MATEMÁTICA COMPUTACIONAL
NOTAS DE AULAS

M. Teresa Diogo
Departamento de Matemática
Instituto Superior Técnico, Lisboa, Portugal

Murilo Tomé
Instituto de Ciências Matemáticas e Computação
USP de São Carlos, Brasil

versão 2014

Índice

1 Conceitos Básicos da Teoria dos Erros

1.1 Introdução

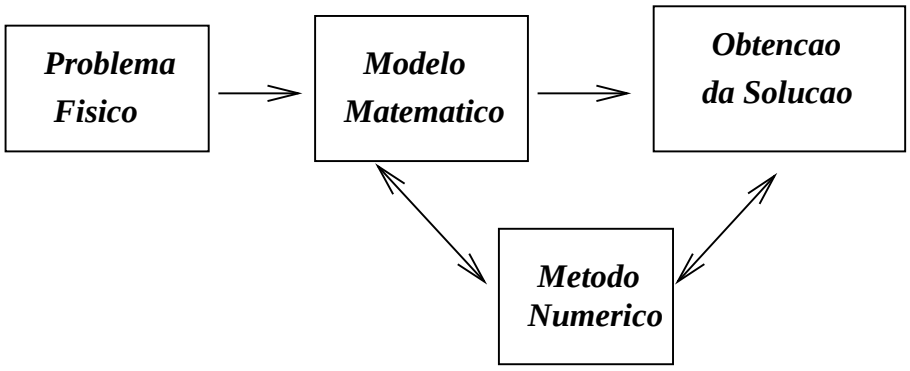
Neste curso de Análise Numérica, vamos estudar métodos que permitem resolver problemas científicos usando um computador. Podemos considerar três fases distintas no processo de resolução. A *primeira fase* consiste em expressar matematicamente o problema científico. Obtem-se assim um modelo matemático, que podemos definir como sendo uma expressão matemática que relaciona grandezas (físicas, de engenharia, de economia, etc). Um exemplo dum modelo matemático será a expressão da segunda lei de Newton do movimento

$$m\frac{d^2}{dt^2}x(t) = F(t) \tag{1.1}$$

que relaciona a posição, num instante t, dum corpo com massa m, com uma força aplicada F.

Depois de obtido o modelo matemático, a *segunda fase* consiste em escolher métodos numéricos que permitam obter, de forma eficiente, uma solução aproximada do problema em questão. A eficiência do método depende não só da precisão que nos pode garantir, mas também da facilidade com que pode ser implementado. Os métodos numéricos que aqui consideraremos serão expressos na forma de um algoritmo: uma sequência de instruções descrevendo um número finito de passos que devem ser executados por uma ordem especificada. Concentrar-nos-emos no estudo de métodos numéricos para resolver problemas no âmbito dos seguintes tópicos: equações não lineares, sistemas de equações lineares, interpolação, aproximação de funções no sentido dos mínimos quadrados e integração. Por exemplo, em relação à equação (1), é possível aproximar $x(t)$, usando técnicas de integração numérica que mais tarde descreveremos.

A *terceira e última fase* do processo que estamos a descrever consiste na implementação do algoritmo no computador. Obtemos assim uma solução numérica do problema inicial.



Tipos de erros num processo de cálculo

Para analisar a qualidade dum resultado numérico, devemos estar cientes das possíveis fontes de erro ao longo de todo o processo. Podemos considerar basicamente três tipos de erros: erros inerentes, erros do método e o erro computacional. Fazemos notar que é o efeito do erro total, o qual engloba todos os tipos de erros originados ao longo dum processo, que afecta a qualidade duma solução aproximada. Há certos problemas nos quais a um "pequeno" erro introduzido num certo passo dos cálculos, corresponde um erro "grande" na solução final. Um problema deste tipo diz-se *mal condicionado*.

Erros inerentes

Em geral, o modelo matemático não traduz exactamente a realidade. A fim de se conseguir um modelo que não seja demasiado complicado e difícil de tratar matematicamente, é muitas vezes necessário impor certas restrições idealistas. Os dados e parâmetros dum problema são muitas vezes resultados de medições experimentais e, portanto, afectados de alguma incerteza.

Podemos também mencionar a impossibilidade de representar exactamente certas constantes matemáticas (por exemplo teremos de substituir π por 3.1416 ou 3.141593, etc.).

Erros do método

Uma outra classe de erros resulta do uso de fórmulas que nos dão valores aproximados e não exactos. Por exemplo, consideremos a série de MacLaurin para representar e^x

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots$$

Se quisermos aproximar o valor e^α, onde α é um número real, teremos de nos limitar a usar um número finito de termos da série

$$e^\alpha \simeq 1 + \alpha + \frac{\alpha^2}{2!} + \cdots + \frac{\alpha^k}{k!}$$

O erro que cometemos é chamado *erro de truncatura* e corresponde neste exemplo ao resto de ordem k da série.

Erro computacional

O erro computacional é devido ao facto de o computador usar apenas um número finito de dígitos para representar os números reais. A maioria dos números e dos resultados de operações aritméticas não podem ser representados exactamente no computador. São assim originados os *erros de arredondamento*. Na secção seguinte consideraremos este tópico em maior detalhe.

1.2 Erro absoluto e erro relativo

Seja x o valor exacto duma grandeza real e seja $\tilde{x}$ uma aproximação de x, ou seja $x \simeq \tilde{x}$. Designa-se por e_x

$$e_x = x - \tilde{x} \tag{1.2}$$

o *erro de $\tilde{x}$ em relação a x*. Ao valor $|e_x|$ chama-se *erro absoluto* de $\tilde{x}$. Se $x \neq 0$, denota-se por δ_x

$$\delta_x = \frac{x - \tilde{x}}{x} \tag{1.3}$$

e chama-se a $|\delta_x|$ o *erro relativo de $\tilde{x}$* .

Ao produto 100 $|\delta_x|$ expresso em percentagem chama-se *percentagem de erro*.

NOTA Atendendo a que

$$\frac{x - \tilde{x}}{\tilde{x}} = \frac{\delta_x}{1 - \delta_x} = \delta_x(1 + \delta_x + {\delta_x}^2 + \cdots) = \delta_x + {\delta_x}^2 + {\delta_x}^3 + \cdots,$$

com δ_x suficientemente pequeno, usa-se na prática

$$\delta_x = \frac{x - \tilde{x}}{\tilde{x}},$$

o que equivale a desprezar os termos de ordem superior a um no desenvolvimento acima.

Observamos que o erro absoluto pode não ser de grande utilidade para informar sobre a qualidade duma aproximação, se não se souber a ordem de grandeza da quantidade a medir. Basta dizer que um erro de um metro na medição da distância da Terra a Saturno seria óptimo, mas se um cirurgião cometesse um erro dessa ordem ao fazer uma incisão seria um desastre possivelmente até fatal ! Incluímos ainda o seguinte exemplo.

Exemplo 1.1 *Consideremos os números $x = 1/3$ e $y = 1/3000$, e as aproximações $\tilde{x} = 0.3333$ e $\tilde{y} = 0.0003$. Tem-se que $\tilde{x}$ e $\tilde{y}$ são valores aproximados de x e y respectivamente, com o mesmo erro $e_x = e_y = 0.0000333\cdots$. Porém a percentagem de erro em $\tilde{x}$ é 0.01% e a percentagem de erro em $\tilde{y}$ é 10%.*

1.3 Representação dos números no computador

1.3.1 Bases

A base decimal é aquela com que estamos mais familiarizados. Todo o número inteiro se pode representar como uma combinação linear de potências de 10, com coeficientes variando

entre 0 e 9. Por exemplo, o número 415 pode-se escrever na forma

$$\begin{aligned} 415 &= 400 + 10 + 5 \\ &= 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0. \end{aligned}$$

o que se representa simbolicamente por $(4\ 1\ 5)_{10}$ (na prática omite-se o 10). De modo geral, usa-se a notação

$$\sigma(b_m b_{m-1} \cdots b_1 b_0)_{10}$$

para representar o valor (inteiro)

$$\sigma(b_m 10^m + b_{m-1} 10^{m-1} + \cdots + b_1 10^1 + b_0 10^0),$$

com $\sigma \in \{+, -\}$ e $0 \leq b_i \leq 9, \quad b_m \neq 0$.

Os números fraccionários podem ser representados na base 10, usando o ponto decimal. Um número entre 0 e 1 pode ser escrito na forma

$$(0.a_1 a_2 \cdots a_n \cdots)_{10}$$

que representa o valor

$$a_1 10^{-1} + a_2 10^{-2} + \cdots + a_{n-1} 10^{-n+1} + a_n 10^{-n} + \cdots$$

Finalmente, um número real qualquer não nulo pode ser representado no sistema decimal por

$$\sigma(b_m b_{m-1} \cdots b_1 b_0 . a_1 a_2 \cdots a_n \cdots)_{10}$$

Por exemplo, 12.34 representa

$$1 \times 10^1 + 2 \times 10^0 + 3 \times 10^{-1} + 4 \times 10^{-2}.$$

Outras bases, tais como 2, 12, 20 e 60 foram usadas por civilizações anteriores para representação dos números. De modo geral, um número real $X \neq 0$ pode ser representado numa base β por

$$X = \sigma(b_m \beta^m + b_{m-1} \beta^{m-1} + \cdots + b_1 \beta^1 + b_0 \beta^0 + a_1 \beta^{-1} + a_2 \beta^{-2} + \cdots + a_{n-1} \beta^{-n+1} + a_n \beta^{-n} + \cdots),$$

ou, simbolicamente, por

$$\sigma(b_m b_{m-1} \cdots b_1 b_0 . a_1 a_2 \cdots a_{n-1} a_n \cdots)_\beta,$$

onde os b_i, a_i são inteiros compreendidos entre 0 e $\beta - 1$.

Exemplo 1.2

$$\begin{aligned} (101)_2 \quad & \textit{representa} \quad 1 \times 2^2 + 1 \times 2^0 = 5 \\ (1100)_2 \quad & \textit{representa} \quad 1 \times 2^3 + 1 \times 2^2 = 12 \\ (0.101)_2 \quad & \textit{representa} \quad 1 \times 2^{-1} + 1 \times 2^{-3} = 0.625 \\ (0.000110011 \cdots)_2 \quad & \textit{representa} \quad 0.1 \end{aligned}$$

1.3.2 Sistemas de vírgula flutuante

Num computador, os números reais são em geral representados em vírgula flutuante na base $\beta = 2$. Outras bases usadas por computadores são $\beta = 8$ e $\beta = 16$. De modo geral, um *número de vírgula flutuante* na base β e com n dígitos é um número que pode ser escrito na forma

$$x = \sigma \ (0.a_1a_2\cdots a_n)_\beta \times \beta^t, \quad \sigma \in \{+,-\}, \tag{1.4}$$

onde $0 \leq a_i \leq \beta - 1$, $i = 1, 2, \cdots, n$, $t_1 \leq t \leq t_2$, sendo t, t_1 e t_2 inteiros.

A $0.a_1a_2\cdots a_n$ chama-se *mantissa* ou *parte fraccionária* e a t o *expoente*. Se $a_1 \neq 0$, o número diz-se *normalizado*. No que se segue consideraremos apenas números normalizados.

Usa-se a notação VF(β, n, t_1, t_2) para designar o conjunto de números da forma (1.4) e ainda o número $x = 0$.

1.3.3 Arredondamentos

Definiremos duas regras para representar números reais no computador.

Consideremos o número real

$$x = \sigma(0.a_1a_2\cdots a_na_{n+1}\cdots)_\beta \times \beta^t, \tag{1.5}$$

onde $a_1 \neq 0$ e β é uma base par.

Para representar este número em vírgula flutuante com n dígitos, podemos simplesmente desprezar os dígitos $a_{n+1}a_{n+2}\cdots$ Este processo é conhecido por *arredondamento por corte*. O número x passa a ser representado por

$$fl(x) = \sigma \ (0.a_1a_2\cdots a_n)_\beta \times \beta^t. \tag{1.6}$$

Um outro modo de representar x em vírgula flutuante é usando o chamado *arredondamento simétrico*. Se $0 \leq a_{n+1} < \beta/2$, procede-se como no arredondamento por corte. Se $\beta/2 \leq a_{n+1} < \beta$ soma-se β^{-n} ao número $(0.a_1a_2\cdots a_n)_\beta$ e normaliza-se. Note-se que, neste caso, pode haver mudança nos dígitos e no expoente (ver terceiro exemplo a seguir).

Resumindo, no arredondamento simétrico

$$fl(x) = \sigma(0.a_1a_2\cdots a_n)_\beta \times \beta^t \qquad 0 \leq a_{n+1} < \beta/2 \tag{1.7}$$

$$fl(x) = fl\left(\sigma((0.a_1a_2\cdots a_n)_\beta + \beta^{-n}) \times \beta^t\right) \qquad \beta/2 \leq a_{n+1} < \beta. \tag{1.8}$$

Exemplo 1.3 *Suponhamos* $\beta = 10$ *e* $n = 2$

$$fl(\frac{2}{3}) = \begin{cases} 0.67 \times 10^0 & \textit{arred. simétrico} \\ 0.66 \times 10^0 & \textit{arred. por corte} \end{cases}$$

$$fl(-838) = \begin{cases} -0.84 \times 10^3 & \textit{arred. simétrico} \\ -0.83 \times 10^3 & \textit{arred. por corte} \end{cases}$$

$$fl(99.5) = \begin{cases} 0.10 \times 10^3 & \textit{arred. simétrico} \\ 0.99 \times 10^2 & \textit{arred. por corte} \end{cases}$$

1.3.4 Erros de arredondamento

O erro de arredondamento introduzido quando x é substituído pelo seu representante $fl(x)$ é a diferença $x - fl(x)$. Consideremos primeiramente o caso dum sistema decimal.

Seja $x = \sigma(0.a_1a_2\cdots a_na_{n+1}\cdots) \times 10^t$, e suponhamos que $fl(x)$ é obtido por arredondamento simétrico.

Seja $a_{n+1} \geq 5$. Usando (1.8), tem-se

$$\begin{aligned} |x - fl(x)| &= |(0.a_1\cdots a_na_{n+1}\cdots) \times 10^t - ((0.a_1\cdots a_n) + (0.0\cdots 01)) \times 10^t| \\ &= |((0.0\cdots 0a_{n+1}a_{n+2}\cdots) - 10^{-n}) \times 10^t| \\ &= |((0.a_{n+1}a_{n+2}\cdots) - 1) \times 10^{t-n}| \\ &\leq 0.5 \times 10^{t-n}. \end{aligned}$$

Se $a_{n+1} < 5$, resulta de (1.7)

$$\begin{aligned} |x - fl(x)| &= |(0.a_1\cdots a_na_{n+1}\cdots) \times 10^t - (0.a_1\cdots a_n) \times 10^t| \\ &= (0.0\cdots 0a_{n+1}a_{n+2}\cdots) \times 10^t \\ &= (0.a_{n+1}a_{n+2}\cdots) \times 10^{-n} \times 10^t \\ &\leq 0.5 \times 10^{t-n}. \end{aligned}$$

Em qualquer dos casos, tem-se o seguinte majorante do erro absoluto de arredondamento simétrico

$$|x - fl(x)| \leq 0.5 \times 10^{t-n}. \tag{1.9}$$

Para obter um majorante do erro relativo, notemos que $a_1 \neq 0$ implica $|x| \geq 0.1 \times 10^t$. Então

$$\frac{|x - fl(x)|}{|x|} \leq \frac{0.5 \times 10^{t-n}}{0.1 \times 10^t} = 5 \times 10^{-n} = 0.5 \times 10^{1-n}. \tag{1.10}$$

De modo análogo se prova que, no caso de arredondamento por corte,

$$|x - fl(x)| \leq 10^{t-n} \tag{1.11}$$

$$\frac{|x - fl(x)|}{|x|} \leq 10^{1-n}. \tag{1.12}$$

A análise efectuada para um sistema VF decimal estende-se ao caso duma base β par. Seja x um número da forma (1.33). No caso de arredondamento por corte, tem-se

$$|x - fl(x)| \leq \beta^{t-n} \tag{1.13}$$

$$\frac{|x - fl(x)|}{|x|} \leq \beta^{1-n}. \tag{1.14}$$

No caso de arredondamento simétrico, obtém-se

$$|x - fl(x)| \leq (1/2)\ \beta^{t-n} \tag{1.15}$$

$$\frac{|x - fl(x)|}{|x|} \leq (1/2)\ \beta^{1-n}. \tag{1.16}$$

É interessante notar que nenhum dos majorantes dos erros relativos depende de x, mas apenas da base β, do número de dígitos n usado pelo computador e, é claro, do processo de arredondamento utilizado.

Dado um sistema VF(β, n, t_1, t_2), define-se

$$U = \begin{cases} \beta^{1-n} & \text{arred. por corte} \\ (1/2)\ \beta^{1-n} & \text{arred. simétrico.} \end{cases}$$

A U chama-se *unidade de arredondamento* do sistema.

Assim, se $fl(x)$ é a representação em VF dum número real x, o seu erro relativo não excede U. É por isso fundamental conhecer a unidade de arredondamento do sistema de vírgula flutuante do computador que se usa.

Exemplo 1.4 *Num computador com o sistema VF(16,6,-64,63), a unidade de arredondamento simétrico é*

$$(1/2)16^{1-6} \simeq 0.476837 \times 10^{-7}.$$

1.3.5 Algarismos significativos

Finalizamos esta secção com o conceito de algarismo significativo. Seja

$$\tilde{x} = \sigma\ (0.a_1a_2\cdots a_n) \times 10^t$$

uma aproximação de x pertencente a VF(10, n, t_1, t_2).

Definição 1.1 *Diz-se que o algarismo a_i de $\tilde{x}$ é significativo se*

$$|e_x| = |x - \tilde{x}| \leq \frac{1}{2}10^{t-i}. \tag{1.17}$$

Concluimos que se

$$|e_x| = |x - \tilde{x}| \leq \frac{1}{2}10^{t-n}, \tag{1.18}$$

os n algarismos de $\tilde{x}$ são significativos.

Exercício
Em vírgula flutuante com 4 dígitos, sejam $x = 0.4537 \times 10^4$ e $\tilde{x} = 0.4501 \times 10^4$. Determine o número de algarismos significativos de $\tilde{x}$.

Tem-se que

$$|e_x| = 0.0036 \times 10^4 = (0.36 \times 10^{-2}) \times 10^4,$$

logo verifica-se (1.17) com $i = 1, 2$. Como além disso

$$|e_x| > (0.5 \times 10^{-3}) \times 10^4,$$

apenas os algarismos 4 e 5 de $\tilde{x}$ são significativos.

1.4 Propagação de erros

Uma questão importante em Análise Numérica é a do modo como um erro num certo passo dos cálculos se vai propagar. Interessa-nos saber se o efeito desse erro se tornará maior ou menor, à medida que vão sendo efectuadas as restantes operações.

1.4.1 Cálculo de valores funcionais usando dados com erros

Seja $f : D \subset \mathcal{R} \to \mathcal{R}$ e seja $\tilde{x}$ uma aproximação de x que verifica $x = \tilde{x} + e_x$. Se pretendemos calcular o valor de $f(x)$ mas apenas conhecemos $\tilde{x}$, então aproximamos $f(x)$ por $f(\tilde{x})$. Em geral, $f(\tilde{x})$ não coincidirá com $f(x)$. Por analogia com as definições da Secção 1.2, designa-se por $e_f = f(x) - f(\tilde{x})$ o erro de $f(\tilde{x})$. Admitindo que todas as operações indicadas na expressão de f podem ser efectuadas exactamente, vamos determinar um limite superior para o erro absoluto $|e_f| = |f(x) - f(\tilde{x})|$. Teremos, assim, uma indicação do efeito do erro e_x.

Se $f'(x)$ existe e é contínua, tem-se, pelo Teorema do valor médio

$$f(x) - f(\tilde{x}) = f'(\xi)(x - \tilde{x}), \tag{1.19}$$

com $\xi \in I = int(x, \tilde{x}) = (min\{x, \tilde{x}\}, max\{x, \tilde{x}\})$, donde

$$|e_f| = |f(x) - f(\tilde{x})| = |f'(\xi)(x - \tilde{x})| \leq \max_{t \in I} |f'(t)||e_x|. \tag{1.20}$$

Na prática conhece-se apenas um majorante de e_x. Se for η esse majorante, tem-se

$$|e_x| \leq \eta \Leftrightarrow \tilde{x} - \eta \leq x \leq \tilde{x} + \eta$$

e a fórmula a usar será

$$|e_f| \leq \max_{t \in [\tilde{x}-\eta, \tilde{x}+\eta]} |f'(t)|\eta. \tag{1.21}$$

Note que

$$I = int(x, \tilde{x}) \subseteq [\tilde{x} - \eta, \tilde{x} + \eta] \Rightarrow \max_{t \in I} |f'(t)| \leq \max_{t \in [\tilde{x}-\eta, \tilde{x}+\eta]} |f'(t)|.$$

A fórmula (1.20) diz respeito ao erro absoluto de f. Pode obter-se um majorante para o erro relativo de f, do modo seguinte

$$|\delta_f| = |\frac{e_f}{f(x)}| \leq \max_{t \in I} |f'(t)| \frac{|e_x|}{|x|} \frac{|x|}{|f(x)|} = \max_{t \in I} |f'(t)| \frac{|x|}{|f(x)|} |\delta_x|$$

As fórmulas de majoração deduzidas têm o inconveniente de exigir o conhecimento dum majorante para a derivada $f'(t)$, o que pode ser difícil.

Exercício. *Suponha que estamos a trabalhar num sistema decimal com 4 dígitos. Determine um majorante do erro (absoluto) que se comete ao aproximar o valor* $\sqrt{x}$ *por* $\sqrt{\tilde{x}}$, *onde* $\tilde{x} = 0.4000 \times 10$ *é um valor aproximado de* x *com 3 algarismos significativos.*

Resulta de (1.21) que

$$|e_f| \leq \max_{t \in [\tilde{x}-\eta, \tilde{x}+\eta]} \frac{1}{2\sqrt{t}} \eta. \tag{1.22}$$

Como $\tilde{x}$ tem 3 algarismos significativos, $|\eta| \leq 0.5 \times 10^{-2}$ e então

$$|e_f| \leq \frac{1}{2\sqrt{3.995}}\ 0.5 \times 10^{-2} \simeq 0.12508 \times 10^{-2}.$$

Se por exemplo $x = 4.003579$, tem-se

$$\sqrt{x} - \sqrt{\tilde{x}} \simeq 0.89455 \times 10^{-3},$$

donde concluimos que, neste caso, a fórmula (1.22) dá uma estimativa do erro absoluto por excesso.

Consideremos agora o seguinte problema mais geral, de uma função de 2 variáveis. Suponhamos que se pretende calcular o valor de $f(x, y)$ usando os valores aproximados $\tilde{x}, \tilde{y}$ em vez de x, y, respectivamente. Supondo que f'_x e f'_y são contínuas, tem-se

$$f(x, y) = f(\tilde{x}, \tilde{y}) + (x - \tilde{x})f'_x(\xi_1, \xi_2) + (y - \tilde{y})f'_y(\xi_1, \xi_2),$$

onde $\xi_1 \in int(x, \tilde{x})$ e $\xi_2 \in int(y, \tilde{y})$. Se $|e_x| \leq \eta_x$ e $|e_y| \leq \eta_y$ e, por outro lado, $max|f'_x| \leq M_x$ e $max|f'_y| \leq M_y$ no intervalo $[\tilde{x} - \eta_x, \tilde{x} + \eta_x] \times [\tilde{y} - \eta_y, \tilde{y} + \eta_y]$, obtém-se a seguinte fórmula

$$|e_f| = |f(x, y) - f(\tilde{x}, \tilde{y})| \leq M_x\eta_x + M_y\eta_y. \tag{1.23}$$

A fórmula acima pode ser generalizada para funções de n variáveis.

1.4.2 Propagação dos erros ao efectuar uma operação aritmética

É de particular importância estudar o modo como um erro se propaga numa sequência de operações aritméticas. Fórmulas para os erros nas quatro operações aritméticas podem ser obtidas através de (1.23). Consideraremos aqui um processo alternativo para deduzir as mesmas.

Adição

Sejam $\tilde{x}$ e $\tilde{y}$ duas aproximações dos valores exactos x e y, com erros respectivos e_x e e_y. Tem-se

$$x + y = \tilde{x} + e_x + \tilde{y} + e_y = (\tilde{x} + \tilde{y}) + (e_x + e_y).$$

O erro na soma, que denotaremos por e_{x+y}, verifica

$$e_{x+y} = e_x + e_y. \tag{1.24}$$

Subtracção

De maneira análoga

$$e_{x-y} = e_x - e_y. \tag{1.25}$$

Multiplicação

Neste caso tem-se

$$\begin{aligned} x.y &= (\tilde{x} + e_x)\ (\tilde{y} + e_y) \\ &= \tilde{x}\tilde{y} + \tilde{x}e_y + \tilde{y}e_x + e_x e_y. \end{aligned}$$

Admitindo que o produto $e_x e_y$ é desprezável em relação aos outros termos, resulta

$$x.y \simeq \tilde{x}\tilde{y} + \tilde{x}e_y + \tilde{y}e_x,$$

donde

$$e_{x.y} \simeq \tilde{x}e_y + \tilde{y}e_x. \tag{1.26}$$

Divisão

Tem-se sucessivamente

$$\begin{aligned} x/y &= \frac{\tilde{x} + e_x}{\tilde{y} + e_y} \\ &= \frac{\tilde{x} + e_x}{\tilde{y}}(\frac{1}{1 + e_y/\tilde{y}}). \end{aligned}$$

Admitindo que $|e_y| < |\tilde{y}|$, podemos escrever

$$\frac{1}{1 + e_y/\tilde{y}} = 1 - e_y/\tilde{y} + (e_y/\tilde{y})^2 - (e_y/\tilde{y})^3 + \cdots$$

Substituindo acima e desprezando os termos que involvem potências de $|e_y/\tilde{y}|$ superiores a um, obtem-se a estimativa

$$\frac{x}{y} \simeq \frac{\tilde{x}}{\tilde{y}} + \frac{e_x}{\tilde{y}} - \frac{\tilde{x}}{(\tilde{y})^2}\ e_y.$$

Daqui resulta

$$e_{x/y} \simeq \frac{1}{\tilde{y}}\ e_x - \frac{\tilde{x}}{(\tilde{y})^2}\ e_y. \tag{1.27}$$

Note-se que as fórmulas (1.24)-(1.27) dão-nos uma indicação sobre o erro no resultado de cada uma das operações aritméticas, admitindo que *não houve arredondamento depois de efectuada a operação.*

Exercício
Determine o número de algarismos significativos que se pode garantir para $\tilde{x}+\tilde{y}$ como aproximação de $x+y$, sabendo que $\tilde{x}=0.425\times 10^3$ e $\tilde{y}=0.326\times 10^3$ têm os três algarismos significativos (despreze os erros de arredondamento).

Resulta de (1.24) e da definição de algarismo significativo (com $t=3$) que

$$\begin{aligned} |e_{x+y}| &\leq |e_x|+|e_y| \\ &\leq (0.5\times 10^{-3}+0.5\times 10^{-3})\times 10^3 \\ &\leq (0.1\times 10^{-2})\times 10^3 = 1. \end{aligned}$$

Concluimos poder garantir dois algarismos significativos para

$$\tilde{x}+\tilde{y}=0.751\times 10^3.$$

As fórmulas (1.24)-(1.27) dizem respeito à propagação dos erros nas quatro operações aritméticas. A partir destas obtêm-se as seguintes estimativas para as quantidades $\delta_{x\triangle y}=e_{x\triangle y}/(x\triangle y)$ ($\triangle$ designa uma operação aritmética).

$$\delta_{x+y} = \frac{e_{x+y}}{x+y} \simeq \frac{\tilde{x}}{\tilde{x}+\tilde{y}}\,\delta_x + \frac{\tilde{y}}{\tilde{x}+\tilde{y}}\,\delta_y \tag{1.28}$$

$$\delta_{x-y} = \frac{e_{x-y}}{x-y} \simeq \frac{\tilde{x}}{\tilde{x}-\tilde{y}}\,\delta_x - \frac{\tilde{y}}{\tilde{x}-\tilde{y}}\,\delta_y \tag{1.29}$$

$$\delta_{x.y} = \frac{e_{x.y}}{xy} \simeq \delta_x+\delta_y \tag{1.30}$$

$$\delta_{x/y} = \frac{e_{x/y}}{x/y} \simeq \delta_x-\delta_y. \tag{1.31}$$

Estas estimativas foram calculadas usando $\tilde{x}$ e $\tilde{y}$ em vez de x e y (ver Nota, parágrafo 1.2).

1.4.3 Propagação dos erros numa sequência de operações aritméticas

É importante compreender claramente o significado das fórmulas que deduzimos na secção anterior. Começamos com dois valores aproximados $\tilde{x}$ e $\tilde{y}$, que contêm erros e_x e e_y. Estes erros podem ser de qualquer dos três tipos considerados na secção 1.1. Nomeadamente, $\tilde{x}$ e $\tilde{y}$ podem ter sido obtidos experimentalmente, contendo por isso erros inerentes; podem ser o resultado de um processo de cálculo a que está associado um certo erro de truncatura; podem ainda ser o resultado de operações aritméticas anteriores e conter erros de arredondamento. Finalmente, $\tilde{x}$ e $\tilde{y}$ poderão conter uma combinação dos três tipos de erros.

As fórmulas (1.28)-(1.31) dão-nos uma indicação sobre o erro no resultado de cada uma das operações aritméticas, como funções de $\tilde{x}$ e $\tilde{y}$, δ_x e δ_y, admitindo que *não houve arredonda-*

mento depois de efectuada a operação. Se no entanto quisermos saber como o erro deste resultado se propaga em ainda outras operações aritméticas subsequentes, teremos de adicionar o *erro de arredondamento* explicitamente.

Exercício
Seja $v = xy + z$ e sejam os números de vírgula flutuante $\tilde{x}$, $\tilde{y}$ e $\tilde{z}$ valores aproximados de x, y e z, com erros relativos $|\delta_x|$, $|\delta_y|$ e $|\delta_z|$, respectivamente. Determine uma estimativa do erro relativo no cálculo de v, num computador com unidade de arredondamento U, e usando os valores aproximados.

O cálculo de v pode ser efectuado usando o seguinte algoritmo.

$$\begin{aligned} u &= xy \\ v &= u + z. \end{aligned} \tag{1.32}$$

O computador começa por efectuar a multiplicação de $\tilde{x}$ por $\tilde{y}$.

Em seguida, o resultado é normalizado. Ou seja, aquilo que se obtém é um valor $\tilde{u}$ dado por

$$\tilde{u} = fl(\tilde{x}\tilde{y}).$$

Usando (1.30), tem-se a seguinte estimativa

$$\delta_u \simeq \delta_x + \delta_y + \delta_{arr1}, \tag{1.33}$$

onde $|\delta_{arr1}|$ é o *erro relativo de arredondamento* correspondente à normalização do produto. Em seguida, é efectuada a adição de $\tilde{u}$ com $\tilde{z}$ e o resultado é normalizado. De modo análogo, resulta de (1.28) que

$$\delta_v \simeq \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\ \delta_u + \frac{\tilde{z}}{\tilde{u} + \tilde{z}}\ \delta_z + \delta_{arr2}, \tag{1.34}$$

onde $|\delta_{arr2}|$ é o *erro relativo de arredondamento* associado à normalização da soma.

Conjugando (33) e (34), obtém-se

$$\delta_{xy+z} \simeq \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_x + \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_y + \frac{\tilde{z}}{\tilde{u} + \tilde{z}}\delta_z + \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_{arr1} + \delta_{arr2},$$

com $|\delta_{arr1}| \leq U$ e $|\delta_{arr2}| \leq U$. Da equação acima pode obter-se uma estimativa para $|\delta_{xy+z}|$. Por simplicidade de escrita, costuma-se omitir o tile. Este procedimento é usual, sempre que for claro, pelo sentido do texto, quando nos referimos a um valor exacto ou a uma sua aproximação.

1.5 Condicionamento e estabilidade

1.5.1 Condicionamento

Devido a arredondamentos, medições inexactas, etc, os dados e parâmetros de um problema estão em geral afectados de erros. É de todo o interesse saber em que medida esses erros podem afectar a solução do problema.

O *condicionamento* de um problema diz respeito à sua sensibilidade a pequenas variações nos dados.

Exemplo 1.5 *Consideremos o sistema de equações lineares*

$$\begin{array}{ll} \frac{1}{20}\ x_1 & +\frac{1}{21}\ x_2 = \frac{41}{420} \\ \frac{1}{21}\ x_1 & +\frac{1}{22}\ x_2 = \frac{43}{462} \end{array} \tag{1.35}$$

com solução exacta $x_1 = x_2 = 1$.

Tem-se

$$\frac{1}{20} = 0.05 \quad \frac{1}{21} = 0.0476190... \quad \frac{1}{22} = 0.0454545...$$

e

$$\frac{41}{420} = 0.09761... \quad \frac{43}{462} = 0.0930736...$$

Consideremos ainda o sistema

$$\begin{array}{ll} 0.0500\ \tilde{x}_1 & +0.0476\ \tilde{x}_2 = 0.0976 \\ 0.0476\ \tilde{x}_1 & +0.0454\ \tilde{x}_2 = 0.0931 \end{array} \tag{1.36}$$

Verificamos que os coeficientes e os termos independentes do sistema (1.36) são aproximações dos correspondentes coeficientes e termos independentes do sistema (3.9) com erros absolutos não excedendo 0.5×10^{-4} e erros relativos não excedendo 0.5×10^{-3}. Ou seja, houve pequenas variações nos dados do problema (3.9).

Contudo, a solução exacta do sistema (1.36) é $\tilde{x}_1 = -0.1226415...$, $\tilde{x}_2 = 2.1792452...$

Calculando os erros relativos de $\tilde{x}_1, \tilde{x}_2$, obtém-se

$$|\delta_{x_1}| = \frac{|x_1 - \tilde{x}_1|}{|x_1|} \simeq 1.123 \Rightarrow 112\% \text{ de erro}$$

e

$$|\delta_{x_2}| = \frac{|x_2 - \tilde{x}_2|}{|x_2|} \simeq 1.18 \Rightarrow 118\% \text{ de erro}$$

Ou seja, uma pequena mudança relativa nos dados correspondeu uma grande mudança relativa na solução.

Um problema cuja solução é muito sensível a variações ou mudanças nos dados e parâmetros diz-se *mal condicionado*. Um problema diz-se *bem condicionado* se pequenas variações nos dados e parâmetros produzem sempre pequenas variações na solução.

Temos então que o sistema (3.9) é um sistema *mal condicionado*.

Exemplo 1.6 *Consideremos o polinómio*

$$p(x) = (x-1)(x-2)(x-3)\cdots(x-19)(x-20) \tag{1.37}$$

O polinómio p tem por raizes $1, 2, 3, \ldots, 20$ e é da forma

$$p(x) = x^{20} + a_{19}\, x^{19} + a_{18}\, x^{18} + \cdots + a_0.$$

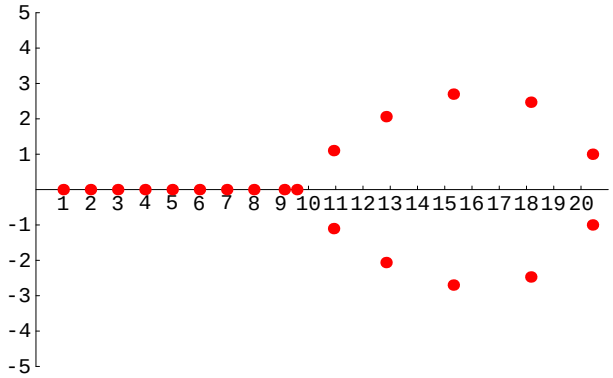
Suponhamos que o coeficiente do termo em x^{19} é afectado por uma pequena mudança 10^{-7} e seja q o polinómio resultante, ou seja,

$$\begin{aligned} q(x) &= x^{20} + (a_{19} + 10^{-7})\, x^{19} + a_{18}\, x^{18} + \cdots + a_0 \\ &= p(x) + (10^{-7}) * x^{19} \end{aligned} \tag{1.38}$$

Para se obterem as raizes de q utilizou-se o Mathematica (comando NSolve[q[x]==0,x]). Obtiveram-se 10 raizes reais para q (1, 2, 3, 4,5, 5.99999, 7.00026, 7.9941, 9.11225, 9.57054) e 5 pares de raizes complexas conjugadas, começando com o par:

$10.9213 + 1.10237\, i$, $\quad 10.9213 - 1.10237\, i$, $\qquad$ até:

$20.422 + 0.999201\, i$, $\; 20.422 - 0.999201\, i$, como mostra o gráfico seguinte. Conclui-se que as raizes de p são extremamente sensíveis a variações nos coeficientes.



Raizes do polinómio q; nas complexas representou-se a parte real e imaginária

Número de condição

Para alguns problemas é possível definir um "número de condição", que se utiliza como indicador de mau-condicionamento.

Suponhamos que pretendemos calcular o valor de uma função f num ponto x. Põe-se a questão: *se x for ligeiramente perturbado, qual o efeito em $f(x)$?*

Se $f'(x)$ existe e é contínua, tem-se, pelo Teorema do valor médio de Lagrange

$$f(x+h) - f(x) = h\ f'(\xi) \approx h\ f'(x) \tag{1.39}$$

com $\quad \xi \in I = (min\{x, x+h\}, max\{x, x+h\})$.

Designando $\tilde{x} = x + h$, vem

$$\frac{f(\tilde{x}) - f(x)}{f(x)} \approx \frac{(\tilde{x} - x)\ f'(x)}{f(x)} = \left[\frac{x\ f'(x)}{f(x)}\right]\left(\frac{\tilde{x} - x}{x}\right) \tag{1.40}$$

Ou seja

$$\frac{f(\tilde{x}) - f(x)}{f(x)} \approx \left[\frac{x\ f'(x)}{f(x)}\right]\left(\frac{\tilde{x} - x}{x}\right) \tag{1.41}$$

Tomando módulos:

$$\left|\frac{f(\tilde{x}) - f(x)}{f(x)}\right| \approx \left|\frac{x\ f'(x)}{f(x)}\right| \left|\frac{\tilde{x} - x}{x}\right|, \tag{1.42}$$

ou seja, em termos dos erros relativos de $f(\tilde{x})$ e de $\tilde{x}$

$$|\delta_{f(x)}| \approx cond_f\, |\delta_x|. \tag{1.43}$$

Acima designámos por $cond_f$ o valor $|x\ f'(x)/f(x)|$, a que se chama número de condição. A fórmula (1.43) permite relacionar o erro relativo de $f(\tilde{x})$ com o erro relativo de $\tilde{x}$. Se $cond_f = 1$ então o erro relativo de $f(\tilde{x})$ é igual ao de x. No caso dum valor superior a 1, o erro relativo de $f(\tilde{x})$ vem maior do o de $\tilde{x}$. Se para certos valores de x o número de condição é muito grande ($\approx$ infinito) temos uma situação de mau condicionamento.

Exemplo 1.7 *Qual o número de condição para a função $f(x) = arcsin(x)$?*

Tem-se

$$\frac{x\ f'(x)}{f(x)} = \frac{x}{\sqrt{1-x^2}\,(arcsin(x))}$$

Para valores de x perto de 1, $\quad arcsin(x) \approx \pi/2$. Então

$$x/(\sqrt{1-x^2}\,(arcsin(x))) \approx 2x/(\pi)\sqrt{1-x^2}$$

e, assim, o número de condição torna-se infinito à medida que x se aproxima de 1. Podemos concluir que pequenos erros relativos em x levam a grandes erros relativos em $arcsin(x)$, para valores de x perto de 1.

1.5.2 Estabilidade de um algoritmo

O conceito de estabilidade diz respeito à propagação dos erros de arredondamento no computador. Um algoritmo diz-se *instável* se a acumulação de erros de arredondamento afecta significativamente o resultado final. Um algoritmo *estável* produz sempre bons resultados (com problemas bem condicionados).

Exemplo 1.8 *Consideremos a sucessão*

$$x_n = \left(\frac{1}{3}\right)^n, \quad n > 0 \tag{1.44}$$

Algoritmo instável

Uma maneira de gerar os termos da sucessão (1.44) é usar a seguinte fórmula de recorrência

$$\begin{cases} x_0 = 1, \qquad x_1 = \dfrac{1}{3} \\ x_{n+1} = \dfrac{10}{3}x_n - x_{n-1}, \qquad n \geq 1. \end{cases} \tag{1.45}$$

Verifica-se facilmente, por indução, que esta relação de recorrência gera a sucessão (1.44). Com efeito, para $n = 0$ e $n = 1$ (1.44) verifica-se. Supondo que (1.44) é válida para $n \leq m$, tem-se

$$\begin{aligned} x_{m+1} &= \frac{10}{3}x_m - x_{m-1} = \frac{10}{3}\left(\frac{1}{3}\right)^m - \left(\frac{1}{3}\right)^{m-1} \\ &= \left(\frac{1}{3}\right)^{m-1}\left[\frac{10}{9} - 1\right] = \left(\frac{1}{3}\right)^{m+1} \end{aligned}$$

A fórmula (1.45) foi implementada em Mathematica e, usando aritmética de 5 e 16 dígitos, respectivamente, produziu os resultados da segunda e terceira colunas da tabela seguinte. Na quarta coluna estão os valores exactos dos termos da sucessão. Os valores estão apresentados com 5 dígitos, depois de arredondamento simétrico.

n	Algor. (1.45)(**5** dígitos)	Algor. (1.45)(**16** dígitos)	**valores exactos**
0	0.10000×10^{1}	0.10000×10^{1}	$x_0 = 0.10000 \times 10^{1}$
1	0.33333×10^{0}	0.33333×10^{0}	$x_1 = 0.33333 \times 10^{0}$
2	0.11110×10^{0}	0.11111×10^{0}	$x_2 = 0.11111 \times 10^{0}$
3	0.37000×10^{-1}	0.37037×10^{-1}	$x_3 = 0.37037 \times 10^{-1}$
4	0.12230×10^{-1}	0.12346×10^{-1}	$x_4 = 0.12346 \times 10^{-1}$
5	0.37660×10^{-2}	0.41152×10^{-2}	$x_5 = 0.41152 \times 10^{-2}$
6	0.32300×10^{-3}	0.13717×10^{-2}	$x_6 = 0.13717 \times 10^{-2}$
7	-0.26893×10^{-2}	0.45725×10^{-3}	$x_7 = 0.45725 \times 10^{-3}$
8	-0.92872×10^{-2}	0.15242×10^{-3}	$x_8 = 0.15242 \times 10^{-3}$
9	-0.28268×10^{-1}	0.50805×10^{-4}	$x_9 = 0.50805 \times 10^{-4}$
10	-0.84939×10^{-1}	0.16935×10^{-4}	$x_{10} = 0.16935 \times 10^{-4}$
	$\vdots$	$\vdots$	$\vdots$
18	-0.5573×10^{3}	-0.54587×10^{-8}	$x_{18} = 0.25812 \times 10^{-8}$
	$\vdots$	$\vdots$	$\vdots$
30	-0.29611×10^{9}	-0.42727×10^{-2}	$x_{30} = 0.48569 \times 10^{-14}$
$\vdots$	$\vdots$	$\vdots$	

Note que de (1.44) se tem $\lim_{n\to\infty} x_n = 0$, sendo isso ilustrado na coluna 4 da tabela. Tal não acontece com os valores obtidos com a fórmula de recorrência (1.45), independentemente do número de dígitos usado nos cálculos (ver colunas 2 e 3 da tabela).

Verificamos que x_1 foi arredondado com erro $|\delta_{x_1}| \leq 0.5 \times 10^{-4}$. Esse erro propagou-se aos valores seguintes, alterando por completo os resultados. Por exemplo, com 5 dígitos, em vez do valor $x_8 = 0.15242 \times 10^{-3}$ obteve-se -0.92872×10^{-2}, o que apresenta um erro absoluto de 0.00943962 e um erro relativo de 59.9, ou seja, cerca de 6000% de erro. Usando maior precisão não resolve o problema, só que o efeito da acumulação de erros aparece mais à frente. Por exemplo, comparando as colunas 3 e 4, vemos que com 16 dígitos os resultados são similares até mais tarde do que no caso de se usarem 5 dígitos . Mas já em vez de $x_{18} = 0.25812 \times 10^{-8}$, obteve-se -0.54587×10^{-8}, a que corresponde um erro relativo de cerca de 3, ou seja, 300% de erro. Prosseguindo, as coisas pioram e o valor -0.42727×10^{-2} é uma aproximação para x_{30} com erro relativo de $\simeq 8.8 \times 10^{11}$.

Pode-se dizer que (1.45), depois de implementado, traduziu-se num algoritmo instável.

Algoritmo estável

Uma outra maneira de gerar os termos x_n de (1.44) é através da fórmula

$$x_0 = 1, \quad x_n = (1/3)x_{n-1}, \quad n > 1 \tag{1.46}$$

Com base nesta fórmula obtém-se um método estável.

Cancelamento subtractivo

O chamado cancelamento subtractivo ocorre quando se subtraem dois números muito próximos e traduz-se num aumento do erro relativo do resultado, o qual se poderá propagar em cálculos posteriores, dando origem a instabilidade.

Exemplo 1.9

Sejam os números $x = 0.3721478693$ e $y = 0.3720230572$ e sejam $fl(x) = 0.37215$ e $fl(y) = 0.37202$, respectivamente, os seus representantes num sistema com 5 dígitos e arredondamento simétrico, com erros relativos dados por $|\delta_x| \simeq 0.6 \times 10^{-7}$ e $|\delta_y| \simeq 0.8 \times 10^{-7}$. Considere-se a diferença $fl(x) - fl(y)$. Tem-se

$$\tilde{z} = fl(x) - fl(y) = 0.00013 = 0.13000 \times 10^{-3}$$

Sendo o valor exacto da diferença $z = x - y = 0.000124812$, vem que o erro relativo de $\tilde{z}$ é

$$|\delta_z| = \frac{|(x-y) - (fl(x) - fl(y))|}{|x-y|} \simeq 0.044$$

o que representa um aumento sgnificativo do erro relativo em relação aos erros das parcelas.

Quando presente num algoritmo, o cancelamento subtractivo pode causar instabilidade numérica.

Exemplo 1.10

Consideremos as fórmulas, matematicamente equivalentes.

$$f(x) = \frac{x}{1 - \sqrt{1-x}} \tag{1.47}$$

$$f(x) = 1 + \sqrt{1-x} \tag{1.48}$$

Obtiveram-se os resultados seguintes para $f(x)$, com x próximo de zero, tendo-se usado um programa em Mthematica e aritmética de 16 dígitos.

x	fórmula(1.47)	fórmula(1.48)
$x = 10^{-11}$	2	2
$x = 10^{-12}$	1.99982	2
$x = 10^{-13}$	1.99274	2
$x = 10^{-14}$	1.95809	2
$x = 10^{-15}$	1.5012	2
$x = 10^{-16}$	0.45036	2

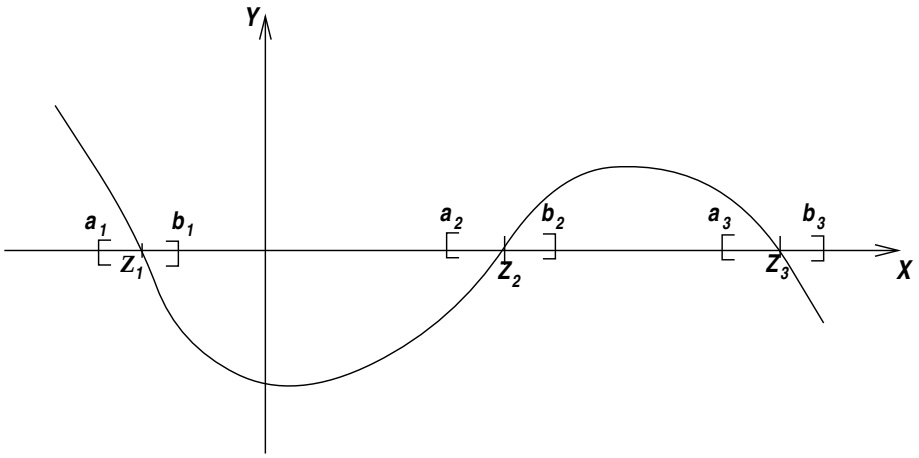
A coluna 3 da tabela mostra que (1.48) dá o valor esperado. A coluna 2 da tabela indica instabilidade quando se usa a fórmula (1.47) para se obter o valor $f(x)$, com x próximo de zero. Isso deve-se ao facto de haver cancelamento subtractivo causado pela subtracção de 2 números próximos de 1, no denominador da fracção; com efeito, tem-se $x \approx 0 \rightarrow \sqrt{1-x} \approx 1$.

2 Resolução de equações não lineares

Dada uma função f definida num intervalo $[a, b]$, pretendemos determinar as *raízes da equação* $f(x) = 0$ (ou os *zeros da função* f), isto é, os valores z tais que $f(z) = 0$. Neste capítulo estudaremos métodos para a obtenção de valores aproximados de z.

2.1 Localização e separação das raízes

Primeiramente há que determinar, para cada raiz, um intervalo que a contenha. Isso pode ser feito recorrendo ao estudo gráfico e teórico de f. Exemplificamos em seguida.



Análise gráfica

Na análise gráfica pode-se esboçar um gráfico de f ou, a partir da equação original $f(x) = 0$, obter uma equação equivalente, para a qual seja mais fácil fazer o gráfico.

Exemplo 2.1 *Pretende-se uma aproximação da raiz positiva da equação*

$$f(x) = \left(\frac{x}{2}\right)^2 - \sin x = 0. \tag{2.1}$$

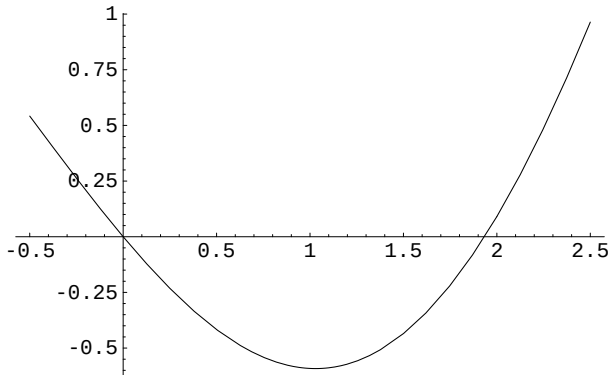


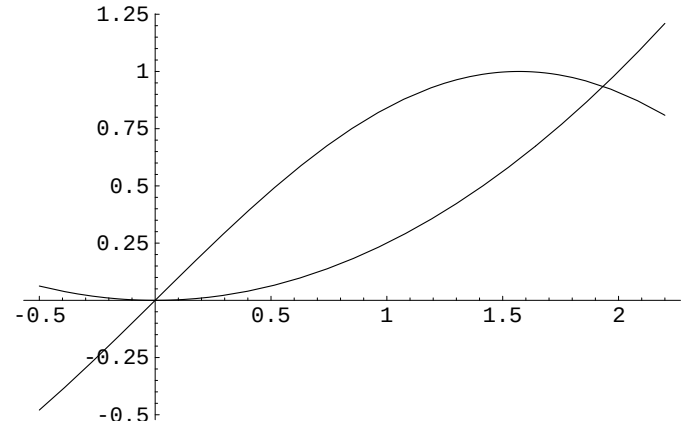
Gráfico da função $f(x) = (x/2)^2 - \sin x$

Neste caso é possível ver pelo gráfico de f que a raiz positiva de $f(x) = 0$ está no intervalo $[\pi/2, 2.0]$. Na verdade, ela parece próxima de 1.9.

Exemplo 2.2 *Uma outra maneira de proceder é reescrever a equação numa forma equivalente. Tem-se*

$$\left(\frac{x}{2}\right)^2 - \sin x = 0 \Longleftrightarrow \left(\frac{x}{2}\right)^2 = \sin x \Longleftrightarrow \begin{cases} y = (x/2)^2 \\ y = \sin x \end{cases}$$

O gráfico sugere que a equação $f(x) = 0$ tem uma raiz no intervalo $[\pi/2, 2.0]$; corresponde à projecção, no eixo dos xx, da intersecção dos gráficos das funções $y = (x/2)^2$ e $y = \sin x$.



Gráficos das funções $y = (x/2)^2$ e $y = \sin x$

Análise teórica

Na análise teórica, são usados os seguintes Teoremas, que são corolários, respectivamente, do Teorema do valor intermédio para funções e do Teorema de Rolle.

Teorema 2.1 *Seja* $f(x) \in C[a,b]$. *Se* $f(a) \times f(b) < 0$ *então existe pelo menos um* $z \in]a,b[$ *tal que* $f(z) = 0$.

Teorema 2.2 *Seja* $f(x) \in C[a,b]$. *Se* $f'(x)$ *existe e tem sinal constante em* $]a,b[$ *então* f *não pode ter mais de um zero em* $]a,b[$.

Para a função $f(x) = (x/2)^2 - \sin x$ do exemplo 2.1, tem-se $f(\pi/2) = (\pi^2)/16 - 1 < 0$ e $f(2) = 1 - \sin 2 > 0$. Logo podemos concluir que f tem pelo menos um zero no intervalo I=$[\pi/2, 2]$. Além disso, tem-se $f'(x) = x/2 - \cos x$ e $f''(x) = 1/2 + \sin x > 0$ em I. Logo $f'(x)$ é monótona crescente em I e, como $f'(\pi/2) = \pi/4 > 0$, então f' não se anula no intervalo I. Conclui-se que a raiz da equação nesse intervalo é única.

2.2 Métodos iterativos para equações não lineares

Escolhida uma aproximação "grosseira" para a raiz z, contida no intervalo obtido na secção 2.1, pretende-se melhorar essa aproximação utilizando um método iterativo.

Dada a aproximação inicial $x_0 \in [a,b]$, um método iterativo (convergente) permite obter sucessivas aproximações para z:

$$x_1, x_2, x_3, \cdots, x_m, \cdots$$

onde $\lim_{m\to\infty} x_m = z$. O processo pára quando uma certa precisão for atingida; são exemplos de critérios de paragem:

1. $|x_m - z| \leq \epsilon$, ϵ dado (i)

ou

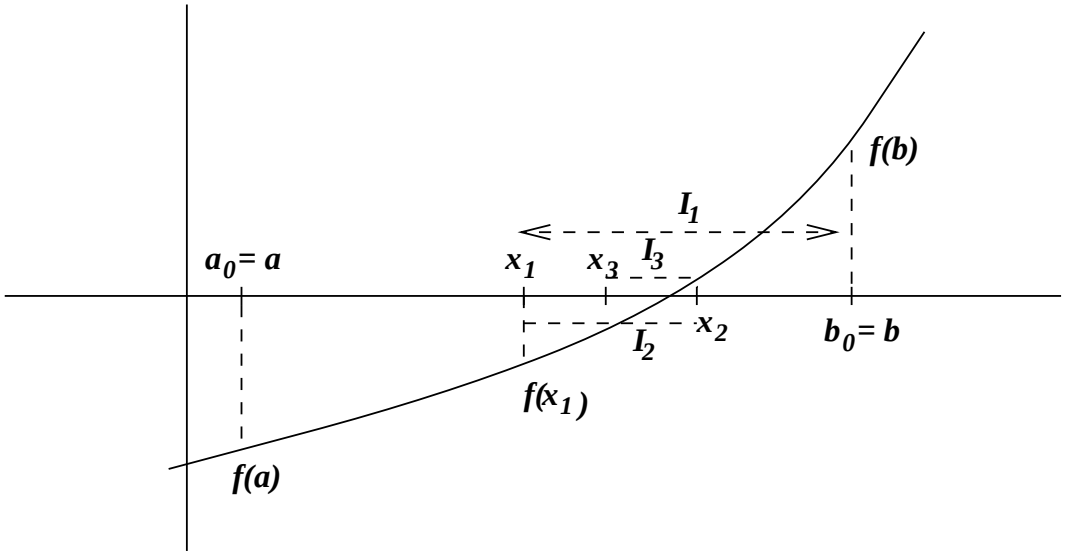
2. $|f(x_m)| \leq \epsilon$ (ii)

Observação É importante notar que os dois critérios acima não são equivalentes, ou seja, que nem sempre é possível satisfazer (i) e (ii) simultâneamente.

2.2.1 Método da bissecção

O método da bissecção é um dos métodos numéricos mais antigos, permitindo uma interpretação geométrica simples.

Seja $f(x)$ contínua em $[a,\ b]$ com $f(a)f(b) < 0$ e suponhamos que z é a única raiz em $[a,\ b]$.



Vai-se construir uma sucessão de intervalos I_k contendo a raiz z, de modo que $I_0 \supset I_1 \supset I_2 \supset I_3 \supset \ldots\ldots \supset I_k \supset \ldots\ldots$ e tomam-se para sucessivas aproximações de z os pontos médios desses intervalos.

Seja $I_0 = [a_0, b_0] = [a,\ b]$ o intervalo dado e seja $x_0 = a$ uma aproximação inicial de z.

Toma-se para aproximação seguinte o ponto médio do intervalo I_0, isto é, faz-se $x_1 = (b+a)/2$. Em seguida analisa-se o sinal da função f nos pontos a, x_1, b para decidir de que lado de x_1 se encontra z. Suponhamos, como é o caso da figura, que $f(x_1) \times f(a) > 0$ e que $f(x_1) \times f(b) < 0$. Então $z \in [x_1, b]$ e define-se $I_1 = [a_1, b_1] = [x_1, b]$. O comprimento de I_1 é agora $(b-a)/2$ e o seu ponto médio é $x_2 = (b_1 + a_1)/2$. Procedendo como para I_0, suponhamos que $f(a_1) \times f(x_2) < 0$ e que $f(x_2) \times f(b_1) > 0$. Então $z \in I_2 = [a_2, b_2] = [a_1, x_2,]$. E o processo continua. Abaixo ilustramos esquematicamente estas ideias.

$$
\begin{aligned}
&I_0 = [a,\ b], && |I_0| = b - a \\
&x_1 = \frac{b+a}{2}, && |I_1| = \frac{b-a}{2} \\
&x_2 = \frac{b_1 + a_1}{2} && |I_2| = \frac{b-a}{2^2} \\
&x_3 = \frac{b_2 + a_2}{2} && |I_3| = \frac{b-a}{2^3} \\
&\quad \vdots \quad \vdots \\
&x_k = \frac{b_{k-1} + a_{k-1}}{2} && |I_k| = \frac{b-a}{2^k}
\end{aligned}
$$

Obtiveram-se intervalos, contendo z, de amplitude cada vez mais pequena. São válidos os seguintes majorantes para os erros das sucessivas aproximações de z.

$$z,\ x_1 \in I_0 \Longrightarrow |e_{x_1}| \ = \ |z - x_1| \leq (b-a)/2$$

$$z,\ x_2 \in I_1 \Longrightarrow |e_{x_2}| \ = \ |z - x_2| \leq (b-a)/2^2$$

$$z,\ x_3 \in I_2 \Longrightarrow |e_{x_3}| \ = \ |z - x_3| \leq (b-a)/2^3$$

$$\vdots$$

de modo geral

$$z,\ x_k \in I_{k-1} \Longrightarrow |e_{x_k}| \ = \ |z - x_k| \leq (b-a)/2^k$$

Da relação

$$0 \leq |e_{x_k}| = |z - x_k| \leq \frac{b-a}{2^k} \quad \longrightarrow \quad 0$$

concluímos que

$x_1, x_2, x_3, \ldots, x_k \ldots \overset{k\to\infty}{\longrightarrow} z$. Temos assim o seguinte resultado.

Teorema 2.3 *Seja $f(x)$ contínua em $[a,\ b]$ com $f(a)f(b) < 0$ e seja z a única raiz de f em $[a,\ b]$. Então o método da bissecção converge e tem-se a seguinte estimativa de erro*

$$|e_{x_k}| = |z - x_k| \leq \frac{b-a}{2^k} \tag{2.2}$$

Qual o número de iterações necessárias para se obter a precisão desejada ?

Dado um número ϵ é possível estimar o número de iterações k de modo a garantir que $|z - x_k| < \epsilon$

Como $|z - x_k| \leq \dfrac{b-a}{2^k}$ basta impor $\dfrac{b-a}{2^k} < \epsilon$, ou seja, $b - a < 2^k\epsilon \iff \dfrac{b-a}{\epsilon} < 2^k \iff \ln\left(\dfrac{b-a}{\epsilon}\right) < k \ln 2 \iff k > \dfrac{\ln\left(\dfrac{b-a}{\epsilon}\right)}{\ln 2}$

Tomando para k o inteiro imediatamente a seguir ao valor acima, temos a garantia de que x_k satisfaz a precisão desejada.

Exemplo 2.3

Finalizamos esta secção com uma tabela que mostra os resultados numéricos obtidos pelo método da bissecção aplicado à equação

$$f(x) = \left(\frac{x}{2}\right)^2 - \sin x = 0, \quad x \in [1.5, 2].$$

Começando com $I_0 = [1.5, 2]$, foram obtidos os intervalos $I_{k-1} = [a_{k-1}, b_{k-1}]$, cujo ponto médio é $x_k = (b_k + a_k)/2$.

Tabela 1: Aproximações x_k para a raiz z, pelo método da bissecção

k	a_{k-1}	b_{k-1}	x_k	$f(x_k)$
1	1.5	2	1.75	< 0
2	1.75	2	1.875	< 0
3	1.875	2	1.9375	> 0
4	1.875	1.9375	1.90625	< 0
5	1.90625	1.9375	1.921875	

2.2.2 Método do ponto fixo

Começaremos por tratar equações da forma

$$g(x) = x, \tag{2.3}$$

já que, como veremos, os *métodos do ponto fixo* foram concebidos para equações deste tipo especial. Os resultados obtidos poderão depois ser aplicados a qualquer equação geral $f(x) = 0$, depois de esta ser reescrita na forma $g(x) = x$. Notemos que a construção de tal $g(x)$ é sempre possível mas não é única. Vejamos exemplos.

Exemplo 2.4

A equação

$$x^3 - 3\,x + 1 = 0. \tag{2.4}$$

é equivalente a

$$i)\quad \underbrace{\frac{1 + x^3}{3}}_{g(x)} = x$$

Atendendo a que

$$x^3 - 3\,x + 1 = 0 \Longleftrightarrow x(x^2 - 3) + 1 = 0,$$

podemos ainda reescrever (2.4) na forma

$$ii)\quad \underbrace{\frac{1}{3 - x^2}}_{g(x)} = x$$

Exemplo 2.5

A equação $x + \ln x = 0$ pode ser reescrita como:

$$i)\quad \underbrace{-\ln x}_{g(x)} = x \quad \text{ou:} \quad ii)\quad \underbrace{e^{-x}}_{g(x)} = x.$$

Exemplo 2.6

A equação $f(x) = 0$ é equivalente a:

$$i)\ \underbrace{x + f(x)}_{g(x)} = x \quad \text{e ainda, de modo geral, } a: \quad ii)\ \underbrace{x + A(x)\, f(x)}_{g(x)} = x,$$

onde $A(x)$ é uma função que nunca se anula.

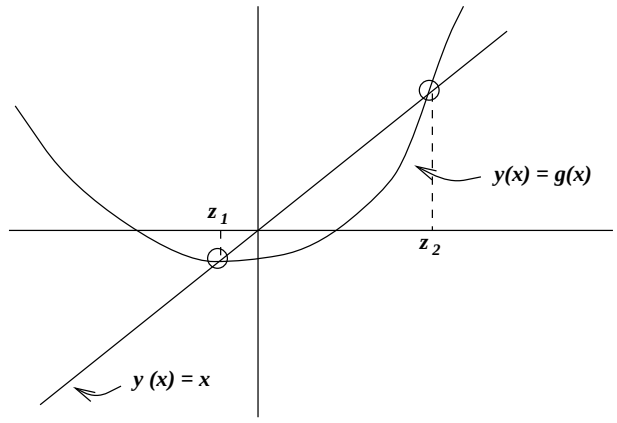
Definição 2.1 *: As raízes de uma equação da forma $g(x) = x$ chamam-se* **pontos fixos** *da função g. Ou seja, z é ponto fixo de g se o valor de g em z for igual ao próprio z.*

Interpretação geométrica

Notemos que se tem

$$x = g(x) \Longleftrightarrow \begin{cases} y = g(x) \\ y = x \end{cases}$$

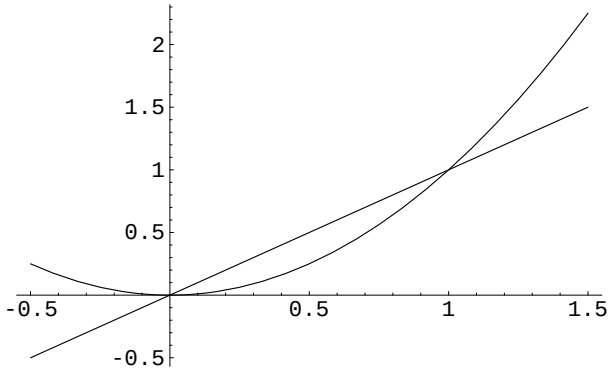
Ou seja, determinar os pontos fixos de g consiste em determinar as projecções, sobre o eixo dos xx, dos pontos de intersecção dos gráficos de $y_1 = g(x)$ e $y_2 = x$.



z_1, z_2 *pontos fixos de* g

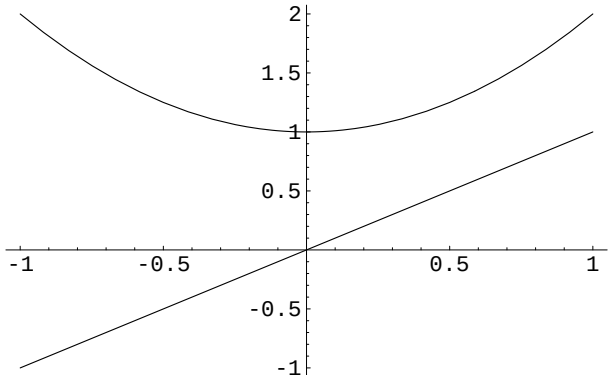
Exemplo 2.7

Seja $g(x) = x^2$. Então os pontos fixos de g são as soluções da equação $x^2 = x$. Trata-se dos valores $x = 0$ e $x = 1$.



Exemplo 2.8

A função $g(x) = x^2 + 1$ não tem pontos fixos, já que a equação $x^2 + 1 = x$ não tem soluções reais.

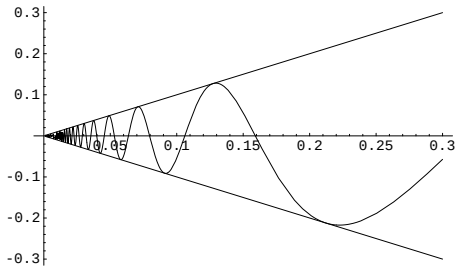


Exemplo 2.9

No caso da função $g(x) = x$, todo o número real é ponto fixo de g. Ou seja, o conjunto dos pontos fixos de g é $\mathcal{R}$.

Exemplo 2.10

Finalmente seja a função $g(x) = x\ Sen(1/x)$, a que corresponde o gráfico abaixo.



A função g tem múltiplos pontos fixos

Voltemos ao principal objectivo desta secção: obter aproximações para as raízes duma equação genérica $f(x) = 0$, depois de a reescrevermos na forma equivalente $x = g(x)$. Como foi dito, as raízes desta última equação são os pontos fixos de g.

Como aproximar um ponto fixo, z, duma função g ? Vamos utilizar um método construído com o auxílio da própria g, que introduzimos a seguir.

Definição 2.2

Dado um certo valor x_0, consideremos a sucessão dada pelo seguinte algoritmo:

$$\begin{aligned} x_1 &= g(x_0) \\ x_2 &= g(x_1) \\ x_3 &= g(x_2) \\ &\vdots \\ x_{m+1} &= g(x_m) \\ &\vdots \end{aligned}$$

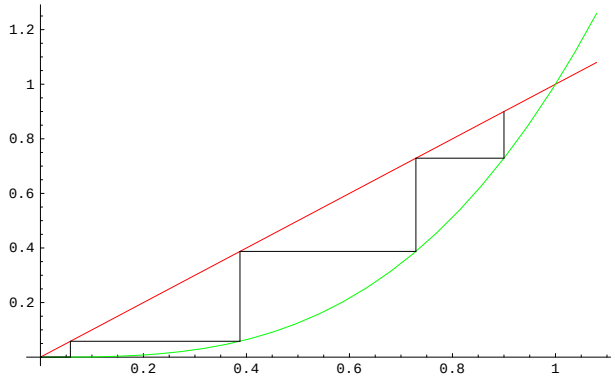
À sucessão assim definida chama-se **sucessão gerada por** g ou **método do ponto fixo associado a** g (ou ainda **iteração do ponto fixo**). Aos termos da sucessão chamamos **iteradas**.

Exemplo 2.11 *Consideremos a sucessão do ponto fixo gerada por* $g(x) = x^3$*, com várias escolhas de* x_0*, e observemos o seu limite.*

- Seja $x_0 = 0.9$. Obtém-se

$$x_1 = g(x_0) = (x_0)^3 = (0.9)^3$$
$$x_2 = g(x_1) = (x_1)^3 = ((0.9)^3)^3 = (0.9)^9$$
$$x_3 = g(x_2) = (x_2)^3 = ((0.9)^9)^3 = (0.9)^{27}$$
$$x_4 = g(x_3) = (x_3)^3 = ((0.9)^{27})^3 = (0.9)^{81}$$
$$\ldots$$

  Obteve-se uma sucessão $x_0 = 0.9, \quad x_{m+1} = (x_m)^3, \quad m = 0, 1, 2...$, que converge para zero. Tem-se a seguinte ilustração gráfica do processo.



Partindo de $x_0 = 0.9$, obtiveram-se, com o auxílio do Mathematica, algumas iteradas:

$$x_0 = 0.9, \quad x_1 = 0.729, \quad x_2 = 0.387420489, \quad x_3 = 0.0581497,$$
$$x_4 = 0.000196627, \quad x_5 = 7.60203 \times 10^{-12}, \quad x_6 = 4.39329 \times 10^{-34},$$
$$x_7 = 8.47946 \times 10^{-101}, \quad x_8 = 6.09683 \times 10^{-301},$$
$$x_9 = 2.26628 \times 10^{-901}, \quad x_{10} = 1.16396 \times 10^{-2702}, ...$$

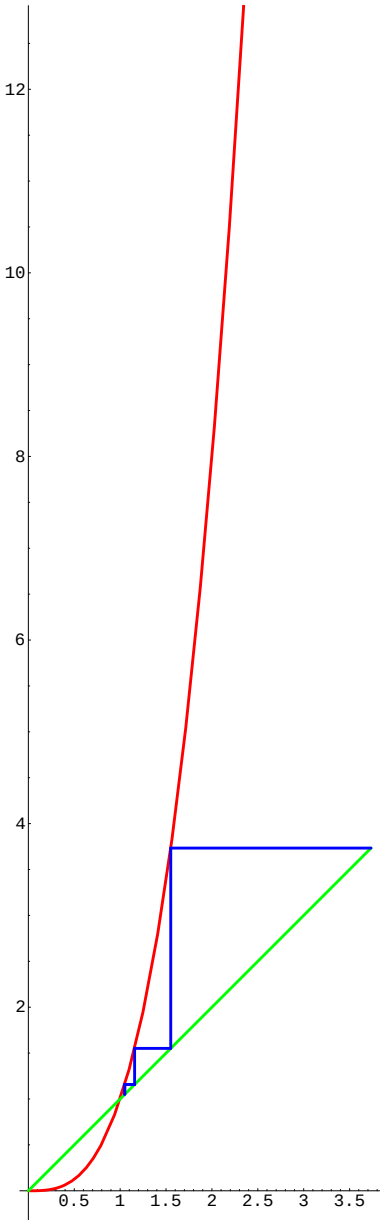
Seja ainda $g(x) = x^3$ mas com a escolha:

- $x_0 = 1$. Então, vem $x_1 = g(x_0) = (1)^3 = 1$, $x_2 = 1$,etc. É simples de constatar que, neste caso, se obtém a sucessão $x_m = 1, m = 0, 1, ...$, cujo limite é 1.

- E se $x_0 = -1$? Facilmente verificamos que vem $x_1 = (-1)^3 = -1$ e, de modo geral, obtém-se $x_m = -1, m = 0, 1, ...$, cujo limite é -1.

- Considere ainda $x_0 = 1.05$. Note que neste caso se obtém
  $x_0 = 1.05, \; x_1 = 1.15763, \quad x_2 = 1.55133, \quad x_3 = 3.73346, \ldots \; x_7 = 2.19279 \times 10^{46}, \quad x_8 = 1.05437 \times 10^{139}$,..
  A sucessão tende para infinito!

Segue-se a ilustração gráfica.



Nos 3 casos de sucessões convergentes, os limites foram: 0, -1 e 1. Esses valores o que representam para g? O teorema a seguir ajuda à resposta.

Provemos agora o resultado seguinte.

Teorema 2.4 *Se uma sucessão gerada por uma função contínua é convergente, o seu limite é um ponto fixo de g.*

Dem Seja $g(x) \in C[a,b]$ e a sucessão associada :

$$x_{m+1} = g(x_m), \quad m = 0, 1, 2, ... \tag{2.5}$$

Suponhamos que existe o limite finito

$$z = \lim_{m\to\infty} x_{m+1}.$$

Então

$$z = \lim_{m\to\infty} x_{m+1} = \lim_{m\to\infty} g(x_m) = g\left(\lim_{m\to\infty} x_m\right) = g(z)$$

onde se usou o facto de g ser contínua. Ou seja, z é ponto fixo de g.

Observação 2.1 *O exemplo 2.11 ilustrou que a sucessão pode ou não ter limite finito (na verdade pode não existir limite). Depende de g e de x_0. Por outro lado, notemos que os limites das sucessões convergentes são pontos fixos de $g(x) = x^3$ (mostre), o que seria de esperar, atendendo ao Teorema 2.4, visto g ser uma função contínua.*

No exemplo considerado conhecem-se os pontos fixos, logo não precisamos de procurar aproximações para eles. O exemplo serviu apenas para ajudar à compreensão deste tópico.

Num caso geral, em que apenas sabemos que para g existe um ponto fixo z pertencente a um certo intervalo I, pretende-se aproximá-lo utilizando a sucessão $x_{m+1} = g(x_m)$, o que se consegue se ela for convergente para esse z. A ideía é começar com uma certa estimativa $x_0 \in I$ para z, e, atendendo ao significado de "limite duma sucessão", é sempre possível calcular iteradas sucessivas $x_1, x_2, x_3...$ até que se chegue "suficientemente próximo" de z (que é o limite). Cada iterada x_k constitui uma aproximação para z com erro absoluto $|z - x_k|$. Veremos que é possível conhecer um majorante para esse erro, o que permite concluir sobre a qualidade da aproximação x_k (perto ou longe do valor exacto z).

Em seguida, consideramos o Teorema do ponto fixo, nos permite responder às questões:

Dada a equação $x = g(x)$,

- em que condições g tem um único ponto fixo z num certo intervalo I?
- como se deve escolher x_0 de modo a que a sucessão $x_{m+1} = g(x_m)$ convirja para esse ponto fixo?

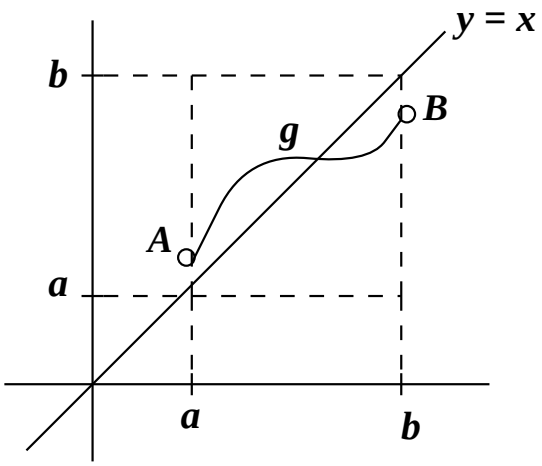
Existência e unicidade

O nosso primeiro objectivo é deduzir condições que garantam que uma dada função g tenha pelo menos um ponto fixo. Ou seja, quando poderemos concluir que o gráfico de g intersecta o gráfico da recta $y = x$ pelo menos uma vez ?

Do exemplo 2.9 concluimos que, para tal, não basta f ser uma função contínua. Mas se juntarmos mais uma condição, obtemos o seguinte resultado.

Teorema 2.5 (Existência) *Seja g definida e contínua em $I = [a, b]$ e tal que $g(I) \subset I$. Então g tem pelo menos um ponto fixo $z \in I$.*

Dem: A condição $g(I) \subset I$ significa que o conjunto de valores de $g(x)$, com $a \leq x \leq b$, está também entre a e b. Ou seja, o gráfico de g está contido no quadrado a tracejado, como mostra a figura seguinte.



Intuitivamente, vemos que para ir de A a B, o gráfico de g tem de cortar o da recta $y = x$ pelo menos uma vez. Provemos esta afirmação matematicamente.

Se $g(a) = a$ ou $g(b) = b$, encontrámos um ponto fixo.

Excluído este caso trivial, seja $g(a) \neq a$ e $g(b) \neq b$. Então $g(a) > a$ e $g(b) < b$, donde $g(a) - a > 0$ e $g(b) - b < 0$. Isto implica que a função $h(x) = g(x) - x$ vai ter sinais contrários em a e b. Sendo h contínua, vai existir um $z \in]a, b[$ tal que $h(z) = 0$. Esse z verifica $g(z) - z = 0$, ou seja, trata-se de um ponto fixo de g. $\square$

Para garantir que o gráfico de g intersecte o da recta $y = x$ uma só vez, o exemplo 2.10 sugere que se imponha alguma restrição sobre a variação de g.

Teorema 2.6 (Existência e unicidade) *Seja g definida e diferenciável em $I = [a, b]$, tal que:*

$$g(I) \subset I \tag{2.6}$$
$$\max_{x\in[a,b]} |g^{'}(x)| = L, \quad 0 < L < 1. \tag{2.7}$$

Então g tem exactamente um ponto fixo em I. Ou seja, existe um único $z \in [a, b]$ tal que $g(z) = z$.

Dem: Sejam z_1 e z_2 pontos fixos de g distintos. Vem que $z_1 = g(z_1)$ e $z_2 = g(z_2)$ e, recorrendo ao teorema de Lagrange, tem-se $z_1 - z_2 = g(z_1) - g(z_2) = g^{'}(\xi)(z_1 - z_2)$, com ξ entre z_1 e z_2. Tomando módulos

$$|z_1 - z_2| \leq \max_{x\in[a,b]} |g^{'}(x)||z_1 - z_2| \leq L|z_1 - z_2|$$

Como $0 < L < 1$, resulta a relação $|z_1 - z_2| < z_1 - z_2|$. Assim, a hipótese $z_1 \neq z_2$ conduziu a um absurdo, pelo que só poderá ter-se $z_1 = z_2$. $\square$

Estabelecemos condições suficientes para que g tenha um único ponto fixo em $I = [a, b]$. Seja z esse ponto fixo, isto é, $g(z) = z$.

Vamos em seguida mostrar que, nas condições do Teorema 2.6, a sucessão (2.5) converge e determinaremos um majorante para o erro ao fim de m passos.

Teorema 2.7 (Teorema do ponto fixo)

Seja g definida no intervalo $I = [a, b]$, tal que:

(i) $g(x) \in C^1[a, b]$
(ii) $g(I) \subset I$
(iii) $\max_{x\in[a,b]} |g^{'}(x)| = L, \quad 0 < L < 1.$

Então
(a) *g tem exactamente um ponto fixo em I. Ou seja, existe um único $z \in [a, b]$ tal que $g(z) = z$.*

(b) *Dada uma qualquer aproximação inicial $x_0 \in [a, b]$, a sucessão $x_{m+1} = g(x_m), \quad m = 0, 1, 2, \ldots$ converge para o ponto fixo z.*

Dem: A parte (a) já foi provada. Vejamos (b). Começamos por notar que, da condição $g(I) \subset I$, resulta que se $x_0 \in I$, então todas as iteradas $x_m, \quad m = 1, 2, \ldots$ também estão contidas em I. Tem-se

$$\begin{aligned} z - x_{m+1} &= g(z) - g(x_m) \quad \text{(por se ter : } z = g(z),\ g(x_m) = x_{m+1}) \qquad (2.8) \\ &= g^{'}(\xi_m)\,(z - x_m), \quad \text{com } \xi_m \text{ entre } z \text{ e } x_m, \qquad (2.9) \end{aligned}$$

onde se usou o Teorema de Lagrange. Tomando módulos e usando (2.8), vem

$$|z - x_{m+1}| \leq L\,|z - x_m| \quad m = 0, 1, 2, \ldots \qquad (2.10)$$

e, sendo $0 < L < 1$, resulta que $|z - x_{m+1}| < |z - x_m|$. Isto permite concluir que x_{m+1} é melhor aproximação para z do que x_m.

Para provar que $x_m \to z$, ou seja, que $z - x_m \longrightarrow 0$, apliquemos sucessivamente (2.10). Vem

$$\begin{aligned} &|z - x_1| \leq L\,|z - x_0| \\ &|z - x_2| \leq L\,|z - x_1| \leq L^2\,|z - x_0| \\ &|z - x_3| \leq L\,|z - x_2| \leq L^2\,|z - x_1| \leq L^3\,|z - x_0| \\ &\text{em geral} \\ &|z - x_m| \leq L^m\,|z - x_0| \quad m = 0, 1, 2, \ldots \end{aligned}$$

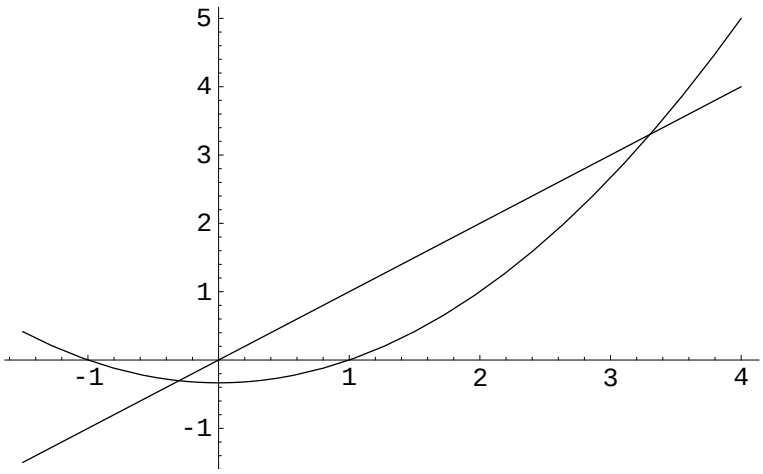
Fazendo $m \longrightarrow \infty$, da relação anterior obtém-se $0 \leq |z - x_m| \leq L^m\,|z - x_0| \to 0$, já que $0 < L < 1 \Longrightarrow L^m \longrightarrow 0$. Então $|z - x_m| \longrightarrow 0 \Longrightarrow x_m \longrightarrow z$. $\square$

Exemplo 2.12

Seja a função $g(x) = (x^2 - 1)/3$. Se g tiver pontos fixos eles são dados resolvendo $\frac{x^2-1}{3} = x$. Tem-se

$$\frac{x^2 - 1}{3} = x \Longleftrightarrow x^2 - 1 - 3x = 0 \Longleftrightarrow x = \frac{3 \pm \sqrt{13}}{2}$$

Da equação anterior concluimos que g tem exactamente 2 pontos fixos. Como $\sqrt{13} \in [3, 4]$, então $\sqrt{13}/2 \in [1.5, 2]$ e uma estimativa grosseira para os pontos fixos de g (ou seja, as raízes da equação $x^2 - 1 - 3x = 0$) será a seguinte: existirá um ponto fixo $z_1 \in [-1, 0]$ e um outro positivo $z_2 \in [3, 4]$.



Consideremos z_1. Verificam-se as condições do Teorema do ponto fixo para o intervalo $I = [-1, 0]$? Em caso afirmativo, além de confirmarmos a localização de z_1, poderemos usar a iteração do ponto fixo para obter um valor aproximado de z_1.

(i) Tem-se $g'(x) = 2x/3$. Então $g \in C^1[-1, 0]$, já que g e g' são funções contínuas.
(ii) $g(I) \subset I$?
Para verificar se $g(I) \subset I$, notemos que, sendo $g'(x) < 0$ para $x \in I$, então g é uma função decrescente em I. Como além disso $g(-1) = 0 \in I$ e $g(0) = -1/3 \in I$, a monotonicidade de g permite concluir que $g(x) \in I$, para qualquer $x \in I$.

(iii) $\max_{x\in[a,b]} |g'(x)| = L, \quad 0 < L < 1$?
Tem-se $|g'(x)| = 2|x|/3$ e, para $-1 \leq x \leq 0$, vem $|g'(x)| \leq 2/3 < 1$. Assim a condição (2.8) é verificada com $L = 2/3$.

Fica assim confirmada a existência de um único ponto fixo $z_1 \in [-1, 0]$ para a função g. Além disso, a sucessão $x_{m+1} = g(x_m) = (x_m^2 - 1)/3$ converge para z_1, qualquer que seja a aproximação inicial $x_0 \in [-1, 0]$. Escolhendo, por exemplo, $x_0 = 0$, obtém-se

$$x_1 = g(x_0) = \frac{0^2 - 1}{3} = -\frac{1}{3} \approx -0.3333333$$
$$x_2 = g(x_1) = \frac{(-1/3)^2 - 1}{3} \approx -0.296296$$
$$x_3 = g(x_2) = \frac{x_2^2 - 1}{3} \approx -0.304069$$
$$x_4 = g(x_3) = \frac{x_3^2 - 1}{3} \approx -0.302514$$
$$x_5 = g(x_4) = \frac{x_4^2 - 1}{3} \approx -0.302828$$
$$x_6 = g(x_5) \approx -0.302765$$
$$x_7 = g(x_6) \approx -0.302778$$
$$x_8 = g(x_7) \approx -0.302775$$

Obs. 1) O Teorema do ponto fixo dá-nos apenas condições suficientes para existência de um único ponto fixo.

Como ilustração, consideremos de novo a função g do exemplo anterior, mas agora o intervalo $I = [3, 4]$. É fácil de provar que a função $f(x) = x^2 - 1 - 3x$ tem um único zero nesse intervalo. Com efeito, tem-se $f(3)f(4) < 0$ e $f'(x) = 2x - 3 > 0$ em I. Então g tem um único ponto fixo em I. Contudo, se tentarmos aplicar o Teorema do ponto fixo à função g no intervalo I, vemos que falham as condições (ii) e (iii):

(ii) $g(3) = 2.666, g(4) = 5$, donde $g(I)$ não está contido em I.

(iii) Sendo $g'(x) = 2x/3$ crescente e $g'(3) = 2 > 1$, tem-se $g'(x) > 1$ para todo o $x \in I$.

Obs. 2) O Teorema do ponto fixo dá-nos condições suficientes para convergência do método iterativo do ponto fixo. Se uma das condições não se verificar não se pode concluir que o método não converge para z.

Se tomarmos para aproximação de z o valor x_m, precisamos estimar o erro $|z - x_m|$. Vamos supor que são verificadas as condições do Teorema (2.7).

I. Erro na iterada x_m, $m > 0$, em função do erro em x_0

Na demonstração do Teorema 2.7, obtivémos a fórmula:

$$|z - x_m| \leq L^m \, |z - x_0|. \tag{2.11}$$

Trata-se de uma fórmula que permite estimar o erro numa certa iterada x_m, *à priori*, isto é, sem ter de se calcular x_m. É preciso aqui uma (boa) estimativa para o erro na iterada inicial.

II. Erro na iterada x_{m+1}, $m > 0$, em função da distância entre as iteradas x_{m+1} e x_m

É válida a seguinte fórmula:

$$|z - x_{m+1}| \leq \frac{L}{1 - L} \, |x_{m+1} - x_m|, \quad m \geq 0. \tag{2.12}$$

Trata-se de uma fórmula de majoração *à posteriori*, isto é, pressupõe termos já calculado x_m e x_{m+1}.

Dem: Tem-se, aplicando (2.10) e depois somando e subtraindo x_{m+1}

$$\begin{aligned} |z - x_{m+1}| &\leq L \, |z - x_m| \\ &= L \, |(z - x_{m+1}) + (x_{m+1} - x_m)| \end{aligned}$$

$$\leq \; L\,|z - x_{m+1}| + L\,|x_{m+1} - x_m|$$

Então

$$|z - x_{m+1}|(1 - L) \leq L|x_{m+1} - x_m|$$

e, sendo $0 < L < 1$, $1 - L > 0$. Dividindo ambos os membros da desigualdade acima por $1 - L$, obtém-se (2.12). □

A fórmula (2.12) pode ser generalizada como se segue.

III. Erro na iterada $x_{m+1},\ m > 0$, em função da distância entre as iteradas x_1 e x_0

$$|z - x_{m+1}| \;\leq\; \frac{L^{m+1}}{1 - L}\,|x_1 - x_0|, \quad m \geq 0. \tag{2.13}$$

Dem: basta usar (2.10) e (2.12) do seguinte modo

$$\begin{aligned} |z - x_{m+1}| &\leq L^m|z - x_1| \qquad \text{(usou-se (2.10))} \\ &\leq L^m\left(\frac{L}{1 - L}|x_1 - x_0|\right) \qquad \text{(usou-se (2.12))}, \end{aligned}$$

donde o resultado. □

Critério de paragem

Qualquer das fórmulas de erro anteriores pode ser usada como teste para se parar o processo iterativo.

Em geral, dado $\epsilon > 0$ (pequeno), o processo iterativo deve parar quando $|z - x_{m+1}| \leq \epsilon$. Da fórmula (2.12) temos:

$$|z - x_{m+1}| \leq \frac{L}{1 - L}|x_{m+1} - x_m|$$

donde vemos que se $\dfrac{L}{1 - L}|x_{m+1} - x_m| < \epsilon$ então teremos $|z - x_{m+1}| \leq \epsilon$. Logo, para que tenhamos um erro inferior a ϵ na iterada x_{m+1} é suficiente que

$$|x_{m+1} - x_m| < \frac{1 - L}{L}\epsilon$$

Obs: Se $L < 1/2$ então $\dfrac{L}{1 - L} < 1$ e nesse caso usa-se o critério de paragem:

$$|x_{m+1} - x_m| < \epsilon.$$

Convergência e Divergência do Método do Ponto Fixo.

A importância da condição $|g'(x)| < 1$ para convergir

Retomemos a igualdade (2.9), que relaciona o erro de x_{m+1} com o de x_m, ou seja:

$|x_{m+1} - z| = |g^{'}(\xi)||x_m - z|$

- Se, $|g'(z)| < 1$, ou equivalentemente, num certo intervalo contendo z, a derivada de g é tal que $|g'(x)| < 1$, então

$$|x_{m+1} - z| = \underbrace{|g^{'}(\xi)|}_{<1} |x_m - z| < |x_m - z|$$

ou seja, a distância de z a x_{m+1} é menor do que a distância de z à iterada anterior x_m. As distâncias diminuem. No teorema 2.8 a seguir, prova-se que é possível ter convergência (as distâncias tendem mesmo para zero) começando com x_0 suficientemente próximo de z.

Teorema 2.8 *: (Teorema da convergência local)* *Seja $g(x)$ diferenciável num intervalo aberto contendo o ponto fixo z de g. Se $|g^{'}(z)| < 1$, então existe um $\epsilon > 0$ tal que a iteração do ponto fixo $x_{m+1} = g(x_m), m = 0, 1, 2, \ldots$ converge para z qualquer que seja a aproximação inicial x_0 que verifica $|z - x_0| < \epsilon$. O ponto fixo z diz-se* **atractor** *para a função g.*

Dem: Por hipótese $|g^{'}(z)| < 1$. Sendo $g^{'}(x)$ contínua num intervalo aberto que contém z, então existe $\epsilon > 0$ tal que

$$x \in I_\epsilon := [z - \epsilon, z + \epsilon] \Rightarrow |g^{'}(x)| \leq K < 1.$$

Tem-se ainda, usando o Teorema do valor médio,

$$\begin{aligned} |g(x) - z| = |g(x) - g(z)| &= |g^{'}(\xi)||x - z| \\ &\leq K\,\epsilon < \epsilon \quad \text{com } \xi \text{ entre } x \text{ e z.} \end{aligned}$$

Mas a desigualdade anterior significa que $g(I_\epsilon) \subset I_\epsilon$. Estão assim verificadas as hipóteses do Teorema (2.7) para o intervalo I_ϵ, pelo que o resultado fica demonstrado. □

Obs.: no Teorema 2.7, conclui-se convergência do método do ponto fixo para x_0 pertencente a um dado intervalo I. Já o Teorema 2.8 apenas nos diz que existe uma vizinhança de z onde a iteração do ponto fixo converge (não indica qual o valor de ϵ).

Usando uma linguagem imprecisa: se $|g'(x)| < 1$ num intervalo contendo z, então temos a garantia de que o método converge se escolhermos x_0 "suficientemente próximo" de z. Diz-se que é um resultado de carácter "local".

- E se, pelo contrário $|g'(z)| > 1$? então

$$|x_{m+1} - z| = \underbrace{|g^{'}(\xi)|}_{>1} |x_m - z| > |x_m - z|$$

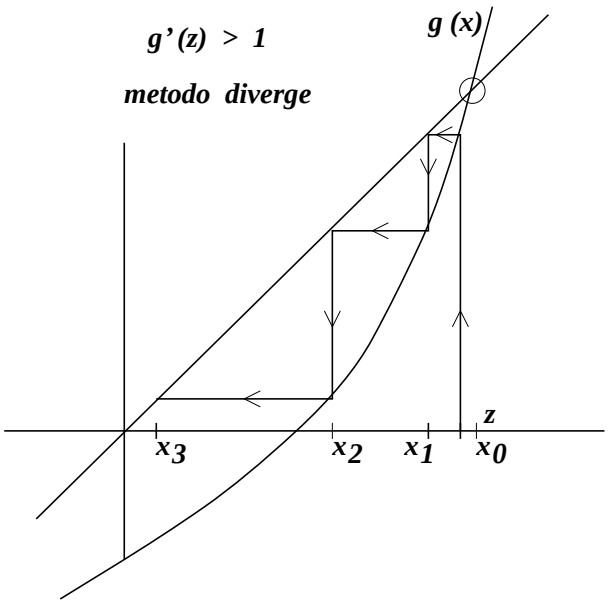
e vemos que a distância de z a x_{m+1} é agora maior do que a distância de z à iterada anterior x_m. As distâncias aumentam e não é possível ter-se convergência da sucessão para z (a não ser que $x_m = z$ a partir de certa ordem).

Acabámos de provar o teorema seguinte:

Teorema 2.9 *(***teorema da divergência***) Seja g uma função diferenciável em I com um ponto fixo $z \in I$. Se $|g'(z)| > 1$, ou equivalentemente, se $|g'(x)| > 1$ numa vizinhança de z, então a sucessão gerada por g não pode convergir para z (a não ser que $x_m = z$ a partir de certa iteração). Diz-se que z é um ponto fixo* **repulsor** *para a função g.*

Exemplo 2.13 *Em relação com a função $g(x) = x^3$ considerada no Exemplo 2.11, como classifica os pontos fixos $z_1 = 0$ e $z_2 = 1.05$? Algum deles é atractor (repulsor)? (Sugestão: calcule $|g'(z_i)|, i = 1, 2$)*

As figuras seguintes ilustram o teorema anterior



Exemplo 2.14

Consideremos a função $g(x) = -ln(x)$. O gráfico abaixo mostra também a recta $y = x$ e a sua intersecção com o gráfico de g. Começando perto do ponto fixo, seja $x_0 = 0.65$. Obtiveram-se as iteradas: $x_1 \simeq 0.430783$, $x_2 \simeq 0.842151$, $x_3 = 0.171796$, $x_4 \simeq 1.76145$, $x_5 \simeq -0.566136$. Note que já não é possível definir x_6.

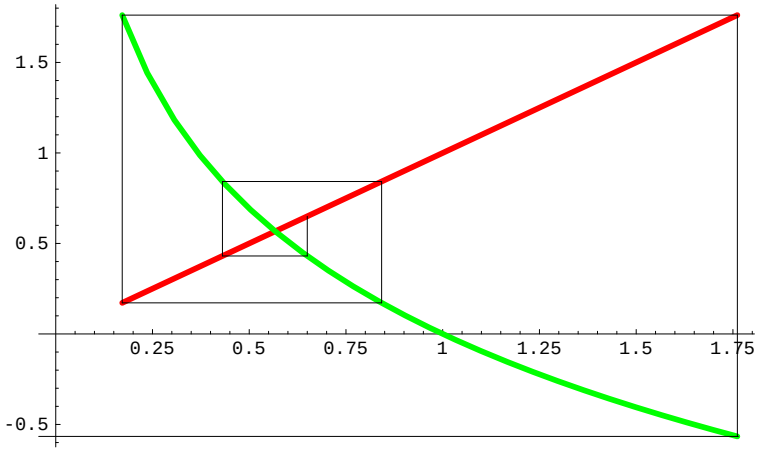
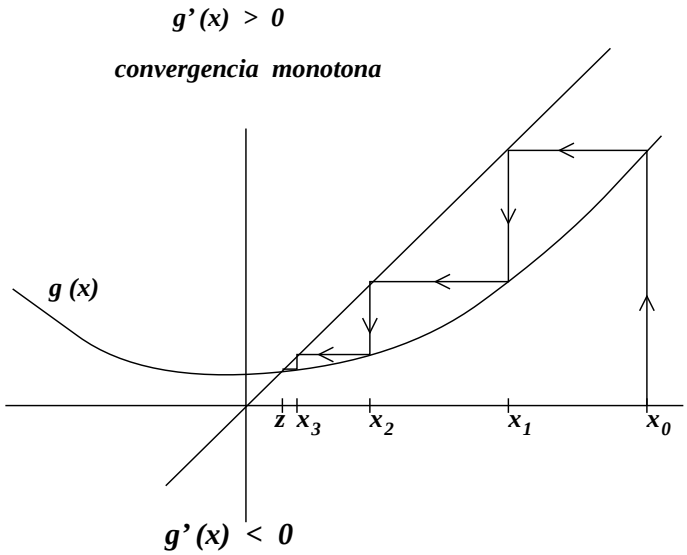
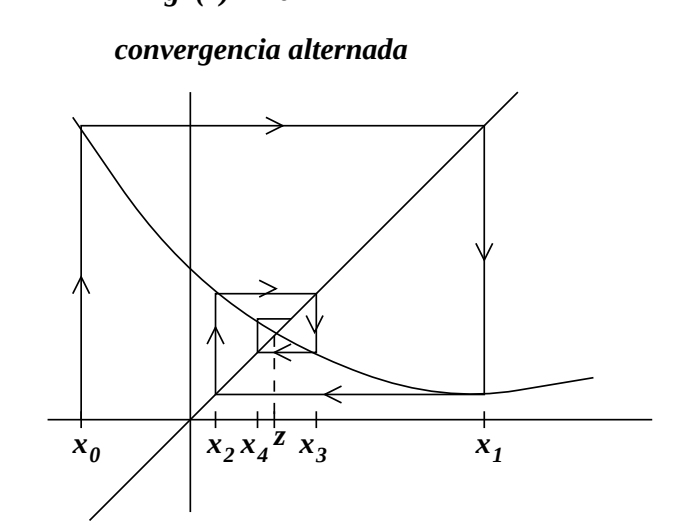


Ilustração Geométrica de convergência monótona e alternada

No caso da sucessão $x_{m+1} = g(x_m)$ convergir, as iteradas podem dispor-se de duas maneiras diferentes em relação ao ponto fixo z (ver figuras seguintes).

- Convergência monótona: $0 \leq g'(x) < 1$.
- Convergência alternada: $-1 < g'(x) \leq 0$.
- Divergência: $|g'(z)| > 1$.





Notemos, que no caso de convergência alternada, o ponto fixo z estará sempre entre duas iteradas consecutivas, x_m e x_{m+1}.

Ordem de Convergência do Método do Ponto Fixo.

Definição 2.3 (ordem de convergência): *Seja* $\{x_n\}$ *uma sucessão que converge para* z *e seja* $e_m := z - x_m$. *Se existirem constantes reais* $p \geq 1$ *e* $K_\infty > 0$ *tais que*

$$\lim_{m\to\infty} \frac{|e_{m+1}|}{|e_m|^p} = K_\infty \tag{2.14}$$

então dizemos que $\{x_m\}$ *converge para* z *com ordem de convergência* p.

K_∞ *chama-se coeficiente assimptótico de convergência*
Se $p = 1$ *e* $K_\infty < 1$, *dizemos que a convergência é linear*
Se $p > 1$ *dizemos que a convergência é supralinear*
Em particular, se $p = 2$ *dizemos que a convergência é quadrática*

Obs: De (2.14), atendendo à definição de limite, resulta a seguinte igualdade assimptótica:

$$|e_{m+1}| \approx |e_m|^p K_\infty, \quad m \text{ suf. grande}$$

A igualdade anterior permite-nos a seguinte conclusão. Como a sucessão $\{x_m\}$ converge para z, temos que a sucessão dos erros e_m tende para 0 quando $m \longrightarrow \infty$. Para m suficientemente grande, e_m estará muito "próximo" de zero. Então, quanto maior for p, mais pequeno será o valor $|e_m|^p$ e, consequentemente, mais pequeno será o erro da iterada seguinte: $e_{m+1} = K_\infty |e_m|^p$.

Somos assim conduzidos ao seguinte conceito.

Rapidez de convergência: rapidez com que os erros decrescem para zero. Depende de K_∞ e p. Quanto maior for p e quanto menor for K_∞, maior é a rapidez de convergência.

Exemplo: Sejam $\{x_m\}$ e $\{\bar{x}_m\}$ duas sucessões satisfazendo

I) $\lim_{m\to\infty} \dfrac{|e_{m+1}|}{|e_m|} = K_\infty > 0$ $\qquad$ ($p = 1$, conv. linear)

II) $\lim_{m\to\infty} \dfrac{|\bar{e}_{m+1}|}{|\bar{e}_m|^2} = \bar{K}_\infty > 0$ $\qquad$ ($p = 2$, conv. quadrática)

a) Suponhamos por simplicidade, que para o método I (conv. linear) se tem:
$|e_{m+1}| \approx |e_m| K_\infty, \quad \forall m.$

Então, para $n = 0, 1, 2, \ldots$ tem-se:

$$|e_1| \approx K_\infty |e_0|$$

$$|e_2| \approx K_\infty |e_1| \approx K_\infty (K_\infty |e_0|) \approx K_\infty^2 |e_0|$$

$$|e_3| \approx K_\infty |e_2| \approx K_\infty (K_\infty^2 |e_0|) \approx K_\infty^3 |e_0|$$

$$\vdots \quad \vdots$$

Em geral obtem-se que: $|e_{m+1}| \approx K_\infty^{(m+1)} |e_0|, \quad n \geq 1 \qquad (1)$

b) Para o método II (quadrático), suponhamos que: $|\bar{e}_{m+1}| \approx \bar{K}_\infty |\bar{e}_m|^2, \quad \forall m$. Então,

$$|\bar{e}_1| \approx \bar{K}_\infty |\bar{e}_0|^2$$

$$|\bar{e}_2| \approx \bar{K}_\infty |\bar{e}_1|^2 \approx \bar{K}_\infty (\bar{K}_\infty |\bar{e}_0|^2)^2 \approx \bar{K}_\infty^3 |\bar{e}_0|^4$$

$$|\bar{e}_3| \approx \bar{K}_\infty |\bar{e}_2^2| \approx \bar{K}_\infty (\bar{K}_\infty^3 |\bar{e}_0|^4)^2 \approx \bar{K}_\infty^7 |\bar{e}_0|^8$$

$$\vdots \quad \vdots$$

Em geral obtém-se: $|e_{m+1}| \approx \bar{K}_\infty^{(2^{m+1}-1)} |\bar{e}_0|^{(2^{m+1})}, \qquad n \geq 1 \qquad (3)$

Suponhamos agora que $K_\infty = \bar{K}_\infty = 0.75$ e $|e_0| = |\bar{e}_0| = 0.5$ e calculemos e_3 e $\bar{e}_3$ utilizando (2) e (3) respectivamente. Então, temos:

$$|e_3| \approx (0.75)^3 |0.5| \approx 0.2109375 \text{ (para o método I)}$$
$$|\bar{e}_3| \approx (0.75)^7 |0.5|^8 \approx \frac{1}{0.75} (0.75 \times 0.5)^8 \approx 5.2 \times 10^{-4} \text{ (para o método II)}$$

Vemos que o erro na terceira iterada com o método quadrático é muito menor que o erro obtido pelo método linear.

Os 2 teoremas seguintes permitem tirar conclusões sobre a ordem de convergência do método do ponto fixo.

Teorema 2.10 *Seja $g(x)$ uma função verificando as condições do teorema do ponto fixo (teor. 2.7) para o intervalo $[a,\ b]$, ou seja,*

- **i)** $g \in C^1[a,\ b]$
- **ii)** $g([a,\ b]) \subset [a,\ b]$
- **iii)** $\max\limits_{x \in [a,\ b]} |g'(x)| \leq L < 1$

Então, a sucessão definida por $x_{m+1} = g(x_m),\ m = 0, 1, \ldots$ converge para $z,\ \forall x_0 \in [a,\ b]$ e tem-se:

$$\lim_{n\to\infty} \frac{|z - x_{m+1}|}{|z - x_m|} = |g'(z)| \tag{2.15}$$

Dem:

A convergência de x_m já foi provada (ver teorema do ponto fixo). Mostremos que o limite acima é válido. Tem-se:

$$z - x_{m+1} = g(z) - g(x_m) = g'(\xi_m)(z - x_m), \quad \xi_m \text{ entre } x_m \text{ e } z, \ \forall m$$

onde se usou o teorema do valor médio de Lagrange.

Então $\dfrac{z - x_{m+1}}{z - x_m} = g'(\xi_m), \forall m$. Note-se que voltamos a usar a igualdade (2.9). Tomando módulos e passando ao limite, vem:

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{|z - x_m|} = \lim_{m\to\infty} |g'(\xi_m)| = {}^{1}|g'(\lim_{m\to\infty} \xi_m)| = {}^{2}|g'(z)|$$

□

Convergência linear

Uma conclusão a tirar da igualdade acima é a seguinte: se $|g'(z)| \neq 0$ a convergência é linear.

Com efeito, neste caso tem-se uma relação do tipo (2.14), com com $p = 1$ e $K_\infty = |g'(z)| > 0$. A convergência é, pois, linear.

Convergência supralinear

E se

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{|z - x_m|} = |g'(z)| = 0 \quad ?$$

Vamos mostrar que a ordem de convergência deixa de ser linear passando a supralinear. Consideremos o seguinte teorema geral que contempla este caso.

Teorema 2.11 (ordem de convergência supralinear do método do ponto fixo): *Seja $z = g(z)$ com $g \in C^p[a,\ b]$, $p \geq 2$ verificando as condições do* **teorema do ponto fixo** *em $[a,\ b]$ e $z \in [a,\ b]$. Se*

$$g'(z) = g''(z) = \ldots = g^{(p-1)}(z) = 0 \ e \ g^{(p)}(z) \neq 0$$

[1] Continuidade de g'

[2] Como ξ_m está entre x_m e z $\forall m$ e $\{x_m\} \longrightarrow z \implies \{\xi_m\} \longrightarrow z$

Então $\lim_{n\to\infty} \dfrac{|e_{m+1}|}{|e_m|^p} = \dfrac{|g^{(p)}(z)|}{p!}$

e portanto a sucessão $\{x_m\}$ *tem ordem de convergência* p *e coeficiente assimptótico de convergência* $K_\infty = \dfrac{|g^{(p)}(z)|}{p!}$.

Dem: Considerando a fórmula de Taylor de g em torno de z vem:

$$g(x) = g(z) \ + \ (x-z)g'(z) \ + \ (x-z)^2\frac{g''(z)}{2!} \ + \ \ldots \ + \ (x-z)^{(p-1)}\,\frac{g^{(p-1)}(z)}{(p-1)!} + (x-z)^p\,\frac{g^{(p)}(\xi)}{p!} \quad , \ \xi \text{ entre } x \text{ e } z.$$

Para $x = x_m$, temos

$$g(x_m) = g(z) \ + \ (x_m-z)g'(z) \ + \ (x_m-z)^2\frac{g''(z)}{2!} \ + \ \ldots \ + \ (x_m-z)^{(p-1)}\,\frac{g^{(p-1)}(z)}{(p-1)!} + (x_m-z)^p\,\frac{g^{(p)}(\xi_m)}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z.$$

Usando as hipóteses $x_{m+1} = g(x_m)$, $g(z) = z$ *e* $g'(z) = g''(z) = \ldots = g^{(p-1)}(z) = 0$, obtém-se

$$x_{m+1} = z \ + \ (x_m-z)^p\,\frac{g^{(p)}(\xi_m)}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z.$$

Daqui vem

$$\begin{aligned} z - x_{m+1} &= -\,(x_m-z)^p\,\frac{g^{(p)}(\xi_m)}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z. \\ &= (-1)^{(p+1)}\,(z-x_m)^p\,\frac{g^{(p)}(\xi_m)}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z. \end{aligned}$$

Então

$$\begin{aligned} &\frac{z-x_{m+1}}{(z-x_m)^p} = (-1)^{(p+1)}\,\frac{g^{(p)}(\xi_m)}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z. \\ \Longrightarrow \ &\frac{|z-x_{m+1}|}{|z-x_m|^p} = \frac{|g^{(p)}(\xi_m)|}{p!} \quad , \ \xi_m \text{ entre } x_m \text{ e } z. \end{aligned}$$

Passando ao limite e usando a continuidade de $g^{(p)}$, vem:

$$\lim_{m\to\infty} \frac{|e_{m+1}|}{|e_m|^p} = \lim_{m\to\infty} \frac{|g^{(p)}(\xi_m)|}{p!}\,\frac{|g^{(p)}(\lim_{m\to\infty}\xi_m)|}{p!} = \frac{|g^{(p)}(z)|}{p!}.$$

Assim, a sucessão $\{x_m\}$ converge para z com ordem de convergência p e factor assimptótico de convergência $K_\infty = \dfrac{|g^{(p)}(z)|}{p!}$ $\square$

2.2.3 Método de Newton

Seja $f \in C^2[a,\ b]$ e suponhamos que $f'(x) \neq 0,\ \forall\ x \in [a,\ b]$, e que existe $z \in [a,\ b]$ tal que $f(z) = 0$. Como vimos anteriormente (teor. 2.11), dada uma sucessão do ponto fixo $x_{m+1} = g(x_m)$ convergente para z, se $g'(z) = 0$ então a convergência é supralinear. Ou seja, nesse caso a ordem é um número $p \geq 2$ (conv. pelo menos quadrática).

Colocamos agora a questão: podemos determinar uma função $A(x)$ (dependente de f) tal que a sucessão do ponto fixo gerada por

$$g(x) = x + A(x)f(x)$$

convirja para z com ordem de convergência $p \geq 2$? A resposta é afirmativa. Com efeito, calculemos $g'(x)$ e imponha-se a condição $g'(z) = 0$. Tem-se

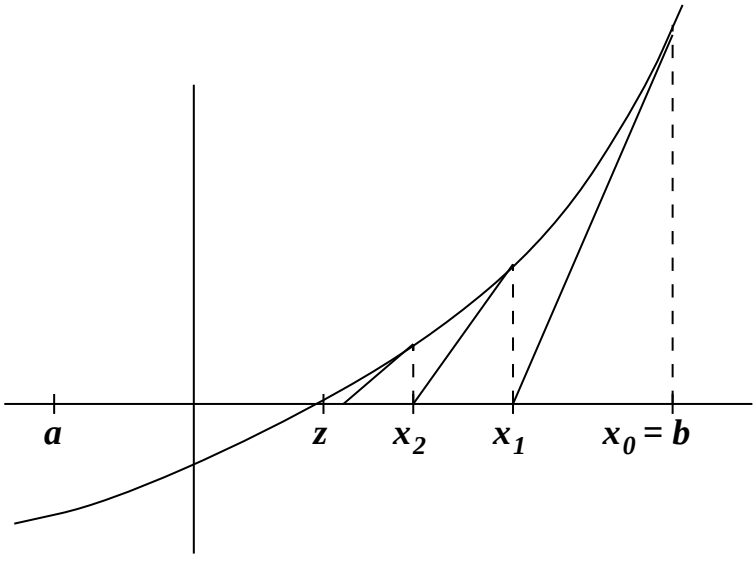
$$g'(x) = 1 + A'(x)f(x) + A(x)f'(x)$$

logo $g'(z) = 1 + A'(z)f(z) + A(z)f'(z) = 1 + A(z)f'(z)$, visto $f(z) = 0$. Impondo $g'(z) = 0$ para se ter convergência supralinear, resulta a condição $A'(z) = -1/f'(z)$. Com a escolha $A(x) = -1/f'(x)$ somos conduzidos ao método

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, \quad m = 0, 1, \ldots \tag{2.16}$$

a que se chama método de Newton.

Interpretação geométrica como "método das tangentes"



O ponto $x_{m+1}, m \geq 0$ é calculado como sendo o ponto de intersecção, com o eixo dos xx, da recta passando por $(x_m, f(x_m))$ com declive $f'(x_m)$. Essa recta é a tangente ao gráfico de f no ponto $(x_m, f(x_m))$) e a sua equação é dada por

$$y = f'(x_m)(x - x_m) + f(x_m)$$

Seja x_{m+1} o ponto de intersecção com o eixo dos xx. Vem

$$0 = f'(x_m)(x_{m+1} - x_m) + f(x_m)$$

donde

$$x_{m+1} - x_m = -\frac{f(x_m)}{f'(x_m)}$$

e obtém-se o método de Newton

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}, m \geq 0$$

Exemplo 2.15

Calcular $\sqrt[3]{100}$ pelo método de Newton.

Solução: Seja $f(x) = x^3 - 100 = 0$ e seja z um zero de f. Então, $z^3 = 100 \implies z = \sqrt[3]{100}$. Para calcularmos z, procedemos como segue:

Seja $I = [4,\ 5]$. Então, $\begin{cases} f(4) = 64 - 100 = -36 \\ f(5) = 125 - 100 = 25 \\ \text{logo, } f(4)f(5) < 0 \overset{Teor1.1}{\Longrightarrow} z \in [4,\ 5] \end{cases}$

Temos que $f'(x) = 3x^2$ e o método de Newton é dado por: $x_{m+1} = x_m - \dfrac{x_m^3 - 100}{3x_m^2}$. Tomando $x_0 = 4$ vem:

$$\begin{cases} x_1 = x_0 - \dfrac{x_0^3 - 100}{3x_0^2} = 4 + \dfrac{36}{48} = 4.750 \\ x_2 = x_1 - \dfrac{x_1^3 - 100}{3x_1^2} = 4.750 - \dfrac{7.171875}{67.68750} = 4.644044 \\ x_3 = x_2 - \dfrac{x_2^3 - 100}{3x_2^2} = 4.644044 - \dfrac{0.158769}{64.701434} = 4.641590 \\ x_4 = x_3 - \dfrac{x_3^3 - 100}{3x_3^2} = 4.641590 - \dfrac{0.0000837}{64.633073} = 4.6415887 \end{cases}$$

Fórmula do erro para o método de Newton:

Teorema 2.12 *Se $f \in C^2[a,\ b]$ então as iteradas do método de Newton satisfazem:*

$$|e_{m+1}| = |z - x_{m+1}| \leq \frac{\max\limits_{x\ \in[a,\ b]} |f''(x)|}{2|f'(x_m)|}(z - x_m)^2$$

Dem:

Comecemos por provar a seguinte relação:

$$|z - x_{m+1}| = (z - x_m)^2 \frac{|f''(\xi_m)|}{2|f'(x_m)|},\ \xi_m \text{ entre } z \text{ e } x_m \tag{2.17}$$

Pela fórmula de Taylor de f em torno de x_m obtemos:

$$f(x) = f(x_m) + (x - x_m)f'(x_m) + \frac{(x - x_m)^2}{2!}f''(\xi_m),\ \xi_m \text{ entre } x \text{ e } x_m$$

Para $x = z$ tem-se:

$$\begin{aligned}
& f(z) = f(x_m) + (z - x_m)f'(x_m) + \frac{(z - x_m)^2}{2!}f''(\xi_m),\ \xi_m \text{ entre } z \text{ e } x_m \\
\implies\ & 0 = f(x_m) + (z - x_m)f'(x_m) + \frac{(z - x_m)^2}{2}f''(\xi_m),\ \xi_m \text{ entre } z \text{ e } x_m \\
\implies\ & 0 = \frac{f(x_m)}{f'(x_m)} + z - x_m + (z - x_m)^2\frac{f''(\xi_m)}{2f'(x_m)},\ \xi_m \text{ entre } z \text{ e } x_m \\
\implies\ & z = \underbrace{x_m - \frac{f(x_m)}{f'(x_m)}}_{x_{m+1}} - (z - x_m)^2\frac{f''(\xi_m)}{2f'(x_m)},\ \xi_m \text{ entre } z \text{ e } x_m \\
\implies\ & z - x_{m+1} = -(z - x_m)^2\frac{f''(\xi_m)}{2f'(x_m)},\ \xi_m \text{ entre } z \text{ e } x_m \\
\implies\ & |z - x_{m+1}| = |(z - x_m)|^2\frac{|f''(\xi_m)|}{2|f'(x_m)|},\ \xi_m \text{ entre } z \text{ e } x_m
\end{aligned}$$

Como em geral não se conhece ξ_m, toma-se um majorante de —f"— e resulta finalmente

$$|z - x_{m+1}| \leq (z - x_m)^2\frac{\max\limits_{x\in[a,b]} |f''(x)|}{2|f'(x_m)|}$$

□

Corolário 2.1 *As iteradas do método de Newton satisfazem*

$$|\underbrace{z - x_{m+1}}_{e_{m+1}}| \leq K|\underbrace{z - x_m}_{e_m}|^2 \quad onde\ K = \frac{\max\limits_{x\in[a,b]}|f''(x)|}{2\min\limits_{x\in[a,b]}|f'(x)|}$$

Número de iterações necessárias p/ obter uma precisão ϵ na iterada x_{m+1}.

Fazendo sucessivamente, $m = 0, 1, 2, \ldots$, na equação $|e_{m+1}| \leq K|e_m|^2$ obtém-se a seguinte relação:

$$|e_{m+1}| \leq K^{(2^{m+1}-1)}|e_0|^{2^{m+1}} \quad \text{onde } e_{m+1} = z - x_{m+1} \quad \text{e} \quad K = \frac{\max\limits_{x\in[a,b]}|f''(x)|}{2\min\limits_{x\in[a,b]}|f'(x)|}\ .$$

OBS: conhecendo-se $|e_0|$ é possível prever os erros em $x_1, x_2, \ldots, x_{m+1}$.

Logo, para que no método de Newton se tenha um erro inferior a ϵ numa iterada x_{m+1}, é suficiente que:

$$K^{(2^{m+1}-1)}|e_0|^{2^{m+1}} < \epsilon \Longrightarrow K^{-1}(K|e_0|)^{2^{m+1}} < \epsilon \Longrightarrow (K|e_0|)^{2^{m+1}} < K\epsilon \Longrightarrow$$

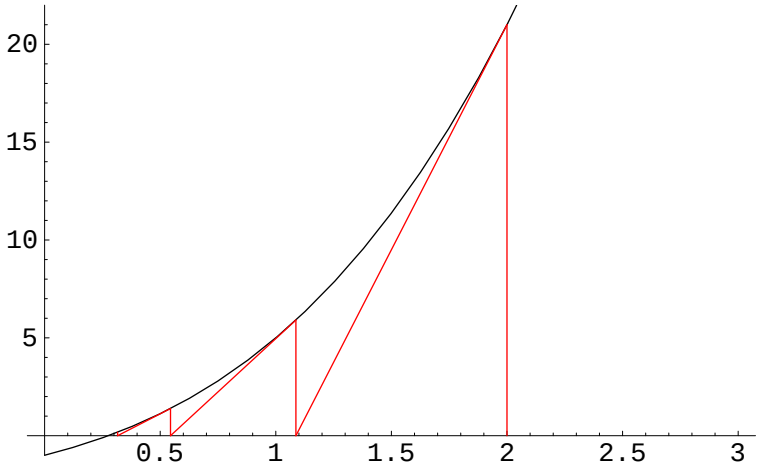
$$2^{m+1}\underbrace{ln(K|e_0|)}_{} < ln(K\epsilon) \Longrightarrow 2^{m+1} > \frac{ln(K\epsilon)}{ln(K|e_0|)} \Longrightarrow m > \frac{ln\left[\dfrac{ln(K\epsilon)}{ln(K|e_0|)}\right]}{\ln(2)} - 1$$

□

Para os resultados que se seguem, é importante relembrar que o método de Newton é um método do ponto fixo, ou seja, a sucessão (2.16) é uma sucessão da forma $x_{m+1} = g(x_m)$, com $g(x) = x - \dfrac{f(x)}{f'(x)}$.

Exemplo 2.16 *Consideremos o método de Newton aplicado à equação $f(x) = 0$ onde $f(x) = x^3 + 2x^2 + 3x - 1$*

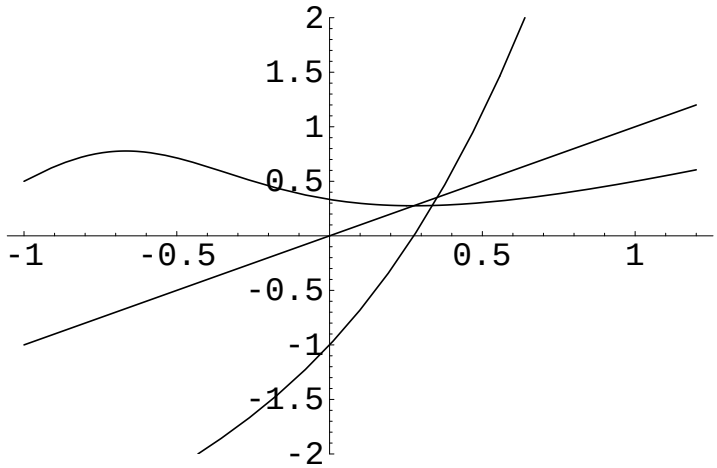
Começamos com uma ilustração geométrica do método de Newton, como "método das tangentes", começando com $x_0 = 2$



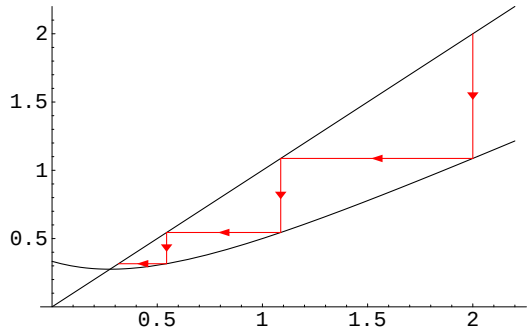
Algumas iteradas são:

2., 1.08695652173913, 0.5445497487285513, 0.315769081825817, 0.2767062213670591, 0.2756828877526929, 0.2756822036512905, 0.275682203650985.

Por outro lado, o método de Newton é um método do ponto fixo com função iteradora $g(x) = x - (f(x)/f'(x)) = x - (x^3 + 2x^2 + 3x - 1)/(3x^2 + 4x + 3)$. Note-se que a raiz da equação $f(x) = 0$ é ponto fixo da função g: tem-se simultaneamente $f(z) = 0$ e $g(z) = z$. Assim, ao aproximarmos o ponto fixo de g pela sucessão $x_{m+1} = g(x_m)$, estamos a aproximar a raiz da equação $f(x) = 0$. A figura abaixo ilustra este facto, mostrando os gráficos de $f(x)$, $g(x)$ e da recta $y = x$.



O gráfico seguinte é uma ilustração geométrica do método de Newton como método do ponto fixo associado à função iteradora $g(x) = x - (f(x)/f'(x)) = x - (x^3 + 2x^2 + 3x - 1)/(3x^2 + 4x + 3)$. Tomou-se para aproximação inicial $x_0 = 2$.



Teorema 2.13 (Convergência local e ordem do Mét. Newton): *Seja* $f \in C^2(I)$, $I = [a,\ b]$ *e suponhamos que* $f'(x) \neq 0,\ \forall\ x \in [a,\ b]$ *e* $\exists\ z \in [a,\ b]$ *tal que* $f(z) = 0$. *Então, existe* $r > 0$ *tq.* $\forall\ x_0 \in I_r = [z - r,\ z + r] \subset I$ *o método de Newton converge para* z *e a ordem de convergência é pelo menos quadrática.*

Dem:

Convergência

Sabendo que o método de Newton é um método do ponto fixo, vamos mostrar que $\exists\ r > 0$ tq. $g(x)$ satisfaz as condições do teorema do ponto fixo em $I_r = [z - r,\ z + r] \subset I$. Isso vai permitir concluir que a sucessão $x_{m+1} = g(x_m)$, ou seja, o método de Newton, converge para z , $\forall x_0 \in I_r$.

Tem-se

$i)\ g(z) = z - \underbrace{\frac{\overbrace{f(z)}^{=0}}{f'(z)}}_{\neq 0} = z$ e $g \in C^1(I)$. (desde que $f \in C^2(I)) \Longrightarrow z$ é ponto fixo de g.

$$ii)\ g'(x) = 1 - \left(\frac{1}{f'(x)} f'(x) - f(x)\frac{f''(x)}{(f'(x))^2}\right) = 1 - \left(1 - f(x)\frac{f''(x)}{(f'(x))^2}\right) = f(x)\frac{f''(x)}{(f'(x))^2}.$$

Portanto, $g'(z) = \underbrace{f(z)}_{=0} \frac{f''(z)}{(f'(z))^2} = 0$

(Note-se que o resultado $g'(z) = 0$ já se sabia ser válido. Porquê ?)

iii) Como $g'(z) = 0$ e g' é continua[3], g está nas condições do Teorema (2.8). Logo a

[3] f é contínua em $\bar{x} \Longleftrightarrow$ dado $\epsilon > 0$ (arbitrário), $\exists\ \delta > 0$ tq. $|f(x) - f(\bar{x})| < \epsilon$, sempre que $|x - \bar{x}| < \delta$

sucessão definida por $x_{m+1} = x_m - \dfrac{f(x_m)}{f'(x_m)}, \quad n = 0, 1, \ldots$ converge para z qualquer que seja a aproximação inicial $x_0 \in I_r$. Ou seja, o método de Newton converge para a solução de $f(x) = 0$ em I_r, $\forall x_0 \in I_r$.

Ordem de convergência

Se aplicarmos limites à igualdade (2.17), vem

$$\lim_{m\to\infty} |z - x_{m+1}| = \lim_{m\to\infty} |(z - x_m)|^2 \frac{|f''(\xi_m)|}{2|f'(x_m)|}$$

donde

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{(z - x_m)^2} = \frac{|f''(z)|}{2|f'(z)|} \tag{2.18}$$

Se $f''(z) \neq 0$, resulta da definição de ordem de convergência que a ordem é 2. Se $f''(z) = 0$, então a ordem será superior a 2. Neste caso, para se determinar a ordem exacta de convergência, investiga-se qual a primeira derivada de g que não se anula em z, como exemplificamos a seguir. □

Observação 2.2 Sobre a ordem de convergência do método de Newton

Sendo o método de Newton um método do ponto fixo, pode-se considerar a aplicação do Teorema geral da ordem de convergência do método do ponto fixo (teor 2.11). Da expressão de $g'(x)$ obtida acima, vem

$$g''(x) = \frac{f(x)}{f'^2(x)} f'''(x) + f''(x) \left[\frac{f'(x)}{f'^2(x)} - 2f(x) \frac{f''(x)}{f'^3(x)} \right]$$

donde

$$g''(z) = \frac{\overbrace{f(z)}^{=0}}{f'^2(z)} f'''(z) + f''(z) \left[\frac{f'(z)}{f'^2(z)} - 2\overbrace{f(z)}^{=0} \frac{f''(z)}{f'^3(z)} \right] = \frac{f''(z)}{f'(z)}$$

Segue-se a conclusão já obtida anteriormente de que se $f''(z) \neq 0$ então a convergência é quadrática. Se $f''(z) = 0$, a ordem será superior a 2. A análise, nesse caso, prosseguiria calculando $g'''(x)$, etc, e a ordem seria p tal que $g^{(p)}(z) \neq 0$.

Condições suficientes de convergência para o método de Newton.

O teorema da convergência local garante convergência do método de Newton para z se escolhermos x_0 "suficientemente próximo" de z. Contudo não diz exactamente onde escolher x_0.

Por outro lado, apesar de o método de Newton ser um método do ponto fixo, ou seja uma sucessão da forma $x_{m+1} = g(x_m)$, é em geral difícil mostrar que g satisfaz as condições do **teorema do ponto fixo**.

Apresentamos as seguintes condições suficientes para que o método convirja:

Critério 1: Seja $f \in C^2[a,\ b]$ satisfazendo:

1. $f(a)f(b) < 0$ (existência da raiz)
2. $f'(x) \neq 0,\ \forall\ x \in\ [a,\ b]$ (unicidade da raiz)
3. $f''(x) \geq 0$ ou $f''(x) \leq 0,\ \forall\ x \in\ [a,\ b]$ ($f(x)$ não muda o sentido da concavidade em $[a,\ b]$)
4. $\dfrac{|f(a)|}{|f'(a)|} < b - a$ e $\dfrac{|f(b)|}{|f'(b)|} < b - a$

Então o método de Newton converge para a solução de $f(x) = 0$, qualquer que seja $x_0 \in [a,\ b]$.

Critério 2: Se as condições 1., 2. e 3. do **critério 1** são válidas e x_0 é escolhido de modo que

$$f(x_0)f''(x) \geq 0\ ,\ \forall\ x \in\ [a,\ b]$$

Então o método de Newton converge para a solução de $f(x) = 0$ e a convergência será monótona.

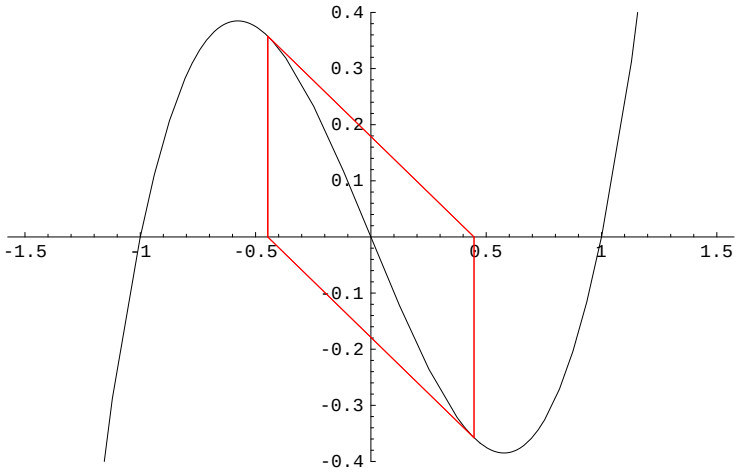
Finalizamos com mais um exemplo

Exemplo 2.17 *Considere a aplicação do método de Newton à equação $x^3 - x = 0$, para aproximar a raiz $z = 0$.*

Começando com $x_0 = -1/\sqrt{5} \simeq -0.4472135954999579$ obtém-se:

-0.4472135954999579, 0.4472135954999579, -0.4472135954999579, 0.4472135954999579, -0.4472135954999579, 0.4472135954999579, -0.4472135954999579, 0.4472135954999579,...

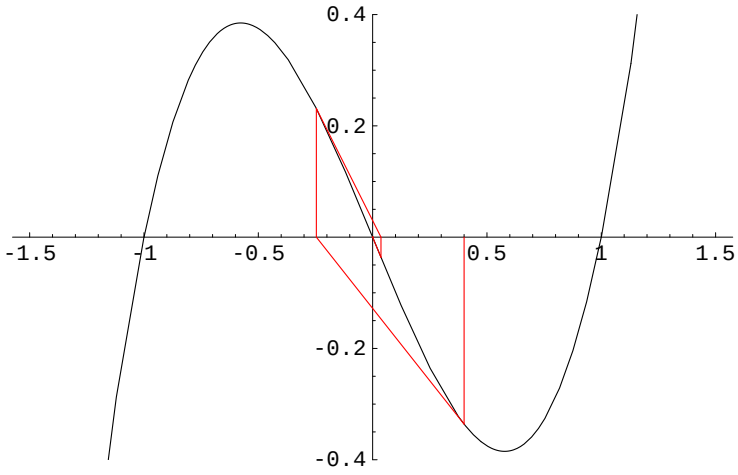
Ou seja, $x_0 = x_2 = x_4 = ...$ e $x_1 = x_3 = x_5 = ...$, o que é ilustrado graficamente na seguinte figura:



Este é um caso em que o método de Newton claramente não converge para a raiz $z = 0$. Então é porque alguma condição das hipóteses dos critérios de convergência falha. Comente.

Por outro lado, começando com $x_0 = 0.4$, eis algumas iteradas:

0.4, -0.24615384615384625, 0.03645668765715054,-0.00009729639046062237, $1.8421296585858235 \times 10^{-12}$, $-1.2502319123287327 \times 10^{-35}$, $3.9084245814817765 \times 10^{-105}$



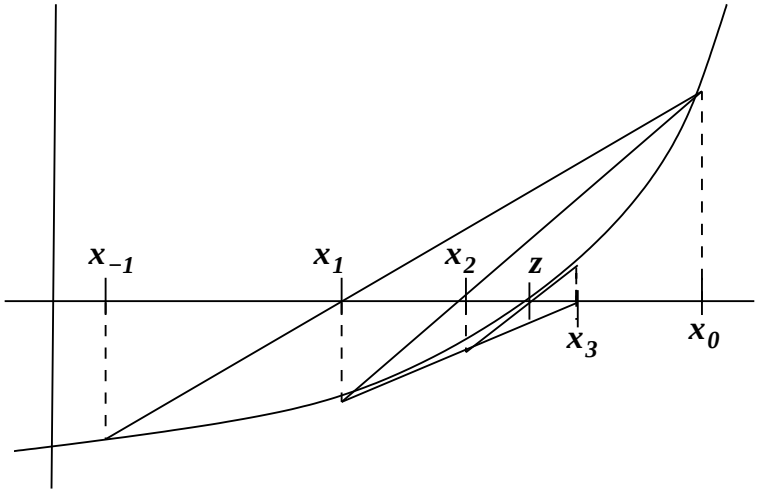
A sucessão acima é uma sucessão convergente para zero. Contudo, ao tentar aplicar os Critérios de convergência, por exemplo no intervalo $[-0.4, 0.4]$, falha uma das condições.

2.2.4 Método da secante

Um dos inconvenientes do método de Newton é o facto de se calcular $f'(x_n)$ a cada iteração, o que pode exigir um grande esforço computacional. Exemplo:

$$\begin{aligned} f(x) &= e^{x^5-2x}\sin(3x)\cos(x^2-1) \\ f'(x) &= e^{x^2}\left(-3\cos(3x)\cos(x^2-1)+\sin(3x)\left(\sin(x^2-1)2x\right)\right) \\ &\quad +\sin(3x)\cos(x^2-1)e^{x^2}(5x^4-2) \end{aligned}$$

O método da secante, em vez de "rectas tangentes", utiliza "rectas secantes" evitando assim o cálculo de $f'(x)$.



Começamos com 2 aproximações iniciais: x_{-1}, x_0. Considera-se os pontos $(x_{-1}, f(x_{-1}))$ e $(x_0, f(x_0))$ e traça-se a recta passando por esses 2 pontos. Da intersecção desta recta com o eixo x obtém-se então o valor de x_1. Tem-se

$$\begin{aligned} & y = \frac{f(x_0)-f(x_{-1})}{(x_0-x_{-1})}(x-x_0)+f(x_0) \\ & y=0 \implies x=x_1 \implies 0=\frac{f(x_0)-f(x_{-1})}{(x_0-x_{-1})}(x_1-x_0)+f(x_0) \\ & \text{donde obtem-se } x_1 = x_0 - f(x_0)\frac{(x_0-x_{-1})}{f(x_0)-f(x_{-1})} \end{aligned}$$

Para o cálculo de x_2, considera-se a recta passando por $(x_0, f(x_0))$ e $(x_1, f(x_1))$. Intersectando esta recta com o eixo x obtém-se x_2:

$$x_2 = x_1 - f(x_1)\frac{(x_1-x_0)}{f(x_1)-f(x_0)}$$

Em geral, x_{n+1} é o zero da recta secante passando por $(x_{n-1}, f(x_{n-1}))$ e $(x_n, f(x_n))$ donde obtem-se a fórmula geral do método da secante:

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

Condições Suficientes de Convergência para o Método da Secante

Critério A1: Se as condições 1., 2., 3. e 4. do critério I do método de Newton forem verificadas então o método da Secante converge para a raiz de $f(x) = 0 \ \forall \ x_{-1}, x_0 \ \in [a,\ b]$.

Critério A2: Se forem verificadas as condições 1., 2. e 3. do critério I do método de Newton e além disso:

4*b*) x_0 e x_{-1} são tais que
$f(x_0)f''(x) \geq 0$
$f(x_{-1})f''(x) \geq 0$

então o método da Secante converge para a única raiz de $f(x) = 0$.

Fórmula do erro do método da Secante

Pode mostrar-se que o erro no método da Secante satisfaz as seguintes fórmulas:

$$z - x_{n+1} = -\frac{f''(\xi_{n1})}{2f'(\xi_{n2})}(z - x_{n-1})(z - x_n) \quad \xi_{n1} \ \in \ int\,(x_n, x_{n-1}, z) \quad \xi_{n2} \in int\,(x_n, x_{n-1})\,.$$

$$|z - x_{n+1}| \leq K|z - x_{n-1}||z - x_n| \qquad \text{onde } K = \frac{\max\limits_{x\in[a,\ b]}|f''(x)|}{2\min\limits_{x\in[a,\ b]}|f'(x)|}$$

Pode ainda mostrar-se que

$$\lim_{n\to\infty}\frac{|z - x_{n+1}|}{|z - x_n|^p} = \left(\frac{|f''(z)|}{2|f'(z)|}\right)^{1/p} \qquad p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618\ldots$$

Então, se $f'(z) \neq 0$ e $f''(z) \neq 0$, a ordem de convergência é $p = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618\ldots$,

Exemplo 2.18

Considere a equação

$$e^{-x^2} - x^2 = 0$$

a) Mostre que essa equação tem 2 duas raizes, z_1 e z_2 onde $z_1 \in [-1,\ 0]$ e $z_2 \in [0,\ 1]$.

b) Calcule z_2 pelo método da secante tendo como aproximações iniciais $x_{-1} = 0$ e $x_0 = 1$. Faça iteradas até que $|f(x_{n+1})| < 10^{-4}$.

1a) Temos que

$$\begin{cases} f(-1) = f(1) = -1 + e^{-1} = -0.63212, \\ f(0) = e^0 = 1.0... \end{cases} \implies \begin{cases} f(-1)f(0) < 0 \\ f(0)f(1) < 0 \end{cases}$$

Como f é contínua, pelo teorema do valor intermediário existe $z_1 \in [-1,\ 0]$ e $z_2 \in [0,\ 1]$ tq. $f(z_1) = f(z_2) = 0$. Agora, $f'(x) = -2xe^{-x^2} - 2x = -2x(\underbrace{1 + e^{-x^2}}_{>0}) = 0 \iff x = 0$. Como $f'(x)$ tem somente 1 zero $\implies f(x)$ tem no máximo 2 zeros. Portanto, z_1 e z_2 são os únicos zeros de $f(x)$.

1b) Pelo método da secante temos: $x_{n+1} = x_n - f(x_n)\dfrac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$. Tomando $x_{-1} = 0$ e $x_0 = 1$, tem-se:

$$\begin{aligned} x_1 &= x_0 - f(x_0)\frac{(x_0 - x_{-1})}{f(x_0) - f(x_{-1})} = 1 + 0.632121 \times \frac{(1-0)}{-0.632121 - 1.0} = 0.61270 \\ x_2 &= 0.61270 - 0.31161 \times \frac{(0.61270 - 1.0)}{0.31161 + 0.63212} = 0.74058 \\ x_3 &= 0.74058 - 0.02940 \times \frac{0.74058 - 0.61270}{0.02940 - 0.31161} = 0.75390 \\ x_4 &= 0.75390 + 0.00191 \times \frac{(0.75390 - 0.74058)}{-0.00191 - 0.02940} = 0.75309 \end{aligned}$$

Como $f(x_4) = 0.0000020$ temos que x_4 tem a precisão desejada.

2.3 EXERCÍCIOS RESOLVIDOS

Concluímos este capítulo com alguns Exercícios provenientes de enunciados de exames.

1. Considere a equação $\sin x - e^{-x} = 0$.

   (a) Prove que esta equação tem uma raiz $z \in [0.5, 0.7]$.
   Solução. *Seja* $f(x) = \sin x - e^{-x}$. *Como* $f(0.5)f(0.7) < 0$, *fica provada a existência de pelo menos uma raiz no intervalo* $I = [0.5, 0.7]$. *Além disso, sendo* $f'(x) = \cos x + e^{-x} > 0$ *em* I, *a raiz é única.*

   (b) Efectue uma iteração pelo método da bissecção e indique um novo intervalo que contenha z.
   Solução. $x_1 = 0.6$*; novo intervalo:* $I_1 = [0.5, 0.6]$.

   (c) Determine o número m de iterações necessárias para garantir $|z - x_m| < 10^{-6}$.
   Solução. *A partir de* $m = 18$ *pode-se garantir aquela precisão.*

2. Considere a equação
$$3x^2 - e^x = 0 \tag{2.19}$$

   (a) Localize graficamente as raízes da equação (4.4) e indique intervalos de comprimento unitário que as contenham.
   Solução. *Reescrevendo a equação dada numa forma equivalente que permita recorrer a um esboço gráfico, tem-se*
$$3x^2 - e^x = 0 \Leftrightarrow 3x^2 = e^x \Leftrightarrow \begin{cases} y = 3x^2 \\ y = e^x \end{cases}$$
   As funções $y = 3x^2$ *e* $y = e^x$ *intersectam-se, aparentemente, em três pontos, sendo as raízes da equação as respectivas projecções sobre o eixo dos xx. O gráfico indica, deste modo, uma raiz negativa e duas positivas. Para os localizarmos, analisemos os sinais de* f *em diversos pontos, sugeridos pelo gráfico. Tem-se* $f(-1) = 2.63 > 0, f(0) = -1 < 0, f(1) = 0.281 > 0, f(3) = 6.31 > 0, f(4) = -6.59 < 0$. *Conclui-se que existe pelo menos uma raiz em cada um dos intervalos:* $[-1, 0]$,$[0, 1]$,$[3, 4]$. *Para confirmarmos o número total de raízes, calculemos as derivadas de* $f(x) = 3x^2 - e^x$. *Tem-se* $f'(x) = 6x - e^x$ *e* $f''(x) = 6 - e^x \Rightarrow$ f'' *tem um só zero em* $x = \ln 6 \simeq 1.79$. *Então* f' *tem, no máximo, 2 zeros e, por conseguinte,* f *tem, no máximo, 3 zeros. Como há no máximo 3 zeros, então existe existe um só em cada um dos intervalos obtidos.*

   (b) Considere as seguintes sucessões
$$(S1) \quad x_{m+1} = \sqrt{\frac{e^{x_m}}{3}} \qquad (S2) \quad x_{m+1} = \ln\left(3x_m^2\right)$$

Mostre que é possível obter aproximações das raízes positivas da equação 4.4 usando, para cada raiz, uma dessas sucessões. Indique, em cada caso, um intervalo onde poderá escolher a iterada inicial x_0.

Solução. *Note que a sucessão (S1)é da forma:*

$$x_{m+1} = g_1(x_m), \quad com \ g_1(x) = \sqrt{\frac{e^x}{3}}. \tag{2.20}$$

Verifique primeiramente qual a relação entre os pontos fixos de g_1 (positivos) e as raízes da equação. O que conclui ? Em seguida, verifique que as condições do Teorema 2.6 (do ponto fixo) são válidas para g_1 no intervalo $[0, 1]$. Quais as conclusões ?

Analogamente, faça um estudo semelhante para a sucessão (S2), relativamente ao intervalo $[3, 4]$.

(c) Efectue duas iterações usando a sucessão (S1) com $x_0 = 1$. Estime o número de algarismos significativos da aproximação obtida.

(d) Será possível usar a sucessão (S1) para aproximar a maior raiz positiva da equação? E poderá usar a sucessão (S2) para aproximar a menor raiz positiva da equação?

Solução. *Consideremos a raiz maior, $z_3 \in I = [3, 4]$ e a sucessão (S1), ou seja, (2.20). Ao tentarmos verificar se as condições do teorema do ponto fixo são satisfeitas por g_1 em I, notamos que $g_1(3) \simeq 2.5875$, $g_1(4) \simeq 4.2661$, ambos fora do intervalo I. Falhando a condição (ii) das hipóteses do Teorema do ponto fixo (Teor. 2.6), não se pode, pois, concluir nada sobre a convergência ou divergência da sucessão (S1).*

Contudo, vemos que $g_1'(x) = (1/2)\sqrt{\frac{e^x}{3}}$ é positiva e crescente. Tem-se $g_1'(3) \simeq 1.29$ e, sendo $z > 3$, então $g_1'(z) > g_1'(3) > 1$. Então, recorrendo ao Teorema 2.8, concluímos que a sucessão $g_1(x_{m+1}) = \sqrt{\frac{e^{x_m}}{3}}$ não pode convergir para z (a não ser escolhendo $x_0 = z$).

De modo análogo, mostre que (S2) não pode ser usada para aproximar a raiz pertencente a $[0, 1]$.

3. Seja a função

$$g(x) = \frac{1}{3}\ln(x^2 + 1)$$

(a) (i)Prove que a sucessão definida por

$$x_{m+1} = \frac{1}{3}\ln(x_m^2 + 1), \quad m = 0, 1, 2, \ldots$$

converge para um número $z \in [-1, 1]$. (ii) Determine z e a ordem de convergência.

Solução. (i) *Comece por provar que g verifica as condições do Teorema do ponto fixo no intervalo $I = [-1, 1]$. Isso permite concluir que existe $z \in I$ tal*

que a sucessão gerada por g converge para z, sendo que z é o único ponto fixo de g nesse intervalo. (ii) *Por um lado, g tem um único ponto fixo em I e, como* $g(0) = (1/3)\ln(0^2+1) = 0$, *isto significa que* $z = 0$ *é esse ponto fixo.*
(iii) *Ordem de convergência. Sabemos que a ordem de convergência é a ordem da primeira derivada de* $g(x)$ *que não se anula em* $x = z$ *(ver Teorema 2.10). Neste caso, tem-se* $z = 0$ *e*

$$g'(x) = \frac{2x}{3(x^2+1)} \Longrightarrow g'(z) = 0$$

Então a convergência é supralinear (se $g'(z) \neq 0$, *seria linear, como se vê atendendo ao Teorema 2.9). Calculemos* $g''(x)$ *para depois obtermos* $g''(z)$. *Vem*

$$g''(x) = \frac{-6(x^2-1)}{9(x^2+1)^2} \Longrightarrow g''(z) = g''(0) = \frac{2}{3} \neq 0$$

Logo, a convergência é quadrática, atendendo a que (pelo Teorema 2.10)

$$\lim_{m\to\infty} \frac{|e_{m+1}}{|e_m|^2} = \frac{|g''(z)|}{2} = \frac{2}{(3)(2)} = 1/3 \neq 0 \tag{2.21}$$

(b) Efectue algumas iterações, começando com $x_0 = 1$, e calcule os quocientes

$$\frac{|e_1|}{(e_0)^2}, \quad \frac{|e_2|}{(e_1)^2}, \frac{|e_3|}{(e_2)^2}, \ldots$$

Os resultados parecem estar de acordo com o que provou na alínea anterior ?
Solução. *Cálculo de algumas iteradas:*

$$\begin{aligned}
&x_0 = 1\\
&x_1 = g(x_0) = g(1) = \frac{1}{3}\ln(x_0^2+1) \simeq 0.2310491\\
&x_2 = g(x_1) \simeq 0.0173358\\
&x_3 = g(x_2) \simeq 0.0001002\\
&x_4 \textit{ já dá zero, usando máquina de calcular}\\
&\ldots
\end{aligned}$$

A sucessão deve convergir para o ponto fixo $z = 0$. *Calculemos agora os erros para cada iterada:*

$$\begin{aligned}
e_0 &= z - x_0 = -x_0 \simeq -1\\
e_1 &= z - x_1 \simeq -0.2310491\\
e_2 &= z - x_2 \simeq -0.0173358\\
e_3 &= z - x_3 \simeq -0.0001002\\
&\ldots
\end{aligned}$$

Resultam então os quocientes

$$\frac{|e_1|}{e_0^2} \simeq 0.2310491$$
$$\frac{|e_2|}{e_1^2} \simeq 0.3247397$$
$$\frac{|e_3|}{e_2^2} \simeq 0.3334110$$
$$\ldots$$

Pela teoria (ver alínea anterior), o limite da sucessão acima é $1/3 \simeq 0.33333...$ *(coeficiente assimptótico da convergência).*

4. Pretende-se determinar, utilizando o método de Newton, a maior das duas raízes positivas da equação
$$-x^3 + 14x - 1 - e^x = 0.$$

(a) Mostre que se x_0 for escolhido no intervalo $I = [2.6, 3]$, estão asseguradas as condições de convergência do método.

Solução. *Verifiquemos que são satisfeitas, para esse intervalo* I*, as condições 1,2,3,4 suficientes para a convergência do método de Newton (Critério 1).*

Vê-se facilmente que $f \in C^2(I)$*, já que é soma de funções com derivadas contínuas de todas as ordens.*

1. $f(2.6) \simeq 4.3603$ *e* $f(3) \simeq -6.0855$*, pelo que a eq. tem uma única raiz em* I*.*

2. $f'(x) = -3x^2 + 14x - e^x$ *é diferente de zero em* I *? Analisemos o seu sinal. Como* $f''(x) = -6x - e^x = -(6x + e^x) < 0$ *em* I*, então* f' *é decrescente em* I*. Sendo* $f'(2.6) = -19.743... < 0$*, vem que* $f'(x) \leq f'(2.6) < 0$ *para* $x \in I = [2.6, 3]$*, logo verifica-se 2.*

3. Já provámos que f'' *tem sempre o mesmo sinal em* I

4.
$$\frac{|f(2.6)|}{|f'(2.6)|} = 0.220843 < (3 - 2.6) \quad \frac{|f(3)|}{|f'(3)|} = 0.183933 < (3 - 2.6)$$

Verificando-se as condições do Critério 1, podemos concluir que o método de Newton (*), aplicado à equação dada, converge para a única raiz da eq. no intervalo $I = [2.6, 3]$*, qualquer que seja* $x_0 \in I$*.*

(*) Ou seja, a sucessão definida por:

$$x_{m+1} = x_m - \frac{-x_m^3 + 14x_m - 1 - e^{x_m}}{-3x_m^2 + 14 - e^{x_m}}, \quad m = 0, 1, 2, ...$$

(b) Calcule um majorante para o erro da segunda iterada (não efectue iterações).

Solução. *Sem efectuar iterações, há que utilizar a fórmula dada no Teorema 2.11, que é da forma:*

$$|z - x_{m+1}| \leq K\,(z - x_m)^2 \tag{2.22}$$

Neste caso, obtenha $K \simeq 0.965$ *e* $|z - x_2| \leq 0.02$.

5. Considere os seguintes métodos para obter um valor aproximado de $z = \sqrt{10}$:

   (a) método de Newton aplicado à função $f(x) = x^2 - 10$. Mostre que se escolher $x_0 = 4^*$ então o método de Newton converge e a ordem é dois. Calcule três iteradas e determine um majorante para o erro de x_3. Quantos algarismos significativos pode garantir ? (* Note que pode concluir convergência se escolher para x_0 qualquer valor ≥ 4)

   Solução. *Pretende-se aproximar* $z = \sqrt{10}$, *que é a raiz positiva da equação* $f(x) = 0$, *ou seja,* $x^2 - 10 = 0$.

   (i) Para provar a convergência começando com um determinado valor de x_0 *(neste caso,* $x_0 = 4$*), tente aplicar o Critério 2 das condições suficientes de convergência do método de Newton. Pode usar o valor 4 para um dos extremos do intervalo onde vai aplicar o Critério 2. (Também pode aplicar o Critério 1?).*

   Para provar que pode concluir convergência se escolher para x_0 *qualquer valor* ≥ 4, *note que fica garantida a convergência desde que* x_0 *seja tal que* $f(x_0) > 0$ *(use o Critério 2).*

   (ii) Para provar que a ordem de convergência é 2, pode-se usar a relação deduzida na parte iii) da dem. do Teorema 2.12 (ver (2.18):

$$\lim_{m\to\infty} \frac{|z - x_{m+1}|}{(z - x_m)^2} = \frac{|f''(z)|}{2|f'(z)|} \tag{2.23}$$

   Se o limite for diferente de zero, sabemos, pela, Definição 2.2, que a ordem é 2. Neste caso, $f''(x) = 2$, *logo diferente de zero para qualquer* x *e, em particular, em* z. *A conclusão sai facilmente.*

   Nota: também poderia usar o facto do método de Newton ser um método do ponto fixo, ou seja, uma sucessão do tipo $x_{m+1} = g(x_m)$. *(Identifique a função* g*). Determinar a ordem equivale a investigar qual a primeira derivada de* g *que não se anula em* z *(ler Obs.2.2).*

   (b) método de Newton aplicado à função $f(x) = x^{-1/2}(x^2 - 10)$. Admitindo que o método converge, mostre que a ordem de convergência é 3.

6. Considere a equação

$$f(x) = x\tan(x) - 1 = 0,$$

   Prove que o método converge para a única raiz no intervalo $[0.8, 0.9]$, quaisquer que sejam as aproximações iniciais x_0, x_{-1} nesse intervalo. Obtenha as três primeiras iteradas para o cálculo da raiz e determine um majorante do erro do resultado obtido.

Solução. *Verifique que o Critério 1 do método da secante é aplicável no intervalo* $[0.8, 0.9]$. *Fazendo* $x_{-1} = 0.8, x_0 = 0.9$, *obtém-se:* $x_1 \simeq 0.856788$, $x_2 \simeq 0.860127$, $x_3 \simeq 0.860335$. *Para obter um majorante do erro de* x_3, *use a fórmula*

$$|z - x_{m+1}| \leq K|z - x_m||z - x_{m-1}|$$

Tem-se $K \simeq 2.063$. *Sugere-se que note o seguinte:* $|z - x_2| < |x_3 - x_2| \simeq 0.207 \times 10^{-3}$ *e* $|z - x_1| < |x_3 - x_1| \simeq 0.355 \times 10^{-2}$. *Vai-se obter o seguinte majorante* $|z - x_3| \leq 0.152 \times 10^{-5}$, *o que garante 5 algarismos significativos na aproximação* x_3.

3 Resolução de Sistemas de equações

Sistemas de equações lineares

Introdução

Muitos problemas provenientes da Física, Matemática, Engenharia, Biologia, Economia, entre outros, envolvem sistemas de equações lineares. Esses sistemas nalguns casos aparecem directamente como parte do modelo matemático do problema real e, noutros casos, surgem em conjunto com a aplicação dum método numérico. Como veremos, a resolução dum sistema de equações não lineares pelo método de Newton conduz a um sistema de equações lineares; também a aplicação de certos métodos numéricos na resolução de equações diferenciais conduz em geral a sistemas lineares da forma:

$$\begin{array}{rccccc} E_1: & a_{11}x_1 & +a_{12}x_2 & +\ldots & +a_{1n}x_n & = b_1 \\ E_2: & a_{21}x_1 & +a_{22}x_2 & +\ldots & +a_{2n}x_n & = b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_n: & a_{n1}x_1 & +a_{n2}x_2 & +\ldots & +a_{nn}x_n & = b_n \end{array}$$

onde $\begin{cases} x_j, j = 1, 2, \ldots, n \text{ são as incógnitas,} \\ a_{i,j}, b_j, i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n, \\ \text{são constantes dadas.} \end{cases}$

O sistema linear acima pode ser representado matricialmente por:

$$A\mathbf{x} = \mathbf{b} \tag{3.1}$$

onde $\begin{cases} A = (a_{i,j}) & \text{matriz dos coeficientes} \\ \mathbf{b} = (b_j) & \text{vector do lado direito} \\ \mathbf{x} = (x_j) & \text{vector solução.} \end{cases}$

Para resolver (3.1), existem duas classes de métodos:

- Métodos Directos: métodos que encontram a solução (exacta) num número finito de operações. Entre os métodos existentes, temos os seguintes: Método de Eliminação de Gauss e os Métodos de factorização LU (Doolittle, Crout e Cholesky).
- Métodos Iterativos: métodos que encontram a solução num número infinito de operações. Mais precisamente, num método iterativo convergente obtém-se uma sucessão de aproximações convergindo para a solução exacta. Aqui estudaremos, em particular, o método de Jacobi e o método de Gauss-Seidel.

Nestas notas começamos com uma revisão do método de Gauss e dos métodos de factorização e descrevemos, em seguida, duas técnicas de pesquisa de pivot. Apresentamos dois

tipos de matrizes especiais e completa-se a primeira parte com uma introdução ao condicionamento e análise do erro de sistemas lineares.

3.1 Métodos directos para sistemas lineares

Método de Eliminação de Gauss

Para obtermos uma solução do sistema (3.1), assumiremos que a matriz A é invertível ($det(A) \neq 0$).

A idéia básica do método de Eliminação de Gauss é transformar o sistema (3.1) num sistema equivalente

$$A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$$

onde $A^{(n)}$ é uma matriz triangular superior. Essa transformação é obtida utilizando as seguintes operações:

$$\begin{cases} i)\ E_i - \lambda E_j \longrightarrow E_i & \{ \text{ substituição de } E_i \text{ por a combinação linear de } E_i \text{ e } E_j \ \} \\ ii)\ \lambda E_i \longrightarrow E_i & \{ \text{ substituição de } E_i \text{ por } \lambda E_i,\ \lambda \neq 0 \ \} \\ iii)\ E_i \longleftrightarrow E_j & \{ \text{ substituição de } E_i \text{ por } E_j \ \} \end{cases}$$

Veremos agora que a aplicação dessas operações ao sistema (3.1) conduz a um sistema equivalente mais fácil de se resolver.

Exemplo 3.1

Seja o sistema linear

$$\begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 4 \\ -1 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}.$$

Para resolvê-lo, procedemos como segue:

1. $a_{11} = 4 \neq 0 \Longrightarrow$ podemos colocar zeros abaixo de a_{11}.

   $$\begin{cases} m_{21} = \dfrac{a_{21}}{a_{11}} = \dfrac{2}{4} = \dfrac{1}{2} \\ m_{31} = \dfrac{a_{31}}{a_{11}} = -\dfrac{1}{4} = -\dfrac{1}{4} \end{cases} \text{ e efectuando as operações } \begin{cases} E_2 - m_{21}E_1 \longrightarrow E_2 \\ E_3 - m_{31}E_1 \longrightarrow E_3 \end{cases}$$

obtém-se o sistema equivalente: $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ onde

$$\underbrace{\left[\begin{array}{ccc} 4 & -9 & 2 \\ 0 & 1/2 & 3 \\ 0 & -1/4 & 5/2 \end{array}\right]}_{A^{(2)}} \vdots \underbrace{\begin{array}{c} 2 \\ 2 \\ 3/2 \end{array}}_{\mathbf{b}^{(2)}}$$

2. $a_{22}^{(2)} = 1/2 \neq 0 \Longrightarrow$ podemos colocar zeros abaixo de $a_{22}^{(2)}$.

   $m_{32} = \dfrac{a_{32}^{(2)}}{a_{22}^{(2)}} = -\dfrac{1/4}{1/2} = -\dfrac{2}{4} = -\dfrac{1}{2}$ e efectuando a operação $E_3 - m_{32}E_2 \longrightarrow E_3$ obtém-se o sistema equivalente $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$:

$$\underbrace{\left[\begin{array}{ccc|c} 4 & -9 & 2 & 2 \\ 0 & 1/2 & 3 & 2 \\ 0 & 0 & 4 & 5/2 \end{array}\right]}_{A^{(2)} \quad \mathbf{b}^{(2)}}$$

   Portanto, o sistema original foi transformado no sistema equivalente:

$$\left[\begin{array}{ccc} 4 & -9 & 2 \\ 0 & 1/2 & 3 \\ 0 & 0 & 4 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right] = \left[\begin{array}{c} 2 \\ 2 \\ 5/2 \end{array}\right]$$

   que tem como solução:

   $x_3 = 5/2/4 = 5/8$
   $1/2x_2 + 3x_3 = 2 \Longrightarrow x_2 = (2 - 3x_3)/1/2 = (2 - 3 \times 5/8)/1/2 = 1/4$
   $4x_1 - 9x_2 + 2x_3 = 2 \Longrightarrow x_1 = 2 + 9x_2 - 2x_3)/4 = (2 + 9 \times 1/4 - 2 \times 5/8)/4 = 3/4$

No caso geral tem-se:

1º Passo: $A\mathbf{x} = \mathbf{b} \longrightarrow A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$

a_{11} é chamado "elemento pivot". Se $a_{11} \neq 0$ então efectuam-se as seguintes operações:

$$\begin{cases} m_{i1} = a_{i1}/a_{11} \\ E_i - m_{i1}E_1 \longrightarrow E_i, \ i = 2, 3, \ldots, n \end{cases}$$

e obtém-se o sistema equivalente:

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} & b_n \end{array}\right] \longrightarrow \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & a_{32}^{(2)} & \ldots & a_{3n}^{(2)} & b_3^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

2º Passo: $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)} \longrightarrow A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$

Se $a_{22}^{(2)} \neq 0$ então calculam-se os multiplicadores

$$m_{i2} = a_{i2}^{(2)}/a_{22}^{(2)},\ i = 3, 4, \ldots, n$$

e efectuam-se as operações:

$$E_i - m_{i2}E_2 \longrightarrow E_i,\ i = 3, 4, \ldots, n$$

Obtém-se o sistema equivalente:

$$[A^{(3)} : \mathbf{b}^{(3)}] = \left[\begin{array}{ccccccc} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} & \vdots & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2n}^{(2)} & \vdots & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3n}^{(3)} & \vdots & b_3^{(3)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & a_{n3}^{(3)} & \ldots & a_{nn}^{(3)} & \vdots & b_n^{(3)} \end{array}\right]$$

Prosseguindo o processo, chegamos, no passo $(n-1)$, ao sistema $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$, onde

$$[A^{(n)} : \mathbf{b}^{(n)}] = \left[\begin{array}{ccccccc} a_{11} & a_{12} & a_{13} & \ldots & \ldots & a_{1n} & \vdots & b_1 \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & \ldots & a_{2n}^{(2)} & \vdots & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \ldots & \ldots & a_{3n}^{(3)} & \vdots & b_3^{(3)} \\ \vdots & \vdots & 0 & \ldots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \ldots & 0 & a_{nn}^{(n)} & \vdots & b_n^{(n)} \end{array}\right]$$

É-se assim conduzido a um sistema linear equivalente ao sistema inicial, cuja matriz é triangular superior. Este sistema pode facilmente ser resolvido, começando com:

$x_n = b_n^{(n)}/a_{nn}^{(n)}$ e calculando-se $x_{n-1}, x_{n-2}, \ldots, x_1$, sucessivamente.

Obs: Se no passo k acontecer que o elemento pivot $a_{kk}^{(k)} = 0$, então procura-se um elemento $a_{lk}^{(k)} \neq 0$ na coluna k, onde $l > k$ e trocam-se as linhas l e k, ou seja, efectua-se a operação $E_k \longleftrightarrow E_l$.

Métodos de Factorização LU

Estes métodos consistem em decompor a matriz A num produto de duas matrizes LU, ou seja, $A = LU$, onde L é uma matriz triangular inferior e U é uma matriz triangular superior.

Suponhamos que existem matrizes L e U tais que $A = LU$, onde L é triangular inferior e U é triangular superior. Então, dado o sistema linear $A\mathbf{x} = \mathbf{b}$, podemos escrever

$$(LU)\mathbf{x} = \mathbf{b} \tag{3.2}$$

Fazendo $U\mathbf{x} = \mathbf{g}$ tem-se que

$$L\mathbf{g} = \mathbf{b}, \tag{3.3}$$

que é um sistema triangular inferior. Resolvendo (3.3) em ordem a $\mathbf{g}$, a solução $\mathbf{x}$ é em seguida obtida resolvendo-se o sistema

$$U\mathbf{x} = \mathbf{g}, \tag{3.4}$$

que é um sistema triangular superior. Portanto, para resolver o sistema $A\mathbf{x} = \mathbf{b}$, basta resolver os dois sistemas triangulares definidos por (3.3) e (3.4).

Existem várias maneiras de se encontrar matrizes L e U de modo que $A = LU$, como veremos a seguir.

Cálculo das matrizes L e U.

Considerando a equação matricial

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}}_{A \text{ (matriz dada)}} = \underbrace{\begin{bmatrix} l_{11} & 0 & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & \dots & l_{nn} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} u_{11} & u_{12} & \dots & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & u_{nn} \end{bmatrix}}_{U}$$

vemos que temos $(n^2 + n)$–incógnitas: l_{ij} e u_{ij}. Por outro lado, pela regra de produto de matrizes, podemos formar n^2-equações. Uma maneira de obter n–equações adicionais é fixar n elementos de L ou U. Em seguida consideramos alguns casos particulares.

Método de Doolittle (ou Factorização de Doolittle)

Este método corresponde a impor que os elementos da diagonal de L sejam iguais a 1, ou seja, $l_{ii} = 1,\ i = 1, 2, \ldots, n$. Temos

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ l_{21} & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \ldots & \ldots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \ldots & \ldots & u_{1n} \\ 0 & u_{22} & u_{23} & \ldots & u_{2n} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & u_{nn} \end{bmatrix}.$$

Efectuando o produto LU e usando a igualdade $A = LU$, elemento a elemento, obtemos as equações:

1a. linha de U : $u_{11} = a_{11}, u_{12} = a_{12}, \ldots, u_{1n} = a_{1n}$.

1a. coluna de L (supondo que $a_{11} \neq 0$)
$$\begin{cases} l_{21}u_{11} = a_{21} \implies l_{21} = a_{21}/u_{11} \\ l_{31}u_{11} = a_{31} \implies l_{31} = a_{31}/u_{11} \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad \vdots \\ l_{n1}u_{11} = a_{n1} \implies l_{n1} = a_{n1}/u_{11} \end{cases}$$

2a. linha de U
$$\begin{cases} l_{21}u_{12} + u_{22} = a_{22} \implies u_{22} = a_{22} - l_{21}u_{12} \\ l_{21}u_{13} + u_{23} = a_{23} \implies u_{23} = a_{23} - l_{21}u_{13} \\ l_{21}u_{14} + u_{24} = a_{24} \implies u_{24} = a_{24} - l_{21}u_{14} \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad\quad\quad \vdots \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad\quad\quad \vdots \\ l_{21}u_{1n} + u_{2n} = a_{2n} \implies u_{2n} = a_{2n} - l_{21}u_{1n} \end{cases}$$

2a. coluna de L
$$\begin{cases} l_{31}u_{12} + l_{32}u_{22} = a_{32} \implies l_{32} = (a_{32} - l_{31}u_{12})/u_{22} \\ l_{41}u_{12} + l_{42}u_{22} = a_{42} \implies l_{42} = (a_{42} - l_{41}u_{12})/u_{22} \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad\quad\quad\quad\quad \vdots \\ \quad\vdots \quad\quad \vdots \quad\quad\quad\quad\quad\quad\quad\quad \vdots \\ l_{n1}u_{12} + l_{n2}u_{22} = a_{n2} \implies l_{n2} = (a_{n2} - l_{n1}u_{12})/u_{22} \end{cases}$$

En geral, tem-se:

$$\text{linha } k \text{ de } U : u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj},\ j = k, k+1, \ldots, n$$

$$\text{coluna } k \text{ de } L : l_{ik} = \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk} \right) / u_{kk}, i = k+1, k+2, \ldots, n$$

Pode mostrar-se que se o método de Eliminação de Gauss for aplicado ao sistema $A\mathbf{x} = \mathbf{b}$ sem troca de linhas, então verifica-se a igualdade $A = LU$, com

$$L = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ m_{21} & 1 & 0 & \dots & \dots & 0 \\ m_{31} & m_{32} & 1 & \ddots & & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ m_{n1} & m_{n2} & \dots & \dots & m_{nn-1} & 1 \end{bmatrix}; \; U = A^{(n)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & \dots & a_{1n} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \ddots & \dots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & a_{nn}^{(n)} \end{bmatrix}$$

Exemplo 3.2

Determine a solução do sistema linear abaixo pelo método de Doolittle.

$$\begin{bmatrix} 2 & 3 & -2 \\ 2 & -16 & 10 \\ 14 & -7 & -7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -3 \\ 0 \end{bmatrix}$$

Solução: Pelo método de Doolittle temos:

$$A\mathbf{x} = \mathbf{b} \Longleftrightarrow LU\mathbf{x} = \mathbf{b} \Longleftrightarrow \begin{cases} L\mathbf{g} = \mathbf{b} \\ U\mathbf{x} = \mathbf{g} \end{cases}$$

onde

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \text{ e } U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

Os elementos das matrizes L e U podem ser obtidos usando as fórmulas:

$$\begin{aligned} u_{kj} &= a_{kj} - \sum_{s=1}^{k-1} l_{ks}u_{sj}, \quad j = k, k+1, \dots, 3 \\ l_{ik} &= \left(a_{ik} - \sum_{s=1}^{k-1} l_{is}u_{sk} \right) / u_{kk}, \quad i = k+1, k+2, \dots, 3 \end{aligned}$$

- Cálculo de L e U:

$$\left| \begin{array}{c} \text{1a. linha de } U: \\ u_{11} = a_{11} \Longrightarrow u_{11} = 2 \\ u_{12} = a_{12} \Longrightarrow u_{12} = 3 \\ u_{13} = a_{13} \Longrightarrow u_{13} = -2 \end{array} \right| \left| \begin{array}{c} \text{1a. coluna de } L: \\ l_{21} = a_{21}/u_{11} = 2/2 = 1 \\ l_{31} = a_{31}/u_{11} = 14/2 = 7 \end{array} \right.$$

$$\left|\begin{array}{c} \text{2a. linha de } U: \\ u_{22} = a_{22} - l_{21}u_{12} \Longrightarrow u_{11} = -16 - 1 \times 3 = -19 \\ u_{23} = a_{23} - l_{21}u_{13} \Longrightarrow u_{23} = 10 - 1 \times -2 = 12 \end{array}\right| \left|\begin{array}{c} \text{2a. coluna de } L: \\ l_{32} = a_{32} - l_{31}u_{12} = \dfrac{-7 - 7 \times 3}{-19} \\ = 28/19 = 1.47368 \end{array}\right|$$

$$\left|\begin{array}{c} \text{3a. linha de } U: \\ u_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23} = -7 - (7 \times -2) - \left(\dfrac{28}{19} \times 12\right) \\ = -7 + 14 - 17.6842 = -10.68421 \end{array}\right|$$

Portanto,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 7 & 1.473681 & 1 \end{bmatrix} \text{ e } U = \begin{bmatrix} 2 & 3 & -2 \\ 0 & -19 & 12 \\ 0 & 0 & -10.68421 \end{bmatrix}$$

Solução dos sistemas lineares $L\mathbf{g} = \mathbf{b}$ e $U\mathbf{x} = \mathbf{g}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 7 & 1.473681 & 1 \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -3 \\ 0 \end{bmatrix} \Longrightarrow \begin{cases} g_1 = 0 \\ g_1 + g_2 = -3 \Longrightarrow g_2 = -3 \\ 7g_1 + 1.473681g_2 + g_3 = 0 \\ \Longrightarrow g_3 = -1.473681g_2 = 4.42104 \end{cases}$$

Portanto, $\mathbf{g} = [0 \ \ -3 \ \ 4.42104]^T$.

$$\begin{bmatrix} 2 & 3 & -2 \\ 0 & -19 & 12 \\ 0 & 0 & -10.68421 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -3 \\ 4.42104 \end{bmatrix}$$

$$\Longrightarrow \begin{cases} x_3 = \dfrac{4.42104}{-10.68421} = -0.413792 \\ -19x_2 + 12x_3 = -3 \Longrightarrow x_2 = \dfrac{-3 - 12x_3}{-19} \\ = \dfrac{-3 - 12 \times -0.413792}{-19} = -0.103448 \\ 2x_1 + 3x_2 - 2x_3 = 0 \Longrightarrow x_1 = (-3x_2 + 2x_3)/2 \\ \Longrightarrow x_1 = (-3 \times -0.103448 + 2 \times -0.413792)/2 = -0.2586206 \end{cases}$$

Método de Crout (ou Factorização de Crout)

Este método corresponde a impor que os elementos da diagonal de U sejam iguais a 1, ou seja, a decomposição é da forma:

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}}_{A \text{ (matriz dada)}} = \underbrace{\begin{bmatrix} l_{11} & 0 & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & \dots & l_{nn} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & u_{12} & \dots & \dots & u_{1n} \\ 0 & 1 & u_{23} & \dots & u_{2n} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}}_{U}$$

A obtenção das matrizes L e U segue o mesmo processo descrito na factorização de Doolittle, só que agora determina-se primeiro a coluna k de L e, em seguida, a linha k de U. As equações para a obtenção de L e U são:

- Coluna k de L: $l_{ik} = a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk},\ i = k, k+1, \dots\ n$
- Linha k de U $u_{kj} = \left(a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj} \right) / l_{kk},\ i = k, k+1, \dots\ n$

Método de Cholesky (ou Factorização de Cholesky)

No caso especial em que A é uma matriz real simétrica definida positiva, a decomposição LU pode ser modificada de maneira que $L = U^T$, ou seja, $A = LL^T$. As equações para obter a matriz L seguem o mesmo procedimento do método LU, ou seja,

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}}_{A \text{ (matriz dada)}} = \underbrace{\begin{bmatrix} l_{11} & 0 & \dots & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & \dots & l_{nn} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} l_{11} & l_{21} & \dots & \dots & l_{n1} \\ 0 & l_{22} & \dots & \dots & l_{n2} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & l_{nn} \end{bmatrix}}_{L^T}$$

Efectuando o produto LL^T e usando a igualdade $A = LU$, elemento a elemento, obtém-se as seguintes equações para a determinação das colunas de L:

$$\begin{aligned} l_{kk} &= \sqrt{a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2} \\ l_{ik} &= \left[a_{ik} - \sum_{r=1}^{k-1} l_{ir} l_{kr} \right] / l_{kk}\ , \quad i = k+1, k+2, \dots, n \end{aligned}$$

Cálculo da Matriz Inversa

Seja A uma matriz não-singular e A^{-1} sua inversa. Então,

$$AA^{-1} = I$$

Seja $\mathbf{x_j} = x_{1j}, x_{2j}, \ldots, x_{nj}$, a coluna j $(j = 1, ...n)$ de A^{-1}. Então temos:

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

Desta igualdade, vemos que o cálculo de A^{-1} equivale a resolver n sistemas lineares

$$\begin{cases} A\mathbf{x_1} = \mathbf{e_1},\ A\mathbf{x_2} = \mathbf{e_2},\ \ldots,\ A\mathbf{x_n} = \mathbf{e_n} \\ \text{onde } \mathbf{e_1} = [1\ 0\ \ldots\ 0]^T,\ \mathbf{e_2} = [0\ 1\ \ldots\ 0]^T,\ \mathbf{e_n} = [0\ \ldots\ 1]^T \end{cases}$$

para a determinação das colunas de A^{-1}. Se a matriz A for factorizada na forma $A = LU$, então faz-se

$(LU)\mathbf{x_1} = \mathbf{e_1},\ (LU)\mathbf{x_2} = \mathbf{e_2},\ \ldots, (LU)\mathbf{x_n} = \mathbf{e_n}$. Ou seja, faz-se a factorização LU uma única vez e resolvem-se os n sistemas lineares cujas soluções são as colunas de A^{-1}.

Técnicas de Pesquisa de Pivot

Devido aos erros de arredondamento, o método de eliminação de Gauss pode conduzir a soluções erróneas. Ou seja, podemos ter problemas de instabilidade numérica. As técnicas de pesquisa de pivot surgem numa tentativa de minorar o efeito da propagação dos erros de arredondamento.

Começamos por apresentar o seguinte exemplo.

Exemplo 3.3

$$\begin{aligned} E_1 &: 0.003000x_1 + 59.14x_2 = 59.17 \\ E_2 &: 5.291x_1 - 6.130x_2 = 47.78 \end{aligned} \tag{3.5}$$

que tem como solução exacta $x_1 = 10.0$ e $x_2 = 1.0$. Suponhamos um computador que usa um sistema $VF(10, 4, -10, 10)$ e arredondamento simétrico. Então, aplicando o método de eliminação de Gauss ao sistema linear (3.5), temos:

$$m_{21} = \frac{5.291}{0.003000} = 1763.66 \approx 0.1764 \times 10^4$$

e, efectuando a operação $E_2 - m_{21}E_1 \longrightarrow E_2$, obtém-se o sistema equivalente:

$$\begin{aligned} 0.003000x_1 + 59.14x_2 &= 59.17 \\ -104300x_2 &= -104400 \end{aligned} \tag{3.6}$$

que tem como solução:

$$x_2 = \frac{-0.1044 \times 10^6}{-0.1043 \times 10^6} = 1.00096 \approx 1.001 \quad \text{e} \quad x_1 = \frac{59.17 - (59.14 \times 1.001)}{0.003000} = -10.0$$

Vemos que um pequeno erro em x_2 : $|\delta x_2| = \dfrac{|1.0 - 1.001|}{1.0} = 0.001(0.1\ \%$ erro) resultou num grande erro em x_1 : $|\delta x_1| = \dfrac{|10.0 - (-10.0)|}{10.0} = 2($ ou seja, 200 % de erro)

Vejamos que não é indiferente a ordem das equações !

Por outro lado, se trocarmos as equações em (3.5) ($E_1 \longleftrightarrow E_2$), tem-se:

$$\begin{aligned} 5.291x_1 \quad -6.130x_2 \quad &= 46.78 \\ 0.00300x_1 \quad +59.14x_2 \quad &= 59.17 \end{aligned} \tag{3.7}$$

e aplicarmos o método de Eliminação de Gauss, obtemos:

$$m_{21} = \frac{0.003000}{5.291} = 0.0005670 \approx 0.0006$$

e a operação $E_2 - m_{21}E_1 \longrightarrow E_2$ conduz agora ao sistema equivalente:

$$\begin{array}{rcrl} 5.291x_1 & - & 6.130x_2 & = 46.78 \\ & & 59.14x_2 & = 59.14 \end{array}$$

que tem como solução $x_1 = 10.0$ e $x_2 = 1.0$!!

Pesquisa Parcial de Pivot

Esta técnica consiste em escolher, para elemento pivot, no k-passo do método de eliminação de Gauss, o elemento de maior valor absoluto na coluna k, ou seja,

$$\begin{cases} \text{Seja } a_{rk}^k \text{ tal que } |a_{rk}^k| = max\{|a_{ik}^k|, i = k, k+1, \ldots, n\} \\ \text{Então, se } r \neq k \text{ troca-se as linhas } E_r^k \text{ e } E_k^k. \end{cases}$$

Notemos que a aplicação desta técnica, no exemplo anterior, ao sistema (3.5) conduz ao sistema (3.7).

obs: Estudos feitos sobre o método de eliminação de Gauss mostram que se os multiplicadores m_{ik} são tais que $|m_{ik}| \leq 1$ então o efeito da propagação dos erros de arredondamento na solução é de algum modo reduzido. Esse é o objectivo da técnica de pesquisa parcial de pivot.

Pesquisa Total de Pivot

Esta técnica consiste em escolher, para elemento pivot, no k-passo do método de eliminação de Gauss, o elemento:

$$|a_{rs}^k| = max\{|a_{ij}^k|, i, j = k, k+1, \ldots, n\}$$

No exemplo anterior, conduziria ao sistema

$$\begin{array}{rcrl} 59.14x_1 & + & 0.00300x_2 & = 59.17 \\ -6.130x_1 & + & 5.291x_2 & = 46.78 \end{array}$$

A pesquisa total de pivot é a técnica que permite maior redução dos erros de arredondamento. Contudo ela requer um maior tempo computacional, sendo, por isso, mais dispendiosa.

Matrizes Especiais

No caso de certo tipo de matrizes, que consideraremos nesta secção, o método de eliminação de Gauss tem um comportamento estável, não sendo necessário recorrer ao uso de pesquisa de pivot.

Definição 3.1 *(Matriz de Diagonal Dominante): Uma matriz $A_{n\times n}$ é chamada "matriz de diagonal dominante" se:*

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \quad i = 1, 2, \ldots, n \text{ diagonal dominante por linhas}$$

$$|a_{jj}| \geq \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \quad j = 1, 2, \ldots, n \text{ diagonal dominante por colunas}$$

com desigualdade estrita $" >"$ válida para pelo menos um índice.

Definição 3.2 *(Matriz de Diagonal Estritamente Dominante): Se nas desigualdades acima o sinal $\geq$ for substituído por $" >"$ (ou seja, desigualdade estrita sempre) então dizemos que a matriz é de diagonal estritamente dominante por linhas (ou por colunas).*

Exemplo 3.4

Considere a matriz

$$A = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & 1 \\ 2 & -2 & 4 \end{bmatrix}$$

Tem-se: $|4| > |-1| + |0|$, $|4| > |-1| + |1|$ e $|4| = |2| + |2|$, donde vemos que esta matriz é de diagonal dominante por linhas, mas não é de diagonal estritamente dominante por linhas. Por outro lado, A é de diagonal estritamente dominante por colunas: $|4| > |-1| + |2|$, $|4| > |-1| + |-2|$ e $|4| > |1| + |0|$

Definição 3.3 *(Matriz Definida Positiva): Seja $A_{n\times n}$ uma matriz simétrica. Se $\mathbf{x}^T A\mathbf{x} > 0$ qualquer que seja o vector não nulo $\mathbf{x} = [x_1, x_2, ..., x_n]^T \in \mathbf{R^n}$, dizemos que A é uma matriz definida positiva.*

Na prática é mais fácil verificar o seguinte.

Teorema 3.1 *Uma matriz $A_{n\times n}$ simétrica é definida positiva se e só se a submatriz A_k, constituída pelas k primeiras linhas e k primeiras colunas de A, verifica:*

$$det(A_k) > 0, \quad k = 1, 2, ..., n.$$

Exemplo 3.5

Para a matriz

$$A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

tem-se

$$A_1 = [2], \quad A_2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad A_3 = A$$

e, sendo A simétrica, com $det(A_1) = 2$, $det(A_2) = 3 > 0$, $det(A_3) = det(A) = 4 > 0$, então A é definida positiva.

Observações

obs1: Se A é simétrica e os valores próprios de A são positivos então A é definida positiva.

obs2: Se A é definida positiva então os elementos da diagonal são positivos.

Teorema 3.2 *Seja A uma matriz de um dos dois tipos seguintes: i) simétrica definida positiva ou ii) de diagonal estritamente dominante por linhas ou por colunas. Então A é não singular e, além disso, o método de Eliminação de Gauss (também o método LU) pode ser aplicado ao sistema linear $A\mathbf{x} = \mathbf{b}$ sem troca de linhas. Ou seja, o processo é estável em relação à propagação dos erros de arredondamento, não sendo preciso usar nenhuma técnica de pesquisa de pivot.*

Tem-se ainda

Teorema 3.3 *Se A é uma matriz simétrica definida positiva então o método de Cholesky pode ser aplicado ao sistema linear $A\mathbf{x} = \mathbf{b}$. (existe uma única matriz triangular inferior L tal que $A = LL^T$).*

3.2 Normas vectoriais e matriciais

Normas vectoriais

Uma norma em $\mathbf{R^n}$ é uma função denotada por $\|\cdot\|$ com valores em $I\!R$, satisfazendo:

N1. $\|\mathbf{x}\| \geq 0,\ \forall\ \mathbf{x}\ \in\ \mathbf{R^n}$

N2. $\|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = 0$

N3. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|,\ \forall\ \alpha\ \in\ I\!R,\ \forall\ \mathbf{x}\ \in\ \mathbf{R^n}$

N4. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{x}\|,\ \forall\ \mathbf{x},\mathbf{y}\ \in\ \mathbf{R^n}$

Em $\mathbf{R^n}$ usaremos as seguintes normas:

I. $\|\mathbf{x}\|_\infty = \max\limits_{1\leq i\leq n}\{|x_i|\}$

II. $\|\mathbf{x}\|_1 = \sum\limits_{i=1}^{n}|x_i|$

III. $\|\mathbf{x}\|_2 = \left(\sum\limits_{i=1}^{n}|x_i|^2\right)^{1/2}$

Exemplo 3.6

seja $\mathbf{x} = [-1\ 2\ 4]^T$. Então:

$\|\mathbf{x}\|_\infty = max\,\{|-1|, |2|, |4|\} = 4$

$\|\mathbf{x}\|_1 = \sum\limits_{i=1}^{3}|x_i| = |-1| + |2| + |4| = 7$

$\|\mathbf{x}\|_2 = \left(\sum\limits_{i=1}^{3}|x_i|^2\right)^{1/2} = (|-1|^2 + |2|^2 + |4|^2)^{1/2} = \sqrt{21}$

Normas de matrizes

Uma norma matricial (ou de matrizes) é uma função definida no conjunto das matrizes quadradas reais com valores em $\mathbb{R}$ ($\|\cdot\| : \mathbf{R^n} \times \mathbf{R^n} \longrightarrow \mathbb{R}$) satisfazendo:

M1. $\|A\| \geq 0, \ \forall \ A$

M2. $\|A\| = 0 \Longleftrightarrow A = 0$

M3. $\|\alpha A\| = |\alpha|\|A\|, \ \forall \ \alpha \ \in \ \mathbb{R}, \ \forall \ A$

M4. $\|A + B\| \leq \|A\| + \|B\|, \ \forall \ A$ e B

M5. $\|AB\| \leq \|A\|\|B\|, \ \forall \ A$ e B

Embora uma norma matricial possa ser definida de várias maneiras, vamos considerar somente normas provenientes de normas de vectores.

Dada uma norma vectorial $\|\cdot\|$, a aplicação

$$\|A\| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \qquad \{ \text{ norma natural ou induzida } \} \tag{3.8}$$

satisfaz as condições **M1**-**M5** para normas de matrizes. Como vemos, a norma de matriz definida deste modo depende da norma de vector adoptada. Vejamos alguns casos particulares:

1. Consideremos a norma vectorial $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n}\{|x_i|\}$. Então tem-se:

   $\|A\|_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$. Nesse caso, pode-se provar que:

   $$\|A\|_\infty = \max_{1\leq i\leq n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\} \qquad \{ \text{ conhecida como "norma das linhas" } \}$$

   Exemplo 3.7

   $$\text{Se } A = \begin{bmatrix} 1 & 3 & -5 \\ 1 & 4 & 2 \\ -6 & 5 & 7 \end{bmatrix}, \ \|A\|_\infty = max \left\{ \begin{array}{l} |1| + |3| + |-5| = 8, \\ |1| + |4| + |2| = 7, \\ |-6| + |5| + |7| = 18 \end{array} \right\} = 18$$

2. Para a norma vectorial $\|\mathbf{x}\|_1 = \sum_{i=1}^{n}\{|x_i|\}$, tem-se:

$$\|A\|_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1}. \text{ e pode-se provar que:}$$

$$\|A\|_1 = \max_{1\leq j\leq n} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\} \qquad \{ \text{ conhecida como "norma das colunas" } \}$$

Exemplo 3.8

Para a matriz acima, tem-se: $\|A\|_1 = max \left\{ \begin{array}{l} |1|+|1|+|-6| = 8, \\ |3|+|4|+|5| = 12, \\ |-5|+|2|+|7| = 14 \end{array} \right\} = 14$

3. Para definirmos a norma $\|A\|_2$ precisamos de introduzir o conceito de raio espectral. Começamos por rever o seguinte.

Definição 3.4 *Definição: Os valores próprios de uma matriz $A_{n\times n}$ são as raízes da equação*

$$det(A - \lambda I) = 0$$

Exemplo 3.9 *Determine os valores próprios da matriz* $A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 2 & 1 \\ -1 & 0 & 0 \end{bmatrix}$

Solução:

$$A - \lambda I = \begin{bmatrix} 1-\lambda & 0 & 1 \\ 2 & 2-\lambda & 1 \\ -1 & 0 & -\lambda \end{bmatrix}$$

Então $det(A - \lambda I) = 0 \Longrightarrow (1-\lambda) \begin{vmatrix} 2-\lambda & 1 \\ 0 & -\lambda \end{vmatrix} + \begin{vmatrix} 2 & 2-\lambda \\ -1 & 0 \end{vmatrix} = 0$

$\Longrightarrow (1-\lambda)\left[(2-\lambda)(-\lambda)\right] + (2-\lambda) = 0$

$\Longrightarrow (2-\lambda)\left[-\lambda(1-\lambda)+1\right] = 0 \Longrightarrow (2-\lambda)\left[\lambda^2 - \lambda + 1\right] = 0$

$\Longrightarrow \lambda_1 = 2,\ \lambda_2 = \dfrac{1+\sqrt{3}i}{2},\ \lambda_3 = \dfrac{1-\sqrt{3}i}{2}$

Definição 3.5 *Seja A uma matriz quadrada. O raio espectral de A, denotado por $\rho(A)$, é definido por:*

$$\rho(A) = \max_{1\leq i\leq m} |\lambda_i| \quad \textit{onde } \lambda_i \quad \textit{é valor próprio de } A$$

Exemplo 3.10

Para a matriz acima tem-se:

$$\rho(A) = \max_{1 \leq i \leq 3} |\lambda_i| = max\{2,\ 1,\ 1\} = 2$$

À norma vectorial $\|\cdot\|_2$ está associada a norma matricial

$$\|A\|_2 = \left[\rho(AA^T)\right]^{1/2}$$

onde $\rho(AA^T)$ é o raio espectral da matriz produto de A por A^T.

Exemplo No caso da matriz anterior, pode verificar-se que os valores próprios de AA^T são: 1, $(11 - \sqrt{105})/2$, $(11 + \sqrt{105})/2$. Então $\rho(AA^T) = (11 + \sqrt{105})/2 \simeq 10.62348$ e $\|A\|_2 \simeq 3.25937$. Note que também se tem $\|A\|_\infty = 5$ e $\|A\|_1 = 4$. Por outro lado, obtivémos $\rho(A) = 2$, que é um valor inferior a qualquer das três normas obtidas.

Te-se a seguinte propriedade do raio espectral.

Teorema 3.4 *i) Qualquer que seja a norma matricial $\|\cdot\|$, induzida por uma norma vectorial, tem-se, para qualquer matriz A:*

$$\rho(A) \leq \|A\|$$

ii) Dada uma matriz A qualquer, para todo o $\epsilon > 0$, existe uma norma induzida $\|\cdot\|$, tal que:

$$\|A\| \leq \rho(A) + \epsilon$$

Ou seja, o raio espectral é o ínfimo de todas as normas induzidas (entre os valores $\rho(A)$ e $\rho(A) + \epsilon$ existe sempre uma norma de A)

3.3 Condicionamento de sistemas lineares

Notação: Seja $\mathbf{x}$ um vector e $\bar{\mathbf{x}}$ uma aproximação para $\mathbf{x}$:

$$\mathbf{e_x} = \mathbf{x} - \bar{\mathbf{x}} \quad \text{é chamado erro de } \bar{\mathbf{x}} \quad \text{(um vector)}$$

$$\|\mathbf{e_x}\| = \|\mathbf{x} - \bar{\mathbf{x}}\| \quad \text{é o erro absoluto de } \bar{\mathbf{x}}$$

$$\|\delta_{\mathbf{x}}\| = \frac{\|\mathbf{e_x}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{\|\mathbf{x}\|}, \text{ se } \mathbf{x} \neq \mathbf{0}, \quad \text{é o erro relativo de } \bar{\mathbf{x}}$$

Quando se pretende resolver um sistema

$$A\mathbf{x} = \mathbf{b}, \tag{3.9}$$

por vezes os elementos de A ou $\mathbf{b}$ (os dados) podem ter de ser arredondados. Também ao utilizarmos o método de eliminação de Gauss, os erros resultantes dos arredondamentos efectuados conduzem a erros na solução final.

Para avaliar até que ponto a solução $\mathbf{x}$ é sensível à propagação dos erros de arredondamento, podemos pensar que o sistema original é substituído por um outro sistema $\bar{A}\,\bar{\mathbf{x}} = \bar{\mathbf{b}}$, onde $\bar{A}$ e $\bar{\mathbf{b}}$ são aproximações de A e $\mathbf{b}$, resultantes de pequenas "perturbações" ou "mudanças" nos elementos de A e $\mathbf{b}$.

Interessa-nos saber se o sistema (3.9) é sensível a pequenas mudanças (perturbações) nos dados.

Dizemos, informalmente, que um sistema linear é *bem condicionado* se a pequenas "perturbações" nos dados correspondem pequenas "perturbações" na solução $\mathbf{x}$. Caso contrário, dizemos que é *mal condicionado*.

No seguinte exemplo estuda-se o efeito de uma pequena perturbação no vector $\mathbf{b}$.

Exemplo 3.11

Consideremos o sistema $A\mathbf{x} = \mathbf{b}$ dado por

$$\left.\begin{array}{rcrl} 1.0001x_1 & - & 2x_2 & = 3.0001 \\ x_1 & + & 2x_2 & = 3.0 \end{array}\right\} \begin{array}{l} \text{cuja solução exacta é} \\ x_1 = 1.0 \text{ e } x_2 = 1.0 \end{array}$$

Agora, se "perturbarmos" o lado direito:

$$\begin{array}{rcrl} 1.0001x_1 & - & 2x_2 & = 3.0003 \\ x_1 & + & 2x_2 & = 3.0 \end{array}$$

este sistema tem por solução: $x_1 = 3.0$ e $x_2 = 0.0$.

Quais foram as mudanças (relativas) em b e em x ?

Calculemos o erro de $\bar{\mathbf{b}}$ e o consequente erro na solução, usando, por exemplo a norma $\|\cdot\|_\infty$.

Tem-se, para o vector $\bar{\mathbf{b}}$,

$$\mathbf{e_b} = \mathbf{b} - \bar{\mathbf{b}} = [3.0001\ 3.0]^T - [3.0003\ 3.0]^T = [-0.0002\ 0.0]^T$$

o que levou a um erro na solução

$$\mathbf{e_x} = \mathbf{x} - \bar{\mathbf{x}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

Calculando os respectivos erros relativos na norma $\|\cdot\|_\infty$ temos:

$$\|\delta_\mathbf{b}\| = \frac{\|\mathbf{e}_b\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{0.0002}{3.0001} \approx 0.000067 \approx 0.007\%$$

e

$$\|\delta_\mathbf{x}\| = \frac{\|\mathbf{e}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{2}{1} = 200\%$$

Assim vemos que uma pequena perturbação relativa em $\mathbf{b}$ conduziu a uma grande perturbação relativa na solução $\mathbf{x}$. O sistema não é bem condicionado. Trata-se dum sistema mal condicionado.

Como identificar um sistema mal condicionado ?

Definição 3.6 *Seja A uma matriz não-singular. O número de condição de A é definido por:*

$$cond(A) = \|A\|\ \|A^{-1}\|, \text{ onde } \|\cdot\| \text{ é uma norma natural de matrizes.}$$

Note-se que o número de condição $cond(A)$ depende da norma utilizada, mas, para as normas induzidas, é sempre maior ou igual a um.

Previsão do erro na solução do sistema

Teorema 3.5 *Seja* $\mathbf{x}$ *a solução exacta do sistema* $A\mathbf{x} = \mathbf{b}$ *e* $\bar{\mathbf{x}}$ *a solução obtida do sistema perturbado* $A\bar{\mathbf{x}} = \bar{\mathbf{b}}$*. Então,*

$$\frac{1}{cond(A)} \frac{\|\mathbf{b} - \bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq cond(A) \frac{\|\mathbf{b} - \bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \tag{3.10}$$

ou

$$\frac{1}{cond(A)} \|\delta_{\mathbf{b}}\| \leq \|\delta_{\mathbf{x}}\| \leq cond(A)\ \|\delta_{\mathbf{b}}\| \tag{3.11}$$

Prova: de $A\mathbf{x} = \mathbf{b}$ e $A\bar{\mathbf{x}} = \bar{\mathbf{b}}$, obtém-se:

$$A\mathbf{x} - A\bar{\mathbf{x}} = \mathbf{b} - \bar{\mathbf{b}} \implies A(\mathbf{x} - \bar{\mathbf{x}}) = \mathbf{b} - \bar{\mathbf{b}} \implies \mathbf{x} - \bar{\mathbf{x}} = A^{-1}(\mathbf{b} - \bar{\mathbf{b}}) \tag{3.12}$$

$$\implies \|\mathbf{x} - \bar{\mathbf{x}}\| = \|A^{-1}(\mathbf{b} - \bar{\mathbf{b}})\| \leq \|A^{-1}\|\ \|\mathbf{b} - \bar{\mathbf{b}}\|$$

$$\implies \|\mathbf{x} - \bar{\mathbf{x}}\| \leq \|A^{-1}\|\ \|\mathbf{b} - \bar{\mathbf{b}}\| \tag{3.13}$$

Por outro lado, $A\mathbf{x} = \mathbf{b} \implies \|\mathbf{b}\| = \|A\mathbf{x}\| \leq \|A\|\ \|\mathbf{x}\|$ donde se tem

$$\frac{1}{\|\mathbf{x}\|} \leq \frac{\|A\|}{\|\mathbf{b}\|} \tag{3.14}$$

Multiplicando (3.13) e (3.14) tem-se:

$$\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A^{-1}\|\ \|A\| \frac{\|\mathbf{b} - \bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \tag{3.15}$$

Agora, de (3.12) tem-se que:

$$\|\mathbf{b} - \bar{\mathbf{b}}\| = \|A(\mathbf{x} - \bar{\mathbf{x}})\| \leq \|A\|\ \|\mathbf{x} - \bar{\mathbf{x}}\| \implies \|\mathbf{x} - \bar{\mathbf{x}}\| \geq \frac{\|\mathbf{b} - \bar{\mathbf{b}}\|}{\|A\|} \tag{3.16}$$

e, dividindo por $||\mathbf{x}||$, obtém-se

$$\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|}{||\mathbf{x}||} \geq \frac{\|\mathbf{b} - \bar{\mathbf{b}}\|}{\|A\|} \frac{1}{||\mathbf{x}||} \tag{3.17}$$

Por outro lado, de $A\mathbf{x} = \mathbf{b}$ vem:

$$\mathbf{x} = A^{-1}\mathbf{b} \implies \|\mathbf{x}\| = \|A^{-1}\mathbf{b}\| \leq \|A^{-1}\|\ \|\mathbf{b}\|$$

donde:

$$\frac{1}{\|\mathbf{x}\|} \geq \frac{1}{\|A^{-1}\|\ \|\mathbf{b}\|} \tag{3.18}$$

Conjugando (3.17) por (3.18), vem:

$$\frac{1}{\|A\|\ \|A^{-1}\|}\frac{\|\mathbf{b}-\bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{x}-\bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \tag{3.19}$$

Finalmente, das equações (3.15) e (3.19) resulta:

$$\frac{1}{\|A\|\ \|A^{-1}\|}\frac{\|\mathbf{b}-\bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{x}-\bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A^{-1}\|\ \|A\|\frac{\|\mathbf{b}-\bar{\mathbf{b}}\|}{\|\mathbf{b}\|}$$

ou

$$\frac{1}{cond(A)}\|\delta_{\mathbf{b}}\| \leq \|\delta_{\mathbf{x}}\| \leq cond(A)\|\delta_{\mathbf{b}}\|$$

Observações: Comecemos por notar que o resultado do teorema anterior compara o erro relativo de $\bar{\mathbf{b}}$ com o erro relativo de $\bar{\mathbf{x}}$, sendo independente do método utilizado para resolver o sistema.

Examinemos a desigualdade que nos dá uma majoração para o erro:

$$\|\delta_{\mathbf{x}}\| \leq cond(A)\ \|\delta_{\mathbf{b}}\|$$

1. Se $cond(A) \approx 1$ (*pequeno*) então, se $\|\delta_{\mathbf{b}}\|$ for pequeno, vem que $\|\delta_{\mathbf{x}}\|$ tambem é pequeno. Neste caso, o sistema é um *sistema bem condicionado*!

2. Se $cond(A) >>> 1$ (*grande*) então $\|\delta_{\mathbf{b}}\|$ pequeno $\not\Longrightarrow$ $\|\delta_{\mathbf{x}}\|$ pequeno. Ou seja, pode haver situações em que resulte um $\|\delta_{\mathbf{x}}\|$ pequeno (ver exemplo abaixo) e outras em que o erro relativo de $\bar{\mathbf{x}}$ possa ser muito maior do que o de $\bar{\mathbf{b}}$. Na verdade é garantido que essas situaes vo-se concretizar para certas escolhas de $\mathbf{b}$ e $\bar{\mathbf{b}}$, só que na prática não se sabe quais são essas escolhas. Um número $cond(A)$ muito grande traduz-se numa grande sensibilidade do sistema a pequenas mudanças relativas em $\mathbf{b}$, ou, por outras palavras, o sistema é *mal condicionado*.

Exemplo 3.12

Para o exemplo anterior temos:

$$A = \begin{bmatrix} 1.0001 & 2 \\ 1 & 2 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 10000 & -10000 \\ -5000 & 5000.5 \end{bmatrix}$$

donde vemos que: $\|A\|_\infty = 3.0001$ e $\|A^{-1}\|_\infty = 20000$. Portando,

$$cond(A)_\infty = 60002 \ >>> 1$$

Previsão:

Usemos a fórmula de majoração, concretizada para a norma $\|\cdot\|_\infty$:

$$\|\delta_{\mathbf{x}}\|_\infty \leq cond_\infty(A)\ \|\delta_{\mathbf{b}}\|_\infty$$

$$\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \leq 60002\frac{\|\mathbf{b}-\bar{\mathbf{b}}\|}{\|\mathbf{b}\|} \simeq 4.02, \quad \text{ou seja} \quad \simeq 400\%$$

Pode concluir-se para este exemplo: quando o erro relativo em $\bar{\mathbf{b}}$ é 0.0067, o erro relativo na solução pode ir até 400%. Na verdade, já tínhamos calculado $\|\delta_{\mathbf{x}}\| = 2$ ($\simeq 200\%$), o que está de acordo com a previsão (que dá um majorante para o erro).

Para melhor ilustrar a incerteza inerente ao mau condicionamento de sistemas, incluímos ainda o seguinte exemplo.

Consideremos de novo o mesmo sistema, mas sujeito a uma perturbação em $\mathbf{b}$ diferente da considerada. Assim, seja o sistema $A\bar{\mathbf{x}} = \bar{\mathbf{b}}$ dado por:

$$\begin{array}{rcrcl} 1.0001\bar{x}_1 & + & 2\bar{x}_2 & = & 3.0011 \\ \bar{x}_1 & + & 2\bar{x}_2 & = & 3.001 \end{array}$$

este sistema tem por solução: $\bar{x}_1 = 1.$ e $\bar{x}_2 = 1.0005$.

É fácil de verificar que $\|\delta_{\mathbf{b}}\| \simeq 0.33\,10^{-1}\%$ e $\|\delta_{\mathbf{x}}\| \simeq 0.5\,10^{-1}\%$. Ou seja, agora a uma pequena perturbação relativa em $\mathbf{b}$ correspondeu também uma pequena perturbação relativa em $\mathbf{x}$.

As observações acima sobre mudanças no vector $\mathbf{b}$ também são válidas quando há mudanças nos elementos da matriz A. Em particular, se compararmos as soluções de $A\mathbf{x} = \mathbf{b}$ e $\bar{A}\bar{\mathbf{x}} = \mathbf{b}$, tem-se, no caso em que o erro relativo de $\bar{A}$ é suficientemente pequeno:

$$\frac{\|\mathbf{x}-\bar{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{cond(A)}{1 - cond(A)\frac{\|A-\bar{A}\|}{\|A\|}}\frac{\|A-\bar{\mathbf{A}}\|}{\|\mathbf{A}\|} \tag{3.20}$$

supondo que $\|A-\bar{\mathbf{A}}\|/\|\mathbf{A}\| < 1/\|A^{-1}\|$

De novo pequenas mudanças em A podem levar a grandes mudanças em $\mathbf{x}$, se $cond(A)$ for grande. Em geral, o software para resolver sistemas lineares tem incorporados métodos para se estimar $cond(A)$, sem ter de se calcular A^{-1}. Quando se suspeita que a matriz é mal condicionada, pode ser usada uma técnica (método da correcção residual) para melhorar a solução aproximada obtida.

Concluímos esta secção com a seguinte observação. Como vimos, as técnicas de pesquisa de pivot são utilizadas para minorar as dificuldades resultantes da propagação dos erros de

arredondamento. Contudo, no caso de sistemas mal condicionados, a sua utilização não vai em princípio melhorar grandemente os resultados, já que o mau condicionamento resulta sempre em instabilidade numérica.

3.4 Métodos iterativos para sistemas lineares

Introdução

Num método iterativo para resolver um sistema de equações lineares $A\mathbf{x} = \mathbf{b}$, começa-se com uma aproximação inicial $\mathbf{x}^{(0)}$ da solução $\mathbf{x}$ e é gerada uma sucessão de aproximações $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$. Os métodos iterativos são em geral usados para sistemas de grande dimensão (por exemplo, se A é da ordem de 10000) onde A tem uma grande percentagem de elementos nulos (matriz esparsa). Sistemas deste tipo surgem por exemplo depois da aplicação de métodos numéricos a problemas de valores na fronteira e equações com derivadas parciais.

Relembremos que, no capítulo 2, para as equações da forma $f(x) = 0$ obtiveram-se métodos iterativos do ponto fixo $x_{k+1} = g(x_k)$, depois de se reescrever a equação dada $f(x) = 0$ na forma $x = g(x)$.

No caso dum sistema linear $A\mathbf{x} = \mathbf{b}$, que é uma equação da forma

$$\underbrace{A\mathbf{x} - \mathbf{b}}_{F(x)} = \mathbf{0}, \tag{3.21}$$

também começamos por obter uma equivalência $\mathbf{x} = G(\mathbf{x})$. Como veremos na secção seguinte, em geral isso pode ser feito através duma decomposição da matriz A, a qual permite reescrever o sistema numa forma equivalente

$$A\mathbf{x} = \mathbf{b} \Longleftrightarrow \mathbf{x} = \underbrace{C\mathbf{x} + \mathbf{d}}_{G(\mathbf{x})}, \tag{3.22}$$

onde C é uma certa matriz apropriada e $\mathbf{d}$ um vector. O método iterativo associado a G será então:

$$\mathbf{x}^{(k+1)} = G(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, ... \tag{3.23}$$

ou,usando (3.22),

$$\mathbf{x}^{(k+1)} = C\mathbf{x}^{(k)} + \mathbf{d}, \quad k = 0, 1, 2, ... \tag{3.24}$$

Seja

$$\mathbf{x}^{(0)} = [x_1^{(0)} \; x_2^{(0)} \; x_3^{(0)} \; \ldots x_n^{(0)}]^T$$

um vector aproximação inicial para

$$\mathbf{x} = [x_1 \; x_2 \; x_3 \ldots x_n]^T,$$

onde

$$\begin{array}{l} x_1^{(0)} \text{é aprox. para } x_1, \\ x_2^{(0)} \text{é aprox. para } x_2, \\ ..... \\ x_n^{(0)} \text{é aprox. para } x_n \end{array}$$

Então construímos aproximações:

$$\begin{aligned}
\mathbf{x}^{(1)} &= [x_1^{(1)}\ x_2^{(1)}\ x_3^{(1)}\ \ldots x_n^{(1)}]^T \\
\mathbf{x}^{(2)} &= [x_1^{(2)}\ x_2^{(2)}\ x_3^{(2)}\ \ldots x_n^{(2)}]^T \\
&\vdots
\end{aligned}$$

aplicando sucessivamente a fórmula (3.24), ou seja, fazendo:

$$\begin{aligned}
\mathbf{x}^{(1)} &= C\mathbf{x}^{(0)} + \mathbf{d} \\
\mathbf{x}^{(2)} &= C\mathbf{x}^{(1)} + \mathbf{d} \\
\mathbf{x}^{(3)} &= C\mathbf{x}^{(2)} + \mathbf{d} \\
&\vdots
\end{aligned}$$

O objectivo é determinar aproximações (iteradas) até se chegar suficientemente próximo de $\mathbf{x}$. Por outras palavras, interessa-nos que a sucessão de aproximações $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(k)}, ...$ seja convergente para $\mathbf{x}$:

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x} \tag{3.25}$$

Ou, doutro modo, interessa-nos que as "distâncias" (medidas por intermédio duma norma) de $\mathbf{x}$ às sucessivas iteradas $\mathbf{x}^{(k)}$ tenda para zero:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0 \tag{3.26}$$

Concluímos esta introdução relembrando que, para as equações $f(x) = 0$, diferentes funções g geravam métodos do ponto fixo (ou sucessões) diferentes. No caso dum sistema, também há diferentes maneiras de reescrevê-lo na forma (3.22), ou seja, são possíveis diferentes escolhas da matriz C (e consequentemente do vector $\mathbf{d}$), por forma a que (3.21) seja equivalente a (3.22). Essas diferentes escolhas conduzem a métodos diferentes para a equação matricial dada. Neste capítulo estudaremos, em particular, os métodos de Jacobi e Gauss-Seidel.

3.4.1 Os Métodos de Jacobi e Gauss-Seidel

Vamos começar por apresentar dois métodos iterativos, utilizando um exemplo concreto.

Exemplo 3.13 *Seja o sistema*

$$\begin{bmatrix} 10 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -8 \\ 6 \end{bmatrix} \quad \begin{bmatrix} \textit{que tem solução exacta:} \\ x_1 = 1,\ x_2 = -2 \\ \textit{e } x_3 = 1. \end{bmatrix}$$

Em forma explícita, temos:

$$\begin{cases} 10x_1 + 2x_2 + x_3 &= 7 \\ x_1 + 5x_2 + x_3 &= -8 \\ 2x_1 + 3x_2 + 10x_3 &= 6 \end{cases} \tag{3.27}$$

O Método de Jacobi

Resolvendo a primeira equação de (3.27) em ordem a x_1, a segunda em ordem a x_2 e a terceira em ordem a x_3, obtém-se o sistema equivalente

$$\begin{array}{rcrcrcl} x_1 &=& (7 &-& 2x_2 &-& x_3)/10 \\ x_2 &=& (-8 &-& x_1 &-& x_3)/5 \\ x_3 &=& (6 &-& 2x_1 &-& 3x_2)/10 \end{array} \tag{3.28}$$

Sendo $\mathbf{x} = [x_1\ x_2\ x_3]^T$ e designando por $g_1(x_1, x_2, x_3)$, $g_2(x_1, x_2, x_3)$, $g_3(x_1, x_2, x_3)$, respectivamente, as expressões do lado direito de (3.28), temos, equivalentemente,

$$\begin{array}{l} x_1 = g_1(x_1, x_2, x_3) \\ x_2 = g_2(x_1, x_2, x_3) \\ x_3 = g_3(x_1, x_2, x_3) \end{array} \quad \Longleftrightarrow \mathbf{x} = G(\mathbf{x}) \tag{3.29}$$

Em seguida associamos a G o método do ponto fixo $\mathbf{x}^{(k+1)} = G(\mathbf{x}^{(k)})$. Isto é, no lado esquerdo das equações substitui-se cada x_i por $x_i^{(k+1)}$ e, no lado direito, por $x_i^{(k)}$. Obtém-se assim o chamado método de Jacobi, definido por

$$\begin{array}{rcrcrcl} x_1^{(k+1)} &=& (7 &-& 2x_2^{(k)} &-& x_3^{(k)})/10 \quad = 0.7 - 0.2x_2^{(k)} - 0.1x_3^{(k)} \\ x_2^{(k+1)} &=& (-8 &-& x_1^{(k)} &-& x_3^{(k)})/5 \quad = -1.6 - 0.2x_1^{(k)} - 0.2x_3^{(k)} \\ x_3^{(k+1)} &=& (6 &-& 2x_1^{(k)} &-& 3x_2^{(k)})/10 \quad = 0.6 - 0.2x_1^{(k)} - 0.3x_2^{(k)}, \end{array} \tag{3.30}$$

com $k = 0, 1, 2, ...$ Suponhamos que se começa com uma aproximação inicial $\mathbf{x}^{(0)} = [0.7\ \ -1.6\ \ 0.6]^T$. Para obter as iteradas seguintes, utiliza-se (3.30), dando-se sucessivos valores a k:

$k = 0$ (cálculo de $\mathbf{x}^{(1)}$)

$$x_1^{(1)} = 0.7 - 0.2x_2^{(0)} - 0.1x_3^{(0)} = -0.2 \times (-1.6) - 0.1 \times (0.6) + 0.7 = 0.9600$$
$$x_2^{(1)} = -1.6 - 0.2x_1^{(0)} - 0.2x_3^{(0)} = -0.20 \times (0.7) - 0.2 \times (0.6) - 1.6 = -1.8600$$
$$x_3^{(1)} = 0.6 - 0.2x_1^{(0)} - 0.3x_2^{(0)} = -0.20 \times (0.7) - 0.3 \times (-1.6) + 0.6 = 0.9400$$

$k = 1$ (cálculo de $\mathbf{x}^{(2)}$)

$$x_1^{(2)} = 0.7 - 0.2x_2^{(1)} - 0.1x_3^{(1)} = -0.2 \times (-1.86) - 0.1 \times (0.94) + 0.7 = 0.9780$$
$$x_2^{(2)} = -1.6 - 0.2x_1^{(1)} - 0.2x_3^{(1)} = -0.20 \times (0.96) - 0.2 \times (0.94) - 1.6 = -1.9800$$
$$x_3^{(2)} = 0.6 - 0.2x_1^{(1)} - 0.3x_2^{(1)} = -0.20 \times (0.96) - 0.3 \times (-1.86) + 0.6 = 0.9660$$

$k = 2$ (cálculo de $\mathbf{x}^{(3)}$)

$$x_1^{(3)} = 0.7 - 0.2x_2^{(2)} - 0.1x_3^{(2)} = -0.2 \times (-1.98) - 0.1 \times (0.966) + 0.7 = 0.9994$$
$$x_2^{(3)} = -1.6 - 0.2x_1^{(2)} - 0.2x_3^{(2)} = -0.20 \times (0.978) - 0.2 \times (0.966) - 1.6 = -1.9988$$
$$x_3^{(3)} = 0.6 - 0.2x_1^{(2)} - 0.3x_2^{(2)} = -0.20 \times (0.978) - 0.3 \times (-1.98) + 0.6 = 0.9984$$

O Método de Gauss-Seidel

O método de Gauss-Seidel pode ser considerado como uma modificação do método de Jacobi. No exemplo que estamos a tratar, observemos o seguinte.

É dada a aproximação inicial $\mathbf{x}^{(0)} = [x_1^{(0)}\ x_2^{(0)}\ x_3^{(0)}]^T$. No cálculo da iterada seguinte, $\mathbf{x}^{(1)}$, pelo método de Jacobi, começa-se por obter a primeira componente $x_1^{(1)}$. Em seguida, para determinar a componente $x_2^{(1)}$, são usados $x_1^{(0)}, x_3^{(0)}$. Ora, no método de Gauss-Seidel, substitui-se $x_1^{(0)}$ por $x_1^{(1)}$, que já se conhece nesta altura. Também no cálculo de $x_3^{(1)}$, o método de Gauss-Seidel usa as aproximações mais recentes $x_1^{(1)}, x_2^{(1)}$, em vez de $x_1^{(0)}, x_2^{(0)}$, respectivamente. Assim, para o método de Gauss-seidel, viria

(expressão de $\mathbf{x}^{(1)}$)

$$x_1^{(1)} = 0.7 - 0.2x_2^{(0)} - 0.1x_3^{(0)} = -0.2 \times (-1.6) - 0.1 \times (0.6) + 0.7 = 0.9600$$
$$x_2^{(1)} = -1.6 - 0.2x_1^{(1)} - 0.2x_3^{(0)} = -0.20 \times (0.9600) - 0.2 \times (0.6) - 1.6 = -1.912$$
$$x_3^{(1)} = 0.6 - 0.2x_1^{(1)} - 0.3x_2^{(1)} = -0.20 \times (0.9600) - 0.3 \times (-1.912) + 0.6 = 0.9816$$

Passemos a deduzir a *fórmula geral* para obtenção de $\mathbf{x}^{(k+1)}$, a partir da expressão (3.30) do método de Jacobi. O método de Gauss-Seidel resulta de fazer algumas modificações:

A expressão da componente $x_1^{(k+1)}$ fica como no método de Jacobi. Na expressão da componente $x_2^{(k+1)}$, substitui-se $x_1^{(k)}$ por $x_1^{(k+1)}$. Na expressão da componente $x_3^{(k+1)}$ substitui-se $x_1^{(k)}$ por $x_1^{(k+1)}$ e $x_2^{(k)}$ por $x_2^{(k+1)}$.

Com base nessas observações, o método de Gauss-Seidel para o exemplo dado é o método iterativo definido por:

$$\begin{array}{rcl} x_1^{(k+1)} & = & (7 - 2x_2^{(k)} - x_3^{(k)})/10 = 0.7 - 0.2x_2^{(k)} - 0.1x_3^{(k)} \\ x_2^{(k+1)} & = & (-8 - x_1^{(k+1)} - x_3^{(k)})/5 = -1.6 - 0.2x_1^{(k+1)} - 0.2x_3^{(k)} \\ x_3^{(k+1)} & = & (6 - 2x_1^{(k+1)} - 3x_2^{(k+1)})/10 = 0.6 - 0.2x_1^{(k+1)} - 0.3x_2^{(k+1)}, \end{array} \tag{3.31}$$

com $k = 0, 1, 2, ...$

Vemos que o método de Gauss-Seidel se distingue do método de Jacobi no facto de utilizar os novos elementos à medida que eles vão sendo calculados. Embora pareça natural pensar que, no caso de ambos convergirem, o método de Gauss-Seidel é "mais rápido" do que o método de Jacobi (no sentido de "chegar próximo" da solução exacta num menor número de iterações), isso nem sempre acontece. Há mesmo casos em que o método de Jacobi converge e o de Gauss-Seidel diverge.

Começámos por apresentar os métodos de Jacobi e Gauss-Seidel duma maneira directa e bastante simples. Contudo, para o estudo da convergência desses métodos, é conveniente recorrer à sua formulação matricial. Isso é o que faremos na secção seguinte, mas para já ilustramos com um exemplo.

Exemplo 3.14 *Consideremos de novo o exemplo (3.13) e mostremos que o método de Jacobi (3.30) é da forma*

$$\mathbf{x}^{(k+1)} = C_J\ \mathbf{x}^{(k)} + \mathbf{d}.$$

Com efeito, reescrevendo a igualdade (3.30) na forma matricial, vem:

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{bmatrix} = \begin{bmatrix} 0 & -0.2 & -0.1 \\ -0.2 & 0 & -0.2 \\ -0.2 & -0.3 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \end{bmatrix} + \begin{bmatrix} 0.7 \\ -1.6 \\ 0.6 \end{bmatrix} \Longleftrightarrow \mathbf{x}^{(k+1)} = C_j\ \mathbf{x}^{(k)} + \mathbf{d}$$

À matriz C_J acima chamamos matriz de iteração do método de Jacobi.

Exemplo 3.15 *Verifique que também é possível expressar o método de Gauss-Seidel (3.31) numa forma $\mathbf{x}^{(k+1)} = C_{GS}\ \mathbf{x}^{(k)} + \mathbf{d}$, onde C_{GS} é uma matriz e $\mathbf{d}$ um vector adequado.*

Veremos que as matrizes C_J e C_{GS} têm um papel importante na convergência dos métodos.

3.4.2 Como construir métodos iterativos: decomposição matricial

Dado o sistema linear $A\mathbf{x} = \mathbf{b}$, vamos agora descrever uma maneira de se obterem métodos iterativos da forma

$$\mathbf{x}^{(k+1)} = C\ \mathbf{x}^{(k)} + \mathbf{d}$$

e, em particular, os métodos de Jacobi e Gauss-Seidel. Começamos por decompor a matriz A na soma de outras duas:

$$A = M + N, \text{ onde } M \text{ é uma matriz invertível.} \tag{3.32}$$

donde resulta

$$A\mathbf{x} = \mathbf{b} \iff (M+N)\mathbf{x} = \mathbf{b} \tag{3.33}$$

Desenvolvendo a igualdade da direita obtém-se sucessivamente

$$A\mathbf{x} = \mathbf{b} \iff M\mathbf{x} = \mathbf{b} - N\mathbf{x} \Longleftrightarrow \mathbf{x} = \underbrace{-M^{-1}N}_{C}\ \mathbf{x} + \underbrace{M^{-1}\mathbf{b}}_{\mathbf{d}} \tag{3.34}$$

Quer dizer, obteve-se a equivalência

$$A\mathbf{x} = \mathbf{b} \iff \mathbf{x} = C\mathbf{x} + \mathbf{d} \tag{3.35}$$

e somos naturalmente conduzidos a métodos iterativos da forma

$$\mathbf{x}^{(k+1)} = C\,\mathbf{x}^{(k)} + \mathbf{d}, \tag{3.36}$$

ou seja,

$$\mathbf{x}^{(k+1)} = \underbrace{-M^{-1}N}_{C}\ \mathbf{x}^{(k)} + M^{-1}\mathbf{b} \tag{3.37}$$

k=0,1,2,... , onde $\mathbf{x}^{(0)}$ é um vector dado (aproximação inicial).

Diferentes escolhas das matrizes M, N resultam em diferentes "rearranjos" do sistema na forma $\mathbf{x} = C\,\mathbf{x} + \mathbf{d}$, que, por sua vez, conduzem a diferentes métodos iterativos. A matriz C é chamada matriz de iteração. Designaremos as matrizes de iteração dos métodos de Jacobi e Gauss-Seidel por C_J e C_{GS}, respectivamente.

O que são as matrizes M, N no caso dos métodos de Jacobi e Gauss-Seidel?

A decomposição A=L+D+U

Nos métodos de Jacobi e Gauss-Seidel a escolha das matrizes M e N é baseada na igualdade $A = L + D + U$, que a seguir descrevemos.

Ao sistema linear $A\mathbf{x} = \mathbf{b}$, da forma:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

associamos as matrizes L, U, D definidas do seguinte modo:

$$L = \begin{bmatrix} 0 & 0 & \dots & \dots & 0 \\ a_{21} & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ a_{n1} & a_{n2} & \dots & a_{nn-1} & 0 \end{bmatrix}; U = \begin{bmatrix} 0 & a_{12} & \dots & \dots & a_{1n} \\ 0 & 0 & a_{23} & \dots & a_{2n} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1n} \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}; D = \begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & a_{nn} \end{bmatrix}$$

Então, tem-se:

$$A = L + D + U \tag{3.38}$$

3.4.3 Formulação matricial dos métodos de Jacobi e Gauss-Seidel

Método de Jacobi

Com base na igualdade (3.38), vem:

$$A\mathbf{x} = \mathbf{bx} \Longleftrightarrow (\underbrace{D}_{M} + \underbrace{L+U}_{N})\mathbf{x} = \mathbf{b} \tag{3.39}$$

Substituindo $M = D$ e $N = L + U$ em (3.37) resulta o método de Jacobi:

$$\mathbf{x}^{(k+1)} = \underbrace{-D^{-1}(L+U)}_{C_J}\mathbf{x}^{(k)} + D^{-1}\mathbf{b}\ , \qquad \mathbf{x}^{(0)} \text{ dado.}$$

Ou seja, o método de Jacobi é um método iterativo da forma (3.37) com: $M = D$ e $N = L+U$ e, consequentemente, matriz de iteração $C_J = -D^{-1}(L + U)$. Estamos a supor que os elementos da diagonal de D são diferentes de zero.

Exemplo 3.16 *Seja o exemplo 3.13 já considerado:* $\begin{bmatrix} 10 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -8 \\ 6 \end{bmatrix}$

Vem

$$M = D = \begin{bmatrix} 10 & & \\ & 5 & \\ & & 10 \end{bmatrix}, \qquad N = L + U = \begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 3 & 0 \end{bmatrix}$$

e a matriz de iteração do método de Jacobi é:

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 1/10 & & \\ & 1/5 & \\ & & 1/10 \end{bmatrix} \begin{bmatrix} 0 & -2 & -1 \\ -1 & 0 & -1 \\ -2 & -3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.2 & -0.1 \\ -0.2 & 0 & -0.2 \\ -0.2 & -0.3 & 0 \end{bmatrix}$$

que é de facto a matriz já obtida no exemplo 3.14.

Dedução da expressão geral do método de Jacobi

É útil deduzir a expressão geral do método de Jacobi, para efeitos de implementação do método no computador.

Da igualdade $D\mathbf{x}^{(k+1)} = -(L+U)\mathbf{x}^{(k)} + \mathbf{b}$ obtemos as equações

$$\begin{array}{rcrcccccccccc}
a_{11}x_1^{(k+1)} & = & - & \left(a_{12}x_2^{(k)} & + & a_{13}x_3^{(k)} & + & a_{14}x_4^{(k)} & + & \ldots & + & a_{1n}x_n^{(k)}\right) & + & b_1 \\
a_{22}x_2^{(k+1)} & = & - & \left(a_{21}x_1^{(k)} & + & a_{23}x_3^{(k)} & + & a_{24}x_4^{(k)} & + & \ldots & + & a_{2n}x_n^{(k)}\right) & + & b_2 \\
a_{33}x_3^{(k+1)} & = & - & \left(a_{31}x_1^{(k)} & + & a_{32}x_2^{(k)} & + & a_{34}x_4^{(k)} & + & \ldots & + & a_{3n}x_n^{(k)}\right) & + & b_3 \\
\vdots & = & & \vdots & + & \vdots & + & \vdots & + & \ldots & + & \vdots & + & \vdots \\
a_{nn}x_3^{(k+1)} & = & - & \left(a_{n1}x_1^{(k)} & + & a_{n2}x_2^{(k)} & + & a_{n3}x_3^{(k)} & + & \ldots & + & a_{nn-1}x_{n-1}^{(k)}\right) & + & b_n
\end{array}$$

e portanto,

$$\begin{array}{rcccccccccc}
x_1^{(k+1)} & = & \left[b_1 - \left(a_{12}x_2^{(k)} \right. & + & a_{13}x_3^{(k)} & + & a_{14}x_4^{(k)} & + & \ldots & + & \left. a_{1n}x_n^{(k)}\right)\right]/a_{11} \\
x_2^{(k+1)} & = & \left[b_2 - \left(a_{21}x_1^{(k)} \right. & + & a_{23}x_3^{(k)} & + & a_{24}x_4^{(k)} & + & \ldots & + & \left. a_{2n}x_n^{(k)}\right)\right]/a_{22} \\
x_3^{(k+1)} & = & \left[b_3 - \left(a_{31}x_1^{(k)} \right. & + & a_{32}x_2^{(k)} & + & a_{34}x_4^{(k)} & + & \ldots & + & \left. a_{3n}x_n^{(k)}\right)\right]/a_{33} \\
\vdots & & \vdots & & \vdots & & \vdots & & & & \vdots \\
x_3^{(k+1)} & = & \left[b_n - \left(a_{n1}x_2^{(k)} \right. & + & a_{n2}x_2^{(k)} & + & a_{n3}x_3^{(k)} & + & \ldots & + & \left. a_{nn-1}x_{n-1}^{(k)}\right)\right]/a_{nn}
\end{array}$$

Em geral, $x_i^{(k+1)}$ pode ser obtido pela fórmula:

$$x_i^{(k+1)} = \left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)} \right) / a_{ii},\ i = 1, 2, \ldots, n \tag{3.40}$$

Exemplo 3.17 *Se aplicarmos a expressão acima ao exemplo 3.13, resultam as equações já obtidas anteriormente:*

$$\begin{array}{rcccccc}
x_1^{(k+1)} & = & [7 & - & (2x_2^{(k)} & + & x_3^{(k)})]/10 \\
x_2^{(k+1)} & = & [-8 & - & (x_1^{(k)} & + & x_3^{(k)})]/5 \\
x_3^{(k+1)} & = & [6 & - & (2x_1^{(k)} & + & 3x_2^{(k)})]/10
\end{array}$$

Na prática, é mais fácil utlizar o processo descrito no exemplo 3.13 para obter a expressão da iterada genérica.

Método de Gauss-Seidel

Com base na igualdade (3.38), associamos as matrizes do seguinte modo:

$$A\mathbf{x} = \mathbf{b} \Longleftrightarrow (\underbrace{D+L}_{M} + \underbrace{U}_{N})\mathbf{x} = \mathbf{b} \tag{3.41}$$

O método de Gauss-Seidel é um método iterativo da forma (3.37), com $M = D + L$ e $N = U$ e, consequentemente, matriz de iteração dada por $C_{GS} = -M^{-1}N = -(D+L)^{-1}U$. Ou seja, é o método definido por

$$\mathbf{x}^{(k+1)} = \underbrace{-(D+L)^{-1}U}_{C_{GS}}\mathbf{x}^{(k)} + (D+L)^{-1}\mathbf{b} \tag{3.42}$$

A fórmula geral para o cálculo de $\mathbf{x}^{(k+1)}$ pode ser obtida de maneira análoga à utilizada para o método de Jacobi, usando a igualdade (equivalente a (3.42)):

$$(D+L)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} - \mathbf{b} \tag{3.43}$$

donde se obtém

$$D\ \mathbf{x}^{(k+1)} = -L\ \mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)} + \mathbf{b} \tag{3.44}$$

Igualando componente a componente, chega-se a

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \right) / a_{ii},\ i = 1, 2, \ldots, n \tag{3.45}$$

Note-se que a expressão acima também pode ser deduzida a partir da expressão geral (3.40) do método de Jacobi, fazendo as modificações seguintes. No método de Gauss-Seidel utilizam-se os valores $x_j^{(k+1)}, j = 1, 2, \ldots, i-1$, que são aproximações actualizadas de $x_j, j = 1, 2, \ldots, i-1$, em vez dos valores $x_j^{(k)}, j = 1, 2, \ldots, i-1$.

3.4.4 Convergência dos métodos iterativos

Teorema 3.6 *(Condição suficiente de convergência) Seja o sistema linear $A\mathbf{x} = \mathbf{b}$ e suponhamos que o mesmo tenha sido transformado no sistema equivalente*

$$\mathbf{x} = C\mathbf{x} + \mathbf{d} \tag{3.46}$$

onde C é uma matriz quadrada e $\mathbf{d}$ é um vector, e consideremos o método iterativo

$$\mathbf{x}^{(k+1)} = C\mathbf{x}^{(k)} + \mathbf{d} \tag{3.47}$$

com $\mathbf{x}^{(0)}$ um vector qualquer do $\mathbf{R^n}$. Se existe alguma norma induzida (ou natural) de matrizes tal que $\|C\| < 1$, então o método iterativo (3.47) converge para a solução $\mathbf{x}$ do sistema (3.46) qualquer que seja o vector $\mathbf{x}^{(0)}$ dado. E têm-se as fórmulas de erro

$$(I) \quad \|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \|C\|\,\|\mathbf{x} - \mathbf{x}^{(k)}\| \tag{3.48}$$

$$(II) \quad \|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \|C\|^{k+1}\,\|\mathbf{x} - \mathbf{x}^{(0)}\| \tag{3.49}$$

$$(III) \quad \|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \frac{\|C\|}{1 - \|C\|}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \tag{3.50}$$

Dem. Seja $\mathbf{x}$ a solução de (3.46) e portanto de $A\mathbf{x} = \mathbf{b}$ e seja

$$\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$$

o erro da iterada $\mathbf{x}^{(k)}$.

Subtraindo (3.46) de (3.47) vem:

$$\mathbf{x} - \mathbf{x}^{(k+1)} = C\mathbf{x} - C\mathbf{x}^{(k)} = C(\mathbf{x} - \mathbf{x}^{(k)}), k = 0, 1, 2, ...$$

ou seja,

$$\mathbf{e}^{(k+1)} = C\mathbf{e}^{(k)} \tag{3.51}$$

Dando sucessivos valores a k obtém-se

$$\begin{array}{ll} \mathbf{e}^{(1)} = & C\mathbf{e}^{(0)} \\ \mathbf{e}^{(2)} = & C\mathbf{e}^{(1)} = C(C\mathbf{e}^{(0)}) = C^2\mathbf{e}^{(0)} \\ \mathbf{e}^{(3)} = & C\mathbf{e}^{(2)} = C(C^2\mathbf{e}^{(0)}) = C^3\mathbf{e}^{(0)} \\ \vdots & \vdots \end{array}$$

e, em geral,

$$\mathbf{e}^{(k+1)} = C^{(k+1)}\mathbf{e}^{(0)} \tag{3.52}$$

Seja $\|\cdot\|$ uma norma de vector e consideremos a norma de matriz associada. A fórmula (I), ou seja,(3.48), resulta de se aplicarem normas à igualdade (3.51):

$$\|\mathbf{e}^{(k+1)}\| = \|C\mathbf{e}^{(k)}\| \leq \|C\|\|\mathbf{e}^{(k)}\| \qquad \text{(I)}$$

Por outro lado, aplicando normas à igualdade (3.52), obtém-se:

$$\|\mathbf{e}^{(k+1)}\| = \|C^{(k+1)}\mathbf{e}^{(0)}\| \leq \|C^{(k+1)}\|\|\mathbf{e}^{(0)}\|$$

e, notando que:

$$\begin{array}{l} \|C^2\| = \|CC\| \leq \|C\|\|C\| = \|C\|^2 \\ \|C^3\| = \|CC^2\| \leq \|C\|\|C^2\| = \|C\|^3 \\ \|C^4\| = \|CC^3\| \leq \|C\|\|C^3\| = \|C\|^4 \\ \vdots \qquad\qquad \vdots \end{array}$$

obtém-se, em geral, para $k \geq 0$: $\|C^{(k+1)}\| \leq \|C\|^{(k+1)}$. Substituindo acima, resulta a fórmula:

$$\|\mathbf{e}^{(k+1)}\| \leq \|C\|^{(k+1)}\|\mathbf{e}^{(0)}\| \qquad \text{(II)}$$

As fórmulas (I) e (II) são válidas para qualquer norma vectorial e correspondente norma matricial induzida.

Convergência

A fórmula (II), juntamente com a hipótese de que existe uma norma tal que $\|C\| < 1$, pode ser utilizada para provar convergência. Com efeito, sendo $\|C\| < 1$, então

$$\lim_{k\to\infty} \|C^{(k+1)}\| = 0 \Longrightarrow \|\mathbf{e}^{(k+1)}\| \overset{k\to\infty}{\longrightarrow} 0$$

ou seja, $\|\mathbf{x} - \mathbf{x}^{(k+1)}\| \overset{k\to\infty}{\longrightarrow} 0 \implies \mathbf{x}^{(k+1)} \overset{k\to\infty}{\longrightarrow} \mathbf{x}$ qualquer que seja $\mathbf{x}^{(0)}$.

Portanto, provámos que o método converge para $\mathbf{x}$ qualquer que seja $\mathbf{x}^{(0)} \in \mathbf{R^n}$

Para finalizar a demonstração do teorema, falta provar a fórmula de erro (III), ou seja, a desigualdade (3.50), que é válida sob a hipótese de que existe uma norma tal que $\|C\| < 1$.

Da equação $\mathbf{e}^{(k+1)} = C\mathbf{e}^{(k)}$ temos:

$$\begin{aligned} &\mathbf{x} - \mathbf{x}^{(k+1)} = C(\mathbf{x} - \mathbf{x}^{(k)}) \\ &\implies \mathbf{x} - \mathbf{x}^{(k+1)} = C(\mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \end{aligned}$$

Aplicando a norma $\|\cdot\|$ vem:

$$\begin{aligned} &\|\mathbf{x} - \mathbf{x}^{(k+1)}\| = \|C(\mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})\| \leq \|C\|\|\mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\ &\leq \|C\| \left(\|\mathbf{x} - \mathbf{x}^{(k+1)}\| + \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|\right) = \|C\|\|\mathbf{x} - \mathbf{x}^{(k+1)}\| + \|C\|\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\ &\implies \|\mathbf{x} - \mathbf{x}^{(k+1)}\| - \|C\|\|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \|C\|\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\ &\implies (1 - \|C\|)\|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \|C\|\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \end{aligned}$$

Finalmente, atendendo a que $\|C\|\| < 1$, vem

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\| \leq \frac{\|C\|}{1 - \|C\|}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \qquad \text{(III)}$$

□

Corolário 3.1 *O método iterativo (3.47) converge para a solução* $\mathbf{x}$ *do sistema (3.46) qualquer que seja o vector* $\mathbf{x}^{(0)}$ *dado, se alguma das condições se verificar:*

$$(1) \quad \|C\|_\infty < 1 \qquad (3.53)$$

$$(2) \quad \|C\|_1 < 1 \qquad (3.54)$$

Observação: Pode não se dar (1) nem (2) e o método ser convergente.

Exemplo 3.18 *Seja de novo o exemplo* $\begin{bmatrix} 10 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -8 \\ 6 \end{bmatrix}$, *cuja solução exacta é* $\mathbf{x} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$. *Mostre que o método de Jacobi converge para a solução exacta do sistema e obtenha um majorante para* $\|\mathbf{x} - \mathbf{x}^{(2)}\|_\infty$

Já obtivémos a matriz de iteração do método de Jacobi:

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -0.2 & -0.1 \\ -0.2 & 0 & -0.2 \\ -0.2 & -0.3 & 0 \end{bmatrix}$$

Então facilmente vemos que:

$$\|C_J\|_\infty = max\{0.3, 0.4, 0.5\} = 0.5 < 1$$

e a convergência fica provada.

Dado que $\mathbf{x}^{(1)} = [0.960 \ \ -1.86 \ \ 0.940]^T$ e $\mathbf{x}^{(2)} = [0.978 \ \ -1.98 \ \ 0.966]^T$ então: $\mathbf{x}^{(2)} - \mathbf{x}^{(1)} = [0.018 \ \ -0.12 \ \ 0.026]^T$. E temos o seguinte majorante para o erro:

$$\|\mathbf{x} - \mathbf{x}^{(2)}\|_\infty \leq \frac{\|C\|_\infty}{1 - \|C\|_\infty} \|\mathbf{x}^{(2)} - \mathbf{x}^{(1)}\|_\infty$$
$$\|\mathbf{x} - \mathbf{x}^{(2)}\|_\infty \leq \frac{0.5}{1 - 0.5} \ (0.12) = 1.0 \times (0.12) = 0.12$$

Teorema 3.7 *(Critério de Conv. Mét. Jacobi) Considere o sistema linear* $A\mathbf{x} = \mathbf{b}$. *Se* A *é uma matriz de diagonal estritamente dominante por linhas ou por colunas então o método de Jacobi converge para a solução de* $A\mathbf{x} = \mathbf{b}$ *qualquer que seja o vector inicial* $\mathbf{x}^{(0)}$.

Dem. Vamos supor primeiramente que A é de diagonal estritamente dominante por linhas. Neste caso, a convergência resulta como corolário do teorema 3.6. A matriz de iteração do método $C_J = -D^{-1}(L+U)$ é dada por

$$C_J = -\begin{bmatrix} 0 & a_{12}/a_{11} & a_{13}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & a_{23}/a_{22} & \dots & a_{2n}/a_{22} \\ \vdots & ... & ... & & \vdots \\ \vdots & ... & ... & & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & a_{n-1n}/a_{nn} & 0 \end{bmatrix}$$

Como A é de diagonal estritamente dominante por linhas

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ \forall i \iff \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \iff \sum_{\substack{j=1 \\ j \neq i}}^{n} |c_{ij}| < 1 \ \forall i$$

Acima designámos por c_{ij} os elementos da matriz C_J e a relação anterior significa que $\|C_J\|_\infty < 1$. Ou seja, a condição de A ser de diagonal estritamente dominante por linhas é equivalente a ter-se $\|C_J\|_\infty < 1$. E estamos nas condições do teorema (3.6). Consequentemente, podemos concluir que o método de Jacobi converge, qualquer que seja o vector inicial $\mathbf{x}^{(0)}$.

Passando agora à condição de A ser de diagonal estritamente dominante por colunas, é interessante notar que ela não implica que $\|C_J\|_1 < 1$. Basta considerar o exemplo:

$$A = \begin{bmatrix} 4 & 4 & -2 \\ 1 & -8 & 1 \\ 1 & 1 & 5 \end{bmatrix}$$

A é de diagonal estritamente dominante por colunas mas facilmente se verifica que para a matriz C_J vem $\|C_J\|_1 > 1$.

Voltando à demonstração do teorema, sendo A de diagonal estritamente dominante por colunas, é possível mostrar que existe uma norma (não necessariamente a norma $\|.\|_1$) tal que $\|C_J\| < 1$ e, do teorema 3.6, de novo se conclui a convergência do método de Jacobi, qualquer que seja o vector inicial $\mathbf{x}^{(0)}$. $\square$

Também é válido um resultado análogo ao teorema (3.7) para o método de Gauss-Seidel, mas cuja demonstração é mais trabalhosa.

Teorema 3.8 *(Critério de Conv. Mét. Gauss-Seidel) Considere o sistema linear $A\mathbf{x} = \mathbf{b}$. Se A é uma matriz de diagonal estritamente dominante por linhas ou por colunas então o método de Gauss-Seidel converge para a solução de $A\mathbf{x} = \mathbf{b}$ qualquer que seja o vector inicial $\mathbf{x}^{(0)}$.*

Exemplo 3.19 *Seja o sistema* $\begin{bmatrix} 4 & -1 & 1 \\ 4 & -8 & 1 \\ -2 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -21 \\ 15 \end{bmatrix}$. *Conclua sobre a convergência ou não dos métodos de Jacobi e Gauss-Seidel.*

A matriz A não é de diagonal estritamente dominante por colunas, pois falha na primeira coluna: $|a_{11}| = 4 < 4 + 2$.

Contudo, A é de diagonal estritamente dominante por linhas: $|a_{11}| = 4 > 1 + 1 = 2; |a_{22}| = 8 > 4 + 1 = 5; |a_{33}| = 5 > 2 + 1 = 3$. Pelos teoremas anteriores, (3.7) e (3.8), podemos concluir que os métodos de Jacobi e Gauss-Seidel convergem para a solução exacta do sistema, $\forall \mathbf{x}^{(0)}$.

Tem-se ainda o resultado seguinte apenas para o método de Gauss-Seidel:

Teorema 3.9 *(Crit. Conv. Gauss-Seidel) Se A é simétrica e definida positiva então o método de Gauss-Seidel aplicado ao sistema linear $A\mathbf{x} = \mathbf{b}$ converge para a solução do mesmo,qualquer que seja o vector inicial $\mathbf{x}^{(0)}$.*

Concluímos com um resultado geral de convergência.

Teorema 3.10 *(condição necessária e suficiente de convergência) O método iterativo definido por*

$$\mathbf{x}^{(k+1)} = C\mathbf{x}^{(k)} + \mathbf{d},$$

converge para a solução do sistema linear $A\mathbf{x} = \mathbf{b}$*, qualquer que seja o vector inicial* $\mathbf{x}^{(0)}$*, se e somente se* $\rho(C) < 1$*.*

Exemplo 3.20 *Considere o sistema*

$$\begin{aligned} x_1 + 3x_2 &= 2 \\ x_1 + 4x_2 &= 3 \end{aligned}$$

Mostre que os métodos de Jacobi e Gauss-Seidel convergem para a solução do sistema, qualquer que seja a aproximação inicial.

Notando que A não é de diagonal estritamente dominante por linhas nem por colunas, nada se conclui sobre a convergência ou divergência dos métodos pelos teoremas (3.7) e (3.8). Calculemos as respectivas matrizes de iteração.

a) Método de Jacobi

Tem-se $C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -3 \\ -1/4 & 0 \end{bmatrix}$

Então $\|C_J\|_\infty = max\{3, 1/4\} = 3 > 1$, como já era de esperar, e $\|C_J\|_1 = max\{1/4, 3\} = 3 > 1$. Também nada se conclui pelo teorema 3.6.

Calculemos o raio espectral de C_J. Tem-se $\rho(C_J) = \max|\lambda_i(C_J)|$ e vem:

$$|C_J - \lambda\, I| = \begin{vmatrix} -\lambda & -3 \\ -1/4 & -\lambda \end{vmatrix} = \lambda^2 - 3/4 = 0 \Longleftrightarrow \lambda = \overset{+}{-} \sqrt{3/4}$$

Logo, $\rho(C_J) \simeq 0.866025 < 1$, pelo que o método converge, $\forall \mathbf{x^{(0)}}$.

b) Faça o estudo da convergência do método de Gauss-Seidel.

Exemplo 3.21 *Considere o sistema de equações*

$$\left\{ \begin{array}{ccccccc} 2x & + & y & + & \epsilon\, cosz & = & -1 \\ x & + & 3y & - & 3\epsilon\, x\, z & = & 0 \\ \epsilon\, x^2 & + & y & + & 3\, z & = & 0 \end{array} \right. \tag{3.55}$$

onde ϵ é um número real conhecido. No caso de $\epsilon = 0$ justifique que o método de Jacobi converge para a solução do sistema e, se tomarmos $\mathbf{x}^{(0)} = \mathbf{0}$ (vector nulo), mostre que as iteradas satisfazem a desigualdade

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \left(\frac{1}{2}\right)^{(k+1)}$$

Para $\epsilon = 0$, o sistema não-linear acima, reduz-se ao sistema linear

$$\left\{ \begin{array}{rcrcrcl} 2x & + & y & & & = & -1 \\ x & + & 3y & & & = & 0 \\ & & y & + & 3\,z & = & 0 \end{array} \right. \tag{3.56}$$

Uma condição suficiente para que o método de Jacobi aplicado ao sistema linear acima convirja para a solução do mesmo, $\forall \mathbf{x}^{(0)}$, é que $\|C_J\| < 1$ onde $\|\cdot\|$ é uma norma de matrizes. Tem-se

$$C_J = -D^{-1}(L+U) = \begin{bmatrix} 0 & -1/2 & 0 \\ -1/3 & 0 & 0 \\ 0 & -1/3 & 0 \end{bmatrix}$$

logo $\|C_J\|_\infty = \max[1/2, 1/3, 1/3] = 1/2 < 1$. Portanto, podemos concluir que o método de Jacobi aplicado ao sistema linear converge para a solução do mesmo, qualquer que seja $\mathbf{x}^{(0)}$.

Por outro lado, se tomarmos $\mathbf{x}^{(0)} = \mathbf{0}$, pelo método de Jacobi obtemos $\mathbf{x}^{(1)} = C_J\mathbf{x}^{(0)} + D^{-1}\mathbf{b} = [-1/2 \;\; 0 \;\; 0]^T$. Em seguida, aplicando a seguinte fórmula do erro

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \frac{C_J^{k+1}}{1 - C_J}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty$$

obtemos

$$\|\mathbf{x} - \mathbf{x}^{(k+1)}\|_\infty \leq \frac{\left(\frac{1}{2}\right)^{k+1}}{1 - \left(\frac{1}{2}\right)}(1/2 - 0) = \left(\frac{1}{2}\right)^{k+1}$$

3.4.5 O método SOR

O método SOR (Successive Over Relaxation) pode ser considerado como uma variante do método de Gauss-Seidel. Segundo o teorema 3.10, é condição necessária e suficiente para que o método de Gauss-Seidel convirja, qualquer que seja a aproximação inicial, que se tenha $\rho(C_{GS}) < 1$, onde $C_{GS} = -(L+D)^{-1}U$. Pode porém ocorrer que $\rho(C_{GS}) \approx 1$ e, nesse caso, a convergência poderá ser muito lenta, pelo que é usual modificar o método de Gauss-Seidel introduzindo um parâmetro de aceleração ω.

Na dedução do método de Gauss-Seidel, considerou-se a decomposição $A = M + N$, com $M = D + L$ e $N = U$. Para o método de SOR considera-se $A = M_\omega + N_\omega$ com M_ω dada por:

$$M_\omega = L + \frac{1}{\omega} D = \frac{1}{\omega}(\omega L + D)$$

A matriz N_ω pode-se obter conjugando as duas igualdades $A = L + D + U$ e $A = M_\omega + N_\omega$, o que conduz à igualdade

$$N_\omega = L + D + U - M_\omega = (1 - \frac{1}{\omega}) D + U$$

Dada uma iterada inicial $\mathbf{x}^{(0)}$, o método iterativo SOR é definido por

$$\mathbf{x}^{(k+1)} = C_{SOR}\mathbf{x}^{(k)} + M_\omega^{-1}\,\mathbf{b} \quad k \geq 0, \tag{3.57}$$

onde $C_{SOR} = -M_\omega^{-1}\,N_\omega$ é a matriz de iteração. Vejamos como obter uma expressão explícita para as componentes do vector $\mathbf{x}^{(k+1)}$.

Tem-se $(M_\omega + N_\omega)\mathbf{x} = \mathbf{b} \Leftrightarrow (M_\omega)\mathbf{x} = -N_\omega\mathbf{x} + \mathbf{b}$ e o método SOR toma a forma

$$M_\omega\,\mathbf{x}^{(k+1)} = -N_\omega\,\mathbf{x}^{(k)} + \mathbf{b},$$

donde se obtém, sucessivamente

$$\begin{aligned} &\frac{1}{\omega}(\omega\,L + D)\,\mathbf{x}^{(k+1)} = \mathbf{b} - (\frac{1}{\omega})\,[(\omega - 1)D + \omega U]\,\mathbf{x}^{(k)} \\ &\Rightarrow (\omega\,L + D)\,\mathbf{x}^{(k+1)} = \omega\,\mathbf{b} - [(\omega - 1)D + \omega U]\,\mathbf{x}^{(k)} \end{aligned} \tag{3.58}$$

Continuando, obtém-se:

$$\begin{aligned} D\,\mathbf{x}^{(k+1)} &= \omega\,\mathbf{b} - \omega L\mathbf{x}^{(k+1)} - [(\omega - 1)D + \omega U]\,\mathbf{x}^{(k)} && (3.59)\\ &= \omega[\mathbf{b} - L\mathbf{x}^{(k+1)} - U\,\mathbf{x}^{(k)}] + (1 - \omega)\,D\mathbf{x}^{(k)} && (3.60) \end{aligned}$$

E chega-se à seguinte fórmula para o método SOR:

$$\mathbf{x}^{(k+1)} = \omega D^{-1}[\mathbf{b} - L\mathbf{x}^{(k+1)} - U\,\mathbf{x}^{(k)}] + (1 - \omega)\mathbf{x}^{(k)}, \quad \text{se } a_{ii} \neq 0. \tag{3.61}$$

Observemos que a última igualdade pode ser reescrita como:

$$\begin{aligned} &\mathbf{x}^{(k+1)} = \omega\,\mathbf{z}^{(k+1)} + (1 - \omega)\,\mathbf{x}^{(k)} \\ &\text{com } \mathbf{z}^{(k+1)} = D^{-1}[\mathbf{b} - L\mathbf{x}^{(k+1)} - \omega U\,\mathbf{x}^{(k)}] \end{aligned}$$

Assim as componentes $x_i^{(k+1)}$ de $\mathbf{x}^{(k+1)}$ podem ser calculadas através das fórmulas:

$$z_i^{(k+1)} = \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right] / a_{ii} \tag{3.62}$$
$$x_i^{(k+1)} = \omega z_i^{(k+1)} + (1-\omega) x_i^k, \quad i = 1, 2, ..., n; k = 0, 1, ...$$

O parâmetro ω é um número real dado e, sob certas condições, pode ser determinado qual o seu valor que conduz à convergência mais rápida. No caso $\omega = 1$ vemos que o método SOR se reduz ao método de Gauss-Seidel.

Teorema 3.11 *Tem-se* $\rho(C_{SOR}) \geq |\omega - 1|, \quad \forall \omega \in R$. *Então* $\rho(C_{SOR}) < 1 \quad$ *se* $\omega \in (0, 2)$.

O teorema diz que o método SOR converge se $0 < \omega < 1$ (esta é pois uma condição necessária para que haja convergência).

Teorema 3.12 *Se* A_n *é simétrica e definida positiva então o método SOR converge para a solução do sistema* $A\mathbf{x} = \mathbf{b}$ *qualquer que seja* $\mathbf{x}^{(0)} \in R^n$.

Exemplo 3.22 *Efectuar uma iteração do método de SOR aplicado ao sistema linear*

$$\begin{bmatrix} 1 & 0 & -1 \\ -3/4 & 2 & 0 \\ -1 & -1/4 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

Começa-se por reescrever o sistema na forma:

$$\begin{array}{rcl} x_1 - x_3 = 2 & \Leftrightarrow & x_1 = 2 + x_3 \\ -\frac{3}{4}x_1 + 2x_2 = 1 & \Leftrightarrow & x_2 = (1 + \frac{3}{4}x_1)/2 \\ -x_1 - \frac{1}{4}x_2 + 2x_3 = 1 & \Leftrightarrow & x_3 = (1 + x_1 + \frac{1}{4}x_2)/2 \end{array}$$

e considera-se a iteração

$$\begin{array}{ll} z_1^{(k+1)} = 2 + x_3^{(k)} & x_1^{(k+1)} = \omega\, z_1^{(k+1)} + (1-\omega)\, x_1^{(k)} \\ z_2^{(k+1)} = \left[1 + \frac{3}{4}\, x_1^{(k+1)}\right]/2 & x_2^{(k+1)} = \omega\, z_2^{(k+1)} + (1-\omega)\, x_2^{(k)} \\ z_3^{(k+1)} = \left[1 + x_1^{(k+1)} + \frac{1}{4}x_2^{(k)}\right]/2 & x_3^{(k+1)} = \omega\, z_3^{(k+1)} + (1-\omega)\, x_3^{(k)} \end{array}$$

Por exemplo, com $\omega = 1.2$ e partindo de $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$, obtém-se:

$$\begin{array}{ll} z_1^{(k+1)} = 2 + 0 = 2 & x_1^{(k+1)} = 1.2\,(2) - 0.2(0) = 2.4 \\ z_2^{(k+1)} = \left[1 + \frac{3}{4}\, 2.4\right]/2 = 1.4 & x_2^{(k+1)} = 1.2\,(1.4) - 0.2(0) = 1.68 \\ z_3^{(k+1)} = \left[1 + 2.4 + \frac{1}{4}1.68\right]/2 = 1.91 & x_3^{(k+1)} = 1.2\,(1.91) - 0.2(0) = 2.292 \end{array}$$

3.5 Sistemas de equações não lineares

Um sistema de n-equações não-lineares a n-incógnitas tem a forma:

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

onde $f_i : \mathbf{R^n} \longrightarrow I\!R, i = 1, 2, \ldots, n$. Definindo $\mathbf{x} = [x_1 \ x_2 \ \ldots \ x_n]^T$ e $\mathbf{F} = [f_1 \ f_2 \ \ldots \ f_n]^T$, podemos escrever o sistema na forma matricial:

$$\mathbf{F}(\mathbf{x}) = 0, \quad \begin{cases} \mathbf{F} : \mathbf{R^n} \longrightarrow \mathbf{R^n} \\ \mathbf{x} \longrightarrow \mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix} \end{cases}$$

Exemplo 3.23 *O sistema não-linear*

$$\begin{array}{l} x_1 + x_2 + x_3 = 4 \\ x_1x_2 + x_1x_3 + x_2x_3 = 1 \\ x_1x_2x_3 = -6 \end{array} \qquad \begin{cases} \textit{tem como solução} \\ \mathbf{x} = [-1 \ \ 2 \ \ 3]^T \end{cases}$$

pode ser escrito como:

$$\mathbf{F}(\mathbf{x}) = 0 \qquad \text{onde}$$

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 - 4 \\ x_1x_2 + x_1x_3 + x_2x_3 - 1 \\ x_1x_2x_3 + 6 \end{bmatrix}$$

3.5.1 O Método de Newton para sistemas não lineares

Para funções $f : I\!R \longrightarrow I\!R$, o método de Newton é dado por:

$$x_{m+1} = g(x_m) = x_m - (f'(x_m))^{-1} f(x_m)$$

No caso de funções vectoriais, $\mathbf{F} : \mathbf{R^n} \longrightarrow \mathbf{R^n}$, a generalização do método de Newton é dada por:

$\mathbf{x}^{(m+1)} = \mathbf{G}(\mathbf{x}^{(m)})$
onde $\mathbf{G}(\mathbf{x}) = \mathbf{x} - [J(\mathbf{x})]^{-1}\mathbf{F}(\mathbf{x})$

e

$$J(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

é a matriz Jacobiana de $\mathbf{F}$, isto é, $J(\mathbf{x}) = \dfrac{\partial f_i}{\partial x_j}, j = 1, 2, \ldots, n; i = 1, 2, \ldots, n.$

Dada uma aproximação inicial $\mathbf{x}^{(0)}$, as iteradas $\mathbf{x}^{(m+1)}$ são calculadas através do processo iterativo:

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \left[J(\mathbf{x}^{(m)})\right]^{-1} \mathbf{F}(\mathbf{x}^{(m)}) \tag{3.63}$$

a que se chama método de Newton para sistemas não-lineares.

Cálculo de $\mathbf{x}^{(m+1)}$

De (3.63) temos:

$$\mathbf{x}^{(m+1)} - \mathbf{x}^{(m)} = -\left[J(\mathbf{x}^{(m)})\right]^{-1} \mathbf{F}(\mathbf{x}^{(m)})$$

Fazendo $\mathbf{d} = \mathbf{x}^{(m+1)} - \mathbf{x}^{(m)}$, vem:

$$\mathbf{d} = -\left[J(\mathbf{x}^{(m)})\right]^{-1} \mathbf{F}(\mathbf{x}^{(m)})$$

e multiplicando ambos os lador por $J(\mathbf{x}^{(m)})$ obtemos:

$$J(\mathbf{x}^{(m)})\ \mathbf{d} = -\mathbf{F}(\mathbf{x}^{(m)}), \tag{3.64}$$

que é um sistema de equações lineares, em que o vector $\mathbf{d}$ é o vector incógnita, que pode ser resolvido pelas técnicas já estudadas: Método de eliminação de Gauss, LU, Jacobi, Gauss-Seidel, etc.

Uma vez obtido $\mathbf{d}$, $\mathbf{x}^{(m+1)}$ é determinado por:

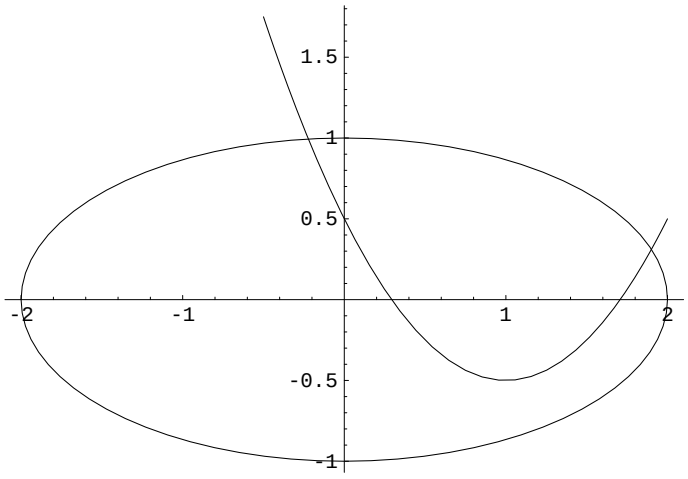
$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{d} \tag{3.65}$$

Obs: Se $J(\mathbf{x}^{(m)})$ for não- singular $\forall m$ e se $\mathbf{x}^{(0)}$ estiver suficientemente perto de $\mathbf{x}$, pode-se provar que $\mathbf{x}^{(m)} \overset{m\to\infty}{\longrightarrow} \mathbf{x}$ é a convergência é quadrática.

Exemplo 3.24 *: começando com uma aproximação inicial* $\mathbf{x}^{(0)} = [0\ 1.0]^T$*, calcule 2 iteradas do método de Newton para o sistema não-linear*

$$\left.\begin{array}{rcl} x_2 - x_1^2 + 2x_1 &=& 0.5 \\ x_1^2 + 4x_2^2 &=& 4 \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{rcl} f_1(x_1,x_2) &=& x_2 - x_1^2 + 2x_1 - 0.5 \\ f_2(x_1,x_2) &=& x_1^2 + 4x_2^2 - 4 \end{array}\right. \begin{array}{c} = \\ = \end{array} \begin{array}{c} 0 \\ 0 \end{array}$$

Graficamente, as soluções do sistema são os pontos de intersecção das duas curvas seguintes (respectivamente, prábola e elipse):



Solução

$$\mathbf{F}(x_1,x_2) = \left[\begin{array}{c} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{array}\right] = \left[\begin{array}{c} x_2 - x_1^2 + 2x_1 - 0.5 \\ x_1^2 + 4x_2^2 - 4 \end{array}\right]$$

Pelo método de Newton temos:

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - \left[J(\mathbf{x}^{(m)})\right]^{-1} \mathbf{F}(\mathbf{x}^{(m)})$$

ou, equivalentemente,

$$\left\{\begin{array}{l} J(\mathbf{x}^{(m)})\mathbf{d} = -\mathbf{F}(\mathbf{x}^{(m)}) \\ \mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{d} \end{array}\right.$$

Cálculo de $J(\mathbf{x}^{(m)})$

$$J(\mathbf{x}^{(m)}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2x_1^{(m)} + 2 & 1 \\ 2x_1^{(m)} & 8x_2^{(m)} \end{bmatrix}$$

O vector $F(\mathbf{x}^{(m)})$

$$\mathbf{F}(\mathbf{x}^{(m)}) = \begin{bmatrix} f_1(x_1^{(m)}, x_2^{(m)}) = x_2^{(m)} - \left(x_1^{(m)}\right)^2 + 2x_1^{(m)} - 0.5 \\ f_2(x_1^{(m)}, x_2^{(m)}) = \left(x_1^{(m)}\right)^2 + 4\left(x_2^{(m)}\right)^2 - 4 \end{bmatrix}$$

Cálculo de $\mathbf{x}^{(1)}$

O vector $F(\mathbf{x}^{(0)})$

Com $\mathbf{x}^{(0)} = [0 \quad 1.0]^T$, vem

$$\mathbf{F}(\mathbf{x}^{(0)}) = \begin{bmatrix} f_1(x_1^{(0)}, x_2^{(0)}) = x_2^{(0)} - \left(x_1^{(0)}\right)^2 + 2x_1^{(0)} - 0.5 = 1 - 0.5 = 0.5 \\ f_2(x_1^{(0)}, x_2^{(0)}) = \left(x_1^{(0)}\right)^2 + 4\left(x_2^{(0)}\right)^2 - 4 = 4 - 4 = 0 \end{bmatrix}$$

Por outro lado,

$$J(x^{(0)}) = \begin{bmatrix} -2x_1^{(0)} + 2 & 1 \\ 2x_1^{(0)} & 8x_2^{(0)} \end{bmatrix}$$

Logo, temos o seguinte sistema linear para o cálculo de $\mathbf{d}$:

$$\begin{bmatrix} 2 & 1 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \quad \begin{cases} \text{que tem como solução:} \\ d_2 = 0 \text{ e } d_1 = -0.25 \end{cases}$$

Portanto,

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{d} = \begin{bmatrix} x_1^{(0)} + d_1 = 0.0 - 0.25 = -0.25 \\ x_2^{(0)} + d_2 = 1.0 + 0.0 = 1.0 \end{bmatrix}$$

Cálculo de $\mathbf{x}^{(2)}$:

$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{d}$
onde $\mathbf{d}$ é solução do sistema linear

$$J(\mathbf{x}^{(1)}) = \begin{bmatrix} -2x_1^{(1)} + 2 & 1 \\ 2x_1^{(1)} & 8x_2^{(1)} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} x_2^{(1)} - \left(x_1^{(1)}\right)^2 + 2x_1^{(1)} - 0.5 \\ \left(x_1^{(1)}\right)^2 + 4\left(x_2^{(1)}\right)^2 - 4 \end{bmatrix}$$

ou seja:

$$\begin{bmatrix} 2.5 & 1 \\ -0.5 & 8 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 0.06255 \\ -0.0625 \end{bmatrix}$$

Podemos resolver o sistema aplicando o método eliminação Gauss (ou por outro meio)

$$\begin{cases} m_{21} = \dfrac{a_{21}}{a_{11}} = \dfrac{-0.5}{2.5} = 0.2 \\ E_2 - m_{21}E_1 \longrightarrow E_2 \end{cases} \Longrightarrow \begin{cases} 2.5d_1 + d_2 = 0.0625 \\ 8.2d_2 = -0.050 \end{cases} \Longrightarrow \begin{cases} d_2 = -0.0061 \\ d_1 = 0.02744 \end{cases}$$

Finalmente,

$$x_1^{(2)} = -0.25 + 0.02744 = -0.22256$$
$$x_2^{(2)} = 1.0 - 0.00610 = 0.99390$$

que são as componentes da iterada $\mathbf{x}^{(2)}$ pretendida.

4 Aproximação de funções

Introdução

Começamos por rever resultados que mostram a importância dos polinómios para aproximar funções contínuas.

Teorema (Weierstrass): Se f é definida e contínua em $[a,\ b]$ então, dado $\epsilon > 0$, existe um polinómio de grau P, definido em $[a,\ b]$, tal que

$$|f(x) - P(x)| < \epsilon, \quad \forall\ x\ \in\ [a,\ b]$$

Polinómio de Taylor: Seja $f(x)$ uma função definida em $[a,\ b]$ e suponhamos que $f(x)\ \in C^{n+1}[a,\ b]$. Então, um polinómio que pode ser usado para aproximar $f(x)$ é o polinómio de Taylor de grau n em torno de x_0:

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2} + \cdots + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!} \tag{4.1}$$

e o erro é dado pelo resto de ordem n:

$$R_n(x) = f(x) - P_n(x) = f^{(n+1)}(\xi_x)\frac{(x - x_0)^{n+1}}{(n+1)!} \qquad \xi_x \ \text{ entre } \ x \text{ e } x_0 \tag{4.2}$$

Exemplo 4.1 *Obtenha o polinómio de Taylor de grau 3 da função $f(x) = \sqrt{1+x}$, em torno de $x_0 = 0$, e calcule uma aproximação para $\sqrt{1.1}$.*

Solução:

$$P_3(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2} + f'''(x_0)\frac{(x - x_0)^3}{3!}$$

$$\begin{aligned}
f(x) &= (1+x)^{1/2}, & f(x_0) &= f(0) = 1 \\
f'(x) &= \frac{1}{2}(1+x)^{-1/2}, & f'(x_0) &= f'(0) = 1/2 \\
f''(x) &= -\frac{1}{4}(1+x)^{-3/2}, & f''(x_0) &= f''(0) = -1/4 \\
f'''(x) &= \frac{3}{8}(1+x)^{-5/2}, & f''(x_0) &= f''(0) = 3/8
\end{aligned}$$

Portanto,

$$P_3(x) = 1 + \frac{1}{2}x + \frac{-1}{4}\frac{1}{2}x^2 + \frac{3}{8}\frac{1}{3!}x^3 \quad = \quad 1 + \frac{1}{2}x + \frac{-1}{8}x^2 + \frac{1}{16}x^3$$

donde obtemos

$$f(1.1) = \sqrt{1.1} \approx P_3(1.1) = 1 + \frac{1}{2}(0.1) + \frac{-1}{8}0.1^2 + \frac{1}{16}0.1^3 = 1.0488125$$

4.1 Interpolação polinomial

Considere a seguinte tabela de dados sobre a população dos EUA:

Ano	1930	1940	1950	1960	1970	1980
Pop. (milhões)	123.203	131.669	150.697	179.323	203.212	226.505

Com base nestes dados, pergunta-se:

i) Pode-se obter uma estimativa para a população em 1965?

i) Qual seria a população no ano 2000?

Podemos responder a estas duas perguntas utlizando uma função interpoladora.

Definição 4.1 *: (***função interpoladora***)*

Seja f uma função definida num intervalo $[a,b]$. Sejam $x_0, x_1, \cdots, x_n$ $n+1$ pontos distintos de $[a,b]$ e defina-se $f_j := f(x_j), \quad j = 0, 1, \ldots, n$. Uma função g tal que

$$g(x_j) = f_j, \qquad j = 0, 1, \ldots, n. \tag{4.3}$$

é chamada função interpoladora. Os pontos x_j são chamados pontos de interpolação ou nós e os valores f_j os valores interpolados.

Em particular, vão ser do nosso interesse funções interpoladoras polinomiais.

Existência e unicidade do polinómio interpolador

Teorema 4.1 *Sejam $x_0, x_1, \cdots, x_n$ $n+1$ pontos distintos do intervalo $[a,b]$ e sejam $f_0, f_1, \cdots, f_n$ os correspondentes valores de uma função f nesses pontos. Existe um único polinómio p_n de grau $\leq n$ que interpola f nos pontos $x_j, j = 0, 1, \cdots, n$, ou seja, tal que:*

$$p_n(x_j) = f_j, \qquad j = 0, 1, \cdots, n. \tag{4.4}$$

Dem.:

Seja

$$p_n(x) := a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \tag{4.5}$$

onde os coeficientes a_j são constantes a determinar. As condições (4.4) traduzem-se nas $n+1$ equações

$$p_n(x_j) = a_0 + a_1 x_j + a_2 x_j^2 + \cdots + a_n x_j^n, \qquad j = 0, 1, \cdots, n.$$

Estas constituem um sistema da forma $\mathbf{Va} = \mathbf{f}$, onde $\mathbf{V}$ é a matriz definida por

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ . & . & . & \cdots & . \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}$$

e onde $\mathbf{a} = [a_0 a_1 \cdots a_n]^T$ e $\mathbf{f} = [f_0 f_1 \cdots f_n]^T$. A matriz $\mathbf{V}$ é uma matriz de Vandermonde e pode verificar-se que o seu determinante é dado por

$$det\mathbf{V} = \prod_{0 \le j < i \le n} (x_i - x_j)$$

Sendo os x_j distintos, será $det\mathbf{V} \neq 0$. Então $\mathbf{V}$ é não singular e o sistema $\mathbf{Va} = \mathbf{f}$ tem uma solução única $\mathbf{a}$. Fica provada a existência dum único polinómio interpolador de grau $\le n$. Para se obter os coeficientes $a_0, a_1, \cdots, a_n$ que devem ser inseridos em (4.5), é necessário resolver o sistema $\mathbf{Va} = \mathbf{f}$. Este processo de determinar uma expressão para p_n é pouco conveniente do ponto de vista computacional. Outros modos de construir p_n serão tratados neste capítulo. □

A fórmula de Lagrange

Vamos em seguida obter uma expressão explícita para o polinómio interpolador.

Para cada $0 \le j \le n$, seja l_j o poliómio de grau n definido por

$$l_j(x) := \frac{(x - x_0)(x - x_1) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)}.$$

Os l_j são chamados *polinómios base de Lagrange*.

Facilmente se verifica que

$$l_j(x_i) = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases} \tag{4.6}$$

Vamos provar que p_n pode ser representado na seguinte forma

$$p_n(x) = \sum_{j=0}^{n} f_j l_j(x).$$

Seja $q(x)$ a expressão do lado direito, isto é

$$q(x) = \sum_{j=0}^{n} f_j l_j(x).$$

Quer-se provar que $p_n(x) = q(x)$. Comecemos por notar que a soma de polinómios de grau n é um polinómio de grau $\leq n$. Além disso, atendendo a (4.6) tem-se, para $0 \leq i \leq n$,

$$\begin{aligned} q(x_i) &= f_0 l_0(x_i) + \cdots + f_n l_n(x_i) \\ &= f_i l_i(x_i) = f_i, \end{aligned}$$

o que prova tratar-se dum polinómio interpolador de grau $\leq n$. Atendendo à unicidade do polinómio interpolador (teorema 2.1), concluimos a igualdade pretendida.

A fórmula

$$p_n(x) = \sum_{j=0}^{n} f_j l_j(x), \tag{4.7}$$

com

$$l_j(x) := \prod_{\substack{0 \leq i \leq n \\ i \neq j}} \frac{(x - x_i)}{(x_j - x_i)} \tag{4.8}$$

é chamada *fórmula de Lagrange* para o polinómio interpolador.

Exemplo 4.2 *Determine o polinómio interpolador de grau ≤ 2 duma certa função f nos seguintes pontos tabelados*

x_i	1	-1	2
$f(x_i)$	0	-3	4

Utilize o polinómio obtido para aproximar o valor $f(0)$.

Solução

Tem-se

$$l_0(x) = \frac{(x+1)(x-2)}{(1+1)(1-2)} = \frac{x^2 - x - 2}{-2}$$

$$l_1(x) = \frac{(x-1)(x-2)}{(-1-1)(-1-2)} = \frac{x^2 - 3x + 2}{6}$$

$$l_2(x) = \frac{(x-1)(x+1)}{(2-1)(2+1)} = \frac{x^2 - 1}{3}$$

A fórmula (4.7) dá-nos então

$$p_2(x) = 0 - 3\frac{x^2 - 3x + 2}{6} + 4\frac{x^2 - 1}{3}.$$

Usamos então o valor $p_2(0) = -7/3$ para aproximar o valor desconhecido $f(0)$. Sem dispormos de mais informação, não podemos avaliar se a aproximação é boa. Note que se

quisermos representar p_2 em potências de x, basta-nos efectuar os cálculos na expressão obtida. Tem-se, neste caso, $p_2(x) = \frac{1}{6}(5x^2 + 9x - 14)$, que está na forma (4.5). $\square$

É importante distinguir entre a *função* p_n e as várias *representações* de p_n. Pelo teorema anterior, sabemos que p_n é único, para cada conjunto de $n+1$ valores. Contudo, poderá haver muitas maneiras de representar explicitamente p_n e cada uma dessas *fórmulas* sugere um dado algoritmo para calcular p_n. Mais adiante estudaremos outra representação para p_n.

Erro de interpolação polinomial

Como pretendemos usar o polinómio interpolador para aproximar a função f em pontos distintos dos *pontos de interpolação* x_j, interessa-nos uma estimativa da diferença $f(x) - p_n(x)$, para $x \in [a, b]$.

Teorema 4.2 *Seja $p_n \in \mathcal{P}_n$ o polinómio interpolador de f em $n+1$ pontos distintos $x_0, x_1, \cdots, x_n$ de $[a, b]$. Se $f \in C^{n+1}[a, b]$, então para cada $x \in [a, b]$ existe um ponto $\xi = \xi(x)$ no intervalo* $\text{int}(x_0, \ldots, x_n, x) := (\min\{x_0, x_1, \cdots, x_n, x\}, \max\{x_0, x_1, \cdots, x_n, x\})$, *tal que*

$$e_n(x) := f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W(x), \tag{4.9}$$

onde

$$W(x) := (x - x_0)(x - x_1) \cdots (x - x_n).$$

Dem.:

Se $x = x_j$, para algum j, o resultado do teorema é trivialmente válido.
Seja x um valor diferente dos x_j e considere-se a função auxiliar F definida por

$$F(t) := f(t) - p_n(t) - CW(t), \tag{4.10}$$

onde

$$C = \frac{f(x) - p_n(x)}{W(x)}.$$

Tem-se

$$F(x_j) = f(x_j) - p_n(x_j) - CW(x_j) = 0, \quad j = 0, 1, \cdots, n,$$

e também

$$F(x) = f(x) - p_n(x) - CW(x) = 0,$$

atendendo à definição de C. A função F tem pelo menos $n+2$ zeros distintos em $[a, b]$. Pelo teorema de Rolle, a derivada F' tem pelo menos $n+1$ zeros no intervalo $I_p(x_0, x_1, \ldots, x_n, x)$.

A segunda derivada terá pelo menos n zeros e, por um processo indutivo, concluimos que a derivada de ordem $(n+1)$ tem pelo menos um zero. Seja ξ esse zero. Derivando $n+1$ vezes a equação (4.10) e fazendo $t = \xi$, obtemos

$$0 = F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - C(n+1)!,$$

o que, depois de substituir C pela sua expressão, conduz ao resultado pretendido. □

É evidente que a equação (4.9) não pode ser usada para calcular o valor exacto do erro $e_n(x)$, visto que $\xi = \xi(x)$ é em geral uma função desconhecida. (Uma excepção é o caso em que a derivada de ordem $(n+1)$ de f é uma constante).

Porém, sendo $f^{(n+1)}$ contínua, existirá uma constante M_{n+1} tal que $|f^{(n+1)}(t)| \leq M_{n+1}$ para qualquer $t \in [a, b]$. Isto, juntamente com a equação (4.9), resulta na fórmula

$$|e_n(x)| = |f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!}|(x - x_0)(x - x_1)\cdots(x - x_n)|, \tag{4.11}$$

a qual permite obter um majorante para o erro de interpolação num certo $x \in [a, b]$, fixado.

Exemplo 4.3

Suponha que no exemplo (4.2) se conhece um majorante para a terceira derivada de f: $|f'''(t)| \leq 1/2$, $\forall t \in [-1, 2]$. Calcule um majorante para o erro absoluto cometido ao aproximar-se $f(0)$ por $p_2(0)$.

Solução Tem-se, por aplicação da fórmula (4.11), com $n = 2$ e $M_3 = 1/2$,

$$|f(0) - p_2(0)| \leq \frac{1/2}{(3)!}|(0 - x_0)(0 - x_1)(0 - x_2)| = \frac{1/2}{(3)(2)}|(0 - 1)(0 + 1)(0 - 2)| \simeq 0.17$$

É claro que em (4.11) tomou-se $x = 0$, que é o ponto onde se pretende estimar o erro. Os pontos de interpolação foram definidos no exemplo 4.2.

Fórmula interpoladora de Newton com diferenças divididas

Uma outra forma de representar o polinómio interpolador é usando diferenças divididas, que a seguir definiremos. Este método, devido a Isaac Newton (1642-1727), permite passar dum polinómio interpolador dum certo grau $n-1$ para o polinómio de grau superior n, a partir do polinómio de grau inferior.

Mais precisamente, sejam $p_{n-1} \in \mathcal{P}_{n-1}$ e $p_n \in \mathcal{P}_n$ os polinómios interpoladores de f nos pontos x_i, para $i = 0, 1, \cdots, n-1$ e $i = 0, 1, \cdots, n$, respectivamente. Vamos provar que é possível escrever

$$p_n(x) = p_{n-1}(x) + C_n(x), \tag{4.12}$$

onde C_n é uma função que a seguir caracterizamos.

Como $C_n \in \mathcal{P}_n$ e $C_n(x_i) = 0, i = 0, 1, \cdots, n-1$, então C_n é da forma

$$C_n(x) = A_n(x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

com A_n constante real. Substituindo em (4.12), vem

$$p_n(x) = p_{n-1}(x) + A_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{4.13}$$

Começando com $p_0(x) = f_0$, obtém-se desta relação

$$\begin{aligned}
p_1(x) &= f_0 + A_1(x - x_0) && (4.14)\\
p_2(x) &= p_1(x) + A_2(x - x_0)(x - x_1) \\
&= f_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) && (4.15)\\
p_3(x) &= p_2(x) + A_3(x - x_0)(x - x_1)(x - x_2) \\
&= f_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + A_3(x - x_0)(x - x_1)(x - x_2) && (4.16)\\
&\vdots \\
p_n(x) &= f_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + A_3(x - x_0)(x - x_1)(x - x_2) + \cdots \\
&+ A_n(x - x_0) \cdots (x - x_{n-1}) && (4.17)
\end{aligned}$$

Substituindo x por $x_1, x_2, \cdots, x_n$, respectivamente nas equações (4.14) ,..., (4.17), obtemos um sistema de equações (de matriz triangular) nas incógnitas $A_1, \cdots, A_n$. É fácil de verificar que este sistema tem uma solução única.

Assim, de $p_1(x_1) = f_1$, sai que

$$A_1 = \frac{f_1 - f_0}{x_1 - x_0}, \tag{4.18}$$

Repare-se que A_1 é o declive da recta que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$. Chama-se a A_1 *diferença dividida de 1ª ordem da função f, associada aos pontos x_0, x_1*, e designa-se por

$$A_1 = f[x_0, x_1] \tag{4.19}$$

Em seguida, substitui-se A_1 pela expressão (4.18) na equação (4.15). Para se determinar A_2, basta fazer $x = x_2$ em (4.15) e usar o facto de $p_2(x_2) = f_2$ (por p_2 interpolar f em x_0, x_1, x_2. Assim, obtém-se

$$f_2 = p_2(x_2) = f_0 + \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_0) + A_2(x_2 - x_0)(x_2 - x_1) \tag{4.20}$$

o que, depois de simplificados os cálculos, conduz a

$$A_2 = \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)} \tag{4.21}$$

Chama-se a A_2 *diferença dividida de 2ª ordem da função* f, *associada aos pontos* x_0, x_1, x_2, e designa-se por

$$A_2 = f[x_0, x_1, x_2] \tag{4.22}$$

Conhecidas A_1 e A_2, de modo análogo se obteria A_3 da equação (4.16). E assim sucessivamente, até obter todas as constantes até A_n.

Notando que se pode reescrever

$$A_1 = \frac{f_0}{x - x_1} + \frac{f_1}{x - x_0}$$

e comparando com a expressão de A_2, é fácil de compreender que a expressão geral de A_n seja dada por

$$A_n = \sum_{j=0}^{n} \frac{f_j}{\prod\limits_{\substack{0 \le i \le n \\ i \neq j}} (x_j - x_i)} \tag{4.23}$$

No seguimento do que foi definido atrás, usa-se a seguinte notação para designar o valor A_n,

$$A_n = f[x_0, x_1, \cdots, x_n] \tag{4.24}$$

a que se chama *diferença dividida de ordem n da função* f, *associada aos pontos* $x_0, x_1, \cdots, x_n$.

De (4.17), podemos ainda concluir: A_n, ou seja, $f[x_0, x_1, \cdots, x_n]$, é *o coeficiente do termo em* x^n *de* p_n.

Com a notação das diferenças divididas, podemos escrever o polinómio interpolador p_n na forma

$$\begin{aligned} p_n(x) \;=\; & f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \\ & \cdots + f[x_0, x_1, \cdots, x_n](x - x_0) \cdots (x - x_{n-1}) \end{aligned} \tag{4.25}$$

a que se chama *fórmula de Newton com diferenças divididas*. Por uma questão de uniformidade, no lugar de $f(x_i)$ usa-se $f[x_i]$, a que se chama diferença de ordem zero.

Exemplos

Resulta de (4.23) e (4.24) que

$$f[x_0, x_1] = \frac{f_0}{x_0 - x_1} + \frac{f_1}{x_1 - x_0}$$

o que confirma a expressão já obtida em (4.18). De modo geral, a diferença dividida de ordem 2, nos pontos x_i, x_{i+1}, é dada por

$$f[x_i, x_{i+1}] = \frac{f_{i+1} - f_i}{x_{i+1} - x_i}.$$

Como exemplo de diferença de ordem 3, tem-se

$$f[x_0, x_1, x_2] = \frac{f_0}{(x_0 - x_1)(x_0 - x_2)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)} + \frac{f_2}{(x_2 - x_0)(x_2 - x_1)}$$

A seguinte propriedade permite relacionar diferenças de uma certa ordem k com diferenças de ordem $k - 1$.

$$f[x_i, x_{i+1}, \cdots, x_{i+k}] = \frac{f[x_{i+1}, \cdots, x_{i+k}] - f[x_i, , \cdots, x_{i+k-1}]}{x_{i+k} - x_i} \tag{4.26}$$

Resulta também de (4.23) que $f[x_0, \cdots, x_n]$ é uma função simétrica de $x_0, \cdots, x_n$, isto é, qualquer permutação dos x_i conduz ao mesmo valor da diferença dividida.

A fórmula (4.26) permite calcular as diferenças divididas usando um esquema recursivo que é em geral apresentado sob a forma duma tabela.

Tabela de diferenças divididas

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ,\ ]$
x_0	$f[x_0]$				
		$f[x_0, x_1]$			
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$		
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$	
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$		$f[x_0, x_1, x_2, x_3, x_4]$
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$	
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$		
		$f[x_3, x_4]$			
x_4	$f[x_4]$				

Na 1^a coluna escrevem-se os pontos x_i e na 2^a coluna os valores $f(x_i), 0 \leq i \leq n$. Depois de obter a terceira coluna da tabela, ou seja, as diferenças de 1^a ordem, por aplicação de (4.26) podem obter-se as diferenças de 2^a ordem na coluna seguinte, e assim sucessivamente. Por exemplo, depois de se calcular

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}, \quad f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1}$$

$$f[x_2, x_3] = \frac{f[x_3] - f[x_2]}{x_3 - x_2}, \quad f[x_3, x_4] = \frac{f[x_4] - f[x_3]}{x_4 - x_3}$$

pode-se obter

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}, \quad f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$$

Exemplo 4.4

Considere-se a função f dada pela tabela

x_i	0	1	2
f_i	1	1	2

Obtenha um valor aproximado de $f(1.5)$ utilizando um polinómio interpolador de grau ≤ 2.

Solução Utilizando a fórmula de Newton para o polinómio interpolador temos:

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)$$

onde $x_0 = 0, x_1 = 1, x_2 = 2$.

Cálculo da diferenças divididas:

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_0, x_1, x_2]$
0	1		
		$f[x_0, x_1] = \dfrac{1-1}{1-0} = 0$	
1	1		$f[x_0, x_1, x_2] = \dfrac{1-0}{2-0} = 0.5$
		$f[x_1, x_2] = \dfrac{2-1}{2-1} = 1$	
2	2		

Logo, $p_2(x) = 1 + 0.5x(x-1)$ e $f(1.5) \approx p_2(1.5) = 1.375$.

Relação entre as diferenças divididas e as derivadas de f

Seja $f \in C^{n+1}[a,b]$ e sejam $x_0, x_1, \cdots, x_n$ pontos distintos de $[a,b]$. Usando a fórmula de Newton com diferenças divididas, podemos escrever a igualdade

$$f(x) = f[x_0] + f[x_0,x_1](x-x_0) + \cdots + f[x_0,x_1,\cdots,x_n](x-x_0)\cdots(x-x_{n-1}) + \frac{f^{(n+1)}(\xi)}{(n+1)!}W(x), \tag{4.27}$$

onde $W(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$ e $\xi \in \text{int}(x_0,\ldots,x_n,x)$. O último termo do lado direito da equação representa o erro de interpolação (ver eq. (4.9)).

Comparando (4.27) com a fórmula de Taylor

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + f^{(n)}(x_0)\frac{(x-x_0)^n}{n!} + f^{(n+1)}(\gamma)\frac{(x-x_0)^{n+1}}{(n+1)!},$$

com γ entre x_0 e x, constatamos uma certa semelhança entre as duas fórmulas, sugerindo a *existência duma relação entre as diferenças divididas e as derivadas de f*. Com efeito, basta notar por exemplo que sendo

$$f[x_0,x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0},$$

resulta do Teorema do valor médio de Lagrange que $f[x_0,x_1]$ coincidirá com o valor da derivada f' nalgum ponto situado entre x_0 e x_1.

É possível generalizar este resultado às diferenças divididas de ordem superior.

Teorema 4.3 *Seja $f \in C^n[a,b]$ e $x_0, x_1, \cdots, x_n$ quaisquer pontos distintos de $[a,b]$. Então*

$$f[x_0,x_1,\cdots,x_n] = \frac{f^{(n)}(\xi)}{n!}, \qquad \xi \in \text{int}(x_0,\ldots,x_n,x_n). \tag{4.28}$$

Nota

Conclui-se de (4.28) que se $f \in \mathcal{P}_n$, as diferenças divididas de f de ordem superior a n são zero.

4.2 Aproximação segundo o critério dos mínimos quadrados

Nesta secção, consideraremos outros tipos de funções aproximadoras para $f(x)$, sem a exigência de que elas interpolem f. Com efeito, em muitas situações não é conveniente que a função aproximadora seja uma função interpoladora.

Por exemplo, suponhamos que a função f traduz uma relação entre duas grandezas físicas, x e $f(x)$.

Seja um conjunto de dados $(x_0, f_0), (x_1, f_1), \ldots, (x_n, f_n)$, onde os valores f_i $i = 0, 1, \ldots, n$ foram obtidos experimentalmente; ou seja, apenas dispomos dos valores $f(x_i) + \epsilon_i$, $i = 0, 1, \ldots, n$, onde os erros ϵ_i são desconhecidos. Neste caso, não seria razoável aproximar f por um polinómio passando por $(x_i, f(x_i) + \epsilon_i)$, $i = 0, 1, \ldots, n$, a não ser que os erros ϵ_i fossem pequenos. Uma solução mais adequada será utilizar uma função aproximadora, que poderá ser polinomial ou não, com a forma (aproximação *linear*)

$$g(x) = \alpha_0\phi_0(x) + \alpha_1\phi_1(x) + \ldots + \alpha_m\phi_m(x) \tag{4.29}$$

que se aproxime o "mais possível" de $f(x)$, sem contudo a imposição de passar pelos pontos (x_i, f_i).

Os parâmetros $\alpha_0, \alpha_1, \cdots, \alpha_m$ são determinados de modo que os desvios

$$d_i = f_i - g_i, \quad i = 0, 1, \cdots, n$$

onde $g_i = g(x_i)$, sejam tão pequenos quanto possível, segundo algum critério. Se impusermos

$$d_i = 0$$

caimos no caso da interpolação (o que não queremos).

Designando por $\overline{d}$, $\overline{f}$ e $\overline{g}$ os vectores de $\mathbb{R}^{n+1}$ com componentes d_i, f_i e g_i e escolhendo em $\mathbb{R}^{n+1}$ uma norma vectorial $\|.\|$, os parâmetros $\alpha_0, \alpha_1, \cdots, \alpha_m$ podem ser determinados de modo que

$$\|\overline{d}\| = \|\overline{f} - \overline{g}\|$$

seja mínima. Como exemplos de normas de $\mathbb{R}^{n+1}$ tem-se:

$$\|\overline{d}\|_1 = \sum_{i=0}^{n} |d_i|, \quad \|\overline{d}\|_\infty = \max_{0\leq i\leq n} |d_i|, \quad \|\overline{d}\|_2 = \left(\sum_{i=0}^{n} |d_i|^2\right)^{1/2}$$

De entre essas normas, a norma euclidiana $\|.\|_2$ é a que conduz a uma técnica mais simples. Então **no método dos mínimos quadrados**, os parâmetros $\alpha_0, \ldots, \alpha_m$ são determinados de modo a **minimizar**

$$||\overline{d}||_2 = \left(\sum_{i=0}^{n}(d_i)^2\right)^{1/2} = \left(\sum_{i=0}^{n}(f_i - g_i)^2\right)^{1/2}$$

Isso equivale a **minimizar** simplesmente a quantidade

$$D = ||\overline{d}||_2^2 = \sum_{i=0}^{n}(f_i - g_i)^2$$

Exemplo 4.5

Considere a seguinte tabela de valores de uma função f

x_i	2	4	6	8
$f(x_i)$	2	11	28	40

Determine a recta $g(x) = \alpha_0 + \alpha_1 x$ que melhor se aproxima de f segundo o critério dos mínimos quadrados. Considerando-se

$$D = \sum_{i=0}^{3}(f_i - \alpha_0 - \alpha_1 x_i)^2$$

como uma função das variáveis α_0 e α_1, pretende-se os valores de α_0 e α_1 que tornam D mínima. O sistema de estacionaridade para D é o seguinte

$$\frac{\partial D}{\partial \alpha_0} = 0; \quad \frac{\partial D}{\partial \alpha_1} = 0 \tag{4.30}$$

Depois de calculadas as derivadas parciais e substituídos os valores dos x_i e dos f_i, em (4.30), obtém-se o sistema (denominado *sistema normal*)

$$\begin{aligned} 4\alpha_0 + 20\alpha_1 &= 81 \\ 20\alpha_0 + 120\alpha_1 &= 536 \end{aligned} \tag{4.31}$$

que tem por solução: $\alpha_0 = -12.5,\ \alpha_1 = 6.55$. Consequentemente $g(x) = -12.5 + 6.55x$ é a recta pretendida.

Sistema de equações normais em termos de produtos internos

No Exemplo 4.5 considerámos um problema simples e fácil de resolver por derivação da "quantidade" a minimizar (a função D). O sistema linear (4.31) a que chegámos chama-se *sistema normal* e, a partir dele, foram obtidos os valores dos coeficientes que fazem parte da expressão de g. O objectivo desta secção é mostrar que o sistema normal pode ser expresso em termos de produtos internos de certos vectores, o que permite, na prática, uma maneira alternativa de chegar a esse sistema sem termos de passar pelo processo de derivação de D.

Para isso, consideremos o caso geral de obter uma função da forma (4.29), onde os valores dos parâmetros $\alpha_0, \ldots, \alpha_n$ pretendidos são os que tornam mínima a função

$$D(\alpha_0, \ldots, \alpha_m) = \sum_{i=0}^{n}(f_i - g_i)^2$$

Pela teoria das funções reais em $\mathbb{R}^{m+1}$, impomos agora as seguintes condições sobre D:

$$\frac{\partial D}{\partial \alpha_j} = 0, \quad j = 0, 1, \ldots, m$$

Fazendo como no exemplo acima, obtém-se o sistema:

$$\begin{aligned}
\frac{\partial D}{\partial \alpha_0} = 0 &\Leftrightarrow \sum_{i=0}^{n} 2[f_i - \alpha_0\phi_0(x_i) - \ldots \alpha_m\phi_m(x_i)](-\phi_0(x_i)) = 0 \\
\frac{\partial D}{\partial \alpha_1} = 0 &\Leftrightarrow \sum_{i=0}^{n} 2[f_i - \alpha_0\phi_0(x_i) - \ldots \alpha_m\phi_m(x_i)](-\phi_1(x_i)) = 0 \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
\frac{\partial D}{\partial \alpha_m} = 0 &\Leftrightarrow \sum_{i=0}^{n} 2[f_i - \alpha_0\phi_0(x_i) - \ldots \alpha_m\phi_m(x_i)](-\phi_m(x_i)) = 0
\end{aligned}$$

que se pode reescrever na forma

$$\begin{aligned}
&\alpha_0 \sum_{i=0}^{n} \phi_0(x_i)\phi_0(x_i) + \alpha_1 \sum_{i=0}^{n} \phi_0(x_i)\phi_1(x_i) + \ldots + \alpha_m \sum_{i=0}^{n} \phi_0(x_i)\phi_m(x_i) = \sum_{i=0}^{n} f_i\phi_0(x_i) \\
&\alpha_0 \sum_{i=0}^{n} \phi_1(x_i)\phi_0(x_i) + \alpha_1 \sum_{i=0}^{n} \phi_1(x_i)\phi_1(x_i) + \ldots + \alpha_m \sum_{i=0}^{n} \phi_1(x_i)\phi_m(x_i) = \sum_{i=0}^{n} f_i\phi_1(x_i) \\
&\qquad\vdots \qquad\qquad\qquad\vdots \qquad\qquad\qquad\vdots \qquad\qquad\vdots \qquad\qquad\vdots \\
&\alpha_0 \sum_{i=0}^{n} \phi_m(x_i)\phi_0(x_i) + \alpha_1 \sum_{i=0}^{n} \phi_m(x_i)\phi_1(x_i) + \ldots + \alpha_m \sum_{i=0}^{n} \phi_m(x_i)\phi_m(x_i) = \sum_{i=0}^{n} f_i\phi_m(x_i)
\end{aligned}$$

Trata-se de um sistema linear de $m+1$ equações nas $m+1$ incógnitas $\alpha_0, \ldots, \alpha_m$, a que se chama *sistema de equações normais* ou *sistema normal*.

Sejam os vectores de $\mathbb{R}^{n+1}$:

$$\overline{\phi}_0 = [\phi_0(x_0), \ldots, \phi_0(x_n)]^T, \overline{\phi}_1 = [\phi_1(x_0), \ldots, \phi_1(x_n)]^T, \ldots, \overline{\phi}_m = [\phi_m(x_0), \ldots, \phi_m(x_n)]^T$$

Com o produto interno usual em $\mathbb{R}^{n+1}$:

$$\overline{u}, \overline{v} \in \mathbb{R}, \quad < \overline{u}, \overline{v} >= \sum_{i=0}^{n} u_i v_i$$

o sistema de equações normais toma a forma

$$\begin{bmatrix} <\overline{\phi}_0,\overline{\phi}_0> & <\overline{\phi}_0,\overline{\phi}_1> & <\overline{\phi}_0,\overline{\phi}_2> & \cdots & <\overline{\phi}_0,\overline{\phi}_m> \\ <\overline{\phi}_1,\overline{\phi}_0> & <\overline{\phi}_1,\overline{\phi}_1> & <\overline{\phi}_1,\overline{\phi}_2> & \cdots & <\overline{\phi}_1,\overline{\phi}_m> \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ <\overline{\phi}_m,\overline{\phi}_0> & <\overline{\phi}_m,\overline{\phi}_1> & <\overline{\phi}_m,\overline{\phi}_2> & \cdots & <\overline{\phi}_m,\overline{\phi}_m> \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} <\overline{f},\overline{\phi}_0> \\ <\overline{f},\overline{\phi}_1> \\ \vdots \\ <\overline{f},\overline{\phi}_m> \end{bmatrix} \tag{4.32}$$

Trata-se de um sistema não homogéneo de $m+1$ equações lineares nas $m+1$ incógnitas $\alpha_k, k = 0, 1, \ldots, m$. Note-se que a matriz do sistema é simétrica. Prova-se que se as funções $\phi_0(x), \phi_1(x), \ldots, \phi_m(x)$ forem linearmente independentes então o sistema tem uma solução única, a qual é ponto de mínimo de D.

Exemplo 4.6

Considere a seguinte tabela de valores de uma função f

x_i	-1	0	1	5
$f(x_i)$	-2.0	-2.0	0	28.0

Determine a função $g(x) = \alpha_0(x-1) + \alpha_1(x-1)^2$ que melhor se ajusta à função tabelada no sentido dos mínimos quadrados.

Solução

Vemos que a função g é do tipo (4.29). Mais precisamente, $g(x) = \alpha_0\phi_0(x) + \alpha_1\phi_1(x)$ onde $\phi_0(x) = x-1$ e $\phi_1(x) = (x-1)^2$. Para que $g(x)$ aproxime $f(x)$ segundo o critério dos mínimos quadrados, os valores dos parâmetros α_0 e α_1 são os que minimizam a função

$$D = \sum_{i=0}^{3} (f_i - \alpha_0(x_i - 1) - \alpha_1(x_i - 1)^2)^2$$

Esses valores são obtidos pela solução do sistema normal (4.32)

$$\begin{bmatrix} (\overline{\phi}_0,\overline{\phi}_0) & (\overline{\phi}_1,\overline{\phi}_0) \\ (\overline{\phi}_0,\overline{\phi}_1) & (\overline{\phi}_1,\overline{\phi}_1) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} (\overline{f},\overline{\phi}_0) \\ (\overline{f},\overline{\phi}_1) \end{bmatrix}$$

onde

$(\overline{\phi}_0,\overline{\phi}_0) = \sum_{i=0}^{3} \phi_0^2(x_i) = \phi_0^2(x_0) + \phi_0^2(x_1) + \phi_0^2(x_2) + \phi_0^2(x_3)$
$(\overline{\phi}_1,\overline{\phi}_0) = \sum_{i=0}^{3} \phi_1(x_i)\phi_0(x_i) = \phi_1(x_0)\phi_0(x_0) + \phi_1(x_1)\phi_0(x_1) + \phi_1(x_2)\phi_0(x_2) + \phi_1(x_3)\phi_0(x_3)$
$(\overline{\phi}_1,\overline{\phi}_1) = \sum_{i=0}^{3} \phi_1^2(x_i) = \phi_1^2(x_0) + \phi_1^2(x_1) + \phi_1^2(x_2) + \phi_1^2(x_3)$
$(\overline{f},\overline{\phi}_0) = \sum_{i=0}^{3} f(x_i)\phi_0(x_i) = f(x_0)\phi_0(x_0) + f(x_1)\phi_0(x_1) + f(x_2)\phi_0(x_2) + f(x_3)\phi_0(x_3)$
$(\overline{f},\overline{\phi}_0) = \sum_{i=0}^{3} f(x_i)\phi_0(x_i) = f(x_0)\phi_1(x_0) + f(x_1)\phi_1(x_1) + f(x_2)\phi_1(x_2) + f(x_3)\phi_1(x_3)$

Logo, temos

$$
\begin{aligned}
(\overline{\phi}_0, \overline{\phi}_0) &= (-2)^2 + (-1)^2 + 0^2 + (4)^2 = 21 \\
(\overline{\phi}_1, \overline{\phi}_0) &= (-2)^3 + (-1)^3 + 0^3 + (4)^3 = 55 \\
(\overline{\phi}_1, \overline{\phi}_1) &= (-2)^4 + (-1)^4 + 0^4 + (4)^4 = 273 \\
(\overline{f}, \overline{\phi}_0) &= (-2)(-2) + (-1)(-2) + 0 + (4)(28) = 118 \\
(\overline{f}, \overline{\phi}_1) &= (-2)^2(-2) + (-1)^2(-2) + 0 + (4)^2(28) = 438
\end{aligned}
$$

e somos conduzidos ao sistema linear

$$
\begin{bmatrix} 21 & 55 \\ 55 & 273 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 118 \\ 438 \end{bmatrix}
$$

que tem como solução: $\alpha_0 = 3.0$ e $\alpha_1 = 1.0$. Finalmente, notemos que poderíamos usar o processo utilizado no Exemplo 4.5 para chegar ao sistema normal.

Exemplo 4.7

Resolva o problema de mínimos quadrados do Exemplo 4.5, utilizando a abordagem vectorial. Ou seja, defina os vectores $\overline{f}, \overline{\phi}_j, j = 0, 1$ e obtenha o sistema normal (4.32). Deverá ser equivalente ao sistema já obtido (4.31).

5 Integração numérica

Seja $f \in C[a,\ b]$ e considere o integral definido

$$I(f) = \int_a^b f(x)dx.$$

Suponha que $f(x)$ é uma "função complicada" (não se conhece uma primitiva) ou que f é conhecida somente num número finito de pontos. Uma aproximação para $I(f)$ pode ser obtida como segue.

Sejam $x_0 = a < x_1 < \cdots < x_n = b,\ \ n+1$ pontos em $[a,\ b]$ e seja $p_n(x)$ o polinómio de grau $\leq n$ interpolador de $f(x)$ nesses pontos. As fórmulas que consideraremos baseiam-se na aproximação

$$I(f) \simeq I(p_n)$$

Se $f \in C^{(n+1)}[a,b]$, então $f(x) = p_n(x) + e_n(x)$, onde $e_n(x)$ é o erro de interpolação, dado por (4.9). Ou seja,

$$f(x) = p_n(x) + (x-x_0)(x-x_1)\cdots(x-x_n)\frac{f^{(n+1)}(\xi(x))}{(n+1)!}, \quad \text{onde} \quad \xi(x) \in [a,\ b].$$

Logo,

$$\int_a^b f(x)dx = \int_a^b p_n(x)dx + \underbrace{\int_a^b e_n(x)dx}_{E_n(f)}, \tag{5.1}$$

onde

$$E_n(f) := \int_a^b e_n(x)dx = \int_a^b (x-x_0)(x-x_1)\cdots(x-x_n)\frac{f^{(n+1)}(\xi(x))}{(n+1)!}dx \tag{5.2}$$

Se desprezarmos o termo $E_n(f)$ obtemos a aproximação

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)dx; \tag{5.3}$$

o termo $E_n(f)$ é o erro cometido ao fazer-se essa aproximação (é um erro de truncatura), também chamado *erro de integração*. Vemos de (5.2) que o erro de integração é o integral do erro de interpolação associado a p_n.

Vamos em seguida obter uma expressão para $\int_a^b p_n(x)dx$.

Utilizando a fórmula de Lagrange do polinómio interpolador, vem:

$$\begin{aligned}\int_a^b p_n(x)dx &= \int_a^b \left(\sum_{k=0}^{n} l_k(x)f(x_k)\right)dx \\ &= \sum_{k=0}^{n}\int_a^b l_k(x)f(x_k)dx = \sum_{k=0}^{n}\underbrace{\left(\int_a^b l_k(x)dx\right)}_{A_k} f(x_k).\end{aligned}$$

Substituindo em (5.3) obtém-se a aproximação

$$\int_a^b f(x)dx \approx \sum_{k=0}^{n} A_k f(x_k) \tag{5.4}$$

Chama-se *fórmula de quadratura* à soma no lado direito; os pontos $x_0, \cdots, x_n$ são os *nós de integração* e os

$$A_k = \int_a^b l_k(x)dx, 0 \leq k \leq n,$$

são chamados os *pesos da fórmula de quadratura.*

5.1 Fórmulas de Newton-Cotes fechadas

As fórmulas de Newton-Cotes fechadas são obtidas através de (5.4), considerando os pontos $x_0, x_1, \cdots, x_n$, igualmente espaçados em $[a,\ b]$, com $x_0 = a$ e $x_n = b$, ou seja,

$$x_j = a + jh, \quad h = \frac{b-a}{n}, \quad j = 0, 1, \cdots, n$$

As fórmulas de Newton-Cotes fechadas mais conhecidas são a "Fórmula dos Trapézios" e a "Fórmula de Simpson".

Fórmula dos Trapézios

Esta fórmula resulta de aproximar $f(x)$ pelo polinómio interpolador em $x_0 = a$ e $x_1 = b$. Nesse caso, fazendo $n = 1$ em (5.4), obtemos:

$$\begin{aligned}\int_a^b f(x)dx &\simeq A_0 f(x_0) + A_1 f(x_1) \quad \text{onde} \quad x_0 = a, x_1 = b \text{ e } h = x_1 - x_0 \\ A_0 &= \int_a^b l_0(x)dx, \qquad A_1 = \int_a^b l_1(x)dx\end{aligned}$$

Cálculo de A_0 e A_1

$$\begin{aligned} A_0 &= \int_a^b \frac{(x-x_1)}{(x_0-x_1)}dx = -\frac{1}{h}\int_a^b (x-x_1)dx = -\frac{1}{h}\frac{(x-x_1)^2}{2}|_{x_0}^{x_1} \\ &= \frac{1}{h}\frac{(x_0-x_1)^2}{2} = \frac{1}{h}\frac{h^2}{2} = \frac{h}{2} \\ A_1 &= \int_a^b \frac{(x-x_0)}{(x_1-x_0)}dx = \frac{1}{h}\int_a^b (x-x_0)dx = \frac{1}{h}\frac{(x-x_0)^2}{2}|_{x_0}^{x_1} \\ &= \frac{1}{h}\frac{(x_1-x_0)^2}{2} = \frac{1}{h}\frac{h^2}{2} = \frac{h}{2} \end{aligned}$$

Logo, a fórmula de quadratura é dada por:

$$T(f) = \frac{h}{2}f(x_0) + \frac{h}{2}f(x_1) = \frac{h}{2}\left[f(x_0) + f(x_1)\right]$$

Erro de integração: $E^T(f) = I(f) - T(f)$

Como vimos, a fórmula dos trapézios foi deduzida aproximando $I(f) = \int_a^b f(x)dx$ pelo integral do polinómio interpolador de $f(x)$ nos pontos $x_0 = a$ e $x_1 = b$. Para obtermos uma expressão do erro cometido procedemos como segue:

$$f(x) = p_1(x) + (x-x_0)(x-x_1)\frac{f''(\xi(x))}{2!}$$

logo,

$$\begin{aligned} &\int_a^b f(x)dx = \int_a^b p_1(x)dx + \int_a^b (x-x_0)(x-x_1)\frac{f''(\xi(x))}{2!}dx \\ &\Longrightarrow \int_a^b f(x)dx - \int_a^b p_1(x)dx = \int_a^b (x-x_0)(x-x_1)\frac{f''(\xi(x))}{2!}dx \\ &\Longrightarrow I(f) - T(f) = \int_a^b (x-x_0)(x-x_1)\frac{f''(\xi(x))}{2!}dx \end{aligned}$$

Portanto o erro é dado por:

$$E^T(f) = \int_a^b (x-x_0)(x-x_1)\frac{f''(\xi(x))}{2!}dx \tag{5.5}$$

É claro que poderíamos ter usado directamente (5.2), con $n = 1$. Para simplificarmos a equação acima, faremos uso do seguinte resultado da Análise:

Teorema do valor médio para integrais: Se $f \in C[a,\ b]$ e g é uma função integrável em $[a,\ b]$ que não muda de sinal em $[a,\ b]$ então existe $z \in (a,\ b)$ tal que

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

No caso do integral (5.5), tem-se $W(x) = (x - x_0)(x - x_1) \leq 0 \ \forall \ x \ \in \ [x_0, \ x_1]$. Aplicando o teor. valor médio para integrais

$$\int_a^b (x - x_0)(x - x_1)\frac{f''(\xi(x))}{2!}dx = \frac{f''(c)}{2!}\int_a^b (x - x_0)(x - x_1)dx, \ z \in [x_0, x_1],$$

onde

$$\begin{aligned}\int_a^b (x - x_0)(x - x_1)dx &= (x - x_0)\frac{(x - x_1)^2}{2}|_{x_0}^{x_1} - \int_a^b \frac{(x - x_1)^2}{2}dx \\ &= -\frac{(x - x_1)^3}{6}|_{x_0}^{x_1} = \frac{(x_0 - x_1)^3}{6} = -\frac{h^3}{6}\end{aligned}$$

Obtem-se assim a seguinte fórmula para o erro da regra dos Trapézios:

$$E^T(f) = -\frac{h^3}{12}f''(c) \quad z \in [a, b] \ .$$

Fórmula de Simpson

A fórmula de Simpson é obtida ao aproximarmos $I(f)$ por $I(p_2)$, onde $p_2(x)$ é o polinómio interpolador de $f(x)$ nos pontos $x_0 = a, x_1 = \frac{a+b}{2}$ e $x_2 = b$. Nesse caso temos:

$$f(x) = p_2(x) + (x - x_0)(x - x_1)(x - x_2)\frac{f^{(3)}(\xi(x))}{3!}, \quad \text{com} \quad \xi(x) \ \in \ [a, \ b],$$

$$\text{com } p_2(x) = l_0(x)f(x_0) + l_1(x)f(x_1) + l_2(x)f(x_2) \ .$$

Logo,

$$\int_a^b f(x)dx = \int_a^b p_2(x)dx + \int_a^b (x - x_0)(x - x_1)(x - x_2)\frac{f^{(3)}(\xi(x))}{3!}, \quad \text{onde} \quad \xi(x) \ \in \ [a, \ b] \tag{5.6}$$

e, desprezando o segundo termo do lado direito, que representa o erro de integração, resulta a seguinte aproximação para o integral de f

$$\int_a^b f(x)dx \approx \int_a^b p_2(x)dx$$

Calculando $\int_a^b p_2(x)dx$ vem:

$$\begin{aligned}\int_a^b p_2(x)dx &= \int_a^b l_0(x)f(x_0)dx + \int_a^b l_1(x)f(x_1)dx + \int_a^b l_2(x)f(x_2)dx \\ &= f(x_0)\int_a^b l_0(x)dx + f(x_1)\int_a^b l_1(x)dx + f(x_2)\int_a^b l_2(x)dx\end{aligned}$$

Cálculo de A_0, A_1, A_2

$$\begin{aligned}
A_0 &= \int_a^b \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}dx = \frac{1}{2h^2}\int_a^b (x-x_1)(x-x_2)dx \\
&= \frac{1}{2h^2}\left[(x-x_1)\frac{(x-x_2)^2}{2}|_{a=x_0}^{b=x_2} - \int_a^b \frac{(x-x_2)^2}{2}dx\right] \\
&= \frac{1}{2h^2}\left[(x-x_1)\frac{(x-x_2)^2}{2}|_{a=x_0}^{b=x_2} - \frac{(x-x_2)^3}{6}|_{a=x_0}^{b=x_2}\right] \\
&= \frac{1}{2h^2}\left[h\frac{4h^2}{2} - \frac{8h^3}{6}\right] = \frac{1}{2h^2}\left[2h^3 - \frac{4}{3}h^3\right] = \frac{1}{2h^2}\left[\frac{2}{3}h^3\right] = \frac{1}{3}h
\end{aligned}$$

$$\begin{aligned}
A_1 &= \int_a^b \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}dx - \frac{1}{h^2}\int_a^b (x-x_0)(x-x_2)dx \\
&= -\frac{1}{h^2}\left[(x-x_0)\frac{(x-x_2)^2}{2}|_{a=x_0}^{b=x_2} - \int_a^b \frac{(x-x_2)^2}{2}dx\right] \\
&= -\frac{1}{h^2}\left[(x-x_0)\frac{(x-x_2)^2}{2}|_{a=x_0}^{b=x_2} - \frac{(x-x_2)^3}{6}|_{a=x_0}^{b=x_2}\right] \\
&= -\frac{1}{h^2}\left[-\frac{8h^3}{6}\right] = \frac{4}{3}h
\end{aligned}$$

$$\begin{aligned}
A_2 &= \int_a^b \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}dx = -\frac{1}{2h^2}\int_a^b (x-x_0)(x-x_1)dx \\
&= \frac{1}{2h^2}\left[(x-x_0)\frac{(x-x_1)^2}{2}|_{a=x_0}^{b=x_2} - \int_a^b \frac{(x-x_1)^2}{2}dx\right] \\
&= \frac{1}{2h^2}\left[(x-x_0)\frac{(x-x_1)^2}{2}|_{a=x_0}^{b=x_2} - \frac{(x-x_1)^3}{6}|_{a=x_0}^{b=x_2}\right] \\
&= \frac{1}{2h^2}\left[2h\frac{h^2}{2} - \left(\frac{h^3}{6} - \left(-\frac{h^3}{6}\right)\right)\right] = \frac{1}{2h^2}\left[h^3 - \frac{1}{3}h^3\right] = \frac{1}{3}h
\end{aligned}$$

Portanto, a fórmula de Simpson é dada por:

$$S(f) = \frac{1}{3}hf(x_0) + \frac{4}{3}hf(x_1) + \frac{1}{3}hf(x_2) = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)].$$

Observamos que esta fórmula é da forma (5.4) com

$$A_0 = \frac{h}{3}, \qquad A_1 = \frac{4h}{3}, \qquad A_2 = \frac{h}{3}\ .$$

Erro de integração

Tal como no caso da fórmula dos trapézios, o erro de integração é dado pelo integral do erro de interpolação (compare-se (5.6)). Ou seja, supondo $f \in C^2[a,b]$,

$$\begin{aligned} E^S(f) &= \int_a^b f(x)dx - \int_a^b p_2(x)dx \\ &= \int_a^b (x-x_0)(x-x_1)(x-x_2)\frac{f^{(3)}(\xi(x))}{3!}, \quad \text{onde} \quad \xi(x) \in [a,\ b] \end{aligned}$$

Pode contudo mostrar-se que, se $f \in C^4[a,b]$, é válida a seguinte fórmula para o erro de integração:

$$E^S(f) = -\frac{h^5}{90}f^4(\xi)\ , \xi \in [a,\ b]$$

Um **majorante para o erro** é então dado por:

$$|E^S(f)| \leq \frac{h^5}{90}\max_{x\in[a,b]}|f^4(x)|$$

5.2 Fórmulas de quadratura compostas

A obtenção de uma fórmula de quadratura para um intervalo $[a,\ b]$ pode ser feita de 2 modos:

1. Tomando um grande número de nós de integração de modo a que a fórmula
$$Q(f) = \sum_{j=0}^{n} A_j f(x_j)$$
represente o integral do polinómio interpolador em $x_0, x_1, \cdots, x_n$. Essa técnica não convém visto que o polinómio terá grau elevado e a interpolação pode não resultar em uma boa aproximação para $f(x)$.

2. Dividir o intervalo $[a,\ b]$ em vários subintervalos $[x_j,\ x_{j+1}]$, aplicar uma fórmula com poucos nós em cada subintervalo e somar os resultados de modo a obter uma aproximação para $I(f)$. Obtém-se assim uma fórmula ou regra de *quadratura composta*.

Sejam $x_0, x_1, \cdots, x_N,\ N+1$ pontos em $[a,\ b]$ com $x_0 = a$ e $x_N = b$. Então,

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{N-1}}^{x_N} f(x)dx = \sum_{j=0}^{N}\int_{x_j}^{x_{j+1}} f(x)dx$$

Assim, para obter uma aproximação para $I(f)$ aplica-se a cada integral $\int_{x_j}^{x_{j+1}} f(x)dx$ uma fórmula de quadratura com poucos pontos e somam-se as contribuições.

Fórmula dos Trapézios composta

Esta fórmula é obtida dividindo o intervalo $[a,\ b]$ em N subintervalos e aproximando cada integral $\int_{x_j}^{x_{j+1}} f(x)dx$ pela regra dos trapézios.

Sejam $x_j = a + jh, \quad x_0 = a, \quad h = \frac{b-a}{N}, \quad j = 0, 1, \cdots, N$. Então,

$$\begin{aligned}
\int_a^b f(x)dx &= \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} f(x)dx \approx \sum_{j=0}^{N} \left[\frac{h}{2}\left(f(x_j) + f(x_{j+1})\right)\right] \\
&= \left[\frac{h}{2}\left(f(x_0) + f(x_1)\right)\right] + \left[\frac{h}{2}\left(f(x_1) + f(x_2)\right)\right] + \cdots + \left[\frac{h}{2}\left(f(x_{N-1}) + f(x_N)\right)\right] \\
&= \frac{h}{2}\left[f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{N-1}) + f(x_N)\right] \\
&= \frac{h}{2}\left[\left(f(x_0) + f(x_N)\right) + \sum_{j=1}^{N-1} 2f(x_j)\right] = T_N(f)
\end{aligned}$$

Erro de integração:

Começamos por notar que o erro $E_N^T(f) = I(f) - T^N(f)$ é a soma dos erros cometidos em cada subintervalo $[x_j,\ x_{j+1}]$. Logo, supondo $f \in C^2[a,\ b]$,

$$E_N^T(f) = I(f) - T_N(f) = \sum_{j=0}^{N-1} -\frac{h^3}{12} f''(\xi_j) \tag{5.7}$$

Majorante para o erro

Um majorante para o erro pode ser facilmente obtido de (5.7). Tomando módulos, vem

$$|E_N^T(f)| \leq \frac{h^3}{12} \sum_{j=0}^{N-1} |f''(\xi_j)| \leq N\ h^3 \max_{x\in[a,b]} |f''(x)| \tag{5.8}$$

Usando a relação $h = (b-a)/N$, também se tem

$$|E_N^T(f)| \leq \frac{(b-a)}{12} h^2 \max_{x\in[a,b]} |f''(x)| \tag{5.9}$$

Fórmula do erro

Vejamos agora como simplificar a fórmula (5.7). Tem-se

$$E_N^T(f) = \sum_{j=0}^{N-1} -\frac{h^3}{12} f''(\xi_j) = -N\frac{h^3}{12} \sum_{j=0}^{N-1} \frac{f''(\xi_j)}{N} \quad \text{onde } \xi_j \ \in\ [x_j,\ x_{j+1}] \tag{5.10}$$

Como $f\ \in\ C^2[a,\ b]$, então existe $m = \min\limits_{x\in[a,b]} f''(x)$ e $M = \max\limits_{x\in[a,b]} f''(x)$. Logo,

$$m \leq f''(\xi_j) \leq M \quad j = 0, 1, \cdots, N-1. \implies Nm \leq \sum_{j=0}^{N-1} f''(\xi_j) \leq NM$$

$$\implies m \leq \sum_{j=0}^{N-1} \frac{f''(\xi_j)}{N} \leq M \implies \min_{x\in[a,b]} f''(x) \leq \sum_{j=0}^{N-1} \frac{f''(\xi_j)}{N} \leq \max_{x\in[a,b]} f''(x)$$

Pelo teorema do valor intermédio, existe $\xi\ \in\ [a,\ b]$ tal que

$$f''(\xi) = \sum_{j=0}^{N-1} \frac{f''(\xi_j)}{N}$$

e substituindo em (5.7) obtemos (visto que $h = (b-a)/N$):

$$E_N^T(f) = -N\frac{h^3}{12} f''(\xi) = -\frac{(b-a)}{12} h^2 f''(\xi) \ \text{ onde } \ \xi\ \in\ [a,\ b]\ . \tag{5.11}$$

Note-se que as fórmulas de majoração (5.8) e (5.9) podem também ser obtidas da equação anterior (5.11).

Fórmula de Simpson composta

Esta fórmula é obtida aplicando a fórmula de Simpson em cada subintervalo da forma $[x_j,\ x_{j+2}]$. Sejam $x_0 = a, x_1, \cdots, x_{N-2}, x_{N-1}, x_N = b$ onde N é um número par. Então,

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \cdots + \int_{x_{N-2}}^{x_N} f(x)dx = \sum_{j=1}^{N/2} \int_{x_{2j}-2}^{x_{2j}} f(x)dx$$

e, aplicando a formula de Simpson a cada integral, obtemos

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{3}[f(x_0)+4f(x_1)+f(x_2)]+\frac{h}{3}[f(x_2)+4f(x_3)+f(x_4)]\\ &\quad +\cdots+\frac{h}{3}[f(x_{N-2})+4f(x_{N-1})+f(x_N)]\\ &= \frac{h}{3}\left[f(x_0)+f(x_N)+4\sum_{j=1}^{N/2}f(x_{2j-1})+2\sum_{j=1}^{N/2-1}f(x_{2j})\right]=S_N(f)\end{aligned}$$

Erro de integração $I(f)-S_N(f)$

De maneira análoga como se fez na fórmula dos trapézios composta, pode-se mostrar que é válida a seguinte fórmula do erro para a fórmula de Simpson composta:

$$E_N^S(f)=-\frac{h^5}{90}\sum_{j=1}^{N/2}f^{(4)}(\xi_i)=-\frac{Nh^5}{180}f^{(4)}(\xi)\qquad \xi\ \in\ [a,\ b].$$

Majorante para o erro

Um majorante para o erro é dado por:

$$|E_N^S(f)|\leq\frac{Nh^5}{180}\max_{x\in[a,b]}|f^{(4)}(x)|$$

5.3 Grau de precisão de uma fórmula de quadratura

Seja a igualdade $I(f)=Q(f)+E^Q(f)$ onde $Q(f)$ é uma fórmula de quadratura que aproxima $I(f)$ e $E^Q(f)$ é o erro de integração. O grau de precisão da fórmula $Q(f)$ é um número r tal que:

$$E^Q(P_r)=0\ \ \forall P_r,\ \text{polinómio de grau}\leq r,\qquad \text{e}\qquad E^Q(x^{r+1})\neq 0$$

Exemplo 1: Ao aproximarmos $I(f)$ pela fórmula dos trapézios temos:

$$\int_a^b f(x)dx=\frac{b-a}{2}[f(a)+f(b)]-\frac{(b-a)^3}{12}f''(\xi)\qquad \xi\ \in\ [a,\ b].$$

Vemos que se $f(x)=P_1(x)$ (um qualquer polinómio de grau ≤ 1) então $f''(x)=0\forall x$. Isto implica que $f''(\xi)=0$ e logo $E^T(f)=0$. Ou seja, a fórmula dos trapézios dá o valor exacto para os polinómios de grau ≤ 1, pelo que o grau de precisão de $T(f)$ é um número $r\geq 1$. Mas se $f(x)=x^2$ temos que

$$E^T(f)=-\frac{(b-a)^3}{12}f''(\xi)=-\frac{(b-a)^3}{12}2\neq 0\qquad I(f)\neq T(f)$$

Ou seja, a fórmula dos trapézios não dá o valor exacto para todo o polinómio de grau ≤ 2 (falha o x^2) e, portanto, o grau de precisão de $T(f)$ é exactamente $r = 1$.

Exemplo 2: Para a regra de Simpson tem-se:

$$\int_a^b f(x)dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] - \left(\frac{b-a}{2}\right)^5 \frac{1}{90} f^{(4)}(\xi) \qquad \xi \in [a,\ b]\ .$$

Como $f^{(4)}(x) = 0$ se $f(x)$ é um polinómio qualquer de grau ≤ 3 então $f^{(4)}(\xi) = 0$ e tem-se $E^S(f) = 0$. Ou seja, a regra $S(f)$ é exacta para polinómios de grau ≤ 3 e portanto o seu grau de precisão é um número $r \geq 3$. Investiguemos se pode ser superior a 3. Se for $f(x) = x^4$ temos

$$f^{(4)}(x) = 4!\ \forall\ x \implies E_S(f) = -\left(\frac{b-a}{2}\right)^5 \frac{1}{90} 4! \neq 0$$

Isso significa que $S(f)$ não é exacta para todo o polinómio de grau 4 (falha no x^4). Então conclui-se que o grau de precisão de $S(f)$ é $r = 3$.

5.4 Métodos dos coeficientes indeterminados

Sejam $x_0, x_1, \cdots, x_n$, $(n+1)$pontos em $[a,\ b]$ e suponhamos conhecidos os valores de f nesses pontos. As fórmulas de quadratura que considerámos, do tipo

$$Q(f) = \sum_{j=0}^{n} A_j f(x_j) \approx \int_a^b f(x)dx. \tag{5.12}$$

foram deduzidas integrando o polinómio interpolador. Isto é, por construção, $Q(f) = I(p_n)$, onde p_n é o polinómio interpolador de grau $\leq n$ de f. O método dos coeficientes indeterminados é uma maneira alternativa de obter os pesos da fórmula de quadratura.

Comecemos por notar que, se f for um polinómio de grau $\leq n$, ou seja $f(x) = P_n(x)$, então a fórmula Q dó valor **exacto** do integral de f. Tem-se assim

$$I(P_n) = Q(P_n),\ \forall P_n \tag{5.13}$$

A igualdade acima resulta da unicidade do polinómio interpolador: sendo f um polinómio de grau $\leq n$, então f coincide com o seu polinómio interpolador p_n. Ou seja, tem-se $f(x) = p_n(x)$, donde $I(f) = I(p_n) = Q(f)$.

Veremos posteriormente que, impondo a condição $I(P_n) = Q(P_n)$, para qualquer polinómio de grau $\leq n$, resulta num sistema de equações lineares nos coeficientes A_j.

Comecemos por provar o seguinte resultado.

Lema 5.1 *A fórmula de quadratura definida por (5.12) é uma aplicação linear.*

Dem.:

Sejam $x_0, x_1, \cdots, x_n, \in [a,\ b]$, f e g duas funções contínuas em $[a,\ b]$ e $\lambda \in I\!R$. Então

$$\begin{aligned} Q(f+g) &= \sum_{j=0}^{n} A_j[(f+g)(x_j)] = \sum_{j=0}^{n} A_j[f(x_j)+g(x_j)] = \sum_{j=0}^{n}[A_j f(x_j) + A_j g(x_j)] \\ &= \sum_{j=0}^{n} A_j f(x_j) + \sum_{j=0}^{n} A_j g(x_j) = Q(f) + Q(g) \\ Q(\lambda f) &= \sum_{j=0}^{n} A_j[(\lambda f)(x_j)] = \sum_{j=0}^{n} A_j \lambda f(x_j) = \lambda \sum_{j=0}^{n} A_j f(x_j) = \lambda Q(f) \end{aligned}$$

Obtenção de $A_0, A_1, \cdots, A_n$

Seja $P_n(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 \cdots + \alpha_n x^n$ um polinómio de grau $\leq n$. Então

$$Q(P_n) = Q(\alpha_0 + \alpha_1 x + \alpha_2 x^2 \cdots + \alpha_n x^n) = \alpha_0 Q(1) + \alpha_1 Q(x) + \alpha_2 Q(x^2) + \cdots + \alpha_n Q(x^n)$$

e

$$I(P_n) = I(\alpha_0 + \alpha_1 x + \alpha_2 x^2 \cdots + \alpha_n x^n) = \alpha_0 I(1) + \alpha_1 I(x) + \alpha_2 I(x^2) + \cdots + \alpha_n I(x^n)$$

Então, para se ter uma fórmula com a propriedade (5.13)

$$Q(P_n) = I(P_n), \forall P_n$$

é necessário e suficiente que verifique:

$$Q(x^k) = I(x^k), \quad k = 0, 1, \cdots, n$$

Ou seja, para que a fórmula Q dê o valor exacto do integral de qualquer polinómio de grau $\leq n$ é necessário e suficiente que dê o valor exacto dos integrais dos monómios $1, x, x^2, \ldots, x^n$. Impondo esta condição somos conduzidos ao sistema de equações:

$$
\begin{aligned}
Q(1) &= \sum_{j=0}^{n} A_j(1) = I(1) = \int_a^b dx = b-a \\
&\Longrightarrow \quad A_0 + A_1 + \cdots + A_n = b-a \\
Q(x) &= \sum_{j=0}^{n} A_j(x_j) = I(x) = \int_a^b x dx = \frac{x^2}{2}|_a^b = \frac{b^2-a^2}{2} \\
&\Longrightarrow \quad A_0x_0 + A_1x_1 + \cdots + A_nx_n = \frac{b^2-a^2}{2} \\
Q(x^2) &= \sum_{j=0}^{n} A_j(x_j^2) = I(x^2) = \int_a^b x^2 dx = \frac{x^3}{3}|_a^b = \frac{b^3-a^3}{3} \\
&\Longrightarrow \quad A_0x_0^2 + A_1x_1^2 + \cdots + A_nx_n^2 = \frac{b^3-a^3}{3} \\
\vdots &= \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
Q(x^n) &= \sum_{j=0}^{n} A_j(x_j^n) = I(x^n) = \int_a^b x^n dx = \frac{x^{n+1}}{n+1}|_a^b = \frac{b^{n+1}-a^{n+1}}{n+1} \\
&\Longrightarrow \quad A_0x_0^n + A_1x_1^n + \cdots + A_nx_n^n = \frac{b^{n+1}-a^{n+1}}{n+1}
\end{aligned}
$$

Pode-se reescrever este sistema numa forma mais compacta

$$
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
x_0 & x_1 & x_2 & \cdots & x_n \\
x_0^2 & x_1^2 & x_2^2 & \cdots & x_n^2 \\
\vdots & \vdots & \vdots & & \vdots \\
\vdots & \vdots & \vdots & & \vdots \\
x_0^n & x_1^n & x_2^n & \cdots & x_n^n
\end{bmatrix}
\begin{bmatrix}
A_0 \\ A_1 \\ A_2 \\ \vdots \\ \vdots \\ A_n
\end{bmatrix}
=
\begin{bmatrix}
b-a \\
\dfrac{b^2-a^2}{2} \\
\dfrac{b^3-a^3}{3} \\
\vdots \\
\vdots \\
\dfrac{b^{n+1}-a^{n+1}}{n+1}
\end{bmatrix}
\tag{5.14}
$$

A matriz acima é conhecida como "Matriz de Vandermonde" e sabe-se que, se os pontos forem distintos, ela é não-singular $\Longrightarrow$ esse sistema tem solução única $A_j,\ j = 0, 1, \cdots, n$.

Exemplo 5.1

Considere-se a função f dada pela tabela

x_i	0	1	2
f_i	1	1	2

a) Pelo método dos coeficientes indeterminados, pretende-se obter uma fórmula de quadratura do tipo

$$Q(f) = A_0 f(0) + A_1 f(1) + A_2 f(2)$$

que permita aproximar o integral $I(f) = \int_{-1}^{3} f(x)dx$.

Obtenha os parâmetros A_0, A_1 e A_2 de modo que $Q(f)$ tenha grau de precisão $r \geq 2$.

Solução: Utilizando o método dos coeficientes indeterminados, pretende-se que $Q(f)$ tenha grau de precisão $r \geq 2$, ou seja, que $I(a_0 + a_1 x + a_2 x^2) = Q(a_0 + a_1 x + a_2 x^2)$, $\forall a_0, a_1, a_2$. Para se ter esta igualdade é necessário e suficiente que

$$Q(1) = I(1) = \int_{-1}^{3} dx, \qquad Q(x) = I(x) = \int_{-1}^{3} x dx, \qquad Q(x^2) = I(x^2) = \int_{-1}^{3} x^2 dx$$

Resulta o sistema de equações

$$\begin{array}{r} A_0 + A_1 + A_2 = 4 \\ A_1 + 2A_2 = 4 \\ A_1 + 4A_2 = \frac{28}{3} \end{array} \left\{ \begin{array}{l} \text{Elimin. Gauss:} \\ m_{32} = 1 \\ (L_3 - m_{32} L_2) \rightarrow L_3 \end{array} \right. \Longrightarrow \quad \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ \frac{16}{3} \end{bmatrix}$$

Obtém-se: $A_0 = \dfrac{8}{3}$, $A_1 = -\dfrac{4}{3}$ e $A_2 = \dfrac{8}{3}$. Logo,

$$Q(f) = \frac{4}{3}\left[2f(0) - f(1) + 2f(2)\right]$$

b) Obtenha o valor aproximado do integral, no caso de f ser a função tabelada.

Solução: Utilizando os valores da tabela dada obtemos:

$$I(f) \approx Q(f) = \frac{4}{3}\left[2(1) - 1(1) + 2(2)\right] = 20/3$$

c) Seja f um polinómio de grau não superior a 3. Prove que o valor obtido pela regra considerada na alínea **a)** é o mesmo que se obteria pela regra de Simpson.

Solução: A fórmula de Simpson é exacta para polinomios de grau ≤ 3 e portanto, basta mostrarmos que o grau de precisão da fórmula obtida é $r \geq 3$. Como $Q(f)$ tem grau de precisão $r \geq 2$ é suficiente mostrarmos que $Q(x^3) = I(x^3)$. De facto, temos

$$Q(x^3) = \frac{4}{3}\left[2(0^3) - (1^3) + 2(2^3)\right] = 20$$

e

$$I(x^3) = \int_{-1}^{3} x^3 dx = \left.\frac{x^4}{4}\right|_{-1}^{3} = \frac{3^4 - (-1)^4}{4} = 20$$

donde se conclui que $Q(f)$ tem grau de precisão $r \geq 3$. Logo, se $f(x) = P_3(x)$, temos:

$$S(P_3) = I(P_3) = Q(P_3)$$

6 Métodos numéricos para problemas de valor inicial

6.1 Introdução

Começamos com alguns exemplos

Exemplo 6.1

A equação

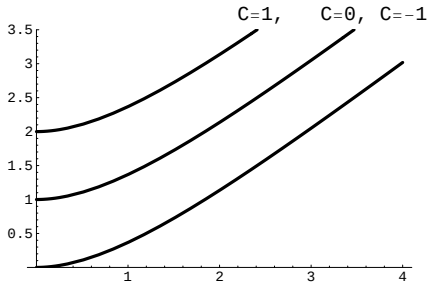
$$\frac{dy}{dx} = 1 - e^{-x} \tag{6.1}$$

é uma equação diferencial pois envolve a derivada $\frac{dy}{dx}$ (ou $y'(x)$) da função incógnita $y(x)$. Primitivando o lado direito, obtém-se

$$y(x) = x + e^{-x} + C, \quad C \in I\!R \tag{6.2}$$

Todas as funções da forma (6.2) são soluções de (6.1), visto que verificam a condição $y'(x) = 1 - e^{-x}$. Elas constituem uma família de soluções: a cada valor de C corresponde uma função que é solução de (6.1). A figura abaixo mostra os seguintes exemplos:

$$\begin{aligned} C = 0 \quad & \Rightarrow y(x) = x + e^{-x} \\ C = 1 \quad & \Rightarrow y(x) = x + e^{-x} + 1 \\ C = -1 \quad & \Rightarrow y(x) = x + e^{-x} - 1 \end{aligned}$$



Exemplo 6.2

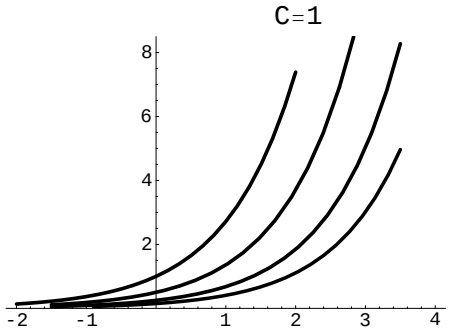
Na equação

$$y'(x) = y(x) \tag{6.3}$$

o lado direito agora também envolve a função incógnita $y(x)$. Pretende-se $y(x)$ tal que a sua derivada $y'(x)$ coincide com a própria $y(x)$. As soluções de (6.3) têm a forma

$$y(x) = C\,e^x, \qquad C \in I\!R \tag{6.4}$$

Com efeito, derivando (6.4) em ordem a x, vem $y'(x) = Ce^x = y(x)$. Graficamente tem-se uma família de curvas descritas por (6.4). A figura seguinte mostra algumas dessas curvas, em particular, o caso da função $y(x) = e^x$, que corresponde a escolher $C = 1$.



Problema de valor inicial

Ao problema da forma

$$\begin{cases} y'(x) = f(x, y(x)), & a \le x \le b \\ y(x_0) = y_0 \end{cases} \tag{6.5}$$

chamamos problema de valor inicial. Uma solução do problema acima será uma função diferenciável $y(x)$ que, substituída em $f(x, y)$, conduz à igualdade $y'(x) = f(x, y(x))$ e tal que $y(x_0) = y_0$. Ou seja, a curva $y = y(x)$ deve verificar a equação diferencial e a condição inicial.

Exemplo 6.3

Associando uma condição inicial à equação (6.3), obtém-se o par

$$\begin{cases} y'(x) = y(x) \\ y(x_0) = y_0 \end{cases} \tag{6.6}$$

Vimos que a solução geral de (6.3) é $y(x) = Ce^x$ (cf. (6.4)). Estamos interessados na curva $y(x)$ que, num determinado ponto x_0, toma o valor y_0. Ela corresponde a uma escolha bem determinada da constante C. Impondo a condição $y(x_0) = y_0$, vem $Ce^{x_0} = y_0$, donde $C = e^{-x_0}y_0$ e, finalmente, a solução do problema (6.6) é $y(x) = y_0\ e^{x-x_0}$. Em particular, a solução que em $x_0 = 1$ toma o valor 1 é $y(x) = e^x$, a qual corresponde à escolha $C = 1$.

Exemplo 6.4

Consideremos a equação

$$y'(x) = -K(y(x) - A), \tag{6.7}$$

onde $y(x)$ representa a temperatura, num instante x, dum objecto que está imerso num meio à temperatura A. A equação (6.7) descreve a evolução da temperatura desse objecto, sendo a taxa de variação da temperatura proporcional à diferença $y - A$.

Suponhamos que é conhecida a temperatura do objecto no instante $t = 0$, dada por $y(0) = y_0$. Tem-se então o problema de valor inicial

$$\begin{cases} y'(x) = -K(y(x) - A) \\ y(0) = y_0 \end{cases} \tag{6.8}$$

que tem por solução

$$y(x) = A + (y_0 - A)\ e^{-Kx} \tag{6.9}$$

A cada valor de y_0 está associada uma curva diferente. Se $y_0 < A$, então o objecto tende a aquecer; se $y_0 > A$, então o objecto tende a arrefecer. Em qualquer dos casos, a temperatura do objecto aproximar-se-á da do meio.

Observação 6.1

É possível provar que a solução dum problema da forma

$$y'(x) = u(x)\, y + v(x), \quad y(x_0) = y_0$$

é a seguinte:

$$y(x) = y_0\, e^{U(x)} + e^{U(x)} \int_{x_0}^{x} v(s) e^{-U(s)} ds, \quad \text{onde } U(x) = \int_{x_0}^{x} u(s) ds \tag{6.10}$$

Na equação do exemplo anterior, tem-se $y'(x) = -K\, y(x) + K\, A$, donde $u(x) = K$ e $v(x) = K\, A$. Aplicando a fórmula (6.10) chega-se a (6.9).

Teorema (Existência e unicidade de solução)

Seja $D = \{(x, y) : a \leq x \leq b, \quad -\infty < y < +\infty\}$ e seja $f(x, y)$ uma função contínua em D. Se existe uma constante positiva K tal que

$$\left|\frac{\partial f(x,y)}{\partial y}\right| \leq K, \qquad \forall (x, y) \in D, \tag{6.11}$$

então o problema de valor inicial (6.5) tem uma única solução $y(x)$, definida em $[a, b]$.

Exemplo 6.5

Seja o problema de valor inicial

$$\begin{cases} y'(x) = 1 + x\, sin(x\, y(x)), \quad 0 \leq x \leq 2 \\ y(0) = 0 \end{cases} \tag{6.12}$$

Definindo $D = \{(x, y) : 0 \leq x \leq 2, \quad -\infty < y < +\infty\}$, tem-se que $f(x, y) = 1 + x\, sin(xy)$ é uma função contínua em D. Como

$$\left|\frac{\partial f(x,y)}{\partial y}\right| = \left|x^2\, cos(xy)\right| \leq 4, \quad (x, y) \in D,$$

então, pelo teorema anterior, conclui-se que o problema tem uma única solução $y(x)$ definida em $[0, 2]$.

6.2 Métodos de Taylor

Em geral não é possível determinar uma expressão explícita para a solução $y(x)$ do problema de valor inicial (6.5).

Nos métodos numéricos que vamos estudar, divide-se o intervalo $[a, b]$ em subintervalos de comprimento h, através de uma "rede" ou "malha" de pontos

$$a = x_0 < x_1 < x_2 < ... < x_N = b. \tag{6.13}$$

Tem-se $h = (b - a)/N$ e $x_i = x_0 + ih, \quad i = 0, 1, ...N$. Obtém-se, não a expressão explícita de uma função que aproxima $y(x)$, mas aproximações para os valores $y(x_i), i =, 1, ..., N$. É costume designar por y_i uma aproximação de $y(x_i)$. Assim, vamos obter um conjunto de valores aproximados $y_1, y_2, ..., y_N$ para $y(x_1), y(x_2), ..., y(x_N)$, respectivamente.

O Método de Euler

Dado o valor inicial $y(x_0) = y_0$, começa-se por obter uma aproximação, y_1, para $y(x_1)$. No método de Euler, traça-se a recta $r(x)$ tangente ao gráfico de $y(x)$ no ponto $(x_0, \underbrace{y(x_0)}_{y_0})$ e toma-se para y_1 o seu valor em $x = x_1$. Como o declive da recta tangente é dado pelo valor da derivada naquele ponto, ou seja, $y'(x_0)$, vem

$$r(x) = y_0 + (x - x_0)y'(x_0) \tag{6.14}$$

Da equação dada em (6.5), $y'(x) = f(x, y(x))$, obtém-se $y'(x_0) = f(x_0, \underbrace{y(x_0)}_{y_0})$, vem

$$r(x) = y_0 + (x - x_0)f(x_0, y_0) \tag{6.15}$$

Fazendo $x = x_1$, então $r(x_1) = y_0 + (x_1 - x_0)f(x_0, y_0)$. Sendo h o espaçamento entre os pontos, obtém-se para aproximação de $y(x_1)$:

$$y_1 = y_0 + h\, f(x_0, y_0) \tag{6.16}$$

Exemplo 6.6

Seja o problema

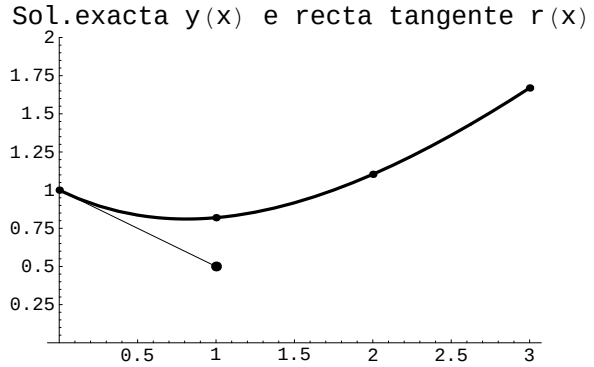
$$\begin{cases} y'(x) = \dfrac{x - y(x)}{2} \quad (= f(x, y(x))), \quad x \in [0, 3], \\ y(0) = 1 \end{cases} \tag{6.17}$$

Tem-se

$$f(x, y) = \frac{x - y}{2} \tag{6.18}$$

Além disso, $x_0 = 0, \quad y_0 = 1$. Se considerarmos uma malha de pontos, com espaçamento $h = 1$ no intervalo $[0, 2]$, então resulta: $x_1 = 1, x_2 = 2, x_3 = 3$.

Neste caso especial, a aplicação da fórmula (6.10), conduz à solução exacta: $y(x) = 3e^{-x/2} - 2 + x$. A figura seguinte mostra o gráfico de $y(x)$ e a recta tangente $r(x)$ definida por (6.14). Sendo $y'(x) = -(3/2)e^{-x/2} + 1$, pode calcular-se o declive da tangente: $y'(0) = -1/2$. Então a recta tangente em $(0, 1)$ é: $r(x) = 1 - x/2$. Fazendo $x = x_1 = 1$ obtém-se $y_1 = 1 - 1/2 = 0.5$.



É claro que, na prática, para obter a aproximação y_1 bastará aplicar directamente a fórmula (6.16), do seguinte modo.

Atendendo a (6.18), vem:

$$f(x_0, y_0) = \frac{x_0 - y_0}{2} = -1/2$$

Logo

$$y_1 = y_0 + h\, f(x_0, y_0) = 1 - \frac{h}{2}$$

Como escolhemos para espaçamento o valor $h = 1$, vem $y_1 = 1/2$ para aproximação de $y(x_1)$.

De maneira análoga a $y_1 = y_0 + h\, f(x_0, y_0)$, se podem definir, sucessivamente

y_2, y_3, ..., y_N:

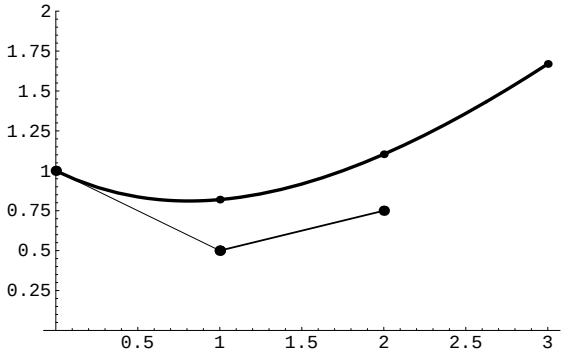
$$\begin{aligned} y_2 &= y_1 + h\, f(x_1, y_1) \\ y_3 &= y_2 + h\, f(x_2, y_2) \\ y_4 &= y_3 + h\, f(x_3, y_3) \\ y_N &= y_{N-1} + h\, f(x_{N-1}, y_{N-1}) \end{aligned}$$

Por exemplo para obter y_2, traça-se uma recta partindo de (x_1, y_1), com declive $f(x_1, y_1)$, e o seu valor em $x = x_2$ é y_2. Notemos que o declive da recta tangente em $(x_1, y(x_1))$, ao gráfico de $y(x)$, é agora $y'(x_1) = f(x_1, y(x_1))$. Como não dispomos do valor exacto $y(x_1)$ usamos a aproximação y_1 obtida anteriormente. Vem

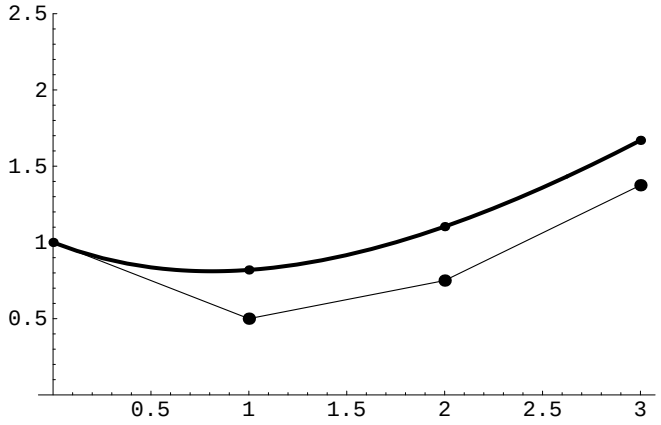
$$y'(x_1) = f(x_1, y(x_1)) \simeq f(x_1, y_1) \tag{6.19}$$

pelo que $f(x_1, y_1)$ representa, em geral, apenas uma aproximação para o declive da tangente. Raciocínio semelhante pode ser aplicado na obtenção dos valores $y_2, y_3, ..., y_N$.

Exemplo 6.7 *A figura seguinte ilustra o que se disse acima em relação ao modo como se obtém y_2 e mostra o segmento de recta partindo de (x_1, y_1), com declive $f(x_1, y_1)$:*



Na figura abaixo, mostram-se os valores exactos $y(1), y(2), y(3)$ (marcados sobre o gráfico de $y(x)$) e os respectivos valores aproximados obtidos pelo método de Euler, com $h = 1$:



Dedução teórica da fórmula geral do método de Euler

A partir de y_1 podem-se obter, sucessivamente, y_2, y_3,.... Supondo que já foram obtidas as aproximações $y_1, \ldots, y_n$, vejamos como obter y_{n+1}. Geometricamente, traçamos uma recta a partir do ponto (x_n, y_n), com declive $f(x_n, y_n)$, e toma-se para y_{n+1} o valor da recta em $x = x_{n+1}$.

Consideremos o desenvolvimento de Taylor para $y(x)$, em torno de x_n:

$$y(x) = y(x_n) + (x - x_n)\ y'(x_n) + \frac{(x - x_n)^2}{2} y''(\xi_n) \tag{6.20}$$

onde $\xi_n \in [x_n, x_{n+1}]$. Fazendo $x = x_{n+1}$, obtém-se

$$y(x_{n+1}) = y(x_n) + (x_{n+1} - x_n)\ y'(x_n) + \frac{(x_{n+1} - x_n)^2}{2} y''(\xi_n) \tag{6.21}$$

ou seja, com $h = x_{n+1} - x_n$,

$$y(x_{n+1}) = y(x_n) + h\ y'(x_n) + \frac{h^2}{2} y''(\xi_n) \tag{6.22}$$

Desprezando o termo de resto $T = \frac{h^2}{2} y''(\xi_n)$, obtém -se uma aproximação para $y(x_{n+1})$:

$$y(x_{n+1}) \simeq y(x_n) + h\ y'(x_n) \tag{6.23}$$

Em seguida, como em princípio só dispomos do valor aproximado y_n, usamo-lo no lugar de $y(x_n)$. Por outro lado, reescrevendo y' em função de f, vem:

$$y'(x_n) = f(x_n, \underbrace{y(x_n)}_{y_n}) \simeq f(x_n, y_n)$$

Desprezando ainda o termo de resto, resulta a aproximação:

$$y(x_{n+1}) \simeq y_n + h\ f(x_n, y_n) \tag{6.24}$$

e somos conduzidos à fórmula geral:

$$y_{n+1} = y_n + h\, f(x_n, y_n) \tag{6.25}$$

Fórmula do erro do método de Euler

Teorema 6.1 *Considere-se o problema de valor inicial (6.5). Sejam $y_1, y_2, ..., y_N$ valores aproximados para $y(x_1), y(x_2), ..., y(x_N)$, respectivamente, obtidos pelo método de Euler.*

Suponhamos que $f(x, y)$ é uma função contínua em $D = \{(x, y) : a \leq x \leq b, \quad -\infty < y < +\infty\}$, e que existe uma constante positiva K tal que

$$\left| \frac{\partial f(x, y)}{\partial y} \right| \leq K, \qquad \forall (x, y) \in D \tag{6.26}$$

Então, como já se disse, o problema de valor inicial (6.5) tem uma única solução $y(x)$, definida em $[a, b]$. Seja, além disso, $y \in C^2[a, b]$ tal que $|y''(x)| \leq M, \quad \forall x \in [a, b]$. Então é válida a seguinte fórmula para o erro global $|y(x_n) - y_n|$:

$$|e_n| = |y(x_n) - y_n| \leq \frac{h\, M}{2K} \left(e^{K(x_n - x_0)} - 1 \right), \quad n = 0, 1, \ldots, N - 1. \tag{6.27}$$

A fórmula (6.27) mostra que o erro num determinado ponto $x = x_0 + nh$ satisfaz:

$$|e_n| = |y(x_n) - y_n| \leq C^* \, h. \tag{6.28}$$

Quando se faz h tender para zero, o erro nesse ponto tende para zero, pelo que nas condições do teorema enunciado, o método de Euler converge. Atendendo ao expoente unitário de h, diz-se que o método de Euler é um método de ordem 1.

Isso tem como consequência prática o seguinte: ao diminuir-se o valor do passo h para $h/2$, espera-se que os erros na solução aproximada venham aproximadamente divididos por 2.

Definição 6.1 *Um método diz-se convergente com ordem p se*

$$e(h) = max_{n=0,\ldots,N-1}|e_n| \leq C^* \, h^p \tag{6.29}$$

O Método de Taylor de ordem 2

O método de Euler foi deduzido considerando o desenvolvimento de Taylor de primeira ordem da solução exacta $y(x)$ e desprezando o respectivo termo de resto de ordem 2.

Para obter métodos de ordens mais elevadas (em que os erros tendem para zero mais rapidamente), consideram-se mais termos no desenvolvimento de Taylor. No método de Taylor de ordem 2 vai-se até ao resto de ordem 3. Vejamos como, em geral, se obtém uma aproximação para $y(x_{n+1})$ a partir da aproximação para $y(x_n)$. Por outras palavras, como obter y_{n+1}, supondo que já foram obtidas $y_1, y_2, ..., y_n$.

Supondo que $y(x)$ tem derivadas contínuas até à terceira ordem, consideremos o desenvolvimento de Taylor de $y(x)$ em torno de x_n:

$$y(x) = y(x_n) + (x - x_n)\; y'(x_n) + \frac{(x - x_n)^2}{2} y''(x_n) + \frac{(x - x_n)^3}{3!} y'''(\xi_n) \tag{6.30}$$

onde $\xi_n \in [x_n, x_{n+1}]$. Fazendo $x = x_{n+1}$, obtém-se

$$y(x_{n+1}) = y(x_n) + (x_{n+1} - x_n)\; y'(x_n) + \frac{(x_{n+1} - x_n)^2}{2} y''(x_n) + \frac{(x_{n+1} - x_n)}{3!} y'''(\xi_n) \tag{6.31}$$

ou seja, com $h = x_{n+1} - x_n$,

$$y(x_{n+1}) = y(x_n) + h\; y'(x_n) + \frac{h^2}{2} y''(x_n) + \frac{h^3}{3!} y'''(\xi_n) \tag{6.32}$$

Desprezando o termo de resto, $T_{n+1} = \dfrac{h^3}{3!} y'''(\xi_n)$, resulta a aproximação:

$$y(x_{n+1}) \simeq y(x_n) + h\ y'(x_n) + \frac{h^2}{2} y''(x_n) \tag{6.33}$$

A seguir, expressemos as derivadas em termos da função f. Tem-se $y'(x) = f(x, y(x))$, donde, tal como no método de Euler, resulta imediatamente $y'(x_n) = f(x, y(x_n))$. Como em geral $y(x_n)$ é desconhecido (apenas dispomos de $y(x_0) = y_0$), usamos, no seu lugar, o valor aproximado y_n.

Para obter a segunda derivada, usamos a regra da derivação da função composta

$$y''(x) = \frac{d}{dx}\left(f(x, y(x))\right) = \frac{\partial f}{\partial x}(x, y(x))\frac{dx}{dx} + \frac{\partial f}{\partial y}(x, y(x))\frac{dy}{dx}$$

Donde

$$\begin{aligned} y''(x_n) &= \frac{\partial f}{\partial x}(x_n, y(x_n)) + \frac{\partial f}{\partial y}(x_n, y(x_n))y'(x_n) && (6.34)\\ &\simeq \frac{\partial f}{\partial x}(x_n, y_n) + \frac{\partial f}{\partial y}(x_n, y_n)y'(x_n) && (6.35) \end{aligned}$$

Obtém-se a expressão geral do método de Taylor de ordem 2

$$y_{n+1} = y_n + h\ f(x_n, y_n) + \frac{h^2}{2}\left[\frac{\partial f}{\partial x}(x_n, y_n) + \frac{\partial f}{\partial y}(x_n, y_n)f(x_n, y_n)\right], \quad n = 0, 1, 2, ..., N-1 \tag{6.36}$$

Poderemos representar esta fórmula de modo mais compacto,

$$y_{n+1} = y_n + h\ y'_n + \frac{h^2}{2} y''_n, \tag{6.37}$$

onde

$$\begin{aligned} y'_n &= f(x_n, y_n),\\ y''_n &= \frac{\partial f}{\partial x}(x_n, y_n) + \frac{\partial f}{\partial y}(x_n, y_n)f(x_n, y_n) \end{aligned}$$

Exemplo 6.8

Obtenha uma aproximação para $y(1)$, com $h = 1$ e $h = 1/2$.

$$\begin{cases} y'(x) = \dfrac{x - y(x)}{2} \\ y(0) = 1 \end{cases}$$

Processo 1. Com o fim de usar a fórmula (6.37), vamos começar por obter as expressões aproximadas para as derivadas.

Tem-se

$$y'(x_n) = \frac{x_n - y(x_n)}{2} \simeq \frac{x_n - y_n}{2} \tag{6.38}$$

$$y''(x) = \left(\frac{x - y(x)}{2}\right)' = \frac{1 - y'(x)}{2} \tag{6.39}$$

$$y''(x_n) = \frac{1 - y'(x_n)}{2} = \frac{1}{2} - \frac{1}{2}\left(\frac{x_n - y(x_n)}{2}\right) \simeq \frac{1}{2} - \frac{1}{2}\left(\frac{1 - y'_n}{2}\right) \tag{6.40}$$

Obtém-se a expressão do método de Taylor de ordem 2

$$y_{n+1} = y_n + h\ \frac{x_n - y_n}{2} + \frac{h^2}{2}\ \left(\frac{1}{2} - \frac{x_n - y_n}{4}\right) \tag{6.41}$$

Processo 2. Note-se que esta fórmula se poderia obter por aplicação de (6.36). Basta calcular as derivadas parciais de $f(x, y) = (x - y)/2$ e substitui-las na expressão do método dada por (6.36):

$$\frac{\partial f}{\partial x}(x, y) = \frac{1}{2}, \qquad \frac{\partial f}{\partial y}(x, y) = -\frac{1}{2}$$

Viria:

$$y_{n+1} = y_n + h\ \frac{x_n - y_n}{2} + \frac{h^2}{2}\ \left(\frac{1}{2} - \frac{1}{2}f(x_n, y_n)\right), \tag{6.42}$$

o que conduz a (6.41).

i) Com $h = 1$, tem-se que o ponto $x = 1$ corresponde a x_1, pelo que uma aproximação para $y(1)$ será y_1. Vem

$$y_1 = y_0 + h\frac{x_0 - y_0}{2} + \frac{h^2}{2}\ \left(\frac{1}{2} - \frac{x_0 - y_0}{4}\right) = y_0 + \frac{x_0 - y_0}{2} + \left(\frac{1}{2} - \frac{x_0 - y_0}{4}\right) = 0.875$$

ii) Com $h = 1/2$, tem-se $x_0 = 0, x_1 = 1/2, x_2 = 1$. O ponto 1 corresponde agora a x_2, pelo que se pretende y_2. Primeiramente é preciso calcular y_1. Note-se que $y_1 \simeq y(x_1)$, ou seja, atendendo a que $x_1 = 1/2$, representa uma aproximação para $y(1/2)$. A fórmula geral do método agora é:

$$y_{n+1} = y_n + \frac{x_n - y_n}{4} + \frac{1}{8}\left(\frac{1}{2} - \frac{x_n - y_n}{4}\right)$$

$$y_1 = y_0 + \frac{x_0 - y_0}{4} + \frac{1}{8}\left(\frac{1}{2} - \frac{x_0 - y_0}{4}\right) = 0.84375$$

$$y_2 = y_1 + \frac{x_1 - y_1}{4} + \frac{1}{8}\left(\frac{1}{2} - \frac{x_1 - y_1}{4}\right) = 0.831055$$

6.3 Métodos de Runge Kutta

Apresentamos agora os métodos de Runge Kutta, que conduzem à mesma precisão que os métodos de Taylor, mas sem o inconveniente de terem de se calcular as derivadas de $y(x)$. Apresentamos dois exemplos de métodos de Runge Kutta, da forma:

$$y_{n+1} = y_n + h(c_1\, V_1 + c_2\, V_2)$$

com
$V_1 = f(x_n, y_n)$
$V_2 = f(x_n + a_2\, h, y_n + b_2\, h\, V_1)$

onde c_1, c_2, a_2, b_2 são determinadas de modo a obter-se a precisão do método de Taylor de ordem 2.

Casos particulares

	c_2	c_1	a_2	b_2
Mét. do ponto médio	1	0	1/2	1/2
Mét. de Heun	1/2	1/2	1	1

Somos conduzidos aos dois seguintes métodos de Runge-Kutta de ordem 2:

Método do ponto médio (ou Euler modificado):

$$y_{n+1} = y_n + hf(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)) \tag{6.43}$$

Método de Heun:

$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + hf(x_n, y_n))] \tag{6.44}$$

Concluímos com um método por vezes designado "método de Runge Kutta clássico", o qual é um **método de Runge-Kutta de ordem 4**:

$$y_{n+1} = y_n + \frac{h}{6}(V_1 + 2V_2 + 2V_3 + V_4)$$

$$V_1 = f(x_n, y_n) \quad V_2 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}V_1)$$

$$V_3 = f(x_n + \frac{h}{2}, y_n + \frac{h}{2}V_2) \quad V_4 = f(x_n + h, y_n + hV_3)$$

Exemplo 6.9

Seja o problema de valor inicial

$$\begin{cases} y'(x) = \dfrac{x^2}{y(x)} \\ y(1) = -1 \end{cases} \tag{6.45}$$

Obtenha uma aproximação para $y(1.2)$, pelo método do ponto médio, com $h = 0.2$.

Tem-se $f(x, y) = x^2/y$, $x_0 = 1$ e $y_0 = -1$.

A fórmula (6.43), com $n = 0$, dá

$$y_1 = y_0 + hf(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}f(x_0, y_0))$$

Sendo $f(x_0, y_0) = x_0^2/y_0 = -1$, tem-se

$$y_1 = -1 + 0.2\, f(1 + \frac{0.2}{2}, -1 + \frac{0.2}{2}(-1))$$

Vemos que precisamos do valor $f(1.1, -1.1) = \dfrac{(1.1)^2}{-1.1}$. Substituindo acima, obtém-se $y_1 = -1 + 0.2(-1.1) = -1.22$

6.4 Sistemas de equações diferenciais e equações diferenciais de ordem n

Um sistema de 2 equações diferenciais de primeira ordem tem a forma genérica:

$$\begin{cases} y_1'(x) = f_1(x, y_1(x), y_2(x)) \\ y_2'(x) = f_2(x, y_1(x), y_2(x)) \\ \text{com as condições iniciais} \\ y_1(x_0) = y_{1,0} \\ y_2(x_0) = y_{2,0} \end{cases}$$

Exemplo 6.10 *Um exemplo concreto é o sistema seguinte*

$$\begin{cases} y_1'(x) = y_1 - 2\,y_2 + 4\cos(x) - 2\sin(x), \quad y_1(0) = 1 \\ y_2'(x) = 3y_1 - 4y_2 + 5\cos(x) - 5\sin(x), \quad y_2(0) = 2 \end{cases}$$

o qual tem por solução $y_1(x) = \cos(x) + \sin(x), \quad y_2(x) = \cos(x)$.

Por outro lado, um problema de valor inicial de ordem 2 tem a forma genérica

$$\begin{cases} y''(x) = g(x, y(x), y'(x)) \\ y(x_0) = y_0 \\ y'(x_0) = y_0' \end{cases}$$

e pode ser reduzido a um sistema de equações diferenciais de ordem 1, definindo novas variáveis, do seguinte modo:

$$\begin{cases} y'(x) = z(x) \quad (= f1(x, y(x), z(x)) \\ z'(x) = g(x, y(x), z(x)) \quad (= f2(x, y(x), z(x)) \\ y(x_0) = y_0 \\ z(x_0) = y_0' \end{cases}$$

Consideremos um exemplo.

Exemplo 6.11

$$\begin{cases} y''(x) = y(x) + e^x, \qquad x \in [0,\, 0.3] \\ y(0) = 1 \\ y'(0) = 0 \end{cases} \Leftrightarrow \begin{cases} y'(x) = z(x) \equiv f_1(x, y(x), z(x)) \\ z'(x) = y(x) + e^x \equiv f_2(x, y(x), z(x)), \qquad x \in [0,\, 0.3] \\ y(0) = 1, z(0) = 0 \end{cases}$$

No caso dum problema de valor inicial de ordem 3, procede-se de modo análogo, o que é ilustrado a seguir.

Exemplo 6.12

$$\begin{cases} y'''(x) - 2xy''(x) + 4y'(x) - x^2y(x) = 1 \\ y(0) = 1, \quad y'(0) = 2, \quad y''(0) = 3 \end{cases}$$

que pode ser convertido no seguinte sistema

$$\begin{cases} y'(x) = z(x) \\ z'(x) = w(x) \qquad (= y''(x)) \\ w'(x) = 2xw(x) + 4z(x) - x^2y(x) + 1 \\ y(0) = 1, z(0) = 2, w(0) = 3 \end{cases}$$

Em geral, um problema de valor inicial de ordem n tem a forma:

$$\begin{cases} y^{(n)} = f(x, y, y', y'', ..., y^{(n-1)}) \\ y(x_0) = y_0 \\ y'(x_0) = y_0' \\ \vdots \\ y^{(n-1)}(x_0) = y_0^{(n-1)} \end{cases}$$

onde $y_0, y_0', \cdots, y_0^{(n-1)}$ são valores dados. Considerando

$u_1(x) = y(x), u_2(x) = y'(x), \cdots, u_n(x) = y^{(n-1)}(x)$, obtém-se um sistema de primeira ordem

$$\begin{cases} u_1'(x) = u_2(x) \\ u_2'(x) = u_3(x) \\ u_3'(x) = u_4(x) \\ \vdots \\ u_{(n)}'(x) = f(x, u_1(x), u_2(x), u_3(x), \cdots, u_n(x)) \end{cases}$$

Definindo os vectores

$$\mathbf{u}(x) = \begin{bmatrix} u_1(x) \\ u_2(x) \\ u_3(x) \\ \vdots \\ u_n(x) \end{bmatrix}, \quad \mathbf{F}(x, \mathbf{u}(x)) = \begin{bmatrix} u_2(x) \\ u_3(x) \\ \vdots \\ f(x, u_1(x), u_2(x), u_3(x), \cdots, u_n(x)) \end{bmatrix}$$

pode-se representar o sistema de equações de primeira ordem na forma vectorial,

$$\mathbf{u}'(x) = \mathbf{F}(x, \mathbf{u}(x)), \quad \mathbf{u}(x_0) = \mathbf{u}_0, \tag{6.46}$$

onde o vector das condições iniciais é dado por:

$\mathbf{u}_0 = [y_0, y_0', \cdots, y_0^{(n-1)}]^T$.

Os métodos numéricos estudados no capítulo 6, para um problema de valor inicial de primeira ordem (uma só função incógnita) podem ser aplicados a um sistema de várias equações usando a sua representação vectorial (6.46).

7 Exercícios propostos

1. Considere a equação
$$f(x) = x^2 - \cos^2(x) = 0. \tag{7.47}$$
   (a) Mostre que a equação (1) tem (apenas) duas raízes.
   (b) Para resolver numericamente a equação (1) vai-se considerar um método do ponto fixo associado a uma função iteradora da forma $g(x) = x + A(x)(\cos(x) - x)$, onde $A(x)$ é uma função que nunca se anula.
      i. Mostre que as raízes da equação (1) e os pontos fixos de g coincidem para $x > 0$.
      ii. Fazendo $A(x) = 1/2$, prove que o método do ponto fixo associado a g converge para a raiz z pertencente ao intervalo $[0, 1]$, qualquer que seja a aproximação inicial x_0 escolhida nesse intervalo.
      iii. Utilizando o método obtido em a) e $x_0 = 1$, calcule uma aproximação x_{k+1} da raiz z, parando a iteração quando $|x_{k+1} - x_k| < 10^{-3}$. Indique ainda uma estimativa para o erro $|z - x_{k+1}|$.
      iv. Determine a ordem de convergência p do método e uma aproximação da constante K_∞, tal que seja verificada a igualdade assimptótica
$$|z - x_{m+1}| \approx K_\infty |z - x_m|^p, \qquad m \text{ suf. grande}$$

2. Considere a equação $F(x) = 0$, onde $F(x) = (x - \alpha)^m h(x)$, ($m > 1$ inteiro), com $h(\alpha) \neq 0$ e tal que h é uma função de classe C^2 num intervalo aberto contendo α. Prove que o método iterativo $x_{m+1} = g(x_m), m = 0, 1, \ldots$, em que
$$g(x) = x - \frac{F(x)}{F'(x)}$$
converge para α se a aproximação inicial x_0 estiver suficientemente próxima de α. Determine a ordem de convergência do método e o factor assimptótico da convergência.

3. Considere a equação
$$\sin(x) + 1 - ax = 0$$
onde a é um número real conhecido.
   (a) Diga, justificando, para que valores de a esta equação tem uma única raiz no intervalo $I = [0, \frac{\pi}{2}]$.
   (b) Para os valores de a considerados na alínea anterior, mostre que o método do ponto fixo, com a função iteradora $g(x) = \dfrac{1 + \sin(x)}{a}$, converge para a z, qualquer que seja $x_0 \in I$.

(c) No caso de $a = 2$, diga qual o número mínimo de iterações do método do ponto fixo que deverá efectuar para garantir que o erro absoluto da aproximação obtida seja inferior a 10^{-3} se x_0 é qualquer número em I.

(d) No caso de $a = 0$, mostre que a equação tem uma única raiz w no intervalo $[-\pi, 0]$. Mostre que, se x_0 estiver suficientemente próximo da raiz, então o método de Newton converge para z, e determine a ordem de convergência.

4. Considere a família de sucessões da forma

$$x_{m+1} = \frac{\lambda\, x_m + 1 - \sin(x_m)}{1+\lambda}, \quad m = 0, 1, \ldots, \tag{7.48}$$

com λ parâmetro real.

(a) Faça $\lambda = 1$ e, usando o teorema do ponto fixo, mostre que a sucessão (1) converge para um certo valor real z, qualquer que seja x_0 pertencente ao intervalo $[0, 1]$. Qual a ordem de convergência deste método iterativo ?

(b) Conclua, utilizando a questão anterior, que z é a única raiz da equação $1 - x - \sin x = 0$ no intervalo $[0, 1]$. Utilizando o método obtido em a) e $x_0 = 1$, determine, sem efectuar iterações, qual o valor de k de modo que se tenha $|z - x_k| < 10^{-5}$.

(c) Como deveria ser λ por forma a sucessão (1) convergir para z o mais rápido possível? Qual a ordem de convergência nesse caso? **[1.0]**

5. Prove que, com $x_0 = 1$, o método de Newton aplicado à equação $1 - x - \sin x = 0$ converge para a única raiz z da equação .

6. Considere a família de funções da forma

$$g_\alpha(x) = \frac{1}{\alpha}\left[(\alpha - 1)x + \frac{78.8}{x}\right], \tag{7.49}$$

onde α é um parâmetro real.

(a) Mostre que os pontos fixos de g_α são as raizes da equação $x^2 - 78.8 = 0$, independentemente do valor do parâmetro α.

(b) Com o objectivo de aproximar a **raiz positiva** z dessa equação, considere a iteração do ponto fixo, $x_{m+1} = g_\alpha(x_m)$, associada a (7.49). A tabela seguinte mostra algumas iteradas das sucessões correspondentes aos valores $\alpha = 3/2$ e $\alpha = 1/2$, com iterada inicial $x_0 = 9$.

	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$x_{m+1} = g_{3/2}(x_m)$	8.837037	8.890356	?	8.78425	8.876441	8.877102
$x_{m+1} = g_{1/2}(x_m)$	8.51111	10.00586	5.74491	21.68807	-14.42104	3.49320

i. No caso de $\alpha = 3/2$, preencha o espaço em branco (obtenha x_3). Diga o que indicam os resultados, no que respeita à convergência ou divergência das sucessões para a raiz z acima referida. **Confirme teoricamente**, em cada caso.

ii. No caso de $\alpha = 3/2$, ob obtenha um majorante para o erro absoluto da iterada x_3.

iii. Como deveria escolher α de modo a obter convergência quadrática, supondo x_0 suf. próximo de z ?

7. Considere a seguinte tabela de valores de uma função $f(x)$

x_i	-2	-1	0	1	2
$f(x_i)$	1	0	2	$\frac{1}{2}$	$-\frac{1}{2}$

(a) Utilizando a fórmula de Newton com diferenças divididas, determine o polinómio de grau ≤ 2, $p_2(x)$, que interpola $f(x)$ nos pontos $x_0 = -2$, $x_2 = 0$ e $x_4 = 2$.

(b) Suponha que pretendemos aproximar o valor $I(f) = \int_{-2}^{2} f(x)dx$ por $\int_{-2}^{2} p_2(x)dx$. Sabendo que as derivadas de f verificam $|f^{(j)}(x)| \leq j/2,\ j = 1, 2, 3, 4$ no intervalo $[-2, 2]$, determine um majorante para o erro de integração. Justifique.

(c) Determine uma aproximação para $I(f) = \int_{-2}^{2} f(x)dx$ usando a regra de Simpson composta e todos os pontos da tabela.

8. Considere-se a função f dada pela tabela

x_i	0	1	2
f_i	1	1	2

a) Obtenha um valor aproximado de $f(1.5)$ utilizando:

i) um spline de grau 1.

ii) um polinómio interpolador de grau 2.

ii) Sabendo que f é um polinómio do tipo

$$f(x) = a\,x^3 + b\,x^2 + c\,x + d,$$

determine uma expressão do erro da última aproximação obtida, em função dos coeficientes de f.

b. Obtenha a função g, do tipo $g(x) = \beta\,x(x-1) + \alpha$, que melhor se ajusta a f (dada na tabela acima), no sentido dos mínimos quadrados.

9. Considere o problema de valor inicial

$$\begin{aligned} &y'(x) = 1 - x + 4y(x), \quad 0 \leq x \leq 1, \\ &y(0) = 1, \end{aligned}$$

com solução exacta $y(x) = x/4 - 3/16 + (19/16)e^{4x}$.

   (a) Obtenha um valor aproximado y_2 para $y(0.2)$ usando o método de Euler com passo $h = 0.1$.

   (b) Recorrendo a um resultado teórico, deduza um majorante para $|y(0.2) - y_2|$. Compare com o valor do erro de facto cometido.

   (c) Utilize o método de Taylor de ordem 2, com $h = 0.1$, para obter uma aproximação de $y(0.2)$. Compare com o resultado obtido em a).

10. Utilize o método de Runge-Kutta de ordem 2 (método do ponto médio) para obter uma aproximação da solução do problema de valor inicial

$$\begin{cases} y'(x) = x + y(x), \quad 0 \leq x \leq 1, \\ y(0) = 0 \end{cases}$$

no ponto $x = 0.1$ com espaçamentos $h = 0.1, 0.05, 0.025$. Sabendo que a solução exacta deste problema é dada por $y(x) = \exp(x) - 1 - x$, compare os resultados obtidos com o valor exacto de $y(0.1)$. Comente.

11. Dado o problema de valor inicial

$$\begin{cases} y'(x) = 1 - \dfrac{y(x)}{x}, \quad 2 \leq x \leq 3, \\ y(2) = 2 \end{cases}$$

determine um valor aproximado de $y(2.1)$ pelo método de Euler com $h = 0.1, 0.05, 0.025$.

12. Dado o problema de valor inicial

$$\begin{cases} y'(x) = 1 - x + 4y(x) \quad 0 \leq x \leq 1, \\ y(0) = 1 \end{cases}$$

obtenha uma aproximação de $y(0.2)$ usando o método de Runge-Kutta de ordem 4, com $h = 0.2$.

Bibliografia

Atkinson, K., *An Introduction to Numerical Analysis*, Wiley & Sons, NY,1978.

Burden, R. & Faires, J., *Numerical Analysis*, 5th ed., PWS Publishing, Boston, 1993.

Chapra, S. & Canale, R., *Numerical Methods for Engineers*, Mc Graw Hill, 2002

Cheney, W. & Kincaid, D. *Numerical Mathematics and Computing*, 2ed ed., Brooks/Cole, Monterey, CA, 1994.

Hildebrand, F.B., *Introduction to Numerical Analysis*, Dover Publications, NY, 1974.

Henrici, P., *Elements of Numerical Analysis*, Wiley & Sons, 1964.

Isaacson, E. & Keller, H. B., *Analysis of Numerical Methods*, Wiley & Sons, 1966.

Johnson, L. W. & Riess, R. D., *Numerical Analysis*, Addison-Wesley, 1977.

Stroud, A. H., *Numerical Quadrature and Solution of Ordinary Differential Equations*, Applied Mathematical Sciences, vol. 10, Springer-Verlag, NY, 1974.

Valença, M. R., *Métodos Numéricos*, Instituto Nacional de Investigação Científica, 1990.