Faculdade de Engenharia da Universidade do Porto

Licenciatura em Engenharia Electrotécnica e de Computadores

# Apontamentos de Análise Numérica

Aníbal Castilho Coimbra de Matos

Setembro de 2005

# Nota introdutória

Estes apontamentos destinam-se a apoiar as aulas da disciplina de Análise Numérica da Licenciatura em Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia de Universidade do Porto.

A sua organização reflecte a forma como tenho vindo a leccionar as aulas teóricas desta disciplina desde o ano lectivo 2001/2002. Estes apontamentos não pretender ser um texto de referência, mas tão só permitir aos alunos um melhor acompanhamento da matéria leccionada. Para um estudo mais aprofundado e sistemático dos assuntos abordados, os alunos são remetidos para as diferentes obras referidas na bibliografia.

As discussões mantidas com o Prof. José Fernando Oliveira, principalmente no ano lectivo 2001/2002, e com o Prof. Carlos Mendonça e Moura, desde então, contribuíram de forma decisiva para a organização destes apontamentos. Por todos os ensinamentos, mas também pela forma como correu e ainda decorre o trabalho realizado em conjunto na leccionação de Análise Numérica, não poderia deixar de manifestar aqui o meu profundo agradecimento a estes meus Professores.

Aníbal Matos, Set/2005

# Conteúdo

# Capítulo 1

# Fundamentos

## 1.1 Introdução

Sempre que se pretende tratar algum problema cuja solução toma a forma do cálculo de um valor numérico é habitual ter de considerar não só conceitos de carácter mais abstracto (que fornecem um modelo consistente para a análise do problema) mas também questões de natureza mais prática relacionadas com os cálculos a efectuar ou com os números necessários à realização de tais cálculos.

**Exemplo 1.1.1.** *Suponha-se que se pretende determinar o volume $V$ de um paralelipípedo a partir dos comprimentos de três arestas $a$, $b$ e $c$, perpendiculares entre si. Neste caso, o modelo abstracto consiste na expressão $V = abc$, que permite calcular o volume a partir dos comprimentos das três arestas. Para aplicar esta expressão é então necessário começar por medir cada uma das arestas. Ora, à medição de cada uma das arestas está associado um erro (erro de medida). Ou seja, o processo de medição fornecerá valores aproximados dos comprimentos das arestas, sendo eventualmente possível obter alguma caracterização dos erros de medida. Ao efectuar, em seguida, o produto das medidas dos três comprimentos ir-se-á obter um valor que apenas poderá ser considerado uma aproximação do volume do paralelipípedo. Obviamente que este valor aproximado terá associado um erro que dependerá dos erros cometidos nos processos de medida.*

A situação descrita neste exemplo de não se conseguir obter um valor numérico exacto para muitos problemas é a mais comum. Esta impossibilidade pode ter origens diversas, de que são exemplos erros associados a processos de medida, modelos abstractos aproximados, ou cálculos efectuados de forma aproximada. Contudo esta situação não é necessariamente má, pois na grande maioria (ou até talvez na totalidade) dos problemas bastará obter um valor numérico suficientemente próximo do valor exacto.

De uma forma simples, pode dizer-se que a Análise Numérica abrange o estudo de métodos

e técnicas que permitam obter soluções aproximadas de problemas numéricos de uma forma eficiente. É por natureza uma disciplina que se situa na fronteira entre a Matemática e a Ciência de Computadores.

Neste capítulo apresentam-se os conceitos fundamentais necessários à compreensão e utilização dos métodos numéricos que irão ser estudados nos capítulos subsequentes.

## 1.2 Valores exactos e aproximados: erros

Consideremos um problema cuja solução é um número real. Este número é designado por **valor exacto** do problema e, no que se segue, será representado por $x$.

Designa-se por **valor aproximado** ou **aproximação**, e representa-se por $x^*$, qualquer valor que se pretende utilizar como solução do problema. Associado a um dado valor aproximado $x^*$ define-se o **erro de aproximação** como a diferença entre o valor exacto e o valor aproximado, isto é,

$$\Delta x^* = x - x^*.$$

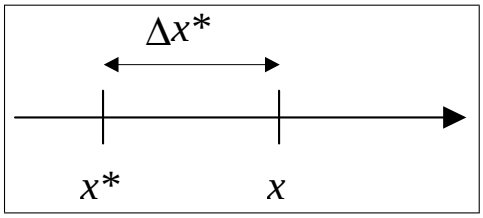



Figura 1.1: Valor exacto e aproximação.

No caso de $x^* < x$, a aproximação diz-se ser **por defeito**, verificando-se então que $\Delta x^* > 0$. No caso de $x^* > x$, a aproximação diz-se ser **por excesso**, tendo-se então que $\Delta x^* < 0$.

**Exemplo 1.2.1.** *É sabido que* $\pi \simeq 3.14159265359$. *Então,*

$$3 \qquad 3.1 \qquad 3.14 \qquad 3.141 \qquad \ldots$$

*são aproximações de* $\pi$ *por defeito e*

$$4 \qquad 3.2 \qquad 3.15 \qquad 3.142 \qquad \ldots$$

*são aproximações de* $\pi$ *por excesso.*

O valor absoluto do erro de aproximação, $|\Delta x^*| = |x - x^*|$, é designado por **erro absoluto**.

Note-se que de um modo geral, não é conhecido o erro $\Delta x^*$ associado a uma dada aproximação $x^*$. De facto, se ambos fossem conhecidos, o valor exacto $x$ poder-se-ia calcular por intermédio da expressão $x = x^* + \Delta x^*$, e então não se utilizaria tal aproximação!

Assim, a situação mais comum é aquela em que se conhece um determinado valor aproximado $x^*$ e um intervalo para o erro de aproximação $\Delta x^*$. Este intervalo é muitas vezes caracterizado a partir de majorantes do erro absoluto. A expressão **erro máximo absoluto** é utilizada para designar um majorante do erro absoluto. É usual indicar o erro máximo absoluto por $\varepsilon$. Então, se $x^*$ for um valor aproximado de $x$ com um erro máximo absoluto $\varepsilon$, verifica-se que $x \in [x^* - \varepsilon, x^* + \varepsilon]$. Neste caso é habitual usar-se a notação $x = x^* \pm \varepsilon$.

**Exemplo 1.2.2.** *Ao escrever-se* $x = 1.23 \pm 0.02$*, pretende dizer-se que* $1.23$ *é uma aproximação de* $x$ *com um erro máximo absoluto de* $0.02$*, ou seja, isto significa que* $x$ *estará no intervalo* $[1.21, 1.25]$.

Outra forma de caracterizar uma aproximação $x^*$ é através do **erro relativo**, que se define por

$$\frac{|\Delta x^*|}{|x|},$$

para valores de $x$ diferentes de zero. Muitas vezes é também considerado o erro relativo aproximado definido por

$$\frac{|\Delta x^*|}{|x^*|}.$$

A noção de erro relativo advém do facto de o mesmo erro absoluto poder ter significados reais diferentes consoante o valor exacto em causa.

Os erros relativos exprimem-se habitualmente em termos percentuais. Por exemplo, um erro relativo de 0.02 é normalmente referido como um erro de 2%.

Define-se também **erro máximo relativo**, normalmente indicado por $\varepsilon'$, como sendo um majorante do erro relativo, isto é,

$$\varepsilon' = \frac{\varepsilon}{|x|},$$

onde $\varepsilon$ representa um erro máximo absoluto. Também aqui é normal trabalhar com a aproximação do erro máximo relativo dada por (notar o abuso de notação)

$$\varepsilon' = \frac{\varepsilon}{|x^*|},$$

valor que é possível calcular com base na aproximação $x^*$ e no erro máximo absoluto $\varepsilon$ conhecido. Assim, dizer que $x^*$ é uma aproximação de $x$ com um erro máximo relativo $\varepsilon'$ é equivalente a dizer que o valor exacto $x$ está no intervalo $[x^*(1-\varepsilon'), x^*(1+\varepsilon')]$. Neste caso, utiliza-se a notação $x = x^* \pm (100\varepsilon')\,\%$.

**Exemplo 1.2.3.** *Ao escrever-se* $x = 1.2 \pm 5\%$*, pretende dizer-se que* $1.2$ *é uma aproximação de* $x$ *com um erro máximo relativo de* $5\%$ *(ou seja,* $0.05$*). Significa isto que o valor exacto* $x$ *estará no intervalo* $[1.2 \cdot (1 - 0.05), 1.2 \cdot (1 + 0.05)]$*, ou seja,* $[1.14, 1.26]$.

Para uma dada aproximação $x^*$, o erro máximo relativo pode ser calculado a partir do erro máximo absoluto conhecido e vice-versa, ainda que de uma forma aproximada. Habitualmente,

os erros máximos quer absolutos quer relativos são indicados com um número reduzido de casas decimais (raramente mais do que duas).

**Exemplo 1.2.4.**
*Seja* $x^* = 3.45$ *com* $\varepsilon = 0.01$. *Então* $\varepsilon' \simeq \frac{0.01}{3.45} \simeq 3 \times 10^{-3}$.
*Seja* $x^* = -2.7$ *com* $\varepsilon' = 0.07$. *Então* $\varepsilon \simeq 0.07 \times 2.7 \simeq 0.19$.

A utilização abusiva do majorante do erro relativo dado por $\frac{\varepsilon}{|x^*|}$ é justificada pelo facto de normalmente se ter que $\varepsilon \ll |x|$, ou, equivalentemente, $\varepsilon' \ll 1$, resultando em que os valores $\frac{\varepsilon}{|x^*|}$ e $\frac{\varepsilon}{|x|}$ sejam muito próximos. Isto será tanto mais verdade quando mais pequeno for $\varepsilon'$.

## 1.3 Algarismos significativos

Um número real $x$ é representado na forma decimal (base 10) pelo seu sinal ($+$ ou $-$) e por uma sequência (finita ou não) de algarismos do conjunto $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ posicionada relativamente ao ponto (ou vírgula) decimal (.), ou seja,

$$x = \pm\, d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} d_{-3} \ldots$$

A necessidade de representar números de diferentes grandezas de uma forma compacta conduziu à introdução da designada **notação científica**, que mais não é do que a introdução na representação de um factor multiplicativo correspondente a uma potência inteira da base de representação, ou seja, de base 10. Assim, teremos

$$x = \pm\, d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} d_{-3} \ldots \times 10^e$$

A parte da representação $d_n d_{n-1} \ldots d_1 d_0 . d_{-1} d_{-2} d_{-3}$ é designada por **mantissa** e o número inteiro $e$ designa-se por **expoente**. A localização do ponto decimal na mantissa pode ser alterada, bastando para tal modificar o valor do expoente de forma correspondente. Por exemplo, o número 10.23 poderá ser também representado por $1.023 \times 10$, $0.1023 \times 10^2$, $102.3 \times 10^{-1}$, etc.. Note-se que mesmo a representação decimal tradicional permite representar certos números de mais do que uma forma (o número 2 pode também ser representado por $1.9999999999\ldots$, embora esta última seja infinita!).

Como na prática apenas podemos utilizar representações finitas e por vezes não queremos ou não podemos utilizar mais do que um dado número de algarismos da mantissa surge a questão de como representar um valor suposto exacto que à partida não será representável. Concretamente, suponhamos que temos um valor com a mantissa $d_1 d_2 \ldots d_n d_{n+1} d_{n+2} \ldots$ (não interessa a localização do ponto decimal, visto que pode ser alterada por simples manipulação do expoente) e que apenas pretendemos utilizar os $n$ primeiros algarismos. Podemos aqui utilizar dois processos: a **truncatura** e o **arredondamento**.

No caso da truncatura, ignoram-se os algarismos da mantissa a partir do índice $n+1$, tendo em atenção que os que correspondam a algarismos inteiros devem ser substituídos por zeros e posteriormente eliminados por alteração de expoente. A representação assim obtida diferirá do valor original menos do que uma unidade da última casa decimal não eliminada.

**Exemplo 1.3.1.** *Ao truncar os números* 123.56 *e* 123.51 *às décimas, obtemos em ambos os casos* 123.5. *Ao truncar o número* 7395 *para as centenas, obteríamos* $73 \times 10^2$.

No caso do arredondamento, o objectivo é escolher o número representável mais próximo do valor original. Para tal, utilizam-se as seguintes regras

1. se $0.d_{n+1}d_{n+2}\ldots > 0.5$ soma-se uma unidade à casa decimal $n$ (e alteram-se se necessário as casas à esquerda desta), ou seja, arredonda-se para **cima**;
2. se $0.d_{n+1}d_{n+2}\ldots < 0.5$ mantém-se a casa decimal $n$, ou seja, arredonda-se para **baixo**;
3. se $0.d_{n+1}d_{n+2}\ldots = 0.5$ arredonda-se para cima ou para baixo de forma a que o algarismo da casa decimal $n$ seja **par** (neste caso é também possível utilizar o arredondamento para cima).

Estas regras asseguram que toda a representação aproximada obtida por arredondamento difere do valor original não mais do que 5 unidades da primeira casa não representada.

**Exemplo 1.3.2.** *Arredondar às décimas os números:* 1.26*,* 1.24*,* 1.25 *e* 1.35.
*De acordo com as regras acima temos:* 1.3*,* 1.2*,* 1.2 *e* 1.4*, respectivamente.*

A utilização da notação $x = x^* \pm \varepsilon$, atrás introduzida, para indicar que $x^*$ é uma aproximação de $x$ com um erro máximo absoluto $\varepsilon$ tende a ser algo extensa e por tal pouco prática. Uma forma de tornar mais simples a representação de aproximações é considerar majorantes do erro absoluto apenas da forma $0.5 \times 10^n$ e representar apenas a aproximação até à casa decimal $10^n$, ficando implícito qual o majorante do erro absoluto. Quando se utiliza esta convenção, os algarismos da mantissa de uma representação, com excepção dos zeros à esquerda, designam-se **algarismos significativos**. É de notar que esta simplificação da notação acarreta uma perda de informação, pois o erro máximo absoluto inicial, $\varepsilon$, será sempre substituído por um seu majorante da forma $0.5 \times 10^n$.

A passagem de uma aproximação da forma $x^* \pm \varepsilon$ para uma representação apenas com algarismos significativos é normalmente efectuada em dois passos: primeiro majora-se $\varepsilon$ por um número da forma $0.5 \times 10^n$, depois arredonda-se $x^*$ para a casa decimal $10^n$.

**Exemplo 1.3.3.** *A aproximação* $2.1 \pm 0.04$ *corresponde a dizer que o valor exacto está no intervalo* $[2.06, 2.14]$. *Esta aproximação representar-se-ia simplesmente por* 2.1*, significando agora que o valor exacto estaria no intervalo* $[2.05, 2.15]$.

O exemplo seguinte mostra que por vezes é necessário considerar um majorante maior de $\varepsilon$, de forma a garantir que todos os valores exactos possíveis estão considerados.

**Exemplo 1.3.4.** *A aproximação* $51.231 \pm 0.023$ *corresponde a dizer que o valor exacto está no intervalo* $[51.208, 51.254]$. *Majorando* $0.023$ *por* $0.05$ *e arredondando* $51.231$ *para as décimas, seríamos levados a utilizar a representação* $51.2$. *Contudo esta representação apenas retrata valores no intervalo* $[51.15, 51.25]$, *não englobando todos os valores iniciais possíveis, sendo por isso inválida. Ter-se-ia então que considerar o majorante* $0.5$ *para o erro absoluto e representar a aproximação apenas por* $51$, *o que indicaria que o valor exacto estaria no intervalo* $[50.5, 51.5]$.

O exemplo acima ilustra como pode ser enorme a perda de informação ao utilizar representações apenas com algarismos significativos. Efectivamente, inicialmente sabia-se que o valor exacto estava num intervalo de largura 0.046 e no fim apenas se pode concluir que ele está num intervalo de largura 1. Para evitar estas situações podem utilizar-se algarismos suplementares, que se sabe não serem significativos, e que são representados entre parêntesis. Assim, a aproximação do exemplo acima representar-se-ia por $51.2(3)$, representando valores no intervalo $[51.18, 51.28]$. É importante não confundir esta notação com a utilizada para representar algarismos que se repetem em dízimas infinitas periódicas!

**Exemplo 1.3.5.** *A tabela seguinte mostra alguns exemplos de aproximações em que apenas se representam os algarismos significativos. Em cada caso, é apresentado o erro máximo absoluto, o menor intervalo em que se garante estar o valor exacto, o número de algarismos significativos, bem como o erro máximo relativo.*

| $x^*$ | $\varepsilon$ | *Intervalo* | *Algarismos significativos* | $\varepsilon'$ |
|---|---|---|---|---|
| 2.24 | 0.005 | $[2.235, 2.245]$ | *3* | $2.2 \times 10^{-3}$ |
| 2.240 | 0.0005 | $[2.2395, 2.2405]$ | *4* | $2.2 \times 10^{-4}$ |
| $1.5 \times 10^2$ | 5 | $[145, 155]$ | *2* | $3.3 \times 10^{-2}$ |
| $1 \times 10^2$ | 50 | $[50, 150]$ | *1* | $5 \times 10^{-1}$ |
| $0.1 \times 10^3$ | 50 | $[50, 150]$ | *1* | $5 \times 10^{-1}$ |

*Note-se a diferença entre as duas primeiras situações, onde se realça a utilização de um zero à direita depois do ponto decimal para significar a existência de mais um algarismo significativo e, logo, de um erro máximo absoluto* 10 *vezes menor.*

Este exemplo ilustra também que o erro máximo relativo diminui à medida que aumenta o número de algarismos significativos de uma aproximação. De facto, existe uma relação entre estas quantidades, como se mostra no teorema abaixo.

**Teorema 1.3.1.** *Uma aproximação com* $n$ *algarismos significativos tem um erro relativo aproximado inferior ou igual a* $5 \times 10^{-n}$.

*Demonstração.* Se $x^*$ é uma aproximação com $n$ algarismos significativos, então $x^*$ é da forma

$$x^* = \pm d_1 d_2 \cdots d_n \times 10^k,$$

para algum $k \in \mathbb{Z}$ e com $d_1 \neq 0$. De acordo com a convenção utilizada, esta aproximação terá um erro máximo absoluto $\varepsilon = 0.5 \times 10^k$ (metade da última casa decimal representada).

O erro máximo relativo (aproximado) $\varepsilon'$ satisfaz

$$\varepsilon' = \frac{\varepsilon}{|x^*|} = \frac{0.5 \times 10^k}{d_1 d_2 \cdots d_n \times 10^k} = \frac{0.5}{d_1 d_2 \cdots d_n}.$$

Como $d_1 \neq 0$ tem-se que $10^{n-1} \leq d_1 d_2 \cdots d_n < 10^n$, concluindo-se finalmente que

$$\varepsilon' \leq \frac{0.5}{10^{n-1}} = 5 \times 10^{-n}.$$ □

## 1.4 Sistemas de vírgula flutuante

A representação mais comum de números reais em sistemas computacionais é realizada em **vírgula flutuante**. Um sistema de vírgula flutuante é caracterizado por 4 parâmetros: a base de representação ($\beta$), o número de dígitos da mantissa ($n$) e os valores máximos e mínimos do expoente ($m$ e $M$, respectivamente). Tal sistema é habitualmente representado por $\mathrm{FP}(\beta, n, m, M)$. Assim, dizer que $x \in \mathrm{FP}(\beta, n, m, M)$ é equivalente a ter

$$x = \pm(0.d_1 d_2 \ldots d_n) \times \beta^e$$

onde $e$ é um inteiro tal que $m \leq e \leq M$, e $d_i$, para $i = 1, \ldots, n$, são dígitos na base $\beta$. Note-se que habitualmente se tem que $m < 0 < M$, de forma a tornar possível representar números com valores absolutos menores e maiores do que a unidade.

Habitualmente, os sistemas computacionais utilizam sistemas de vírgula flutuante de base 2, de forma a que apenas seja necessário utilizar os dígitos "0" e "1".

Obviamente que um sistema de vírgula flutuante apenas permite representar um subconjunto finito de números reais. Nestes sistemas, o conjunto de expoentes permitidos limita a gama de valores representáveis e o número de dígitos da mantissa caracteriza a precisão com que se podem aproximar números que não tenham representação exacta.

Diz-se ainda que um sistema de vírgula flutuante se encontra **normalizado** se apenas permitir representações de números cujo primeiro algarismo da mantissa seja diferente de zero, isto é, $d_1 \neq 0$, isto para além de permitir a representação do número zero.

Independentemente de se tratar de um sistema normalizado ou não, qualquer sistema de vírgula flutuante terá a si associado o número diferente de zero com menor valor absoluto representável bem como o número com o maior valor absoluto representável.

Quando se utiliza um sistema de vírgula flutuante, as operações aritméticas serão realizadas sobre números representáveis nesse sistema. Contudo, em muitas situações o resultado da operação não terá representação exacta nesse sistema. Desta forma o valor fornecido pelo sistema computacional será um valor aproximado (tipicamente obtido por arredondamento ou truncatura). Os erros resultantes de tais aproximações serão analisados na secção seguinte.

Situações há, todavia, em que o resultado de uma dada operação se encontra fora da gama de valores representáveis, seja porque o seu valor absoluto é não nulo mas inferior ao menor valor absoluto representável, seja porque o seu valor absoluto é superior ao maior valor absoluto representável. A primeira destas situações é designada por **underflow** e a segunda por **overflow**. Nestes casos não é aconselhável utilizar um número do sistema de vírgula flutuante para representar o resultado, pois o erro relativo de tal aproximação poderá ser arbitrariamente elevado. Por tal motivo, é comum os sistemas computacionais tratarem as situações de overflow e underflow como situações de erro. Refira-se também que muitos sistemas computacionais não sinalizam a ocorrência de underflow, limitando-se a fornecer o valor 0 como resultado da operação em causa.

**Exemplo 1.4.1.** *Consideremos um hipotético sistema de vírgula flutuante* $\mathrm{FP}(10, 3, -10, 30)$ *normalizado. Sejam ainda os números*

$$x = 0.200 \times 10^{-8}$$
$$y = 0.400 \times 10^{-5}$$
$$z = 0.600 \times 10^{28}$$

*todos com representação exacta neste sistema.*

*O resultado da operação* $x \times y$ *é*

$$0.8 \times 10^{-14}.$$

*Este resultado não é representável no sistema considerado por o expoente ser inferior ao menor expoente representável. De facto o menor número positivo representável é* $0.1 \times 10^{-10}$*. Assim a operação* $x \times y$ *resulta numa situação de underflow.*

*O resultado da operação* $z/x$ *é*

$$0.3 \times 10^{37}.$$

*Este valor é superior ao maior valor (positivo) representável no sistema considerado, que é,* $0.999 \times 10^{30}$*. Verifica-se assim que a operação* $z/x$ *resulta numa situação de overflow.*

Do exposto acima, pode facilmente concluir-se que a implementação de um sistema de vírgula flutuante pode ser bastante complexa, sendo necessário definir, para além dos parâmetros $(\beta, n, m, M)$, os algoritmos que implementam as operações aritméticas básicas, a forma como são aproximados os resultados que não possuem representação exacto, o tratamento de situações de

underflow e overflow, entre outros. Assim, diferentes versões de um mesmo sistema de vírgula flutuante $\mathrm{FP}(\beta, n, m, M)$, podem diferir em termos de implementação de arredondamentos, tratamento de excepções, entre outros. De tal, facto resulta que as mesmas operações aritméticas, com os mesmos dados de entrada, possam produzir resultados diferentes, mesmo quando à partida se crê estar a usar o mesmo sistema de vírgula flutuante. Este facto pode ser bastante desvantajoso, nomeadamente em termos de repetibilidade de resultados, portabilidade de código de computação numérica e validação de resultados. Como resposta a estas desvantagens surgiu em 1985 a norma IEEE 754 que define formatos para precisões simples, dupla e estendida, bem como directrizes de implementação dos procedimentos de cálculo, arredondamentos e tratamento de excepções. Esta norma tem vindo a ser adoptada pelos fabricantes de sistemas computacionais.

## 1.5 Aritmética em representações finitas

O cálculo de uma expressão envolvendo múltiplas operações aritméticas realizadas utilizando representações finitas deve ser efectuado com algum cuidado. De facto, a necessidade de guardar resultados intermédios, obviamente utilizando uma representação finita, faz com que se cometam diversos erros de arredondamento desses resultados intermédios, erros esses que se podem ir acumulando à medida que os cálculos progridem, podendo resultar em elevados erros no resultado final.

Um dos pontos a considerar advém do facto de operações aritméticas que habitualmente gozam de associatividade (como a soma e a multiplicação) poderem perder essa propriedade quando se trabalha em representações finitas. O exemplo seguinte ilustra este efeito.

**Exemplo 1.5.1.** *Calcular* $0.5 + 0.024 + 0.012$ *utilizando* 2 *dígitos em vírgula flutuante.*

*a) Somando da esquerda para a direita*

$$\begin{aligned}(0.50 \times 10^0 + 0.24 \times 10^{-1}) + 0.12 \times 10^{-1} &= (0.50 \times 10^0 + 0.02 \times 10^0) + 0.12 \times 10^{-1} \\ &= 0.52 \times 10^0 + 0.01 \times 10^0 \\ &= 0.53 \times 10^0\end{aligned}$$

*b) Somando da direita para a esquerda*

$$\begin{aligned}0.50 \times 10^0 + (0.24 \times 10^{-1} + 0.12 \times 10^{-1}) &= 0.50 \times 10^0 + 0.36 \times 10^{-1} \\ &= 0.50 \times 10^0 + 0.04 \times 10^0 \\ &= 0.54 \times 10^0\end{aligned}$$

*Utilizando aritmética exacta o resultado seria sempre* $0.536$.

Este exemplo mostra que ao somar números de magnitudes diferentes poderão ser "perdidos" algarismos menos significativos do número de menor magnitude, sendo o resultado afectado de um erro.

Este problema poderá ocorrer também ao somar sequencialmente um elevado número de parcelas de magnitudes semelhantes e com o mesmo sinal: de facto, a magnitude da soma parcial poderá tornar-se elevada face à das parcelas, originando erros no processo de soma. Tal efeito pode tornar-se muito nefasto, fazendo com que o resultado final obtido com aritmética finita esteja muito longe do verdadeiro valor. Por exemplo, se numa máquina com 4 dígitos de mantissa tentarmos somar sequencialmente um milhão de parcelas de valor 1, obtemos como resultado final o valor $10^4$, e não $10^6$! Efectivamente, nessa máquina hipotética, a soma de $10^4$ com 1 resulta em $10^4$. Este problema poderá ser evitado quer utilizando máquinas com precisão (leia-se número de dígitos da mantissa) suficiente, ou então, organizando os cálculos de uma forma alternativa, por exemplo, somando as parcelas duas a duas, e depois tais somas novamente duas as duas, etc.

Outro caso que é necessário ter em atenção é a subtracção de dois números quase iguais. Aqui, o resultado poderá ter um erro máximo absoluto da sua ordem de grandeza, originando um erro relativo elevado. Este fenómeno de perda de algarismos significativos é designado por **cancelamento subtractivo**.

**Exemplo 1.5.2.** *Efectuar a subtracção* $2.034 - 2.016$ *utilizando* 3 *dígitos em vírgula flutuante.*

***Resolução***
*Em primeiro lugar é necessário representar os números em questão apenas com* 3 *dígitos. Arredondando os dois números dados para* 3 *algarismos obtém-se* 2.03 *e* 2.02*, respectivamente. O resultado aproximado da subtracção, utilizando os números arredondados é* $x^* = 0.01$*.*

*O valor exacto da subtracção é* 0.018*, pelo que o erro absoluto de* $x^*$ *é* 0.008 *e o seu erro relativo é* 44%*, aproximadamente.*

O cancelamento subtractivo pode levar a resultados com elevados erros relativos que são sempre indesejáveis. No entanto, é por vezes possível dispor os cálculos de forma a evitar tal cancelamento.

**Exemplo 1.5.3.** *Seja* $x \gg 1$ *e* $y = \sqrt{x+1} - \sqrt{x}$*. O cálculo de y pela expressão dada pode originar um erro relativo elevado devido ao cancelamento subtractivo. Contudo, a expressão equivalente*

$$y = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

*permite calcular y, evitando tal fenómeno.*

## 1.6 Propagação de erros no cálculo de funções

Nesta secção iremos analisar como se propagam os erros de aproximação no cálculo de funções. Abordaremos primeiro o caso de uma função real de variável real e posteriormente o caso de uma função real de variável vectorial.

Seja então $f : \mathbb{R} \to \mathbb{R}$. A situação que iremos tratar pode descrever-se do seguinte modo: conhecendo uma aproximação $x^*$ de $x$, que valor $y^*$ considerar para aproximar $y = f(x)$ e como relacionar os erros de aproximação de $x^*$ e de $y^*$?

No caso de a função $f$ ser contínua verifica-se que à medida que $x^*$ se aproxima de $x$ mais o valor $f(x^*)$ se aproxima de $f(x)$. Nesta situação, que é a mais usual, pode utilizar-se o valor $y^* = f(x^*)$ como aproximação de $y = f(x)$.

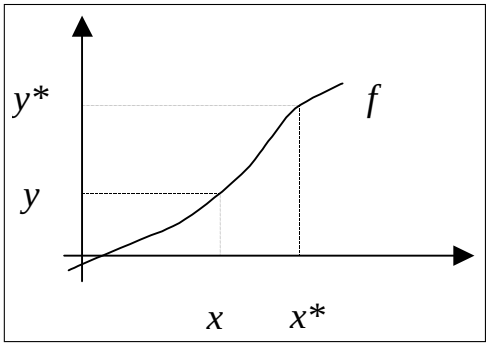



Figura 1.2: $f(x^*)$ aproximação de $f(x)$.

Para além da determinação do valor aproximado de $y^* = f(x^*)$, interessa também caracterizar o erro cometido nesta aproximação, ou melhor, relacionar este erro com o erro de aproximação de $x$ por $x^*$. É claro que o erro $\Delta y^* = y - y^*$ dependerá do erro $\Delta x^* = x - x^*$ e também da função $f$ em questão. De facto, o erro de aproximação $\Delta y^*$ é obtido pela expressão

$$\Delta y^* = y - y^* = f(x) - f(x^*) = f(x^* + \Delta x^*) - f(x^*).$$

Se a função $f$ for continuamente diferenciável, a aplicação do teorema do valor médio permite escrever

$$f(x^* + \Delta x^*) - f(x^*) = f'(\bar{x}) \cdot \Delta x^*$$

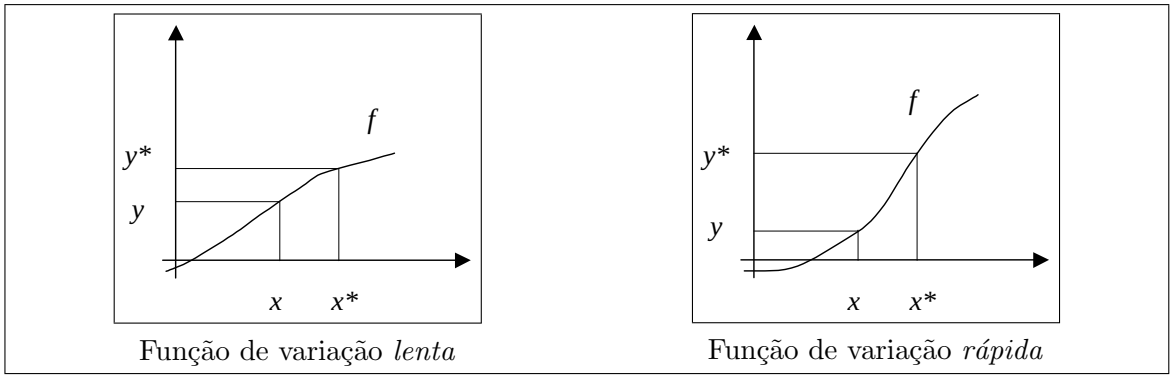



Figura 1.3: Influência de $f$ na propagação de erros.

para algum $\bar{x}$ entre $x^*$ e $x^* + \Delta x^*$. Obtém-se então que

$$\Delta y^* = f'(\bar{x}) \cdot \Delta x^*,$$

ou ainda,

$$|\Delta y^*| = |f'(\bar{x})| \cdot |\Delta x^*|. \tag{1.6.1}$$

Sendo $\varepsilon_x$ um majorante para $|\Delta x^*|$ conclui-se que $|\Delta y^*| \leq |f'|_{|\max} \cdot \varepsilon_x$. Então, o valor

$$\varepsilon_y = |f'|_{|\max} \cdot \varepsilon_x$$

é um majorante para o erro absoluto da aproximação $y^*$ de $y$. Nesta expressão, o valor máximo de $|f'|$ é determinado no intervalo $[x^* - \varepsilon_x, x^* + \varepsilon_x]$.

**Exemplo 1.6.1.** *Calcular um valor aproximado de* $y = \sin x$ *e o correspondente erro máximo absoluto quando* $x \approx 0.57$ *(isto é,* $x = 0.57 \pm 0.005$*).*

***Resolução***

*Um valor aproximado será* $\bar{y} = \sin \bar{x} = \sin 0.57 \simeq 0.5396$*.*

*O erro máximo absoluto será*

$$\varepsilon_y \leq \max_x \left|\frac{dy}{dx}\right| \cdot \varepsilon_x = \max_x |\cos x| \cdot \varepsilon_x$$

*No intervalo em questão, a função* cos *é positiva e decrescente. Então*

$$\varepsilon_y \leq \cos(0.57 - 0.005) \cdot 0.005 \simeq 4.2 \times 10^{-3}$$

*Finalmente tem-se que* $y = 0.5396 \pm 4.2 \times 10^{-3}$*, ou ainda,* $y \approx 0.54 \pm 5 \times 10^{-3}$*.*

Partindo da equação (1.6.1) pode escrever-se que

$$\frac{|\Delta y^*|}{|y|} = \left|\frac{f'(\bar{x})x}{y}\right| \cdot \frac{|\Delta x^*|}{|x|}$$

permitindo obter o majorante para o erro relativo de $y^* = f(x^*)$ dado por

$$\varepsilon'_y = \left|f'(x) \cdot \frac{x}{f(x)}\right|_{\max} \cdot \varepsilon'_x$$

onde $\varepsilon'_x = \frac{\varepsilon_x}{|x|}$, e o máximo de $\left|\frac{xf'(x)}{f(x)}\right|$ é determinado no intervalo $[x^* - \varepsilon_x, x^* + \varepsilon_x]$.

Dados $x \in \mathbb{R}$ e uma função $f$, o **número de condição** de $f$ em $x$ é definido como sendo

$$\left|\frac{xf'(x)}{f(x)}\right|.$$

Este valor pode ser utilizado para avaliar a perda ou o ganho de algarismos significativos no cálculo de uma função, uma vez que caracteriza a ampliação ou redução do erro relativo. Quando o número de condição for reduzido a função diz-se **bem condicionada**. Quando o número de condição for elevado a função diz-se **mal condicionada** e o erro relativo é amplificado.

**Exemplo 1.6.2.** *Quantos dígitos significativos se podem perder no cálculo da função* $y = \tan(x)$ *quando* $x$ *está próximo de* $1$*? E quando* $x$ *está próximo de* $1.5$*?*

***Resolução***
*Como* $\frac{dy}{dx} = 1 + \tan^2(x)$ *tem-se que*

$$\left|\frac{dy}{dx} \cdot \frac{x}{y}\right|_{x=1} = \left|\frac{(1+\tan^2(x)) \cdot x}{\tan(x)}\right|_{x=1} = \frac{1+\tan^2(1)}{\tan(1)} \approx 2.2 > 1$$

*podendo perder-se um dígito significativo.*

*Repetindo os cálculos para* $x = 1.5$*, obter-se-ia* $\left|\frac{dy}{dx} \cdot \frac{x}{y}\right| \approx 21$*, concluindo-se que em tal caso se poderiam perder até* $2$ *dígitos significativos.*

Passemos agora a analisar o caso em que $y$ depende de diversas variáveis, isto é, quando $y = f(x_1, x_2, \ldots, x_n)$, onde $f$ é uma função de $\mathbb{R}$ em $\mathbb{R}^n$, que se considera continuamente diferenciável.

Para cada $i = 1, \ldots, n$, seja $x_i^*$, um valor aproximado de $x_i$, com erro máximo absoluto $\varepsilon_{x_i}$. Nestas condições verifica-se que

$$y^* = f(x_1^*, x_2^*, \ldots, x_n^*)$$

será um valor aproximado de $y = f(x_1, x_2, \ldots, x_n)$ com erro máximo absoluto

$$\varepsilon_y = \sum_{i=1}^{n} \left( \left|\frac{\partial f}{\partial x_i}\right|_{\max} \cdot \varepsilon_{x_i} \right),$$

onde cada um dos máximos das derivadas parciais de $f$ em relação às diversas variáveis independentes é determinado em $\prod_{i=1}^{n} [x_i - \varepsilon_{x_i}, x_i + \varepsilon_{x_i}]$.

É também possível obter o erro relativo máximo para $y^*$ dado por

$$\varepsilon'_y = \sum_{i=1}^{n} \left( \left|\frac{\partial f}{\partial x_i} \cdot \frac{x_i}{f}\right|_{\max} \cdot \varepsilon'_{x_i} \right).$$

Nesta expressão, considera-se que $\varepsilon'_{x_i}$ é um majorante do erro relativo de $x_i^*$, para $i = 1, \ldots, n$. As maximizações são ainda realizadas no conjunto indicado acima, tomando-se agora $\varepsilon_{x_i} = \varepsilon'_{x_i}|x_i|$.

**Exemplo 1.6.3.** *O erro máximo absoluto no cálculo de* $s = a + b$ *pode ser obtido a partir dos erros máximos absolutos em* $a$ *e* $b$ *da seguinte forma*

$$\varepsilon_s = \left|\frac{\partial s}{\partial a}\right|_{\max} \cdot \varepsilon_a + \left|\frac{\partial s}{\partial b}\right|_{\max} \cdot \varepsilon_b = \varepsilon_a + \varepsilon_b.$$

**Exemplo 1.6.4.** *O erro máximo relativo no cálculo de* $w = xyz$*, pode ser obtido a partir dos erros máximos relativos em* $x$*,* $y$ *e* $z$ *da seguinte forma*

$$\begin{aligned}
\varepsilon'_w &= \left|\frac{\partial w}{\partial x} \cdot \frac{x}{w}\right|_{\max} \cdot \varepsilon'_x + \left|\frac{\partial w}{\partial y} \cdot \frac{y}{w}\right|_{\max} \cdot \varepsilon'_y + \left|\frac{\partial w}{\partial z} \cdot \frac{z}{w}\right|_{\max} \cdot \varepsilon'_z \\
&= \left|yz \cdot \frac{x}{xyz}\right|_{\max} \cdot \varepsilon'_x + \left|xz \cdot \frac{y}{xyz}\right|_{\max} \cdot \varepsilon'_y + \left|xy \cdot \frac{z}{xyz}\right|_{\max} \cdot \varepsilon'_z \\
&= \varepsilon'_x + \varepsilon'_y + \varepsilon'_z.
\end{aligned}$$

A terminar esta exposição é conveniente salientar a importância de nas expressões de propagação de erros absolutos e relativos se considerar o valor máximo possível para o factor de amplificação (ou redução do erro). Efectivamente, só esta maximização garante que se conseguem obter majorantes para os erros nas variáveis dependentes a partir dos erros nas variáveis independentes. Contudo, em análises mais simplificadas da propagação de erros apenas se considera o valor de tal factor num ponto (normalmente o valor aproximado da variável independente). Este tipo de análise é por vezes suficiente pois nem sempre interessa conhecer um majorante do erro, mas apenas a sua ordem de grandeza.

## 1.7 Cálculo de séries e erro de truncatura

Por vezes a determinação de um certo valor envolve a realização de uma sucessão infinita de operações. O erro cometido quando se toma uma aproximação resultante da realização de um número finito de operações designa-se **erro de truncatura**.

Um dos casos em que se surge o erro de truncatura é no caso da aproximação da soma $S$ de uma série convergente $\sum_{i=0}^{\infty} a_i$ pela soma parcial $S_n = \sum_{i=0}^{n} a_i$. Neste caso, o erro de truncatura será $R_n = S - S_n$.

No caso geral não é simples determinar o número de termos a somar para calcular o valor da série com um dado erro máximo pretendido. Há contudo um tipo de séries, as séries alternadas, em que esta tarefa é bastante simples, como refere o teorema seguinte.

**Teorema 1.7.1.** *Considere-se a sucessão $\{a_n\}_{n=0}^{\infty}$ decrescente e de termos não negativos, ou seja, $a_0 \geq a_1 \geq \ldots \geq a_n \geq \ldots \geq 0$. Estão a série $\sum_{i=0}^{\infty}(-1)^i a_i$ é convergente para um número $S$. Verifica-se ainda que a soma parcial $S_n = \sum_{i=0}^{n}(-1)^i a_i$ verifica a relação*

$$|S - S_n| \leq a_{n+1},$$

*ou seja, o erro de truncatura é, em valor absoluto, inferior ou igual ao primeiro termo não considerado.*

**Exemplo 1.7.1.** *A série alternada*

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$$

*é convergente para o valor $\frac{\pi}{4}$. Determinar quantos termos são necessários para calcular este valor com um erro inferior a $10^{-4}$.*

***Resolução***

*O termo geral desta série é $\frac{(-1)^n}{2n+1}$, para $n = 0, 1, \ldots$. Para se garantir o erro pretendido, o primeiro termo a não considerar deverá satisfazer*

$$\frac{1}{2n+1} \leq 10^{-4}$$

*ou seja,* $n \geq 4999.5$*. Como* $n$ *é inteiro far-se-á* $n = 5000$*, pelo que se deverão somar os termos de* $0$ *até* $4999$*.*

O erro de truncatura é particularmente importante quando se efectua a aproximação de uma função por polinómios de Taylor, reduzindo assim o seu cálculo à realização de operações de soma, subtracção, multiplicação e divisão, que são as operações aritméticas elementares à custa das quais todos os cálculos numéricos são realizados.

O desenvolvimento de Taylor de uma função $f$ em torno do ponto $x_0$ permite escrever

$$f(x) = \underbrace{f(x_0) + f'(x_0)(x - x_0) + \cdots + f^{(n)}(x_0)\frac{(x - x_0)^n}{n!}}_{P_{x_0,n}(x)} + R_{x_0,n}(x)$$

onde $R_{x_0,n}(x) = f^{(n+1)}(x_0 + (x - x_0)\theta)\frac{(x-x_0)^{n+1}}{(n+1)!}$ para algum $\theta \in [0, 1]$.

O erro de truncatura na aproximação $f(x) \approx P_{x_0,n}(x)$ é dado pelo resto de Taylor $R_{x_0,n}(x)$. Se se verificar que $R_{x_0,n}(x) \xrightarrow[n\to+\infty]{} 0$ a aproximação por polinómios de Taylor pode ser tão boa quanto se queira, bastando para tal considerar um número suficientemente elevado de termos.

**Exemplo 1.7.2.** *Considere aproximações da função* $e^x$ *no intervalo* $[-2, 2]$ *dadas por polinómios de Taylor. Qual deverá ser o grau do polinómio a utilizar se se pretender que o erro absoluto devido à truncatura da série seja inferior a* $5 \times 10^{-5}$*?*

***Resolução***

*O desenvolvimento de Taylor em torno de* $0$ *é*

$$e^x = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!} + R_n(x),$$

*onde* $R_n(x) = e^{\theta x}\frac{x^{n+1}}{(n+1)!}$*, para* $\theta \in [0, 1]$*.*

*O erro absoluto devido à truncatura pode ser majorado da seguinte forma*

$$\varepsilon_{trunc} = |R_n(x)| = \left| e^{\theta x}\frac{x^{n+1}}{(n+1)!} \right| \leq 8\frac{2^{n+1}}{(n+1)!}$$

*uma vez que* $\theta \in [0, 1]$ *e* $x \in [-2, 2]$*.*

*Calculando estes majorantes para alguns valores de* $n$*, obtêm-se os seguintes valores*

| $n$ | $8\frac{2^{n+1}}{(n+1)!}$ |
|---|---|
| 10 | $4.1 \times 10^{-4}$ |
| 11 | $6.8 \times 10^{-5}$ |
| 12 | $1.1 \times 10^{-6}$ |
| 13 | $1.5 \times 10^{-7}$ |

*Conclui-se então que para* $n = 12$ *se tem* $\varepsilon_{trunc} \leq 1.0 \times 10^{-5}$*, devendo-se portanto utilizar um polinómio de grau* $12$*.*

# Capítulo 2

# Equações Não Lineares

## 2.1 Introdução

Neste capítulo iremos estudar alguns métodos para a resolução numérica de equações algébricas não lineares, isto é, equações que se possam escrever na forma $f(x) = 0$, onde $f$ é uma função real de variável real. Todo o valor $s$ que anula $f$, isto é, tal que $f(s) = 0$, designa-se por **zero** da função $f$ ou **solução** da equação $f(x) = 0$.

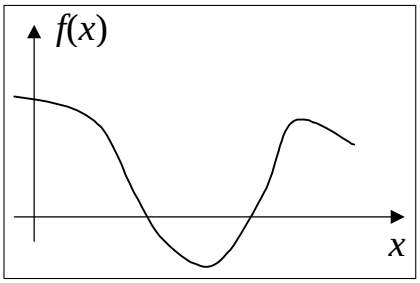



Figura 2.1: Zeros de uma função

Perante uma equação do tipo $f(x) = 0$, antes de tentar aplicar um qualquer método de resolução, é importante garantir que de facto a equação tenha solução, ou seja, que existe um real $s$ tal que $f(s) = 0$. Muitas vezes importa também determinar se a solução é única, ou se existem diferentes soluções e, neste caso, saber qual ou quais importa determinar.

Os métodos de resolução de uma equação do tipo $f(x) = 0$ podem dividir-se em dois grandes grupos: métodos **directos** e métodos **iterativos**.

Nos primeiros, a equação é resolvida por intermédio de expressões que envolvem a função $f$. As soluções da equação são determinadas de uma forma exacta após um número finito de operações (supondo a utilização de aritmética exacta). Estes métodos apenas se aplicam a alguns tipos de problemas. Um exemplo é a fórmula resolvente de equações do 2º grau.

Os métodos iterativos caracterizam-se por gerarem sucessões convergentes para as soluções da equação a resolver. Estes métodos distinguem-se entre si pela forma como são geradas as sucessões de soluções aproximadas. Os métodos iterativos são aplicáveis vastas gamas de problemas.

Contrariamente aos métodos directos, que exigem formas bem específicas da função $f$ (por exemplo, funções afins, quadráticas, etc.), a aplicação de métodos iterativos exige apenas a satisfação de condições sobre propriedades mais gerais da função $f$, como sejam continuidade, monotonia, diferenciabilidade, ou limites inferiores ou superiores de derivadas.

Tipicamente, a aplicação de um método iterativo parte de uma **estimativa inicial**, $x_0$, da solução a determinar. Por aplicação de um procedimento bem definido, vão sendo gerados os termos de uma sucessão de estimativas $\{x_n\}$ que se pretende que convirja para a solução $s$ pretendida. Em cada **iteração** é calculado um termo da sucessão, ou seja, uma **nova estimativa**, $x_k$, à custa da estimativa anterior, $x_{k-1}$, por intermédio de uma regra que caracteriza o método. Este processo iterativo é terminado assim que a estimativa $x_k$ satisfaz um dado **critério de paragem** (por exemplo $x_k$ estar próximo de $s$ ou $f(x_k)$ ser próximo de 0) ou após um número máximo de iterações ou tempo de processamento.

| | | | |
|---|---|---|---|
| valor inicial | $x_0$ | : | $f(x_0) \neq 0$ |
| 1ª iteração | $\downarrow$ | | |
| | $x_1$ | : | $f(x_1) \neq 0$ |
| 2ª iteração | $\downarrow$ | | |
| | $x_2$ | : | $f(x_2) \neq 0$ |
| | $\downarrow$ | | |
| | $\vdots$ | | |
| iteração $k$ | $\downarrow$ | | |
| | $x_k$ | : | $f(x_k) \approx 0$ ou $x_k \approx s$ |

Figura 2.2: Aplicação de um método iterativo

Quando se pretendem determinar múltiplas soluções de uma equação, será necessário aplicar o método iterativo para cada uma das soluções a calcular. Estas aplicações deverão necessariamente partir de estimativas iniciais $x_0$ diferentes.

A aplicação bem sucedida de um método iterativo para a determinação de uma solução da equação $f(x) = 0$ envolve um conjunto de questões que interessa analisar. A mais importante destas prende-se com a convergência da sucessão das estimativas $\{x_n\}$ gerada pelo método. Como iremos ver, é possível estabelecer condições, associadas a cada método, que uma vez satisfeitas garantem que a sucessão gerada converge para a solução da equação pretendida. Estas condições são designadas por condições suficientes de convergência. É claro que existem situações em que

os métodos produzem sucessões convergentes para a solução sem que as condições suficientes sejam satisfeitas, mas ... será que vale a pena arriscar?

Outro aspecto a considerar é já referido critério de paragem. Uma vez que é normal apenas se garantir que a sucessão $\{x_n\}$ converge para a solução $s$, não é de supor que se tenha $x_k = s$ a partir de uma dada iteração. O critério de paragem não é mais do que uma regra, a avaliar em cada iteração, que permite decidir se se pára na estimativa mais actual ou se continua a calcular novas estimativas. Em abstracto devemos terminar a aplicação do método iterativo assim que a estimativa da iteração $k$, $x_k$, esteja suficientemente próxima da solução $s$, de acordo com uma tolerância definida. Note-se que que como $s$ não é conhecido, para aplicar este tipo de critério será necessário proceder a uma majoração do erro de aproximação. Uma possibilidade é terminar a aplicação do método assim que $f(x_k)$ seja suficientemente próximo de zero. Como se verá, é muitas vezes possível relacionar o erro de aproximação $s - x_k$ com o valor de $f(x_k)$. Nas implementações computacionais dos métodos iterativos é ainda usual estabelecer um número máximo de iterações ao fim das quais o método é terminado, mesmo que não se verifique qualquer outro critério de paragem.

Finalmente, mas não de menor importância, há a considerar a maior ou menor rapidez de convergência da sucessão $\{x_n\}$ para a solução pretendida $s$. De uma forma simplista, a rapidez de convergência é medida através da evolução do erro de aproximação $e_k = s - x_k$ em função do índice de iteração $k$. Como iremos ver, esta evolução depende do método aplicado e também das propriedades da função $f$ que define a equação $f(x) = 0$.

Antes de iniciar a exposição dos diferentes métodos iterativos, apresenta-se um resultado que relaciona o valor de uma função num ponto com a distância desse ponto ao zero da função, que será único nas condições do teorema.

**Teorema 2.1.1.** *Seja $f$ uma função continuamente diferenciável no intervalo $[a, b]$. Suponha-se que $m_1 = \min_{\xi \in [a,b]} |f'(\xi)| > 0$ e também que existe $s \in [a, b]$ tal que $f(s) = 0$. Então*

$$|s - x| \leq \frac{|f(x)|}{m_1} \quad \forall x \in [a, b].$$

*Demonstração.* Sendo $x \in [a, b]$, o teorema do valor médio permite afirmar que

$$f(s) - f(x) = f'(\xi)(s - x)$$

para algum $\xi$ entre $x$ e $s$. Então $\xi \in [a, b]$ e, uma vez que $f(s) = 0$, verifica-se

$$|f(x)| = |f'(\xi)| \cdot |s - x| \geq m_1 \cdot |s - x|,$$

obtendo-se o resultado pretendido, pois $m_1 > 0$. □

Repare-se que a partir deste teorema é imediata a obtenção de um critério de paragem. De facto se se parar a aplicação do método assim que $|f(x_k)| \leq m_1 \cdot \varepsilon$, garante-se que o erro absoluto da aproximação $x_k$ está majorado por $\varepsilon$. Contudo é necessário conhecer um minorante em valor absoluto não nulo da derivada da função ($m_1$).

## 2.2 Método das bissecções sucessivas

Consideremos uma função $f$ contínua um intervalo $[a, b]$ e tal que $f(a)$ e $f(b)$ possuem sinais diferentes. O teorema dos valores intermédios permite afirmar que existe um número $s$ no intervalo $[a, b]$ tal que $f(s) = 0$. Para simplificar a exposição vamos supor que tal número $s$ é único.

O método das bissecções sucessivas parte do intervalo inicial $[a, b]$ que se sabe conter o zero de $f$, suposto único. Em cada iteração é produzido um intervalo com metade do comprimento do intervalo actual. Para tal, divide-se o intervalo actual a meio e escolhe-se o subintervalo esquerdo ou direito de forma a que a função tenha sinais diferentes nos extremos do subintervalo escolhido. Ou seja, sendo $[a_n, b_n]$ o intervalo na iteração $n$, calcula-se $x_{n+1} = \frac{a_n+b_n}{2}$. O valor $x_{n+1}$ substitui $a_n$ ou $b_n$ consoante $f(x_{n+1})f(b_n) < 0$ ou $f(x_{n+1})f(a_n) < 0$. Desta forma, assegura-se que $s \in [a_n, b_n]$ em qualquer iteração.

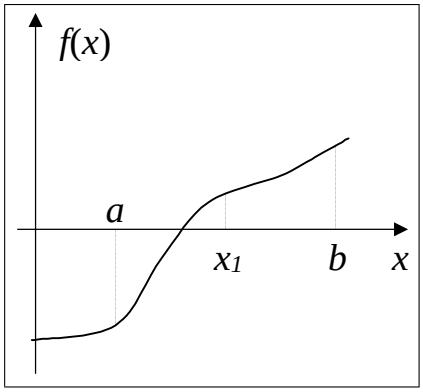



Figura 2.3: Bissecções sucessivas

**Método das bissecções sucessivas**

| | |
|---|---|
| **Inicialização** | $[a_0, b_0] = [a, b]$ |
| **Repetir** | 1. $x_{n+1} = \frac{a_n+b_n}{2}$;<br>2. **Se** $f(x_{n+1})f(a_n) < 0$<br>**Então** $a_{n+1} = a_n$; $b_{n+1} = x_{n+1}$;<br>**Senão** $a_{n+1} = x_{n+1}$; $b_{n+1} = b_n$; |
| **Até** | *verificar critério de paragem* |

O teorema seguinte estabelece condições suficientes para a convergência do método das bissecções sucessivas.

**Teorema 2.2.1.** *Seja $f$ contínua em $[a, b]$ tal que $f(a)f(b) \leq 0$, e seja $s$ o único zero de $f$ nesse intervalo. Então, o método das bissecções sucessivas gera uma sucessão convergente para $s$.*

*Demonstração.* A sucessão $\{a_n\}$ é crescente e limitada e a sucessão $\{b_n\}$ é decrescente e limitada,

pelo que são ambas convergentes.

Como se verifica que $b_n - a_n = \frac{b-a}{2^n}$, conclui-se que $\lim a_n = \lim b_n = z$, para algum $z \in [a, b]$. Como $x_{n+1} = \frac{a_n+b_n}{2}$ tem-se também que $\lim x_n = z$.

A aplicação do método garante que $f(a_n)f(b_n) \leq 0$, para todo o $n$. Então, como $f$ é contínua tem-se que $[f(z)]^2 \leq 0$, o que implica que $f(z) = 0$, ou seja, $z = s$, uma vez que $s$ é, por hipótese, o único zero de $f$ em $[a, b]$. □

Uma vez que $s \in [a_n, b_n]$ e $x_{n+1} = \frac{a_n+b_n}{2}$, verifica-se facilmente que

$$|s - x_{n+1}| \leq \frac{b_n - a_n}{2} = \frac{b-a}{2^{n+1}}$$

Pode então afirmar-se que o erro absoluto da estimativa $x_n$ está majorado por

$$\frac{b-a}{2^n}.$$

O número de iterações suficientes para garantir um erro absoluto não superior a $\delta$ pode ser calculado fazendo $\frac{b-a}{2^n} \leq \delta$ obtendo-se o valor

$$\boxed{n \geq \log_2 \frac{b-a}{\delta}.}$$

O exemplo seguinte ilustra a aplicação deste método.

**Exemplo 2.2.1.** *Determinar uma aproximação com um erro absoluto inferior a* $5 \times 10^{-3}$ *da (única) solução da equação* $1 + x + e^x = 0$ *que se sabe estar no intervalo* $[-2, -1]$.

***Resolução***

*Verificação de condições de convergência*
*A função* $f(x) = 1 + x + e^x$ *é monótona,* $f(-2) = -0.865 < 0$, *e* $f(-1) = 0.368 > 0$.

*Determinação do número de iterações*
*Como se pretende uma precisão de* $5 \times 10^{-3}$ *deve-se escolher* $n$ *tal que*

$$n > \log_2 \frac{-1 - (-2)}{5 \times 10^{-3}} \Rightarrow n > 7.6 \Rightarrow n = 8$$

*Efectuando* 8 *iterações a partir de* $[-2, -1]$ *tem-se um erro máximo absoluto de* $\frac{1}{2^8} \approx 4 \times 10^{-3}$.

*Iterações*
*Partindo do intervalo* $[-2, -1]$, *temos na primeira iteração,*

$$x_1 = \frac{-2 + (-1)}{2} = -1.5$$
$$f(-1.5) = -0.277$$

*Como* $f(-1.5) \cdot f(-2) > 0$ *o novo intervalo será* $[-1.5, -1]$.

*Na segunda iteração temos*

$$x_2 = \frac{-1.5 + (-1)}{2} = -1.25$$
$$f(-1.25) = 0.037$$

*Como* $f(-1.25) \cdot f(-1.5) < 0$ *o novo intervalo será* $[-1.5, -1.25]$.

*A tabela seguinte apresenta os valores resultantes da aplicação do método para as* 8 *iterações necessárias.*

| $n$ | $a_n$ | $f(a_n)$ | $b_n$ | $f(b_n)$ | $x_{n+1}$ | $f(x_{n+1})$ |
|---|---|---|---|---|---|---|
| 0 | $-2.000$ | $-0.865$ | $-1.000$ | $+0.368$ | $-1.500$ | $-0.277$ |
| 1 | $-1.500$ | $-0.277$ | $-1.000$ | $+0.368$ | $-1.250$ | $+0.037$ |
| 2 | $-1.500$ | $-0.277$ | $-1.250$ | $+0.037$ | $-1.375$ | $-0.122$ |
| 3 | $-1.375$ | $-0.122$ | $-1.250$ | $+0.037$ | $-1.313$ | $-0.043$ |
| 4 | $-1.313$ | $-0.043$ | $-1.250$ | $+0.037$ | $-1.281$ | $-0.004$ |
| 5 | $-1.281$ | $-0.004$ | $-1.250$ | $+0.037$ | $-1.266$ | $+0.016$ |
| 6 | $-1.281$ | $-0.004$ | $-1.266$ | $+0.016$ | $-1.273$ | $+0.006$ |
| 7 | $-1.281$ | $-0.004$ | $-1.273$ | $+0.006$ | $-1.277$ | $+0.001$ |

*Solução*
*A solução da equação será* $s = -1.277 \pm 4 \times 10^{-3}$*, ou seja,* $s \in [-1.281, -1.273]$.

## 2.3 Método da falsa posição (*regula falsi*)

O método da falsa posição (também designado por *regula falsi*) permite também determinar o zero (suposto único) de uma função $f$ contínua num intervalo $[a, b]$ que toma valores com sinais opostos nos extremos desse intervalo. A hipótese de existência de apenas um zero em $[a, b]$ visa apenas facilitar a exposição.

Este método é análogo ao método das bissecções, sendo em cada iteração o intervalo $[a_n, b_n]$ dividido em duas partes. No entanto, a divisão do intervalo é feita no ponto $x_{n+1}$, correspondente à intersecção com o eixo dos $xx$ da recta que passa pelos pontos $(a_n, f(a_n))$ e $(b_n, f(b_n))$. Partindo da equação

$$y = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n}(x - a_n)$$

da recta que une os referidos pontos, conclui-se facilmente que

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

É de notar que sendo $f(a_n)f(b_n) < 0$ se tem que $x_{n+1} \in ]a_n, b_n[$.

Na iteração seguinte é utilizado o subintervalo $[a_n, x_{n+1}]$ ou o subintervalo $[x_{n+1}, b_n]$, consoante se verifique que $f(a_n)f(x_{n+1}) < 0$ ou $f(x_{n+1})f(b_n) < 0$. No caso (difícil de detectar) de $f(x_{n+1}) = 0$, a aplicação do método pararia nessa iteração!

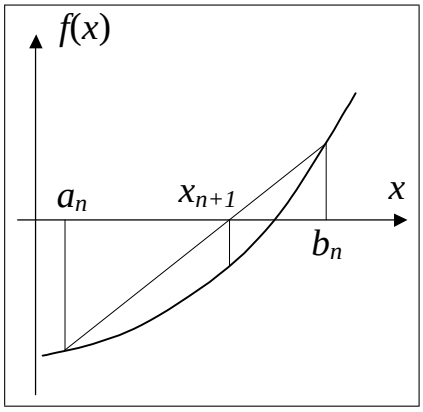



Figura 2.4: Método da falsa posição

O método da falsa posição corresponde a aproximar a função pela recta secante nos extremos do intervalo e a utilizar o zero de recta como estimativa do zero da função (daí o seu nome). Esta aproximação é tanto mais razoável quanto mais o gráfico de $f$ se aproximar de uma recta, ou seja, $f'$ variar pouco (isto no caso de $f$ ser diferenciável).

**Método da falsa posição**

| | |
|---|---|
| **Inicialização** | $[a_0, b_0] = [a, b]$ |
| **Repetir** | 1. $x_{n+1} = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}$;<br>2. **Se** $f(x_{n+1}) f(a_n) < 0$<br>**Então** $a_{n+1} = a_n;\ b_{n+1} = x_{n+1}$;<br>**Senão** $a_{n+1} = x_{n+1};\ b_{n+1} = b_n$; |
| **Até** | *verificar critério de paragem* |

O teorema seguinte estabelece condições suficientes para a convergência do método da falsa posição.

**Teorema 2.3.1.** *Se a função $f$ for contínua e estritamente monótona no intervalo $[a, b]$ e se $f(a)f(b) \leq 0$, então o método da falsa posição produz uma sucessão convergente para o único zero de $f$ nesse intervalo.*

No método da falsa posição, não é possível, de um modo geral, determinar antecipadamente um número de iterações que garanta uma dada precisão na aproximação do zero da função. Assim, quando se pretende determinar o valor do zero com um dado erro máximo absoluto é necessário calcular estimativas do erro ao longo das iterações para verificar a satisfação da precisão requerida.

O teorema seguinte apresenta uma forma de determinar um majorante do erro de aproximação.

**Teorema 2.3.2.** *Seja $f$ uma função continuamente diferenciável no intervalo $[a, b]$ e tal que $f(a)f(b) \leq 0$. Definam-se $m_1 = \min_{\xi \in [a,b]} |f'(\xi)|$ e $M_1 = \max_{\xi \in [a,b]} |f'(\xi)|$, e suponha-se que*

$m_1 > 0$.

*Então, o erro de aproximação de $s$, único zero de $f$ em $[a, b]$, pela estimativa $x_{n+1}$ satisfaz a relação*

$$|s - x_{n+1}| \leq \frac{M_1 - m_1}{m_1}|x_{n+1} - x_n|.$$

O estabelecimento de um critério de paragem com base no majorante do erro definido atrás, pode ser feito como se indica em seguida. Após a determinação da estimativa $x_{n+1}$ (de acordo com a expressão do método da falsa posição) é calculado do majorante do erro absoluto de $x_{n+1}$

$$\varepsilon_{n+1} = \frac{M_1 - m_1}{m_1}|x_{n+1} - x_n|$$

parando-se a aplicação do método assim que este majorante seja inferior a um dado valor pretendido. Para a utilização deste critério de paragem é necessário determinar os valores $m_1$ e $M_1$ antes de iniciar a aplicação do método. É também importante notar que por vezes a estimativa do erro dada por este majorante poderá ser algo pessimista, sendo o erro absoluto em $x_{n+1}$ bastante inferior a $\varepsilon_{n+1}$.

Alternativamente, pode também estabelecer-se um critério de paragem com base no majorante do erro de aproximação fornecido pelo teorema 2.1.1.

O exemplo seguinte ilustra a aplicação do método da falsa posição à equação já resolvida pelo método das bissecções sucessivas.

**Exemplo 2.3.1.** *Utilizar o método da falsa posição para determinar uma aproximação, com um erro absoluto inferior a $5 \times 10^{-3}$, do (único) zero da função $f(x) = 1 + x + e^x$.*

***Resolução***

*Convergência e intervalo inicial*

*$f$ é estritamente monótona e $f(-2)f(-1) < 0$, logo o método converge.*

*Estimação do erro*

$f'(x) = 1 + e^x$

$m_1 = \min_{x\in[-2,-1]} |f'(x)| = 1.1353$

$M_1 = \min_{x\in[-2,-1]} |f'(x)| = 1.3679$

$\Rightarrow \quad \varepsilon_n = 0.205\,|x_{n+1} - x_n|$

*Iterações Para o intervalo $[-2, -1]$ temos*

$$f(-2) = -0.865$$
$$f(-1) = 0.368$$

*pelo que teremos*

$$x_1 = \frac{(-2) \cdot f(-1) - (-1) \cdot f(-2)}{f(-1) - f(-2)} = -1.298.$$

*Como o critério de paragem exige o conhecimento de duas estimativas consecutivas devemos prosseguir as iterações. Sendo* $f(x_1) = -2.55 \times 10^{-2}$*, o novo intervalo será* $[-1.298, -1]$*.*

*Para a segunda iteração temos então*

$$x_2 = \frac{(-1.298) \cdot f(-1) - (-1) \cdot f(-1.298)}{f(-1) - f(-1.298)} = -1.297.$$

*O majorante o erro de aproximação será*

$$\varepsilon_2 = 0.205|x_2 - x_1| = 4 \times 10^{-3}.$$

*Como* $\varepsilon_2 \leq 5 \times 10^{-3}$*, o critério de paragem está satisfeito, pelo o valor aproximado da solução será* $x_2$*.*

*A tabela seguinte apresenta os valores relevantes das iterações efectuadas.*

| $n$ | $a_n$ | $f(a_n)$ | $b_n$ | $f(b_n)$ | $x_{n+1}$ | $f(x_{n+1})$ | $\varepsilon_{n+1}$ |
|---|---|---|---|---|---|---|---|
| 0 | $-2.000$ | $-0.865$ | $-1.000$ | $+0.368$ | $-1.298$ | $-2.55 \times 10^{-2}$ | – |
| 1 | $-1.298$ | $-0.026$ | $-1.000$ | $+0.368$ | $-1.279$ | $-8.22 \times 10^{-4}$ | $+4.0 \times 10^{-3}$ |

*Solução*

*A solução aproximada será então* $s \simeq -1.279$, *com um erro absoluto máximo de* $4.0 \times 10^{-3}$*.*

Neste caso, o método da falsa posição demonstrou ser bastante mais eficiente que o método das bissecções sucessivas. No entanto, este comportamento nem sempre se verifica, como se pode constatar pelo exemplo seguinte.

**Exemplo 2.3.2.** *A determinação do zero de* $x + e^{x^5} - 5$ *no intervalo* $[0, 1.3]$ *pelo método da falsa posição, com um erro máximo de* $5 \times 10^{-3}$*, conduziu aos seguintes resultados (onde a majoração do erro foi efectuada de acordo com o teorema 2.1.1).*

| $n$ | $a_n$ | $f(a_n)$ | $b_n$ | $f(b_n)$ | $x_{n+1}$ | $f(x_{n+1})$ | $\varepsilon_{n+1}$ |
|---|---|---|---|---|---|---|---|
| 0 | $+0.000$ | $-4.000$ | $+1.300$ | $+37.274$ | $+0.126$ | $-3.87$ | $+3.87$ |
| 1 | $+0.126$ | $-3.874$ | $+1.300$ | $+37.274$ | $+0.237$ | $-3.76$ | $+3.76$ |
| 2 | $+0.237$ | $-3.763$ | $+1.300$ | $+37.274$ | $+0.334$ | $-3.66$ | $+3.66$ |
| 3 | $+0.334$ | $-3.662$ | $+1.300$ | $+37.274$ | $+0.420$ | $-3.57$ | $+3.57$ |
| 4 | $+0.420$ | $-3.566$ | $+1.300$ | $+37.274$ | $+0.497$ | $-3.47$ | $+3.47$ |
| 5 | $+0.497$ | $-3.472$ | $+1.300$ | $+37.274$ | $+0.566$ | $-3.37$ | $+3.37$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| 50 | $+1.065$ | $-0.008$ | $+1.300$ | $+37.274$ | $+1.065$ | $-6.64 \times 10^{-3}$ | $+6.64 \times 10^{-3}$ |
| 51 | $+1.065$ | $-0.007$ | $+1.300$ | $+37.274$ | $+1.065$ | $-5.54 \times 10^{-3}$ | $+5.54 \times 10^{-3}$ |
| 52 | $+1.065$ | $-0.006$ | $+1.300$ | $+37.274$ | $+1.065$ | $-4.63 \times 10^{-3}$ | $+4.63 \times 10^{-3}$ |

*Analisando os resultados, verifica-se que o extremo superior do intervalo permanece constante e o extremo inferior converge para o zero de* $f$*.*

*Aplicando o método das bissecções sucessivas ao mesmo problema, garante-se o mesmo erro máximo apenas em* 9 *iterações!*

| $n$ | $a_n$ | $f(a_n)$ | $b_n$ | $f(b_n)$ | $x_{n+1}$ | $f(x_{n+1})$ |
|---|---|---|---|---|---|---|
| 0 | +0.000 | −4.000 | +1.300 | +37.274 | +0.650 | −3.227 |
| 1 | +0.650 | −3.227 | +1.300 | +37.274 | +0.975 | −1.611 |
| 2 | +0.975 | −1.611 | +1.300 | +37.274 | +1.138 | +2.853 |
| 3 | +0.975 | −1.611 | +1.138 | +2.853 | +1.056 | −0.220 |
| 4 | +1.056 | −0.220 | +1.138 | +2.853 | +1.097 | +0.990 |
| 5 | +1.056 | −0.220 | +1.097 | +0.990 | +1.077 | +0.323 |
| 6 | +1.056 | −0.220 | +1.077 | +0.323 | +1.066 | +0.038 |
| 7 | +1.056 | −0.220 | +1.066 | +0.038 | +1.061 | −0.094 |
| 8 | +1.061 | −0.094 | +1.066 | +0.038 | +1.064 | −0.029 |

A convergência lenta do método da falsa posição patente no exemplo acima está relacionada com o facto de um dos extremos do intervalo que contém a solução permanecer inalterado, à medida que o outro extremo vai convergindo (lentamente) para a solução pretendida, como se ilustra na figura.

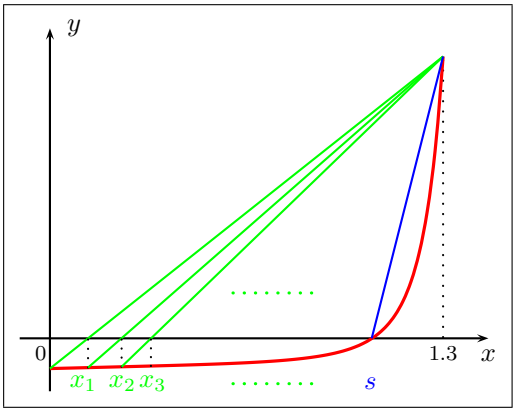



Figura 2.5: Convergência lateral do método da falsa posição

O teorema seguinte justifica este tipo de comportamento característico do método da falsa posição.

**Teorema 2.3.3.** *Se a função $f$ for estritamente monótona e duplamente diferenciável no intervalo $[a, b]$, se $f(a)f(b) \leq 0$ e se o sinal de $f''$ não variar em $[a, b]$, então a sucessão produzida pelo método da falsa posição converge monotonamente para o zero de $f$ nesse intervalo. Também se verifica que um dos extremos do intervalo permanece inalterado.*

Sempre que se verifica este comportamento, o método da falsa posição não fornece uma sucessão de intervalos com largura a convergir para zero, contrariamente ao método das bissecções sucessivas. Assim, é não é possível obter uma estimativa para o erro de aproximação apenas com base na largura do intervalo utilizado em cada iteração. Note-se que este efeito de convergência lateral verifica-se em muitas situações, pois sendo $f''$ contínua, se $f''(s) \neq 0$, então existe uma vizinhança de $s$ em que $f''$ não troca de sinal!

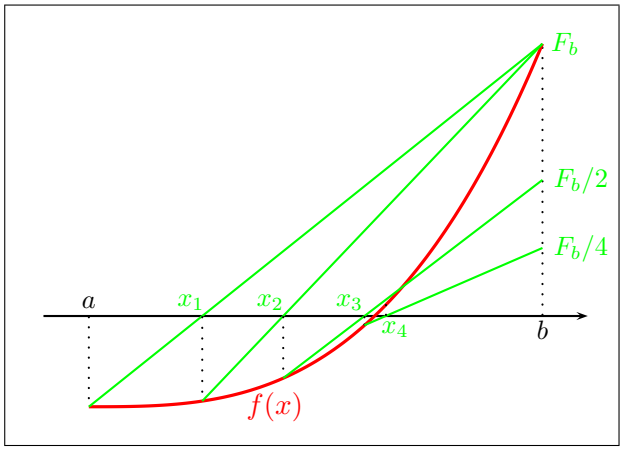



Figura 2.6: Método da falsa posição modificado

O método da **falsa posição modificado** constitui uma alternativa ao método da falsa posição que procura evitar este tipo de comportamento. Este método é em tudo análogo ao da falsa posição, excepto que sempre que $f(x_n)f(x_{n+1}) > 0$ o valor da ordenada do extremo do intervalo que se mantém constante é divido por 2. Procura-se desta forma evitar que um dos extremos do intervalo permaneça fixo durante todo o processo iterativo.

O teorema seguinte apresenta condições suficientes para a convergência do método da falsa posição modificado. Estas condições são em tudo análogas às apresentadas para o método da falsa posição.

**Teorema 2.3.4.** *Se $f$ for contínua, estritamente monótona e tiver sinais contrários nos extremos de um intervalo $[a, b]$, a sucessão produzida pelo método da falsa posição modificado converge para o zero de $f$ em $[a, b]$.*

O estabelecimento de um critério de paragem com base no erro de aproximação da estimativa $x_n$ pode ser feito recorrendo mais uma vez ao majorante fornecido pelo teorema 2.1.1.

**Método da falsa posição modificado**

| | |
|---|---|
| **Inicialização** | $[a_0, b_0] = [a, b];\ F_a = f(a_0);\ F_b = f(b_0)$ |
| **Repetir** | 1. $x_{n+1} = \frac{a_n F_b - b_n F_a}{F_b - F_a}$;<br>2. **Se** $f(x_{n+1})f(a_n) < 0$<br>**Então** $a_{n+1} = a_n;\ b_{n+1} = x_{n+1};\ F_b = f(x_{n+1})$;<br>**Se** $f(x_{n+1})f(x_n) > 0$ **Então** $F_a = \frac{F_a}{2}$;<br>**Senão** $a_{n+1} = x_{n+1};\ b_{n+1} = b_n;\ F_a = f(x_{n+1})$;<br>**Se** $f(x_{n+1})f(x_n) > 0$ **Então** $F_b = \frac{F_b}{2}$; |
| **Até** | *verificar critério de paragem* |

Apresenta-se em seguida a aplicação deste método ao exemplo anterior. Como se pode verificar, o efeito da convergência lateral, lenta por natureza, foi agora eliminado.

**Exemplo 2.3.3.** *Utilizar o método da falsa posição modificado para determinar uma aproximação, com um erro absoluto inferior a* $5 \times 10^{-3}$*, do zero de* $f(x) = x + e^{x^5} - 5$ *no intervalo* $[0, 1.3]$.

***Resolução***

*Convergência*

$f$ *é estritamente monótona e* $f(0)f(1.3) < 0$*, logo o método converge.*

*Estimação do erro*

$f'(x) = 1 + 5x^4 e^{x^5} \quad \Rightarrow \quad \min_{x\in[0,1.3]} |f'(x)| = 1 \quad \Rightarrow \quad \varepsilon_n = |f(x_n)| \leq \delta \quad \Rightarrow \quad |x_n - s| \leq \delta$

*Iterações*

| $n$ | $a_n$ | $F_a$ | $b_n$ | $F_b$ | $x_{n+1}$ | $f(x_{n+1})$ | $\varepsilon_{n+1}$ |
|---|---|---|---|---|---|---|---|
| 0 | +0.000 | −4.000 | +1.300 | +37.274 | +0.126 | −3.87 | +3.87 |
| 1 | +0.126 | −3.874 | +1.300 | +37.274 | +0.237 | −3.76 | +3.76 |
| 2 | +0.237 | −3.763 | +1.300 | +18.637 | +0.415 | −3.57 | +3.57 |
| 3 | +0.415 | −3.572 | +1.300 | +9.318 | +0.660 | −3.21 | +3.21 |
| 4 | +0.660 | −3.206 | +1.300 | +4.659 | +0.921 | −2.14 | +2.14 |
| 5 | +0.921 | −2.138 | +1.300 | +2.330 | +1.102 | +1.20 | +1.20 |
| 6 | +0.921 | −2.138 | +1.102 | +1.198 | +1.037 | $-6.39 \times 10^{-1}$ | $+6.39 \times 10^{-1}$ |
| 7 | +1.037 | −0.639 | +1.102 | +1.198 | +1.060 | $-1.29 \times 10^{-1}$ | $+1.29 \times 10^{-1}$ |
| 8 | +1.060 | −0.129 | +1.102 | +0.599 | +1.067 | $+6.65 \times 10^{-2}$ | $+6.65 \times 10^{-2}$ |
| 9 | +1.060 | −0.129 | +1.067 | +0.066 | +1.065 | $-1.61 \times 10^{-3}$ | $+1.61 \times 10^{-3}$ |

## 2.4 Método iterativo simples

O método iterativo simples, também designado por iteração de ponto fixo, é um método de importância fundamental e simultaneamente de grande simplicidade.

Para aplicar este método à resolução de uma equação do tipo $f(x) = 0$, é necessário em primeiro lugar obter uma equação equivalente a esta que tenha a forma

$$x = F(x),$$

onde $F$ será uma nova função a determinar de modo que as duas equações sejam equivalentes.

Em seguida, escolhe-se um valor inicial $x_0$ e gera-se a sucessão $\{x_n\}$ por intermédio da relação de recorrência

$$x_{n+1} = F(x_n)$$

para $n = 0, 1, \ldots$. A função $F$ é por vezes designada **função de recorrência**.

A justificação do funcionamento deste método reside no seguinte argumento. Se a sucessão $\{x_n\}$ convergir, para um dado valor $s$, e se a função de recorrência $F$ for contínua, verifica-se então que $s = F(s)$, ou seja, que $s$ é um **ponto fixo** da função $F$. Uma vez que por hipótese se tem que $f(x) = 0 \Leftrightarrow x = F(x)$, conclui-se finalmente que $f(s) = 0$, ou seja, que o método iterativo simples, quando convergente, produz sucessões que convergem para zeros da função $f$.

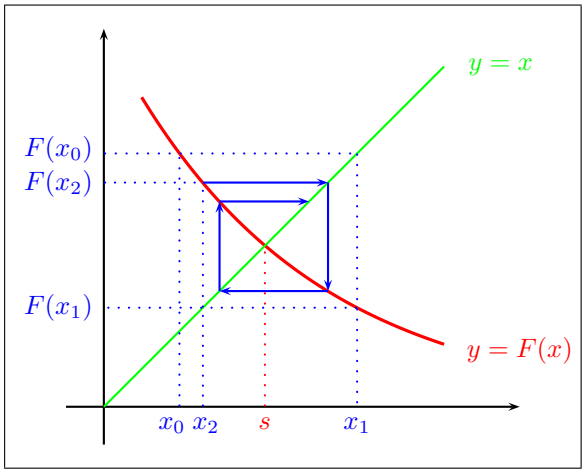



Figura 2.7: Método iterativo simples

A implementação deste método é muito simples, bastando para tal encontrar uma função de recorrência $F$ e um valor inicial $x_0$.

**Método iterativo simples**

| | |
|---|---|
| **Inicialização** | Escolher $x_0$ |
| **Repetir** | $x_{n+1} = F(x_n)$ |
| **Até** | *verificar critério de paragem* |

Habitualmente, a função de recorrência $F$ é obtida por manipulação algébrica da equação $f(x) = 0$ de forma a isolar num dos membros a variável $x$. Por exemplo, para aplicar este método na resolução da equação $x - e^{-x} = 0$ poder-se-ia passar para a equação equivalente $x = e^{-x}$, obtendo-se a função de recorrência $F(x) = e^{-x}$. Poder-se-ia também passar da equação $x = e^{-x}$ para a equação $x = -\ln(x)$, obtendo-se a função de recorrência $\tilde{F}(x) = -\ln(x)$, válida para $x > 0$.

É de referir que para uma dada equação $f(x) = 0$ se pode obter uma infinidade de funções de recorrência $F$. Para isso, basta notar que $f(x) = 0 \Leftrightarrow x = x + rf(x)$ para qualquer $r \neq 0$, tendo-se $F(x) = x + rf(x)$.

Dependendo da função de recorrência $F$ e do valor inicial $x_0$, o método pode ter diferentes comportamentos, alguns dos quais se podem observar na figura 2.4. Como se pode verificar, o método nem sempre converge e, quando converge, a sucessão gerada pode ser monótona ou não. Uma vez que há grande liberdade na escolha da função de iteração, é importante conhecer algum tipo de critério que permita avaliar se uma dada função de recorrência (juntamente com um ponto inicial) gerará ou não uma sucessão convergente para a solução $s$ pretendida.

O teorema seguinte apresenta condições que garantem a convergência do método iterativo simples. Este teorema fornece um critério que permite seleccionar funções de recorrência desejáveis,

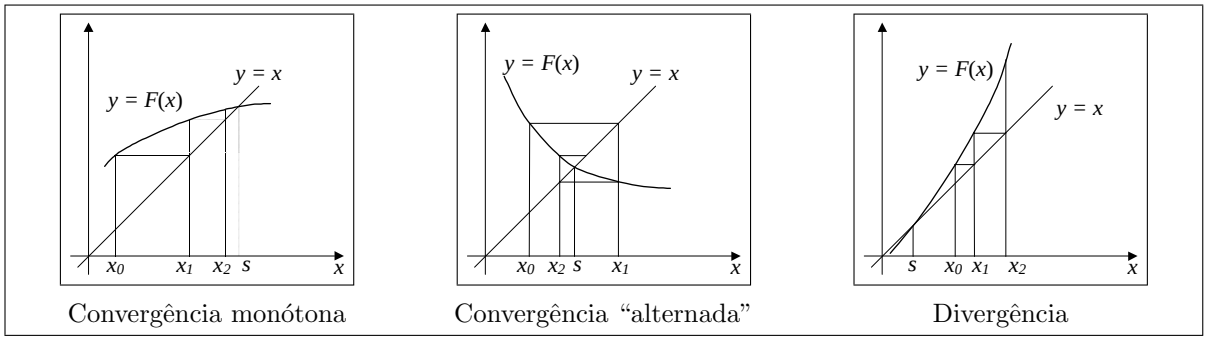



Figura 2.8: Diferentes comportamentos do método iterativo simples

isto é, tais que o método convirja.

**Teorema 2.4.1.** *Se $F$ for continuamente diferenciável em $[a,b]$, $\max_{x\in[a,b]} |F'(x)| < 1$ e existir $s \in [a,b]$ tal que $s = F(s)$, então, para qualquer valor inicial $x_0 \in [a,b]$, a sucessão gerada pelo método iterativo simples converge para $s$.*

*Demonstração.* Seja $L = \max_{x\in[a,b]} |F'(x)|$. Por hipótese temos $L < 1$. Como $s = F(x)$ e $x_1 = F(x_0)$ temos

$$x_1 - s = F(x_0) - F(s) = F'(\xi_0) \cdot (x_0 - s)$$

para algum $\xi_0 \in [a,b]$. Como $x_2 = F(x_1)$ temos

$$x_2 - s = F(x_1) - F(s) = F'(\xi_1) \cdot (x_1 - s) = F'(\xi_1) \cdot F'(\xi_0) \cdot (x_0 - s)$$

para $\xi_0, \xi_1 \in [a,b]$. Continuando este raciocínio conclui-se que

$$x_n - s = F'(\xi_{n-1}) \cdot F'(\xi_{n-2}) \cdots F'(\xi_0) \cdot (x_0 - s)$$

onde $\xi_0, \ldots \xi_{n-1} \in [a,b]$. Então

$$|x_n - s| = |F'(\xi_{n-1})| \cdot |F'(\xi_{n-2})| \cdots |F'(\xi_0)| \cdot |x_0 - s| \leq L^n \cdot |x_0 - s|.$$

Como $0 \leq L < 1$, então $L^n \to 0$ e logo $|x_n - s| \to 0$, ou seja $x_n \to s$. □

Este teorema permite afirmar que se a função de recorrência for tal que $|F'(s)| < 1$, o método iterativo simples converge desde que o valor inicial $x_0$ esteja suficientemente próximo de $s$. Das muitas (infinitas!) possibilidades de escolha de $F$ é necessário seleccionar uma que verifique $|F'(x)| < 1$ numa vizinhança da solução.

Uma vez analisada a questão da convergência, vamos agora estudar o comportamento do erro de aproximação, de forma a se poder estabelecer um critério de paragem.

Na demonstração do teorema acima obteve-se a expressão

$$|x_n - s| \leq L^n \cdot |x_0 - s|, \tag{2.4.1}$$

onde $L = \max_{x\in[a,b]} |F'(x)|$, que se supõe ser inferior a 1. Esta expressão fornece um majorante do erro de aproximação de $x_{n+1}$ com base no erro de aproximação de $x_0$. Ora este último não é habitualmente conhecido e um seu majorante conhecido pode ser bastante pessimista, pelo que será interessante encontrar outra expressão para o erro de aproximação. No entanto, a expressão (2.4.1) permite desde já prever que quanto mais próximo de zero for $L$, mais rapidamente convergirá para zero o erro de aproximação, pelo que menos iterações serão necessárias para alcançar uma dada precisão pretendida.

Para obter uma expressão para o erro de aproximação de $x_{n+1}$, vamos partir novamente da aplicação do teorema do valor médio para a função $F$ no intervalo de extremos $x_n$ e $s$, garantindo-se a existência de $\xi_n$ nesse intervalo tal que $F(x_n)-F(s) = F'(\xi_n)(x_n-s)$. Agora pode escrever-se

$$\begin{aligned}
x_{n+1} - s &= F'(\xi_n)\cdot(x_n - s)\\
x_{n+1} - s &= F'(\xi_n)\cdot(x_n - s - x_{n+1} + x_{n+1})\\
|x_{n+1} - s| &= |F'(\xi_n)|\cdot|x_{n+1} - s + x_n - x_{n+1}|\\
|x_{n+1} - s| &\leq L\cdot|x_{n+1} - s + x_n - x_{n+1}|\\
|x_{n+1} - s| &\leq L\cdot(|x_{n+1} - s| + |x_n - x_{n+1}|)\\
(1-L)\cdot|x_{n+1} - s| &\leq L\cdot|x_n - x_{n+1}|\\
|x_{n+1} - s| &\leq \frac{L}{1-L}\cdot|x_n - x_{n+1}|,
\end{aligned}$$

onde $L = \max_x |F'(x)|$ se supõe menor do que 1.

O valor $\varepsilon_{n+1} = \frac{L}{1-L}|x_{n+1} - x_n|$ constitui assim um majorante do erro em $x_{n+1}$, majorante esse que pode ser calculado após a determinação de $x_{n+1}$. Se se pretender determinar $s$ com um erro absoluto inferior a um dado $\delta$, definido à partida, dever-se-á terminar a aplicação do método assim que $\varepsilon_{n+1} \leq \delta$. Para utilizar este critério de paragem, é apenas necessário determinar $L$ antes de iniciar a aplicação do método. Note-se no entanto que esta determinação de $L$ é muitas vezes necessária para garantir a convergência do método.

**Exemplo 2.4.1.** *Utilizar o método iterativo simples para determinar uma aproximação, com um erro absoluto inferior a $5\times10^{-5}$, do (único) zero da função $f(x) = 1 + x + e^x$, que se sabe estar no intervalo $[-2,-1]$.*

***Resolução***

*Função de iteração e valor inicial*

*Fazendo $F(x) = -1 - e^x$ tem-se que $f(x) = 0 \Leftrightarrow x = F(x)$.*

*Como $F'(x) = -e^x$, verifica-se que $L = \max_{x\in[-2,-1]} |F'(x)| = 0.3679 < 1$.*
*Escolhendo $x_0 = -2$ garante-se a convergência do método.*

*A função $\tilde{F}(x) = \ln(-1-x)$ não poderá ser utilizada pois tem-se que $\max_x |\tilde{F}'(x)| > 1$ em qualquer vizinhança da solução!*

*Estimação do erro*
$\varepsilon_{n+1} = \frac{L}{1-L} |x_{n+1} - x_n| = 0.582 \cdot |x_{n+1} - x_n|$

*Critério de paragem*
*Estabelecendo o critério de paragem* $\varepsilon_{n+1} \leq 5 \times 10^{-5}$*, garante o erro máximo pretendido.*

*Iteração* 1
$x_1 = F(x_0) = -1 - e^{x_0} = -1.13534$
$\varepsilon_1 = 0.582 \cdot |x_1 - x_0| = 5 \times 10^{-1}$*. Como* $\varepsilon_1 \nleq 5 \times 10^{-5}$ *continua-se a aplicação do método.*

*Iteração* 2
$x_2 = F(x_1) = -1 - e^{x_1} = -1.32131$
$\varepsilon_2 = 0.582 \cdot |x_2 - x_1| = 1.1 \times 10^{-1}$*. Como* $\varepsilon_2 \nleq 5 \times 10^{-5}$ *continua-se a aplicação do método.*

*Iterações*
*A tabela seguinte apresenta os resultados da aplicação do método até à satisfação do critério de paragem.*

| $n$ | $x_n$ | $x_{n+1} = F(x_n)$ | $\varepsilon_{n+1}$ |
|---|---|---|---|
| 0 | $-2.00000$ | $-1.13534$ | $+5.0 \times 10^{-1}$ |
| 1 | $-1.13534$ | $-1.32131$ | $+1.1 \times 10^{-1}$ |
| 2 | $-1.32131$ | $-1.26678$ | $+3.2 \times 10^{-2}$ |
| 3 | $-1.26678$ | $-1.28174$ | $+8.7 \times 10^{-3}$ |
| 4 | $-1.28174$ | $-1.27756$ | $+2.4 \times 10^{-3}$ |
| 5 | $-1.27756$ | $-1.27872$ | $+6.8 \times 10^{-4}$ |
| 6 | $-1.27872$ | $-1.27839$ | $+1.9 \times 10^{-4}$ |
| 7 | $-1.27839$ | $-1.27848$ | $+5.2 \times 10^{-5}$ |
| 8 | $-1.27848$ | $-1.27846$ | $+1.5 \times 10^{-5}$ |

*Solução*
*A estimativa obtida será* $s \simeq -1.27846$*, com um erro absoluto inferior a* $2 \times 10^{-5}$*.*

## 2.5 Método de Newton

O método de Newton é um dos métodos mais poderosos para resolver equações do tipo $f(x) = 0$. Tal como no caso do método iterativo simples (de que pode ser considerado um caso particular), este método parte de uma estimativa inicial $x_0$ e gera uma sucessão $\{x_n\}$ de uma forma recorrente.

Cada novo valor da sucessão, $x_{n+1}$, é determinado como sendo a abcissa do ponto de intersecção com o eixo dos $xx$ da recta tangente ao gráfico da função no ponto $(x_n, (f(x_n))$, ou seja, no ponto correspondente ao valor anterior da sucessão.

A expressão de recorrência que permite determinar $x_{n+1}$ em função de $x_n$ obtém-se facilmente

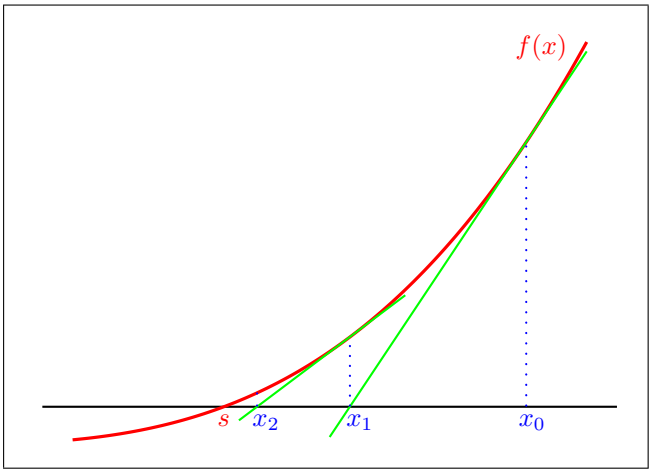



Figura 2.9: Método de Newton

notando que a recta tangente ao gráfico de $f$ no ponto $(x_n, (f(x_n))$ pode ser descrita pela equação

$$y = f(x_n) + f'(x_n) \cdot (x - x_n).$$

De acordo com o exposto atrás, esta recta passará também pelo ponto $(x_{n+1}, 0)$. Substituindo na equação da recta este ponto e resolvendo a equação obtida em ordem a $x_{n+1}$ obtém-se

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

que será então a expressão de recorrência do método de Newton. Refira-se que neste método se tem também que $x_{n+1} = F(x_n)$ para a função de recorrência

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

Note-se ainda que se $f'(x) \neq 0$ se tem que

$$f(x) = 0 \Leftrightarrow x = x - \frac{f(x)}{f'(x)}.$$

**Método de Newton**

| | |
|---|---|
| **Inicialização** | Escolher $x_0$ |
| **Repetir** | $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ |
| **Até** | *verificar critério de paragem* |

Antes de apresentar condições que garantem a convergência do método de Newton, mostram-se graficamente na figura 2.5 algumas situações em que o método não produz sucessões convergentes para a solução da equação que se pretende calcular.

O teorema apresentado em seguida fornece condições suficientes para a convergência do método de Newton. Estas condições não são, em geral, necessárias, isto é, há situações em que elas não

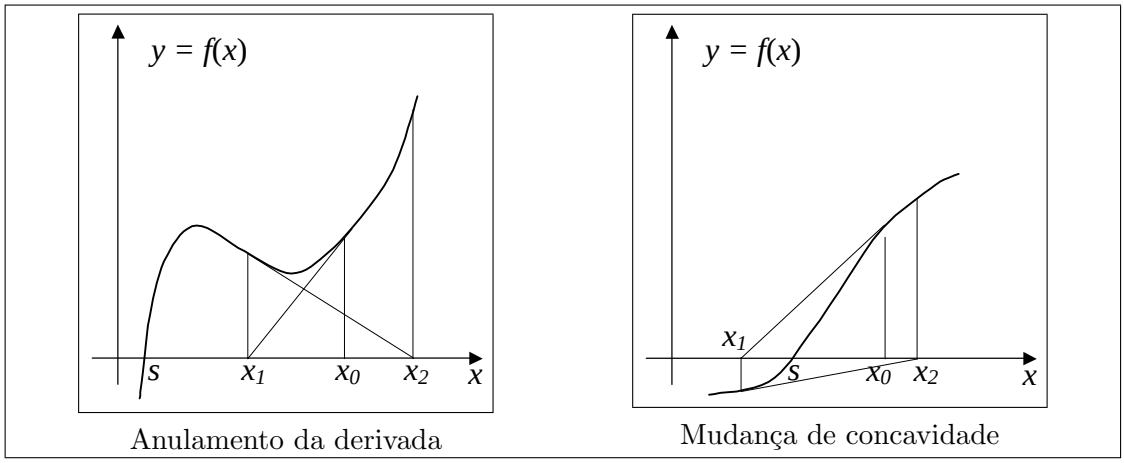



Figura 2.10: Alguns comportamentos indesejáveis do método de Newton

se verificam e o método converge. Refira-se também que é possível estabelecer outras condições suficientes de convergência.

**Teorema 2.5.1.** *Seja $f \in C^2([a,b];\mathbb{R})$ tal que $f'(x) \neq 0$, e $f''(x) \leq 0$ ou $f''(x) \geq 0$ em $[a,b]$. Seja ainda $s$ o (único) zero de $f$ em $[a,b]$. Então a sucessão gerada pelo método de Newton converge para $s$ sempre que o ponto inicial $x_0 \in [a,b]$ satisfizer $f(x_0)f''(x_0) \geq 0$. Mais ainda, a sucessão gerada é monótona.*

*Demonstração.*

Consideremos o caso $f' > 0$ e $f'' \geq 0$ (nos outros casos a demonstração é semelhante).

1. Seja então $x_0 \in [a,b]$ tal que $f(x_0) \geq 0$, por forma a que $f(x_0)f''(x_0) \geq 0$.
2. Como $f$ é crescente tem-se então que $x_0 \geq s$.
3. Como $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$, tem-se ainda que $x_1 \leq x_0$.
4. O desenvolvimento de Taylor de $f$ em torno do ponto $x_0$ permite escrever

$$f(s) = f(x_0) + f'(x_0)(s - x_0) + \frac{f''(\xi_0)}{2}(s - x_0)^2,$$

   para algum $\xi_0$ entre $x_0$ e $s$. Como, por hipótese, $f(s) = 0$, tem-se

$$s - x_0 = -\frac{f(x_0)}{f'(x_0)} - \frac{f''(\xi_0)}{2f'(x_0)}(s - x_0)^2 \leq -\frac{f(x_0)}{f'(x_0)}$$

   atendendo aos sinais de $f'$ e $f''$. Como $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$, então $x_1 \geq s$ e também $f(x_1) \geq 0$.
5. Supondo que $x_n \geq s$, e argumentando como atrás, é possível concluir que $x_{n+1} \leq x_n$ que $x_{n+1} \geq s$ e ainda que $f(x_{n+1}) \geq 0$.

6. Acabou de se mostrar, por indução, que $\{x_n\}$ é decrescente e limitada inferiormente por $s$. Então $\{x_n\}$ é convergente, para um dado valor $z$, no intervalo $[a, b]$.

7. Como $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n}$ e $f$ e $f'$ são funções contínuas, então no limite tem-se $z = z - \frac{f(z)}{f'(z)}$, ou ainda $f(z) = 0$.

8. Sendo este zero único (devido à monotonia estrita de $f$) conclui-se finalmente que $z = s$.

□

Vamos agora determinar a evolução do erro de aproximação para as estimativas geradas pelo método de Newton. Na exposição que se segue supõe-se que $\{x_n\} \subset [a, b]$.

1. Mais uma vez, do desenvolvimento de Taylor de $f$ em torno de $x_n$, garante-se a existência de $\xi_n$ entre $x_n$ e $x_{n+1}$ tal que

$$f(x_{n+1}) = f(x_n) + f'(x_n)(x_{n+1} - x_n) + \frac{f''(\xi_n)}{2}(x_{n+1} - x_n)^2.$$

2. Da expressão de recorrência do método de Newton, $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, podemos concluir que $f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0$, verificando-se assim que

$$f(x_{n+1}) = \frac{f''(\xi_n)}{2}(x_{n+1} - x_n)^2. \tag{2.5.1}$$

3. Do desenvolvimento de Taylor de $f$ em torno de $s$, garante-se a existência de $\zeta_n$ entre $x_{n+1}$ e $s$, tal que

$$f(x_{n+1}) = f(s) + f'(\zeta_n)(x_{n+1} - s).$$

   Uma vez que $f(s) = 0$, esta expressão pode ser escrita na forma

$$f(x_{n+1}) = f'(\zeta_n)(x_{n+1} - s). \tag{2.5.2}$$

4. Combinando agora as expressões (2.5.1) e (2.5.2), pode escrever-se

$$f'(\zeta_n)(x_{n+1} - s) = \frac{f''(\xi_n)}{2}(x_{n+1} - x_n)^2,$$

   ou ainda,

$$|f'(\zeta_n)||x_{n+1} - s| = \frac{|f''(\xi_n)|}{2}|x_{n+1} - x_n|^2.$$

5. Definindo agora $M_2 = \max_{x\in[a,b]} |f''(x)|$ e $m_1 = \min_{x\in[a,b]} |f'(x)|$, e supondo que $m_1 > 0$, pode afirmar-se que

$$|x_{n+1} - s| \leq \frac{M_2}{2m_1}|x_{n+1} - x_n|^2,$$

   expressão esta que poderá ser utilizada para determinar o majorante do erro de aproximação de $x_{n+1}$, dado por

$$\varepsilon_{n+1} = \frac{M_2}{2m_1}|x_{n+1} - x_n|^2,$$

**Exemplo 2.5.1.** *Utilizar o método de Newton para determinar uma aproximação, com um erro absoluto inferior a* $5 \times 10^{-6}$*, do (único) zero da função* $f(x) = 1 + x + e^x$*, que se sabe estar no intervalo* $[-2, -1]$*.*

***Resolução***

*Condições de convergência*

$f'(x) = 1 + e^x \quad \rightarrow \quad f' > 0$

$f''(x) = e^x \quad \rightarrow \quad f'' > 0$

*O método converge desde que* $x_0$ *esteja à direita do zero, garantindo* $f(x_0)f''(x_0) > 0$*. Então, escolhendo* $x_0 = -1$*, garante-se a convergência do método.*

*Estimação do erro*

*Utilizando a estimativa do erro de aproximação atrás deduzida temos*

$$m_1 = \min_{x \in [-2,-1]} |f'(x)| = 1 + e^{-2} = 1.1353$$

$$M_2 = \max_{x \in [-2,-1]} |f''(x)| = e^{-1} = 0.3679$$

$$\frac{M_2}{2m_1} = 0.162$$

*pelo que* $\varepsilon_{n+1} = 0.162 \cdot |x_{n+1} - x_n|^2$ *será um majorante do erro de* $x_{n+1}$*.*

*Critério de paragem*

*De acordo com a majoração do erro o critério de paragem a utilizar será* $\varepsilon_{n+1} \leq 5 \times 10^{-6}$*.*

*Iteração 1*

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = -1.26894$$

$$\varepsilon_1 = 0.162 \cdot |x_1 - x_0|^2 = 1.2 \times 10^{-1}$$

*Como* $\varepsilon_1 \not\leq 5 \times 10^{-6}$*, devemos prosseguir as iterações.*

*Iteração 2*

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -1.27485$$

$$\varepsilon_2 = 0.162 \cdot |x_2 - x_1|^2 = 1.5 \times 10^{-5}$$

*Como* $\varepsilon_2 \not\leq 5 \times 10^{-6}$*, devemos prosseguir as iterações.*

*Iterações*

*A tabela seguinte resume a aplicação do método.*

| $n$ | $x_n$ | $f(x_n)$ | $f'(x_n)$ | $x_{n+1}$ | $\varepsilon_{n+1}$ |
|---|---|---|---|---|---|
| 0 | $-1.00000$ | $+3.68 \times 10^{-1}$ | $+1.368$ | $-1.26894$ | $+1.2 \times 10^{-1}$ |
| 1 | $-1.26894$ | $+1.22 \times 10^{-2}$ | $+1.281$ | $-1.27845$ | $+1.5 \times 10^{-5}$ |
| 2 | $-1.27845$ | $+1.27 \times 10^{-5}$ | $+1.278$ | $-1.27846$ | $+1.6 \times 10^{-11}$ |

*Solução*
*A solução aproximada será* $s \simeq -1.27846$ *(com todos os algarismos exactos).*

Neste exemplo verificou-se que o método de Newton apresentou uma convergência bastante mais rápida do que os métodos anteriores, conseguindo-se uma precisão maior num menor número de iterações. Como será discutido mais à frente, o método de Newton é na generalidade dos casos um método de convergência mais rápida. Note-se no entanto que a sua aplicação exige o cálculo de valores da derivada da função e também que as condições para a sua convergência podem ser mais difíceis de verificar.

A terminar a exposição sobre o método de Newton, apresenta-se em seguida um teorema que fornece outras condições suficientes para a convergência do método de Newton. Este teorema justifica a constatação de que o método de Newton, de uma forma geral, é convergente desde que parta de uma estimativa inicial $x_0$ suficientemente próxima da solução $s$ a determinar.

**Teorema 2.5.2.** Sendo $f \in C^2([a,b];\mathbb{R})$ e $s$ um zero de $f$ em $[a,b]$, tal que $f'(s) \neq 0$, então existe $\delta > 0$ tal que a sucessão $\{x_n\}$ gerada pelo método de Newton converge para $s$ sempre que $x_0 \in [s-\delta, s+\delta]$.

## 2.6 Método da secante

O método da secante é semelhante ao método de Newton, com a diferença de que a recta tangente ao gráfico da função é substituída (como o próprio nome indica) pela recta secante nos dois últimos pontos. Este método obriga a que em cada iteração sejam guardadas as duas últimas estimativas da solução a determinar.

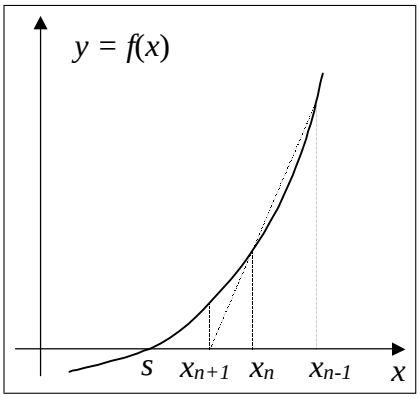



Figura 2.11: Método da secante

A recta que passa pelos pontos $(x_{n-1}, f(x_{n-1}))$ e $(x_n, f(x_n))$ é descrita pela equação

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

Como a estimativa $x_{n+1}$ é dada pela abcissa da intersecção desta recta com o eixo dos $xx$, tem-se que o ponto $(x_{n+1}, 0)$ é um ponto desta recta. Fazendo esta substituição rapidamente se conclui que $x_{n+1}$ será dado pela expressão

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

onde se pode notar a semelhança com a expressão de recorrência do método da falsa posição ou, equivalente, pela expressão

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

que salienta a sua semelhança ao método de Newton.

Na aplicação do método da secante não se garante que $f$ tome em $x_n$ e $x_{n-1}$ valores com sinais opostos. Assim, o ponto $x_{n+1}$ poderá não estar entre $x_n$ e $x_{n+1}$. Este método poderá não convergir quando aplicado a problemas em que o método da falsa posição converge.

**Método da secante**

| | |
|---|---|
| **Inicialização** | Escolher $x_{-1}$ e $x_0$ |
| **Repetir** | $x_{n+1} = \frac{x_{n-1}f(x_n)-x_n f(x_{n-1})}{f(x_n)-f(x_{n-1})}$ |
| **Até** | *verificar critério de paragem* |

O seguinte resultado (que não será aqui demonstrado) fornece condições suficientes para a convergência do método da secante. É de notar a semelhança entre estas condições e as condições do teorema 2.5.1 relativo ao método de Newton.

**Teorema 2.6.1.** *Seja $f \in C^2([a,b];\mathbb{R})$ tal que $f'(x) \neq 0$, e $f''(x) \leq 0$ ou $f''(x) \geq 0$ em $[a,b]$. Seja ainda $s$ o (único) zero de $f$ em $[a,b]$. Então a sucessão gerada pelo método da secante converge para $s$ sempre que os pontos iniciais $x_{-1}, x_0 \in [a,b]$ satisfizerem $f(x_{-1})f''(x_{-1}) \geq 0$ e $f(x_0)f''(x_0) \geq 0$. Mais ainda, a sucessão gerada é monótona.*

De uma forma semelhante ao que foi efectuado para o método de Newton, é possível obter um majorante do erro de aproximação com base nas distâncias entre estimativas consecutivas, tal como se mostra em seguida.

1. Consideremos a função polinomial de grau 2

$$p(x) = f(x_{n-1}) + \frac{f(x_n)-f(x_{n-1})}{x_n - x_{n-1}} \cdot (x - x_{n-1}) + \frac{f(x_{n+1})}{(x_{n+1}-x_{n-1})\cdot(x_{n+1}-x_n)} \cdot (x - x_{n-1})(x - x_n).$$

2. Como

$$p(x_{n-1}) = f(x_{n-1})$$
$$p(x_n) = f(x_n)$$
$$p(x_{n-1}) = f(x_{n-1})$$

conclui-se que

$$f(x) - p(x) \text{ tem pelo menos 3 zeros}$$
$$f'(x) - p'(x) \text{ tem pelo menos 2 zeros}$$
$$f''(x) - p''(x) \text{ tem pelo menos 1 zero}$$

e, portanto, $f''(\xi_n) = p''(\xi_n)$, para algum $\xi_n$.

3. Como $p''(x) = \frac{2f(x_{n+1})}{(x_{n+1}-x_n)(x_{n+1}-x_{n-1})}$, então

$$f(x_{n+1}) = \frac{f''(\xi_n)}{2} \cdot (x_{n+1} - x_n)(x_{n+1} - x_{n-1}). \tag{2.6.1}$$

4. Sendo $s$ tal que $f(s) = 0$, pode dizer-se que

$$f(x_{n+1}) = f'(\zeta_n) \cdot (x_{n+1} - s) \tag{2.6.2}$$

para algum $\zeta_n$.

5. Combinando as expressões (2.6.1) e (2.6.2) obtém-se a expressão

$$x_{n+1} - s = \frac{f''(\xi_n)}{2f'(\zeta_n)} \cdot (x_{n+1} - x_n)(x_{n+1} - x_{n-1}).$$

6. Considerando, como anteriormente, $M_2 = \max_{x\in[a,b]} |f''(x)|$ e $m_1 = \min_{x\in[a,b]} |f'(x)|$, e supondo-se ainda que $m_1 > 0$, resulta

$$|x_{n+1} - s| \leq \frac{M_2}{2m_1} \cdot |x_{n+1} - x_n| \cdot |x_{n+1} - x_{n-1}|,$$

pelo que o valor $\varepsilon_{n+1}$ definido por

$$\varepsilon_{n+1} = \frac{M_2}{2m_1} \cdot |x_{n+1} - x_n| \cdot |x_{n+1} - x_{n-1}|$$

é um majorante do erro de aproximação de $x_{n+1}$.

**Exemplo 2.6.1.** *Utilizar o método da secante para determinar uma aproximação, com um erro absoluto inferior a* $5 \times 10^{-6}$*, do (único) zero da função* $f(x) = 1 + x + e^x$*, que se sabe estar no intervalo* $[-2, -1]$.

***Resolução*** *(comparar com exemplo do método de Newton)*
*Condições de convergência*

$f'(x) = 1 + e^x \quad \rightarrow \quad f' > 0$
$f''(x) = e^x \quad \rightarrow \quad f'' > 0$
*O método converge desde que* $x_{-1}$ *e* $x_0$ *sejam tais que* $f(x_{-1})f''(x_{-1}) > 0$ *e* $f(x_0)f''(x_0) > 0$.
*Então, escolhendo*
$x_{-1} = -1 \quad \rightarrow \quad f(x_{-1}) = 0.3679$
$x_0 = -1.1 \quad \rightarrow \quad f(x_0) = 0.2329$
*garante-se a convergência do método.*

*Estimação do erro*
*Utilizando a estimativa do erro de aproximação atrás deduzida temos*

$$m_1 = \min_{x\in[-2,-1]} |f'(x)| = 1 + e^{-2} = 1.1353$$
$$M_2 = \max_{x\in[-2,-1]} |f''(x)| = e^{-1} = 0.3679$$
$$\frac{M_2}{2m_1} = 0.162$$

*pelo que* $\varepsilon_{n+1} = 0.162 \cdot |x_{n+1} - x_n| \cdot |x_{n+1} - x_{n-1}|$ *será um majorante do erro de* $x_{n+1}$.

*Critério de paragem*
*De acordo com a majoração do erro o critério de paragem a utilizar será* $\varepsilon_{n+1} \leq 5 \times 10^{-6}$.

*Iteração 1*

$$x_1 = \frac{x_{-1}f(x_0) - x_0 f(x_{-1})}{f(x_0) - f(x_{-1})} = -1.27249$$
$$\varepsilon_1 = 0.162 \cdot |x_1 - x_0| \cdot |x_1 - x_{-1}| = 7.6 \times 10^{-3}$$

*Como* $\varepsilon_1 \not\leq 5 \times 10^{-6}$, *devemos prosseguir as iterações.*

*Iteração 2*

$$x_2 = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} = -1.27834$$
$$\varepsilon_2 = 0.162 \cdot |x_2 - x_1| \cdot |x_2 - x_0| = 1.7 \times 10^{-4}$$

*Como* $\varepsilon_2 \not\leq 5 \times 10^{-6}$, *devemos prosseguir as iterações.*

*Iterações*
*A tabela seguinte resume a aplicação do método.*

| $n$ | $x_{n-1}$ | $x_n$ | $x_{n+1}$ | $f(x_{n+1})$ | $\varepsilon_{n+1}$ |
|---|---|---|---|---|---|
| 0 | $-1.00000$ | $-1.10000$ | $-1.27249$ | $+7.65 \times 10^{-3}$ | $7.6 \times 10^{-3}$ |
| 1 | $-1.10000$ | $-1.27249$ | $-1.27834$ | $+1.55 \times 10^{-4}$ | $1.7 \times 10^{-4}$ |
| 2 | $-1.27249$ | $-1.27834$ | $-1.27846$ | $+1.01 \times 10^{-7}$ | $1.2 \times 10^{-7}$ |

*Solução*
*A estimativa obtida é* $s \simeq -1.27846$ *(com todos os algarismos exactos).*

## 2.7 Ordem de convergência

Após a apresentação dos diferentes métodos iterativos vamos agora analisar a sua rapidez de convergência. Esta rapidez pode ser medida através da noção de ordem de convergência de um método iterativo, que se expõe em seguida.

Comecemos por considerar um método iterativo com função de recorrência $F$ e um valor $s$ que seja ponto fixo da função $F$, isto é, tal que $F(s) = s$. Suponha-se também que $F$ é uma função de classe $C^p$ numa vizinhança do ponto $s$, tal que

$$\begin{aligned} &F^{(p)}(s) \neq 0 \\ &F'(s) = \ldots = F^{(p-1)}(s) = 0 \quad (\text{se } p > 1) \end{aligned}$$

ou seja, que todas as suas derivadas até à ordem $p-1$ se anulam no ponto fixo $s$ e a derivada de ordem $p$ é não nula nesse ponto.

Suponha-se também que $\{x_n\}$ é uma sucessão, convergente para $s$, gerada por este método, isto é, tal que $x_{n+1} = F(x_n)$.

Do desenvolvimento de Taylor de $F$ em torno de $s$ obtém-se

$$\begin{aligned} F(x_n) &= F(s) + F'(s)(x_n - s) + \ldots + \frac{F^{(p-1)}(s)}{(p-1)!}(x_n - s)^{p-1} + \frac{F^{(p)}(\xi_n)}{p!}(x_n - s)^p \\ &= s + \frac{F^{(p)}(\xi_n)}{p!}(x_n - s)^p \end{aligned}$$

para algum $\xi_n$ entre $x_n$ e $s$.

Como $x_{n+1} = F(x_n)$, pode ainda escrever-se $x_{n+1} - s = \frac{F^{(p)}(\xi_n)}{p!}(x_n - s)^p$. Definindo, para cada $n$, $\Delta_n = s - x_n$ (erro em $x_n$), obtém-se

$$\Delta_{n+1} = -(-1)^p \frac{F^{(p)}(\xi_n)}{p!}\Delta_n^p.$$

Como $\xi_n \to s$, para $n$ suficientemente elevado verifica-se $\Delta_{n+1} \simeq -(-1)^p \frac{F^{(p)}(s)}{p!}\Delta_n^p$, ou seja,

$$\Delta_{n+1} \propto \Delta_n^p,$$

pelo que o erro na iteração $n+1$ é proporcional à potência de ordem $p$ do erro na iteração $n$.

Nesta situação diz-se que o método iterativo tem **convergência de ordem** $p$. Quando $p = 1$ a convergência diz-se **linear** ou de 1ª ordem. Quando $p = 2$ a convergência diz-se **quadrática** ou de 2ª ordem.

**Exemplo 2.7.1.** *Considere dois métodos iterativos $A$ e $B$, para os quais se tem $\Delta_{n+1} = 10^{-2}\Delta_n$ e $\Delta_{n+1} = \Delta_n^2$, respectivamente. Supondo que em ambos os casos se tem que $\Delta_0 = 10^{-1}$, determine a evolução do erro para as primeiras 6 iterações de aplicação de cada método.*

*Resolução*

| $n$ | $\Delta_n$ *(mét. A)* | $\Delta_n$ *(mét. B)* |
|---|---|---|
| 0 | $10^{-1}$ | $10^{-1}$ |
| 1 | $10^{-3}$ | $10^{-2}$ |
| 2 | $10^{-5}$ | $10^{-4}$ |
| 3 | $10^{-7}$ | $10^{-8}$ |
| 4 | $10^{-9}$ | $10^{-16}$ |
| 5 | $10^{-11}$ | $10^{-32}$ |
| 6 | $10^{-13}$ | $10^{-64}$ |

Como se ilustra neste exemplo, quanto maior for a ordem de convergência de um método iterativo, mais rapidamente o erro de aproximação converge para zero.

Vamos agora analisar a ordem de convergência de alguns dos métodos estudados. No caso do método iterativo simples não se impõe qualquer condição sober a nulidade da função de recorrência $F$. Trata-se portanto, no caso geral, de um método de convergência linear ou de 1ª ordem. Pode também mostrar-se que quer o método das bissecções quer o método da falsa posição são métodos de convergência linear.

Vamos agora analisar a ordem de convergência do método de Newton. Como já visto, a função de recorrência deste método é $F(x) = x - \frac{f(x)}{f'(x)}$. Derivando esta função obtém-se

$$F'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}.$$

Sendo $s$ um zero de $f$ tal que $f'(s) \neq 0$ (condição habitualmente imposta na aplicação do método de Newton), tem-se que $F'(s) = 0$. O cálculo de $F''$ fornece (verifique!)

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

Então, tem-se que $F''(s) = \frac{f''(s)}{f'(s)}$, que será, em geral, não nulo. Conclui-se assim que o método de Newton tem uma convergência quadrática, ou seja, é de um método de 2ª ordem. Explica-se deste modo o comportamento do método de Newton, que habitualmente permite obter soluções com menores erros de aproximação em menos iterações.

É também possível definir a noção de ordem de convergência de uma sucessão. Suponha-se então que $\{e_n\}$ é uma sucessão convergente para 0. Se existir uma constante $p$, maior do que zero, tal que

$$\lim_{n \to +\infty} \frac{|e_{n+1}|}{|e_n|^p} = K,$$

onde $0 < K < +\infty$, diz-se que a sucessão $\{e_n\}$ tem **ordem de convergência** $p$. Repare-se que a partir da definição de limite, se pode concluir que para valores de $n$ suficientemente elevados se tem que $|e_{n+1}| \propto |e_n|^p$, de uma forma análoga à definição de ordem de convergência

de um método iterativo. Se $p = 1$ a convergência diz-se **linear**. Se $p > 1$ a convergência diz-se **supralinear**. Se $p = 2$ a convergência diz-se **quadrática**.

Esta definição de ordem de convergência permite considerar ordens não inteiras, generalizando de algum modo a noção de ordem de convergência de um método iterativo. Pode mostrar-se que, de uma forma geral, os erros de aproximação do método da secante apresentam uma convergência de ordem $\frac{1+\sqrt{5}}{2}$ ($\approx 1.618$). Trata-se portanto de um método supralinear.

## 2.8 Localização de zeros

Cada aplicação de um método iterativo permite (mediante a satisfação de certas condições) determinar o valor de um zero de uma função. Ao se pretender calcular vários zeros, será necessário aplicar o ou os métodos iterativos quantos os zeros a determinar. Assim, antes de iniciar a aplicação de um método é necessário proceder a uma análise preliminar para estabelecer qual ou quais os zeros a determinar, bem como eventualmente a sua localização aproximada. Este processo é designado por **separação dos zeros** e consiste na determinação de intervalos disjuntos, cada um contendo um zero da função.

Note-se, por outro lado, que a verificação de condições suficientes de convergência de métodos iterativos exige certas propriedades da função e das suas derivadas, as quais deverão ser satisfeitas num dado intervalo ao qual se aplica o método, ou que contenha a estimativa inicial para a sua aplicação.

A determinação de intervalos contendo um e só zero da função e que satisfazendo condições suficientes de convergência, pode ser feita de uma forma mais ou menos automática, mas sempre recorrendo a uma ou mais das seguintes abordagens

- cálculo de valores da função,
- estudo do gráfico da função,
- análise de propriedades da função.

O esboço do gráfico da função $f$ permite em muitas situações determinar de uma forma visual intervalos disjuntos, cada um dos quais contendo apenas um zero de $f$.

O gráfico da função pode ser obtido utilizando meios computacionais, através de um estudo analítico das propriedades de $f$, ou mesmos ambos os processos de uma forma conjugada.

Por vezes, em vez de se analisar directamente o gráfico da função $f$, a equação $f(x) = 0$ é reescrita na forma equivalente $g(x) = h(x)$, onde $g$ e $h$ são funções cujos gráficos são mais simples de estudar do que o gráfico da função $f$. O estudo dos zeros de $f$ fica então reduzido à análise das intersecções dos gráficos de $g$ e $h$.

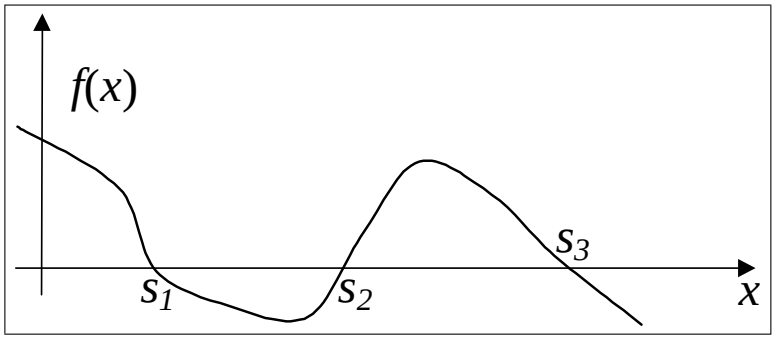



Figura 2.12: Gráfico de $f$ mostrando a localização dos zeros

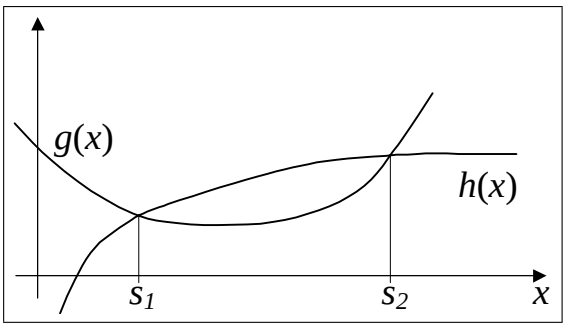



Figura 2.13: Soluções de $g(x) = h(x)$.

Os métodos analíticos de separação dos zeros de $f$ baseiam-se principalmente na determinação de intervalos de monotonia de $f$ e no cálculo e ordenação dos números de Rolle de $f$. Relembremos que se designam por **números de Rolle** de uma função $f : D \to \mathbb{R}$ os pontos fronteira de $D$ e os zeros da função $f'$.

Os dois teoremas apresentados abaixo constituem a justificação teórica dos métodos analíticos de separação de zeros.

**Teorema 2.8.1.** *Se $f$ é estritamente monótona em $[a, b]$, $f$ tem no máximo um zero em $[a, b]$.*

**Teorema 2.8.2.** *Se $f$ é diferenciável, entre dois números de Rolle consecutivos existe quando muito um zero de $f$.*

## 2.9 Raízes de polinómios

A determinação directa de raízes de polinómios (ou seja dos zeros das funções polinomiais correspondentes) só é possível de efectuar no caso geral para polinómios de grau não superior a 4. Assim, a determinação de raízes de polinómios de grau superior a 4 (ou até mesmo de grau 3 ou 4) terá na maioria das situações de ser efectuada por métodos iterativos.

Os métodos estudados anteriormente para a determinação de zeros de funções podem também ser

utilizados na determinação de raízes reais de polinómios. Contudo, é sabido que os polinómios (mesmo de coeficientes reais) podem ter raízes complexas.

Nesta secção apresentam-se resultados sobre a localização de raízes de polinómios e métodos especialmente dedicados à determinação das suas raízes, sejam estas reais ou complexas. O objecto de estudo nesta secção será um polinómio de grau $n$ com todos os coeficientes reais, ou seja,

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

onde $a_i \in \mathbb{R},\ i = 0, \ldots, n$, e $a_n \neq 0$.

Comecemos por relembrar o seguinte resultado sobre as raízes de um polinómio.

**Teorema 2.9.1.** *Um polinómio $p$ de grau $n$ tem exactamente $n$ raízes (contando com a multiplicidade). Estas raízes podem ser reais ou complexas. Se os coeficientes do polinómio forem todos reais as raízes complexas surgem em pares conjugados.*

Sendo $s \in \mathbb{C}$ tem-se que $p(x) = (x-s)q(x) + r$, onde $q$ é um polinómio de grau $n-1$ (designado quociente) e $r$ designa-se por resto. O quociente $q(x) = b_{n-1}x^{n-1} + \ldots + b_1 x + b_0$ e o resto $r$ podem ser obtidos por divisão polinomial ou pela **regra de Ruffini**. Os cálculos desta regra são habitualmente dispostos na forma de uma tabela, tal como se apresenta em seguida.

| | | | | | | |
|---|---|---|---|---|---|---|
| | $a_n$ | $a_{n-1}$ | $\ldots$ | $a_1$ | | $a_0$ |
| $s$ | | $sb_{n-1}$ | $\ldots$ | $sb_1$ | | $sb_0$ |
| | $b_{n-1}$ | $b_{n-2}$ | $\ldots$ | $b_0$ | $\|\|$ | $r$ |

Os coeficientes do quociente e o valor do resto são determinados de acordo com as expressões $b_{n-1} = a_n$, $b_i = a_{i+1} + sb_{i+1}$, para $i = n-2, \ldots, 0$, e $r = a_0 + sb_0$. Uma consequência directa da regra de Rufinni é o facto de o resto da divisão de $p(x)$ por $x - s$ ser $p(s)$. Pode assim obter-se o seguinte resultado.

**Teorema 2.9.2.** *Se o resto da divisão de $p(x)$ por $x - s$ for o polinómio nulo então $s$ é raiz de $p(x)$. Mais ainda, as restantes raízes de $p(x)$ são as raízes do polinómio quociente.*

A aplicação sucessiva da regra de Ruffini permite ainda demonstrar o seguinte resultado.

**Teorema 2.9.3.** *Sejam $r_1, r_2, \ldots, r_n$ as $n$ raízes do polinómio de grau $n$, $p(x) = a_n x^n + \cdots + a_1 x + a_0$, contando com eventuais multiplicidades. Então, $p(x)$ pode ser escrito como*

$$p(x) = a_n(x - r_1)(x - r_2)\cdots(x - r_n).$$

Consideremos agora a divisão do polinómio $p(x)$, de grau $n \geq 2$, por um polinómio de grau 2 da forma $x^2 - \alpha x - \beta$. Facilmente se pode concluir a seguinte igualdade

$$p(x) = (x^2 - \alpha x - \beta)q(x) + (rx + s),$$

onde $q(x) = b_{n-2}x^{n-2} + b_{n-3}x^{n-3} + \ldots + b_1 x + b_0$ é um polinómio de grau $n-2$ designado por quociente, e o polinómio $rx + s$ é designado por resto.

Os coeficientes dos polinómios quociente e resto podem ser obtidos de uma forma expedita dispondo os cálculos como se mostra na tabela

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $a_n$ | $a_{n-1}$ | $a_{n-2}$ | $\ldots$ | $a_2$ | | $a_1$ | $a_0$ |
| $\beta$ | | | $\beta b_{n-2}$ | $\ldots$ | $\beta b_2$ | | $\beta b_1$ | $\beta b_0$ |
| $\alpha$ | | $\alpha b_{n-2}$ | $\alpha b_{n-3}$ | $\ldots$ | $\alpha b_1$ | | $\alpha b_0$ | |
| | $b_{n-2}$ | $b_{n-3}$ | $b_{n-4}$ | $\ldots$ | $b_0$ | $\|\|$ | $r$ | $s$ |

onde se verificam as relações

$$
\begin{aligned}
&b_{n-2} = a_n,\\
&b_{n-3} = a_{n-1} + \alpha b_{n-2},\\
&b_i = a_{i+2} + \alpha b_{i+1} + \beta b_{i+2}, \quad \text{para } i = n-4, n-3, \ldots, 0,\\
&r = a_1 + \alpha b_0 + \beta b_1, \text{ e}\\
&s = a_0 + \beta b_0.
\end{aligned}
$$

O seguinte resultado é uma consequência da divisão polinomial indicada acima.

**Teorema 2.9.4.** *Se o resto da divisão de $p(x) = a_n x^n + \cdots + a_1 x + a_0$ (onde $a_n \neq 0$) por $x^2 - \alpha x - \beta$ for o polinómio nulo então, as raízes de $x^2 - \alpha x - \beta$ são também raízes de $p(x)$. As restantes raízes de $p(x)$ são as raízes do polinómio quociente.*

O resultado seguinte fornece uma expressão geral para as raízes racionais de polinómios de coeficientes inteiros.

**Teorema 2.9.5.** *Seja $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, com $a_i \in \mathbb{Z}$, $a_n \neq 0$ e $a_0 \neq 0$. Então, toda a raiz racional de $p(x)$ é da forma*

$$\pm \frac{d_0}{d_n}$$

*onde $d_0$ é um divisor de $a_0$ e $d_n$ é um divisor de $a_n$.*

Tal como no caso geral de funções com múltiplos zeros, é por vezes importante obter informação sobre a localização das raízes de um polinómio sem as determinar. Enunciam-se em seguida alguns resultados que podem ser utilizados para obter tal informação. Refira-se que existe um grande número de resultados sobre a localização de raízes de polinómios, optando-se por apresentar aqui alguns dos considerados de aplicação mais imediata.

**Teorema 2.9.6** (Regra dos sinais de Descartes I)**.** *O número de raízes reais positivas de um polinómio $p(x)$ é igual, ou menor pela diferença de um número par, ao número de mudanças de sinal dos seus coeficientes não nulos.*

É imediato verificar que as raízes do polinómio $m(x) = p(-x)$ são simétricas das raízes de $p(x)$, pelo que facilmente se obtém o seguinte corolário.

**Corolário** (Regra dos sinais de Descartes II)**.** *O número de raízes reais negativas de um polinómio $p(x)$ é igual, ou menor pela diferença de um número par, ao número de mudanças de sinal dos coeficientes não nulos de $p(-x)$.*

**Teorema 2.9.7.** *Seja $p(x)$ um polinómio cujos coeficientes satisfazem*

$$a_n > 0,\ a_{n-1} \geq 0,\ , \ldots ,\ a_{m+1} \geq 0,\ a_m < 0$$

*ou seja, $a_m$ é o primeiro coeficiente negativo de $\{a_n, a_{n-1}, \ldots, a_1, a_0\}$. Então os zeros reais de $p$ são majorados por*

$$1 + \left[ \max_{a_k<0} \left| \frac{a_k}{a_n} \right| \right]^{\frac{1}{n-m}}.$$

**Teorema 2.9.8.** *Todos os zeros do polinómio $p(x)$ situam-se no interior do círculo (no plano complexo) centrado na origem e de raio*

$$1 + \max_{0\leq k\leq n-1} \left| \frac{a_k}{a_n} \right|.$$

Os teoremas 2.9.2 e 2.9.4 permitem definir uma estratégia sistemática para a determinação de todas as raízes de um polinómio $p(x)$, de grau $n$. Esta estratégia consiste em obter uma raiz $s$ (ou um par de raízes) de cada vez, por aplicação de um método iterativo. Após a obtenção de uma raiz, o polinómio considerado é dividido por $x - s$ (ou por $x^2 - \alpha x - \beta$ no caso de um par de raízes), aplicando-se em seguida novamente um método iterativo mas agora ao polinómio quociente e assim sucessivamente até se chegar a um polinómio cujas raízes se determinem por um método directo.

É importante ter em atenção a propagação de erros de arredondamento, os quais de uma forma geral vão aumentando à medida que se vão obtendo novas raízes e calculando os polinómios quociente. Por forma a diminuir estes erros, após a obtenção de todas as raízes, é por vezes utilizado um procedimento de refinamento das raízes, que consiste em aplicar um método iterativo partindo das estimativas das soluções determinadas anteriormente mas utilizando directamente o polinómio original $p(x)$.

Em seguida serão apresentados dois métodos iterativos para a obtenção das raízes de um polinómio: o método de Newton e o método de Lin.

A aplicação do método de Newton é em tudo ao caso geral apresentado para a determinação de zeros de funções. Tal como então, a expressão de recorrência é

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k)}.$$

A principal diferença é que agora se pode escolher a estimativa inicial $x_0 \in \mathbb{C}$, podendo determinar-se directamente uma raiz complexa. No entanto isto obriga a efectuar operações em aritmética complexa.

De uma forma geral, $x_0$ é escolhido como uma das soluções de

$$a_n x^2 + a_{n-1}x + a_{n-2} = 0$$

ou seja, considerando apenas os três termos de ordem mais elevada de $p(x)$.

**Exemplo 2.9.1.** *Determinar todas as raízes do polinómio* $p(x) = x^4 + 2x^3 + 10x^2 + 24x + 80$ *aplicando o método de Newton.*

***Resolução***

*Derivada e fórmula de recorrência*

$p'(x) = 4x^3 + 6x^2 + 20x + 24$

$x_{k+1} = x_k - \dfrac{p(x_k)}{p'(x_k)}$

*Determinação do ponto inicial*

$x^2 + 2x + 10 = 0 \quad \Rightarrow \quad x = -1 \pm 3j$

$x_0 = -1 + 3j$

*Obtenção do primeiro par de raízes*

*Iteração 1:*

$p(x_0) = x_0^4 + 2x_0^3 + 10x_0^2 + 24x_0 + 80 = 56 + 72j$

$p'(x_0) = 4x_0^3 + 6x_0^2 + 20x_0 + 24 = 60 - 48j$

$x_1 = x_0 - \frac{p(x_0)}{p'(x_0)} = -0.98 + 1.81j$

*Iteração 2:*

$p(x_1) = x_1^4 + 2x_1^3 + 10x_1^2 + 24x_1 + 80 = 43.45 + 23.00j$

$p'(x_1) = 4x_1^3 + 6x_1^2 + 20x_1 + 24 = 25.40 + 12.07j$

$x_2 = x_1 - \frac{p(x_1)}{p'(x_1)} = -2.73 + 1.74j$

*Continuando a aplicar o processo iterativo até que* $x_{n+1}$ *esteja suficiente próximo de* $x_n$*, obtêm-se os resultados apresentados na tabela seguinte.*

| $n$ | $x_n$ | $p(x_n)$ | $p'(x_n)$ |
|---|---|---|---|
| 0 | $-1.00 + 3.00j$ | $56.00 + 72.00j$ | $60.00 - 48.00j$ |
| 1 | $-0.98 + 1.81j$ | $43.35 + 23.00j$ | $25.40 + 12.07j$ |
| 2 | $-2.73 + 1.74j$ | $-2.57 - 69.73j$ | $13.53 + 111.88j$ |
| 3 | $-2.11 + 1.79j$ | $8.26 - 15.13j$ | $32.70 + 63.12j$ |
| 4 | $-1.97 + 1.99j$ | $1.84 + 0.91j$ | $47.11 + 54.20j$ |
| 5 | $-2.00 + 2.00j$ | $-0.02 - 0.02j$ | $48.01 + 56.03j$ |
| 6 | $-2.00 + 2.00j$ | $\approx 0$ | |

*A raiz obtida será então* $r_1 = -2+2j$. *Obtém-se então imediatamente a raiz* $r_2 = r_1^* = -2-2j$

*Determinação das restantes raízes*

*Fazendo* $m(x) = (x-r_1)(x-r_2) = (x+2-2j)(x+2+2j) = (x^2+4x+8)$ *e dividindo* $p(x)$ *por* $m(x)$ *obtém-se o polinómio* $q(x) = x^2-2x+10$. *As raízes deste polinómio, obtidas directamente, são* $1 \pm 3j$.

*Resultado*

*As raízes de* $p(x)$ *são* $-2 \pm 2j$ *e* $1 \pm 3j$.

O **método de Lin** permite obter raízes complexas de um polinómio efectuando apenas operações em aritmética real. Este método consiste em construir sucessões $\{p_i\}$ e $\{q_i\}$ convergentes para $\bar{p}$ e $\bar{q}$ de forma a que as raízes de $x^2 + \bar{p}x + \bar{q}$ sejam raízes de $p(x)$, ou seja, que o polinómio $x^2 + \bar{p}x + \bar{q}$ seja divisor de $p(x)$.

Em cada iteração é realizada a divisão polinomial

$$\frac{p(x)}{x^2 + p_i x + q_i} = q(x) + \frac{rx + s}{x^2 + p_i x + q_i}.$$

Esta divisão é parada após a obtenção do quociente $q(x)$, fazendo-se então as substituições $p_i \to p_{i+1}$ e $q_i \to q_{i+1}$, sendo então determinados $p_{i+1}$ e $q_{i+1}$ de modo a anular o resto $rx+s$. Este processo é repetido até termos consecutivos das sucessões $\{p_i\}$ e $\{q_i\}$ se tornarem suficientemente próximos.

A aplicação deste método é facilitada dispondo os cálculos da divisão polinomial da seguinte forma

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $a_n$ | $a_{n-1}$ | $\dots$ | $a_2$ | | $a_1$ | $a_0$ |
| $-q_i$ | | | $\dots$ | $-q_i b_2$ | | $-q_i b_1$ | $\boxed{-q_{i+1} b_0}$ |
| $-p_i$ | | $-p_i b_{n-2}$ | $\dots$ | $-p_i b_1$ | | $\boxed{-p_{i+1} b_0}$ | |
| | $b_{n-2}$ | $b_{n-3}$ | $\dots$ | $b_0$ | $\|\|$ | 0 | 0 |

Pode assim concluir-se que a determinação de $p_{i+1}$ é $q_{i+1}$ é feita resolvendo

$$\begin{cases} a_1 - q_i b_1 - p_{i+1} b_0 = 0 \\ a_0 - q_{i+1} b_0 = 0 \end{cases}$$

**Exemplo 2.9.2.** *Determinar os zeros de* $p(x) = x^4 - 6x^3 + 18x^2 - 24x + 16$ *pelo método de Lin.*

***Resolução***

*Inicialmente escolheu-se* $p_0 = 1$ *e* $q_0 = 1$.

*No quadro seguinte apresenta-se a divisão polinomial até obter o quociente.*

| | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | $-6$ | 18 | | $-24$ | 16 |
| $-1$ | | | $-1$ | | 7 | $-24q_1$ |
| $-1$ | | $-1$ | 7 | | $-24p_1$ | |
| | 1 | $-7$ | 24 | $\|\|$ | 0 | 0 |

*O sistema de equações a resolver para anular o resto será*

$$\begin{cases} -24 + 7 - 24p_1 = 0 \\ 16 - 24q_1 = 0 \end{cases}$$

*resultando* $p_1 = -0.70833$ *e* $q_1 = 0.66667$ .

*As operações da segunda iteração do método encontram-se na tabela abaixo.*

| | | | | | | |
|---|---|---|---|---|---|---|
| | 1 | $-6$ | 18 | | $-24$ | 16 |
| $-0.66667$ | | | $-0.66667$ | | 3.52778 | $-13.58507q_2$ |
| 0.70833 | | 0.70833 | $-3.74826$ | | $-13.58507p_2$ | |
| | 1 | $-5.29167$ | 13.58507 | $\|\|$ | 0 | 0 |

*Agora o sistema de equações a resolver será*

$$\begin{cases} -24 + 3.52778 - 13.58507p_2 = 0 \\ 16 - 13.58507q_2 = 0 \end{cases}$$

*resultando* $p_2 = -1.50696$ *e* $q_2 = 1.17776$.

*... após mais algumas iterações conclui-se que* $p_i \to -2$ *e* $q_i \to 2$. *Assim, conclui-se que o* $p(x)$ *é divisível por* $x^2 - 2x + 2$. *As raízes de* $x^2 - 2x + 2$, *que são* $1 \pm j$, *são também raízes de* $p(x)$.

*Finalmente, dividindo* $p(x)$ *por* $x^2 - 2x + 2$ *obtém-se o polinómio* $x^2 - 4x + 8$, *cujas raízes são* $2 \pm 2j$. *Conclui-se assim que as raízes de* $p(x)$ *são* $1 \pm j$ *e* $2 \pm 2j$.

# Capítulo 3

# Normas de vectores e matrizes

## 3.1 Introdução

Neste capítulo apresentam-se alguns resultados sobre normas em espaços vectoriais que irão ser necessários no tratamento de problemas de sistemas de equações.

## 3.2 Normas de vectores

Comecemos por relembrar que uma norma num espaço vectorial real $V$ é uma função que associa a cada elemento $x \in V$ um número real, representado por $\|x\|$, que verifica as seguintes condições

1. $\|x\| \geq 0 \quad \forall x \in V \text{ e } \|x\| = 0 \Rightarrow x = 0,$
2. $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall \alpha \in \mathbb{R}, \forall x \in V,$
3. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in V.$

A noção de norma está associada ao tamanho de um vector. Habitualmente, quando $V = \mathbb{R}^n$, é utilizada a norma euclidiana que se define por

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

para todo o vector $x = (x_1, x_2, \ldots, x_n)$ de $\mathbb{R}^n$. No entanto, podem definir-se outras normas, que sejam mais úteis em certas situações. Alguns exemplos de normas em $\mathbb{R}^n$, onde $x = (x_1, x_2, \ldots, x_n)$, são

$\rightarrow$ norma 1 $\quad \displaystyle\sum_{i=1}^{n} |x_i|$

$\rightarrow$ norma $\infty$ $\quad \displaystyle\max_{1 \leq i \leq n} |x_i|$

$\rightarrow$ norma $p$ $\quad \left( \displaystyle\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \quad (\text{com } p \geq 1)$

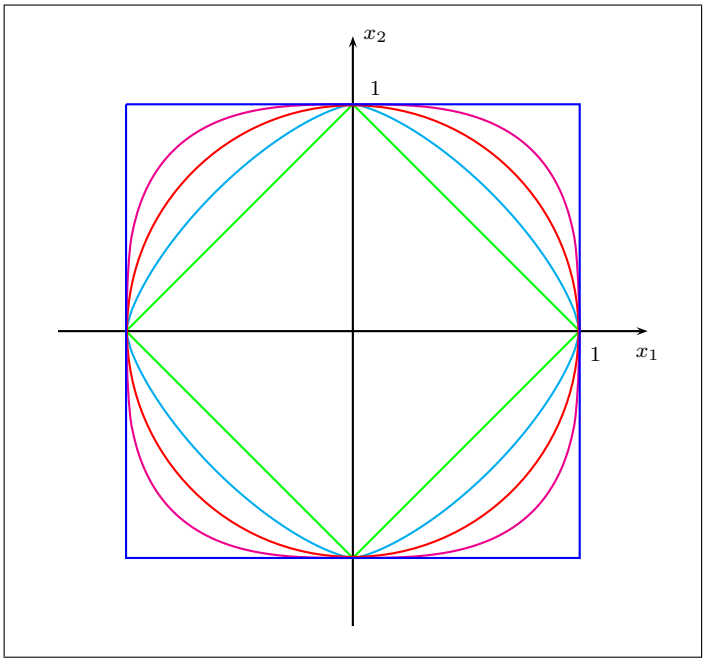



Figura 3.1: Visualização de diferentes normas em $\mathbb{R}^2$. De dentro para fora aparecem as linhas $\|x\|_1 = 1$, $\|x\|_{1.4} = 1$, $\|x\|_2 = 1$, $\|x\|_3 = 1$ e $\|x\|_\infty = 1$, respectivamente.

Embora diferentes, todas as normas em $\mathbb{R}^n$ são de alguma forma equivalentes, no sentido apresentado no seguinte teorema.

**Teorema 3.2.1.**
*Sejam $\|\cdot\|_\alpha$ e $\|\cdot\|_\beta$ duas normas definidas em $\mathbb{R}^n$. Então existem constantes $k_1, k_2 > 0$ tais que*

$$k_1 \|x\|_\alpha \leq \|x\|_\beta \leq k_2 \|x\|_\alpha, \quad \forall x \in \mathbb{R}^n.$$

**Exemplo 3.2.1.** *Consideremos as normas $\|\cdot\|_2$ e $\|\cdot\|_\infty$, definidas em $\mathbb{R}^n$. Das suas definições temos*

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad e \qquad \|x\|_\infty = \max_{1\leq i\leq n} |x_i| = |x_{i_0}|$$

*para algum $1 \leq i_0 \leq n$. Destas expressões conclui-se facilmente que*

$$\|x\|_\infty = |x_{i_0}| = \sqrt{x_{i_0}^2} \leq \sqrt{\sum_{i=1}^{n} x_i^2}$$

*e também que*

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \leq \sqrt{n \cdot x_{i_0}^2} = \sqrt{n} \cdot |x_{i_0}|$$

*resultando finalmente*

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty.$$

## 3.3 Normas de matrizes

O espaço das matrizes quadradas $\mathbb{R}^{n\times n}$ é em si mesmo um espaço vectorial real (de dimensão $n\times n$) no qual se podem obviamente definir diversas normas. No entanto, têm particular interesse normas que resultem da consideração dos elementos deste espaço como sendo operadores lineares de $\mathbb{R}^n$ em $\mathbb{R}^n$.

Seja então $\|\cdot\|$ uma qualquer norma definida em $\mathbb{R}^n$. É possível definir uma norma em $\mathbb{R}^{n\times n}$, que por simplicidade se representa também por $\|\cdot\|$, pela expressão

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

para qualquer $A \in \mathbb{R}^{n\times n}$. Esta norma em $\mathbb{R}^{n\times n}$ designa-se por **norma induzida** pela norma definida em $\mathbb{R}^n$. Da definição de norma induzida resulta imediatamente, para qualquer $A \in \mathbb{R}^{n\times n}$,

1. $\forall x \in \mathbb{R}^n \quad \|Ax\| \leq \|A\|\,\|x\|$,
2. $\exists x \in \mathbb{R}^n \setminus \{0\} \quad \|Ax\| = \|A\|\,\|x\|$,
3. $\|A\| = \max_{\|x\|=1} \|Ax\|$.

Algumas propriedades importantes de qualquer norma induzida são ainda

1. $\|AB\| \leq \|A\|\,\|B\|, \quad \forall A, B \in \mathbb{R}^{n\times n}$ e
2. $\|I\| = 1$ (onde $I$ é a matriz identidade).

É de referir que diferentes normas em $\mathbb{R}^n$ conduzem a diferentes normas induzidas. Por exemplo, teremos

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$$
$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$
$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$$

A consideração de diversas normas justifica-se não só por haver situações em que interessa utilizar uma dada norma em particular como também pelo facto das normas induzidas de matrizes não apresentarem todas as mesmas facilidades de cálculo. Como mostram os dois resultados seguintes, as normas induzidas $\|\cdot\|_1$ e $\|\cdot\|_\infty$ são de cálculo extremamente simples.

**Teorema 3.3.1.** *Seja $A \in \mathbb{R}^{n\times n}$ de elemento genérico $a_{ij}$. Então verifica-se*

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|,$$

*ou seja, o máximo das somas por colunas dos valores absolutos dos elementos de $A$.*

*Demonstração.* Sendo $x \in \mathbb{R}^n$ qualquer tem-se

$$\|Ax\|_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}x_j| \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}x_j| = \sum_{j=1}^{n} \left( |x_j| \sum_{i=1}^{n} |a_{ij}| \right)$$
$$\leq \sum_{j=1}^{n} |x_j| \cdot \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| = \|x\|_1 \cdot \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Seja agora $j_0$ tal que $\sum_{i=1}^{n} |a_{ij_0}| = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$, ou seja, o (ou um) índice de coluna que corresponda à maior soma de valores absolutos.
Seja também $\bar{x}$ o vector de $\mathbb{R}^n$ tal que

$$\bar{x}_j = \begin{cases} 1 & \text{se } j = j_0 \\ 0 & \text{se } j \neq j_0 \end{cases}$$

Então, $\|\bar{x}\|_1 = 1$ e

$$\|A\bar{x}\|_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{ij}\bar{x}_j| = \sum_{i=1}^{n} |a_{ij_0}| = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| = \|\bar{x}\|_1 \cdot \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|.$$

Desta forma, conclui-se que $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$. □

**Teorema 3.3.2.** *Seja $A \in \mathbb{R}^{n\times n}$ de elemento genérico $a_{ij}$. Então verifica-se*

$$\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|,$$

*ou seja, o máximo das somas por linhas dos valores absolutos dos elementos de A.*

*Demonstração.* Sendo $x \in \mathbb{R}^n$ qualquer tem-se

$$\|Ax\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1}^{n} a_{ij}x_j| \leq \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|\,|x_j| \leq \max_{i=1,\ldots,n} \left( \max_{1\leq j\leq n} |x_j| \cdot \sum_{j=1}^{n} |a_{ij}| \right)$$
$$= \max_{1\leq j\leq n} |x_j| \cdot \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| = \|x\|_\infty \cdot \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Seja agora $i_0$ tal que $\sum_{j=1}^{n} |a_{i_0j}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$.
Seja também $\bar{x}$ tal que

$$\bar{x}_j = \begin{cases} 1 & \text{se } a_{i_0j} \geq 0 \\ -1 & \text{se } a_{i_0j} < 0 \end{cases}$$

Então $\|\bar{x}\|_\infty = 1$ e $a_{i_0j}\bar{x}_j = |a_{i_0j}|$. Logo

$$\|A\bar{x}\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1}^{n} a_{ij}\bar{x}_j| \geq |\sum_{j=1}^{n} a_{i_0j}\bar{x}_j| = \sum_{j=1}^{n} |a_{i_0j}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| = \|\bar{x}\|_\infty \cdot \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Desta forma, conclui-se que $\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$. □

**Exemplo 3.3.1.** *Seja*

$$A = \begin{bmatrix} -2 & 0 & 1 & 6 \\ -3 & -1 & 2 & 4 \\ 2 & 1 & -1 & 1 \\ 3 & -2 & 2 & 5 \end{bmatrix}$$

*então*

$$\|A\|_1 = \max\{10, 4, 6, 16\} = 16, \ \ e$$
$$\|A\|_\infty = \max\{9, 10, 5, 12\} = 12.$$

A norma 1 e a norma $\infty$ são efectivamente as de cálculo mais simples. A norma induzida $\|\cdot\|_2$ é já de cálculo mais trabalhoso, verificando-se que

$$\|A\|_2 = \sqrt{\rho(A^{\mathrm{T}}A)}$$

onde $\rho$ é o raio espectral. O **raio espectral** de uma matriz quadrada define-se como sendo o máximo dos módulos dos valores próprios da matriz. Assim, sendo $C \in \mathbb{R}^{n\times n}$ o seu raio espectral $\rho(C)$ é dado por

$$\rho(C) = \max_{1\leq i\leq n} |\lambda_i|,$$

onde $\lambda_1, \ldots, \lambda_n$ são os valores próprios de $C$. De forma conclui-se que o cálculo da norma induzida $\|\cdot\|_2$ exige a determinação de valores próprios.

O seguinte teorema estabelece uma relação entre o raio espectral de uma matriz e as normas induzidas dessa matriz, permitindo considerar o raio espectral de uma matriz como o ínfimo das normas induzidas dessa mesma matriz.

**Teorema 3.3.3.** *Para qualquer norma induzida $\|\cdot\|$ e para qualquer $A \in \mathbb{R}^{n\times n}$ verifica-se que*

$$\rho(A) \leq \|A\|.$$

*Dada uma matriz $A \in \mathbb{R}^{n\times n}$ e um $\varepsilon > 0$, existe uma norma induzida $\|\cdot\|$ tal que*

$$\|A\| \leq \rho(A) + \varepsilon.$$

# Capítulo 4

# Sistemas de Equações Não Lineares

## 4.1 Introdução

Neste capítulo iremos abordar o problema de resolução numérica de sistemas de equações não lineares. Um sistema de $n$ equações nas $n$ variáveis $x_1, x_2, \ldots, x_n$ pode ser escrito na forma

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{cases}$$

onde $f_1, f_2, \ldots, f_n$ são funções de $\mathbb{R}^n$ em $\mathbb{R}$.

Utilizando uma notação mais compacta, podemos definir o vector $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ e a função $F : \mathbb{R}^n \to \mathbb{R}^n$ de acordo com

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) \end{bmatrix}$$

podendo agora o sistema de equações ser escrito como

$$F(x) = 0.$$

**Exemplo 4.1.1.** *O sistema de equações*

$$\begin{cases} 4x_1x_2^2 - 2x_1^2x_2 + 2 = 0 \\ 2x_1 - 4x_2 + \sqrt{x_1x_2} - 3 = 0 \end{cases}$$

*pode ser reescrito na forma* $F(x) = 0$ *definindo a função*

$$F : \mathbb{R}^n \to \mathbb{R}^n$$
$$x \mapsto \begin{bmatrix} 4x_1x_2^2 - 2x_1^2x_2 + 2 \\ 2x_1 - 4x_2 + \sqrt{x_1x_2} - 3 \end{bmatrix}$$

Na quase totalidade das situações não existem métodos directos para a resolução de sistemas de equações não lineares, sendo necessário recorrer a métodos iterativos. Nas secções seguintes iremos estudar dois métodos iterativos para a resolução de sistemas de equações não lineares. Trata-se em ambos os casos de extensões de métodos já estudados para a resolução de uma equação não linear. Refira-se também que por vezes é possível por manipulação algébrica das diferentes de equações proceder à eliminação de variáveis reduzindo o número de equações a resolver e eventualmente ficando apenas com uma equação não linear para resolver. Este procedimento simplifica o problema e deve ser realizado sempre que possível.

## 4.2 Método iterativo simples (iteração de ponto fixo)

Analogamente ao caso unidimensional, o método iterativo simples baseia-se na possibilidade de escrever o sistema de equações $F(x) = 0$ num outro equivalente da forma

$$x = G(x)$$

onde $G : \mathbb{R}^n \to \mathbb{R}^n$, ou seja,

$$\begin{cases} x_1 = g_1(x_1, x_2, \ldots, x_n) \\ x_2 = g_2(x_1, x_2, \ldots, x_n) \\ \qquad\qquad \vdots \\ x_n = g_n(x_1, x_2, \ldots, x_n) \end{cases}$$

onde $g_1, g_2, \ldots, g_n$ são as componentes de $G$.

O método interativo simples consiste então em gerar uma sucessão de pontos em $\mathbb{R}^n$ por intermédio da relação de recorrência

$$x_{(k+1)} = G(x_{(k)}), \qquad k = 0, 1, \ldots,$$

a partir de um ponto inicial $x_{(0)}$. Pretende-se que esta sucessão de pontos em $\mathbb{R}^n$ convirja para um ponto fixo $s$ da função $G$, isto é, tal que $s = G(s)$ que será portanto solução do sistema original, ou seja, tal que $F(s) = 0$.

Este método é totalmente análogo ao método iterativo simples já estudado, sendo agora necessário calcular em cada iteração as novas estimativas de todas as variáveis.

**Exemplo 4.2.1.** *Reescrevendo o sistema*

$$\begin{cases} 4x_1 - \ln(x_1x_2) - 8 = 0 \\ 2x_1 - 4x_2 + \sqrt{x_1x_2} - 3 = 0 \end{cases}$$

*na forma equivalente*

$$\begin{cases} x_1 = \dfrac{\ln(x_1x_2) + 8}{4} \\ x_2 = \dfrac{2x_1 + \sqrt{x_1x_2} - 3}{4} \end{cases}$$

*obtém-se a seguinte expressão de recorrência*

$$\begin{cases} x_{1,(k+1)} = \dfrac{\ln\left(x_{1,(k)}x_{2,(k)}\right) + 8}{4} \\ \\ x_{2,(k+1)} = \dfrac{2x_{1,(k)} + \sqrt{x_{1,(k)}x_{2,(k)}} - 3}{4} \end{cases}$$

*Partindo da estimativa inicial* $x_{1,(0)} = 1.5$, $x_{2,(0)} = 1$, *temos na primeira iteração*

$$x_{1,(1)} = \frac{\ln(x_{1,(0)}x_{2,(0)}) + 8}{4} = 2.1.0137$$
$$x_{2,(1)} = \frac{2x_{1,(0)} + \sqrt{x_{1,(0)}x_{2,(0)}} - 3}{4} = 0.30619$$

*e continuando a aplicar o método, obtêm-se os seguintes resultados*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $g_1(x_{1,(k)}, x_{2,(k)})$ | $g_2(x_{1,(k)}, x_{2,(k)})$ |
|---|---|---|---|---|
| 0 | 1.50000 | 1.00000 | 2.10137 | 0.30619 |
| 1 | 2.10137 | 0.30619 | 1.88976 | 0.50122 |
| 2 | 1.88976 | 0.50122 | 1.98643 | 0.43819 |
| 3 | 1.98643 | 0.43819 | 1.96531 | 0.47646 |
| 4 | 1.96531 | 0.47646 | 1.98357 | 0.47457 |
| 5 | 1.98357 | 0.47457 | 1.98489 | 0.48434 |
| 6 | 1.98489 | 0.48434 | 1.99015 | 0.48757 |
| 7 | 1.99015 | 0.48757 | 1.99247 | 0.49134 |
| 8 | 1.99247 | 0.49134 | 1.99469 | 0.49359 |
| 9 | 1.99469 | 0.49359 | 1.99611 | 0.49541 |
| 10 | 1.99611 | 0.49541 | 1.99721 | 0.49666 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Como se passa com todos os métodos iterativos, é importante analisar a convergência do método iterativo simples. O seguinte resultado fornece condições suficientes para a convergência do método iterativo simples. É de notar a semelhança entre estas condições e as apresentadas para o caso unidimensional.

**Teorema 4.2.1.** *Seja* $D \subset \mathbb{R}^n$ *um conjunto fechado e convexo. Seja* $G : D \to \mathbb{R}^n$ *de classe* $C^1$ *e seja* $\|\cdot\|$ *uma norma em* $\mathbb{R}^n$. *Se*

*i)* $\|J_G(x)\| \leq L < 1 \qquad \forall x \in D$

*ii)* $G(D) \subset D$

*então*

*i) existe um e só um* $z \in D$ *tal que* $z = G(z)$

*ii) o método iterativo simples converge para* $z$*, qualquer que seja* $x^{(0)} \in D$

*iii) verifica-se que*

$$\|z - x_{(k+1)}\| \leq \frac{L}{1-L}\|x_{(k+1)} - x_{(k)}\|$$

O exemplo seguinte ilustra a aplicação deste teorema na resolução de um sistema de equações não lineares.

**Exemplo 4.2.2.** *Utilizando o método iterativo simples, determinar a solução do sistema de equações*

$$\begin{cases} 4x_1 - \cos(x_1 + x_2) = 4 \\ 3x_2 - \sin(x_1 + x_2) = 6 \end{cases}$$

*com um erro máximo, na norma* 1*, de* $10^{-5}$*.*

***Resolução***

*Obtenção da função de recorrência*

*Este sistema pode ser reescrito na forma*

$$\begin{cases} x_1 = 1 + \frac{1}{4}\cos(x_1 + x_2) \\ x_2 = 2 + \frac{1}{3}\sin(x_1 + x_2) \end{cases} \tag{4.2.1}$$

*ou ainda,*

$$G(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} 1 + \frac{1}{4}\cos(x_1 + x_2) \\ 2 + \frac{1}{3}\sin(x_1 + x_2) \end{bmatrix}$$

*Condições de convergência*

$$J_G(x) = \begin{bmatrix} -\frac{1}{4}\sin(x_1 + x_2) & -\frac{1}{4}\sin(x_1 + x_2) \\ \frac{1}{3}\cos(x_1 + x_2) & \frac{1}{3}\cos(x_1 + x_2) \end{bmatrix}$$

*Então,*

$$\begin{aligned} \|J_G(x)\|_1 &= \max\{\tfrac{1}{4}|\sin(x_1 + x_2)| + \tfrac{1}{3}|\cos(x_1 + x_2)|, \tfrac{1}{4}|\sin(x_1 + x_2)| + \tfrac{1}{3}|\cos(x_1 + x_2)|\} \\ &\leq \frac{1}{4} + \frac{1}{3} = \frac{7}{12} \end{aligned}$$

*pelo que definindo* $L = \frac{7}{12}$ *tem-se* $\|J_G(x)\|_1 \leq L < 1$ *para qualquer* $(x_1, x_2) \in \mathbb{R}^2$*. Conclui-se assim que o sistema tem uma solução única e que o método iterativo simples com a expressão*

*de recorrência dada por (4.2.1) converge para essa solução, qualquer que seja o ponto inicial escolhido.*

*Critério de paragem*
*Temos ainda que*

$$\|x_{(k+1)} - s\|_1 \leq \frac{L}{1-L} \|x_{(k+1)} - x_{(k)}\|_1 = 1.4 \|x_{(k+1)} - x_{(k)}\|_1,$$

*sendo então*

$$\varepsilon_{k+1} = 1.4 \|x_{(k+1)} - x_{(k)}\|_1$$

*um majorante da norma do erro de aproximação* $\|x_{(k+1)} - s\|_1$.

*Assim, se* $\varepsilon_{k+1} \leq 10^{-5}$ *tem-se que* $\|x_{(k+1)} - s\|_1 \leq 10^{-5}$.

*Estimativa inicial*
*Uma vez que o método converge globalmente, escolheu-se arbitrariamente o ponto inicial* $x_{1,(0)} = 1$, $x_{2,(0)} = 1$.

*Iteração 1*

$$x_{1,(1)} = 1 + \tfrac{1}{4}\cos(x_{1,(0)} + x_{2,(0)}) = 0.89596$$
$$x_{2,(1)} = 2 + \tfrac{1}{3}\sin(x_{1,(0)} + x_{2,(0)}) = 2.30310$$

*Como* $\varepsilon_1 = 1.4 \|x_{(1)} - x_{(0)}\| = 2.0 \not\leq 5 \times 10^{-5}$, *continua-se com a iteração* 2.

*Resultados*
*A tabela seguinte apresenta os resultados da aplicação do método até à satisfação do critério de paragem.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $g_1(x_{1,(k)}, x_{2,(k)})$ | $g_2(x_{1,(k)}, x_{2,(k)})$ | $\varepsilon_{k+1}$ |
|---|---|---|---|---|---|
| 0 | 1.00000 | 1.00000 | 0.89596 | 2.30310 | 2.0 |
| 1 | 0.89596 | 2.30310 | 0.75041 | 1.98085 | $6.5 \times 10^{-1}$ |
| 2 | 0.75041 | 1.98085 | 0.77075 | 2.13297 | $2.4 \times 10^{-1}$ |
| 3 | 0.77075 | 2.13297 | 0.75704 | 2.07854 | $9.5 \times 10^{-2}$ |
| 4 | 0.75704 | 2.07854 | 0.76161 | 2.10042 | $3.7 \times 10^{-2}$ |
| 5 | 0.76161 | 2.10042 | 0.75971 | 2.09198 | $1.4 \times 10^{-2}$ |
| 6 | 0.75971 | 2.09198 | 0.76043 | 2.09529 | $5.7 \times 10^{-3}$ |
| 7 | 0.76043 | 2.09529 | 0.76015 | 2.09400 | $2.2 \times 10^{-3}$ |
| 8 | 0.76015 | 2.09400 | 0.76026 | 2.09450 | $8.6 \times 10^{-4}$ |
| 9 | 0.76026 | 2.09450 | 0.76021 | 2.09431 | $3.4 \times 10^{-4}$ |
| 10 | 0.76021 | 2.09431 | 0.76023 | 2.09438 | $1.3 \times 10^{-4}$ |
| 11 | 0.76023 | 2.09438 | 0.76022 | 2.09435 | $5.1 \times 10^{-5}$ |
| 12 | 0.76022 | 2.09435 | 0.76023 | 2.09436 | $2.0 \times 10^{-5}$ |
| 13 | 0.76023 | 2.09436 | 0.76023 | 2.09436 | $7.8 \times 10^{-6}$ |

*Solução*
*O ponto obtido* $x_1 = 0.76023$, $x_2 = 2.09436$ *será então a solução procurada.*

As condições suficientes de convergência enunciadas no teorema 4.2.1 permitem guiar a escolha da função de iteração $G$, bem como do ponto inicial $x_{(0)}$. Devemos assim escolher uma função $G$ tal que $\|J_G(z)\| < 1$, para alguma norma induzida, onde $z$ é a solução pretendida. Nestas condições é possível garantir a convergência do método qualquer que seja o ponto inicial $x_{(0)}$ suficientemente próximo de $z$, ou seja, tal que $\|x_{(0)} - z\| < \varepsilon$ para $\varepsilon > 0$ suficientemente pequeno.

## 4.3 Método de Newton

O método de Newton para a resolução de sistemas de equações é também uma generalização do método já estudado para o caso de apenas uma equação. Consideremos novamente o sistema de equações $F(x) = 0$. Supondo que a matriz jacobiana $J_F(x)$ é não singular, este sistema é ainda equivalente a $J_F(x)^{-1}F(x) = 0$, ou ainda a

$$x = x - [J_F(x)]^{-1}F(x).$$

O método de Newton consiste em utilizar esta expressão como relação de recorrência para gerar uma sucessão de pontos $\{x_{(k)}\}$ que se pretende convergente para a solução $z$ do sistema de equações. Os termos da sucessão são calculados a partir

$$x_{(k+1)} = x_{(k)} - [J_F(x_{(k)})]^{-1}F(x_{(k)}), \qquad k = 1, 2, \ldots$$

sendo o ponto inicial $x_{(0)}$ convenientemente escolhido.

Para obter $x_{(k+1)}$ é necessário determinar

$$J_F(x_{(k)}) = \left[ \begin{array}{ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{array} \right]_{x_{(k)}}$$

sendo em seguida calculado $v_{(k)} = [J_F(x_{(k)})]^{-1}F(x_{(k)})$. Este cálculo efectua-se resolvendo o seguinte sistema de equações lineares

$$J_F(x_{(k)})\, v_{(k)} = F(x_{(k)}).$$

Finalmente, obtém-se $x_{(k+1)}$ a partir da expressão

$$x_{(k+1)} = x_{(k)} - v_{(k)}.$$

O seguinte teorema apresenta condições suficientes para a convergência do método de Newton. Tal como no caso unidimensional, verifica-se que este método apresenta uma convergência quadrática desde que a matriz jacobiana avaliada na solução do sistema de equações seja não singular.

**Teorema 4.3.1.** *Sejam $F$ de classe $C^2$ e $z$ tal que $F(z) = 0$. Se $\det(J_F(z)) \neq 0$ então a sucessão gerada pelo método de Newton é convergente para $z$ qualquer que seja o ponto inicial $x_{(0)}$ suficientemente próximo de $z$. Verifica-se ainda que existe uma constante positiva $c$ tal que*

$$\|z - x_{(k+1)}\| \leq c \, \|z - x_{(k)}\|^2,$$

*ou seja a convergência é quadrática.*

O exemplo seguinte ilustra a aplicação do método de Newton na resolução de um sistema de equações não lineares.

**Exemplo 4.3.1.** *Voltemos ao sistema de equações*

$$\begin{cases} 4x_1x_2^2 - 2x_1^2x_2 + 2 = 0 \\ 2x_1 - 4x_2 + \sqrt{x_1x_2} - 3 = 0 \end{cases}$$

*Definindo a função*

$$F(x) = \begin{bmatrix} 4x_1x_2^2 - 2x_1^2x_2 + 2 \\ 2x_1 - 4x_2 + \sqrt{x_1x_2} - 3 \end{bmatrix},$$

*obtém-se a matriz jacobiana*

$$J_F(x) = \begin{bmatrix} 4x_2^2 - 4x_1x_2 & 8x_1x_2 - 2x_1^2 \\ 2 + \frac{1}{2}\sqrt{\frac{x_2}{x_1}} & -4 + \frac{1}{2}\sqrt{\frac{x_1}{x_2}} \end{bmatrix}.$$

*A expressão de recorrência do método de Newton tomará para este caso a forma*

$$\begin{bmatrix} x_{1,(k)} \\ x_{2,(k)} \end{bmatrix} = \begin{bmatrix} x_{1,(k-1)} \\ x_{2,(k-1)} \end{bmatrix} - \begin{bmatrix} v_{1,(k-1)} \\ v_{2,(k-1)} \end{bmatrix}.$$

*onde*

$$\begin{bmatrix} 4x_{2,(k-1)}^2 - 4x_{1,(k-1)}x_{2,(k-1)} & 8x_{1,(k-1)}x_{2,(k-1)} - 2x_{1,(k-1)}^2 \\ 2 + \frac{1}{2}\sqrt{\frac{x_{2,(k-1)}}{x_{1,(k-1)}}} & -4 + \frac{1}{2}\sqrt{\frac{x_{1,(k-1)}}{x_{2,(k-1)}}} \end{bmatrix} \begin{bmatrix} v_{1,(k-1)} \\ v_{2,(k-1)} \end{bmatrix} = \begin{bmatrix} 4x_{1,(k-1)}x_{2,(k-1)}^2 - 2x_{1,(k-1)}^2x_{2,(k-1)} + 2 \\ 2x_{1,(k-1)} - 4x_{2,(k-1)} + \sqrt{x_{1,(k-1)}x_{2,(k-1)}} - 3 \end{bmatrix}.$$

*Iniciando as iterações no ponto $x_{1,(0)} = 1.5$ e $x_{2,(0)} = 1$ obtém-se*

$$F(x_{(0)}) = \begin{bmatrix} 3.5 \\ -2.77526 \end{bmatrix}$$

*e também*

$$J_F(x_{(0)}) = \begin{bmatrix} -2 & 7.5 \\ 2.40825 & -3.38763 \end{bmatrix}.$$

*Tem-se então que*

$$\begin{bmatrix} -2 & 7.5 \\ 2.40825 & -3.38763 \end{bmatrix} \begin{bmatrix} v_{1,(0)} \\ v_{2,(0)} \end{bmatrix} = \begin{bmatrix} 3.5 \\ -2.77526 \end{bmatrix}.$$

*Resolvendo este sistema obtém-se*

$$\begin{bmatrix} v_{1,(0)} \\ v_{2,(0)} \end{bmatrix} = \begin{bmatrix} -0.79366 \\ 0.25502 \end{bmatrix}$$

*resultando então*

$$\begin{bmatrix} x_{1,(1)} \\ x_{2,(1)} \end{bmatrix} = \begin{bmatrix} 2.29366 \\ 0.74498 \end{bmatrix}.$$

*Continuando a aplicar o método obtêm-se os resultados constantes na tabela seguinte*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $f_1(x_{(k)})$ | $f_2(x_{(k)})$ | $v_{1,(k)}$ | $v_{2,(k)}$ | $x_{2,(k+1)}$ | $x_{2,(k+1)}$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1.50000 | 1.00000 | 3.50000 | −2.77526 | −0.79366 | 0.25502 | 2.29366 | 0.74498 |
| 1 | 2.29366 | 0.74498 | −0.74662 | −0.08540 | 0.36026 | 0.29097 | 1.93340 | 0.45401 |
| 2 | 1.93340 | 0.45401 | 0.19989 | −0.01235 | −0.06661 | −0.04616 | 2.00000 | 0.50017 |
| 3 | 2.00000 | 0.50017 | 0.00000 | −0.00050 | 0.00000 | 0.00017 | 2.00000 | 0.50000 |

A convergência quadrática do método de Newton é patente neste exemplo em que se obtém a solução do sistema em 3 iterações com um erro inferior a $10^{-5}$. Esta característica de elevada rapidez é uma das grandes vantagens do método de Newton. Entre as suas maiores desvantagens inclui-se o elevado número de operações necessárias à execução de cada iteração do método (nomeadamente a resolução de um sistema de equações lineares) e também a necessidade de recorrer ao cálculo de derivadas das funções que definem o sistema de equações. Deve ainda referir-se que uma das maiores dificuldades na aplicação deste método é a garantia da sua convergência. De facto, em muitas situações não existem à partida estimativas iniciais suficientemente próximas da solução que garantam a convergência do método de Newton. Tendo em vista ultrapassar as principais desvantagens e dificuldades deste método podem ser utilizadas algumas modificações do seu funcionamento.

Por exemplo, para diminuir o peso computacional do método, é habitual não recalcular a matriz jacobiana (e obviamente a sua inversa) todas as iterações. Este procedimento reduz, na maioria dos casos, a rapidez de convergência do método (avaliada em número de iterações) mas as iterações serão mais rápidas, resultando muitas vezes num menor esforço total para a obtenção da solução com uma dada precisão. Muitas vezes ainda, as derivadas parciais que compõem a matriz jacobiana são aproximadas por quocientes de diferenças finitas.

Para garantir a convergência do método para um maior conjunto de pontos iniciais é frequente alterar a expressão de recorrência do método para

$$x_{(k)} = x_{(k-1)} - \alpha_{k-1} \cdot [J_F(x_{(k-1)})]^{-1} F(x_{(k-1)}),$$

onde o valor positivo $\alpha_{k-1}$, designado por **passo**, é escolhido, em cada iteração, de forma a que

$$\|F(x_{(k)})\| < \|F(x_{(k-1)})\|,$$

sendo aqui utilizada $\|F\|$ como "medida da distância à solução do sistema".

# Capítulo 5

# Sistemas de Equações Lineares

## 5.1 Introdução

Neste capítulo iremos abordar a resolução de sistemas de equações lineares. De uma forma geral poderemos ter um sistema $m$ equações a $n$ incógnitas como o representado abaixo.

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{cases}$$

Este sistema, utilizando uma notação matricial, pode ainda ser escrito na forma

$$Ax = b$$

onde se tem que

| | |
|---|---|
| $A \in \mathbb{R}^{m\times n}$ | é a matriz dos coeficientes, de elementos $a_{ij}$, |
| $b \in \mathbb{R}^m$ | é o vector dos termos independentes, de elementos $b_i$, |
| $x \in \mathbb{R}^n$ | é o vector de incógnitas, de elementos $x_j$. |

Este estudo incidirá sobre os designados **sistemas de Cramer**, ou seja, sistemas de $n$ equações a $n$ incógnitas possíveis e determinados, isto é, com solução única. Nestes sistemas tem-se que $A \in \mathbb{R}^{n\times n}$, verificando-se ainda que $\det A \neq 0$. Este tipo de sistemas pode ser resolvido pela **regra de Cramer**, verificando-se que

$$x_i = \frac{\det A_i}{\det A}, \qquad i = 1, \ldots, n$$

onde $A_i$ é a matriz que se obtém substituindo a coluna $i$ de $A$ pelo vector coluna $b$. Esta expressão, embora de aspecto simples, é geralmente pouco atractiva para a determinação da solução de um sistema. De facto, o cálculo de um determinante de ordem $n$, a partir da definição,

requer $(n-1)n!$ multiplicações e $n!-1$ somas ou subtracções. Por exemplo, para calcular um determinante de ordem 10 seriam necessárias mais de 40 milhões de operações aritméticas, as quais, para além de demorarem um tempo não desprezável a realizar, podem conduzir a resultados sem qualquer utilidade, devido a erros de arredondamento.

Embora seja possível calcular determinantes de modo muito mais eficiente do que a partir da definição, existem outros métodos que permitem obter a solução do sistema com a realização de um menor número de operações do que as necessárias à aplicação da regra de Cramer.

Os principais objectivos deste capítulo serão estudar métodos que permitam resolver numericamente sistemas de $n$ equações a $n$ incógnitas de modo **eficiente**, isto é, executando um *pequeno* número de operações aritméticas, e **eficaz**, isto é, fornecendo boas aproximações da solução exacta, bem como analisar algumas questões numéricas associadas aos sistemas de equações lineares.

## 5.2 Eliminação gaussiana

A eliminação gaussiana é um método directo de resolução de uma sistemas de equações lineares pois fornece a solução exacta do sistema num número finito de operações, quando se utiliza aritmética exacta.

Comecemos por recordar que se o sistema a resolver estiver numa forma triangular

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n &= b_2 \\ \vdots & \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn}x_n &= b_n \end{cases}$$

a obtenção da solução é imediata. Da última equação obtém-se imediatamente o valor de $x_n$ por

$$x_n = \frac{b_n}{a_{nn}}.$$

Substituindo o valor de $x_n$ na penúltima equação obtém-se

$$a_{n-1,n-1}x_{n-1} + a_{n-1,n}\frac{b_n}{a_{nn}} = b_{n-1} \quad \Leftrightarrow \quad x_{n-1} = \frac{b_{n-1} - a_{n-1,n}\frac{b_n}{a_{nn}}}{a_{n-1,n-1}}.$$

Substituindo agora os valores de $x_n$ e $x_{n-1}$ na antepenúltima equação obtém-se o valor de $x_{n-2}$ e assim sucessivamente até obter os valores de todas as outras incógnitas.

De uma forma geral, o valor de $x_i$ obtém-se a partir da equação $i$, conhecidos os valores de $x_j$, para $j = i+1, \ldots, n$, ou seja

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}$$

Este processo é possível de aplicar se e só se $a_{ii} \neq 0$, $\forall i$, condição que é equivalente a $\det A \neq 0$, como deverá ser para que o sistema tenha solução única.

O método de Gauss, ou de eliminação gaussiana, consiste em transformar o sistema original num outro equivalente que seja triangular superior. Este processo é realizado em etapas sucessivas. Na etapa $j$ são anulados os coeficientes $a_{ij}$, com $i > j$, ou seja, a variável $x_j$ é eliminada nas equações $i > j$. Esta eliminação é feita por **pivotação**, ou seja, para cada $i > j$ a equação $i$ é substituída pela sua soma com múltiplo da equação $j$, de modo a anular o elemento $a_{ij}$.

Na etapa $j$, a equação $j$ é designada por **equação pivot** e o elemento $a_{jj}$ é designado por **elemento pivot**. O múltiplo $m_{ij}$ da equação $j$ a somar à equação $i$ deverá ser

$$m_{ij} = -\frac{a_{ij}}{a_{jj}}.$$

Caso o elemento pivot $a_{jj}$ seja nulo, a equação $j$ deverá ser trocada com uma equação $i$, com $i > j$, tal que $a_{ij} \neq 0$.

**Exemplo 5.2.1.** *Resolver o sistema de equações por eliminação gaussiana.*

$$\begin{cases} 2x_1 + 3x_2 - x_3 = 5 \\ 4x_1 + 4x_2 - 3x_3 = 3 \\ -2x_1 + 3x_2 - x_3 = 1 \end{cases}$$

***Resolução***

***1ª etapa**: equação pivot: 1ª, elemento pivot: $a_{11} = 2$*

- *a equação pivot, multiplicada por $m_{21} = -\frac{4}{2} = -2$, é somada à 2ª equação, anulando o elemento $a_{21}$*
- *a equação pivot, multiplicada por $m_{22} = -\frac{-2}{2} = 1$, é somada à 3ª equação, anulando o elemento $a_{31}$*

*Após a 1ª etapa o sistema a resolver será*

$$\begin{cases} 2x_1 + 3x_2 - x_3 = 5 \\ -2x_2 - x_3 = -7 \\ 6x_2 - 2x_3 = 6 \end{cases}$$

***2ª etapa**: equação pivot: 2ª, elemento pivot: $a_{22} = -2$*

- *a equação pivot, multiplicada por $m_{32} = -\frac{6}{-2} = 3$, é somada à 3ª equação, anulando o elemento $a_{32}$*

*Após a 2ª etapa o sistema a resolver será*

$$\begin{cases} 2x_1 + 3x_2 - x_3 = 5 \\ -2x_2 - x_3 = -7 \\ -5x_3 = -15 \end{cases}$$

*Este é um sistema triangular superior cuja solução se determina facilmente por substituição inversa, resultando*

$$\begin{cases} x_1 = 1 \\ x_2 = 2 \\ x_3 = 3 \end{cases}$$

As dificuldades de utilização do método de eliminação gaussiana aparecem apenas quando se utiliza aritmética com precisão finita com os inerentes erros de arredondamento. O exemplo seguinte ilustra estas dificuldades.

**Exemplo 5.2.2.** *Resolver o sistema seguinte com aritmética de* 4 *dígitos.*

$$\begin{cases} 0.0002x_1 + 1.672x_2 = 1.673 \\ 1.336x_1 - 2.471x_2 = 4.209 \end{cases}$$

*Nota: A solução exacta deste sistema é* $x_1 = 5$, $x_2 = 1$.

***Resolução***

*Sendo* $m_{21} = -\frac{1.336}{2\times 10^{-4}} = -6680$, *o coeficiente de* $x_2$ *na equação* 2 *será*

$$-6680 \times 1672 - 2.471 = -1.117 \times 10^4 - 2.471 = -1.117 \times 10^4$$

*e o termo independente será*

$$-6680 \times 1.673 + 4.209 = -1.118 \times 10^4 + 4.209 = -1.118 \times 10^4$$

*obtendo-se o sistema*

$$\begin{cases} 2 \times 10^{-4}x_1 + 1.672x_2 = 1.673 \\ - 1.117 \times 10^4 x_2 = -1.118 \times 10^4 \end{cases}$$

*Agora,* $x_2$ *determina-se facilmente por*

$$x_2 = \frac{1.118}{1.117} = 1.001$$

*Substituindo este valor na equação* 1 *obtém-se*

$$x_1 = \frac{1.673 - 1.672 \times 1.001}{2.000 \times 10^{-4}} = \frac{1.673 - 1.674}{2.000 \times 10^{-4}} = \frac{-1.000 \times 10^{-4}}{2.000 \times 10^{-4}} = -5.000$$

*pelo que a solução obtida é*

$$\begin{cases} x_1 = -5.000 \\ x_2 = 1.001 \end{cases}$$

*Resolvamos agora o sistema, com a ordem das equações alterada, ou seja,*

$$\begin{cases} 1.336x_1 - 2.471x_2 = 4.209 \\ 2.0000 \times 10^{-4}x_1 + 1.672x_2 = 1.673 \end{cases}$$

*Sendo* $m_{21} = -\frac{2.0000\times10^{-4}}{1.336} = -1.497 \times 10^{-4}$*, o coeficiente de* $x_2$ *na equação* 2 *de agora, será*

$$-1.497 \times 10^{-4} \times (-2.471) + 1.672 = 3.700 \times 10^{-4} + 1.672 = 1.672$$

*e o termo independente desta mesma equação fica*

$$-1.497 \times 10^{-4} \times 4.209 + 1.673 = -6.301 \times 10^{-4} + 1.672 = 1.672$$

*obtendo-se o sistema*

$$\begin{cases} 1.336x_1 - 2.471x_2 = 4.209 \\ \qquad\qquad 1.672x_2 = 1.672 \end{cases}$$

*A solução assim obtida será*

$$\begin{cases} x_2 = \frac{1.672}{1.672} = 1.000 \\ x_1 = \frac{4.209+2.471\times1.000}{1.336} = 5.000 \end{cases} \qquad \textit{que é a solução exacta!}$$

*Mesmo que no cálculo de* $x_1$ *se tivesse usado* $x_2 = 1.001$ *obter-se-ia*

$$x_1 = \frac{4.209 + 2.471 \times 1.001}{1.336} = 5.002$$

*quando no primeiro caso se obteve* $x_1 = -5.000$*. Qual a razão de tão grande diferença?*

Neste exemplo, após a redução do sistema a uma forma triangular superior e ao cálculo de $x_2$ a partir da última equação, o valor de $x_1$ é obtido por

$$x_1 = \frac{b_1}{a_{11}} - \frac{a_{12}}{a_{11}}x_2,$$

onde os elementos da matriz de coeficientes e do vector de termos independentes se referem ao sistema triangular superior obtido. Se o valor de $x_2$ usado nesta expressão estiver afectado de um erro absoluto $\varepsilon$, então $x_1$ virá afectado de um erro, em valor absoluto, dado por

$$\left|\frac{a_{12}}{a_{11}}\right| \varepsilon.$$

Note-se que no primeiro caso se tinha

$$\left|\frac{a_{12}}{a_{11}}\right| = \left|\frac{1.672}{2 \times 10^{-4}}\right| = 8360,$$

enquanto no segundo este quociente era

$$\left|\frac{a_{12}}{a_{11}}\right| = \left|\frac{2.471}{1.336}\right| = 1.850,$$

interessando portanto que $\left|\frac{a_{12}}{a_{11}}\right|$ seja o menor possível.

Generalizando agora este resultado, conclui-se facilmente da expressão de cálculo de $x_i$ por substituição inversa

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}$$

que estando os valores $x_j$ afectados de erros, então $x_i$ também estará, de acordo com a expressão

$$\varepsilon_{x_i} \leq \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|}\varepsilon_{x_j}.$$

De forma a diminuis a influência dos erros de $x_j$, para $j > i$, no cálculo de $x_i$, interessa que os quocientes $\frac{|a_{ij}|}{|a_{ii}|}$ sejam pequenos.

A obtenção de valores pequenos para tais quocientes pode ser garantida usando as designadas **estratégias de escolha de pivot**. Estas estratégias tiram partindo da possibilidade de escolha, numa qualquer etapa $j$ da eliminação gaussiana, quer da equação pivot a utilizar (troca de linhas) quer da variável pivot a utilizar (troca de colunas).

A **estratégia parcial de pivot (ou pivotação parcial)** apenas permite a troca de linhas de acordo com o seguinte procedimento

1. Na etapa $j$ é escolhida a equação pivot $k$ $\quad (j \leq k \leq n)$

   (a) calculam-se os valores $\quad d_i = \max_{i \leq l \leq n} |a_{il}| \quad i = j, \ldots, n;$

   (b) calculam-se os quocientes $\quad \dfrac{|a_{ij}|}{d_i} \qquad i = j, \ldots, n;$

   (c) selecciona-se para pivot equação $k$ como sendo aquela em que

   $$\frac{|a_{kj}|}{d_k} \qquad \text{é máximo.}$$

2. Troca-se a equação $k$ com a $j$.

3. Realiza-se a eliminação.

**Exemplo 5.2.3.** *Aplicando a estratégia parcial de pivot ao exemplo anterior obtém-se*

$$\begin{cases} 2.000 \times 10^{-4} x_1 + 1.672 x_2 = 1.673 \\ 1.336 x_1 - 2.471 x_2 = 4.209 \end{cases}$$

*pelo que* $\left|\frac{a_{11}}{d_1}\right| = 1.196 \times 10^{-4}$ *e* $\left|\frac{a_{21}}{d_2}\right| = 0.5406$*, concluindo-se que a equação pivot deve ser a segunda!*

Outra forma possível de escolha do elemento pivot é a designada **estratégia total de pivot (ou pivotação total)** que se pode resumir nos seguintes passos

1. Na etapa $j$ escolhe-se o elemento pivot $a_{kl}$ $\quad (j \leq k, l \leq n)$

   $\rightarrow$ $|a_{kl}|$ é máximo.

2. Troca-se a equação $j$ com a equação $k$.

3. Troca-se a variável $x_j$ com a variável $x_l$.

4. Realiza-se a eliminação.

**Exemplo 5.2.4.** *Voltando ainda ao exemplo anterior*

$$\begin{cases} 2.000 \times 10^{-4} x_1 + 1.672 x_2 = 1.673 & \rightarrow d_1 = 1.672 \\ 1.336 x_1 - 2.471 x_2 = 4.209 & \rightarrow d_2 = 2.471 \end{cases}$$

*verifica-se que* $\max_{1 \leq i,j \leq 2} |a_{ij}| = 2.471$*, para* $i = 2$ *e* $j = 2$*. Então deve trocar-se a primeira equação com a segunda (trocas de linhas) e a variável* $x_1$ *com* $x_2$ *(troca de colunas). Neste caso o sistema ficaria*

$$\begin{cases} -2.471 x_2 + 1.336 x_1 = 4.209 \\ 1.672 x_2 + 2.000 \times 10^{-4} x_1 = 1.673 \end{cases}$$

*devendo agora eliminar-se* $x_2$ *da segunda equação.*

Como é fácil de entender, a estratégia de pivotação total é computacionalmente mais "cara" pois exige troca de colunas, isto para além da troca de linhas. Em termo de qualidade dos resultados, ou seja, diminuição da propagação dos erros numéricos resultantes de arredondamentos, pode demonstrar-se que a pivotação total conduz a melhores resultados. Contudo, verifica-se também que a pivotação parcial produz resultados suficientemente bons na maioria das situações.

## 5.3 Erro e resíduo de uma solução aproximada

Como em todos os problemas de resolução numérica, também na resolução dos sistemas de equações lineares se coloca a questão da qualidade da solução aproximada obtida por via numérica.

Sejam $A \in \mathbb{R}^{n \times n}$ (invertível) e $b \in \mathbb{R}^n$ e considere-se o sistema de equações $Ax = b$. Designando por $\bar{x}$ a solução exacta e sendo $\tilde{x}$ uma solução aproximada definem-se

- erro da solução aproximada: $e = \bar{x} - \tilde{x}$,
- resíduo da solução aproximada: $r = b - A\tilde{x}$,

que são ambos elementos de $\mathbb{R}^n$.

A questão que aqui se coloca é a da estimação do erro de aproximação $e$. Note-se que este erro não se pode calcular directamente uma vez que não dispomos da solução exacta $\bar{x}$. Se este valor estivesse disponível teríamos o nosso problema resolvido, e nem precisaríamos de estimar erros de soluções aproximadas! Resta-nos então tentar obter estimativas para este erro. Uma das possibilidades será utilizar o resíduo atrás definido. Repare-se que erro e resíduo estão relacionados, pois $r = A\bar{x} - A\tilde{x} = A(\bar{x} - \tilde{x}) = Ae$.

Se $\tilde{x} = \bar{x}$ então o erro é nulo, e o resíduo também será nulo. Por outro lado se o resíduo for nulo, o erro também o será (e a solução será exacta). E quando $\tilde{x} \neq \bar{x}$, será que a um erro pequeno corresponde um resíduo pequeno? E a um resíduo pequeno, corresponderá um erro pequeno?

**Exemplo 5.3.1.** *O sistema*

$$\begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

*tem como solução exacta* $\bar{x} = [1\ 1]^{\mathrm{T}}$.

*Para a solução aproximada* $\tilde{x} = [1.01\ 1.01]^{\mathrm{T}}$ *tem-se* $e = [-0.01\ -0.01]^{\mathrm{T}}$ *e* $r = [-0.02\ -0.02]^{\mathrm{T}}$. *O erro relativo é de* 1% *em cada componente e o resíduo relativo é também de* 1% *em cada componente.*

*Para a solução aproximada* $\hat{x} = [2\ 0]^{\mathrm{T}}$ *tem-se* $e = [-1\ 1]^{\mathrm{T}}$ *e* $r = [-0.02\ 0.02]$. *O erro relativo é agora de* 100% *em cada componente, sendo o resíduo relativo de apenas* 1% *em cada componente.*

**Exemplo 5.3.2.** *O sistema*

$$\begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$

*tem como solução exacta* $\bar{x} = [100\ -100]$.

*Para a solução aproximada* $\tilde{x} = [101\ -99]$ *tem-se* $e = [-1\ -1]$ *e* $r = [-2\ -2]$.
*O erro relativo é de* 1% *em cada componente e o resíduo relativo é agora de* 100% *em cada componente.*

Nestes exemplos, os erros e resíduos foram comparados usando valores "relativos". Estes valores foram determinados relativamente à componente máxima da solução, no caso do erro, e à componente máxima do vector de termos independentes, no caso do resíduo. Como estes exemplos ilustram, nem sempre erros pequenos correspondem a resíduos pequenos nem resíduos pequenos a erros pequenos. Vamos então analisar a relação entre erro e resíduo de uma solução aproximada. Do exposto atrás pode escrever-se

$$\begin{array}{cc} r = Ae & e = A^{-1}r \\ \Downarrow & \Downarrow \\ \|r\| = \|Ae\| \leq \|A\|\,\|e\| & \|e\| = \|A^{-1}r\| \leq \|A^{-1}\|\,\|r\| \end{array}$$

concluindo-se que

$$\frac{\|r\|}{\|A\|} \leq \|e\| \leq \|A^{-1}\|\,\|r\|. \tag{5.3.1}$$

Por outro lado, tem-se que

$$\begin{array}{cc} b = A\bar{x} & \bar{x} = A^{-1}b \\ \\ \|b\| = \|A\bar{x}\| \leq \|A\| \cdot \|\bar{x}\| & \|\bar{x}\| = \|A^{-1}b\| \leq \|A^{-1}\| \cdot \|b\| \end{array}$$

concluindo-se também que

$$\frac{\|b\|}{\|A\|} \leq \|\bar{x}\| \leq \|A^{-1}\| \cdot \|b\|,$$

ou, de forma equivalente

$$\frac{1}{\|A^{-1}\| \cdot \|b\|} \leq \frac{1}{\|\bar{x}\|} \leq \frac{\|A\|}{\|b\|}. \tag{5.3.2}$$

Das expressões (5.3.1) e (5.3.2) pode ainda concluir-se que

$$\frac{1}{\|A\| \cdot \|A^{-1}\|} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|\bar{x}\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|r\|}{\|b\|}.$$

O valor $\|A\| \cdot \|A^{-1}\|$ que aparece nesta última expressão é designado por **número de condição** da matriz $A$ e habitualmente representado por $\text{cond}(A)$. É de notar que o número de condição de uma matriz depende obviamente da norma escolhida. Agora, a relação entre erro e resíduo pode ser escrita como

$$\frac{1}{\text{cond}(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|\bar{x}\|} \leq \text{cond}(A) \frac{\|r\|}{\|b\|},$$

onde $\frac{\|e\|}{\|\bar{x}\|}$ pode ser interpretado como o erro relativo e $\frac{\|r\|}{\|b\|}$ como o resíduo relativo.

Notando que para toda a matriz $A$ invertível se tem $I = AA^{-1}$ conclui-se que

$$1 = \|I\| \leq \|A\| \cdot \|A^{-1}\|$$

verificando-se então que $\text{cond}(A) \geq 1$.

Diz-se que a matriz $A$ é **bem condicionada** quando $\text{cond}(A) \simeq 1$. Nesta situação, o erro relativo $\frac{\|e\|}{\|\bar{x}\|}$ será da mesma ordem de grandeza do resíduo relativo $\frac{\|r\|}{\|b\|}$. Se $\text{cond}(A) \gg 1$ a matriz diz-se **mal condicionada**. Em tais casos, a relação entre erro relativo e resíduo relativo obtida atrás é pouco informativa. A erros pequenos podem corresponder resíduos grandes e resíduos pequenos podem corresponder a erros grandes.

O cálculo de $\text{cond}(A)$ pela definição implica a determinação de $A^{-1}$, o que pode não ser muito prático. Uma alternativa para estimar $\text{cond}(A)$ será utilizar a seguinte propriedade

$$\frac{1}{\text{cond}(A)} = \min_{B \text{ singular}} \left( \frac{\|A - B\|}{\|A\|} \right).$$

Escolhendo então uma matriz $B$ singular obtém-se um minorante para $\text{cond}(A)$ dado por

$$\text{cond}(A) \geq \frac{\|A\|}{\|A - B\|}.$$

Este minorante será tanto melhor quanto mais "próxima" de $A$ for a matriz $B$ utilizada. Podemos também concluir que o número de condição de $A$ será tanto maior quanto mais $A$ estiver próxima de uma matriz singular.

**Exemplo 5.3.3.** *A matriz dos coeficientes dos sistemas dos exemplos 5.3.1 e 5.3.2 era*

$$A = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix}.$$

*Escolhendo a matriz singular*

$$B = \begin{bmatrix} 0.99 & 0.99 \\ 0.99 & 0.99 \end{bmatrix}$$

*conclui-se, na norma $\infty$, que*

$$\text{cond}(A) \geq \frac{\|A\|_\infty}{\|A - B\|_\infty} = \frac{2}{0.02} = 100.$$

*Na verdade, tem-se neste caso que* $\text{cond}(A) = 100$*, como se pode comprovar calculando-o pela pela definição. Então, para aqueles sistemas de equações, verifica-se a relação*

$$0.01 \times \frac{\|r\|_\infty}{\|b\|_\infty} \leq \frac{\|e\|_\infty}{\|\bar{x}\|_\infty} \leq 100 \times \frac{\|r\|_\infty}{\|b\|_\infty}$$

*pelo que o resíduo relativo não fornece grande informação sobre o erro relativo e vice-versa, tal como então se tinha verificado.*

## 5.4 Perturbações no sistema de equações

Em muitas situações, os elementos da matriz de coeficientes $A$ ou do vector de termos independentes $b$ estão sujeitos a erros. Estes erros podem resultar do facto de tais elementos serem obtidos a partir de medições (sempre sujeitas a erros) ou de cálculos que originem erros de arredondamento (ou outros). Estas considerações tornam relevante a análise da sensibilidade da solução do sistema de equações $Ax = b$ face a perturbações, quer na matriz $A$, quer no vector $b$.

O resultado apresentado em seguida afirma que "variações relativas" nos termos independentes aparecem multiplicadas pelo número de condição de $A$ como "variações relativas" na solução do sistema. O majorante aqui apresentado pode ser, por vezes, bastante pessimista.

**Teorema 5.4.1.** *Considere-se o sistema de equações $Ax = b$, onde se supõe que $A \in \mathbb{R}^{n\times n}$ é não singular e $b \in \mathbb{R}^n$ é não nulo. Seja $\bar{x}$ a solução deste sistema, isto é, $\bar{x} = A^{-1}b$. Seja também $\tilde{b} \in \mathbb{R}^n$ e represente-se por $\tilde{x}$ a solução do sistema (perturbado) $Ax = \tilde{b}$, ou seja, $\tilde{x} = A^{-1}\tilde{b}$. Então verifica-se que*

$$\frac{\|\bar{x} - \tilde{x}\|}{\|\bar{x}\|} \leq \text{cond}(A)\frac{\|b - \tilde{b}\|}{\|b\|}.$$

*Demonstração.* Dado que $\bar{x} - \tilde{x} = A^{-1}(b - \tilde{b})$, obtém-se a relação

$$\|\bar{x} - \tilde{x}\| \leq \|A^{-1}\| \cdot \|b - \tilde{b}\|$$

Por outro lado, tem-se $b = A\bar{x}$, e logo $\|b\| \leq \|A\| \cdot \|\bar{x}\|$, ou ainda

$$\frac{1}{\bar{x}} \leq \|A\|\frac{1}{\|b\|}$$

Multiplicando termo a termos estas desigualdades obtém-se a relação

$$\frac{\|\bar{x} - \tilde{x}\|}{\|\bar{x}\|} \leq \|A\| \cdot \|A^{-1}\|\frac{\|b - \tilde{b}\|}{\|b\|}$$

que é equivalente à relação pretendida, pois $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$. □

**Exemplo 5.4.1.** *Considere-se o sistema de equações $Ax = b$, onde*

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 4 & 3 & 1 \\ 2 & 2 & 3 \end{bmatrix} \qquad e \qquad b = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}.$$

*A solução deste sistema é $\bar{x} = [-0.2 \;\; 1 \;\; -0.2]^{\mathrm{T}}$. Considerando o novo termo independente $\tilde{b} = [1.1 \;\; 2.2 \;\; 0.9]^{\mathrm{T}}$, obtém-se a solução $\tilde{x} = [-0.62 \;\; 1.7 \;\; -0.42]^{\mathrm{T}}$.*

*A "variação relativa" nos termos independentes, medida na norma $\infty$, é*

$$\frac{\|b - \tilde{b}\|_\infty}{\|b\|_\infty} = \frac{0.2}{2} = 0.1,$$

*enquanto a "variação relativa" nas soluções, medida na mesma norma, é*

$$\frac{\|\bar{x} - \tilde{x}\|_\infty}{\|\bar{x}\|_\infty} = \frac{0.7}{1} = 0.7,$$

*ou seja, 7 vezes superior. Neste caso tem-se que* $\mathrm{cond}(A) = 48$ *na norma $\infty$.*

Consideremos agora perturbações na matriz dos coeficientes. O resultado seguinte relaciona "variações relativas" na matriz dos coeficientes com "variações relativas" na solução do sistema. Mais uma vez, o factor de amplificação do majorante aqui apresentado é o número de condição da matriz $A$. É de notar que em algumas situações esta estimativa pode ser bastante pessimista.

**Teorema 5.4.2.** *Considere-se o sistema de equações $Ax = b$, onde se supõe que $A \in \mathbb{R}^{n\times n}$ é não singular e $b \in \mathbb{R}^n$ é não nulo. Seja $\bar{x}$ a solução deste sistema, isto é, $\bar{x} = A^{-1}b$.*

*Seja também $\tilde{A} \in \mathbb{R}^{n\times n}$, não singular, e represente-se por $\tilde{x}$ a solução do sistema (perturbado) $\tilde{A}x = b$, ou seja, $\tilde{x} = \tilde{A}^{-1}b$.*

*Então verifica-se que*

$$\frac{\|\bar{x} - \tilde{x}\|}{\|\tilde{x}\|} \leq \mathrm{cond}(A)\frac{\|\tilde{A} - A\|}{\|A\|}.$$

*Demonstração.* As hipóteses do teorema permitem escrever

$$\bar{x} = A^{-1}b = A^{-1}\tilde{A}\tilde{x} = A^{-1}(A + \tilde{A} - A)\tilde{x} = A^{-1}(\tilde{A} - A)\tilde{x} + \tilde{x}$$

ou seja,

$$\bar{x} - \tilde{x} = A^{-1}(\tilde{A} - A)\tilde{x}.$$

Então, verifica-se que $\|\bar{x} - \tilde{x}\| \leq \|A^{-1}\| \cdot \|\tilde{A} - A\| \cdot \|\tilde{x}\|$. Ou ainda,

$$\frac{\|\bar{x} - \tilde{x}\|}{\|\tilde{x}\|} \leq \|A^{-1}\| \cdot \|A\| \frac{\|\tilde{A} - A\|}{\|A\|} = \mathrm{cond}(A)\frac{\|\tilde{A} - A\|}{\|A\|}$$

como se pretendia mostrar. □

**Exemplo 5.4.2.** *Considere-se o sistema de equações $Ax = b$, onde*

$$A = \begin{bmatrix} 1 & 5 & 10 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{bmatrix} \qquad e \qquad b = \begin{bmatrix} 16 \\ -5 \\ 1 \end{bmatrix},$$

*cuja solução é $\bar{x} = [1\ 1\ 1]^{\mathrm{T}}$.*

*Considere-se também a matriz $\tilde{A}$, definida por*

$$\tilde{A} = \begin{bmatrix} 1 & 5 & 10 \\ 0 & 1 & -6 \\ 0 & 0 & 1.1 \end{bmatrix}$$

*A solução do sistema $\tilde{A}x = b$ é $\tilde{x} = \left[\frac{51}{11}\ \ \frac{5}{11}\ \ \frac{10}{11}\right]^{\mathrm{T}}$. A perturbação na matriz dos coeficientes é*

$$\tilde{A} - A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}.$$

*Neste caso, a variação relativa na matriz dos coeficientes é, na norma $\infty$,*

$$\frac{\|\tilde{A} - A\|_\infty}{\|A\|_\infty} = \frac{0.1}{16} = \frac{1}{160}.$$

*A variação relativa na solução será*

$$\frac{\|\bar{x} - \tilde{x}\|_\infty}{\|\tilde{x}\|_\infty} = \frac{\frac{40}{11}}{\frac{51}{11}} = \frac{40}{51},$$

*ou seja, $\frac{6400}{51}$ (cerca de 125) vezes maior. Neste caso tem-se que $\mathrm{cond}(A) = 736$ na norma $\infty$.*

## 5.5 Métodos iterativos

Vamos agora estudar métodos iterativos para a resolução de sistemas de equações lineares. Consideremos novamente um sistema de equações $Ax = b$. De uma forma geral, os métodos iterativos consistem na substituição do sistema original por um outro equivalente, da forma

$$x = Gx + d,$$

onde $G \in \mathbb{R}^{n\times n}$ e $d \in \mathbb{R}^n$, e na geração de uma sucessão $\{x_{(k)}\} \subset \mathbb{R}^n$ pela expressão de recorrência

$$x_{(k+1)} = Gx_{(k)} + d \qquad\qquad k = 0, 1, \ldots,$$

a partir de um valor inicial $x_{(0)} \in \mathbb{R}^n$. Obviamente que se pretende que a sucessão $\{x_{(k)}\}$ seja convergente para $A^{-1}b$, que é o valor procurado.

Dado o sistema de equações, onde $a_{ii} \neq 0 \; \forall i$,

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

resolvendo cada equação $i$ em ordem à variável $x_i$, obtém-se o sistema equivalente

$$\begin{cases} x_1 = & -\frac{a_{12}}{a_{11}}x_2 & -\frac{a_{13}}{a_{11}}x_3 & - \cdots & -\frac{a_{1n}}{a_{11}}x_n & + & \frac{b_1}{a_{11}} \\ x_2 = -\frac{a_{21}}{a_{22}}x_1 & & -\frac{a_{23}}{a_{22}}x_3 & - \cdots & -\frac{a_{2n}}{a_{22}}x_n & + & \frac{b_2}{a_{22}} \\ \vdots & & & & & & \vdots \\ x_n = -\frac{a_{n1}}{a_{nn}}x_1 & -\frac{a_{n2}}{a_{nn}}x_2 & -\frac{a_{n3}}{a_{nn}}x_3 & - \cdots & & + & \frac{b_n}{a_{nn}} \end{cases}$$

Definindo $B \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}^n$ respectivamente por

$$b_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}} & \text{se } i \neq j \\ 0 & \text{se } i = j \end{cases} \qquad i,j = 1,\ldots,n, \quad \text{e}$$

$$c_i = \frac{b_i}{a_{ii}} \qquad i = 1,\ldots,n,$$

este último sistema pode ser escrito como $x = Bx + c$.

O **método iterativo de Jacobi** é caracterizado por utilizar a expressão de recorrência

$$x_{(k+1)} = Bx_{(k)} + c$$

ou, de forma equivalente para cada uma das variáveis,

$$x_{i,(k+1)} = \sum_{j=1}^{n} \left[ b_{ij}\, x_{j,(k)} \right] + c_i,$$

isto para $i = 1,\ldots,n$.

O seguinte exemplo ilustra a aplicação do método de Jacobi .

**Exemplo 5.5.1.** *Aplicar o método de Jacobi para resolver o sistema*

$$\begin{bmatrix} 3 & -1 & 1 \\ 0 & 2 & 1 \\ 1 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}.$$

***Resolução***

*Expressões de recorrência*

*Isolando uma variável em cada uma das equações, obtêm-se as expressões de recorrência*

$$\begin{cases} x_{1,(k+1)} = & & \frac{1}{3}x_{2,(k)} & -\frac{1}{3}x_{3,(k)} & + & 1 \\ x_{2,(k+1)} = & & & -\frac{1}{2}x_{3,(k)} & + & \frac{3}{2} \\ x_{3,(k+1)} = & -\frac{1}{4}x_{1,(k)} & +\frac{1}{2}x_{2,(k)} & & + & \frac{3}{4} \end{cases}$$

*Estimativa inicial*

*Escolhamos* $x_0 = [0 \quad 0 \quad 0]^{\mathrm{T}}$.

*Iteração 1*

$$\begin{cases} x_{1,(1)} = \frac{1}{3} \times 0 - \frac{1}{3} \times 0 + 1 = 1 \\ x_{2,(1)} = -\frac{1}{2} \times 0 + \frac{3}{2} = 1.5 \\ x_{3,(1)} = -\frac{1}{4} \times 0 + \frac{1}{2} \times 0 + \frac{3}{4} = 0.75 \end{cases}$$

*Iteração 2*

$$\begin{cases} x_{1,(2)} = \frac{1}{3} \times 1.5 - \frac{1}{3} \times 0.75 + 1 = 1.25 \\ x_{2,(2)} = -\frac{1}{2} \times 0.75 + \frac{3}{2} = 1.125 \\ x_{3,(2)} = -\frac{1}{4} \times 1 + \frac{1}{2} \times 1.5 + \frac{3}{4} = 1.25 \end{cases}$$

*Resultados*

*Continuando a aplicação do método, obtêm-se as seguintes estimativas*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1.0000 | 1.5000 | 0.7500 |
| 2 | 1.2500 | 1.1250 | 1.2500 |
| 3 | 0.9583 | 0.8750 | 1.0000 |
| 4 | 0.9583 | 1.0000 | 0.9479 |
| 5 | 1.0174 | 1.0260 | 1.0104 |
| 6 | 1.0052 | 0.9948 | 1.0087 |
| 7 | 0.9954 | 0.9957 | 0.9961 |
| 8 | 0.9999 | 1.0020 | 0.9990 |
| 9 | 1.0010 | 1.0005 | 1.0010 |
| 10 | 0.9998 | 0.9995 | 1.0000 |
| 11 | 0.9998 | 0.9999 | 0.9998 |

*que convergem para a solução* $[1 \quad 1 \quad 1]^{\mathrm{T}}$.

Analisando a expressão de recorrência do método de Jacobi, verifica-se a determinação da nova estimativa de uma variável utiliza as estimativas da iteração anterior das outras variáveis. Considerando que as novas estimativas são determinadas sequencialmente, ou seja, primeiro $x_1$, depois $x_2$ e assim sucessivamente até $x_n$, verifica-se que quando se vai calcular a nova estimativa de $x_i$ já se dispõe de novos valores para as variáveis $x_j$, como $j = 1, \ldots, i-1$.

O **método iterativo de Gauss-Seidel** tira partido deste facto, utilizando no cálculo da nova estimativa de uma variável sempre a última estimativa disponível das variável necessárias. Assim, podemos caracterizar o método de Gauss-Seidel pela expressão de recorrência

$$x_{i,(k+1)} = \sum_{j=1}^{i-1} \left[ b_{ij}\, x_{j,(k+1)} \right] + \sum_{j=i+1}^{n} \left[ b_{ij}\, x_{j,(k)} \right] + c_i,$$

para $i = 1, \ldots, n$. Pretende-se com esta alteração obter uma maior rapidez de convergência para a solução pretendida.

A aplicação do método de Gauss-Seidel encontra-se ilustrada no exemplo seguinte.

**Exemplo 5.5.2.** *Aplicar o método de Gauss-Seidel para resolver o sistema*

$$\begin{bmatrix} 3 & -1 & 1 \\ 0 & 2 & 1 \\ 1 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}.$$

***Resolução***

*Expressões de recorrência*

*As expressões de recorrência são agora as seguintes*

$$\begin{cases} x_{1,(k+1)} = & & \frac{1}{3}x_{2,(k)} & -\frac{1}{3}x_{3,(k)} & + & 1 \\ x_{2,(k+1)} = & & & -\frac{1}{2}x_{3,(k)} & + & \frac{3}{2} \\ x_{3,(k+1)} = & -\frac{1}{4}x_{1,(k+1)} & +\frac{1}{2}x_{2,(k+1)} & & + & \frac{3}{4} \end{cases}$$

*Estimativa inicial*

*Escolhamos* $x_0 = [0 \quad 0 \quad 0]^{\mathrm{T}}$.

*Iteração 1*

$$\begin{cases} x_{1,(1)} = \frac{1}{3} \times 0 - \frac{1}{3} \times 0 + 1 = 1 \\ x_{2,(1)} = -\frac{1}{2} \times 0 + \frac{3}{2} = 1.5 \\ x_{3,(1)} = -\frac{1}{4} \times 1 + \frac{1}{2} \times 1.5 + \frac{3}{4} = 1.25 \end{cases}$$

*Iteração 2*

$$\begin{cases} x_{1,(2)} = \frac{1}{3} \times 1.5 - \frac{1}{3} \times 1.25 + 1 = 1.0833 \\ x_{2,(2)} = -\frac{1}{2} \times 1.25 + \frac{3}{2} = 0.875 \\ x_{3,(2)} = -\frac{1}{4} \times 1.0833 + \frac{1}{2} \times 0.875 + \frac{3}{4} = 0.9167 \end{cases}$$

*Resultados*

*Continuando a aplicação do método, obtêm-se as seguintes estimativas*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1.0000 | 1.5000 | 1.2500 |
| 2 | 1.0833 | 0.8750 | 0.9167 |
| 3 | 0.9861 | 1.0417 | 1.0243 |
| 4 | 1.0058 | 0.9878 | 0.9925 |
| 5 | 0.9985 | 1.0038 | 1.0023 |
| 6 | 1.0005 | 0.9989 | 0.9993 |
| 7 | 0.9999 | 1.0003 | 1.0002 |
| 8 | 1.0000 | 0.9999 | 0.9999 |

*que convergem para a solução* $[1 \quad 1 \quad 1]^{\mathrm{T}}$.

Em ambos os exemplos atrás apresentados verifica-se que as sucessões geradas pelos métodos iterativos convergem para a solução do sistema procurada. No entanto este comportamento nem sempre se verifica, como se mostra no seguinte exemplo.

**Exemplo 5.5.3.** *Aplicar o método de Jacobi e também o método de Gauss-Seidel para resolver o sistema*

$$\begin{bmatrix} 1 & -1 & 1 \\ 0 & 2 & -1 \\ 1 & -2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

***Resolução***

*Aplicando o método de Jabobi, partindo de* $x_0 = [0 \quad 0 \quad 0]^{\mathrm{T}}$*, obtém-se uma sucessão que não converge para a solução (única)* $\bar{x} = [1 \quad 1 \quad 1]^{\mathrm{T}}$*, como se pode ver pela tabela seguinte.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1.0000 | 0.5000 | 0.5000 |
| 2 | 1.0000 | 0.7500 | 0.5000 |
| 3 | 1.2500 | 0.7500 | 0.7500 |
| 4 | 1.0000 | 0.8750 | 0.6250 |
| 5 | 1.2500 | 0.8125 | 0.8750 |
| 6 | 0.9375 | 0.9375 | 0.6875 |
| 7 | 1.2500 | 0.8438 | 0.9688 |
| 8 | 0.8750 | 0.9844 | 0.7188 |
| 9 | 1.2656 | 0.8594 | 1.0469 |
| ... | ... | ... | ... |

*Aplicando agora o método de Gauss-Seidel e partindo também de* $x_0 = [0 \quad 0 \quad 0]^{\mathrm{T}}$*, obtém-se uma sucessão que converge para a solução do sistema, como se pode observar pela tabela seguinte.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 1.5000 | 0.5000 | 0.5000 |
| 2 | 1.0000 | 0.7500 | 0.7500 |
| 3 | 1.0000 | 0.8750 | 0.8750 |
| 4 | 1.0000 | 0.9375 | 0.9375 |
| 5 | 1.0000 | 0.9688 | 0.9688 |
| 6 | 1.0000 | 0.9844 | 0.9844 |
| 7 | 1.0000 | 0.9922 | 0.9922 |
| 8 | 1.0000 | 0.9961 | 0.9961 |
| 9 | 1.0000 | 0.9980 | 0.9980 |
| ... | ... | ... | ... |

Este exemplo mostra que é necessário, como seria de esperar, obter condições que garantam a convergência dos métodos iterativos estudados. As condições que iremos estudar são casos particulares de uma resultado mais geral sobre convergência de métodos iterativo de expressão de recorrência

$$x_{(k+1)} = Gx_{(k)} + d,$$

que apresentamos em seguida.

**Teorema 5.5.1.** *Sejam $G \in \mathbb{R}^{n\times n}$ e $d \in \mathbb{R}^n$. Se para alguma norma induzida se verificar $\|G\| < 1$, então*

1. *existe uma e uma só solução $\bar{x} \in \mathbb{R}^n$ da equação*

$$x = Gx + d,$$

2. *a sucessão $\{x_{(k)}\}$, gerada pela expressão de recorrência*

$$x_{(k+1)} = Gx_{(k)} + d, \qquad k = 0, 1, \ldots,$$

   *converge para $\bar{x}$, qualquer que seja o ponto inicial $x_{(0)}$,*

3. *o erro de aproximação de $\bar{x}$ por $x_{(k+1)}$, $\bar{x} - x_{(k+1)}$, satisfaz*

$$\|\bar{x} - x_{(k+1)}\| \leq \frac{\|G\|}{1 - \|G\|}\|x_{(k+1)} - x_{(k)}\|, \qquad k = 0, 1, \ldots .$$

*Demonstração.*
**1.** A equação $x = Gx + d$ é equivalente a $(I - G)x = d$, que terá uma e uma só solução se a matriz $I - G$ for não singular.

Suponha-se que $I - G$ é singular. Então existe $\tilde{x} \neq 0$ (em $\mathbb{R}^n$) tal que $(I - G)\tilde{x} = 0$, ou ainda $\tilde{x} = G\tilde{x}$. Logo, para a norma considerada, verifica-se que

$$\|\tilde{x}\| = \|G\tilde{x}\| \leq \|G\| \cdot \|\tilde{x}\|,$$

concluindo-se imediatamente que $\|G\| \geq 1$. Como este facto contraria a hipótese $\|G\| < 1$, a matriz $I - G$ terá de ser não singular, como se pretendia mostrar.

**2.** Como $\bar{x} = G\bar{x} + d$ e $x^{(k+1)} = Gx^{(k)} + d$, $\forall k$, verifica-se que

$$\bar{x} - x_{(k+1)} = G\bar{x} + d - (Gx_{(k)} + d) = G(\bar{x} - x_{(k)}), \qquad k = 0, 1, \ldots .$$

Aplicando sucessivamente esta expressão, conclui-se que

$$\bar{x} - x_{(k+1)} = G(\bar{x} - x_{(k)}) = G^2(\bar{x} - x_{(k-1)}) = \cdots = G^k(\bar{x} - x_{(0)}), \quad k = 0, 1, \ldots .$$

podendo então escrever-se que $\|\bar{x} - x_{(k)}\| \leq \|G^k\| \cdot \|\bar{x} - x_{(0)}\|$.

Por outro lado, tem-se que

$$\|G^k\| = \| \overbrace{G \times G \times \cdots \times G}^{k \text{ vezes}} \| \leq \overbrace{\|G\| \times \|G\| \times \cdots \times \|G\|}^{k \text{ vezes}} = \|G\|^k.$$

Como $\|G\| < 1$, pode afirmar-se que $\lim_{k\to+\infty} \|G\|^k = 0$, resultando então que

$$\lim_{k\to+\infty} \|\bar{x} - x_{(k)}\| = 0,$$

como se pretendia mostrar.

**3.** Partindo da expressão

$$\bar{x} - x_{(k+1)} = G(\bar{x} - x_{(k)}),$$

válida para $k = 0, 1, \ldots$, como visto atrás, pode concluir-se que

$$\bar{x} - x_{(k+1)} = G(\bar{x} - x_{(k+1)} + x_{(k+1)} - x_{(k)}) = G(\bar{x} - x_{(k+1)}) + G(x_{(k+1)} - x_{(k)}).$$

Desta expressão resulta que

$$\begin{aligned} \|\bar{x} - x_{(k+1)}\| &\leq \|G(\bar{x} - x_{(k+1)})\| + \|G(x_{(k+1)} - x_{(k)})\| \\ &\leq \|G\| \cdot \|\bar{x} - x_{(k+1)}\| + \|G\| \cdot \|x_{(k+1)} - x_{(k)}\|, \end{aligned}$$

que pode ser reescrita como

$$(1 - \|G\|)\, \|\bar{x} - x_{(k+1)}\| \leq \|G\| \cdot \|x_{(k+1)} - x_{(k)}\|.$$

Dado que $\|G\| < 1$, tem-se $1 - \|G\| > 0$, obtendo-se imediatamente a expressão pretendida. □

Seja novamente $A \in \mathbb{R}^{n\times n}$. Diz-se que matriz $A$ é **estritamente diagonalmente dominante por linhas** quando se verifica

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

ou seja, quando para cada linha da matriz se verifica que o valor absoluto do elemento da diagonal é superior à soma dos valores absolutos de todos os outros elementos.

O resultado seguinte fornece condições suficientes para a convergência do método de Jacobi. No entanto, estas condições não são necessárias para a convergência do método. Isto é, há casos em que estas condições não se verificam e o método converge.

**Teorema 5.5.2.** *Sejam $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$. Se a matriz $A$ for estritamente diagonalmente dominante por linhas então a sucessão gerada pelo método de Jacobi converge para a única solução do sistema de equações $Ax = b$, designada $\bar{x}$, qualquer que seja o ponto inicial $x_{(0)}$.*

*Demonstração.* A expressão de recorrência do método de Jacobi é

$$x_{(k+1)} = Bx_{(k)} + c,$$

onde $B$ e $c$ são obtidos à custa de $A$ e $b$, de acordo com as expressões vistas atrás.

Sendo $A$ estritamente diagonalmente dominante por linhas, verifica-se que todos os elementos da sua diagonal são não nulos. Logo, a matriz $B$ e o vector $c$ estão bem definidos.

Tem-se também, para qualquer $i = 1, \ldots, n$, que

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < 1,$$

concluindo-se imediatamente que $\|B\|_\infty < 1$.

Aplicando agora o resultado sobre convergência de métodos iterativos, pode afirmar-se que a equação $x = Bx + c$ tem uma e uma só solução $\bar{x}$, e também que o método de Jacobi converge para $\bar{x}$, qualquer que seja o ponto inicial $x_{(0)}$.

Este teorema fica demonstrado notando que a equação $x = Bx + c$ é equivalente a $Ax = b$, pelo que $\bar{x}$ é a única solução desta última equação. □

Como corolário deste resultado tem-se que toda a matriz quadrada estritamente diagonalmente dominante por linhas é não singular.

Este resultado, ao fornecer condições suficientes para a convergência do método de Jacobi, indica como proceder para garantir que a aplicação deste método fornecerá uma sucessão convergente. De facto, se a matriz $A$ dos coeficientes do sistema não for estritamente diagonalmente dominante por linhas não há garantia da convergência do método. Em tais situações dever-se-á proceder a uma prévia manipulação de $A$ de forma a satisfazer as condições de convergência. Esta manipulação pode passar pela troca de linhas da matriz (que corresponde à troca de ordem de equações), ou troca de colunas (que corresponde à troca da ordem das variáveis), ou ainda à realização de outras operações sobre a matriz que mantenham a equivalência do sistema de equações.

É também imediato concluir da validade da seguinte expressão para a majoração da norma do erro em $x_{(k+1)}$

$$\|\bar{x} - x_{(k+1)}\| \leq \frac{\|B\|_\infty}{1 - \|B\|_\infty} \|x_{(k+1)} - x_{(k)}\|.$$

**Exemplo 5.5.4.** *Aplicando o método de Jacobi, obter uma solução aproximada do sistema de equações, com um erro máximo absoluto em cada variável de* $5 \times 10^{-3}$.

$$\begin{cases} 4x_1 - 2x_2 + x_3 &= 3 \\ x_1 - x_2 + 3x_3 &= 3 \\ -x_1 + 3x_2 &= 2 \end{cases}$$

***Resolução***

*Uma vez que a matriz dos coeficientes não é estritamente diagonalmente dominante por linhas, torna-se necessário efectuar operações sobre a matriz previamente à aplicação do método. Assim, trocando a segunda equação com a terceira obtém-se o sistema equivalente*

$$\begin{bmatrix} 4 & -2 & 1 \\ -1 & 3 & 0 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix}$$

*cuja matriz de coeficientes já é estritamente diagonalmente dominante por linhas, garantindo a convergência do método de Jacobi.*

*A expressão de recorrência do método de Jacobi é* $x_{(k)} = Bx_{(k-1)} + c$*, tendo-se aqui que*

$$B = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix} \qquad e \qquad c = \begin{bmatrix} \frac{3}{4} \\ \frac{2}{3} \\ 1 \end{bmatrix}.$$

*Sendo* $e_{(k)}$ *o erro na iteração* $k$*, e uma vez que* $\|B\|_\infty = \frac{3}{4}$*, verifica-se a estimativa*

$$\|e_{(k)}\|_\infty \leq \frac{\frac{3}{4}}{1-\frac{3}{4}} \|x_{(k)} - x_{(k-1)}\|_\infty = 3 \|x_{(k)} - x_{(k-1)}\|_\infty$$

*Garantir um erro máximo absoluto em cada variável de* $5 \times 10^{-3}$ *na iteração* $k$ *é equivalente a ter* $\|e_{(k)}\|_\infty \leq 5 \times 10^{-3}$*. Para tal, bastará impor* $\varepsilon_k = 3 \|x_{(k)} - x_{(k-1)}\|_\infty \leq 5 \times 10^{-3}$*, que será a condição de paragem do método.*

*Partindo da condição inicial nula, obtiveram-se os resultados apresentados na tabela ao lado.*

*De acordo com a estimativa do erro, parou-se a aplicação do método assim que* $\varepsilon_k \leq 5 \times 10^{-3}$*.*

*A solução do sistema é* $x_1 = x_2 = x_3 = 1$*, obtendo-se na iteração* 10 *erros máximos absolutos em todas as variáveis inferiores a* $5 \times 10^{-4}$*, pelo que a estimativa do erro utilizada é, neste caso, algo conservadora.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ | $\varepsilon_k$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | – |
| 1 | 0.75000 | 0.66667 | 1.00000 | 3 |
| 2 | 0.83333 | 0.91667 | 0.97222 | $7.5 \times 10^{-1}$ |
| 3 | 0.96528 | 0.94444 | 1.02778 | $4.0 \times 10^{-1}$ |
| 4 | 0.96528 | 0.98843 | 0.99306 | $1.3 \times 10^{-1}$ |
| 5 | 0.99595 | 0.98843 | 1.00772 | $9.2 \times 10^{-2}$ |
| 6 | 0.99228 | 0.99865 | 0.99749 | $3.1 \times 10^{-2}$ |
| 7 | 0.99995 | 0.99743 | 1.00212 | $2.3 \times 10^{-2}$ |
| 8 | 0.99818 | 0.99998 | 0.99916 | $8.9 \times 10^{-3}$ |
| 9 | 1.00020 | 0.99939 | 1.00060 | $6.0 \times 10^{-3}$ |
| 10 | 0.99955 | 1.00007 | 0.99973 | $2.6 \times 10^{-3}$ |

Passemos agora ao método de Gauss-Seidel. O teorema seguinte fornece condições de convergência para este método.

**Teorema 5.5.3.** *Sejam $A \in \mathbb{R}^{n\times n}$ e $b \in \mathbb{R}^n$. Se a matriz $A$ for estritamente diagonalmente dominante por linhas então a sucessão gerada pelo método de Gauss-Seidel converge para a única solução do sistema de equações $Ax = b$, qualquer que seja o ponto inicial $x_{(0)}$.*

Estas condições de convergência do método de Gauss-Seidel são semelhantes às apresentadas para o método de Jacobi. Tal como então, trata-se apenas de condições suficientes, ou seja, há situações em que estas condição não se verificam e o método de Gauss-Seidel converge.

A análise aqui apresentada não permite concluir qual dos métodos (Jacobi ou Gauss-Seidel) possui uma convergência mais rápida. Contudo, é frequente o método de Gauss-Seidel convergir mais rapidamente que o método de Jacobi.

**Exemplo 5.5.5.** *Aplicando o método de Gauss-Seidel, obter uma solução aproximada do sistema de equações. Terminar o método assim que a diferença entre duas estimativas consecutivas seja inferior ou igual a $10^{-3}$, em todas as variáveis.*

$$\begin{cases} x_1 - 4x_3 & = -3 \\ 4x_2 - 2x_3 & = 2 \\ 4x_1 - 2x_2 & = 2 \end{cases}$$

***Resolução***

*A matriz dos coeficientes do sistema não é estritamente diagonalmente dominante por linhas. No entanto, trocando a primeira equação com a terceira obtém-se o sistema equivalente*

$$\begin{bmatrix} 4 & 0 & -2 \\ 0 & 4 & -2 \\ 1 & 0 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ -3 \end{bmatrix}$$

*cuja matriz de coeficientes é estritamente diagonalmente dominante por linhas, condição suficiente para a convergência do método de Gauss-Seidel.*

*As expressões de recorrência serão*

$$\begin{cases} x_{1,(k)} = \frac{1}{2}x_{3,(k-1)} & + & \frac{1}{2} \\ x_{2,(k)} = \frac{1}{2}x_{3,(k-1)} & + & \frac{1}{2} \\ x_{3,(k)} = \frac{1}{4}x_{1,(k)} & + & \frac{3}{4} \end{cases}$$

*sendo a condição de paragem definida por $\|x_{(k)} - x_{(k-1)}\|_\infty \leq 10^{-3}$.*

*Partindo da condição inicial nula, obtêm-se os resultados apresentados na tabela seguinte.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ | $\|x_{(k)} - x_{(k-1)}\|_\infty$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | – |
| 1 | 0.50000 | 0.50000 | 0.87500 | $8.8 \times 10^{-1}$ |
| 2 | 0.93750 | 0.93750 | 0.98438 | $4.4 \times 10^{-1}$ |
| 3 | 0.99219 | 0.99219 | 0.99805 | $5.5 \times 10^{-2}$ |
| 4 | 0.99902 | 0.99902 | 0.99976 | $6.8 \times 10^{-3}$ |
| 5 | 0.99988 | 0.99988 | 0.99997 | $8.5 \times 10^{-4}$ |

## 5.6 Relaxação dos métodos de Jacobi e Gauss-Seidel

A expressão de recorrência do método de Jacobi é

$$x_{i,(k+1)} = \frac{1}{a_{ii}} \left[ b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_{j,(k)} \right]$$

que pode ainda ser escrita na forma

$$x_{i,(k+1)} = x_{i,(k)} + \frac{1}{a_{ii}} \left[ b_i - \sum_{j=1}^{n} a_{ij} x_{j,(k)} \right],$$

a qual evidencia que cada nova estimativa de $x_i$ é obtida somando à estimativa anterior um dado valor, que não é mais do que o resíduo da equação $i$ dividido pelo termo $a_{ii}$.

A relaxação do método de Jacobi consiste em tomar uma constante $\omega > 0$ e utilizar a expressão de recorrência

$$x_{i,(k+1)} = x_{i,(k)} + \omega \cdot \frac{1}{a_{ii}} \left[ b_i - \sum_{j=1}^{n} a_{ij} x_{j,(k)} \right].$$

com o objectivo de alterar a convergência do método.

No caso do método de Gauss-Seidel, a expressão de recorrência

$$x_{i,(k+1)} = \frac{1}{a_{ii}} \left[ b_i - \sum_{j=1}^{i-1} a_{ij} x_{j,(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_{j,(k)} \right]$$

pode também tomar a forma

$$x_{i,(k+1)} = x_{i,(k)} + \frac{1}{a_{ii}} \left[ b_i - \sum_{j=1}^{i-1} a_{ij} x_{j,(k+1)} - \sum_{j=i}^{n} a_{ij} x_{j,(k)} \right].$$

A relaxação deste método é de igual modo efectuada considerando um valor $\omega > 0$ e utilizando agora a expressão de recorrência

$$x_{i,(k+1)} = x_{i,(k)} + \omega \cdot \frac{1}{a_{ii}} \left[ b_i - \sum_{j=1}^{i-1} a_{ij} x_{j,(k+1)} - \sum_{j=i}^{n} a_{ij} x_{j,(k)} \right].$$

De uma maneira geral, a consideração de $\omega < 1$, designada por **sub-relaxação**, permite tornar convergente um método que inicialmente não o seria, enquanto a consideração de $\omega > 1$, designada por **sobre-relaxação**, permite acelerar a convergência de um método, podendo contudo torná-lo divergente!

A sobre-relaxação do método de Gauss-Seidel é habitualmente designada por **método das sobre-relaxações sucessivas** (successive over relaxation – SOR) sendo bastante utilizada na resolução de sistemas de equações lineares por métodos iterativos. Efectivamente, em muitas situações é possível determinar, em função da matriz $A$ de coeficientes do sistema, o valor óptimo do parâmetro $\omega$ que conduz a uma maior rapidez de convergência.

**Exemplo 5.6.1.** *Compare o desempenho dos métodos de Gauss-Seidel e SOR com* $\omega = 1.25$ *na resolução do sistema de equações*

$$\begin{cases} 4x_1 + 3x_2 & = 24 \\ 3x_1 + 4x_2 - x_3 & = 30 \\ -x_2 + 4x_3 & = -24 \end{cases}$$

*cuja solução é* $x_1 = 3$, $x_2 = 4$, $x_3 = -5$. *Em ambos os casos partir de* $x_{1,(0)} = x_{2,(0)} = x_{3,(0)} = 1$ *e efectuar* 8 *iterações.*

***Resolução***

*As expressões de recorrência do método de Gauss-Seidel são*

$$\begin{aligned} x_{1,(k+1)} &= x_{1,(k)} + \frac{1}{4}\left[24 - 4x_{1,(k)} - 3x_{2,(k)}\right] \\ x_{2,(k+1)} &= x_{2,(k)} + \frac{1}{4}\left[30 - 3x_{1,(k+1)} - 4x_{2,(k)} + x_{3,(k)}\right] \\ x_{3,(k+1)} &= x_{3,(k)} + \frac{1}{4}\left[-24 + x_{2,(k+1)} - 4x_{3,(k)}\right] \end{aligned}$$

*Partindo* $x_{1,(0)} = x_{2,(0)} = x_{3,(0)} = 1$ *obtêm-se os resultados apresentados na tabela seguinte.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 1.00000 | 1.00000 | 1.00000 |
| 1 | 5.25000 | 3.81250 | −5.04688 |
| 2 | 3.14063 | 3.88281 | −5.02930 |
| 3 | 3.08789 | 3.92676 | −5.01831 |
| 4 | 3.05493 | 3.95422 | −5.01144 |
| 5 | 3.03433 | 3.97139 | −5.00715 |
| 6 | 3.02146 | 3.98212 | −5.00447 |
| 7 | 3.01341 | 3.98882 | −5.00279 |
| 8 | 3.00838 | 3.99302 | −5.00175 |

*As expressões de recorrência do método SOR com* $\omega = 1.25$ *são*

$$x_{1,(k+1)} = x_{1,(k)} + \frac{1.25}{4}\left[24 - 4x_{1,(k)} - 3x_{2,(k)}\right]$$
$$x_{2,(k+1)} = x_{2,(k)} + \frac{1.25}{4}\left[30 - 3x_{1,(k+1)} - 4x_{2,(k)} + x_{3,(k)}\right]$$
$$x_{3,(k+1)} = x_{3,(k)} + \frac{1.25}{4}\left[-24 + x_{2,(k+1)} - 4x_{3,(k)}\right]$$

*Partindo* $x_{1,(0)} = x_{2,(0)} = x_{3,(0)} = 1$ *obtêm-se os resultados apresentados na tabela seguinte.*

| $k$ | $x_{1,(k)}$ | $x_{2,(k)}$ | $x_{3,(k)}$ |
|---|---|---|---|
| 0 | 1.00000 | 1.00000 | 1.00000 |
| 1 | 6.10000 | 3.61000 | −6.31700 |
| 2 | 2.73100 | 3.92500 | −4.75910 |
| 3 | 3.12130 | 3.97810 | −5.05475 |
| 4 | 2.99545 | 3.99205 | −4.99144 |
| 5 | 3.00807 | 3.99690 | −5.00264 |
| 6 | 3.00118 | 3.99877 | −4.99984 |
| 7 | 3.00087 | 3.99951 | −5.00018 |
| 8 | 3.00027 | 3.99980 | −5.00002 |

*Comparando os resultados constata-se facilmente que a sucessão produzida pelo método SOR converge muito mais rapidamente para a solução do problema.*

# Capítulo 6

# Aproximação dos Mínimos Quadrados

## 6.1 Introdução

O problema de aproximação que será estudado neste capítulo pode ser descrito como se segue. Dado um conjunto de pares ordenados $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$, pretende-se determinar uma função **aproximante** $g$ tal que $g(x_i)$ **seja próximo** de $y_i$, para $i = 1, 2, \ldots, n$.

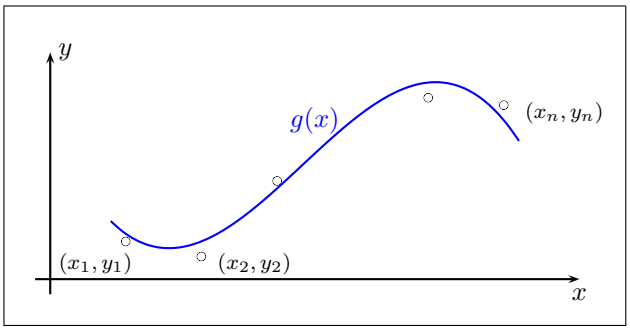



Figura 6.1: Aproximação.

É de notar que contrariamente ao problema de interpolação, no qual se pretendia determinar uma função que tomasse valores bem definidos num conjunto de pontos determinados, no problema de aproximação apenas se exige que os valores da função aproximante estejam próximos dos valores dados. Muitas vezes os valores $y_i$ estão afectados por erros, não fazendo muito sentido "obrigar" a função $g$ a satisfazer as condições $g(x_i) = y_i$.

De um modo semelhante ao que se passa com os problemas de interpolação, ao tratar um problema de aproximação será necessário abordar algumas questões tais como a escolha da classe de funções aproximantes a utilizar, o critério de aproximação que permitirá seleccionar a

"melhor" função aproximante dentro da classe de funções consideradas e ainda a forma de obter tal função, partindo dos dados do problema.

## 6.2 Funções aproximantes e desvios

De um modo geral, poderemos dizer que a classe de funções aproximantes estará parametrizada por um conjunto de valores $c_1, c_2, \ldots, c_k$, isto é, toda função aproximante $g$ poderá ser escrita na forma

$$g(x) = F(x; c_1, c_2, \ldots, c_k).$$

Assim, a resolução de um dado problema de aproximação consistirá em determinar os valores $c_1, c_2, \ldots, c_k$ que definem a função que melhor aproxima os dados $(x_i, y_i)_{i=1}^n$, de acordo com um dado critério definido à partida.

Por exemplo, se se pretender aproximar os pontos dados por uma linha recta será natural que a classe de funções aproximantes seja da forma

$$F(x; c_1, c_2) = c_1 + c_2 x,$$

sendo $c_1$ e $c_2$ os valores a determinar; se se pretender aproximar os pontos por uma parábola, teremos

$$F(x; c_1, c_2) = c_1 + c_2 x + c_3 x^2,$$

sendo agora $c_1$, $c_2$ e $c_3$os valores a determinar.

O critério de selecção da melhor função deverá traduzir o maior ou menor grau de aproximação dos valores da função aproximante aos valores dados. Desta forma, para cada conjunto de valores $c_1, c_2, \ldots, c_k$ definem-se os **desvios** como sendo as diferenças entre cada um dos valores $y_i$ dados e o respectivo valor da função aproximante $F(x_i; c_1, c_2, \ldots, c_k)$, isto é,

$$d_i = y_i - F(x_i; c_1, c_2, \ldots, c_k), \qquad i = 1, 2, \ldots, n.$$

Será então natural que o critério de comparação de funções aproximantes que permite decidir qual delas é melhor seja baseado nestes desvios. Este critério deverá conduzir a funções aproximantes que tornem tais desvios "pequenos", em valor absoluto. Alguns critérios possíveis serão

1. minimizar $\displaystyle\sum_{i=1}^{n} |d_i|$
2. minimizar $\displaystyle\max_{1\leq i\leq n} |d_i|$
3. minimizar $\displaystyle\sum_{i=1}^{n} d_i^2$

É de referir que em qualquer dos casos a minimização consistirá em encontrar o conjunto de valores $c_1, c_2, \ldots, c_k$ que tornem mínimo o critério em causa, pois os desvios considerados, e logo a função a minimizar, dependem destes parâmetros.

Os dois primeiros critérios acima apresentados conduzem, de um modo geral, à resolução de sistemas de equações não lineares para determinar os parâmetros que definem a melhor função aproximante. Tal facto constitui assim uma desvantagem destes critérios.

No terceiro caso, a determinação da melhor função é efectuada resolvendo um sistema de equações lineares nos parâmetros $c_1, \ldots, c_k$, sempre que a classe de funções aproximantes seja definida por

$$F(x; c_1, c_2, \ldots, c_k) = c_1\phi_1(x) + c_2\phi_2(x) \cdots + c_k\phi_k(x),$$

onde $\phi_1(x), \phi_2(x), \ldots, \phi_k(x)$ são funções dadas. Neste caso, temos o designado **métodos dos mínimos quadrados**, que será estudado nas secções seguintes.

## 6.3 Aproximação dos mínimos quadrados

Dados os pares $(x_i, y_i)$, com $i = 1, \ldots, n$, e as funções $\phi_1, \ldots, \phi_k$, a aproximação dos mínimos quadrados consiste em determinar os parâmetros $c_1, \ldots, c_k$ que tornam mínima a quantidade

$$e(c_1, \ldots, c_k) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - ( \, c_1\phi_1(x_i) + \ldots + c_k\phi_k(x_i) \, )]^2 = \sum_{i=1}^{n} \left[ y_i - \sum_{l=1}^{k} c_l\phi_l(x_i) \right]^2$$

ou seja, que minimizam a soma dos quadrados dos desvios. Este é um problema de minimização em $\mathbb{R}^k$. Não se colocando qualquer restrição sobre os parâmetros, para que $e(c_1, \ldots, c_k)$ seja mínimo será necessário que

$$\nabla e = 0,$$

ou, equivalentemente,

$$\frac{\partial e}{\partial c_j} = 0, \quad j = 1, \ldots, k.$$

Calculando estas derivadas parciais obtém-se

$$\begin{aligned}
\frac{\partial e}{\partial c_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial c_j} \left( y_i - \sum_{l=1}^{k} c_l\phi_l(x_i) \right)^2 = \sum_{i=1}^{n} \left[ (-2) \left( y_i - \sum_{l=1}^{k} c_l\phi_l(x_i) \right) \cdot \frac{\partial}{\partial c_j} \sum_{l=1}^{k} c_l\phi_l(x_i) \right] \\
&= -2 \sum_{i=1}^{n} \left[ \left( y_i - \sum_{l=1}^{k} c_l\phi_l(x_i) \right) \cdot \phi_j(x_i) \right] = -2 \left[ \sum_{i=1}^{n} y_i\phi_j(x_i) - \sum_{i=1}^{n} \sum_{l=1}^{k} c_l\phi_l(x_i)\phi_j(x_i) \right] \\
&= -2 \left[ \sum_{i=1}^{n} y_i\phi_j(x_i) - \sum_{l=1}^{k} c_l \sum_{i=1}^{n} \phi_l(x_i)\phi_j(x_i) \right]
\end{aligned}$$

E então, como se pretende ter $\frac{\partial e}{\partial c_j} = 0$, resulta

$$\sum_{l=1}^{k} c_l \sum_{i=1}^{n} \phi_l(x_i)\phi_j(x_i) = \sum_{i=1}^{n} y_i\phi_j(x_i),$$

obtendo-se, finalmente, o sistema de equações

$$\begin{cases} c_1 \sum_{i=1}^{n} \phi_1^2(x_i) + c_2 \sum_{i=1}^{n} \phi_1(x_i)\phi_2(x_i) + \cdots + c_k \sum_{i=1}^{n} \phi_1(x_i)\phi_k(x_i) = \sum_{i=1}^{n} y_i \phi_1(x_i) \\ c_1 \sum_{i=1}^{n} \phi_2(x_i)\phi_1(x_i) + c_2 \sum_{i=1}^{n} \phi_2^2(x_i) + \cdots + c_k \sum_{i=1}^{n} \phi_2(x_i)\phi_k(x_i) = \sum_{i=1}^{n} y_i \phi_2(x_i) \\ \cdots\cdots\cdots \\ c_1 \sum_{i=1}^{n} \phi_k(x_i)\phi_1(x_i) + c_2 \sum_{i=1}^{n} \phi_k(x_i)\phi_2(x_i) + \cdots + c_k \sum_{i=1}^{n} \phi_k^2(x_i) = \sum_{i=1}^{n} y_i \phi_k(x_i) \end{cases}$$

Este sistema de $k$ equações lineares em $k$ incógnitas permite obter as constantes $c_1, \ldots, c_k$ que caracterizam a melhor função aproximante no sentido dos mínimos quadrados. Vamos para já supor que este sistema tem solução única. A análise de existência e unicidade de solução deste sistema será abordada mais tarde.

Se pretendermos aproximar os pontos por uma recta, as funções aproximantes serão da forma $g(x) = c_1 + c_2 x$. Teremos então $k = 2$ e as funções $\phi_1(x) = 1$ e $\phi_2(x) = x$. Neste caso, o sistema de equações a resolver toma a forma

$$\begin{cases} c_1 \sum_{i=1}^{n} 1 + c_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ c_1 \sum_{i=1}^{n} x_i + c_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Estes somatórios são facilmente determinados organizando os cálculos numa tabela como a seguinte.

| $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ |
|---|---|---|---|
| $x_1$ | $y_1$ | $x_1^2$ | $x_1 y_1$ |
| $x_2$ | $y_2$ | $x_2^2$ | $x_2 y_2$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $x_n$ | $y_n$ | $x_n^2$ | $x_n y_n$ |
| $\sum x_i$ | $\sum y_i$ | $\sum x_i^2$ | $\sum x_i y_i$ |

Se se pretender efectuar uma aproximação por uma parábola, as funções aproximantes serão da forma $g(x) = c_1 + c_2 x + c_3 x^2$. Então, dever-se-á ter $k = 3$ e as funções $\phi_1(x) = 1$, $\phi_2(x) = x$ e $\phi_3(x) = x^2$. O sistema de equações a resolver é o seguinte.

$$\begin{cases} c_1 \sum_{i=1}^{n} 1 + c_2 \sum_{i=1}^{n} x_i + c_3 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ c_1 \sum_{i=1}^{n} x_i + c_2 \sum_{i=1}^{n} x_i^2 + c_3 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ c_1 \sum_{i=1}^{n} x_i^2 + c_2 \sum_{i=1}^{n} x_i^3 + c_3 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

**Exemplo 6.3.1.** *Determine a aproximação dos mínimos quadrados aos pontos da tabela.*

| $x$ | 1 | 2 | 4 | 5 | 7 | 8 | 10 |
|---|---|---|---|---|---|---|---|
| $y$ | 1 | 2 | 4 | 4 | 5 | 6 | 7 |

*a) Por uma recta.*

*b) Por uma parábola.*

*c) Por uma recta que minimize o erro em $x$.*

***Resolução***

*a) A função aproximante será $F(x) = c_1 + c_2 x$, sendo $c_1$ e $c_2$ calculados resolvendo o sistema*

$$\begin{cases} c_1 \sum_{i=1}^{7} 1 + c_2 \sum_{i=1}^{7} x_i = \sum_{i=1}^{7} y_i \\ c_1 \sum_{i=1}^{7} x_i + c_2 \sum_{i=1}^{7} x_i^2 = \sum_{i=1}^{7} x_i y_i \end{cases}$$

*Na tabela abaixo encontram-se os cálculos necessários à completa definição deste sistema de equações.*

| | $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ |
|---|---|---|---|---|
| | 1 | 1 | 1 | 1 |
| | 2 | 2 | 4 | 4 |
| | 4 | 4 | 16 | 16 |
| | 5 | 4 | 25 | 20 |
| | 7 | 5 | 49 | 35 |
| | 8 | 6 | 64 | 48 |
| | 10 | 7 | 100 | 70 |
| $\sum$ | 37 | 29 | 259 | 194 |

*O sistema a resolver será*

$$\begin{cases} 7c_1 + 37c_2 = 29 \\ 37c_1 + 259c_2 = 194 \end{cases}$$

*resultando $c_1 = 0.75$ e $c_2 = 0.64189$. A recta aproximante será então*

$$y = 0.75 + 0.64189x.$$

*b) A função aproximante será $F(x) = c_1 + c_2 x + c_3 x^2$, sendo $c_1$, $c_2$ e $c_3$ determinados por*

$$\begin{cases} c_1 \sum_i 1 + c_2 \sum_i x_i + c_3 \sum_i x_i^2 = \sum_i y_i \\ c_1 \sum_i x_i + c_2 \sum_i x_i^2 + c_3 \sum_i x_i^3 = \sum_i x_i y_i \\ c_1 \sum_i x_i^2 + c_2 \sum_i x_i^3 + c_3 \sum_i x_i^4 = \sum_i x_i^2 y_i \end{cases}$$

*Os coeficientes do sistema determinam-se a partir dos cálculos expostos na seguinte tabela.*

| | $x_i$ | $y_i$ | $x_i^2$ | $x_i^3$ | $x_i^4$ | $x_i y_i$ | $x_i^2 y_i$ |
|---|---|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 2 | 2 | 4 | 8 | 16 | 4 | 8 |
| | 4 | 4 | 16 | 64 | 256 | 16 | 64 |
| | 5 | 4 | 25 | 125 | 625 | 20 | 100 |
| | 7 | 5 | 49 | 343 | 2401 | 35 | 245 |
| | 8 | 6 | 64 | 512 | 4096 | 48 | 384 |
| | 10 | 7 | 100 | 1000 | 10000 | 70 | 700 |
| $\sum$ | 37 | 29 | 259 | 2053 | 17395 | 194 | 1502 |

*O sistema de equações a resolver será então*

$$\begin{cases} 7c_1 + 37c_2 + 259c_3 = 29 \\ 37c_1 + 259c_2 + 2053c_3 = 194 \\ 259c_1 + 2053c_2 + 17395c_3 = 1502 \end{cases}$$

*resultando* $c_1 = 0.288690$, $c_2 = 0.890625$ *e* $c_3 = -0.023065$. *A parábola que aproxima os pontos será portanto*

$$y = 0.288690 - 0.890625x + 0.023065x^2.$$

*c) Agora tem-se* $F(y) = c_1 + c_2 y$. *As constantes* $c_1$ *e* $c_2$ *são calculadas resolvendo o sistema*

$$\begin{cases} c_1 \sum_{i=1}^{7} 1 + c_2 \sum_{i=1}^{7} y_i = \sum_{i=1}^{7} x_i \\ c_1 \sum_{i=1}^{7} y_i + c_2 \sum_{i=1}^{7} y_i^2 = \sum_{i=1}^{7} y_i x_i \end{cases}$$

*Os cálculos necessários à definição deste sistema apresentam-se na tabela seguinte.*

| | $x_i$ | $y_i$ | $y_i^2$ | $y_i x_i$ |
|---|---|---|---|---|
| | 1 | 1 | 1 | 1 |
| | 2 | 2 | 4 | 4 |
| | 4 | 4 | 16 | 16 |
| | 5 | 4 | 16 | 20 |
| | 7 | 5 | 25 | 35 |
| | 8 | 6 | 36 | 48 |
| | 10 | 7 | 49 | 70 |
| $\sum$ | 37 | 29 | 147 | 194 |

*Assim, o sistema de equações a resolver será*

$$\begin{cases} 7c_1 + 29c_2 = 37 \\ 29c_1 + 147c_2 = 194 \end{cases}$$

*do qual resultam os valores* $c_1 = -0.99468$ *e* $c_2 = 1.51596$. *Agora, a recta aproximante será*

$$x = -0.99468 + 1.51596y.$$

## 6.4 Redução a problemas de mínimos quadrados

Por vezes interessa considerar funções aproximantes $F(x; c_1, \ldots, c_k)$ que não podem ser escritas na forma $F(x; c_1, \ldots, c_k) = c_1\phi_1(x) + \ldots + c_k\phi_k(x)$, isto é, como uma combinação linear de funções dadas. Nestes casos, a aplicação do método dos mínimos quadrados para a determinação dos parâmetros $c_1, \ldots, c_k$ que definem a melhor função aproximante resulta na resolução de um sistema de equações não lineares.

Contudo, em diversas situações é possível transformar estes problemas em outros cuja resolução seja mais fácil. Considere-se então que a classe de funções aproximantes é da forma

$$F(x; c_1, \ldots, c_k),$$

onde $c_1, \ldots, c_k$ são os parâmetros a determinar e suponha-se que existe uma função $g$ tal que

$$g(F(x; c_1, \ldots, c_k)) = b_1\phi_1(x) + \cdots + b_k\phi_k(x),$$

onde $\phi_1, \ldots, \phi_k$ são funções conhecidas, e os parâmetros $b_1, \ldots, b_k$ se relacionam com os parâmetros $c_1, \ldots, c_k$ por intermédio das relações $b_1 = \psi_1(c_1), \ldots, b_k = \psi_k(c_k)$, para funções $\psi_1, \ldots, \psi_k$, também conhecidas. Isto equivale a transformar a classe de funções dada numa outra em que os parâmetros a determinar sejam os coeficientes de uma combinação linear de funções conhecidas.

Nestas situações, é possível determinar a função aproximante que minimiza a soma dos quadrados dos **desvios modificados** resolvendo um sistema de equações lineares. Estes desvios modificados definem-se por

$$g(y_i) - g(F(x_i; c_1, \ldots, c_k)) = g(y_i) - [b_1\phi_1(x_i) + \cdots + b_k\phi_k(x_i)].$$

O problema que se está agora a resolver consiste em determinar a função

$$b_1\phi_1(x) + \cdots + b_k\phi_k(x)$$

que melhor aproxima os pontos $(x_i, g(y_i))$ no sentido dos mínimos quadrados. Este problema reduz-se à resolução de uma sistema de equações lineares nos parâmetros $b_1, \ldots, b_k$. Uma vez resolvido este problema será necessário determinar os parâmetros $c_1, \ldots, c_k$ que caracterizam a função aproximante pretendida. Para tal serão utilizadas as relações

$$c_j = \psi_j^{-1}(b_j), \qquad j = 1, \ldots, k.$$

Refira-se que esta abordagem de determinação da melhor função aproximante não permite determinar a função que minimiza a soma dos quadrados dos desvios, mas sim a soma dos quadrados dos desvios modificados, perdendo-se algum do significado do conceito de melhor função aproximante. Todavia, a vantagem obtida com a redução do problema original à simples resolução de um sistema de equações lineares compensa em muitas situações tal perda. Claro está que

a principal dificuldade desta abordagem está na determinação da função $g$, isto é, da transformação dos dados do problema que permite obter a classe de funções aproximantes como uma combinação linear de funções conhecidas.

**Exemplo 6.4.1.** *Aproximar por uma função da forma $y = ax^b$ os pontos*

| $x$ | 1 | 1.2 | 1.6 | 2 |
|---|---|---|---|---|
| $y$ | 1 | 1.3 | 1.4 | 1.7 |

***Resolução***

*Aplicando uma transformação logarítmica aos valores $y = ax^b$ obtém-se*

$$\ln(y) = \ln(a) + b\ln(x).$$

*Desta forma, minimizando a soma dos quadrados dos desvios dos logaritmos de $y$, obtém-se um problema cuja solução é determinada resolvendo um sistema de equações lineares. Para tal basta tomar $\phi_1(x) = 1$ e $\phi_2(x) = \ln(x)$. Agora tem-se que $\ln(y) = b_1\phi_1(x) + b_2\phi_2(x)$, sendo $b_1 = \ln(a)$ e $b_2 = b$ as constantes a determinar.*

*As constantes $b_1$ e $b_2$ são calculadas resolvendo as equações*

$$\begin{cases} b_1 \sum_{i=1}^4 1 + b_2 \sum_{i=1}^4 \ln(x_i) = \sum_{i=1}^4 \ln(y_i) \\ b_1 \sum_{i=1}^4 \ln(x_i) + b_2 \sum_{i=1}^4 \ln^2(x_i) = \sum_{i=1}^4 \ln(y_i)\ln(x_i) \end{cases}$$

*Calculando os somatórios indicados, obtém-se o sistema*

$$\begin{cases} 4b_1 + 1.34547b_2 = 1.12946 \\ 1.34547b_1 + 0.73460b_2 = 0.57378 \end{cases}$$

*cuja solução é $b_1 = 0.05144$ e $b_2 = 0.68741$. Então $a = e^{b_1} = 1.05247$ e $b = b_2 = 0.68741$. A função aproximante será*

$$y = 1.05247x^{0.68741}.$$

*Na tabela seguinte apresentam-se os valores de $y$ dados, bem como os valores obtidos com a função aproximante determinada.*

| $x$ | 1 | 1.2 | 1.6 | 2 |
|---|---|---|---|---|
| $y$ | 1 | 1.3 | 1.4 | 1.7 |
| $1.05247x^{0.68741}$ | 1.052 | 1.193 | 1.454 | 1.695 |

## 6.5 Aproximação em espaços vectoriais e mínimos quadrados

O problema de aproximação dos mínimos quadrados que temos vindo a abordar pode ser incluído num problema mais geral de aproximação em espaços vectoriais. Esta inclusão permite não

só perspectivar extensões do problema considerado, bem como sistematizar o estudo de tais problemas.

Consideremos então um espaço vectorial real $V$ no qual se encontra definido um produto interno que representaremos por $\langle \cdot, \cdot \rangle$. Seja ainda $\| \cdot \|$ a norma em $V$ induzida pelo produto interno considerado, isto é,

$$\|v\| = \sqrt{\langle v, v \rangle}\,, \qquad v \in V.$$

Tomemos um conjunto de vectores de $V$, $\{v_1, v_2, \ldots, v_k\}$, que por simplicidade de tratamento suporemos linearmente independentes. Seja ainda $u$ um qualquer vector de $V$ e consideremos o problema de determinar a combinação linear $c_1v_1 + c_2v_2 + \cdots + c_kv_k$ que melhor aproxima $u$ no sentido de tornar mínimo

$$\|u - (c_1v_1 + c_2v_2 + \cdots + c_kv_k)\|^2.$$

Este problema mais não é do que o de determinar o elemento do subespaço de $V$ gerado pelos vectores $v_1, v_2, \ldots, v_k$ que se encontra mais próximo do vector $u$. Uma vez que os elementos de tal subespaço se encontram parametrizados por $c_1, c_2, \ldots, c_k$, trata-se de um problema de minimização em $\mathbb{R}^k$.

O teorema seguinte estabelece um conjunto de condições que têm de ser satisfeitas pelo elemento minimizante, as quais permitirão determinar tal elemento.

**Teorema 6.5.1.** *Considere-se o conjunto $\{v_1, v_2, \ldots, v_k\}$ de vectores de $V$ linearmente independentes e um vector $u \in V$. A combinação linear $c_1v_1 + c_2v_2 + \cdots + c_kv_k$ que torna mínimo o valor*

$$\|u - (c_1v_1 + c_2v_2 + \cdots + c_kv_k)\|^2$$

*satisfaz as relações*

$$\langle v_j, u - (c_1v_1 + c_2v_2 + \cdots + c_kv_k) \rangle = 0, \qquad j = 1, 2, \ldots, k.$$

Este resultado tem uma interpretação geométrica bastante simples, pois afirma que a diferença entre o vector $u$ e a combinação linear $c_1v_1 + c_2v_2 + \cdots + c_kv_k$ que melhor o aproxima é ortogonal a cada um dos vectores $v_1, v_2, \ldots, v_k$ que geram o subespaço considerado.

Destas relações de ortogonalidade é possível concluir que

$$\langle v_j, u \rangle - \langle v_j, c_1v_1 + c_2v_2 + \cdots + c_kv_k \rangle = 0$$

ou seja,

$$c_1\langle v_j, v_1 \rangle + c_2\langle v_j, v_2 \rangle + \ldots + c_k\langle v_j, v_k \rangle = \langle v_j, u \rangle$$

para $j = 1, 2, \ldots, k$. Obtém-se então o seguinte sistema de $k$ equações lineares nas $k$ incógnitas

$c_1, c_2, \ldots, c_k$.

$$\begin{bmatrix} \langle v_1, v_1\rangle & \langle v_1, v_2\rangle & \ldots & \langle v_1, v_k\rangle \\ \langle v_2, v_1\rangle & \langle v_2, v_2\rangle & \ldots & \langle v_2, v_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_k, v_1\rangle & \langle v_k, v_2\rangle & \ldots & \langle v_k, v_k\rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} \langle v_1, u\rangle \\ \langle v_2, u\rangle \\ \vdots \\ \langle v_k, u\rangle \end{bmatrix}.$$

Habitualmente estas equações são designadas por **equações normais**. Este sistema de equações tem solução única uma vez que se está a supor que os vectores $v_1, v_2, \ldots, v_k$ são linearmente independentes.

Voltemos agora ao problema original de aproximação dos mínimos quadrados que consiste em determinar a combinação linear

$$c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_k\phi_k(x),$$

que minimiza a soma dos quadrados dos desvios relativos aos pares $(x_i, y_i)_{i=1}^n$.

Considerem-se os vectores de $\mathbb{R}^n$, $\bar{\phi}_1, \bar{\phi}_2, \ldots, \bar{\phi}_k$ e $\bar{y}$ definidos por

$$\bar{\phi}_1 = \begin{bmatrix} \phi_1(x_1) \\ \phi_1(x_2) \\ \vdots \\ \phi_1(x_n) \end{bmatrix}, \ \bar{\phi}_2 = \begin{bmatrix} \phi_2(x_1) \\ \phi_2(x_2) \\ \vdots \\ \phi_2(x_n) \end{bmatrix}, \ \ldots, \ \bar{\phi}_k = \begin{bmatrix} \phi_k(x_1) \\ \phi_k(x_2) \\ \vdots \\ \phi_k(x_n) \end{bmatrix}, \text{ e } \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Consideremos o produto interno usual definido em $\mathbb{R}^n$ e a respectiva norma induzida, isto é,

$$\langle u, v\rangle = \sum_{i=1}^n u_i v_i, \quad \forall u, v \in \mathbb{R}^n$$

$$\|u\| = \sqrt{\langle u, u\rangle} = \left(\sum_{i=1}^n u_i^2\right)^{1/2}, \quad \forall u \in \mathbb{R}^n.$$

O problema de aproximação dos mínimos quadrados é então equivalente ao problema de determinar a combinação linear $c_1\bar{\phi}_1 + c_2\bar{\phi}_2 + \cdots + c_k\bar{\phi}_k$ que torna mínimo o valor

$$\|\bar{y} - (c_1\bar{\phi}_1 + c_2\bar{\phi}_2 + \cdots + c_k\bar{\phi}_k)\|^2.$$

Trata-se então de um problema de aproximação em espaços vectoriais como o acima apresentado. Desta forma, conclui-se que os valores $c_1, c_2, \ldots, c_k$, que caracterizam a solução do problema, são determinados resolvendo o seguinte sistema de equações.

$$\begin{bmatrix} \langle \bar{\phi}_1, \bar{\phi}_1\rangle & \langle \bar{\phi}_1, \bar{\phi}_2\rangle & \ldots & \langle \bar{\phi}_1, \bar{\phi}_k\rangle \\ \langle \bar{\phi}_2, \bar{\phi}_1\rangle & \langle \bar{\phi}_2, \bar{\phi}_2\rangle & \ldots & \langle \bar{\phi}_2, \bar{\phi}_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \bar{\phi}_k, \bar{\phi}_1\rangle & \langle \bar{\phi}_k, \bar{\phi}_2\rangle & \ldots & \langle \bar{\phi}_k, \bar{\phi}_k\rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} \langle \bar{\phi}_1, \bar{y}\rangle \\ \langle \bar{\phi}_2, \bar{y}\rangle \\ \vdots \\ \langle \bar{\phi}_k, \bar{y}\rangle \end{bmatrix}.$$

Atendendo ao produto interno considerado em $\mathbb{R}^n$, este sistema de equações toma a forma

$$\begin{bmatrix} \sum_{i=1}^{n} \phi_1(x_i)\phi_1(x_i) & \sum_{i=1}^{n} \phi_1(x_i)\phi_2(x_i) & \dots & \sum_{i=1}^{n} \phi_1(x_i)\phi_k(x_i) \\ \sum_{i=1}^{n} \phi_2(x_i)\phi_1(x_i) & \sum_{i=1}^{n} \phi_2(x_i)\phi_2(x_i) & \dots & \sum_{i=1}^{n} \phi_2(x_i)\phi_k(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} \phi_k(x_i)\phi_1(x_i) & \sum_{i=1}^{n} \phi_k(x_i)\phi_2(x_i) & \dots & \sum_{i=1}^{n} \phi_k(x_i)\phi_k(x_i) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \phi_1(x_i)y_i \\ \sum_{i=1}^{n} \phi_2(x_i)y_i \\ \vdots \\ \sum_{i=1}^{n} \phi_k(x_i)y_i \end{bmatrix}$$

que mais não é do que aquele anteriormente obtido.

Esta abordagem da aproximação dos mínimos quadrados permite agora analisar facilmente a questão da existência e unicidade de solução. Assim, pode concluir-se que este problema tem solução única se os vectores $\bar{\phi}_1, \bar{\phi}_2, \dots, \bar{\phi}_k$, atrás definidos, forem linearmente independentes. Neste caso diz-se que as funções $\phi_1, \dots, \phi_k$ são **linearmente independentes nos pontos** $x_1, x_2, \dots, x_n$. Daqui resulta naturalmente que o número de pontos $n$ deverá ser sempre superior ou igual ao número de funções consideradas $k$.

# Capítulo 7

# Interpolação

## 7.1 Introdução

O problema de interpolação consiste em, dado um conjunto de pares ordenados $(x_0, y_0)$, $(x_1, y_1)$, ..., $(x_n, y_n)$, determinar uma função $g$, designada **função interpoladora**, tal que

$$g(x_i) = y_i, \quad i = 0, 1, \ldots, n.$$

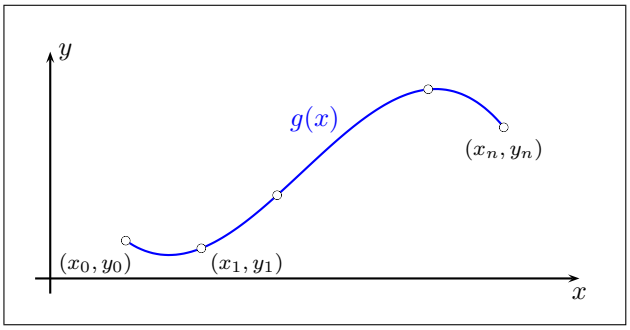



Figura 7.1: Interpolação.

Os valores $x_0, x_1, \ldots, x_n$ designam-se por **nós de interpolação** e devem satisfazer a condição $i \neq j \Rightarrow x_i \neq x_j$, ou seja, serem todos diferentes. Os correspondentes valores $y_0, y_1, \ldots, y_n$ designam-se por **valores nodais**.

Perante um dado problema de interpolação será necessário ter em consideração diversas questões, das quais se destacam a escolha da classe de funções interpoladoras a utilizar e a forma de determinar concretamente a função (ou uma função) interpoladora.

O problema de interpolação tem aplicações em diversas situações como sejam

- o cálculo de funções fornecidas por tabelas quando se pretende avaliar a função em pontos não tabelados (muito importante no passado!).

- quando apenas se conhecem os valores de uma função em certos pontos, por exemplo resultantes de medidas experimentais, e se pretende avaliar a função em novos pontos (sem repetir experiências ou medições ...).
- a aproximação de funções cujo cálculo seja complexo ou exija grande esforço.
- a base de muitos métodos numéricos.

O estudo de problemas de interpolação aqui apresentado centra-se na interpolação polinomial (funções interpoladoras polinomiais), abordando ainda a interpolação polinomial segmentada (splines polinomiais).

## 7.2 Interpolação polinomial

Comecemos por relembrar que uma função $p$ diz-se **polinomial de grau** $n$ se puder ser escrita na forma

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

onde $n \in \mathbb{N}_0$ e $a_n \neq 0$, excepto quando $n = 0$ em que $a_n$ pode ser nulo. Neste último caso o polinómio diz-se **nulo**, e o seu grau é, por convenção, $-\infty$.

Das justificações para a utilização de funções interpoladoras polinomiais podemos destacar as que se seguem.

- O cálculo dos valores de funções polinomiais é feito com um número finito de multiplicações e somas.
- As operações de derivação e primitivação de funções polinomiais são simples e podem ser facilmente realizadas de forma automática.
- As funções polinomiais são de classe $C^\infty$.
- As funções polinomiais aproximam tanto quanto se queira qualquer função contínua num intervalo finito (ver resultado abaixo).

Por abuso de linguagem, uma função polinomial é aqui identificada com o polinómio que a caracteriza.

**Teorema 7.2.1** (Weierstrass)**.** *Seja $[a, b]$ um intervalo real e $f$ uma função contínua em $[a, b]$. Então, qualquer que seja $\varepsilon > 0$, existe uma função polinomial $p$ tal que*

$$\max_{x \in [a,b]} |f(x) - p(x)| < \varepsilon.$$

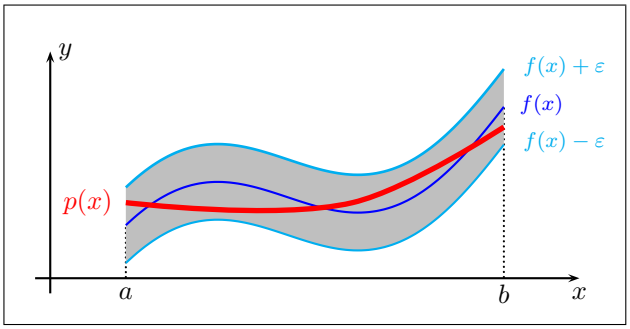



Figura 7.2: Teorema de Weierstrass.

Este teorema afirma a existência de polinómios que aproximam tanto quanto se queira qualquer função contínua (num intervalo limitado). No entanto, não fornece nenhuma indicação sobre como determinar tais polinómios, em função dependendo de uma aproximação $\varepsilon$ pretendida. É de referir, no entanto, que em determinadas circunstâncias, a interpolação polinomial produz polinómios aproximantes.

Um dado polinómio $p$ (leia-se função polinomial) pode ser apresentado de diversas formas. Na **forma de potências simples** será escrito como

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

Na **forma de potências centradas** será agora escrito como

$$p(x) = \bar{a}_0 + \bar{a}_1 (x - c) + \bar{a}_2 (x - c)^2 + \cdots + \bar{a}_n (x - c)^n,$$

onde $c$ é uma constante, designada por centro. Na **forma de Newton** será escrito como

$$p(x) = \tilde{a}_0 + \tilde{a}_1 (x - c_1) + \tilde{a}_2 (x - c_1)(x - c_2) + \cdots + \tilde{a}_n (x - c_1) \cdots (x - c_n),$$

onde os $c_i$ $(i = 1, \ldots, n)$ são constantes, designadas por centros.

O cálculo do valor de um polinómio $p$ num ponto $x$ pode ser efectuado de uma forma eficiente (reduzindo o número de operações aritméticas a realizar) empregando o designado **algoritmo de Horner**.

Para a forma de potências simples, $p(x) = a_0 + a_1 x + \cdots + a_n x^n$, tem-se

$y = a_n$
**Para** $i = n - 1$ **até** $0$ **fazer**
$\quad y = a_i + y \cdot x$

Para a forma de Newton, $p(x) = a_0 + a_1 (x - c_1) + \cdots + a_n (x - c_1) \cdots (x - c_n)$, tem-se

$y = a_n$
**Para** $i = n - 1$ **até** $0$ **fazer**
$\quad y = a_i + y \cdot (x - c_{i+1})$

Em ambos os casos $p(x)$ é dado pelo valor final de $y$.

## 7.3 Polinómio interpolador: unicidade e existência

Nesta secção iremos mostrar que para um conjunto de nós distintos $(x_i)_{i=0}^n$ e respectivos valores nodais $(y_i)_{i=0}^n$ quaisquer, existe um e um só polinómio $p$ de grau menor ou igual a $n$ tal que $p(x_i) = y_i$, para $i = 0, \ldots, n$.

Comecemos por relembrar o seguinte resultado sobre factorização de polinómios, que será utilizado posteriormente.

**Teorema 7.3.1.** *Se $z_1, z_2, \ldots, z_k$ forem zeros distintos do polinómio $p$, então*

$$p(x) = (x - z_1) \cdot (x - z_2) \cdots (x - z_k) \cdot r(x)$$

*onde $r$ é também um polinómio.*

O resultado seguinte afirma que se existir um polinómio interpolador de grau menor ou igual a $n$ então ele é único.

**Teorema 7.3.2** (Unicidade do polinómio interpolador)**.** *Sejam $p$ e $q$ polinómios, de grau inferior ou igual a $n$, que tomam os mesmos valores num conjunto de nós $x_0, x_1, \ldots, x_n$ distintos. Então estes polinómios são iguais.*

*Demonstração.* Seja $d$ o polinómio diferença entre $p$ e $q$, isto é

$$d(x) = p(x) - q(x)$$

Este polinómio terá grau inferior ou igual a $n$.

Como $p$ e $q$ tomam valores iguais em $x_0, x_1, \ldots, x_n$, é imediato concluir que $x_0, x_1, \ldots, x_n$ são raízes distintas de $d$. Então pode escrever-se

$$d(x) = (x - x_0) \cdot (x - x_1) \cdots (x - x_n) \cdot r(x)$$

para algum polinómio $r$. Seja $m$ o grau de $r$ e suponha-se que $m \geq 0$.

Então o grau de $d$ seria $n + 1 + m$, contrariando o facto do grau de $d$ ser inferior ou igual a $n$.

Conclui-se assim que não se pode ter $m \geq 0$.

A alternativa é $r$ ser o polinómio nulo e, consequentemente, $d$ ser também o polinómio nulo, ou seja, ou polinómios $p$ e $q$ serem iguais. □

Passemos agora à questão da existência do polinómio interpolador. Se o polinómio, de grau menor ou igual a $n$, $p(x) = a_0 + a_1 x + \ldots + a_n x^n$, interpolar os valores $y_i$ nos nós $x_i$ $(i = 0, \ldots, n)$

distintos, então os seus coeficientes terão de verificar

$$\begin{cases} a_0 + a_1 x_0 + \ldots + a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + \ldots + a_n x_1^n = y_1 \\ \ldots \\ a_0 + a_1 x_n + \ldots + a_n x_n^n = y_n \end{cases}$$

que não é mais do que um sistema de $n+1$ equações lineares nas $n+1$ incógnitas $a_0, a_1, \ldots, a_n$.

A existência do polinómio $p$, é então equivalente à existência de solução deste sistema de equações. Esta questão pode ser avaliada analisando o determinante da matriz dos coeficientes do sistema. Este, designa-se por **determinante de Vandermonde** nos $n+1$ pontos $x_0, x_1, \ldots, x_n$ e é dado por

$$v(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^{n-1} & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} & x_n^n \end{vmatrix}.$$

O valor deste determinante pode calculado como se segue. Comecemos por subtrair à última coluna deste determinante a penúltima coluna multiplicada por $x_0$. Obtém-se assim o determinante equivalente

$$v(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^{n-1} & 0 \\ 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} & x_1^{n-1}(x_1 - x_0) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} & x_n^{n-1}(x_n - x_0) \end{vmatrix}.$$

Subtraindo agora à penúltima coluna a ante-penúltima coluna multiplicada por $x_0$, resulta

$$v(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^{n-2} & 0 & 0 \\ 1 & x_1 & x_1^2 & \ldots & x_1^{n-2} & x_1^{n-2}(x_1 - x_0) & x_1^{n-1}(x_1 - x_0) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-2} & x_n^{n-2}(x_n - x_0) & x_n^{n-1}(x_n - x_0) \end{vmatrix}.$$

Repetindo este processo até subtrair à segunda coluna a primeira coluna multiplicada por $x_0$, obtém-se

$$v(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & x_1 - x_0 & x_1(x_1 - x_0) & \ldots & x_1^{n-2}(x_1 - x_0) & x_1^{n-1}(x_1 - x_0) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_n - x_0 & x_n(x_n - x_0) & \ldots & x_n^{n-2}(x_n - x_0) & x_n^{n-1}(x_n - x_0) \end{vmatrix}.$$

Desenvolvendo este determinante pela primeira linha, chega-se a

$$v(x_0, x_1, \ldots, x_n) = \begin{vmatrix} x_1 - x_0 & x_1(x_1 - x_0) & \ldots & x_1^{n-2}(x_1 - x_0) & x_1^{n-1}(x_1 - x_0) \\ x_2 - x_0 & x_2(x_2 - x_0) & \ldots & x_2^{n-2}(x_2 - x_0) & x_2^{n-1}(x_2 - x_0) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_n - x_0 & x_n(x_n - x_0) & \ldots & x_n^{n-2}(x_n - x_0) & x_n^{n-1}(x_n - x_0) \end{vmatrix}.$$

Colocando agora em evidência na primeira linha $x_1 - x_0$, na segunda linha $x_2 - x_0$, e assim sucessivamente, até $x_n - x_0$ na última linha, tem-se ainda que

$$v(x_0, x_1, \ldots, x_n) = (x_1 - x_0) \cdot (x_2 - x_0) \cdots (x_n - x_0) \cdot \begin{vmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{vmatrix}.$$

pelo que se pode escrever

$$v(x_0, x_1, \ldots, x_n) = \left[\prod_{j=1}^{n}(x_j - x_0)\right] \cdot v(x_1, \ldots, x_n),$$

onde $v(x_1, \ldots, x_n)$ é o determinante de Vandermonde nos $n$ pontos $x_1, \ldots, x_n$.

Repetindo o processo acima para o determinante $v(x_1, \ldots, x_n)$, depois para $v(x_2, \ldots, x_n)$ e assim sucessivamente, obtém-se a expressão

$$v(x_0, x_1, \ldots, x_n) = \left[\prod_{j=1}^{n}(x_j - x_0)\right] \cdot \left[\prod_{j=2}^{n}(x_j - x_1)\right] \cdot \ \ldots \ \cdot \left[\prod_{j=n}^{n}(x_j - x_{n-1})\right]$$

concluindo-se então que $v(x_0, x_1, \ldots, x_n)$ será não nulo desde que os nós $x_i$ sejam todos diferentes.

Verifica-se deste modo que o sistema de equações que permite obter os coeficientes do polinómio interpolador é possível (e determinado), podendo então afirmar-se que existe um polinómio de grau não superior a $n$ que interpola os valores $(y_i)_{i=0}^n$ nos nós distintos $(x_i)_{i=0}^n$.

Uma vez mostrada a existência e unicidade do polinómio interpolador, interessa agora encontrar formas de o determinar. Uma possibilidade é resolver o sistema de equações de interpolação

$$\sum_{j=0}^{n} a_j x_i^j = y_i, \qquad i = 0, 1, \ldots, n.$$

Esta abordagem, embora simples, não é aconselhável, pois exige um número elevado de cálculos. Por outro lado, a resolução deste sistema pode acarretar elevados erros numéricos devidos à utilização de aritmética finita, que pioram à medida que $n$ cresce.

Nas secções seguintes serão estudados processos mais eficientes de determinar o polinómio interpolador. Interessa aqui realçar que os processos apresentados constituem diferentes formas de obter o mesmo polinómio interpolador (dado um mesmo conjunto de nós e respectivos valores nodais).

## 7.4 Forma de Lagrange

Consideremos novamente um conjunto de nós distintos $(x_i)_{i=0}^n$. Os polinómios (de grau $n$) definidos pela expressão

$$L_k(x) = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{x - x_i}{x_k - x_i}, \quad k = 0, 1, \ldots, n,$$

designam-se por **polinómios de Lagrange**, relativos aos nós $x_0, x_1, \ldots, x_n$.

Facilmente se conclui que estes polinómios verificam as relações $L_k(x_j) = \delta_{kj}$, onde $\delta_{kj}$ é o designado **delta de Kronecker**, ou seja

$$\delta_{kj} = \begin{cases} 1 & \text{se } k = j, \\ 0 & \text{se } k \neq j. \end{cases}$$

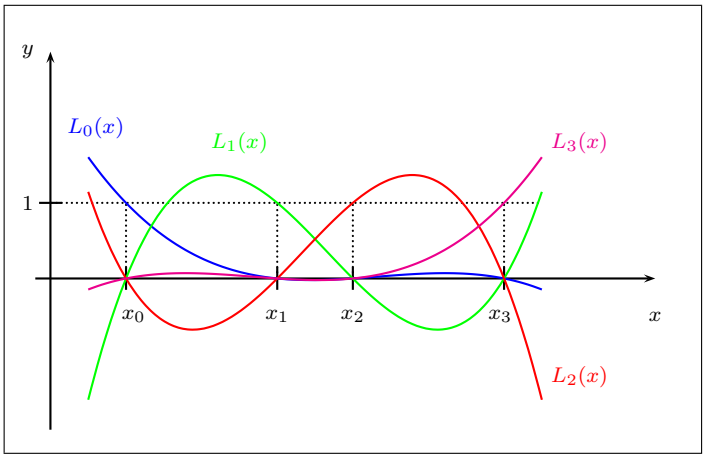



Figura 7.3: Polinómios de Lagrange (exemplo com 4 nós).

O polinómio interpolador na **forma de Lagrange** é obtido como uma combinação linear dos polinómios de Lagrange relativos aos nós em questão. Os coeficientes desta combinação linear serão os valores nodais a interpolar, como se refere no seguinte teorema.

**Teorema 7.4.1.** *O polinómio $p$, de grau menor ou igual a $n$, que interpola o conjunto de valores $y_0, y_1, \ldots, y_n$ nos nós distintos $x_0, x_1, \ldots, x_n$ é dado por*

$$p(x) = \sum_{k=0}^{n} y_k L_k(x).$$

*Demonstração.* Como $p$ é a soma de polinómios de grau $n$ ou nulos, conclui-se que o grau de $p$ é menor ou igual a $n$. Por outro lado, para cada nó $x_i$ tem-se que

$$p(x_i) = \sum_{k=0}^{n} y_k L_k(x_i) = \sum_{k=0}^{n} y_k \delta_{ki} = y_i$$

pelo que $p$ interpola os valores nodais. □

O exemplo seguinte ilustra a obtenção do polinómio interpolador na forma de Lagrange.

**Exemplo 7.4.1.** *Determinar o polinómio de grau menor ou igual a* 3 *que interpola os valores*

| $x$ | $-1$ | $0$ | $2$ | $3$ |
|---|---|---|---|---|
| $y$ | $6$ | $-12$ | $18$ | $24$ |

***Resolução***

*Inicialmente calculam-se os polinómios de Lagrange relativos aos nós de interpolação.*

$$\begin{aligned}
L_0(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \\
&= \frac{(x-0)(x-2)(x-3)}{(-1-0)(-1-2)(-1-3)} = -\frac{1}{12}x(x-2)(x-3) \\
L_1(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
&= \frac{(x-(-1))(x-2)(x-3)}{(0-(-1))(0-2)(0-3)} = \frac{1}{6}(x+1)(x-2)(x-3) \\
L_2(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \\
&= \frac{(x-(-1))(x-0)(x-3)}{(2-(-1))(2-0)(2-3)} = -\frac{1}{6}(x+1)x(x-3) \\
L_3(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
&= \frac{(x-(-1))(x-0)(x-2)}{(3-(-1))(3-0)(3-2)} = \frac{1}{12}(x+1)x(x-2)
\end{aligned}$$

*O polinómio interpolador na forma de Lagrange será*

$$\begin{aligned}
p(x) &= 6\cdot L_0(x) + (-12)\cdot L_1(x) + 18\cdot L_2(x) + 24\cdot L_3(x) \\
&= -\frac{1}{2}x(x-2)(x-3) - 2(x+1)(x-2)(x-3) \\
&\quad -3(x+1)x(x-3) + 2(x+1)x(x-2)
\end{aligned}$$

*Este polinómio escrito na forma de potências simples fica*

$$p(x) = -\frac{7}{2}x^3 + \frac{29}{2}x^2 - 12.$$

A forma de Lagrange do polinómio interpolador é bastante fácil de determinar. Além disso, se os nós de interpolação se mantiverem fixos, mas algum ou alguns dos valores nodais for alterado, não é necessário recalcular os polinómios $L_k$, mas somente a sua combinação linear. Por outro lado, quando se altera ou adiciona um nó é necessário recalcular todos os polinómios $L_k$, desaproveitando todos os cálculos entretanto efectuados.

## 7.5 Forma de Aitken-Neville

A **forma de Aitken-Neville** permite calcular o valor do polinómio interpolador num ponto $x$ de uma forma recursiva, considerando sucessivamente mais nós de interpolação e respectivos valores nodais.

Sejam $m$ um inteiro entre 0 e $n$, $k$ um inteiro entre 0 e $n-m$, e defina-se $p_{m,k}$ como o polinómio de grau menor ou igual a $k$ que interpola os valores $(y_i)_{i=m}^{m+k}$ nos nós $(x_i)_{i=m}^{m+k}$. A obtenção do polinómio interpolador na forma de Aitken-Neville é ilustrada na figura seguinte, onde o polinómio $p_{m,k+1}$ é construído à custa dos polinómios $p_{m,k}$ e $p_{m+1,k}$.

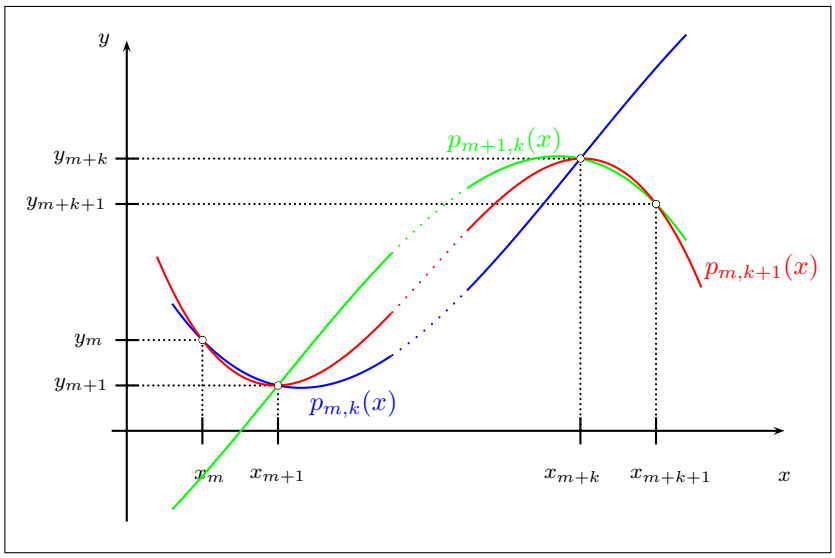



Figura 7.4: Forma de Aitken-Neville.

O teorema seguinte apresenta a expressão que permite o cálculo do polinómio interpolador na forma de Aitken-Neville.

**Teorema 7.5.1.** *Dados $m$ e $k$, e os polinómios $p_{m,k}$ e $p_{m+1,k}$, o polinómio $p_{m,k+1}$ satisfaz a relação*

$$p_{m,k+1}(x) = \frac{(x - x_{m+k+1}) \cdot p_{m,k}(x) + (x_m - x) \cdot p_{m+1,k}(x)}{x_m - x_{m+k+1}}.$$

*Demonstração.* Como $p_{m,k}$ e $p_{m+1,k}$ são polinómios de grau não superior a $k$, o polinómio $q$ definido por

$$q(x) = \frac{(x - x_{m+k+1}) \cdot p_{m,k}(x) + (x_m - x) \cdot p_{m+1,k}(x)}{x_m - x_{m+k+1}}$$

terá grau menor ou igual a $k+1$. Para mostrar que $q \equiv p_{m,k+1}$ resta então verificar que $q(x_i) = y_i$, para $i = m, m+1, \ldots, m+k, m+k+1$.

Seja $i$ um inteiro tal que $m+1 \leq i \leq m+k$. Então $p_{m,k}(x_i) = y_i$ e $p_{m+1,k}(x_i) = y_i$. Calculando

$q(x_i)$ obtém-se

$$q(x_i) = \frac{(x_i - x_{m+k+1}) \cdot y_i + (x_m - x_i) \cdot y_i}{x_m - x_{m+k+1}} = \frac{(x_m - x_{m+k+1}) \cdot y_i}{x_m - x_{m+k+1}} = y_i.$$

Por outro lado, como $p_{m,k}(x_m) = y_m$ e $p_{m+1,k}(x_{m+k+1}) = y_{m+k+1}$, tem-se respectivamente que

$$q(x_m) = \frac{(x_m - x_{m+k+1}) \cdot y_m}{x_m - x_{m+k+1}} = y_m \quad \text{e}$$
$$q(x_{m+k+1}) = \frac{(x_m - x_{m+k+1}) \cdot y_{m+k+1}}{x_m - x_{m+k+1}} = y_{m+k+1},$$

concluindo-se portanto que $q(x_i) = y_i$ para $i = m, m+1, \ldots, m+k, m+k+1$, como se pretendia mostrar. □

A aplicação repetida da expressão (7.5.1) para um dado ponto $x$, permite avaliar o valor do polinómio interpolador nesse ponto *sem determinar os coeficientes do polinómio.*

A forma de Aitken-Neville é muitas vezes também designada por **interpolação linear iterada**. De facto, a expressão (7.5.1) corresponde a uma generalização da expressão

$$\frac{(x - x_1) \cdot y_0 + (x_0 - x) \cdot y_1}{x_0 - x_1}$$

que permite calcular o valor em $x$ da função linear que interpola $y_0$ em $x_0$ e $y_1$ em $x_1$.

A expressão de recorrência da forma de Aitken-Neville pode ainda ser escrita como

$$p_{m,k+1}(x) = \frac{\begin{vmatrix} p_{m,k}(x) & x - x_m \\ p_{m+1,k}(x) & x - x_{m+k+1} \end{vmatrix}}{x_m - x_{m+k+1}}.$$

Para avaliar o polinómio que interpola $(y_i)_{i=0}^n$ nos nós $(x_i)_{i=0}^n$, em $x$, é necessário calcular

$$\begin{aligned} &p_{i,0}(x), \quad i = 0, \ldots, n, \\ &p_{i,1}(x), \quad i = 0, \ldots, n-1, \\ &\ldots \quad \text{e, finalmente,} \\ &p_{0,n}(x) = p(x). \end{aligned}$$

Uma vez que $p_{i,0}(x) \equiv y_i$, é habitual utilizar a notação

$$\begin{aligned} p_{i,0}(x) &= y_i(x) \\ p_{i,1}(x) &= y_{i,i+1}(x) \\ p_{i,2}(x) &= y_{i,i+1,i+2}(x) \\ &\ldots \end{aligned}$$

tendo-se então

$$y_{01}(x) = \frac{\begin{vmatrix} y_0 & x - x_0 \\ y_1 & x - x_1 \end{vmatrix}}{x_0 - x_1},\ y_{12}(x) = \frac{\begin{vmatrix} y_1 & x - x_1 \\ y_2 & x - x_2 \end{vmatrix}}{x_1 - x_2}, \ldots$$

$$y_{012}(x) = \frac{\begin{vmatrix} y_{01}(x) & x - x_0 \\ y_{12}(x) & x - x_2 \end{vmatrix}}{x_0 - x_2}, \ldots$$

$$\ldots$$

**Exemplo 7.5.1.** *Determinar, em $x = 1$, o valor do polinómio de grau menor ou igual a 3 que interpola os valores da seguinte tabela.*

| $x$ | $-1$ | $0$ | $2$ | $3$ |
|---|---|---|---|---|
| $y$ | $6$ | $-12$ | $18$ | $24$ |

***Resolução***

*Interpolando linearmente entre cada dois pontos consecutivos, obtêm-se os valores $y_{i,i+1}$*

$$y_{01}(1) = \frac{\begin{vmatrix} 6 & 1+1 \\ -12 & 1-0 \end{vmatrix}}{-1-0} = -30,\ y_{12}(1) = \frac{\begin{vmatrix} -12 & 1-0 \\ 18 & 1-2 \end{vmatrix}}{0-2} = 3,\ y_{23}(1) = \frac{\begin{vmatrix} 18 & 1-2 \\ 24 & 1-3 \end{vmatrix}}{2-3} = 12$$

*Segue-se a interpolação dos valores obtidos acima para obter os valores $y_{i,i+1,i+2}$*

$$y_{012}(1) = \frac{\begin{vmatrix} -30 & 1+1 \\ 3 & 1-2 \end{vmatrix}}{-1-2} = -8, \qquad y_{123}(1) = \frac{\begin{vmatrix} 3 & 1-0 \\ 12 & 1-3 \end{vmatrix}}{0-3} = 6$$

*Finalmente, obtém-se o valor $y_{0123}$ pretendido*

$$y_{0123}(1) = \frac{\begin{vmatrix} -8 & 1+1 \\ 6 & 1-3 \end{vmatrix}}{-1-3} = -1.$$

A principal característica que distingue a forma de Aitken-Neville prende-se com o facto de permitir calcular o valor do polinómio interpolador num dados ponto sem calcular os seus coeficientes. Esta forma permite ainda adicionar e retirar nós nos "extremos" reutilizando os cálculos já efectuados. Contudo, exige a repetição dos cálculos se houver alteração dos valores nodais.

## 7.6 Forma de Newton

Consideremos novamente os nós de interpolação distintos $x_0, x_1, \ldots, x_n$. Definam-se os polinómios $W_i$, para $i = 0, 1, \ldots, n-1$, designados polinómios nodais, da seguinte forma

$$\begin{aligned} W_0(x) &= x - x_0 \\ W_1(x) &= (x - x_0)(x - x_1) \\ &\ldots \\ W_{n-1}(x) &= (x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned}$$

O polinómio interpolador $p$ pode ser escrito na **forma de Newton** tomando como centros os nós distintos $x_0, x_1, \ldots, x_n$, isto é,

$$p(x) = a_0 + a_1 W_0(x) + \cdots + a_n W_{n-1}(x),$$

ficando a sua determinação reduzida ao cálculo dos coeficientes $a_0, a_1, \ldots, a_n$.

Partindo do polinómio interpolador $p$, escrito na forma Newton acima indicada, definam-se os polinómios $p_0, p_1, \ldots, p_n$ por intermédio de

$$\begin{aligned} p_0(x) &= a_0, \\ p_1(x) &= a_0 + a_1 W_0(x), \\ p_2(x) &= a_0 + a_1 W_0(x) + a_2 W_1(x), \\ &\ldots \end{aligned}$$

Estes polinómios podem obter-se de uma forma recursiva fazendo

$$\begin{aligned} &p_0(x) = a_0 \qquad \text{e} \\ &p_k(x) = p_{k-1}(x) + a_k W_{k-1}(x), \quad k = 1, \ldots, n. \end{aligned}$$

Note-se que o polinómio $p_k$ apenas depende dos valores $a_0, \ldots, a_k$ e também que o polinómio interpolador $p$ será dado por $p_n$.

O teorema seguinte mostra como se devem calcular os valores dos coeficientes $a_0, a_1, \ldots, a_n$ do polinómio interpolador na forma de Newton.

**Teorema 7.6.1.** *Fazendo $a_0 = y_0$ e*

$$a_k = \frac{y_k - p_{k-1}(x_k)}{W_{k-1}(x_k)}, \quad k = 1, \ldots, n,$$

*então o polinómio $p_k$ interpola os valores $(y_j)_{j=0}^k$ nos nós $(x_j)_{j=0}^k$, isto para $k = 0, 1, \ldots, n$.*

*Demonstração.* Esta demonstração será feita por indução.

Como $p_0(x) = y_0$, é óbvio que este polinómio interpola $y_0$ em $x_0$.

Suponha-se agora que $p_{k-1}$ interpola os valores $(y_j)_{j=0}^{k-1}$ nos nós $(x_j)_{j=0}^{k-1}$. Como se viu atrás, $p_k(x) = p_{k-1}(x) + a_k W_{k-1}(x)$. Da definição dos polinómios $W_0, W_1, \ldots, W_{n-1}$, tem-se que $W_{k-1}(x_j) = 0,\ \forall j = 0, 1, \ldots, k-1$, concluindo-se então que $p_k(x_j) = p_{k-1}(x_j) = y_j,\ \forall j = 0, 1, \ldots, k-1$. Por outro lado, tem-se que

$$p_k(x_k) = p_{k-1}(x_k) + \frac{y_k - p_{k-1}(x_k)}{W_{k-1}(x_k)} W_{k-1}(x_k) = y_k,$$

concluindo-se finalmente que $p_k$ interpola os valores $(y_j)_{j=0}^{k}$ nos nós $(x_j)_{j=0}^{k}$. □

Do processo de determinação dos coeficientes do polinómio na forma de Newton, conclui-se que a consideração de novos nós apenas exige o cálculo dos coeficientes adicionais, aproveitando os cálculos entretanto já efectuados. Embora seja habitual ordenar os nós de interpolação, tal não é necessário, podendo estes ser considerados por qualquer ordem.

**Exemplo 7.6.1.** *Determinar, na forma de Newton, o polinómio de grau menor ou igual a* 2 *que interpola os valores da seguinte tabela.*

| $x$ | $-1$ | $2$ | $3$ |
|---|---|---|---|
| $y$ | $1$ | $3$ | $5$ |

***Resolução*** *Começando com o nó* 0 *tem-se que* $p_0(x) = y_0 = 1$*, polinómio que interpola o primeiro ponto da tabela.*

*Passando a agora ao nó* 1*, e fazendo* $W_0(x) = x - x_0 = x + 1$*, obtém-se*

$$p_1(x) = p_0(x) + \frac{y_1 - p_0(x_1)}{W_0(x_1)} W_0(x) = 1 + \frac{3-1}{2+1}(x+1) = 1 + \frac{2}{3}(x+1).$$

*Usando finalmente o nó* 3*, e como* $W_0(x) = (x - x_0)(x - x_1) = (x+1)(x-1)$*, tem-se*

$$p_2(x) = p_1(x) + \frac{y_2 - p_1(x_2)}{W_1(x_2)} W_1(x) = 1 + \frac{2}{3}(x+1) + \frac{5 - (1 + \frac{2}{3}(3+1))}{(3+1)(3-2)}(x+1)(x-2)$$

*Sendo então o polinómio interpolador* $p(x) = p_2(x)$ *dado por*

$$p(x) = 1 + \frac{2}{3}(x+1) + \frac{1}{3}(x+1)(x-2).$$

## 7.7 Diferenças divididas e diferenças finitas

Sendo $m$ e $k$ inteiros não negativos, defina-se $p_{m,k}$ como o polinómio de grau menor ou igual a $k$ que interpola os valores $(y_i)_{i=m}^{m+k}$ nos nós $(x_i)_{i=m}^{m+k}$. Na forma de Newton, este polinómio será

$$p_{m,k}(x) = a_{m,0} + a_{m,1}(x - x_m) + \cdots + a_{m,k}(x - x_m) \cdots (x - x_{m+k-1})$$

A construção do polinómio interpolador na forma de Newton permite concluir que o coeficiente $a_{m,j}$, para $j = 0, 1, \ldots, k$, apenas depende dos valores $(y_i)_{i=m}^{m+j}$ e dos nós $(x_i)_{i=m}^{m+j}$. Este coeficiente representa-se por

$$a_{m,j} = y[x_m, \ldots, x_{m+j}]$$

e designa-se por **diferença dividida** (de ordem $j$ nos nós $x_m, \ldots, x_{m+j}$).

O teorema seguinte estabelece uma relação entre diferenças divididas que permite efectuar o seu cálculo de um modo recursivo.

**Teorema 7.7.1.** *As diferenças dividas satisfazem $y[x_j] = y_j$, com $0 \leq j \leq n$, e*

$$y[x_m, \ldots, x_{k+1}] = \frac{y[x_{m+1}, \ldots, x_{k+1}] - y[x_m, \ldots, x_k]}{x_{k+1} - x_m}$$

*com $0 \leq m \leq k \leq n-1$.*

*Demonstração.* $y[x_j] = y_j$, uma vez que o polinómio constante $p_{j,0}(x) = y_j$ interpola $y_j$ em $x_j$.

Sejam $p_{m,k-m}$ e $p_{m+1,k-m}$ os polinómios, de graus menores ou iguais a $k-m$, que interpolam $(y_j)_{j=m}^{k}$ em $(x_j)_{j=m}^{k}$ e $(y_j)_{j=m+1}^{k+1}$ em $(x_j)_{j=m+1}^{k+1}$, respectivamente. Então, o polinómio $q$ definido pela expressão

$$q(x) = \frac{x - x_m}{x_{k+1} - x_m} p_{m+1,k-m}(x) + \frac{x_{k+1} - x}{x_{k+1} - x_m} p_{m,k-m}(x)$$

interpola $y_m, \ldots, y_{k+1}$ em $x_m, \ldots, x_{k+1}$ e tem grau menor ou igual a $k-m+1$ (notar a semelhança entre esta expressão e a relação de recorrência da forma de Aitken-Neville do polinómio interpolador). Da unicidade do polinómio interpolador tem-se que $q \equiv p_{m,k-m+1}$. Igualando os coeficientes do termo $x^{k-m+1}$ obtém-se

$$a_{m,k-m+1} = \frac{a_{m+1,k-m} - a_{m,k-m}}{x_{k+1} - x_m},$$

ou ainda, usando diferenças divididas,

$$y[x_m, \ldots, x_{k+1}] = \frac{y[x_{m+1}, \ldots, x_{k+1}] - y[x_m, \ldots, x_k]}{x_{k+1} - x_m}.$$ □

A utilização de diferenças divididas permite escrever o polinómio interpolador na forma de Newton como

$$p(x) = y[x_0] + y[x_0, x_1](x - x_0) + \cdots + y[x_0, x_1, \ldots, x_n](x - x_0) \cdots (x - x_{n-1})$$

onde

$$\begin{aligned}
&y[x_0] = y_0 \\
&y[x_0, x_1] = \frac{y[x_1] - y[x_0]}{x_1 - x_0} \\
&y[x_0, x_1, x_2] = \frac{y[x_1, x_2] - y[x_0, x_1]}{x_2 - x_0} \\
&\ldots
\end{aligned}$$

Os cálculos das diferenças divididas podem ser organizados de um modo expedito dispondo-os numa tabela como se mostra abaixo (exemplo com 4 nós).

| $x$ | $y[\cdot]$ | $y[\cdot,\cdot]$ | $y[\cdot,\cdot,\cdot]$ | $y[\cdot,\cdot,\cdot,\cdot]$ |
|---|---|---|---|---|
| $x_0$ | $y_0$ | | | |
| | | $y[x_0,x_1]$ | | |
| $x_1$ | $y_1$ | | $y[x_0,x_1,x_2]$ | |
| | | $y[x_1,x_2]$ | | $y[x_0,x_1,x_2,x_3]$ |
| $x_2$ | $y_2$ | | $y[x_1,x_2,x_3]$ | |
| | | $y[x_2,x_3]$ | | |
| $x_3$ | $y_3$ | | | |

O exemplo seguinte ilustra a utilização de diferenças divididas para a obtenção do polinómio interpolador na forma de Newton.

**Exemplo 7.7.1.** *Determinar, na forma de Newton, o polinómio de grau menor ou igual a* 3 *que interpola os valores da seguinte tabela.*

| $x$ | $-1$ | $0$ | $2$ | $3$ |
|---|---|---|---|---|
| $y$ | $6$ | $-12$ | $18$ | $24$ |

***Resolução***

*A tabela das diferenças divididas correspondente aos valores dados é*

| $x$ | $y[\cdot]$ | $y[\cdot,\cdot]$ | $y[\cdot,\cdot,\cdot]$ | $y[\cdot,\cdot,\cdot,\cdot]$ |
|---|---|---|---|---|
| $-1$ | $\boxed{6}$ | | | |
| | | $\boxed{-18}$ | | |
| $0$ | $-12$ | | $\boxed{11}$ | |
| | | $15$ | | $\boxed{-\frac{7}{2}}$ |
| $2$ | $18$ | | $-3$ | |
| | | $6$ | | |
| $3$ | $24$ | | | |

*E então o polinómio interpolador será*

$$p(x) = 6 - 18(x+1) + 11(x+1)x - \tfrac{7}{2}(x+1)x(x-2).$$

Para além das diferenças divididas, podem também definir-se as designadas diferenças finitas. A **diferença finita** de ordem $k \in \mathbb{N}_0$ e passo $h > 0$ da função $f$ representa-se por $\Delta_h^k f$ e o seu valor no ponto $x$ é

$$\begin{aligned} &\Delta_h^0 f(x) = f(x), \\ &\Delta_h^{k+1} f(x) = \Delta_h^k f(x+h) - \Delta_h^k f(x), \qquad k = 0, 1, \ldots \end{aligned}$$

Em particular, tem-se que

$$\Delta_h^1 f(x) = f(x+h) - x(x),$$
$$\Delta_h^2 f(x) = \Delta_h^1 f(x+h) - \Delta_h^1 f(x) = [f(x+2h) - f(x+h)] - [f(x+h) - f(x)],$$
$$\dots$$

Sempre que não haja ambiguidade quanto ao valor do passo $h$, as diferenças finitas representam-se simplesmente por $\Delta^0 f, \Delta^1 f, \Delta^2 f, \dots$

Quando os nós de interpolação se encontram igualmente espaçados, isto é, quando existe um valor $h$ tal que $x_{i+1} - x_i = h$, para $i = 0, 1, \dots, n-1$, as diferenças finitas dos valores nodais $(y_i)_{i=0}^n$ são dadas por

$$\begin{aligned}
&\Delta^0 y_i = y_i && i = 0, \dots, n \\
&\Delta^1 y_i = \Delta y_i = y_{i+1} - y_i && i = 0, \dots, n-1 \\
&\Delta^2 y_i = \Delta^1 y_{i+1} - \Delta^1 y_i && i = 0, \dots, n-2 \\
&\dots
\end{aligned}$$

O resultado apresentado em seguida estabelece uma relação entre as diferenças finitas e as diferenças divididas dos valores nodais correspondentes a nós igualmente espaçados.

**Teorema 7.7.2.** *A diferença dividida de ordem $k$ dos valores nodais $y$ nos nós $h$-equidistantes $x_i, x_{i+1}, \dots, x_{i+k}$ satisfaz*

$$y[x_i, \dots, x_{i+k}] = \frac{1}{k! h^k} \Delta^k y_i$$

*Demonstração.* Sendo $k = 0$ verifica-se que $y[x_i] = y_i = \Delta^0 y_i$, por definição.

A relação de recorrência entre as diferenças divididas permite escrever

$$y[x_i, \dots, x_{i+k+1}] = \frac{y[x_{i+1}, \dots, x_{i+k+1}] - y[x_i, \dots, x_{i+k}]}{x_{i+k+1} - x_i}$$

Supondo a validade da relação a mostrar para $k$, tem-se

$$y[x_i, \dots, x_{i+k+1}] = \frac{\frac{1}{k! h^k} \Delta^k y_{i+1} - \frac{1}{k! h^k} \Delta^k y_i}{(k+1)h}$$

Da definição das diferenças finitas tem-se $\Delta^{k+1} y_i = \Delta^k y_{i+1} - \Delta^k y_i$, obtendo-se

$$y[x_i, \dots, x_{i+k+1}] = \frac{1}{k! h^k} \frac{1}{(k+1)h} \Delta^{k+1} y_i = \frac{1}{(k+1)! h^{k+1}} \Delta^{k+1} y_i$$

ou seja, a validade da expressão dada para $k+1$.

Desta forma, o resultado fica demonstrado por indução. □

Tal como no caso das diferenças divididas, é também vantajoso dispor os cálculos das diferenças finitas numa tabela.

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\dots$ | $\Delta^{n-1} y$ | $\Delta^n y$ |
|---|---|---|---|---|---|---|
| $x_0$ | $y_0$ | | | | | |
| | | $\Delta y_0$ | | | | |
| $x_1$ | $y_1$ | | $\Delta^2 y_0$ | | | |
| | | $\Delta y_1$ | | $\dots$ | | |
| $\dots$ | $\dots$ | | $\dots$ | | $\Delta^{n-1} y_0$ | |
| | | $\dots$ | | $\dots$ | | $\Delta^n y_0$ |
| $\dots$ | $\dots$ | | $\dots$ | | $\Delta^{n-1} y_1$ | |
| | | $\Delta y_{n-2}$ | | $\dots$ | | |
| $x_{n-1}$ | $y_{n-1}$ | | $\Delta^2 y_{n-2}$ | | | |
| | | $\Delta y_{n-1}$ | | | | |
| $x_n$ | $y_n$ | | | | | |

Caso os nós de interpolação sejam equidistantes é então possível obter o polinómio $p$, de grau menor ou igual a $n$, que interpola os valores $(y_i)_{i=0}^n$ nos nós $h$-equidistantes $(x_i)_{i=0}^n$ na forma de Newton utilizando diferenças finitas. Este polinómio será dado por

$$p(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2h^2}(x - x_0)(x - x_1) + \\ + \dots + \frac{\Delta^n y_0}{n! h^n}(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

ou, numa forma compacta,

$$p(x) = \sum_{k=0}^{n} \left[ \frac{\Delta^k y_0}{k! h^k} \prod_{i=0}^{k-1} (x - x_i) \right].$$

**Exemplo 7.7.2.** *Determinar o polinómio $p$, de grau menor ou igual a 3, que interpola os valores da seguinte tabela.*

| $x$ | $-1$ | $1$ | $3$ | $5$ |
|---|---|---|---|---|
| $y$ | $2$ | $5$ | $3$ | $1$ |

***Resolução***

*A tabela das diferenças finitas dos valores nodais é*

| $x$ | $y$ | $\Delta y$ | $\Delta^2 y$ | $\Delta^3 y$ |
|---|---|---|---|---|
| $-1$ | $\boxed{2}$ | | | |
| | | $\boxed{3}$ | | |
| $1$ | $5$ | | $\boxed{-5}$ | |
| | | $-2$ | | $\boxed{5}$ |
| $3$ | $3$ | | $0$ | |
| | | $-2$ | | |
| $5$ | $1$ | | | |

*Pelo que o polinómio interpolador será*

$$p(x) = 2 + \frac{3}{2}(x+1) - \frac{5}{2\times 4}(x+1)(x-1) + \frac{5}{6\times 8}(x+1)(x-1)(x-3).$$

Como fica patente pelos exemplos apresentados, a determinação dos valores nodais obriga a refazer todos os cálculos na determinação do polinómio da forma de Newton. Refira-se também que estes cálculos se tornam mais simples se os nós estiverem igualmente espaçados (utilizando diferenças finitas em vez de diferenças divididas).

## 7.8 Interpolação directa e inversa

Sejam $f : [a, b] \to \mathbb{R}$ , $(x_i)_{i=0}^n$ nós distintos pertencentes a $[a, b]$ e $y_i = f(x_i)$ para $i = 0, 1, \ldots, n$.

A **interpolação directa** de $f$ nos nós $(x_i)_{i=0}^n$ consiste em determinar o polinómio $p$ (de grau menor ou igual a $n$) que verifica $p(x_i) = y_i$ para $i = 0, 1, \ldots, n$.

Se $f$ admitir inversa em $[a, b]$ então a **interpolação inversa** de $f$ consiste em determinar um polinómio $q$ (de grau menor ou igual a $n$) tal que

$$q(y_i) = x_i, \qquad i = 0, 1, \ldots, n.$$

Agora $(y_i)_{i=0}^n$ são os nós de interpolação e $(x_i)_{i=0}^n$ são os valores nodais, da função $f^{-1}$ a interpolar.

Uma das aplicações da interpolação inversa é a determinação de zeros de funções, como se ilustra no exemplo seguinte.

**Exemplo 7.8.1.** *Determinar um valor aproximado do zero de* $f(x) = \frac{3}{2}\sin(x) - e^{-x}$ *em* $[0, 1]$.

***Resolução***

*Um zero* $s$ *é, por definição, um valor tal que* $f(s) = 0$.

*Tem-se que* $f(0) = -1$ *e* $f(1) = 0.89433$. *Como* $f$ *é estritamente crescente em* $[0, 1]$ *(porquê?) então* $f$ *admite inversa nesse intervalo. Logo conclui-se que*

$$f(s) = 0 \Leftrightarrow s = f^{-1}(0)$$

*Utilizando interpolação inversa de* $f$ *e calculando o valor de um polinómio interpolador de* $f^{-1}$ *em* $0$ *obter-se-á um valor aproximado do zero de* $f$.

*Escolhendo alguns nós em* $[0, 1]$ *e calculando os valores nodais obtém-se*

| $x$ | 0 | 0.4 | 0.6 | 1 |
|---|---|---|---|---|
| $y = f(x)$ | $-1.00000$ | $-0.08619$ | 0.29815 | 0.89433 |

*Utilizando a forma de Newton calculada a partir das diferenças divididas*

| $y$ | $x$ | $x[,]$ | $x[,,]$ | $x[,,,]$ |
|---|---|---|---|---|
| $-1.00000$ | 0 | | | |
| | | 0.43773 | | |
| $-0.08619$ | 0.4 | | 0.06366 | |
| | | 0.52037 | | 0.04745 |
| 0.28815 | 0.6 | | 0.15356 | |
| | | 0.67094 | | |
| 0.89433 | 1 | | | |

*O polinómio interpolador fica*

$$\begin{aligned} p(y) &= 0 + 0.43773(y+1) + 0.06366(y+1)(y+0.08619) + \\ &\quad + 0.04745(y+1)(y+0.08619)(y-0.28815) \end{aligned}$$

*E então,* $s = f^{-1}(0) \approx p(0) = 0.44200$ *(verificando-se que* $f(0.44200) = -0.00113$*).*

## 7.9 Dupla interpolação

Consideremos o problema descrito em seguida. Conhecidos os valores $z_{ij} = f(x_i, y_j)$ de uma função $f : \mathbb{R}^2 \to \mathbb{R}$, onde $(x_i)_{i=0}^n$ são distintos, e $(y_j)_{i=0}^m$ são também distintos, pretende-se obter um valor aproximado de $f(\bar{x}, \bar{y})$. Este é um problema de interpolação em $\mathbb{R}^2$, mas que pode ser "resolvido" utilizando interpolação em $\mathbb{R}$. Para tal poder-se-á aplicar o método designado por **dupla interpolação**, que consiste em efectuar interpolações polinomiais independentes nas duas variáveis, uma de cada vez. Estas interpolações podem ser efectuadas de duas formas alternativa. A primeira alternativa consiste em realizar as operações

1. interpolando em $x$, obtém-se para cada $j$ o polinómio $p_j$ que interpola os valores $(z_{ij})_{i=0}^n$ nos nós $(x_i)_{i=0}^n$;
2. posteriormente, determina-se o polinómio $q$ que interpola os valores $p_j(\bar{x})$ nos nós $(y_j)_{j=0}^m$;

sendo o valor procurado $q(\bar{y})$. Para esta alternativa será vantajoso dispor os cálculos como se mostra na seguinte tabela.

| $f(x,y)$ | $y_0$ | $\cdots$ | $y_l$ | $\bar{y}$ | $y_{l+1}$ | $\cdots$ | $y_m$ |
|---|---|---|---|---|---|---|---|
| $x_0$ | $z_{00}$ | $\cdots$ | $z_{0l}$ | | $z_{0,l+1}$ | $\cdots$ | $z_{0m}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_k$ | $z_{k0}$ | $\cdots$ | $z_{kl}$ | | $z_{k,l+1}$ | $\cdots$ | $z_{km}$ |
| $\bar{x}$ | $p_0(\bar{x})$ | $\cdots$ | $p_l(\bar{x})$ | $\boxed{q(\bar{y})}$ | $p_{l+1}(\bar{x})$ | $\cdots$ | $p_m(\bar{x})$ |
| $x_{k+1}$ | $z_{k+1,0}$ | $\cdots$ | $z_{k+1,l}$ | | $z_{k+1,l+1}$ | $\cdots$ | $z_{k+1,m}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_n$ | $z_{n0}$ | $\cdots$ | $z_{nl}$ | | $z_{n,l+1}$ | $\cdots$ | $z_{nm}$ |

A segunda alternativa consiste em realizar as operações

1. interpolando em $y$, obtém-se para cada $i$ o polinómio $q_i$ que interpola os valores $(z_{ij})_{j=0}^{m}$ nos nós $(y_j)_{j=0}^{m}$;
2. posteriormente, determina-se o polinómio $p$ que interpola os valores $q_i(\bar{y})$ nos nós $(x_i)_{i=0}^{n}$.

sendo agora o valor procurado $p(\bar{x})$. Neste caso dever-se-ão dispor os cálculos como se mostra na seguinte tabela.

| $f(x,y)$ | $y_0$ | $\cdots$ | $y_l$ | $\bar{y}$ | $y_{l+1}$ | $\cdots$ | $y_m$ |
|---|---|---|---|---|---|---|---|
| $x_0$ | $z_{00}$ | $\cdots$ | $z_{0l}$ | $q_0(\bar{y})$ | $z_{0,l+1}$ | $\cdots$ | $z_{0m}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_k$ | $z_{k0}$ | $\cdots$ | $z_{kl}$ | $q_k(\bar{y})$ | $z_{k,l+1}$ | $\cdots$ | $z_{km}$ |
| $\bar{x}$ | | | | $\boxed{p(\bar{y})}$ | | | |
| $x_{k+1}$ | $z_{k+1,0}$ | $\cdots$ | $z_{k+1,l}$ | $q_{k+1}(\bar{y})$ | $z_{k+1,l+1}$ | $\cdots$ | $z_{k+1,m}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $x_n$ | $z_{n0}$ | $\cdots$ | $z_{nl}$ | $q_n(\bar{y})$ | $z_{n,l+1}$ | $\cdots$ | $z_{nm}$ |

**Exemplo 7.9.1.**

*Considere a seguinte tabela de alguns valores de $z(x,y)$ conhecidos.*

| | | $y$ | | | |
|---|---|---|---|---|---|
| $z$ | | 1 | 2 | 4 | 6 |
| $x$ | 1 | 10 | 15 | 18 | 22 |
| | 2 | 7 | 12 | 15 | 20 |
| | 5 | 5 | 8 | 10 | 14 |

1. *Interpolando linearmente em $x$ e em $y$, estime o valor de $z(4,5)$*
   (a) *interpolando primeiro em $x$.*
   (b) *interpolando primeiro em $y$.*
2. *Estime agora $z(4,5)$ utilizando interpolação linear em $x$ e quadrática em $y$ e interpolando primeiro em $x$.*

***Resolução***

1. *Interpolação linear em $x$ e $y$, escolhendo para cada variável os dois pontos mais próximos.*
   (a) *Interpolando primeiro em $x$*

$$z_{01}(4,4) = \frac{\begin{vmatrix} z(2,4) & 4-2 \\ z(5,4) & 4-5 \end{vmatrix}}{2-5} = 11.6667$$

$$z_{01}(4,6) = \frac{\begin{vmatrix} z(2,6) & 4-2 \\ z(5,6) & 4-5 \end{vmatrix}}{2-5} = 16$$

| | $z$ | $y$ 1 | 2 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | 1 | 10 | 15 | 18 | | 22 |
| $x$ | 2 | 7 | 12 | 15 | | 20 |
| | 4 | | | | □ | |
| | 5 | 5 | 8 | 10 | | 14 |

*Interpolando agora em y os valores calculados, obtém-se*

$$z_{01}(4,5) = \frac{\begin{vmatrix} z_{01}(4,4) & 5-4 \\ z_{01}(4,6) & 5-6 \end{vmatrix}}{4-6} = \boxed{13.833.}$$

*(b) Interpolando primeiro em y*

$$z_{01}(2,5) = \frac{\begin{vmatrix} z(2,4) & 5-4 \\ z(2,6) & 5-6 \end{vmatrix}}{4-6} = 17.5$$

$$z_{01}(5,5) = \frac{\begin{vmatrix} z(5,4) & 5-4 \\ z(5,6) & 5-6 \end{vmatrix}}{4-6} = 12$$

| | $z$ | $y$ 1 | 2 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | 1 | 10 | 15 | 18 | | 22 |
| $x$ | 2 | 7 | 12 | 15 | | 20 |
| | 4 | | | | □ | |
| | 5 | 5 | 8 | 10 | | 14 |

*Interpolando agora em x os valores calculados, obtém-se*

$$z_{01}(4,5) = \frac{\begin{vmatrix} z_{01}(2,5) & 4-2 \\ z_{01}(5,6) & 4-5 \end{vmatrix}}{2-5} = \boxed{13.833.}$$

*A obtenção do mesmo valor fazendo as interpolações nas duas variáveis por ordem diferente terá sido coincidência?*

2. *Interpolação linear em x e quadrática em y, escolhendo para cada variável os pontos mais próximos.*

   *Interpolando primeiro em x*

$$z_{01}(4,2) = \frac{\begin{vmatrix} z(2,2) & 4-2 \\ z(5,2) & 4-5 \end{vmatrix}}{2-5} = 9.3333$$

$$z_{01}(4,4) = \frac{\begin{vmatrix} z(2,4) & 4-2 \\ z(5,4) & 4-5 \end{vmatrix}}{2-5} = 11.6667$$

$$z_{01}(4,6) = \frac{\begin{vmatrix} z(2,6) & 4-2 \\ z(5,6) & 4-5 \end{vmatrix}}{2-5} = 16$$

| | | $y$ | | | | |
|---|---|---|---|---|---|---|
| $z$ | | 1 | 2 | 4 | 5 | 6 |
| $x$ | 1 | 10 | 15 | 18 | | 22 |
| | 2 | 7 | 12 | 15 | □ | 20 |
| | 4 | | | | | |
| | 5 | 5 | 8 | 10 | | 14 |

*Interpolando agora em y os valores calculados, obtém-se*

$$z_{01}(4,5) = \frac{\begin{vmatrix} z_{01}(4,2) & 5-2 \\ z_{01}(4,4) & 5-4 \end{vmatrix}}{2-4} = 12.8333 \qquad z_{12}(4,5) = \frac{\begin{vmatrix} z_{01}(4,4) & 5-4 \\ z_{01}(4,6) & 5-6 \end{vmatrix}}{4-6} = 13.8333$$

$$z_{012}(4,5) = \frac{\begin{vmatrix} z_{01}(4,5) & 5-2 \\ z_{12}(4,5) & 5-6 \end{vmatrix}}{2-6} = \boxed{13.5833.}$$

***Nota:*** *Em todos os cálculos foi utilizada a forma de Aitken-Neville uma vez que em cada caso apenas é necessário calcular o valor do polinómio interpolador num ponto.*

## 7.10 Erro de interpolação

Se os valores nodais a interpolar corresponderem a valores de uma dada função $f$, pode ser interessante analisar em que medida o polinómio interpolar se aproxima da função, obviamente que em pontos distintos dos nós de interpolação.

O resultado apresentado em seguida generaliza o conhecido teorema do valor médio que permite concluir a existência de um ponto onde a tangente ao gráfico da de uma função é paralela a uma dada recta secante. De facto, fazendo $k = 1$ no enunciado do resultado abaixo obtém-se directamente aquele teorema pois $f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0}$. (Relembremos que as diferenças divididas dos valores da função $f$ são representadas por $f[\ldots].$)

**Teorema 7.10.1.** *Sejam $f \in C^k([a,b];\mathbb{R})$ e $(x_i)_{i=0}^k$ um conjunto de nós distintos em $[a,b]$. Então existe $\xi \in [a,b]$ tal que*

$$f[x_0, x_1, \ldots, x_k] = \frac{1}{k!} f^{(k)}(\xi).$$

*Demonstração.* Seja $p$ o polinómio de grau menor ou igual a $k$ que interpola $f$ nos nós distintos $(x_i)_{i=0}^k$. Então, a função $e = f - p$ tem pelo menos $k+1$ zeros distintos em $[a,b]$. Logo

$$\begin{array}{ll} e' = f' - p' & \text{tem pelo menos } k \text{ zeros distintos em } [a,b], \\ e^{(2)} = f^{(2)} - p^{(2)} & \text{tem pelo menos } k-1 \text{ zeros distintos em } [a,b], \\ \ldots & \\ e^{(k)} = f^{(k)} - p^{(k)} & \text{tem pelo menos 1 zero em } [a,b], \end{array}$$

ou seja, existe $\xi \in [a,b]$ tal que $f^{(k)}(\xi) = p^{(k)}(\xi)$.

Designando por $a_k$ o coeficiente de $x^k$ em $p$ verifica-se que $p^{(k)}(x) \equiv k! a_k$.

Da forma de Newton do polinómio interpolador verifica-se que $a_k = f[x_0, x_1, \ldots, x_k]$, concluindo-se então que $k! f[x_0, x_1, \ldots, x_k] = f^{(k)}(\xi)$, como pretendido. □

O teorema seguinte permite estimar o erro cometido ao aproximar uma função $f$ por um polinómio interpolador dessa função, habitualmente designado por **erro de interpolação**.

**Teorema 7.10.2.** *Sejam $f \in C^{n+1}([a,b];\mathbb{R})$ e $p$ o polinómio de grau menor ou igual a $n$ que interpola $f$ nos nós distintos $(x_i)_{i=0}^n$, pertencentes a $[a,b]$. Então, para qualquer $x \in [a,b]$ existe $\xi \in [a,b]$ tal que*

$$e(x) \equiv f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) W_n(x),$$

*onde $W_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$.*

*Demonstração.* Seja $\bar{x}$ um qualquer elemento de $[a,b]$.

Se $\bar{x} = x_i$ para algum $i$, o erro é nulo e o teorema é verdadeiro, pois $W_n(x_i) = 0$, $\forall i$.

Suponha-se agora que $\bar{x}$ é distinto de $(x_i)_{i=0}^n$. O polinómio $q$ de grau menor ou igual a $n+1$, que interpola $f$ nos nós $x_0, x_1, \ldots, x_n$ e $\bar{x}$, pode ser escrito como (relembrar a forma de Newton)

$$q(x) = p(x) + f[x_0, x_1, \ldots, x_n, \bar{x}] W_n(x).$$

Desta expressão resulta que $f(\bar{x}) = q(\bar{x}) = p(\bar{x}) + f[x_0, x_1, \ldots, x_n, \bar{x}] W_n(\bar{x})$.

Como já visto, existe $\xi \in [a,b]$ tal que $f[x_0, x_1, \ldots, x_n, \bar{x}] = \frac{1}{(n+1)!} f^{(n+1)}(\xi)$, obtendo-se finalmente que

$$e(\bar{x}) = f(\bar{x}) - p(\bar{x}) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) W_n(\bar{x})$$

como se pretendia mostrar. □

Na expressão do erro de interpolação

$$e(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) W_n(x),$$

o ponto $\xi$ (dependente de $x$ e dos nós de interpolação) é desconhecido, sendo usual considerar uma das seguintes majorações do erro absoluto

$$|e(x)| \leq \frac{1}{(n+1)!} \cdot \max_{z\in[a,b]} |f^{(n+1)}(z)| \cdot |W_n(x)|,$$

ou

$$|e(x)| \leq \frac{1}{(n+1)!} \cdot \max_{z\in[a,b]} |f^{(n+1)}(z)| \cdot \max_{z\in[a,b]} |W_n(z)|.$$

Supondo os nós ordenados de forma crescente (o que não acarreta qualquer perda de generalidade) e sendo $h$ o espaçamento máximo entre dois nós consecutivos, a majoração de $|W_n|$ no intervalo $[x_0, x_n]$ conduz ainda ao seguinte majorante do erro absoluto

$$|e(x)| \leq \frac{h^{n+1}}{4(n+1)} \cdot \max_{z\in[a,b]} |f^{(n+1)}(z)|$$

que é válida para todo o $x \in [x_0, x_n]$.

**Exemplo 7.10.1.** *Pretende-se construir uma tabela da função $f(x) = \tan(x)$ no intervalo $[0, \frac{\pi}{4}]$ com nós equidistantes, por forma a que o erro absoluto cometido quando se interpola linearmente nesta tabela não exceda $5 \times 10^{-5}$. Qual o espaçamento mínimo entre os nós?*

***Resolução***
*O erro máximo absoluto na interpolação linear entre nós consecutivos será*

$$\varepsilon \leq \tfrac{h^2}{4\times 2} \cdot \max |f''|$$

*Tem-se ainda que $f''(x) = [\tan(x)]'' = [1+\tan^2(x)]' = 2\tan(x)(1+\tan^2(x))$, cujo valor máximo em $[0, \frac{\pi}{4}]$ é 4 (para $x = \frac{\pi}{4}$). Para obter o erro máximo desejado bastará impor a condição*

$$\tfrac{h^2}{4\times 2} \times 4 \leq 5 \times 10^{-5}$$

*obtendo-se $h \leq 10^{-2}$, o que corresponde a um número de intervalos superior a $\frac{\pi}{4\times 10^{-2}} \approx 78.5$, ou seja, será usada uma tabela com 80 pontos (incluindo os extremos) espaçados de $\frac{\pi}{4\times 79}$.*

Uma questão que surge com alguma naturalidade é a de saber se à medida que se aumenta o número de nós de interpolação, o polinómio interpolador "converge" para a função a interpolar, ou seja se o erro de interpolação diminui à medida que o grau do polinómio interpolador aumenta.

**Exemplo 7.10.2.** *Seja $f : [-1, 1] \to \mathbb{R}$ definida por*

$$f(x) = \frac{1}{1+25x^2}.$$

*Tomando como nós de interpolação os pontos $-1 + \frac{i}{2}$, $(i = 0, \ldots, 4)$, obtém-se o polinómio interpolador*

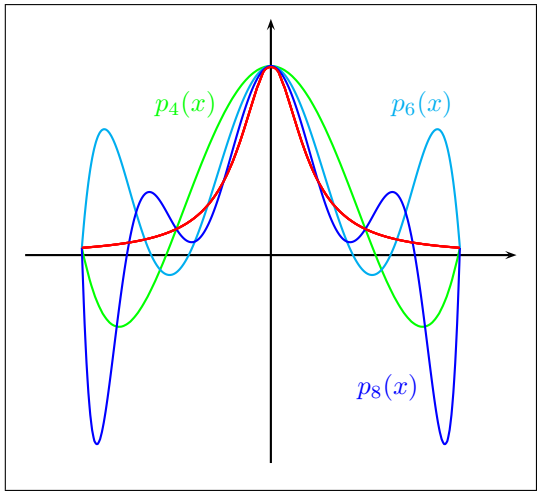
$$p_4(x) = \frac{1250}{377}x^4 - \frac{3225}{754}x^2 + 1.$$

*Interpolando nos nós* $-1+\frac{i}{3}$, $(i=0,\ldots,6)$, *obtém-se o polinómio interpolador*

$$p_6(x) = -\frac{1265625}{96356}x^6 + \frac{2019375}{96356}x^4 - \frac{211600}{24089}x^2 + 1.$$

*Interpolando agora nos nós* $-1+\frac{i}{4}$, $(i=0,\ldots,8)$, *obtém-se o polinómio interpolador*

$$p_8(x) = \frac{200000000}{3725137}x^8 - \frac{383000000}{3725137}x^6 + \frac{228601250}{3725137}x^4 - \frac{98366225}{7450274}x^2 + 1.$$




*Função interpolada e polinómios interpoladores.*

*Como se pode depreender da análise da figura, aumentando o número de nós e mantendo-os equidistantes verifica-se que os polinómios interpoladores apresentam cada vez maiores oscilações. Este comportamento continua a manter-se continuando a aumentar o número de nós. Verifica-se assim que o polinómios interpoladores não se aproximam cada vez mais da função a interpolar como seria desejável.*

Neste exemplo, à medida que o número de nós aumenta, o erro de interpolação não converge para 0, verificando-se que os polinómios interpoladores apresentam "oscilações" de amplitudes crescentes. Este comportamento deve-se à habitualmente designada **rigidez dos polinómios**, que se traduz no eventual aparecimento de oscilações quando se obriga um polinómio a passar por determinados pontos.

Este tipo de comportamento é bastante indesejável quando se pretendem utilizar polinómios interpoladores para aproximar funções. Analisando a expressão do erro de interpolação pode concluir-se que este comportamento pode ser causado quer pelo aumento dos valores das derivadas de ordem superior da função $f$ a interpolar, quer pelo aumento dos valores dos polinómios nodais $W_i$. Se, para um dado problema de aproximação por interpolação polinomial, os valores que tomam as derivadas de $f$ são algo que não se pode contornar, já os polinómios nodais podem alterados bastando para isso alterar a localização dos nós de interpolação. Na verdade, é

possível escolher os nós de interpolação de forma a que os polinómios nodais $W_i$ tomem valores tão pequenos quanto possível. Outra forma de evitar este comportamento será utilizar funções interpoladoras não polinomiais.

## 7.11 Polinómios de Chebyshev e nós de interpolação

A escolha dos nós de interpolação de forma a garantir que os polinómios nodais tomam valores pequenos deverá ser realizada fazendo-os coincidir com as raízes dos designados polinómios de Chebyshev, como se expõe em seguida.

Sendo $x \in [-1, 1]$ e $n = 0, 1, \ldots$, defina-se a função

$$T_n(x) = \cos(n \arccos x).$$

Facilmente se verifica que $T_n(x)$ é uma função polinomial de grau $n$!!

Fazendo $\theta = \arccos x$ pode escrever-se

$$\begin{aligned} T_n(x) &= \cos(n\theta) \\ T_{n+1}(x) &= \cos((n+1)\theta)) = \cos(\theta)\cos(n\theta) - \sin(\theta)\sin(n\theta) \\ T_{n-1}(x) &= \cos((n-1)\theta)) = \cos(\theta)\cos(n\theta) + \sin(\theta)\sin(n\theta) \end{aligned}$$

verificando-se então

$$T_{n+1}(x) + T_{n-1}(x) = 2\cos(theta)\cos(n\theta) = 2xT_n(x)$$

obtendo-se a expressão de recorrência

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

que juntamente com as condições

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \end{aligned}$$

permite concluir que $T_n(x)$ é de facto uma função polinomial em $[-1, 1]$.

Os polinómios $T_n(x)$ são designados **polinómios de Chebyshev**. Alguns destes polinómios são

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \end{aligned}$$

Uma das características desta família de polinómios é o facto de para $n \geq 1$, o coeficiente de $x^n$ em $T_n(x)$ ser $2^{n-1}$, isto é,

$$T_n(x) = 2^{n-1}x^n + \cdots$$

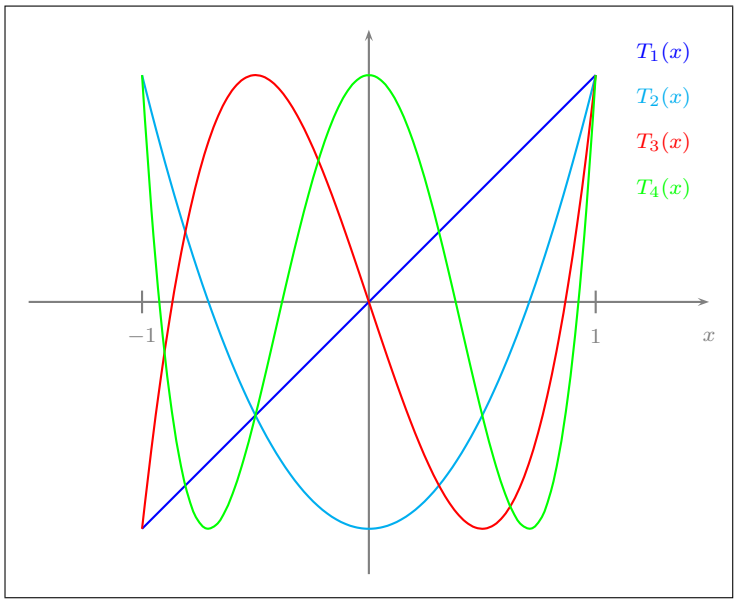



Figura 7.5: Polinómios de Chebyshev.

O seguinte resultado estabelece a localização das raízes dos polinómios de Chebyshev e dos seus valores extremos.

**Teorema 7.11.1.** *O polinómio $T_n(x)$ tem $n$ raízes simples em*

$$\bar{x}_k = \cos\left[\frac{(2k+1)\pi}{2n}\right] \qquad k = 0, 1, \ldots, n-1$$

*e toma valores extremos em*

$$\bar{x}'_k = \cos\left[\frac{k\pi}{n}\right] \quad com \quad T_n(\bar{x}'_k) = (-1)^k \qquad k = 0, 1, \ldots, n.$$

Para $n \geq 1$, o polinómio de grau $n$ $2^{1-n}T_n(x)$ tem coeficiente de $x^n$ unitário (diz-se polinómio **mónico**) e, de acordo com o resultado anterior, satisfaz a condição

$$\max_{x\in[-1,1]} |2^{1-n}T_n(x)| = \frac{1}{2^{n-1}}.$$

Se $P_n(x)$ for um polinómio mónico de grau $n$, tem-se

$$\max_{x\in[-1,1]} |P_n(x)| \geq \frac{1}{2^{n-1}}.$$

Consideremos agora a interpolação de uma função $f$ por um polinómio de grau $n$ nos nós $(x_i)_{i=0}^n$ de $[-1,1]$. Como já visto, o erro de interpolação será dado, para $x \in [-1,1]$, por

$$e(x) = f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\zeta) W_n(x)$$

onde $\zeta \in [-1,1]$ e $W_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$. Esta expressão realça a dependência do erro de interpolação relativamente aos nós de interpolação, que poderão não estar à partida definidos.

Uma vez que $W_n(x)$ é um polinómio mónico de grau $n+1$, ter-se-á

$$\max_{x \in [-1,1]} |W_n(x)| \geq \frac{1}{2^n}$$

sendo este valor o menor possível quando $W_n(x) = \frac{1}{2^n} T_{n+1}(x)$, ou seja, quando os nós de interpolação forem os zeros de $T_{n+1}(x)$, isto é

$$x_i = \cos \frac{(2i+1)\pi}{2(n+1)} \qquad i = 0, 1, \ldots, n.$$

De acordo com o exposto acima, se $p(x)$ for o polinómio de grau menor ou igual a $n$ que interpola $f(x)$ nos nós que são as raízes de $T_{n+1}(x)$ então o erro de interpolação pode ser majorado pela expressão

$$\max_{x \in [-1,1]} |p(x) - f(x)| \leq \frac{1}{2^n (n+1)!} \max_{z \in [-1,1]} |f^{(n+1)}(z)|.$$

Caso se pretenda aproximar $f$ por um polinómio interpolador de grau $n$ num intervalo $[a,b]$ que não o $[-1,1]$, os nós de interpolação que conduzem conduzem ao menor valor máximo de $W_n(x)$ no intervalo $[a,b]$ serão os pontos

$$\tilde{x}_i = \frac{1}{2} \left[ (b-a) \bar{x}_i + a + b \right]$$

onde $\bar{x}_i$ $(i = 0, 1, \ldots, n)$ são os zeros de $T_{n+1}(x)$.

## 7.12 Interpolação polinomial segmentada (splines)

Consideremos novamente a questão de interpolar uma função $f$ num intervalo $[a,b]$. Em diversas situações de utilização de polinómios interpoladores não se verifica a convergência para 0 do erro de interpolação à medida que se consideram mais nós, isto é, polinómios de mais grau mais elevado. Por outro lado, nem sempre é vantajoso do trabalhar com polinómios de grau elevados, pois a sua avaliação num ponto utilizando aritmética finita está sujeita a erros de arredondamento.

Uma alternativa será utilizar funções interpoladoras que não sejam de classe $C^\infty$. Particularmente interessante é a utilização de funções polinomiais por segmentos, isto é, funções que em cada subintervalo sejam definidas por um polinómio, mas que em diferentes subintervalos possam ser definidas por diferentes polinómios.

**Definição 7.12.1.** *Uma função $S$ diz-se um* ***spline polinomial*** *de grau $m$ (onde $m \in \mathbb{N}$), relativo aos nós $a = x_0 < x_1 < \cdots < x_n = b$, quando*

1. *$S$ coincide com um polinómio $S_i$ de grau menor ou igual a $m$ em cada subintervalo $[x_{i-1}, x_i]$, $i = 1, \ldots, n$.*
2. *$S \in C^{m-1}([a, b]; \mathbb{R})$.*

Dados os nós $x_0 < x_1 < \cdots < x_n$, a definição do spline é feita à custa dos polinómios $S_i$, que caracterizam $S$ nos diferentes intervalos $[x_{i-1}, x_i]$. Sendo as funções polinomiais de classe $C^\infty$, a condição 2 é sempre válida no interior de cada subintervalo, pelo que apenas é necessário verificá-la nos nós $x_1, \ldots, x_{n-1}$.

Dado um conjunto de nós $x_0 < x_1 < \cdots < x_n$ e os valores nodais $y_0, y_1, \ldots, y_n$ respectivos, a **interpolação por splines de grau** $m$ consiste em encontrar um spline $S$ de grau $m$ relativo aos nós $x_0 < x_1 < \cdots < x_n$ tal que

$$S(x_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

Tal como no caso da interpolação polinomial também agora se colocam algumas questões importantes às quais interessa responder, das quais se destacam as seguintes

- Será que existe spline interpolador?
- Será que o spline interpolador é único?
- Como se determinam os polinómios $S_i$ que definem o spline?
- Como se estima o erro na interpolação por splines de uma função?

Estas questões serão de alguma forma endereçadas no estudo que se segue sobre interpolação por splines.

### Spline de grau 1 ou linear

O spline $S$ coincide em cada subintervalo $[x_{i-1}, x_i]$ com o segmento de recta que passa pelos pontos $(x_{i-1}, y_{i-1})$ e $(x_i, y_i)$. Ou seja, os polinómios $S_i$, definidores do spline, satisfazem

$$\begin{aligned} S_i(x_{i-1}) &= y_{i-1} && i = 1, \ldots, n, \\ S_i(x_i) &= y_i && i = 1, \ldots, n. \end{aligned}$$

de onde resultam $2n$ equações. Sendo cada $S_i$ um polinómio de de grau 1 o spline é definido por $2n$ coeficientes. Resulta daqui a existência e unicidade deste spline. Também facilmente se concluir que os polinómios definidores do spline serão dados por

$$S_i(x) = y_{i-1}\frac{x_i - x}{h_i} + y_i\frac{x - x_{i-1}}{h_i}$$

para $i = 1, 2, \ldots, n$. (**Nota:** nesta expressão e no que se segue, define-se $h_i = x_i - x_{i-1}$, para $i = 1, 2, \ldots, n$.)

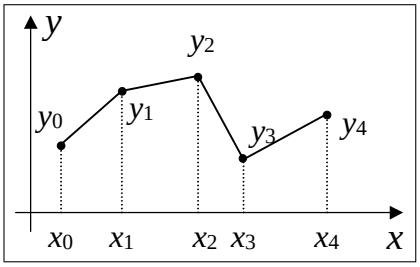



Figura 7.6: Spline linear.

Caso os valores nodais $y_i$ sejam dados por uma função, isto é, $y_i = f(x_i)$, onde $f$ é uma função de classe $C^2$, pode concluir-se que o erro de interpolação por um spline de grau 1 é majorado por

$$|e| \leq \frac{1}{8} \cdot |f''|\Big|_{\max} \cdot h^2$$

com $h = \max\{h_i : \ 1 \leq i \leq n\}$.

Esta expressão obtém-se directamente a partir da majoração do erro de interpolação polinomial para polinómios de grau menor ou igual a um.

### Spline de grau 2 ou quadrático

O spline coincide em cada intervalo $[x_{i-1}, x_i]$ com um arco de parábola. Estes arcos ligam-se de forma contínua, deverão passar pelos valores a interpolar e assegurar a continuidade da primeira derivada nos nós $x_1, x_2, \ldots, x_{n-1}$.

As condições a impor aos polinómios $S_i$, definidores do spline $S$, serão

$$S_i(x_{i-1}) = y_{i-1} \qquad i = 1, \ldots, n, \tag{7.12.1}$$

$$S_i(x_i) = y_i \qquad i = 1, \ldots, n, \tag{7.12.2}$$

$$S_i'(x_i) = S_{i+1}'(x_i) \qquad i = 1, \ldots, n-1, \tag{7.12.3}$$

que resultam em $3n - 1$ equações a satisfazer pelos coeficientes dos $S_i$. Neste caso, o spline será definido por $3n$ coeficientes. Conclui-se então que o spline quadrático não é único, pelo que será necessário impor uma condição adicional

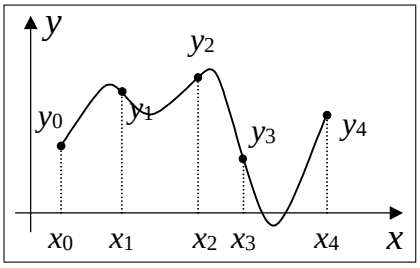



Figura 7.7: Spline quadrático.

Definido os polinómios $S_i$, $i = 1, \ldots, n$, por

$$S_i(x) = y_{i-1} + m_{i-1} \cdot (x - x_{i-1}) + \frac{M_i}{2} \cdot (x - x_{i-1})^2$$

garante-se, por construção,a satisfação de (7.12.1). Deste modo será necessário determinar os valores $m_i$ e $M_i$, para $i = 1, \ldots, n$, para definir completamente o spline.

Partindo de (7.12.2) e (7.12.3), é possível determinar os valores $m_i$ e $M_i$ de uma forma recorrente por intermédio das expressões

$$\begin{aligned} m_i &= 2 \cdot \frac{y_i - y_{i-1}}{h_i} - m_{i-1} && i = 1, \ldots, n, \\ M_i &= \frac{m_i - m_{i-1}}{h_i} && i = 1, \ldots, n. \end{aligned}$$

sendo necessário definir o valor adicional $m_0$, que corresponde a estipular a derivada do spline em $x_0$.

É de salientar o facto dos splines quadráticos serem pouco utilizados, por habitualmente apresentarem um comportamento com grandes oscilações.

### Spline de grau 3 ou cúbico

Em $[x_{i-1}, x_i]$ o spline $S$ coincide com um polinómio de grau menor ou igual a 3. Estas funções polinomiais ligam-se de forma contínua, deverão passar pelos valores a interpolar e assegurar a continuidade da primeira e segunda derivadas nos nós $x_1, x_2, \ldots, x_{n-1}$.

As condições a impor aos polinómios $S_i$, definidores do spline $S$, serão

$$\begin{aligned} &S_i(x_{i-1}) = y_{i-1} && i = 1, \ldots, n, \\ &S_i(x_i) = y_i && i = 1, \ldots, n, \\ &S_i'(x_i) = S_{i+1}'(x_i) && i = 1, \ldots, n-1, \\ &S_i''(x_i) = S_{i+1}''(x_i) && i = 1, \ldots, n-1. \end{aligned}$$

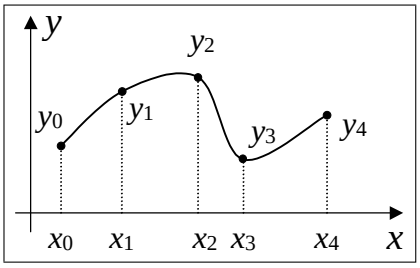



Figura 7.8: Spline cúbico.

Temos agora $4n-2$ condições e $4n$ coeficientes que definem o spline. Assim, o spline cúbico não será único, sendo necessárias duas condições adicionais.

Definindo $M_i = S''(x_i)$, para $i = 0, 1, \ldots, n$, a continuidade da segunda derivada fica assegurada fazendo-se

$$S_i''(x) = M_{i-1}\frac{x_i - x}{h_i} + M_i\frac{x - x_{i-1}}{h_i}.$$

Os parâmetros $M_i$ são habitualmente designados por **momentos**. Integrando esta expressão duas vezes obtém-se

$$S_i(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + \alpha_i x + \beta_i$$

onde $\alpha_i$ e $\beta_i$ são constantes de integração. Definindo $c_i = \alpha_i x_{i-1} + \beta_i$ e $d_i = \alpha_i x_i + \beta_i$, tem-se ainda

$$S_i(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + c_i\frac{x_i - x}{h_i} + d_i\frac{x - x_{i-1}}{h_i}$$

Impondo agora as condições $S_i(x_{i-1}) = y_{i-1}$ e $S_i(x_i) = y_i$, conclui-se que

$$c_i = y_{i-1} - \frac{M_{i-1}h_i^2}{6} \qquad \text{e} \qquad d_i = y_i - \frac{M_i h_i^2}{6}.$$

Substituindo estes valores, conclui-se que os polinómios $S_i$ podem ser representados por

$$S_i(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + \left(y_{i-1} - \frac{M_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(y_i - \frac{M_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}$$

Impondo a continuidade da primeira derivada nos nós interiores conclui-se que

$$\frac{h_i}{6}M_{i-1} + \frac{h_i + h_{i+1}}{3}M_i + \frac{h_{i+1}}{6}M_{i+1} = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}$$

para $i = 1, 2, \ldots, n-1$, obtendo-se deste modo um sistema de $n-1$ equações lineares com $n+1$ incógnitas, que são os momentos $M_0, M_1, \ldots, M_n$.

Habitualmente, as duas condições a impor para definir univocamente o spline são $M_0 = 0$ e $M_n = 0$ (anulamento da segunda derivada no primeiro e no último nó). Neste caso, diz-se que o spline é **natural**.

Os splines cúbicos são bastante utilizados como funções interpoladoras. Tendo por base polinómios de grau 3 são funções de fácil avaliação num ponto e também garantem a continuidade da segunda derivada. Às suas propriedades há ainda a juntar a descrita no resultado seguinte.

**Teorema 7.12.1.** *Sejam os nós $a = x_0 < \cdots < x_n = b$ e os valores nodais $y_0, \ldots, y_n$. Então, de todas as funções $g \in C^2([a,b];\mathbb{R})$ que interpolam estes pontos, o spline cúbico natural é a única que torna mínimo o valor de*

$$\int_a^b [g''(x)]^2 dx.$$

Caso os valores nodais obedeçam a $y_i = f(x_i)$, onde $f$ é uma função de classe $C^4$, o erro de interpolação por um spline cúbico é majorado por

$$|e| \leq \frac{5}{384} \cdot |f^{(4)}|_{\max} \cdot h^4.$$

**Exemplo 7.12.1.** *Interpolar a função*

$$f(x) = \frac{1}{1+25x^2} \qquad x \in [-1,1]$$

*por splines polinomiais, utilizando 7 pontos do intervalo $[-1,1]$ equidistantes.*

***Resolução***

*Os valores a interpolar serão*

| $x$ | $-1$ | $-\frac{2}{3}$ | $-\frac{1}{3}$ | $0$ | $\frac{1}{3}$ | $\frac{2}{3}$ | $1$ |
|---|---|---|---|---|---|---|---|
| $y$ | $\frac{1}{26}$ | $\frac{9}{109}$ | $\frac{9}{34}$ | $1$ | $\frac{9}{34}$ | $\frac{9}{109}$ | $\frac{1}{26}$ |

*Interpolando por um spline linear obtém-se*

$$\begin{aligned}
S_1(x) &= 0.17078 + 0.13232x, && x \in [-1, -\tfrac{2}{3}] \\
S_2(x) &= 0.44684 + 0.54641x, && x \in [-\tfrac{2}{3}, -\tfrac{1}{3}] \\
S_3(x) &= 1 + 2.20588x, && x \in [-\tfrac{1}{3}, 0] \\
S_4(x) &= 1 - 2.20588x, && x \in [0, \tfrac{1}{3}] \\
S_5(x) &= 0.44684 - 0.54641x, && x \in [\tfrac{1}{3}, \tfrac{2}{3}] \\
S_6(x) &= 0.17078 - 0.13232x, && x \in [\tfrac{2}{3}, 1]
\end{aligned}$$

*Interpolando por um spline quadrático (e considerando $m_0 = 0$) obtém-se*

| $i$ | 0 | 1 | 3 | 2 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $m_i$ | 0 | 0.26464 | 0.82818 | 3.58359 | $-7.99535$ | 6.90253 | $-7.16717$ |
| $M_i$ | – | 0.79393 | 1.69061 | 8.26622 | $-34.73681$ | 44.69364 | $-42.20910$ |

$$
\begin{aligned}
S_1(x) &= 0.43543 + 0.79393x + 0.39697x^2, && x \in [-1, -\tfrac{2}{3}] \\
S_2(x) &= 0.63469 + 1.39171x + 0.84530x^2, && x \in [-\tfrac{2}{3}, -\tfrac{1}{3}] \\
S_3(x) &= 1 + 3.58359x + 4.13311x^2, && x \in [-\tfrac{1}{3}, 0] \\
S_4(x) &= 1 + 3.58359x - 17.36841x^2, && x \in [0, \tfrac{1}{3}] \\
S_5(x) &= 5.41280 - 22.89323x + 22.34682x^2, && x \in [\tfrac{1}{3}, \tfrac{2}{3}] \\
S_6(x) &= -13.89892 + 35.04193x - 21.10455x^2, && x \in [\tfrac{2}{3}, 1]
\end{aligned}
$$

*A interpolação por um spline cúbico natural ($M_0 = 0$ e $M_6 = 0$) passa pela resolução do seguinte sistema de equações*

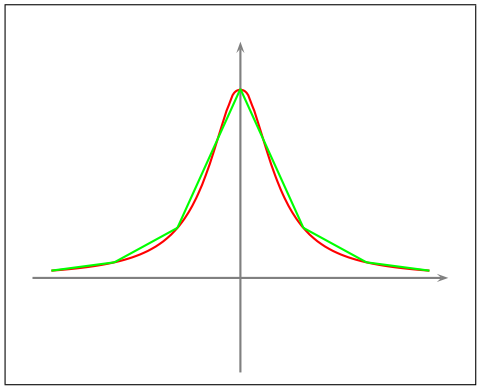
$$
\begin{bmatrix}
\frac{2}{9} & \frac{1}{18} & 0 & 0 & 0 \\
\frac{1}{18} & \frac{2}{9} & \frac{1}{18} & 0 & 0 \\
0 & \frac{1}{18} & \frac{2}{9} & \frac{1}{18} & 0 \\
0 & 0 & \frac{1}{18} & \frac{2}{9} & \frac{1}{18} \\
0 & 0 & 0 & \frac{1}{18} & \frac{2}{9}
\end{bmatrix}
\times
\begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \end{bmatrix}
=
\begin{bmatrix} \frac{9975}{24089} \\ \frac{3075}{1853} \\ -\frac{75}{17} \\ \frac{3075}{1853} \\ \frac{9975}{24089} \end{bmatrix}
$$

*cuja solução é*

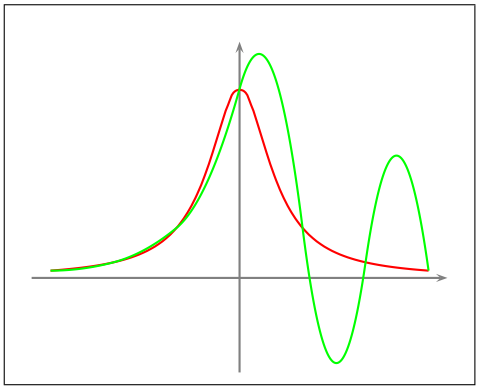
$$
\begin{aligned}
[M_1 \quad M_2 \quad M_3 \quad M_4 \quad M_5]^T &= \\
&= [-1.81814 \quad 14.72616 \quad -27.21602 \quad 14.72616 \quad -1.81814]^{\mathrm{T}}.
\end{aligned}
$$

*Os polinómios definidores do spline cúbico serão*

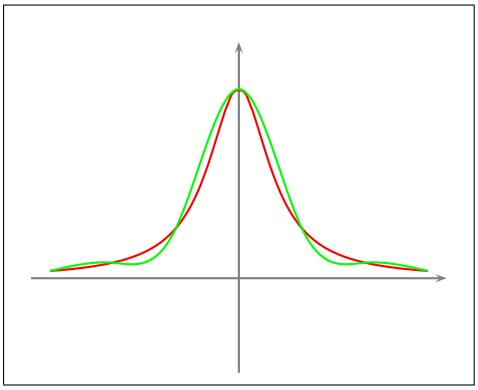
$$
\begin{aligned}
S_1(x) &= -0.63728 - 2.49388x - 2.72721x^2 - 0.90907x^3, && x \in [-1, -\tfrac{2}{3}] \\
S_2(x) &= 2.08308 + 9.74775x + 15.63523x^2 + 8.27215x^3, && x \in [-\tfrac{2}{3}, -\tfrac{1}{3}] \\
S_3(x) &= 1 - 13.60801x^2 - 20.97109x^3, && x \in [-\tfrac{1}{3}, 0] \\
S_4(x) &= 1 - 13.60801x^2 + 20.97109x^3, && x \in [0, \tfrac{1}{3}] \\
S_5(x) &= 2.08308 - 9.74775x + 15.63523x^2 - 8.27215x^3, && x \in [\tfrac{1}{3}, \tfrac{2}{3}] \\
S_6(x) &= -0.63728 + 2.49388x - 2.72721x^2 + 0.90907x^3, && x \in [\tfrac{2}{3}, 1]
\end{aligned}
$$



*Spline linear*



*Spline quadrático*



*Spline cúbico*

*Como se pode verificar, os splines linear e cúbico constituem boas aproximações da função $f$, este último com propriedades de continuidade das duas primeiras derivadas. É de relembrar que a aproximação desta função por polinómios interpoladores em nós equidistantes se torna muito problemática.*

# Capítulo 8

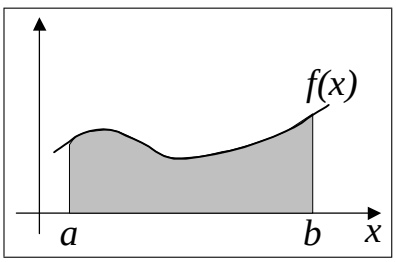
# Integração Numérica

## 8.1 Introdução

Em diversas aplicações é necessário calcular o integral definido de uma função $f$ para a qual não se conhece uma expressão explícita de uma primitiva, tal primitiva é de obtenção dispendiosa ou quando não se conhece uma expressão para a própria função. Nestas situações, pode ser utilizada a designada **integração numérica** que consiste em aproximar

$$I(f) = \int_a^b f(x)dx,$$

utilizando apenas valores da função $f$ num conjunto finito de pontos no intervalo $[a, b]$.




De uma forma geral, pode dizer-se que os métodos de integração numérica consistem em aproximar a função $f$ por outra função $g$ cuja primitivação seja simples de realizar. Desta forma, o integral de $f$ será aproximado por

$$I(f) \simeq I(g) = \int_a^b g(x)dx.$$

O erro cometido neste processo, representado por $E(f)$, é dado por

$$E(f) = I(f) - I(g) = I(f - g)$$

uma vez que a integração é um operador linear. Assim, a aproximação será tanto melhor quanto melhor a função $g$ aproximar $f$ no intervalo $[a, b]$.

## 8.2 Regras de integração básicas e compostas

Dado que as funções polinomiais são simples de integrar, a utilização de polinómios interpoladores com funções aproximantes constitui uma abordagem interessante ao problema de integração numérica.

As **regras de integração básicas** consistem em aproximar o integral de $f$ em $[a,b]$ pelo integral de um polinómio interpolador de $f$ num conjunto de nós em $[a,b]$. Designemos por $p_n$ o polinómio de grau menor ou igual a $n$ que interpola $f$ nos nós $x_0 < x_1 < \cdots < x_n$, pertencentes a $[a,b]$. Representando este polinómio na forma de Lagrange, obtém-se

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x)$$

onde $L_i$ são os polinómios de Lagrange relativos aos nós considerados. Então

$$I(p_n) = \int_a^b p_n(x)dx = \int_a^b \left( \sum_{i=0}^{n} f(x_i) L_i(x) \right) = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)dx$$

Definindo, para $i = 0, 1, \ldots, n$, $A_i = \int_a^b L_i(x)dx$, verifica-se que

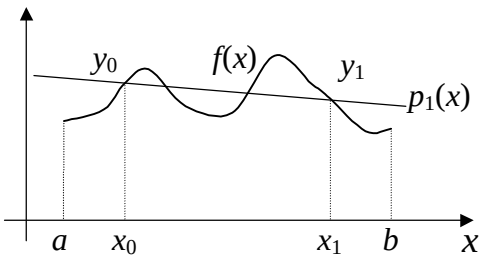
$$I(p_n) = \sum_{i=0}^{n} A_i f(x_i)$$

e logo o integral de $f$ será aproximado da seguinte forma

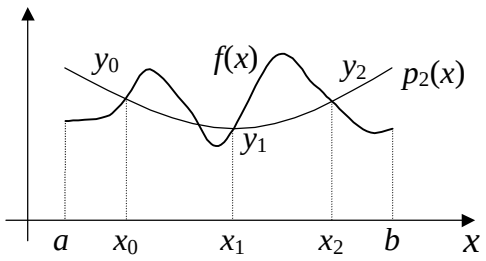
$$I(f) \simeq \sum_{i=0}^{n} A_i f(x_i)$$

ou seja, por uma combinação linear dos valores de $f$ nos nós. Os coeficientes desta combinação linear, também designados por **pesos**, apenas dependem dos nós escolhidos.

É claro que escolhendo diferentes números de nós e diferentes localizações destes se obtêm diferentes regras de integração. A aplicação das diferentes regras consiste numa primeira fase em determinara os pesos $A_i$, que apenas dependem dos nós escolhidos, e posteriormente em efectuar a combinação linear dos valores da função nos nós, de acordo com a expressão acima.




Polinómio interpolador em 2 nós




Polinómio interpolador em 3 nós

Diz-se que uma regra de integração é de **grau ou exactidão** $n$ se integrar exactamente todos os polinómios de grau menor ou igual a $n$ e existir pelo menos um polinómio de grau $n+1$ que não é integrado exactamente.

Uma consequência imediata desta definição é o facto de toda a regra de integração que resulte da aproximação de $f$ por um polinómio interpolador em $n+1$ nós ser de exactidão maior ou igual a $n$.

Relembrando que o erro na aproximação de $f$ pelo polinómio interpolador $p_n$ é dado por

$$e(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x]W_n(x),$$

onde $W_n(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$, conclui-se que o erro de integração, também designado por **erro de truncatura**, será

$$E(f) = \int_a^b e(x)dx = \int_a^b f[x_0, \ldots, x_n, x]W_n(x)dx.$$

**Nota**: As diferenças divididas de $f$ utilizadas nesta expressão deverão ser entendidas num sentido mais geral uma vez que $x$ não é necessariamente distinto dos nós considerados. No entanto, apenas importa referir aqui que sendo $f$ de classe $C^{n+1}$ se tem que $x \mapsto f[x_0, x_1, \ldots, x_n, x]$ é contínua e que existe $\xi \in [a,b]$ tal que $f[x_0, x_1, \ldots, x_n, x] = \frac{1}{(n+1)!}f^{(n+1)}(\xi)$.

A utilização de polinómios interpoladores de maior grau conduz a regras de integração básicas de maior exactidão. No entanto, os polinómios interpoladores podem apresentar comportamentos pouco desejáveis, em termos de aproximação da função interpolada. Por tal motivo, as regras básicas de integração com polinómios de grau elevado não são vulgarmente utilizadas, pois nem sempre se consegue reduzir o erro de integração quando se aumenta o grau do polinómio interpolador.

Para diminuir o erro de integração sem aumentar o grau dos polinómios interpoladores utilizam-se **regras de integração compostas**. Estas consistem em dividir o intervalo $[a,b]$ em sub-intervalos $[a_0, a_1]$, $[a_1, a_2]$, ..., $[a_{n-1}, a_n]$ (onde $a_0 = a$ e $a_n = b$). Em cada subintervalo $[a_{i-1}, a_i]$, $f$ é interpolada por um polinómio $p_i$, sendo o integral de $f$ em $[a,b]$ aproximado pela soma dos integrais dos polinómios interpoladores, cada um no subintervalo respectivo, ou seja,

$$I(f) = \int_a^b f(x)dx = \sum_{i=1}^{n} \int_{a_{i-1}}^{a_i} f(x)dx \simeq \sum_{i=1}^{n} \int_{a_{i-1}}^{a_i} p_i(x)dx.$$

O erro de interpolação neste tipo de regras pode ser controlado pela largura $h_i$ de cada sub-intervalo $[a_{i-1}, a_i]$, ou seja, $h_i = a_i - a_{i-1}$. Muitas vezes consideram-se sub-intervalos de igual largura, isto é, $h_i = h = \frac{b-a}{n}$, $\forall i$. Claro está que a diminuição das larguras dos subintervalos implica o aumento do seu número e logo o aumento do número de operações aritméticas na aplicação da regra.

Em seguida serão estudadas duas regras de integração compostas, a **regra dos trapézios** e a **regra de Simpson**. Em cada um dos casos será deduzida a expressão que permite calcular o valor aproximado do integral de $f$, sendo também estudado o erro de integração cometido.

## 8.3 Regra dos trapézios

Nesta regra, a função $f$ é aproximada em cada subintervalo pela função polinomial de grau menor ou igual a 1 que interpola $f$ nos extremos desse subintervalo.

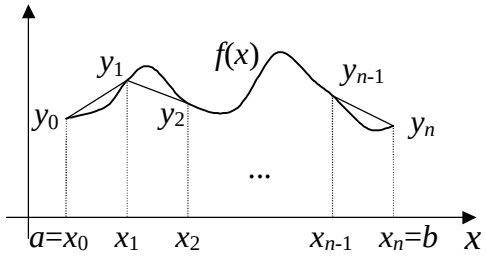



Figura 8.1: Regra dos trapézios.

Considerando $n$ subintervalos do intervalo original $[a, b]$, verifica-se que a largura de cada subintervalo é dada por $h = \frac{b-a}{n}$, sendo os extremos destes subintervalos os pontos $x_i = a + ih$, para $i = 0, 1, \ldots, n$. Designando por $y_i$ o valor de $f$ em $x_i$, o polinómio, de grau 1, que interpola $f$ nos pontos $x_i$ e $x_{i+1}$ é dado por

$$p_i(x) = y_i + \frac{y_{i+1} - y_i}{h}(x - x_i).$$

Integrando o polinómio $p_i$ subintervalo $[x_i, x_{i+1}]$ obtém-se (a área do trapézio!)

$$\int_{x_i}^{x_{i+1}} p_i(x)dx = \frac{h}{2}(y_i + y_{i+1}).$$

Somando estes valores para todos os subintervalos obtém-se

$$\sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} p_i(x)dx = \sum_{i=1}^{n} \frac{h}{2}(y_i + y_{i+1}) = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n),$$

pelo que a expressão que permite o cálculo aproximado do integral de $f$ em $[a, b]$ pela regra dos trapézios será

$$I(f) \simeq \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \cdots + 2y_{n-1} + y_n).$$

Passemos agora ao estudo do erro de truncatura. No intervalo $[x_i, x_{i+1}]$, o erro de aproximação de $f$ por $p_i$ é dado pela expressão (relembrar o erro de interpolação!)

$$e_i(x) = f(x) - p_i(x) = f[x_i, x_{i+1}, x](x - x_i)(x - x_{i+1}).$$

Então, o erro de aproximação de $\int_{x_i}^{x_{i+1}} f(x)dx$ por $\int_{x_i}^{x_{i+1}} p_i(x)dx$, $E_i$, será

$$E_i = \int_{x_i}^{x_{i+1}} e_i(x)dx = \int_{x_i}^{x_{i+1}} f[x_i, x_{i+1}, x](x - x_i)(x - x_{i+1})dx.$$

Assumindo que $f$ é de classe $C^2$, pode concluir-se que a função $x \mapsto f[x_i, x_{i+1}, x]$ é contínua. Por outro lado, a função $x \mapsto (x - x_i)(x - x_{i+1})$ não muda de sinal em $[x_i, x_{i+1}]$, sendo sempre não positiva. Então, existe $\zeta_i \in [x_i, x_{i+1}]$ tal que

$$\int_{x_i}^{x_{i+1}} f[x_i, x_{i+1}, x](x - x_i)(x - x_{i+1})dx = f[x_i, x_{i+1}, \zeta_i] \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1})dx$$

Efectuando a mudança de variável $z = x - x_i$, conclui-se que

$$\int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1})dx = \int_0^h z(z - h)dz = -\frac{h^3}{6}.$$

Pode ainda afirmar-se que $f[x_i, x_{i+1}, \zeta_i] = \frac{1}{2}f''(\xi_i)$, para algum $\xi_i \in [x_i, x_{i+1}]$. Desta forma pode escrever-se que o erro $E_i$ será

$$E_i = -\frac{1}{12}f''(\xi_i)h^3.$$

O erro de truncatura da regra dos trapézios obtém-se agora por

$$E(f) = \sum_{i=0}^{n-1} E_i = \sum_{i=0}^{n-1} \left( -\frac{1}{12}f''(\xi_i)h^3 \right) = -\frac{h^3}{12} \sum_{i=0}^{n-1} f''(\xi_i).$$

Dado que $f$ é suposta de classe $C^2$, é possível concluir a existência de $\xi \in [a, b]$ tal que

$$\sum_{i=0}^{n-1} f''(\xi_i) = nf''(\xi).$$

Então, a expressão do erro de truncatura da regra dos trapézios será

$$E(f) = -\frac{h^2}{12}(b - a)f''(\xi),$$

uma vez que $nh = b - a$. Como o ponto $\xi \in [a, b]$ é desconhecido, é usual utilizar o majorante do erro de truncatura em valor absoluto dado por

$$|E(f)| \leq \frac{h^2}{12}(b - a) \max_{z \in [a,b]} |f''(z)|.$$

Em algumas situações os valores $y_i$ estão eles mesmo afectados de erros de arredondamento que irão provocar um erro de arredondamento na aplicação da regra dos trapézios. Considerando que cada $y_i$ tem um erro absoluto máximo $\varepsilon$, o erro de arredondamento $\varepsilon_a$ satisfará a condição

$$\begin{aligned}
\varepsilon_a &\leq \sum_{i=0}^{n} \frac{\partial}{\partial y_i} \left[ \frac{h}{2}(y_0 + 2y_1 + \cdots + 2y_{n-1} + y_n) \right] \cdot \varepsilon \\
&= \frac{h}{2}(\varepsilon + 2\varepsilon + \cdots + 2\varepsilon + \varepsilon) \\
&= \frac{h}{2} \cdot 2n\varepsilon \\
&= (b - a)\varepsilon.
\end{aligned}$$

Um majorante para o **erro absoluto total**, na aplicação da regra dos trapézios será então

$$|E(f)| + \varepsilon_a.$$

**Exemplo 8.3.1.** *Sendo $f(x) = e^{-x^2}$, calcular um valor aproximado de $\int_0^1 f(x)dx$, utilizando a regra dos trapézios com 20 subintervalos e obter um majorante para o erro cometido (considere que os valores de $f$ são exactos). Qual o erro máximo absoluto admissível para os valores de $f$ se se pretender que o erro de arredondamento não seja superior ao erro de truncatura?*

***Resolução***
*Sendo $n = 20$, temos $h = \frac{1}{20}$ e a função será avaliada nos pontos $x_i = ih$, para $i = 0, 1, \ldots, 20$. O valor aproximado do integral será então*

$$\begin{aligned}\int_0^1 e^{-x^2}dx &\simeq \frac{\frac{1}{20}}{2}\left[e^0 + 2e^{-\left(\frac{1}{20}\right)^2} + \cdots + 2e^{-\left(\frac{19}{10}\right)^2} + e^{-1}\right] \\ &= \frac{\frac{1}{20}}{2}\left[e^0 + 2\sum_{i=1}^{19} e^{-\left(\frac{i}{20}\right)^2} + e^{-1}\right] \\ &= 0.7467.\end{aligned}$$

*Uma vez que $f''(x) = (4x^2 - 2)e^{-x^2}$ é majorada em valor absoluto no intervalo $[0, 1]$ por 2, conclui-se que o erro de truncatura será majorado por*

$$\frac{h^2}{12}(b-a)|f''|_{\max} = \frac{(1/20)^2}{12} \times 2 \simeq 4.2 \times 10^{-4},$$

*garantindo-se assim que o resultado apresentado terá 3 algarismos exactos.*

*Se se pretender que $\varepsilon_a \leq |E(f)|$ dever-se-á impor que*

$$(b-a)\varepsilon \leq 4.2 \times 10^{-4}$$

*pelo que $\varepsilon = 4.2 \times 10^{-4}$ será o erro máximo absoluto permitido no cálculo de cada valor de $f$, pois $b - a = 1$.*

## 8.4 Regra de Simpson

Na regra de Simpson a função $f$ é aproximada por polinómios de grau menor ou igual a 2, cada um dos quais interpolando $f$ em três pontos igualmente espaçados.

Agora, o número $n$ de subintervalos deverá ser **par**, pois cada parábola interpoladora é definida em dois subintervalos consecutivos. Definido novamente $h = \frac{b-a}{n}$, os extremos dos subintervalos serão os pontos $x_i = a + ih$, para $i = 0, 1, \ldots, n$. Designemos ainda por $y_i$ o valor de $f$ em $x_i$.

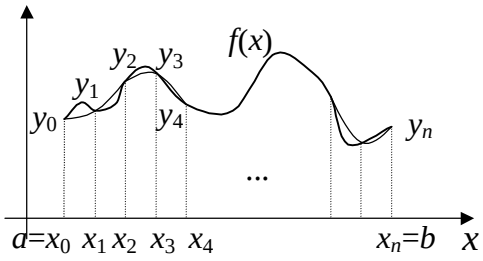



Figura 8.2: Regra de Simpson.

Seja também $p_i$ o polinómio de menor grau que interpola $f$ nos pontos $x_{i-1}$, $x_i$ e $x_{i+1}$, isto para $i = 1, 3, \ldots, n-1$. Tem-se então que

$$\begin{aligned} p_i(x) &= y_{i-1} + f[x_{i-1}, x_i](x - x_{i-1}) + f[x_{i-1}, x_i, x_{i+1}](x - x_{i-1})(x - x_i) \\ &= y_{i-1} + \frac{y_i - y_{i-1}}{h}(x - x_{i-1}) + \frac{\frac{y_{i+1}-y_i}{h} - \frac{y_i - y_{i-1}}{h}}{2h}(x - x_{i-1})(x - x_i). \end{aligned}$$

Integrando $p_i$ em $[x_{i-1}, x_{i+1}]$ obtém-se, efectuando a mudança de variável $z = x - x_{i-1}$,

$$\begin{aligned} \int_{x_{i-1}}^{x_{i+1}} p_i(x)dx &= \int_0^{2h} \left[ y_{i-1} + \frac{y_i - y_{i-1}}{h} z + \frac{y_{i+1} - 2y_i + y_{i-1}}{2h^2}(z^2 - hz) \right] dz \\ &= y_{i-1}2h + \frac{y_i - y_{i-1}}{h} 2h^2 + \frac{y_{i+1} - 2y_i + y_{i-1}}{2h^2} \left( \frac{8h^3}{3} - 2h^3 \right) \\ &= \frac{h}{6} \left( 12y_{i-1} + 12y_i - 12y_{i-1} + 2y_{i+1} - 4y_i + 2y_{i-1} \right) \\ &= \frac{h}{3} \left( y_{i-1} + 4y_i + y_{i+1} \right). \end{aligned}$$

Somando estes integrais para todos os sub-intervalos $[x_{i-1}, x_{i+1}]$, com $i = 1, 3, \ldots, n-1$, de forma a cobrir todo o intervalo $[a, b]$, obtém-se

$$\begin{aligned} \sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} \int_{x_{i-1}}^{x_{i+1}} p_i(x)dx &= \sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} \left[ \frac{h}{3} \left( y_{i-1} + 4y_i + y_{i+1} \right) \right] \\ &= \frac{h}{3} \left( y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 4y_{n-1} + y_n \right) \end{aligned}$$

resultando então a seguinte expressão para a regra de Simpson

$$I(f) \simeq \frac{h}{3} \left( y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \cdots + 4y_{n-1} + y_n \right).$$

Passemos agora ao estudo do erro de truncatura. No intervalo $[x_{i-1}, x_{i+1}]$, o erro de aproximação da função $f$ pelo polinómio $p_i$ é

$$e_i(x) = f(x) - p_i(x) = f[x_{i-1}, x_i, x_{i+1}, x](x - x_{i-1})(x - x_i)(x - x_{i+1}).$$

Então, o erro de aproximação de $\int_{x_{i-1}}^{x_{i+1}} f(x)dx$ por $\int_{x_{i-1}}^{x_{i+1}} p_i(x)dx$, $E_i$, será

$$E_i = \int_{x_{i-1}}^{x_{i+1}} e_i(x)dx = \int_{x_{i-1}}^{x_{i+1}} f[x_{i-1}, x_i, x_{i+1}, x](x - x_{i-1})(x - x_i)(x - x_{i+1})dx.$$

Supondo $f$ de classe $C^4$, demonstra-se ainda que

$$E_i = -\frac{h^5}{90} f^{(4)}(\xi_i)$$

para algum $\xi_i \in [x_{i-1}, x_{i+1}]$.

O erro de truncatura da regra de Simpson obtém-se agora por

$$E(f) = \sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} E_i = \sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} \left( -\frac{h^5}{90} f^{(4)}(\xi_i) \right) = -\frac{h^5}{90} \sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} f^{(4)}(\xi_i).$$

É também agora possível assegurar a existência de $\xi \in [a, b]$ tal que

$$\sum_{\substack{i=1 \\ i \text{ ímpar}}}^{n-1} f^{(4)}(\xi_i) = \frac{n}{2} f^{(4)}(\xi)$$

e como $nh = b - a$, a expressão do erro de truncatura da regra de Simpson fica

$$E(f) = -\frac{h^4}{180}(b - a) f^{(4)}(\xi).$$

Sendo o ponto $\xi \in [a, b]$ desconhecido, é usual utilizar o majorante do erro de truncatura em valor absoluto dado por

$$|E(f)| \leq \frac{h^4}{180}(b - a) \max_{z \in [a,b]} |f^{(4)}(z)|.$$

Supondo que os valores $y_i$ estão eles mesmo afectados de erros de arredondamento, cada um dos quais majorado em valor absoluto por $\varepsilon$, o erro de arredondamento $\varepsilon_a$ na regra de Simpson satisfará a condição

$$\begin{aligned}
\varepsilon_a &\leq \sum_{i=0}^{n} \frac{\partial}{\partial y_i} \left[ \frac{h}{3}(y_0 + 4y_1 + 2y_2 + 4y_3 \cdots + 4y_{n-1} + y_n) \right] \cdot \varepsilon \\
&= \frac{h}{3}(\varepsilon + 4\varepsilon + 2\varepsilon + 4\varepsilon + \cdots + 4\varepsilon + \varepsilon) \\
&= \frac{h}{3} \cdot \left( \varepsilon + \tfrac{n}{2} 4\varepsilon + \left( \tfrac{n}{2} - 1 \right) 2\varepsilon + \varepsilon \right) = \frac{h}{3} \cdot 3n\varepsilon \\
&= (b - a)\varepsilon.
\end{aligned}$$

Um majorante para o **erro absoluto total**, na aplicação da regra de Simpson, será então

$$|E(f)| + \varepsilon_a.$$

**Exemplo 8.4.1.** *Sendo $f(x) = e^{-x^2}$, calcular um valor aproximado de $\int_0^1 f(x)dx$, utilizando a regra de Simpson com 12 sub-intervalos e obter um majorante para o erro cometido (considerando que os valores de $f$ são exactos).*

***Resolução***

*Sendo $h = \frac{1}{12}$, a função será avaliada nos pontos $x_i = ih$, para $i = 0, 1, \ldots, 12$.*

*O valor aproximado do integral, pela regra de Simpson, será então*

$$\begin{aligned}\int_0^1 e^{-x^2}dx &\simeq \frac{\frac{1}{12}}{3}\left[e^0 + 4e^{-\left(\frac{1}{12}\right)^2} + 2e^{-\left(\frac{2}{12}\right)^2} + 4e^{-\left(\frac{3}{12}\right)^2} + \cdots + 2e^{-\left(\frac{10}{12}\right)^2} + 4e^{-\left(\frac{11}{12}\right)^2} + e^{-1}\right] \\ &= \frac{\frac{1}{12}}{3}\left[e^0 + 4\sum_{j=0}^{5} e^{-\left(\frac{2j+1}{12}\right)^2} + 2\sum_{j=0}^{4} e^{-\left(\frac{2j+2}{12}\right)^2} + e^{-1}\right] \\ &= 0.746825\end{aligned}$$

*Calculando $f^{(4)}(x)$ obtém-se*

$$f^{(4)}(x) = (16x^4 - 48x^2 + 12)e^{-x^2}$$

*que é majorada em valor absoluto no intervalo $[0, 1]$ por 12.*

*Conclui-se então que o erro de truncatura será majorado por*

$$\frac{h^4}{180}(b-a)|f^{(4)}|\Big|_{\max} = \frac{(1/12)^4}{180} \times 12 \simeq 3.2 \times 10^{-6}$$

*pelo que o resultado apresentado terá 5 algarismos exactos.*

## 8.5 Integração de Romberg

Nesta secção iremos apresentar uma técnica que permite obter resultados de maior precisão a partir de diversas aplicações da regra dos trapézios. O método exposto designa-se por **integração de Romberg** e constitui um caso particular da técnica designada por extrapolação de Richardson.

Consideremos o problema de aproximar o integral $I = \int_a^b f(x)dx$ por aplicação da regra dos trapézios. Sendo $h$ um valor positivo, tal que $\frac{b-a}{h}$ seja inteiro, designemos por $T(h)$ o valor aproximado de $I$ dado pela regra dos trapézios com subintervalos de largura $h$. É possível mostrar que

$$I = T(h) + K_1h^2 + K_2h^4 + K_3h^6 + \cdots, \tag{8.5.1}$$

onde $K_1, K_2, K_3, \ldots$ são constantes independentes de $h$. Desta expressão pode concluir-se que o erro de truncatura de $T(h)$ é de ordem 2, ou seja, converge para 0 à mesma taxa que $h^2$. Esta mesma conclusão podia já ser obtida a partir da expressão anteriormente obtida para o erro de

truncatura da regra dos trapézios. No entanto, a expressão acima permite ainda concluir que no erro de truncatura apenas aparecem termos com expoente de $h$ par.

Aplicando agora a regra dos trapézios com subintervalos de largura $\frac{h}{2}$ temos então que

$$I = T(h/2) + K_1(h/2)^2 + K_2(h/2)^4 + K_3(h/2)^6 + \cdots ,$$

ou ainda,

$$I = T(h/2) + \frac{K_1}{4}h^2 + \frac{K_2}{16}h^4 + \frac{K_3}{64}h^6 + \cdots . \tag{8.5.2}$$

Multiplicando por 4 a equação (8.5.2) e subtraindo-lhe a equação (8.5.1) obtém-se a equação

$$3I = 4T(h/2) - T(h) + K_2\left(\frac{1}{4} - 1\right)h^4 + K_3\left(\frac{1}{16} - 1\right)h^6 + \cdots .$$

Definindo, $T_1(h)$ por intermédio da expressão

$$T_1(h) = \frac{4T(h/2) - T(h)}{3}$$

pode concluir-se que $I = T_1(h) + K_1'h^4 + K_2'h^6 + \cdots$, pelo que $T_1(h)$ é uma aproximação de $I$ com um erro de truncatura de ordem 4.

Utilizando agora um procedimento semelhante para eliminar o termo em $h^4$ na expressão do erro de $T_1(h)$, define-se

$$T_2(h) = \frac{16T_1(h/2) - T_1(h)}{15}$$

e conclui-se facilmente que $I = T_2(h) + K_1''h^6 + K_2''h^8 + \cdots$, pelo que $T_2(h)$ é uma aproximação de $I$ com um erro de truncatura de ordem 6. Continuando este processo, podemos definir

$$T_3(h) = \frac{64T_2(h/2) - T_2(h)}{63}$$

concluindo-se que $I = T_3(h) + K_1'''h^8 + K_2'''h^{10} + \cdots$, sendo então o erro de truncatura de $T_3(h)$ de ordem 8.

De uma forma geral, podemos definir a aproximação $T_n(h)$ de uma forma recursiva por intermédio de

$$T_n(h) = \frac{4^nT_{n-1}(h/2) - T_{n-1}(h)}{4^n - 1},$$

concluindo-se que esta aproximação terá um erro de truncatura de ordem $2n + 2$.

Esta técnica de obtenção de aproximações de $I$ com ordens de erro cada vez mais elevadas permite em muitas circunstâncias obter valores aproximados do integral de uma função com elevada precisão e sem grandes esforços computacionais.

**Exemplo 8.5.1.** *Obter uma estimativa de* $\int_0^1 \frac{dx}{1+x^2}$ *com erro de ordem 8, utilizando um valor de inicial* $h = 0.25$.

***Resolução***

*Aplicando a regra dos trapézios com* $h = 0.25$, $h = 0.125$, $h = 0.0625$ *e* $h = 0.03125$, *obtêm-se os seguintes valores.*

| $h$ | $T(h)$ |
|---|---|
| 0.25 | 0.7827941176471 |
| 0.125 | 0.7847471236228 |
| 0.0625 | 0.7852354030103 |
| 0.03125 | 0.7853574732937 |

*Os valores extrapolados, obtidos por integração de Romberg, encontram-se na tabela seguinte*

| $h$ | $T(h)$ | $T_1(h)$ | $T_2(h)$ | $T_3(h)$ |
|---|---|---|---|---|
| 0.25 | 0.7827941176471 | 0.7853981256147 | 0.7853981652856 | 0.7853981633975 |
| 0.125 | 0.7847471236228 | 0.7853981628062 | 0.7853981634270 | — |
| 0.0625 | 0.7852354030103 | 0.7853981633882 | — | — |
| 0.03125 | 0.7853574732937 | — | — | — |

*O valor exacto deste integral é* $\frac{\pi}{4}$, *sendo aproximadamente* $0.78539816339744830963\ldots$, *pelo que o erro de truncatura de* $T_3(0.25)$ *é cerca de* $10^{-13}$. *Este erro é substancialmente inferior ao erro de truncatura de* $T(0.03125)$, *que é o valor obtido pela regra dos trapézios com maior precisão utilizada. Refira-se ainda que cada um dos cálculos dos valores* $T_1$, $T_2$ *e* $T_3$ *requer apenas* 3 *operações aritméticas, pelo que o maior esforço na obtenção de* $T_3(0.25)$ *está no cálculo de* $T(0.03125)$. *Por curiosidade, refira-se que para obter um valor de precisão semelhante a* $T_3(0.25)$ *por simples aplicação de uma regra dos trapézios exigiria um valor de* $h$ *de cerca de* $10^{-6}$, *ou seja, cerca de um milhão de subintervalos! Para efectuar tal cálculo seria necessário um número de operações aritméticas muito mais elevado. Para além do esforço na realização de tal cálculo deveriam ainda ser considerados eventuais erros de arredondamento resultantes da utilização de aritmética com precisão finita.*

## 8.6 Quadratura gaussiana

Nos métodos estudados atrás, os nós encontravam-se igualmente espaçados, sendo a sua localização apenas dependente do número de nós considerados e, claro, do intervalo de integração utilizado. Vamos agora estudar um método, designado por quadratura gaussiana, em que os nós não se encontram igualmente espaçados, sendo a sua localização um parâmetro de escolha.

A quadratura gaussiana consiste em efectuar a aproximação

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} c_i f(x_i)$$

sendo os nós $x_1, x_2, \ldots, x_n$ de $[a, b]$ e os coeficientes $c_1, c_2, \ldots, c_n$ escolhidos de forma a que a integração seja exacta para a maior classe de polinómios possível.

Havendo $2n$ parâmetros a definir e sendo um polinómio de grau $2n-1$ definido por um conjunto de $2n$ coeficientes, é de esperar que a quadratura gaussiana de ordem $n$ permita integrar com exactidão polinómios de grau até $2n-1$.

Comecemos por analisar o caso $n=2$. Para simplificar, é habitual considerar que a integração será efectuada no intervalo $[-1,1]$. Pretende-se então determinar $c_1$, $c_2$, $x_1$ e $x_2$ de modo que a relação

$$\int_{-1}^{1} f(x)dx = c_1 f(x_1) + c_2 f(x_2) \tag{8.6.1}$$

se verifique quando $f(x)$ for um polinómio de grau menor ou igual a $2\times 2-1=3$, ou seja,

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

onde $a_0$, $a_1$, $a_2$ e $a_3$ são constantes arbitrárias.

A linearidade da operação de integração permite afirmar então que a integração deverá ser exacta para as funções 1, $x$, $x^2$ e $x^3$. Substituindo na relação (8.6.1) $f(x)$ por cada uma destas funções, obtêm-se as seguintes relações

$$\begin{aligned}
c_1\cdot 1 + c_2\cdot 1 &= \int_{-1}^{1} 1dx = 2\\
c_1\cdot x_1 + c_2\cdot x_2 &= \int_{-1}^{1} xdx = 0\\
c_1\cdot x_1^2 + c_2\cdot x_2^2 &= \int_{-1}^{1} x^2dx = \frac{2}{3}\\
c_1\cdot x_1^3 + c_2\cdot x_2^3 &= \int_{-1}^{1} x^3dx = 0
\end{aligned}$$

Considerando que $x_2 > x_1$, a única solução deste sistema de equações é

$$c_1 = 1, \quad c_2 = 1, \quad x_1 = -\tfrac{\sqrt{3}}{3}, \quad x_2 = -\tfrac{\sqrt{3}}{3}.$$

Assim, conclui-se que a expressão

$$\int_{-1}^{1} f(x)dx \approx f\left(-\tfrac{\sqrt{3}}{3}\right) + f\left(\tfrac{\sqrt{3}}{3}\right)$$

é exacta sempre que $f(x)$ é substituída por um polinómio de grau inferior ou igual a 3.

O procedimento seguido anteriormente por ser aplicado para qualquer valor de $n$, sendo que obviamente teremos que resolver um sistema de $2n$ equações a $2n$ incógnitas. No entanto, é possível mostrar que tanto os nós $x_i$ como os valores dos coeficientes $c_i$ podem ser obtidos a partir dos designados **polinómios de Legendre**. Estes polinómios, aqui referenciados por $P_0(x)$, $P_1(x)$, ... verificam as propriedades

1. Para cada $n$, $P_n(x)$ é um polinómio de grau $n$.
2. $\int_{-1}^{1} f(x)P_n(x)dx = 0$ se $f(x)$ é um polinómio de grau $< n$.

Os polinómios de Legendre podem ser obtidos explicitamente pela expressão

$$P_n(x) = \frac{1}{2^n\, n!}\, \frac{d^n}{dx^n}(x^2-1)^n,$$

chegando-se assim facilmente à conclusão que os primeiros polinómios de Legendre serão

$$P_0(x) = 1, \qquad P_1(x) = x,$$
$$P_2(x) = \tfrac{1}{2}(3x^2-1), \qquad P_3(x) = \tfrac{1}{2}(5x^3-3x),$$
$$P_4(x) = \tfrac{1}{8}(35x^4-30x^2+3) \qquad \text{e} \qquad P_5(x) = \tfrac{1}{8}(63x^5-70x^3+15x).$$

Na figura 8.3 podem ver-se os gráficos dos polinómios $P_1(x)$ a $P_5(x)$.

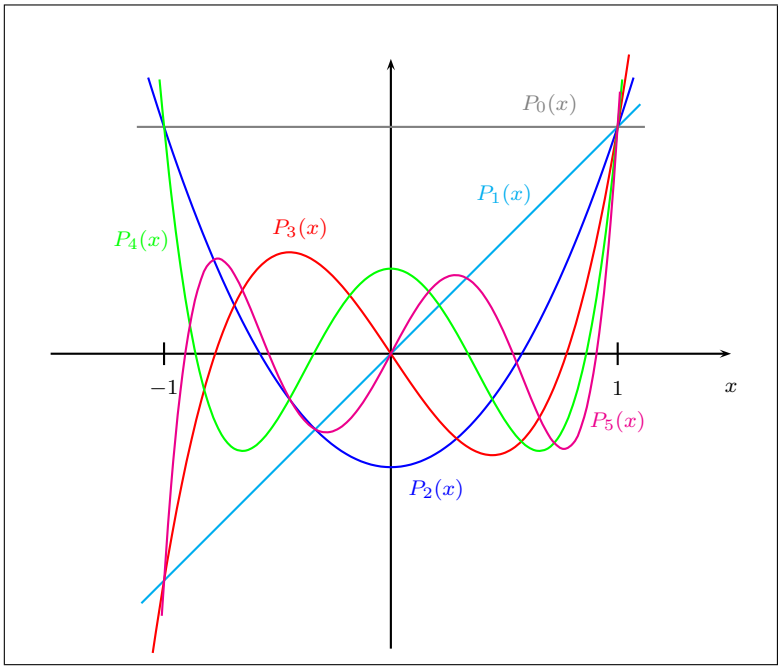



Figura 8.3: Polinómios de Legendre.

Uma propriedade interessante dos polinómios de Legendre é o facto do polinómio $P_n(x)$ ter exactamente $n$ raízes no interior do intervalo $[-1,1]$. Estas raízes serão exactamente os nós das regras de quadratura gaussiana, tal como se afirma no teorema seguinte.

**Teorema 8.6.1.**
*Sejam $x_1, x_2, \ldots, x_n$ as raízes do $n$-ésimo polinómio de Legendre $P_n(x)$ e, para cada $i = 1, 2, \ldots, n$, seja $c_i$ dado por*

$$c_i = \int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\; dx.$$

*Se $f(x)$ for um polinómio de grau $< 2n$, então*

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} c_i P(x_i).$$

Com base neste resultado é possível determinar os nós e os coeficientes das regras de integração gaussiana de diferentes ordens. Na tabela seguinte apresentam-se os valores destes nós e coeficientes para alguns casos.

| $n$ | $x_{n,i}$ | $c_{n,i}$ |
|---|---|---|
| 2 | $-0.5773502692$ | 1.0000000000 |
| | 0.5773502692 | 1.0000000000 |
| 3 | $-0.7745966692$ | 0.5555555556 |
| | 0.0000000000 | 0.8888888889 |
| | 0.7745966692 | 0.5555555556 |
| 4 | $-0.8611363116$ | 0.3478548451 |
| | $-0.3399810436$ | 0.6521451549 |
| | 0.3399810436 | 0.6521451549 |
| | 0.8611363116 | 0.3478548451 |
| 5 | $-0.9061798459$ | 0.2369268850 |
| | $-0.5384693101$ | 0.4786286705 |
| | 0.0000000000 | 0.5688888889 |
| | 0.5384693101 | 0.4786286705 |
| | 0.9061798459 | 0.2369268850 |

Em toda a dedução das regras de integração gaussiana considerou-se que a integração estava a ser efectuada no intervalo $[-1, 1]$. Ora no caso geral pretende-se calcular o integral $\int_a^b f(x)dx$ num intervalo $[a, b]$ genérico. Para tal há que realizar uma mudança de variável de forma a obter-se uma integração sobre o intervalo $[-1, 1]$. Utilizando a mudança de variável

$$t = \frac{2x - a - b}{b - a} \quad \Leftrightarrow \quad x = \frac{1}{2}[(b - a)t + a + b]$$

facilmente se verifica que $t = -1 \Leftrightarrow x = a$ e $t = 1 \Leftrightarrow x = b$ . Desta forma obtém-se a relação

$$\int_a^b f(x)dx = \int_{-1}^1 f\left(\frac{(b - a)t + a + b}{2}\right) \frac{b - a}{2}\, dt$$

sendo agora possível aplicar as regras de quadratura gaussiana atrás expostas.

**Exemplo 8.6.1.** *Utilizar quadratura gaussiana com* 2, 3, 4 *e* 5 *nós para aproximar* $\int_0^1 \frac{dx}{1+x^2}$.

***Resolução*** *O primeiro passo consiste em efectuar a mudança de variável*

$$t = 2x - 1 \quad \Leftrightarrow \quad x = \frac{t + 1}{2},$$

*no integral que se pretende calcular. Desta mudança resulta*

$$\int_0^1 \frac{dx}{1 + x^2} = \int_{-1}^1 \frac{2}{4 + (t + 1)^2}\, dt.$$

*Considerando a função* $g(t) = \frac{2}{4+(t+1)^2}$, *teremos então*

$$n = 2 \quad \rightarrow \quad \int_0^1 \frac{dx}{1+x^2} \approx g(-0.5773502692) + g(0.5773502692)$$
$$= \underline{0.78}68852458$$

$$n = 3 \quad \rightarrow \quad \int_0^1 \frac{dx}{1+x^2} \approx 0.5555555556 \cdot g(-0.7745966692) + 0.8888888889 \cdot g(0)$$
$$+ 0.5555555556 \cdot g(0.7745966692)$$
$$= \underline{0.7852}670352$$

$$n = 4 \quad \rightarrow \quad \int_0^1 \frac{dx}{1+x^2} \approx 0.3478548451 \cdot g(-0.8611363116) + 0.6521451549 \cdot g(-0.3399810436)$$
$$+ 0.6521451549 \cdot g(0.3399810436) + 0.3478548451 \cdot g(0.8611363116)$$
$$= \underline{0.7854}029762$$

$$n = 5 \quad \rightarrow \quad \int_0^1 \frac{dx}{1+x^2} \approx 0.2369268850 \cdot g(-0.9061798459) + 0.4786286705 \cdot g(-0.5384693101)$$
$$+ 0.5688888889 \cdot g(0) + 0.4786286705 \cdot g(0.5384693101)$$
$$+ 0.2369268850 \cdot g(0.9061798459)$$
$$= \underline{0.7853981}602$$

*Nota: Em cada um dos casos, apresentam-se sublinhados os algarismos correctos.*

Uma das vantagens dos métodos de quadratura gaussiana face aos outros métodos de integração numérica aqui estudados reside no facto de habitualmente fornecerem maior exactidão para o mesmo número de avaliações da função, como de alguma forma o exemplo acima ilustra.

# Capítulo 9

# Equações Diferenciais Ordinárias: problemas de valor inicial

## 9.1 Introdução

Muitos problemas de interesse em engenharia (e também noutros domínios) são modelizados recorrendo a equações diferenciais, quer ordinárias quer em derivadas parciais. De um modo geral, a resolução de uma equação diferencial consiste em determinar a função que satisfaz tal equação e simultaneamente obedece a um conjunto de condições adicionais habitualmente designadas por condições fronteira.

A maioria das equações diferenciais não admite soluções que se possam caracterizar por expressões analíticas. Nestes casos, a caracterização da solução da equação diferencial poderá ser feita de uma forma aproximada, por exemplo através de um desenvolvimento em série ou calculando de forma aproximada o valor da solução num conjunto finito de valores da variável independente.

É de notar, contudo, que existem importantes classes de equações diferenciais para as quais é possível determinar expressões analíticas das suas soluções. Uma destas classes é as equações diferencias ordinárias lineares de coeficientes constantes, que permitem modelizar sistemas lineares e invariantes no tempo.

Neste capítulo serão estudados métodos numéricos que permitem obter soluções (aproximadas) equações diferenciais ordinárias. No caso geral, procuraremos determinar a função $x$ que satisfaz a equação diferencial de ordem $n$

$$x^{(n)} = f(t, x, x', x'', \ldots, x^{(n-1)})$$

no intervalo $[t_0, T]$. Trataremos apenas os designados **problemas de valor inicial**, nos quais

a função $x$ deverá também satisfazer as condições iniciais

$$\begin{aligned} x(t_0) &= x_{0,0} \\ x'(t_0) &= x_{0,1} \\ &\vdots \\ x^{(n-1)}(t_0) &= x_{0,n-1} \end{aligned}$$

onde $x_{0,0}, x_{0,1}, \ldots, x_{0,n-1}$ são valores conhecidos.

Após a apresentação de algumas noções de base, serão estudados métodos para a resolução de equações diferenciais de ordem 1. Em seguida abordar-se-á o caso de sistemas de equações diferenciais de ordem 1, tratando-se por fim o caso geral das equações diferenciais de ordem $n$.

## 9.2 Solução numérica de equações diferenciais

Os métodos numéricos de resolução de equações diferenciais que serão estudados produzem valores de soluções aproximadas num conjunto finito de pontos da variável independente. Tal conjunto de pontos será aqui representado de uma forma geral por $\{t_i\}_{i=0}^N$. Dir-se-á ainda que este conjunto forma uma **malha** do intervalo $[t_0, T]$ se

$$t_0 < t_1 < \cdots < t_N = T.$$

Os pontos $t_i$ são designados por **nós da malha**. As distâncias

$$h_i = t_i - t_{i-1}, \qquad i = 1, \ldots, N,$$

designam-se por **passos** da malha. A malha diz-se **uniforme** se todas estas distâncias forem iguais. Também é usual designar por **passo** da malha o valor

$$h = \max_{1 \leq i \leq N} h_i.$$

Desta forma, a resolução numérica de uma equação diferencial consiste em definir uma malha $\{t_i\}_{i=0}^N$ no intervalo $[t_0, T]$ e em seguida calcular os valores $\{x_i\}_{i=0}^N$ da solução aproximada nos nós da malha.

Os métodos em que o cálculo de $x_i$ é feito apenas usando informação do intervalo $[t_{i-1}, t_i]$ designam-se por **métodos de passo simples**. Os que recorrem a informação fora deste intervalo para determinar $x_i$ designam-se por **métodos multi-passo**, ou de passo múltiplo. Aqui, apenas se estudarão métodos de passo simples.

É de notar que a solução aproximada obtida apenas estará definida nos nós $t_i$. Para obter valores em pontos intermédios, pode utilizar-se interpolação (por exemplo linear) entre cada dois nós consecutivos, como se mostra na figura.

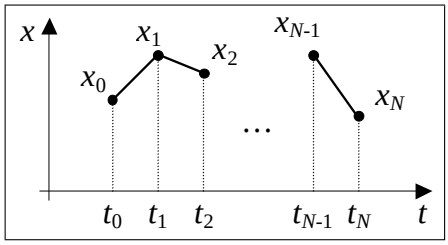



Figura 9.1: Solução aproximada.

Uma vez que as soluções de equações diferenciais são funções, e os métodos numéricos produzem soluções aproximadas, é importante ter uma forma de medir a **distância** entre duas funções. Esta distância permitirá assim medir o afastamento das soluções aproximadas produzidas pelos métodos numéricos estudados relativamente à solução exacta.

Dada uma função contínua $v$ definida no intervalo $[t_0, T]$, a **norma máximo** de $v$, representada por $\|v\|$, é definida por

$$\|v\| = \max_{t\in[t_0,T]} |v(t)|.$$

A distância entre $v$ e $w$, funções definidas e contínuas no intervalo $[t_0, T]$, é definida por

$$\|v - w\| = \max_{t\in[t_0,T]} |v(t) - w(t)|.$$

Claramente que estas funções serão iguais se e só se $\|v - w\| = 0$.

## 9.3 Equações diferenciais ordinárias de ordem 1

Antes de iniciarmos o estudos dos métodos numéricos de resolução de equações diferenciais de ordem 1, vamos relembrar resultados que garantem a existência e unicidade de solução para tais equações. Note-se que só fará sentido obter soluções aproximadas (utilizando métodos numéricos) de uma dada equação diferencial, quando a solução da equação existir e for única.

Seja então $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ uma função dada e consideremos a equação diferencial

$$x'(t) = f(t, x(t))$$

no intervalo $[t_0, T]$l. Como já referido, o problema de valor inicial associado a esta equação consiste em determinar a sua solução, sendo dado o valor que a função solução deverá no ponto $t_0$.

O seguinte teorema fornece condições suficiente para a existência e unicidade de solução para este problema de valor inicial.

**Teorema 9.3.1.** *Seja $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ uma função com as propriedades*

1. *$f$ é contínua em $[t_0, T]$ com respeito ao primeiro argumento;*

2. *$f$ é Lipschitz contínua com respeito ao segundo argumento, isto é, existe uma constante $L \geq 0$ (designada constante de Lipschitz) tal que*

$$|f(t, x_1) - f(t, x_2)| \leq L|x_1 - x_2|, \qquad \forall t \in [t_0, T], \forall x_1, x_2 \in \mathbb{R}.$$

*Então, o problema de valor inicial referido possui uma solução única. Mais ainda, a solução deste problema é uma função continuamente diferenciável em $[t_0, T]$.*

Uma das características desejáveis de um método numérico de solução de equações diferenciais é que produza soluções tão aproximadas da solução exacta quanto se pretenda, sendo tal aproximação normalmente controlada através do passo $h$ da malha de pontos considerada. Esta característica é designada por **convergência**, tal como se descreve em seguida.

Seja $x$ a solução exacta de uma dada equação diferencial e $x_h$ a solução aproximada produzida por aplicação de um dado método quando se utiliza uma malha de pontos de passo $h$ (por simplicidade consideramos aqui malhas uniformes). Seja ainda $e_h = x - x_h$ a função erro associada à solução aproximada $x_h$. Se se verificar que

$$\lim_{h \to 0} \|x_h - x\| \equiv \lim_{h \to 0} \|e_h\| = 0$$

para todas as soluções de equações diferenciais que verifiquem as condições de existência e unicidade acima enunciadas e para todas as condições iniciais tais que $\lim_{t \to 0} |e_h(t_0)| = 0$, então diz-se que tal método numérico de resolução de equações diferenciais é **convergente**. Diz-se ainda que um método convergente possui **ordem de convergência** igual a $p > 0$ se

$$\|x_h - x\| \leq ch^p$$

para todo o $h$ suficientemente pequeno, onde $c \in ]0, +\infty[$ é uma constante independente de $h$, mas dependente da função $f$ que caracteriza a equação diferencial.

Consideremos novamente a equação diferencial

$$x'(t) = f(t, x(t)), \qquad t \in [t_0, T].$$

Consideremos dois pontos consecutivos $t_i$ e $t_{i+1}$ de uma malha. Por facilidade de notação, no que se segue estes pontos serão simplesmente designados por $t$ e $t+h$ (onde $h$ representa obviamente o passo da malha, que se supõe uniforme). Integrando então a equação diferencial entre $t$ e $t+h$, obtém-se a relação

$$x(t + h) = x(t) + \int_t^{t+h} f(\xi, x(\xi)) d\xi.$$

Conclui-se assim que o valor da solução exacta $u$ no ponto $t + h$ poderia ser calculado somando ao valor da solução exacta em $t$ o valor do integral de $f(\xi, u(\xi))$ em $[t, t + h]$. Uma vez que nos

problemas de valor inicial, o valor $x(t_0) = x_0$ é conhecido, todos os valores pretendidos poderiam ser obtidos, bastando para tal considerar uma malha adequada.

A principal dificuldade reside no facto do integral acima envolver a própria função a determinar, o que impede o seu cálculo de um modo explícito. Os métodos numéricos de resolução de equações diferenciais caracterizam-se por prescindir do cálculo do valor exacto do integral, calculando-o de uma forma aproximada, como se indica em seguida.

Defina-se $F(t, x)$ como

$$F(t, x) = \frac{1}{h} \int_t^{t+h} f(\xi, x(\xi)) d\xi$$

e seja $F_h(t, x)$ um valor aproximado de $F(t, x)$, ou seja,

$$F_h(t, x) \approx \frac{1}{h} \int_t^{t+h} f(\xi, x(\xi)) d\xi,$$

que será diferente consoante o método de resolução aproximada da equação diferencial empregue, como se verá posteriormente. Representando por $T_h(t, x)$ o erro associado a esta aproximação, designado por **erro de truncatura**, tem-se

$$F(t, x) = F_h(t, x) + T_h(t, x).$$

A equação $x(t + h) = x(t) + \int_t^{t+h} f(\xi, x(\xi)) d\xi$ pode agora ser escrita como

$$\frac{x(t + h) - x(t)}{h} = F(t, x) = F_h(t, x) + T_h(t, x). \tag{9.3.1}$$

Fazendo $h \to 0$,que corresponde a aumentar o número de pontos da malha, e assumindo a existência dos limites tem-se que

$$x'(t) = \lim_{h \to 0} F_h(t, x) + \lim_{h \to 0} T_h(t, x).$$

Então, se o erro de truncatura tender para 0 com $h$, é legítimo supor que a eliminação deste termo em (9.3.1) conduza a equações *próximas* da equação original. Representando por $x_h$ a solução de (9.3.1) quando se despreza o erro de truncatura, e sendo $x_i = x_h(t_i)$ os valores nodais de $x_h$, verifica-se que estes satisfazem a seguinte relação de recorrência

$$x_{i+1} = x_i + hF_h(t_i, x_i), \qquad i = 0, 1, \ldots, N - 1.$$

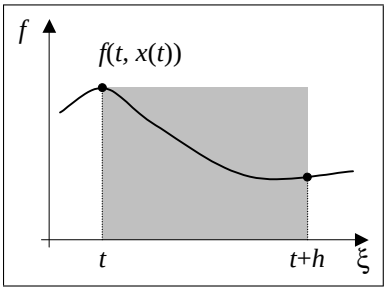
Diferentes escolhas da função $F_h$ conduzem a diferentes métodos para resolução numérica do problema de valor inicial, como veremos nas secções seguintes.

## 9.4 Métodos de Euler

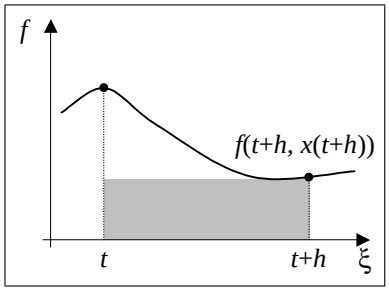
Uma forma simples de aproximar o integral

$$\int_t^{t+h} f(\xi, x(\xi)) d\xi$$

consiste em utilizar áreas dos rectângulos de base $h$ e alturas dadas pelos valores da função a integrar nos dois extremos do intervalo. Os métodos de Euler são baseados exactamente nestas duas aproximações do integral em causa.




Rectângulo à esquerda




Rectângulo à direita

Figura 9.2: Aproximações do rectângulos à esquerda e à direita.

O **método de Euler progressivo**, também designado simplesmente por **método de Euler**, consiste em utilizar a aproximação do rectângulo à esquerda. Apresenta-se em seguida a dedução da expressão de recorrência deste método, bem como uma expressão do erro de truncatura a ele associado. Do desenvolvimento de Taylor de $x(\cdot)$ temos

$$x(t+h) = x(t) + hx'(t) + \tfrac{h^2}{2}x''(\xi)$$

para algum $\xi \in [t, t+h]$.

Da equação diferencial original temos que

$$\begin{aligned} x'(t) &= f(t, x(t)) \\ x''(\xi) &= f'(\xi, x(\xi)) = \left(\tfrac{\partial f}{\partial t} + \tfrac{\partial f}{\partial x} f\right)\Big|_{\xi} \end{aligned}$$

Então

$$x(t+h) = x(t) + h\left[x'(t) + \tfrac{h}{2}x''(\xi)\right] = x(t) + h\left[f(t, x(t)) + \tfrac{h}{2}f'(\xi, x(\xi))\right],$$

e a aproximação fornecida por este método consiste em considerar

$$x(t+h) \simeq x(t) + hf(t, x(t))$$

correspondendo a ter

- $F_h(t, x) = f(t, x(t))$, e
- $T_h(t, x) = \frac{h}{2}f'(\xi, x(\xi))$.

Pode assim afirmar-se que a expressão de recorrência para a determinação dos valores nodais da solução aproximada $x_h$ será

$$x_{i+1} = x_i + hf(t_i, x_i), \qquad i = 0, 1, \ldots, N-1,$$

sendo $x_0 = x(t_0)$ a condição inicial.

**Exemplo 9.4.1.** *Utilizar o método de Euler com passo constante $h = 0.1$ para obter uma solução aproximada de*

$$x' = 1 + t - x, \quad t \in [0, 1] \qquad \textit{com} \qquad x(0) = 1.$$

***Resolução***

*Uma vez que $f(t_i, x_i) = 1 + t_i - x_i$, obtém-se a expressão de recorrência*

$$x_{i+1} = x_i + 0.1 \times (1 + t_i - x_i)$$

*para $i = 0, 1, \ldots, 9$. A condição inicial será obviamente $x_0 = x(0) = 1$. Partindo então desta condição e aplicando a expressão de recorrência obtida, determinam-se os valores apresentados na seguinte tabela.*

| $t_i$ | $x_i$ | $x_{i+1}$ |
|---|---|---|
| 0.0 | 1.0000 | 1.0000 |
| 0.1 | 1.0000 | 1.0100 |
| 0.2 | 1.0100 | 1.0290 |
| 0.3 | 1.0290 | 1.0561 |
| 0.4 | 1.0561 | 1.0905 |
| 0.5 | 1.0905 | 1.1314 |
| 0.6 | 1.1314 | 1.1783 |
| 0.7 | 1.1783 | 1.2305 |
| 0.8 | 1.2305 | 1.2874 |
| 0.9 | 1.2874 | 1.3487 |
| 1.0 | 1.3487 | — |

Passemos agora ao **método de Euler regressivo** que consiste em aproximar o integral

$$\int_t^{t+h} f(\xi, x(\xi)) d\xi$$

pelo valor do rectângulo à direita. Considerando o desenvolvimento e Taylor, agora a partir do ponto $t + h$, temos

$$x(t) = x(t + h) - hx'(t + h) + \tfrac{h^2}{2} x''(\xi)$$

para algum $\xi \in [t, t + h]$. De uma forma análoga ao efectuado atrás obtemos

$$\begin{aligned} x(t + h) &= x(t) + hx'(t + h) - \tfrac{h^2}{2} x''(\xi) \\ x(t + h) &= x(t) + h \left[ x'(t + h) - \tfrac{h}{2} x''(\xi) \right] \\ x(t + h) &= x(t) + h \left[ f(t + h, x(t + h)) - \tfrac{h}{2} f'(\xi, x(\xi)) \right] \end{aligned}$$

No método de Euler regressivo utiliza-se a aproximação

$$x(t + h) \simeq x(t) + hf(t + h, x(t + h))$$

o que corresponde a considerar

- $F_h(t,x) = f(t+h, x(t+h))$, e
- $T_h(t,x) = -\frac{h}{2} f'(\xi, x(\xi))$.

Do exposto conclui-se que a expressão de recorrência para determinação dos valores nodais da solução aproximada $x_h$ será

$$x_{i+1} = x_i + hf(t_{i+1}, x_{i+1}), \qquad i = 0, 1, \ldots, N-1,$$

sendo $x_0 = x(t_0)$ a condição inicial.

É de notar que neste método, o valor $u_{i+1}$ é definido de uma forma implícita. Podendo $f$ ser uma função não linear, não será possível em muitas situações obter uma expressão explícita para $x_{i+1}$. De um modo geral tem-se que

$$x_{i+1} = \phi_i(x_{i+1})$$

onde $\phi_i(x) = x_i + hf(t_{i+1}, x)$. Interessa aqui analisar algumas questões importantes. Por um lado, a existência e unicidade de solução desta equação e, por outro, o modo de resolver esta equação. A forma desta equação sugere a utilização do método iterativo simples, cuja condição de convergência é

$$\left| \frac{d\phi_i(x)}{dx} \right| = h \left| \frac{\partial f(t_{i+1}, x)}{\partial x} \right| < 1,$$

que se verifica desde que $h$ seja suficientemente pequeno (pois $f$ é Lipschitz contínua em $x$). Se esta condição se verificar é possível garantir a existência e unicidade de solução da equação que determina $x_{i+1}$.

**Exemplo 9.4.2.** *Utilizar o método de Euler regressivo com passo constante $h = 0.1$ para obter uma solução aproximada de*

$$x' = 1 + t - x, \quad t \in [0,1] \qquad \textit{com} \qquad x(0) = 1.$$

***Resolução***

*Uma vez que $f(t_{i+1}, x_{i+1}) = 1 + t_{i+1} - x_{i+1}$, obtém-se a expressão de recorrência*

$$x_{i+1} = x_i + 0.1 \times (1 + t_{i+1} - x_{i+1})$$

*para $i = 0, 1, \ldots, 9$. Neste caso, o valor de $x_{i+1}$ pode obter-se de uma forma explícita por*

$$x_{i+1} = \frac{x_i + 0.1 \times (1 + t_{i+1})}{1.1}.$$

*Utilizando a condição inicial, será obviamente $x_0 = x(0) = 1$, e aplicando a expressão de*

*recorrência acima obtêm-se os valores indicados na tabela abaixo.*

| $t_i$ | $x_i$ | $x_{i+1}$ |
|---|---|---|
| 0.0 | 1.0000 | 1.0091 |
| 0.1 | 1.0091 | 1.0264 |
| 0.2 | 1.0264 | 1.0513 |
| 0.3 | 1.0513 | 1.0830 |
| 0.4 | 1.0830 | 1.1209 |
| 0.5 | 1.1209 | 1.1645 |
| 0.6 | 1.1645 | 1.2132 |
| 0.7 | 1.2132 | 1.2665 |
| 0.8 | 1.2665 | 1.3241 |
| 0.9 | 1.3241 | 1.3855 |
| 1.0 | 1.3855 | — |

O erro de truncatura em qualquer dos métodos de Euler pode ser majorado por

$$\|T_h\| = \frac{h}{2} \sup_{t\in[t_0,T]} |f'(t, x(t))|$$

Sendo $f$ de classe $C^1$, as condições do teorema sobre existência e unicidade de solução permitem concluir que $f'(\cdot, x(\cdot))$ é contínua, pelo que o supremo acima é finito. Assim, o erro de truncatura dos métodos de Euler satisfaz

$$\|T_h\| \leq ch,$$

onde $c$ não depende de $h$, embora dependa dos dados que caracterizam o problema de valor inicial: a função $f$, o intervalo $[t_0, T]$, e o valor $x_0$.

## 9.5 Métodos de Taylor

Os métodos de Taylor de resolução numérica de equações diferencias caracterizam-se por aproximarem o integral $\int_t^{t+h} f(\xi, x(\xi))d\xi$ por polinómios de Taylor. As expressões de recorrência destes métodos, bem como os seus erros de truncatura obtêm-se facilmente como se mostra em seguida. Consideremos o desenvolvimento de Taylor

$$x(t+h) = x(t) + hx'(t) + \tfrac{h^2}{2}x''(t) + \cdots + \tfrac{h^p}{p!}x^{(p)}(t) + \tfrac{h^{p+1}}{(p+1)!}x^{(p+1)}(\xi)$$

onde $\xi \in [t, t+h]$. Da equação diferencial temos

$$x(t+h) = x(t) + hf(t,(x(t))) + \tfrac{h^2}{2}f'(t, x(t)) + \cdots + \tfrac{h^p}{p!}f^{(p-1)}(t, x(t)) + \tfrac{h^{p+1}}{(p+1)!}f^{(p)}(\xi, x(\xi)).$$

O **método de Taylor de ordem** $p$ é caracterizado pela aproximação

$$x(t+h) \simeq x(t) + hf(t,(x(t))) + \tfrac{h^2}{2}f'(t, x(t)) + \cdots + \tfrac{h^p}{p!}f^{(p-1)}(t, x(t))$$

o que corresponde a ter

$$F_h(t,x) = f(t,x(t)) + \frac{h}{2}f'(t,x(t)) + \cdots + \frac{h^{p-1}}{p!}f^{(p-1)}(t,x(t)).$$

Então, o erro de truncatura deste método será

$$T_h(t,x) = \frac{h^p}{(p+1)!}f^{(p)}(\zeta, x(\zeta)), \qquad \zeta \in [t, t+h].$$

A expressão de recorrência do método de Taylor de ordem $p$ será assim

$$x_{i+1} = x_i + hf(t_i,x_i) + \frac{h^2}{2}f'(t_i,x_i) + \cdots + \frac{h^p}{p!}f^{(p-1)}(t_i,x_i)$$

para $i = 0, 1, \ldots, N-1$. Obviamente que o método de Taylor de ordem 1 não é senão o método de Euler progressivo.

Sendo válidas as hipóteses sobre existência e unicidade de solução do problema de valor inicial e verificando-se também que $f$ é de classe $C^p$, verifica-se que a função $t \to f(t,x(t))$ é também de classe $C^p$ no intervalo $[t_0, T]$. Pode, assim, afirmar-se que

$$\|T_h\| \leq \|f^{(p)}(\cdot, x(\cdot))\| \frac{h^p}{(p+1)!}.$$

A aplicação da expressão de recorrência dos métodos de Taylor (assim como a avaliação do erro de truncatura) necessita que sejam obtidas expressões para as derivadas da função $t \to f(t,x(t))$ num ponto $(t,x(t))$ genérico. Estas derivadas podem ser obtidas da seguinte forma

$$\begin{aligned} f' &= f_t + f_x x' = f_t + f_x f \\ f'' &= f_{tt} + 2f_{tx}f + f_{xx}f^2 + f_x f_t + f_x^2 f \\ &\cdots \end{aligned}$$

onde $f_t = \frac{\partial f}{\partial t}$ e $f_x = \frac{\partial f}{\partial x}$. Excepto se $f$ tiver uma forma muito simples, o cálculo destas derivadas rapidamente se torna bastante complexo, pelo que os métodos de Taylor de ordem elevada não são em geral de utilização muito prática.

**Exemplo 9.5.1.** *Usando o método de Taylor de ordem* 2*, com passo* 0.1*, obter uma solução aproximada de*

$$x' = 1 + t - x, \quad t \in [0,1] \qquad \textit{com} \qquad x(0) = 1.$$

***Resolução***

*A expressão de recorrência é* $x_{i+1} = x_i + hf(t_i,x_i) + \frac{h^2}{2}f'(t_i,x_i)$*, pelo que é necessário calcular* $f'$*, obtendo-se*

$$f'(t,x) = 1 - 1 \times (1 + t - x) = x - t.$$

*Assim, a expressão de recorrência é para este problema*

$$x_{i+1} = x_i + 0.1 \times (1 + t_i - x_i) + 0.005 \times (x_i - t_i)$$

*onde* $i = 0, 1, \ldots, 9$.
*Partindo da condição inicial* $x_0 = x(0) = 1$, *obtêm-se os seguintes resultados*

| $t_i$ | $x_i$ | $x_{i+1}$ |
|---|---|---|
| 0.0 | 1.0000 | 1.0050 |
| 0.1 | 1.0050 | 1.0190 |
| 0.2 | 1.0190 | 1.0412 |
| 0.3 | 1.0412 | 1.0708 |
| 0.4 | 1.0708 | 1.1071 |
| 0.5 | 1.1071 | 1.1494 |
| 0.6 | 1.1494 | 1.1972 |
| 0.7 | 1.1972 | 1.2500 |
| 0.8 | 1.2500 | 1.3072 |
| 0.9 | 1.3072 | 1.3685 |
| 1.0 | 1.3685 | — |

## 9.6 Consistência e convergência

Um método de resolução numérica de equações diferenciais diz-se **consistente** se

$$\lim_{h \to 0} \|T_h\| = 0$$

e diz-se que a sua **ordem de consistência** é $p > 0$ se

$$\|T_h\| \leq ch^p,$$

para todo o $h$ suficiente pequeno e $c > 0$, independente de $h$.

Desta definição e do exposto atrás resulta imediatamente que ambos os métodos de Euler têm ordem de consistência igual a 1e também que o método de Taylor de ordem $p$ tem ordem de consistência $p$ (daí o seu nome!).

Note-se, contudo, que o erro de truncatura (e logo a ordem de consistência) apenas caracterizam o erro local em cada passo e não o erro global de aproximação da solução exacta $x$ por $x_h$. Em muitas situações interessa analisar não o erro de truncatura (que apenas fornece informação local), mas o erro de aproximação global, definido por

$$e_h = x - x_h.$$

Em particular, interessa saber se este erro converge para zero à medida que $h$ vai para zero e, em caso afirmativo, qual a ordem de convergência do método. Apresenta-se em seguida um resultado que relaciona a ordem de consistência e a ordem de convergência de métodos que satisfazem certas condições.

**Definição 9.6.1.** *Um método de passo simples diz-se satisfazer a condição de Lipschitz se* $F_h$ *verificar*

$$|F_h(t, v) - F_h(t, w)| \leq L_h |v - w|, \qquad t \in [t_0, T],$$

*para todo o $h > 0$ suficientemente pequeno, onde $L_h$ é independente de $h$.*

O resultado seguinte apresenta condições de equivalência entre os conceitos de consistência e convergência e estabelece uma estimativa para o erro de aproximação.

**Teorema 9.6.1.** *Se um método de passo simples satisfizer a condição de Lipschitz então será consistente se e só se for convergente.*

*Mais ainda, para $h$ suficientemente pequeno, verifica-se que*

$$|e_h(t)| \leq e^{L_h(t-t_0)}|e_0| + \frac{\|T_h\|}{L_h}[e^{L_h(t-t_0)} - 1], \qquad t \in [t_0, T],$$

*onde $e_0 = x(t_0) - x_h(t_0)$.*

Se $f \in C^p$ e as hipóteses deste teorema se verificarem, então os métodos de Taylor de ordem (de consistência) $p$ têm ordem de convergência $p$, razão pela qual os métodos de Taylor de ordem mais elevada têm associados erros que convergem mais rapidamente para zero, isto com a diminuição do passo $h$.

É importante referir aqui, ainda que de uma forma informal, que a utilização de passos $h$ muito reduzidos, embora aparentemente benéfica por levar a erros mais baixo, é muitas vezes fonte de erros que se podem tornar extremamente elevados. De facto, quanto menor for o valor de $h$ maior número de cálculos será necessário para determinar a solução da equação diferencial no intervalo dado. Dado que os cálculos são sempre (ou quase sempre) realizados em aritmética finita, verifica-se que quanto mais cálculos se tiverem de efectuar maiores serão os erros devidos à utilização da aritmética finita. Para um problema concreto que se pretende resolver com um dado método numérico numa máquina com uma dada precisão finita verifica-se que existe normalmente um valor "óptimo" de $h$ que conduz ao menor erro global na solução aproximada. Para valores de $h$ superiores o erro aumenta por aumentar o erro devido a se utilizar um método aproximado de solução, enquanto para valores menores de $h$ o erro aumenta por aumentarem os erros devidos à aritmética finita.

Esta discussão indicia que de uma forma geral os métodos de maior ordem permitirão melhores resultados pois os erros de truncatura e logo os de aproximação diminuem mais rapidamente com a diminuição do passo $h$.

## 9.7 Métodos de Runge-Kutta

Como já foi visto atrás, o aumento da ordem de consistência dos métodos de Taylor é efectuado à custa do esforço de cálculo de derivadas de ordem superior da função $f$.

Os métodos conhecidos genericamente por **métodos de Runge-Kutta** foram desenvolvidos de forma a possuírem ordens de consistência superiores a 1 e a não necessitarem do cálculo de

derivadas de $f$ para a sua aplicação. É também importante referir que os métodos de Runge-Kutta gozam ainda da propriedade de possuírem ordem de convergência igual à sua ordem de consistência.

De uma forma geral, a obtenção do valor aproximado $x_{i+1}$ no instante $t_{i+1}$ é feita avaliando a função $f$ em pontos "intermédios" entre $(t_i, x_i)$ e $(t_{i+1}, x_{i+1})$. A selecção de tais pontos "intermédios" e da expressão de cálculo de $x_{i+1}$ são efectuadas de modo a garantir a ordem de consistência pretendida.

De uma forma geral os métodos de Runge-Kutta (explícitos) permitem obter o valor de $x_{i+1}$ efectuando os seguinte cálculos

$$\begin{aligned}
F_1 &= f(t_i, x_i) \\
F_2 &= f(t_i + \alpha_2 h, x_i + h\beta_{21} F_1) \\
F_3 &= f(t_i + \alpha_3 h, x_i + h(\beta_{31} F_1 + \beta_{32} F_2)) \\
&\dots \\
F_s &= f(t_i + \alpha_s h, x_i + h(\beta_{s,1} F_1 + \beta_{s,2} F_2 + \ldots + \beta_{s,s-1} F_{s-1})) \\
x_{i+1} &= x_i + h(w_1 F_1 + w_2 F_2 + \ldots + w_s F_s)
\end{aligned}$$

Nestas expressões, $s$ é um inteiro que traduz o número de estágios e $\alpha_j$, $\beta_{jk}$ e $w_j$ são parâmetros a determinar de modo a garantir a ordem de consistência desejada. Para ordens de consistência até 4 verifica-se sem possível obter métodos com número de estágios igual à ordem de consistência. Apresentam-se em seguida os métodos de ordem 2, indicando-se o processo de determinação dos coeficientes, e também o método de ordem 4 mais utilizado.

### Métodos de Runge-Kutta de 2ª ordem

Estes métodos utilizam apenas um ponto intermédio entre $t_i$ e $t_{i+1}$. O valor de $x_{i+1}$ é calculado com as seguintes expressões

$$\begin{aligned}
F_1 &= f(t_i, x_i) \\
F_2 &= f(t_i + \alpha_2 h, x_i + h\beta_{21} F_1) \\
x_{i+1} &= x_i + h(w_1 F_1 + w_2 F_2)
\end{aligned}$$

onde $w_1$, $w_2$, $\alpha_2$ e $\beta_{21}$, são determinados de modo a se obter a maior ordem de consistência possível, como indicado em seguida.

De acordo com as expressões acima tem-se neste caso que

$$F_h(t, x) = w_1 f(t, x(t)) + w_2 f(t + \alpha_2 h, x(t) + h\beta_{21} F_1)$$

sendo então o erro de truncatura dado por

$$T_h(t, x) = \frac{1}{h} \int_t^{t+h} f(\xi), x(\xi)) d\xi - w_1 f(t, x(t)) - w_2 f(t + \alpha_2 h, x(t) + h\beta_{21} F_1)$$

Efectuando o desenvolvimento em série de Taylor de $T_h(t,x)$, obtém-se a seguinte expressão

$$\begin{aligned} T_h(t,x) = {} & (1-w_1-w_2)f \\ & + h\left[\left(\tfrac{1}{2}-\alpha_2 w_2\right)f_t\right) + \left(\tfrac{1}{2}-\beta_{21}w_2\right)f_x f\right)\Big] \\ & + h^2\left[\tfrac{1}{6}(f_{tt}+2f_{tx}f+f_{xx}f^2+f_t f_x+f_x^2 f) - w_2\left(\tfrac{\alpha_2^2}{2}f_{tt}+\alpha_2\beta_{21}f_{tx}f+\tfrac{\beta_{21}^2}{2}f_{xx}f^2\right)\right] \\ & + O(h^3) \end{aligned}$$

Analisando esta expressão conclui-se que de uma forma geral não será possível anular o termo em $h^2$. Todavia, é possível anular os termos de ordem inferior a $h^2$, garantindo-se assim uma ordem de consistência 2. Para tal, basta que se verifiquem as seguintes igualdades

$$\begin{aligned} w_1 + w_2 &= 1 \\ \alpha_2 w_2 &= \tfrac{1}{2} \\ \beta_{21} w_2 &= \tfrac{1}{2} \end{aligned}$$

Como facilmente se constata, este sistema de equações possui diferentes conjuntos de soluções, cada um deles correspondendo a um método numérico para a resolução da equação diferencial. As escolhas mais comuns resultam nos dois métodos abaixo indicados. É importante referir que a designação destes métodos não é consensual, variando de autor para autor.

O **método de Euler modificado** é obtido fazendo $w_1 = w_2 = \frac{1}{2}$, $\alpha_2 = 1$ e $\beta_{21} = 1$. O cálculo de $x_{i+1}$ por este método será então feito de acordo com

$$\begin{aligned} F_1 &= f(t_i, x_i) \\ F_2 &= f(t_i + h, x_i + hF_1) \\ x_{i+1} &= x_i + \tfrac{h}{2}(F_1 + F_2). \end{aligned}$$

O **método de Heun** é obtido fazendo $w_1 = \frac{1}{4}$, $w_2 = \frac{3}{4}$, $\alpha_2 = \beta_{21} = \frac{2}{3}$. O cálculo de $x_{i+1}$ por este método será então feito de acordo com

$$\begin{aligned} F_1 &= f(t_i, x_i) \\ F_2 &= f(t_i + \tfrac{2}{3}h, x_i + \tfrac{2}{3}hF_1) \\ x_{i+1} &= x_i + \tfrac{h}{4}(F_1 + 3F_2). \end{aligned}$$

### Métodos de Runge-Kutta de 4ª ordem

O método de Runge-Kutta de 4ª ordem abaixo indicado é um dos mais utilizados. A sua popularidade advém do seu bom compromisso entre esforço computacional requerido e precisão alcançada. Os valores dos coeficientes utilizados obtêm-se de forma a garantir que o erro de truncatura é de ordem $h^4$, ou seja, que se trata efectivamente de um método de ordem de consistência 4.

As expressões que permitem determinar o valor $x_{i+1}$ por este método são as seguintes

$$\begin{aligned} F_1 &= f(t_i, x_i) \\ F_2 &= f(t_i + \tfrac{h}{2}, x_i + \tfrac{h}{2}F_1) \\ F_3 &= f(t_i + \tfrac{h}{2}, x_i + \tfrac{h}{2}F_2) \\ F_4 &= f(t_i + h, x_i + hF_3) \\ x_{i+1} &= x_i + \tfrac{h}{6}(F_1 + 2F_2 + 2F_3 + F_4) \end{aligned}$$

que se devem aplicar para $i = 0, 1, \ldots, N-1$.

**Exemplo 9.7.1.** *Aplicar o método de Runge-Kutta de 4ª ordem com passo* 0.1 *para obter uma solução aproximada de*

$$x' = 1 + t - x, \quad t \in [0,1] \qquad \textit{com} \qquad x(0) = 1.$$

***Resultados***

| $t_i$ | $x_i$ | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $x_{i+1}$ |
|---|---|---|---|---|---|---|
| 0.0 | 1.00000 | 0.00000 | 0.05000 | 0.04750 | 0.09525 | 1.00484 |
| 0.1 | 1.00484 | 0.09516 | 0.14040 | 0.13814 | 0.18135 | 1.01873 |
| 0.2 | 1.01873 | 0.18127 | 0.22221 | 0.22016 | 0.25925 | 1.04082 |
| 0.3 | 1.04082 | 0.25918 | 0.29622 | 0.29437 | 0.32974 | 1.07032 |
| 0.4 | 1.07032 | 0.32968 | 0.36320 | 0.36152 | 0.39353 | 1.10653 |
| 0.5 | 1.10653 | 0.39347 | 0.42380 | 0.42228 | 0.45124 | 1.14881 |
| 0.6 | 1.14881 | 0.45119 | 0.47863 | 0.47726 | 0.50346 | 1.19659 |
| 0.7 | 1.19659 | 0.50341 | 0.52824 | 0.52700 | 0.55071 | 1.24933 |
| 0.8 | 1.24933 | 0.55067 | 0.57314 | 0.57201 | 0.59347 | 1.30657 |
| 0.9 | 1.30657 | 0.59343 | 0.61376 | 0.61274 | 0.63216 | 1.36788 |
| 1.0 | 1.36788 | — | — | — | — | — |

O exemplo seguinte sintetiza os resultados dos exemplos anteriores, podendo constatar-se as diferenças entre eles e verificar o desempenho superior (como esperado) do método de Runge-Kutta de 4ª ordem.

**Exemplo 9.7.2.** *Na tabela seguinte apresentam-se os resultados obtidos nos exemplos anteriores com os diferentes métodos, bem como a solução exacta do problema que é* $x(t) = t + e^{-t}$.

| $t_i$ | *Euler prog.* | *Euler reg.* | *Taylor 2* | *R-K 4* | *Sol. exacta* |
|---|---|---|---|---|---|
| 0.0 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| 0.1 | 1.000000 | 1.009091 | 1.005000 | 1.004838 | 1.004837 |
| 0.2 | 1.010000 | 1.026446 | 1.019025 | 1.018731 | 1.018731 |
| 0.3 | 1.029000 | 1.051315 | 1.041218 | 1.040818 | 1.040818 |
| 0.4 | 1.056100 | 1.083013 | 1.070802 | 1.070320 | 1.070320 |
| 0.5 | 1.090490 | 1.120921 | 1.107076 | 1.106531 | 1.106531 |
| 0.6 | 1.131441 | 1.164474 | 1.149404 | 1.148812 | 1.148812 |
| 0.7 | 1.178297 | 1.213158 | 1.197210 | 1.196586 | 1.196585 |
| 0.8 | 1.230467 | 1.266507 | 1.249975 | 1.249329 | 1.249329 |
| 0.9 | 1.287420 | 1.324098 | 1.307228 | 1.306570 | 1.306570 |
| 1.0 | 1.348678 | 1.385543 | 1.368541 | 1.367880 | 1.367879 |

## 9.8 Sistemas de equações diferenciais

Dadas as funções $f_1, f_2, \ldots, f_n$, de $\mathbb{R}^{1+n}$ em $\mathbb{R}$, um sistema de equações diferenciais de ordem 1 é definido por

$$\begin{cases} x_1'(t) &= f_1(t, x_1(t), x_2(t), \ldots, x_n(t)) \\ x_2'(t) &= f_2(t, x_1(t), x_2(t), \ldots, x_n(t)) \\ & \vdots \\ x_n'(t) &= f_n(t, x_1(t), x_2(t), \ldots, x_n(t)) \end{cases}$$

O **problema de valor inicial** consiste agora em determinar funções $x_1, x_2, \ldots, x_n$, de um intervalo $[t_0, T]$ em $\mathbb{R}$, que satisfazem estas equações diferenciais e as condições

$$x_1(t_0) = x_{1,0}, \quad x_2(t_0) = x_{2,0}, \quad \ldots, \quad x_n(t_0) = x_{n,0},$$

para $x_{1,0}, x_{2,0}, \ldots, x_{n,0} \in \mathbb{R}$ dados.

Numa notação mais compacta, o sistema de equações diferenciais representa-se por

$$\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t))$$

onde $\mathbf{f} : \mathbb{R}^{1+n} \to \mathbb{R}^n$ é definida por $\mathbf{f} = [f_1 \quad f_2 \quad \ldots \quad f_n]^{\mathrm{T}}$ e $\mathbf{x}$ é a função de $\mathbb{R}$ em $\mathbb{R}^n$, definida por $\mathbf{x} = [x_1 \quad x_2 \quad \ldots \quad x_n]^{\mathrm{T}}$.

O problema de valor inicial consiste em determinar a função $\mathbf{x}$ que satisfaz esta equação diferencial vectorial num intervalo $[t_0, T]$ e a condição inicial

$$\mathbf{x}(t_0) = \mathbf{x}_0,$$

para algum $\mathbf{x}_0 \in \mathbb{R}^n$. Importa referir aqui que é possível estabelecer condições de existência e unicidade de solução para este problema análogas às formuladas no caso de uma equação diferencial escalar.

Os métodos numéricos de resolução aproximada de problemas de valor inicial estudados para o caso escalar (uma equação) podem ser aplicados de uma forma imediata ao caso vectorial (sistema de equações). Considerando uma malha $\{t_i\}_{i=0}^N$ de passo $h$ no intervalo $[t_0, T]$, sendo $\mathbf{x}_h$ uma solução aproximada do problema de valor inicial, de um modo geral, os valores $\mathbf{x}_i = \mathbf{x}_h(t_i)$ podem ser obtidos pela expressão de recorrência

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{F}_h(t_i, \mathbf{x}_i),$$

para $i = 0, 1, \ldots, N-1$, sendo também habitual considerar $\mathbf{x}_h(t_0) = \mathbf{x}_0$. É de notar a semelhança entre esta expressão de recorrência e a expressão geral utilizada no caso escalar.

A função $\mathbf{F}_h$ define-se em termos de $\mathbf{f}$, de forma análoga ao caso escalar. A principal diferença face ao caso escalar reside no facto dos valores $\mathbf{x}_i$ a determinar por via numérica serem elementos de $\mathbb{R}^n$, sendo em cada passo necessário calcular $n$ números reais.

Exceptuando o método de Euler regressivo que é de extensão mais complexa para o caso vectorial, todos os outros métodos são de imediata adaptação:

- o método de Euler progressivo conduz à expressão de recorrência

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i).$$

- o método de Taylor de ordem 2 tem por expressão de recorrência

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i) + \frac{h^2}{2}\mathbf{f}'(t_i, \mathbf{x}_i).$$

- ...

É de notar agora que o cálculo de $\mathbf{f}', \mathbf{f}'', \ldots$ pode ser bastante complexo, pois cada componente de $\mathbf{f}$ depende de $t$ quer directamente quer indirectamente através das componentes de $\mathbf{x}$.

**Exemplo 9.8.1.** *Considere o seguinte problema de valor inicial*

$$\begin{cases} u_1' = u_1 u_2 \\ u_2' = t + u_1 - u_2 \end{cases} \qquad t \in [0,1],$$
$$u_1(0) = 1, \qquad u_2(0) = 0.$$

***a)*** *Determinar uma solução aproximada pelo método de Euler progressivo com passo* 0.1.
***b)*** *Determinar uma solução aproximada pelo método de Taylor de ordem* 2 *com passo* 0.1.

***Resolução***
***a)*** *Definam-se* $f_1$ *e* $f_2$ *por*

$$f_1(t, u_1, u_2) = u_1 u_2$$
$$f_2(t, u_1, u_2) = t + u_1 - u_2$$

*A expressão do método de Euler progressivo*

$$\mathbf{u}_{i+1} = \mathbf{u}_i + h\mathbf{F}_h(t_i, \mathbf{u}_i)$$

*toma neste caso a forma*

$$u_{1,i+1} = u_{1,i} + hf_1(t_i, u_{1,i}, u_{2,i})$$
$$u_{2,i+1} = u_{2,i} + hf_2(t_i, u_{1,i}, u_{2,i})$$

*ou ainda*

$$u_{1,i+1} = u_{1,i} + 0.1 \times u_{1,i}u_{2,i}$$
$$u_{2,i+1} = u_{2,i} + 0.1 \times (t_i + u_{1,i} - u_{2,i})$$

*para* $i = 0, 1, \ldots, 9$*, com as condições iniciais* $u_{1,0} = u_1(0) = 1$ *e* $u_{2,0} = u_2(0) = 0$*.*

*A tabela abaixo apresenta os resultados obtidos.*

| $t_i$ | $u_{1,i}$ | $u_{2,i}$ | $u_{1,i+1}$ | $u_{2,i+1}$ |
|---|---|---|---|---|
| 0.0 | 1.0000 | 0.0000 | 1.0000 | 0.1000 |
| 0.1 | 1.0000 | 0.1000 | 1.0100 | 0.2000 |
| 0.2 | 1.0100 | 0.2000 | 1.0302 | 0.3010 |
| 0.3 | 1.0302 | 0.3010 | 1.0612 | 0.4039 |
| 0.4 | 1.0612 | 0.4039 | 1.1041 | 0.5096 |
| 0.5 | 1.1041 | 0.5096 | 1.1603 | 0.6191 |
| 0.6 | 1.1603 | 0.6191 | 1.2322 | 0.7332 |
| 0.7 | 1.2322 | 0.7332 | 1.3225 | 0.8531 |
| 0.8 | 1.3225 | 0.8531 | 1.4353 | 0.9801 |
| 0.9 | 1.4353 | 0.9801 | 1.5760 | 1.1156 |
| 1.0 | 1.5760 | 1.1156 | — | — |

***b)*** *A expressão do método de Taylor de ordem* 2 *é*

$$\mathbf{u}_{i+1} = \mathbf{u}_i + h\mathbf{f}(t_i, \mathbf{u}_i) + \frac{h^2}{2}\mathbf{f}'(t_i, \mathbf{u}_i)$$

*sendo então necessário determinar* $f_1'$ *e* $f_2'$*. Estas funções obtêm-se de acordo com*

$$f_1'(t, u_1, u_2) = u_2 u_1' + u_1 u_2' = u_1 u_2^2 + u_1 \cdot (t + u_1 - u_2)$$
$$f_2'(t, u_1, u_2) = 1 + u_1' - u_2' = 1 + u_1 u_2 - (t + u_1 - u_2)$$

*As expressões de recorrência tomam então a forma*

$$u_{1,i+1} = u_{1,i} + 0.1 \times u_{1,i}u_{2,i} + 0.005 \times \left(u_{1,i}u_{2,i}^2 + u_{1,i} \cdot (t_i + u_{1,i} - u_{2,i})\right)$$
$$u_{2,i+1} = u_{2,i} + 0.1 \times (t_i + u_{1,i} - u_{2,i}) + 0.005 \times (1 + u_{1,i}u_{2,i} - (t_i + u_{1,i} - u_{2,i}))$$

*devendo ser determinada para* $i = 0, 1, \ldots, 9$ *com as condições iniciais* $u_{1,0} = 1$ *e* $u_{2,0} = 0$*.*

*A tabela abaixo apresenta os valores obtidos.*

| $t_i$ | $u_{1,i}$ | $u_{2,i}$ | $u_{1,i+1}$ | $u_{2,i+1}$ |
|---|---|---|---|---|
| 0.0 | 1.0000 | 0.0000 | 1.0050 | 0.1000 |
| 0.1 | 1.0050 | 0.1000 | 1.0202 | 0.2010 |
| 0.2 | 1.0202 | 0.2010 | 1.0461 | 0.3038 |
| 0.3 | 1.0461 | 0.3038 | 1.0838 | 0.4094 |
| 0.4 | 1.0838 | 0.4094 | 1.1349 | 0.5187 |
| 0.5 | 1.1349 | 0.5187 | 1.2016 | 0.6327 |
| 0.6 | 1.2016 | 0.6327 | 1.2871 | 0.7525 |
| 0.7 | 1.2871 | 0.7525 | 1.3955 | 0.8797 |
| 0.8 | 1.3955 | 0.8797 | 1.5328 | 1.0158 |
| 0.9 | 1.5328 | 1.0158 | 1.7073 | 1.1632 |
| 1.0 | 1.7073 | 1.1632 | — | — |

## 9.9 Equações diferenciais de ordem $n$

Consideremos agora o problema de determinar a função $x : \mathbb{R} \to \mathbb{R}$ que é solução de uma dada equação diferencial de ordem $n$

$$x^{(n)}(t) = f(t, x(t), x'(t), \ldots, x^{(n-1)}(t))$$

num dado intervalo $[t_0, T]$ e satisfaz as condições iniciais

$$\begin{aligned} x(t_0) &= x_{0,0} \\ x'(t_0) &= x_{0,1} \\ &\cdots \\ x^{(n-1)}(t_0) &= x_{0,n-1} \end{aligned}$$

para $x_{0,0}, x_{0,1}, \ldots, x_{0,n-1} \in \mathbb{R}$ dados.

A resolução numérica deste problema é obtida transformando a equação diferencial de ordem $n$ num sistema de $n$ equações diferenciais de ordem 1, como se indica em seguida. Considerando as variáveis dependentes (isto é, as funções) $x_1, x_2, \ldots, x_n$ definidas por

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= x'(t) \\ &\cdots \\ x_n(t) &= x^{(n-1)}(t) \end{aligned}$$

conclui-se facilmente que $x_i'(t) = x_{i+1}(t)$ para $i = 1, 2, \ldots, n-1$.

Utilizando estas novas funções tem-se ainda que

$$x_n'(t) = \left[x^{(n-1)}\right]'(t) = x^{(n)}(t) = f(t, x_1(t), x_2(t), \ldots, x_n(t))$$

O sistema de equações diferenciais de ordem 1 toma então a forma

$$\begin{aligned} x_1'(t) &= x_2(t) \\ x_2'(t) &= x_3(t) \\ &\cdots \\ x_n'(t) &= f(t, x_1(t), x_2(t), \ldots, x_n(t)) \end{aligned}$$

devendo a sua solução satisfazer as condições iniciais

$$x_1(t_0) = x_{0,0}, \quad x_2(t_0) = x_{0,1}, \quad \ldots, x_n(t_0) = x_{0,n-1}.$$

Os métodos para resolver uma equação diferencial de ordem $n$ serão assim os mesmos que se utilizam para resolver um sistema de equações diferenciais de ordem 1.

**Exemplo 9.9.1.** *Determinar, pelo método de Euler progressivo com passo* 0.05*, uma solução aproximada de*

$$\theta'' + 10\sin\theta = 0,\ t \in [0, 0.5], \qquad \theta(0) = 0.1, \quad \theta'(0) = 0.$$

***Resolução***

*Definindo* $x_1 = \theta$ *e* $x_2 = \theta'$*, obtém-se o sistema de equações diferenciais*

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= -10\sin(x_1) \end{aligned}$$

*As expressões de recorrência serão*

$$\begin{aligned} x_{1,i+1} &= x_{1,i} + 0.05 \times x_{2,i} \\ x_{2,i+1} &= x_{2,i} - 0.05 \times 10\sin(x_{1,i}) \end{aligned}$$

*com* $x_{1,0} = 0.1$ *e* $x_{2,0} = 0$*.*

*Aplicando sucessivamente estas expressões, obtêm-se os valores apresentados na tabela seguinte.*

| $t_i$ | $x_{1,i} = \theta_i$ | $x_{2,i}$ | $x_{1,i+1}$ | $x_{2,i+1}$ |
|---|---|---|---|---|
| 0.00 | 0.1000 | 0.0000 | 0.1000 | −0.0499 |
| 0.05 | 0.1000 | −0.0499 | 0.0975 | −0.0998 |
| 0.10 | 0.0975 | −0.0998 | 0.0925 | −0.1485 |
| 0.15 | 0.0925 | −0.1485 | 0.0851 | −0.1947 |
| 0.20 | 0.0851 | −0.1947 | 0.0754 | −0.2372 |
| 0.25 | 0.0754 | −0.2372 | 0.0635 | −0.2748 |
| 0.30 | 0.0635 | −0.2748 | 0.0498 | −0.3066 |
| 0.35 | 0.0498 | −0.3066 | 0.0344 | −0.3314 |
| 0.40 | 0.0344 | −0.3314 | 0.0179 | −0.3486 |
| 0.45 | 0.0179 | −0.3486 | 0.0004 | −0.3576 |
| 0.50 | 0.0004 | −0.3576 | — | — |

# Bibliografia

[1] R. Burden, J. Faires, “Numerical Analysis”, Brooks Cole, 2001.

[2] W. Cheney, D. Kincaid, “Numerical Mathematics and Computing”, Thomson Learning, 2004.

[3] S. Conte, C. de Boor, “Elementary Numerical Analysis: an Algorithmic Approach”, McGraw-Hill, 1987.

[4] H. Pina, “Métodos Numéricos”, McGraw-Hill, 1995.