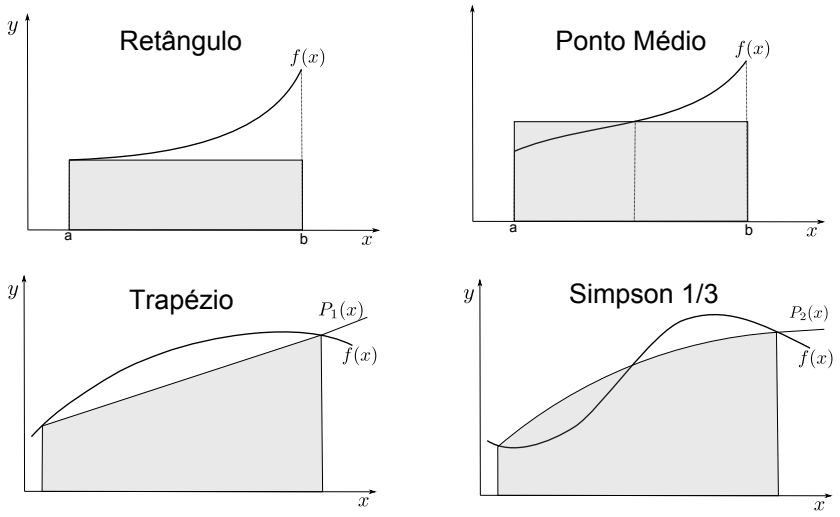
$$h = \frac{x_2 - x_0}{2}$$

Pela fórmula é preciso calcular o valore de $f(x)$ em x_0, x_1 e x_2.

$$\begin{aligned}f(x_)) = f(0) &= e^0 \cos(0) = 1 \\ f(x_1) = f(0.6) &= e^{0.6}\cos(0.6) = 1.50 \\ f(x_2) = f(1.2) &= e^{1.2}\cos(1.2) = 1.20\end{aligned}$$

assim

$$I = \frac{0.6}{3}\left[1 + 4(1.50) + 1.2\right] = 0.2(8.2) = 1.64$$



7.1.5 Regra 3/8 de Simpson

Vamos usar agora um polinômio interpolador $P_3(x)$ de grau 3 para $f(x)$. Assim

$$\int_a^b f(x)\ dx \approx \int_{a=x_0}^{b=x_3} P_3(x)\ dx$$

Novamente usando o polinômio interpolador na forma de Newton-Gregory temos

$$\int_{x_0}^{x_3} P_3\ dx = \int_0^3 \left[f(x_0) + u\Delta f(x_0) + u(u-1)\frac{\Delta^2 f(x_0)}{2} + u(u-1)(u-2)\frac{\Delta^3 f(x_0)}{6}\right] h\ du$$

de onde obtemos após integrar de forma analítica em u, a regra 3/8 de Simpson dada por

$$\boxed{I_s^{3/8} = \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right]} \tag{7.6}$$

Exemplo Podemos calcular novamente $\int_0^{1.2} e^x \cos(x)\ dx$, agora pela regra 3/8 de Simpson. Para tal, sabemos que $h = \frac{1.2-0}{3} = 0.4$ e assim calculamos

$$f(0) = 1, \qquad f(0) = 1.37$$
$$f(0.8) = 1.55, \qquad f(1.2) = 1.2$$

logo

$$I = \frac{3(0.4)}{8}\left[1 + 3(1.37) + 3(1.55) + 1.2\right] = 1.6465$$

□

7.1.6 Resumo

7.2 Análise do erro

Vamos considerar agora o erro cometido ao usar as regras de quadratura apresentadas até agora. Em todos os casos aproximamos $f(x)$ por um polinômio interpolador $P_n(x)$ de grau n no intervalo $[a, b]$, e então calculamos a integral de P_n como aproximação para a integral.

Logo o erro cometido é dado por

$$E = \int_a^b [f(x) - P_n(x)]\ dx$$

Como vimos no estudo de interpolação, o erro é dado por

$$f(x) - P_n(x) = (x - x_0)(x - x_1)\dots(x - x_n)\frac{f^{(n+1)}\ (\eta(x))}{(n+1)!}$$

onde $\eta(x)$ é um ponto entre $[a, b]$ e $x_0, \dots, x_n$ são os pontos de interpolação.

Assim de forma geral temos que

$$\boxed{E = \tfrac{1}{(n+1)!}\int_a^b (x - x_0)(x - x_1)\dots(x - x_n)f^{(n+1)}(\eta)\ dx} \tag{7.7}$$

Antes de continuar vamos enunciar um resultado do qual faremos uso na dedução das fórmulas dos erros cometidos na integração numérica.

Teorema 13 (Teorema Valor Médio para Integrais)**.** Sejam $h(x)$ e $g(x)$ funções contínuas em $[a, b]$ tal que $h(x)$ não muda de sinal, então existe $\xi \in [a, b]$ tal que

$$\int_a^b h(x)g(x)\ dx = g(\xi)\int_a^b h(x)\ dx$$

Vamos aplicar a Equação 7.7 para alguns casos particulares.

7.2.1 Erro na Regra do Retângulo

Nesse caso $n = 0$ e $x_0 = a$, portanto

$$E_R = \int_a^b (x - a)f'(\eta(x))\ dx$$

Aplicando o teorema do valor médio para integrais temos

$$\begin{aligned} E_R &= \int_a^b (x - a)f'(\eta(x))\ dx = f'(\xi)\int_a^b x - a\ dx \\ &= f'(\xi)\left[\tfrac{x^2}{2} - ax\right]\Big|_a^b = f'(\xi)\left[\tfrac{b^2}{2} - ab - \tfrac{a^2}{2} + a^2\right] \\ &= \frac{f'(\xi)}{2}\left[b^2 - 2ab + a^2\right] \end{aligned}$$

isto é

$$\boxed{E_R = \frac{f'(\xi)}{2}(b - a)^2}$$

$$\boxed{E_R = \frac{f'(\xi)}{2}(b-a)^2}$$

Devido a dificuldade de determinar o ponto ξ, em geral trabalhamos com um limitante superior para o erro, o qual é dado por

$$\boxed{|E_R| \leq \frac{M_1}{2}(b-a)^2}$$

onde M_1 é um limitante para $|f'(x)|$ em $[a,b]$, isto é

$$M_1 = \max_{a \leq x \leq b} |f'(x)|$$

7.2.2 Erro na Regra do Trapézio

Para a regra do trapézio temos $n = 1$ e $x_0 = a$ e $x_1 = b$, assim temos

$$E_T = \frac{1}{2}\int_a^b (x-a)(x-b)f''(\eta(x))\ dx$$

Como $(x-a)(x-b)$ não muda de sinal, usamos o teorema do valor médio para integrais e obtemos

$$E_T = \frac{f''(\xi)}{2}\int_a^b (x-a)(x-b)\ dx$$

que após integração resulta em

$$\boxed{E_T = -\frac{f''(\xi)}{12}(b-a)^3} \quad \Rightarrow \quad \boxed{|E_T| \leq \frac{M_2}{12}(b-a)^3}$$

onde M_2 é um limitante para a segunda derivada de $|f''(x)|$ no intervalo $[a,b]$.

7.2.3 Erro na Regra do Ponto Médio

Podemos proceder como nos casos anteriores para obter uma estimativa para o erro na **regra do ponto médio**, entretanto podemos obter uma estimativa do erro melhor do que desta forma.

Seja $m = (a+b)/2$ e vamos tomar a expansão em série de Taylor de $f(x)$ em torno do ponto m, isto é

$$f(x) = f(m) + f'(m)(x-m) + \tfrac{f''(\eta(x))}{2}(x-m)^2$$

E ainda nesse caso $n = 0$ e $P_0(x) = f(m)$, assim

$$\underbrace{f(x) - P_0(m)}_{\text{erro}} = f'(m)(x-m) + \tfrac{f''(\eta(x))}{2}(x-m)^2$$

integrando temos

$$E_M = \int_a^b f'(m)(x-m) + \tfrac{f''(\eta(x))}{2}(x-m)^2 \, dx$$

Continuando

$$\begin{aligned} E_M &= \int_a^b f'(m)(x-m) + \tfrac{f''(\eta(x))}{2}(x-m)^2 \, dx \\ &= \int_a^b f'(m)(x-m) \, dx + \frac{1}{2}\int_a^b f''(\eta(x))(x-m)^2 \, dx \\ &= f'(m) \underbrace{\int_a^b (x-m) \, dx}_{=0} + \frac{1}{2} f''(\xi) \int_a^b (x-m)^2 \, dx \end{aligned}$$

assim

$$\boxed{E_M = \frac{f''(\xi)}{24}(b-a)^3}$$

7.2.4 Erro na Regra do Ponto Médio e Simpson 1/3

Com a fórmula do erro anterior podemos obter o seguinte limitante

$$\boxed{|E_M| \leq \frac{M_2}{24}(b-a)^3}$$

onde M_2 é um limitante para $|f''(x)|$ em $[a,b]$.

O erro para a regra 1/3 de Simpson é dado por

$$\boxed{E_S = -\frac{f^{(4)}(\xi)}{2880}(b-a)^5} \quad \Rightarrow \quad \boxed{|E_S| \leq \frac{M_4}{2880}(b-a)^5}$$

onde M_4 é um limitante para $|f^{(4)}(x)|$ em $[a,b]$.

7.2.5 Resumo

- Retângulo ($n = 0$)

$$\boxed{E_R = \frac{f'(\xi)}{2}(b-a)^2} \qquad \boxed{|E_R| \leq \frac{M_1}{2}(b-a)^2}$$

- Ponto Médio ($n = 0$)

$$\boxed{E_M = \frac{f''(\xi)}{24}(b-a)^3} \qquad \boxed{|E_M| \leq \frac{M_2}{24}(b-a)^3}$$

- Trapézio ($n = 1$)

$$\boxed{E_T = -\frac{f''(\xi)}{12}(b-a)^3} \qquad \boxed{|E_T| \leq \frac{M_2}{12}(b-a)^3}$$

- Simpson 1/3 ($n = 2$)

$$\boxed{E_S = -\frac{f^{(4)}(\xi)}{2880}(b-a)^5} \qquad \boxed{|E_S| \leq \frac{M_4}{2880}(b-a)^5}$$

Exemplo Calcule

$$\ln 2 = \int_1^2 \frac{1}{x}\, dx \approx 0.69314718$$

usando as regras

- ponto médio
- trapézio
- Simpson 1/3
- Simpson 3/8

e faça uma análise do erro. □

7.3 Fórmulas Repetidas

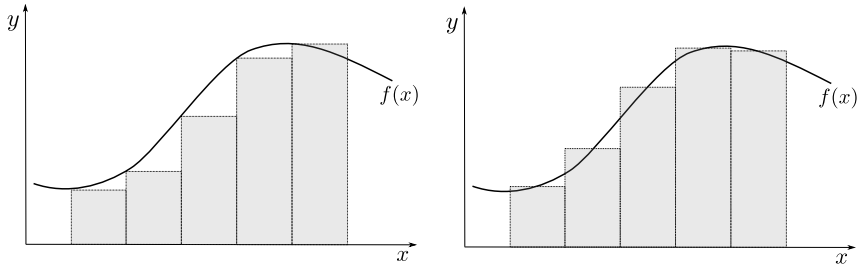
Quando o intervalo é grande, pode não ser conveniente aumentar o grau do polinômio interpolador para estabelecer outras regras de interpolação.

- se o intervalo é grande, o erro é grande
- fórmulas complicadas
- problemas com interpolação de alta ordem para pontos igualmente espaçados

Uma idéia alternativa é dividir o intervalo original em diversos subintervalos e aplicar a regra de integração em cada subintervalo.

Essas são as chamadas regras **repetidas**, **generalizadas** ou **compostas**:

- Regra do retângulo repetida
- Regra do ponto médio repetida
- Regra do trapézio repetida
- Regra de 1/3 de Simpson repetida
- Regra do 3/8 de Simpson repetida



7.3.1 Regra do retângulo e do ponto médio repetidas

Para começar vamos aplicar a idéia às regras do retângulo e do ponto médio. Dividimos o intervalo $[a, b]$ em m subintervalos, com $x_0 = a$ e $x_m = b$ e $x_i = a + ih$ para $i = 0, \ldots, m$. Então

$$I = \int_a^b f(x)\ dx = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\ dx$$

Se aplicarmos a regra do retângulo a cada subintervalo, temos a regra do retângulo repetida, isto é

$$\boxed{I_R^R = \sum_{i=1}^{m} hf(x_{i-1})}$$

e para a regra do ponto médio temos

$$\boxed{I_M^R = \sum_{i=1}^{m} hf\left(\frac{x_{i-1} + x_i}{2}\right)}$$

7.3.2 Regra do trapézio repetida

A fórmula de integração da regra do trapézio é

$$I_T = \frac{h}{2}[f(x_0) + f(x_1)] \tag{7.8}$$

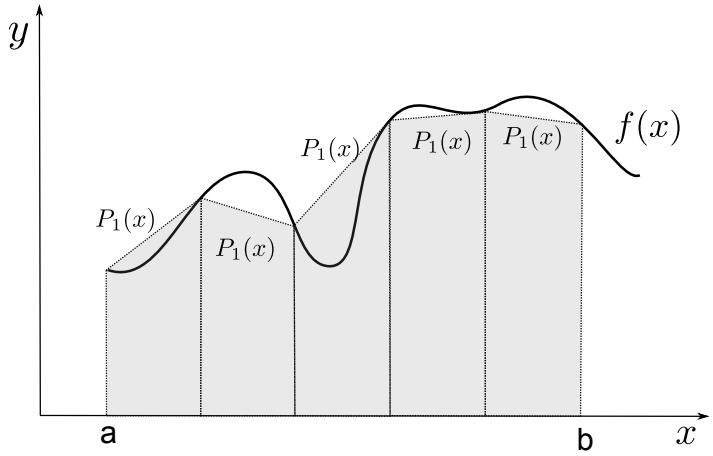
Subdividindo o intervalo de integração $[a, b]$ em m subintervalos iguais e usando a Equação 7.8 a cada 2 pontos, temos

$$\begin{aligned} I_T^R &= \frac{h}{2}[f(x_0) + f(x_1)] + \frac{h}{2}[f(x_1) + f(x_2)] + \ldots + \frac{h}{2}[f(x_{m-1}) + f(x_m)] \\ &= \frac{h}{2}\left[f(x_0) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{m-1}) + f(x_m)\right] \end{aligned}$$

e assim temos

$$\boxed{I_T^R = \frac{h}{2}\sum_{i=0}^{m} c_i f(x_i)} \tag{7.9}$$

onde $c_0 = c_m = 1$ e $c_i = 2$ para $i = 1, \ldots, m-1$.



7.3.3 Regra de 1/3 de Simpson repetida

De forma similar vamos deduzir a versão repetida da fórmula 1/3 de Simpson

$$I_S = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] \tag{7.10}$$

Dividindo o intervalo $[a, b]$ em m (múltiplo de 2) subintervalos iguais e aplicando a Equação 7.10 a cada 3 pontos temos

$$\begin{aligned} I_S^R &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] + \frac{h}{3}\left[f(x_2) + 4f(x_3) + f(x_4)\right] \\ &+ \ldots + \frac{h}{3}\left[f(x_{m-2}) + 4f(x_{m-1}) + f(x_m)\right] \\ &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + + \ldots + 4f(x_{m-1}) + f(x_m)\right] \end{aligned}$$

e assim

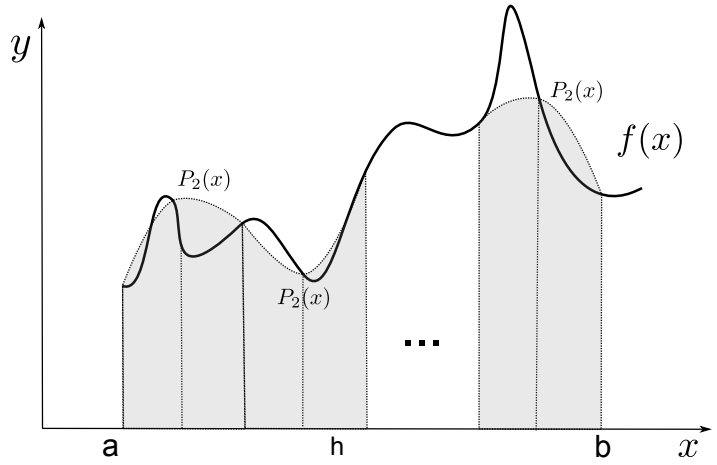
$$\boxed{I_S^R = \frac{h}{3}\sum_{i=0}^{m} c_i f(x_i)} \tag{7.11}$$

onde $c_0 = c_m = 1$, $c_i = 4$ se i for impar e $c_i = 2$ se i for par.

7.3.4 Regra de 3/8 de Simpson repetida

Considerando

$$I_{S3/8} = \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right] \tag{7.12}$$



Subdividindo o intervalo em m (agora múltipli de 3) subintervalos, e aplicando a Equação 7.12 a cada 4 pontos, temos

$$\begin{aligned} I^R_{S3/8} &= \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right] \\ &+ \frac{3h}{8}\left[f(x_3) + 3f(x_4) + 3f(x_5) + f(x_6)\right] \\ &+ \ldots + \frac{3h}{8}\left[f(x_{m-3}) + 3f(x_{m-2}) + 3f(x_{m-1}) + f(x_m)\right] \\ &= \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + 2f(x_3) + 3f(x_4) + \ldots + f(x_m)\right] \end{aligned}$$

que de forma geral pode ser escrita como

$$\boxed{I^R_{S3/8} = \frac{3h}{8}\sum_{i=0}^{m} c_i f(x_i)} \tag{7.13}$$

7.4 Análise do erro para fórmulas repetidas

Todos os limitantes para o erro cometidos que vimos para as fórmulas (simples) envolvem alguma potência do tamanho do intervalo $(b-a)$, e a menos que este seja pequeno, os limitantes não serão pequenos.
Como vimos, na prática é comum dividir o intervalo em subintervalos e fazer uso das regras

repetidas.
Vamos agora considerar o erro cometido para as fórmulas repetidas. Para isso, basta somar

o erro cometido em cada subintervalo.
No que segue, consideramos que o espaçamento em cada subintervalo i é o mesmo, isto é,

$h_i = h$.

7.4.1 Regra do Retângulo

Nesse caso, temos que o erro para o caso simples, no intervalo $[a, b]$ é dado por

$$E_R = \frac{f'(\xi)}{2}(b-a)^2$$

Como no caso das fórmulas repetidas, cada subintervalo $[x_{i-1}, x_i]$ com $i = 1, \ldots, m$, tem o mesmo tamanho h, somando o erro de cada subintervalo temos

$$E_R^R = \sum_{i=1}^{m} \frac{f'(\xi_i)}{2}h^2 = f'(\xi)\sum_{i=1}^{m}\frac{h^2}{2} = f'(\xi)\frac{h^2}{2}m$$

Como $h = (b-a)/m$, temos que $m = (b-a)/h$ e assim

$$E_R^R = f'(\xi)\frac{h^2}{2}\frac{(b-a)}{h} \quad \Rightarrow \quad \boxed{E_R^R = \frac{f'(\xi)}{2}(b-a)h}$$

7.4.2 Regra do Retângulo

E temos o seguinte limitante

$$|E_R^R| \leq \frac{M_1}{2}(b-a)h$$

onde $M_1 = \max_{a\leq x\leq b}|f'(x)|$. Note que M_1 é um limitante para a primeira derivada em todo o intervalo $[a, b]$. Poderíamos obter um limitante melhor considerando limitantes para $|f'(x)|$ em cada subintervalo.

7.4.3 Regra do Ponto Médio

Para a regra do ponto médio já mostramos que

$$E_M = \frac{f''(\xi)}{24}(b-a)^3$$

agora para a aplicação da fórmula repetida, temos

$$\begin{aligned} E_M^R &= \sum_{i=1}^{m}\frac{f''(\xi_i)}{24}h^3 = f''(\xi)\sum_{i=1}^{m}\frac{h^3}{24} \\ &= f''(\xi)\frac{h^3}{24}m = f''(\xi)\frac{h^3}{24}\frac{(b-a)}{h} \\ E_M^R &= f''(\xi)\frac{h^2}{24}(b-a) \end{aligned}$$

Limitante superior para o erro

$$\boxed{|E_M^R| \leq \frac{M_2}{24}h^2(b-a)}$$

7.4.4 Regra do Trapézio e 1/3 de Simpson

De forma similar obtemos para a regra do trapézio

$$E_T^R = -\frac{f''(\xi)}{12}(b-a)h^2 \quad \Rightarrow \quad \boxed{|E_T^R| \leq \frac{M_2}{12}(b-a)h^2}$$

e para a regra 1/3 de Simpson temos

$$E_S^R = -\frac{f^{(4)}(\xi)}{180}(b-a)h^4 \quad \Rightarrow \quad \boxed{|E_S^R| \leq \frac{M_4}{180}(b-a)h^4}$$

Exemplo Usando a regra 1/3 de Simpson obter a integral

$$\int_0^{1.2} e^x \cos(x)\ dx$$

com 3 casas decimais corretas (erro menor do que 10^{-3}), sabendo que o valor exato da integral é 1.648774427. **Solução** Quantos intervalos devem ser usados para calcular de forma aproximada o valor da integral

$$\int_0^{1.2} e^x \cos(x)\ dx$$

usando a regra dos Trapézios repetida de forma que o erro seja menor do que 0.5×10^{-3}. □

Exemplo Quantos intervalos devem ser usados para aproximar $\int_0^1 e^{x^{-2}}\ dx$ para que a aproximação pela regra do Trapézio repetida tenha erro menor do que 10^{-5}?

Exemplo Determine h para que a regra do ponto médio forneça o valor de

$$\int_{0.2}^{0.8} \sin(x)\ dx$$

com erro inferior a 10^{-5}.

7.5 Método dos Coeficientes Indeterminados

Como vimos podemos obter uma regra para integração numérica ao integrar o polinômio interpolador do integrando.

Veremos agora uma forma alternativa de derivar as fórmulas de integração numérica até então estudadas.

Como vimos as fórmulas de integração numérica são do tipo

$$\int_a^b f(x)\ dx \approx \sum w_i f(x_i)$$

onde w_i são constantes e $f(x_i)$ são os valores da função f nos pontos x_i.

Ex: regra do Trapézio, Simpson 1/3

$$I_T = \frac{h}{2}[f(x_0) + f(x_1)], \qquad I_S = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)]$$

No método dos coeficientes indeterminados, consideramos os pontos $x_0, \ldots, x_n$ dados e buscamos determinar os coeficientes $w_0, \ldots, w_n$ de tal forma que a fórmula de integração numérica

$$\int_a^b f(x)\ dx \approx \sum w_i f(x_i)$$

seja exata para certos tipos de funções, como por exemplo quando $f(x)$ é um polinômio de grau $\leq n$.

Vamos ilustrar a idéia do método através de um exemplo.

Procuramos uma fórmula

$$\int_a^b f(x)\ dx \approx w_0 f(x_0) + w_1 f(x_1) + w_2 f(x_2) \tag{7.14}$$

que será exata para todos os polinômios de grau ≤ 2, isto é, quando $f(x) = c_0 + c_1 x + c_2 x^2$ então a fórmula de integração numérica fornece o valor exato da integral.

Como

$$\begin{aligned}\int_a^b f(x)\ dx &= \int_a^b [c_0 + c_1 x + c_2 x^2]\ dx \\ &= c_0 \int_a^b dx + c_1 \int_a^b x\ dx + c_2 \int_a^b x^2\ dx\end{aligned}$$

Isto é, exigir que a fórmula integre a função $f(x)$ (nesse exemplo um polinômio de grau 2) exatamente é o mesmo que exigir que a fórmula integre as funções base 1, x e x^2 exatamente.

Assim usando a Equação (7.14), temos

$$\begin{aligned} f(x) = 1 \quad &\Rightarrow \quad w_o 1 + w_1 1 w_2 1 = \int_a^b dx \\ f(x) = x \quad &\Rightarrow \quad w_o x_0 + w_1 x_1 w_2 x_2 = \int_a^b x\ dx \\ f(x) = x^2 \quad &\Rightarrow \quad w_o x_0^2 + w_1 x_1^2 w_2 x_2^2 = \int_a^b x^2\ dx \end{aligned}$$

Calculando as integrais

$$\begin{aligned}\int_a^b dx &= (b-a) \\ \int_a^b x\ dx &= \frac{(b^2 - a^2)}{2} \\ \int_a^b x^2\ dx &= \frac{(b^3 - a^3)}{3}\end{aligned}$$

assim considerando que $x_0 = a$, $x_1 = (a+b)/2$ e $x_2 = b$, podemos escrever as equações de forma matricial como

$$\begin{bmatrix} 1 & 1 & 1 \\ a & \frac{a+b}{2} & b \\ a^2 & \left(\frac{a+b}{2}\right)^2 & b^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \frac{b^3-a^3}{3} \end{bmatrix}$$

Resolvendo esse sistema de equações lineares, chegamos a solução (coeficientes da fórmula de integração):

$$w_0 = \frac{b-a}{6}, \quad w_1 = \frac{2(b-a)}{3}, \quad w_2 = \frac{b-a}{6}$$

que reconhecemos como a regra de Simpson $1/3$

$$I_s = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)]$$
$$\Rightarrow \quad \frac{b-a}{6}f(x_0) + \frac{4(b-a)}{6}f(x_1) + \frac{(b-a)}{6}f(x_2)$$

pois $h = (b-a)/2$. □

Exemplo Encontre a fórmula

$$\int_0^1 f(x)\ dx \approx w_0 f(x_0) + w_1 f(x_1)$$

que é exata para funções da forma

$$f(x) = ae^x + b\cos\left(\tfrac{\pi x}{2}\right)$$

Solução do Exemplo Para $f(x) = \cos\left(\frac{\pi x}{2}\right)$ temos

$$\int_0^1 f(x)\ dx = \int_0^1 \cos\left(\frac{\pi x}{2}\right)\ dx = \frac{2\sin\left(\frac{\pi x}{2}\right)}{\pi}\bigg|_0^1 = \frac{2}{\pi}$$

E assim temos a equação

$$w_0 f(0) + w_1 f(1) = \int_0^1 \cos\frac{\pi x}{2}\ dx = \frac{2}{\pi}$$

como $f(0) = \cos(0) = 1$ e $f(1) = \cos(\pi/2) = 0$ temos

$$\boxed{w_0 = \frac{2}{\pi}} \tag{7.15}$$

Para $f(x) = e^x$ temos

$$\int_0^1 e^x\ dx = e^x\Big|_0^1 = e - 1 \tag{7.16}$$

e assim

$$w_0 f(0) + w_1 f(1) = e - 1 \quad \Rightarrow \quad w_0 + w_1 e = e - 1$$

de onde obtemos

$$\boxed{w_1 = 1 - \frac{1}{e} - \frac{2}{\pi e}} \tag{7.17}$$

E assim para

$$I = w_0 f(x_0) + w_1 f(x_1)$$

obtemos a fórmula

$$I = \frac{2}{\pi} f(x_0) + \left(1 - \frac{1}{e} - \frac{2}{\pi e}\right) f(x_1)$$

que integra

$$\int_{x_0}^{x_1} \left[a e^x + b \cos(\pi x/2)\right] \, dx$$

para quaisquer valores de a e b de forma exata.
□

7.6 Grau de precisão

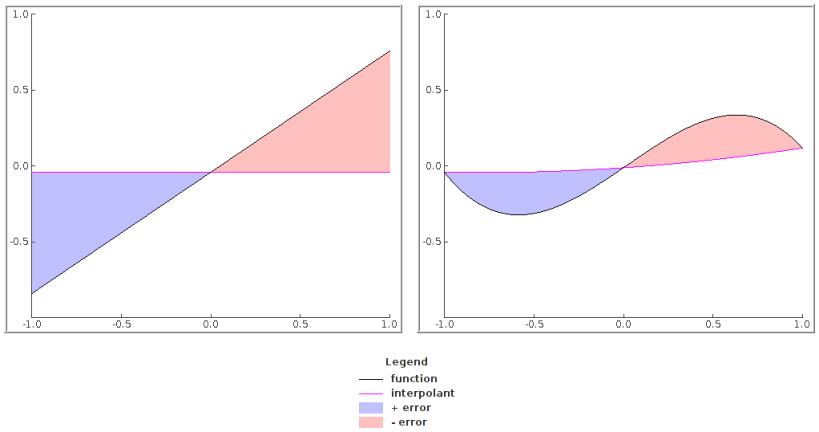
Definição 13 (Grau de precisão)**.** Dizemos que uma regra de Newton-Cotes de n pontos tem grau de precisão (ou é de grau polinomial) d se ela é exata (i.e. o erro cometido é zero) para todo polinômio de grau d, mas não é exata para algum polinômio de grau $d+1$.

Como uma regra de n pontos de Newton-Cotes é baseada em um polinômio interpolador de grau $n-1$, é de se esperar que esta tenha grau pelo menos $n-1$, e de fato tem, pois isto foi definido em sua construção (met. coeficientes indeterminados). Sendo assim podemos esperar que

- Ponto médio: 1 ponto (x_0) $\Rightarrow$ grau 0
- Trapézio: 2 pontos (x_0, x_1) $\Rightarrow$ grau 1
- Simpson 1/3: 3 pontos (x_0, x_1, x_2) $\Rightarrow$ grau 2
- Simpson 3/8: 4 pontos (x_0, x_1, x_2, x_3) $\Rightarrow$ grau 3

Entretanto, vimos na análise do erro cometido que o erro para a regra do ponto médio depende da segunda derivada do integrando, a qual é nula para polinômios lineares e constantes (logo erro é zero). Isso implica que a regra do ponto médio de fato integra polinômios lineares de forma exata e portanto tem grau 1.

De forma similar vimos que o erro na regra de Simpson 1/3 depende da derivada quarta do integrando, e portanto esta regra integra exatamente polinômios de grau ≤ 3. Sendo assim



- Ponto médio: 1 ponto (x_0) $\Rightarrow$ **grau 1**
- Trapézio: 2 pontos (x_0, x_1) $\Rightarrow$ **grau 1**
- Simpson 1/3: 3 pontos (x_0, x_1, x_2) $\Rightarrow$ **grau 3**
- Simpson 3/8: 4 pontos (x_0, x_1, x_2, x_3) $\Rightarrow$ **grau 3**

Cancelamento do erro

Em geral, em uma regra de Newton-Cotes com n pontos (n ímpar), temos um grau de precisão extra além do grau do polinômio interpolador.

Esse fenômeno ocorre devido ao cancelamento de erros negativos e positivos, como ilustra na Figura a seguir para os métodos do ponto médio e Simpson 1/3.

Em geral, uma regra de n pontos de Newton-Cotes é de grau

- $n-1$, se n é par
- n, se n é ímpar

7.7 Mudança de intervalo

Seja $x \in [a, b]$. Podemos fazer a seguinte mudança de variável

$$x(t) = \frac{(b-a)t}{2} + \frac{b+a}{2}, \qquad t \in [-1, 1]$$

Qualquer que seja $x \in [a, b]$, existe $t \in [-1, 1]$ tal que $x = x(t)$. Sendo assim

$$\frac{dx}{dt} = x'(t) = \frac{b-a}{2} \quad \Rightarrow \quad dx = \frac{b-a}{2} dt$$

logo usando $x = x(t)$ e $dx = x'(t)\ dt$ temos

$$I = \int_a^b f(x)\ dx = \int_{-1}^{1} f(x(t))\ x'(t)\ dt = \int_{-1}^{1} F(t)\ dt$$

onde $F(t) = f(x(t))\ x'(t) = f\left(t\frac{(b-a)}{2} + \frac{b+a}{2}\right)\frac{b-a}{2}$

7.8 Quadratura de Gauss

Como vimos, as regras de integração de Newton-Cotes são simples e efetivas, mas possuem algumas desvantagens:

- uso de muitos pontos para interpolação de alta ordem pode gerar alguns problemas;
- requer a avaliação de $f(x)$ nos pontos do extremo do intervalo, onde geralmente ocorrem singularidades;
- não possuem um grau de precisão tão alto quanto poderiam.

Veremos que algumas dessas desvantagens são contornadas pela quadratura de Gauss (ou quadratura Gaussiana).

Estamos interessados em obter uma fórmula de integração numérica a ser avaliada em n pontos, a qual é dada por

$$I = \int_a^b f(x)\ dx = w_0 f(x_0) + w_1 f(x_1) + \ldots + w_{n-1} f(x_{n-1}),$$

onde agora temos que determinar os coeficientes w_i e os pontos x_i para $i = 0, \ldots, n-1$ de forma a obter a **melhor precisão possível**.

Logo, temos as seguintes incógnitas $x_0, x_1, \ldots, x_{n-1}$ e $w_0, w_1, \ldots, w_{n-1}$, isto é, um total de $2n$ incógnitas a serem determinadas. Sendo assim, podemos esperar que as regras que iremos obter sejam capazes de integrar exatamente polinômios de grau $\leq 2n-1$ uma vez que estes são definidos por $2n$ parâmetros.

Na quadratura de Gauss trabalha-se com integrais definidas no intervalo $[-1, 1]$ da seguinte forma

$$I = \int_{-1}^{1} F(t)dt \approx w_0 F(t_0) + w_1 F(t_1),$$

onde ambos os pontos t_0 e t_1 e os pesos w_0 e w_1 devem ser determinados de modo que a regra seja exata para polinômios de grau ≤ 3, pois com 2 pontos tem que se determinar 4 parâmetros: t_0, t_1, w_0 e w_1.

De forma geral, uma fórmula de quadratura de Gauss com n pontos $t_0, t_1, \ldots, t_{n-1}$, tem grau de precisão polinomial dado por $2n-1$. Por exemplo, se tivermos 2 pontos, isto é, t_0 e t_1, a quadratura de Gauss tem precisão $2(2) - 1 = 4 - 1 = 3$. Com 3 pontos, tem-se grau de precisão dado por $2(3) - 1 = 5$.

Quadratura de Gauss com 2 pontos

Vamos deduzir o caso

$$I = \int_{-1}^{1} F(t)\ dt = w_0 F(t_0) + w_1 F(t_1),$$

usando o método dos coeficientes indeterminados. Queremos encontrar w_0, w_1, t_0 e t_1 isto é, 4 parâmetros, logo, a regra de integração que vamos deduzir deve integrar exatamente um polinômio de grau ≤ 3. Sendo assim, podemos escrever

$$F(t) = c_0\phi_0(t) + c_1\phi_1(t) + c_2\phi_2(t) + c_3\phi_3(t)$$

onde as funções base são: $\phi_j(t) = t^j$. Agora basta exigir que a regra que queremos encontrar, i.e., $w_0F(t_0) + w_1F(t_1)$ integre exatamente cada uma das funções base.

Primeiro temos que exigir que a regra integre $\phi_0(t)$ exatamente. Neste caso, considere $F(t) = \phi_0(t)$, e assim

$$w_0\phi_0(t_0) + w_1\phi_0(t_1) = \int_{-1}^{1} \phi_0(t)\ dt,$$

como $\phi_0(t) = 1$ isto resulta em

$$w_0 1 + w_1 1 = \int_{-1}^{1} 1\ dt.$$

De forma similar, repetimos o processo para ϕ_1, ϕ_2 e ϕ_3. Ao final do procedimento encontramos as seguintes equações

$$\begin{aligned}
\phi_0(t) = 1 \quad &\Rightarrow \quad w_0\ 1 + w_1\ 1 = \int_{-1}^{1} dt = t\Big|_{-1}^{1} = 2 \\
\phi_1(t) = t \quad &\Rightarrow \quad w_0t_0 + w_1t_1 = \int_{-1}^{1} t\ dt = \frac{t^2}{2}\bigg|_{-1}^{1} = 0 \\
\phi_2(t) = t^2 \quad &\Rightarrow \quad w_0t_0^2 + w_1t_1^2 = \int_{-1}^{1} t^2\ dt = \frac{t^3}{3}\bigg|_{-1}^{1} = \tfrac{2}{3} \\
\phi_3(t) = t^3 \quad &\Rightarrow \quad w_0t_0^3 + w_1t_1^3 = \int_{-1}^{1} t^3\ dt = \frac{t^4}{4}\bigg|_{-1}^{1} = 0
\end{aligned}$$

que é um **sistema de equações não-lineares** para determinar w_0, w_1, t_0 e t_1, isto é

$$\begin{aligned}
w_0 + w_1 &= 2, \\
w_0t_0 + w_1t_1 &= 0, \\
w_0t_0^2 + w_1t_1^2 &= 2/3, \\
w_0t_0^3 + w_1t_1^3 &= 0.
\end{aligned}$$

Em geral precisamos recorrer a método numéricos para resolver sistemas de **equações não-lineares** como, por exemplo, o método de Newton. Neste caso, em particular, é simples encontrar uma solução considerando $t_0 = -t_1$, o que resulta em

$$-w_0t_1 + w_1t_1 = 0 \quad \Rightarrow \quad t_1(w_1 - w_0) = 0 \quad \Rightarrow \quad w_0 = w_1$$

assim

$$w_0 + w_1 = 2 \quad \Rightarrow \quad \boxed{w_0 = w_1 = 1}$$

e ainda temos que

$$t_0^2 + t_1^2 = \frac{2}{3} \quad \Rightarrow \quad 2t_1^2 = \frac{2}{3} \quad \Rightarrow \quad \boxed{t_1 = \frac{\sqrt{3}}{3}}$$

e como $t_0 = -t_1$, resulta em $\boxed{t_0 = -\sqrt{3}/3}$. Logo, os pesos e pontos são dados por

$$w_0 = 1, \quad w_1 = 1, \quad t_0 = -\tfrac{\sqrt{3}}{3}, \quad t_1 = \tfrac{\sqrt{3}}{3},$$

pesos		pontos	
w_0	0.555	t_0	$-\sqrt{\frac{3}{5}}$
w_1	0.888	t_1	0
w_2	0.555	t_2	$\sqrt{\frac{3}{5}}$

e assim define-se a seguinte regra de integração numérica

$$I = \int_{-1}^{1} F(t)\ dt \approx w_0 F(t_0) + w_1 F(t_1) = F\left(-\frac{\sqrt{3}}{3}\right) + F\left(\frac{\sqrt{3}}{3}\right)$$

que é chamada de quadratura de Gauss com 2 pontos e é exata para polinômios de grau ≤ 3.

Como vimos, uma fórmula de quadratura de Gauss com apenas 2 pontos é capaz de integrar polinômios de grau até 3, enquanto que as fórmulas de Newton-Cotes com 2 pontos (Regra do Trapézio) integram apenas polinômios de grau 1.

Quadratura de Gauss com 3 pontos

Para o caso com 3 pontos ($t_0, t_1, t_2 \to n = 2$) temos $2n + 1 = 5$ e portanto essa quadratura de Gauss é capaz de integrar exatamente polinômios de grau ≤ 5.

$$I = \int_{-1}^{1} F(t)\ dt = w_0 F(t_0) + w_1 F(t_1) + w_2 F(t_2)$$

Considerando $\phi_0 = 1, \phi_1 = t, \phi_2 = t^2, \phi_3 = t^3, \phi_4 = t^4$ e $\phi_5 = t^5$

$$w_0 + w_1 + w_2 = \int_{-1}^{1} dt = 2$$

$$w_0 t_0 + w_1 t_1 + w_2 t_2 = \int_{-1}^{1} t\ dt = 0$$

$$w_0 t_0^2 + w_1 t_1^2 + w_2 t_2^2 = \int_{-1}^{1} t^2\ dt = 2/3$$

$$w_0 t_0^3 + w_1 t_1^3 + w_2 t_2^3 = \int_{-1}^{1} t^3\ dt = 0$$

$$w_0 t_0^4 + w_1 t_1^4 + w_2 t_2^4 = \int_{-1}^{1} t^4\ dt = 2/5$$

$$w_0 t_0^5 + w_1 t_1^5 + w_2 t_2^5 = \int_{-1}^{1} t^5\ dt = 0$$

A solução do sistema fornece Em geral as fórmulas de quadratura Gaussiana são dadas em forma de tabelas com os coeficientes (pesos) w_i e pontos t_i a serem usados na fórmula

$$I = \int_{-1}^{1} F(t)\ dt \approx \sum_{i=0}^{n} w_i F(t_i),$$

e como vimos essas regras de integração utilizam $n + 1$ pontos (e pesos) e, portanto, tem grau de precisão $2n + 1$ por construção.

Exemplos

A seguir dois exemplos ilustram a aplicação da técnica de quadratura de Gauss.

Exemplo 52

Calcule $I = \int_1^3 3e^x \, dx$ usando a quadratura Gaussiana com 2 pontos.
Inicialmente é necessário realizar a mudança de intervalo de $[-2, 0]$ para $[-1, 1]$, isto é

$$x(t) = \frac{(b-a)t}{2} + \frac{b+a}{2} = t + 2$$

logo

$$x'(t) = \frac{dx}{dt} = 1 \quad \Rightarrow \quad dx = dt$$

assim

$$\int_1^3 3e^x \, dx = \int_{-1}^1 3e^{(t+2)} \, 1 \, dt$$

Precisamos avaliar $F(t) = 3e^{(t+2)}$ em $t = -\sqrt{3}/3$ e $t = \sqrt{3}/3$:

$$F(-0.577350) = 3e^{(-0.577350+2)} = 12.444292$$
$$F(0.577350) = 3e^{(0.577350+2)} = 39.486647$$

Assim calculamos a integral de forma aproximada como

$$I = F(-0.577350) + F(0.577350) = 51.930938$$

Se usarmos uma regra com 3 pontos temos

$$I = \frac{5}{9}F(-\sqrt{\frac{3}{5}}) + \frac{8}{9}F(0) + \frac{5}{9}F(\sqrt{\frac{3}{5}}) = 52.1004$$

Compare a aproximação numérica com o valor exato da integral: $3[e^3 - e] = 52.1018$.

Exemplo 53

Calcular a integral $I = \int_{-2}^0 (x^2 - 1) \, dx$ com a quadratura de Gauss de 2 pontos. **Solução do Exemplo** Mudança de intervalo

$$x(t) = \frac{(0-(-2))t}{2} + \frac{(0-2)}{2} = t - 1$$
$$x'(t) = \frac{dx}{dt} = 1 \quad \Rightarrow \quad dx = dt$$

e portanto

$$\int_{-2}^{0}(x^2-1)\,dx = \int_{-1}^{1}[(t-1)^2-1]\;1\;dx$$
$$= \int_{-1}^{1} t^2-2t+1-1\;dt = \int_{-1}^{1}[t^2-2t]\;dt$$

A aproximação da integral é dada por

$$I = F\left(-\frac{\sqrt{3}}{3}\right)+F\left(\frac{\sqrt{3}}{3}\right) = 1.488-0.821 = 0.66666$$

a qual pode ser comparada com o valor exato que é

$$\int_{-1}^{1}[t^2-2t]\;dt = \frac{t^3}{3}\Big|_{-1}^{1} - t^2\Big|_{-1}^{1} = \frac{2}{3} = 0.66666$$

De onde podemos ver que de fato a quadratura de Gauss de 2 pontos integra polinômios de grau ≤ 3 de forma exata.
□