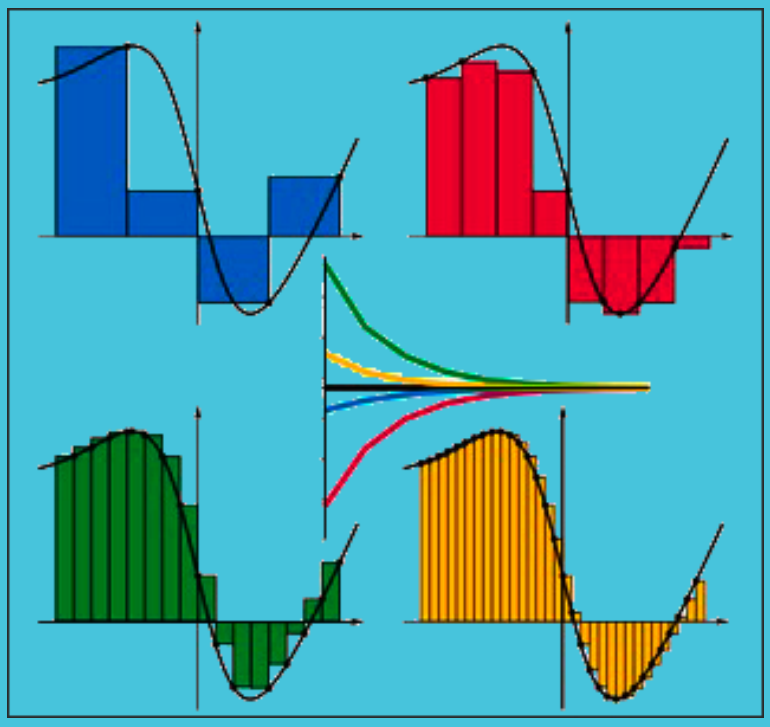
Métodos Numéricos

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO



Mário João Pires Fernandes Garcia Monteiro

DEPARTAMENTO DE FÍSICA E ASTRONOMIA DA
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
&
CENTRO DE ASTROFÍSICA DA UNIVERSIDADE DO PORTO

Versão: 30 Janeiro 2016

A figura na capa representa os quatro métodos de Riemann para estimar o integral da curva. Retirada da Wikipedia, do artigo http://en.wikipedia.org/wiki/Riemann_sum

Sumário

Neste conjunto de notas pretende-se dar uma visão geral dos métodos numéricos, das suas condições de aplicabilidade e das suas limitações. O tratamento de cada tópico é feito a nível introdutório, incidindo no significado geométrico sempre que possível.

Pretende-se transmitir conhecimentos que permitam a classificação e/ou a adequação de problemas numéricos a problemas tipo, a escolha do método numérico mais adequado e a compreensão dos diagnósticos de erro e diagnósticos de solução.

Este documento não deve ser usado como bibliografia única da disciplina de *Métodos Numéricos* mas apenas como referência dos tópicos abordados e nomenclatura usada. Assim, a sua utilização pressupõe uma leitura complementar para cada tópico, que deverá incluir - pelo menos - a consulta das referências bibliográficas indicadas no texto.

© 2000-2015 Mário J. P. F. G. Monteiro

O copyright das figuras, provenientes de outras fontes, pertence aos respectivos autores/editoras.


This work is licensed under the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0
International License [CC BY-NC-ND 4.0]

Historial

A primeira versão desta sebenta foi elaborada no ano lectivo de 2000/2001, baseando-se na sebenta de *Análise Numérica* do Prof. Manuel Rogério Silva (DMA/FCUP). Ao longo dos anos o conteúdo foi revisto e alargado, incluindo-se novos tópicos e extendendo os existentes, de forma a reforçar uma abordagen mais prática fortemente direccionada para a aplicação da análise numérica nas ciências e em engenharia.

Ao longo dos anos a sebenta tem tido correcções, adições e alterações significativas, contando para tal com contribuições dos alunos e dos colegas que participaram ao longo dos anos nesta disciplina. De referir em particular as correcções e sugestões do Prof. Jorge Filipe Gameiro e do Prof. Daniel Folha.

Conteúdo

Capítulo 1

Erros numéricos

Neste capítulo apresenta-se os aspectos básicos ligados a erros no cálculo numérico de quantidades. Devido à omnipresença deste tipo de erros sempre que se recorre ao cálculo numérico torna-se importante compreender como tal tipo de erros influenciam a resposta obtida. Por outro lado, quando lidamos com quantidades medidas temos de incluir na determinação dos resultados a incerteza que resulta da presença de erros nos dados. Assim é fundamental dispôr de ferramentas que estabeleçam, majorando, a incerteza no resultado de forma a qualificar o valor obtido.

1.1 Introdução

A necessidade de considerar erros numéricos é uma consequência do facto de nunca podermos representar computacionalmente todos os números reais com precisão infinita.

Um caso tipíco é por exemplo escrever um terço:

$$\frac{1}{3} = 0.3333333333333(3)\ . \tag{1.1.1}$$

Número este que nunca poderá ser escrito completamente. Daí que seja necessário não incluir uma parte do número, que neste caso seria por exemplo

$$0.0000000000333(3)\ , \tag{1.1.2}$$

se escrevessemos

$$\frac{1}{3} \sim 0.3333333333\ . \tag{1.1.3}$$

Existem duas formas de "medir" o erro associado a um número. Consideremos então que

$$\begin{cases} X & \text{- é o valor exacto de uma quantidade,} \\ x & \text{- um valor aproximado com que representamos essa quantidade} \end{cases} \tag{1.1.4}$$

Logo, x representa X com erro $\delta x \equiv X - x$, pois é uma aproximação.

1.1.1 Erro absoluto

Um deles, o **erro absoluto**, é o simples modulo do valor da parte do número que ignoramos, isto é,

$$0.0000000000333(3)\ . \tag{1.1.5}$$

Ou seja, por definição, é

$$\text{Erro absoluto} \equiv |X-x| = |\delta x| \,. \tag{1.1.6}$$

Pelo que podemos afirmar que se x é um valor aproximado de X, com um erro absoluto majorado por Δx, então

$$x-\Delta x \leq X \leq x+\Delta x \,. \tag{1.1.7}$$

Como desconhecemos o valor exacto de $|X-x|$, este é usualmente substituído por um majorante que representamos como sendo Δx.

1.1.2 Erro relativo

O outro é o chamado **erro relativo** que mede o valor relativo do erro absoluto quando comparado com o valor do número considerado. Neste caso tal erro seria

$$\frac{\text{erro absoluto}}{1/3} = 0.0000000000999(9) \,. \tag{1.1.8}$$

Logo, por definição, é

$$\text{Erro relativo} \equiv \frac{\Delta x}{|X|} = \left|\frac{X-x}{X}\right| = \left|\frac{\delta x}{X}\right| \,. \tag{1.1.9}$$

Note que dois números com o mesmo erro absoluto podem ter os seus erros relativos muito diferentes.

Exemplo 1.1.1: Consideremos os seguintes dois valores e uma representação aproximada destes;

X	x	Δx	$\frac{\Delta x}{\lvert X\rvert}$
π	3.14159	0.000003	0.000001
$\frac{1}{3003}$	0.00033	0.000004	0.02

Embora os erros absolutos sejam da mesma ordem de grandeza, os seus erros relativos são muito diferentes, indicando que no segundo caso a aproximação considerada é significativamente pior que no primeiro. Tal acontece porque usamos "menos" algarismos ao escrever o valor aproximado no segundo caso, daí estarmos a perder mais informação sobre o valor exacto - tal facto é indicado pelo valor do erro relativo.

1.1.3 Algarismos significativos

Definem-se como algarismos significativos de um número o primeiro algarismo que não é zero (analisando a partir da esquerda) bem como todos aqueles que estão à direita deste.

Exemplo 1.1.2: Os seguintes números têm todos 4 algarismos significativos:

$$010.23, \quad 0.0001000, \quad 2000, \quad 0.05234 \quad \text{e} \quad 1.064\times 10^4 \,.$$

O erro relativo de um número está directamente relacionado com a quantidade de algarismos significativos usados para o escrever.

Por exemplo, se um número X é aproximado, por arredondamento, por x com m algarismos significativos, então

$$\frac{\Delta x}{|X|} \leq 5\times 10^{-m} \,. \tag{1.1.10}$$

1.2 Erros de arredondamento

Erros de arredondamento são os que resultam da necessidade de representar números reais com um número finito de algarismos significativos.

Exemplo 1.2.1: Seja

$$X = \frac{1}{3} \qquad \text{com} \qquad x = 0.333333\ .$$

Então, o erro de arredondamento cometido ao escrever X é

$$\Delta x = 0.000000333(3)\ .$$

1.2.1 Regras de arredondamento

São normalmente consideradas duas regras básicas a usar no processo de arredondamento de um número. Estas são:

a) Se o primeiro algarismo a remover for inferior a 5 mantém-se o valor do último algarismo a conservar

b) Se o primeiro algarismo a remover for superior, ou igual, a 5 aumenta-se de uma unidade o último algarismo a conservar

Exemplo 1.2.2: Vejamos o caso de querermos arrendondar a quatro casas decimais:

Número:	21.0345209	21.0123689	0.0020001	0.0000123
Após arredondamento:	21.0345	21.0124	0.0020	0.0000

Note que então, se dissermos que x=12.572 é um valor aproximado de X, tendo sido obtido por arredondamento, então

$$12.5715 \leq X \leq 12.5725\ . \tag{1.2.1}$$

1.2.2 Representação dos erros

No caso de pretendermos representar o erro absoluto de um número não se deverá usar as regras de arredondamento mas sim indicar um majorante do erro. Isto é, se δx=$X-x$ tem um número infinito de algarismos (o que em geral acontece) então o erro absoluto deve ser representado por Δx, tal que

$$|X-x| \leq \Delta x\ , \tag{1.2.2}$$

de forma a garantir que

$$x-\Delta x \leq X \leq x+\Delta x\ . \tag{1.2.3}$$

Exemplo 1.2.3: Seja X=π. Se escrevermos um valor aproximado x=3.1416 então o erro cometido será

$$X - x = -0.000007346...$$

Daí que representemos o erro absoluto quando representamos X por x como sendo

$$|X - x| \leq 0.000008$$

De facto em presença de um error, passamos a representar um valor por um intervalo. Ou seja, dizemos que

$$X = x \pm \Delta x \qquad \text{ou} \qquad X \in [x - \Delta x, x + \Delta x]\,. \tag{1.2.4}$$

No caso do error relativo, que por definição é dado a partir do valor de X - que é desconhecido - temos de o representar mais uma vez por um majorante;

$$\left|\frac{\delta x}{X}\right| \leq \frac{\Delta x}{|x| - \Delta x}\,. \tag{1.2.5}$$

Exemplo 1.2.4: Considerando o exemplo anterior temos que

$$\pi = 3.1416 \pm 0.000008\,.$$

Então podemos escrever que

$$\pi \in [3.141592, 3.141608]\,,$$

sendo o error relativo majorado por

$$\frac{\Delta x}{|X|} \leq \frac{0.000008}{3.141592} \leq 3 \times 10^{-6}\,.$$

1.3 Erros de truncatura

Os erros de truncatura resultam da interrupção de algoritmos infinitos ou do uso de formulas aproximadas que representam um valor.

Exemplo 1.3.1: Uma forma de calcular o $\cos(c)$ é usando a série

$$\cos(c) \equiv 1 - \frac{c^2}{2!} + \frac{c^4}{4!} - \frac{c^6}{6!} + \ldots + (-1)^k \frac{c^{2k}}{(2k)!} + \ldots$$

Se calcularmos então o valor de $\cos(\pi/3)$ (que sabemos ser exactamente 0.5) usando apenas os primeiros três termos, temos que (para $c{=}\pi/3$);

$$\cos(c) \sim 1 - \frac{\pi^2}{18} + \frac{\pi^4}{1944} = 0.5017962\ldots\,,$$

logo o erro cometido ao truncar a série infinita para o "cos" no terceiro termo é de 0.0017962...

De forma a podermos majorar o erro de truncatura de desenvolvimentos em série de funções recorre-se às propriedades das séries, podendo-se em alguns casos encontrar um majorante para a soma dos termos desprezados. Isto é, seja

$$Y = \sum_{n=1}^{\infty} u_n = u_1 + u_2 + u_3 + \ldots + u_k + u_{k+1} + \ldots\,, \tag{1.3.1}$$

e consideremos um valor aproximado y dado pelo soma dos primeiros k termos;

$$y = \sum_{n=1}^{k} u_n\,. \tag{1.3.2}$$

O erro cometido ao aproximar Y pelo valor y é então

$$\begin{aligned} \delta y &= Y - y = R_k \\ &\equiv \sum_{n=k+1}^{\infty} u_n = u_{k+1} + u_{k+2} + \dots . \end{aligned} \tag{1.3.3}$$

O nosso objectivo é agora estabelecer critérios que nos permitam majorar o resto R_k para uma determinada série de forma a estimar $\Delta y \geq |R_k|$.

1.3.1 Critério da série alternada

Vejamos o caso de uma série alternada da forma

$$Y = \sum_{n=1}^{\infty} (-1)^n u_n \tag{1.3.4}$$

cujo valor aproximado é calculado a partir da soma dos k primeiros termos;

$$y = \sum_{n=1}^{k} (-1)^n u_n \, . \tag{1.3.5}$$

Sendo a série convergente então o erro de truncatura, que corresponde à soma dos termos desprezados, será majorado de acordo com

$$R_k = \sum_{n=k+1}^{\infty} (-1)^n u_n \leq |u_{k+1}| \, . \tag{1.3.6}$$

Exemplo 1.3.2: Consideremos então a série para o cálculo de $\cos(c)$;

$$\cos(c) \equiv \sum_{n=0}^{\infty} (-1)^n \frac{c^{2n}}{(2n)!} \, .$$

Então, aproximando o valor de $Y = \cos(c)$ pela soma y dos primeiros $k+1$ termos, isto é

$$y = \sum_{n=0}^{k} (-1)^n \frac{c^{2n}}{(2n)!} \, ,$$

cometemos um erro δy majorado por

$$\Delta y = |R_k| \leq \left| \frac{c^{2k+2}}{(2k+2)!} \right| \, ,$$

de acordo com a expressão (1.3.6) para o erro de truncatura.

Exemplo 1.3.3: Vejamos o caso de querermos o valor de $Y = \mathrm{e}^{-c^2}$ $(0 \leq c \leq 1)$, com um erro absoluto de truncatura majorado por ε_t. Como

$$Y = \mathrm{e}^{-c^2} = \sum_{n=0}^{\infty} (-1)^n \frac{c^{2n}}{n!} \, ,$$

então temos que

$$|R_k| \leq |u_{k+1}| = \frac{c^{2k+2}}{(k+1)!} \leq \varepsilon_t \, .$$

Ou seja, a última desigualdade desta expressão, permite-nos determinar o número mínimo de termos (valor de k) que é necessário somar para obter o valor y com um erro de truncatura inferior a ε_t.

1.3.2 Critério de d'Alembert

Considermos uma série de termo geral u_n (com $u_i \cdot u_{i+1} \geq 0$ para qualquer $i \in \{1,2,3,...\}$) e tal que

$$\frac{u_{n+1}}{u_n} \leq \alpha < 1 \, . \tag{1.3.7}$$

Então

$$\begin{aligned} R_k &= u_{k+1} + u_{k+2} + u_{k+3} + ... \\ &\leq \alpha u_k + \alpha^2 u_k + \alpha^3 u_k + ... = u_k \sum_{i=1}^{\infty} \alpha^i = u_k \, \frac{\alpha}{1-\alpha} \, , \end{aligned} \tag{1.3.8}$$

ou seja, teremos que

$$|R_k| \leq \Delta y \leq |u_k| \, \frac{\alpha}{1-\alpha} \, . \tag{1.3.9}$$

Exemplo 1.3.4: A exponencial de c pode ser calculada através da seguinte série:

$$Y = e^c = \sum_{n=0}^{\infty} \frac{c^n}{n!} \qquad ; \, 0 \leq c \leq 1 \, .$$

Pelo que neste caso

$$\frac{u_{n+1}}{u_n} = c \, \frac{1}{n+1} \leq \frac{1}{n+1} \, ,$$

que será sempre menor que $1/2$ para $n \geq 1$. Assim, o erro de aproximar Y por y, em que

$$y = 1 + \sum_{n=1}^{k} \frac{c^n}{n!} \, , \quad \text{é dado por} \quad \Delta y \leq \frac{c^k}{k!} \, \frac{\alpha}{1-\alpha} \, ,$$

onde $\alpha = 1/2$.

Exemplo 1.3.5: Calculemos então o valor do número de Napier e, com um erro inferior a $\varepsilon = 5 \times 10^{-4}$. Como já vimos o erro de truncatura (se considerarmos k termos) será majorado de acordo com

$$\Delta y_t \leq \frac{1}{k!} \, .$$

Mas necessitamos ainda incluir o efeito dos erros de arredondamento no cáculo de cada um dos k primeiros termos da série. Se estabelecermos que uma fracção de ε será absorvida para erros de arredondamento, e o restante devido á truncatura da série no termo $k+1$, podemos estabelecer quantos termos temos de somar e com que precisão cada um deles deve ser escrito. Seja então,

$$\varepsilon = \varepsilon_t + \varepsilon_a \qquad \text{com} \qquad \varepsilon_t = 3 \times 10^{-4} \quad \text{e} \quad \varepsilon_a = 2 \times 10^{-4} \, ,$$

implicando que k é tal que

$$\varepsilon_t \geq \frac{1}{k!} \geq \Delta y_t \, .$$

Logo $k \geq 7$. Sendo assim, o erro máximo de arredondamento de cada um dos termos a somar será de $\varepsilon_{ap} = \varepsilon_a / 7 = 2.8 \times 10^{-5}$.

Exemplo 1.3.6: Façamos então o cálculo;

n	$1/n!$	Valor a usar
0	1	1.00000
1	1	1.00000
2	0.5	0.50000
3	0.166666666(6)	0.16667
4	0.041666666(6)	0.04167
5	0.008333333(3)	0.00833
6	0.001388888(8)	0.00139
7	0.000198412...	0.00020
8	0.000024801...	0.00002
	Total	2.71828

Logo, o resultado é

$$e = 2.718 \pm 0.0005 \, .$$

Lembrando que o valor exacto é 2.718281828..., pode-se verificar que de facto o resultado obtido está dentro da precisão requerida.

1.3.3 Critério de Cauchy

Consideremos agora o caso de o termo geral da série u_n ser tal que

$$(u_n)^{1/n} \leq \alpha < 1 \, . \tag{1.3.10}$$

Então $u_n \leq \alpha^n$, de onde resulta que

$$R_k \leq \alpha^k \sum_{i=1}^{\infty} \alpha^i = \frac{\alpha^{k+1}}{1-\alpha} \, . \tag{1.3.11}$$

Expressão esta que nos permite majorar o erro de truncatura.

Exemplo 1.3.7: Como vimos anteriormente a exponencial de c pode ser calculada através da seguinte série:

$$Y = e^c = \sum_{n=0}^{\infty} \frac{c^n}{n!} \qquad ; 0 \leq c \leq 1 \, .$$

Pelo que, para aplicação do critério de Cauchy precisamos encontrar α tal que

$$(u_n)^{1/n} = \frac{c}{(n!)^{1/n}} \leq \alpha \, .$$

Como,

$$n! \geq 2^{n-1} \qquad \Rightarrow \qquad (n!)^{1/n} \geq 2^{\frac{n-1}{n}} \geq \sqrt{2} \, ,$$

se $n \geq 2$, temos que $\alpha = 1/\sqrt{2}$ quando $n \geq 2$. Assim, o erro de aproximar Y por y, em que

$$y = 1 + c + \sum_{n=2}^{k} \frac{c^n}{n!} \, ,$$

é dado por

$$\Delta y \leq \frac{\alpha^{k+1}}{1-\alpha} \, ,$$

onde $\alpha = 1/\sqrt{2}$.

1.4 Avaliação de funções

Procuremos agora uma forma de estimar o erro de uma quantidade no caso em que esta é dada através de uma expressão envolvendo num parâmetro do qual apenas temos um valor aproximado, isto é, com erro.

1.4.1 Funções de um parâmetro

Seja Y o valor que uma função f toma em X. Isto é,

$$Y = f(X) \, . \tag{1.4.1}$$

Se agora tivermos um valor x que aproxima X (onde $\delta x{=}X{-}x$), então o valor y que obtemos a partir de $f(x)$ está relacionado com Y através da relação,

$$\begin{aligned} \delta y &= Y{-}y = f(X) - f(x) = f(x{+}\delta x) - f(x) \\ &= f'(\xi)\ \delta x \end{aligned} \tag{1.4.2}$$

onde $\xi \in [x, x{+}\delta x] {\equiv} I$. Em que se usou o teorema do valor médio, pressupondo que as condições de aplicabilidade são verificadas. Caso assim não seja torna-se necessário considerar o termo seguinte na expansão de $f(x{+}\delta x)$.

Podemos também escrever uma expressão aproximada para estimar o erro na avaliação de Y, que é

$$\Delta y \leq \mathrm{Max}_{a \in I} \left| f'(a) \right| \ \Delta x \, . \tag{1.4.3}$$

Exemplo 1.4.1: Dispomos de um valor aproximado para X dado por $x{=}1.20$, que é conhecido com um erro $\Delta x{\leq}0.01$. Determinemos então o erro com que podemos conhecer um valor Y, se este é obtido a partir de X recorrendo à seguinte função;

$$Y = f(X) \equiv X \ \log X \, .$$

Usando a expressão (1.4.3), e já que

$$f'(x) = \log x + 1 \, ,$$

temos que

$$\Delta y \leq \mathrm{Max}_{a \in [1.19, 1.21]} |\log a + 1| \times \Delta x = (\log 1.21 + 1) \times 0.01 \leq 0.02 \, .$$

Temos então que $y{=}1.20{\times}\log 1.20{=}0.22$ representa Y com um erro $\Delta y{\leq}0.02$.

1.4.2 Funções de vários parâmetros

Se tivermos uma quantidade Z calculada através do uso de uma função f de dois valores X e Y, isto é

$$Z = f(X, Y) \, , \tag{1.4.4}$$

então se x e y são valores aproximados de X e Y, respectivamente, temos que um valor aproximado de Z é

$$z = f(x, y) \, . \tag{1.4.5}$$

O erro desta aproximação é dado por, onde $\delta x{=}X{-}x$ e $\delta y{=}Y{-}y$,

$$\begin{aligned} \delta z &\equiv Z{-}z = f\,(x{+}\delta x, y{+}\delta y) - f(x, y) \\ &= \frac{\partial f}{\partial x}(\xi_{xx}, \xi_{xy})\ \delta x + \frac{\partial f}{\partial y}(\xi_{yx}, \xi_{yy})\ \delta y \, , \end{aligned} \tag{1.4.6}$$

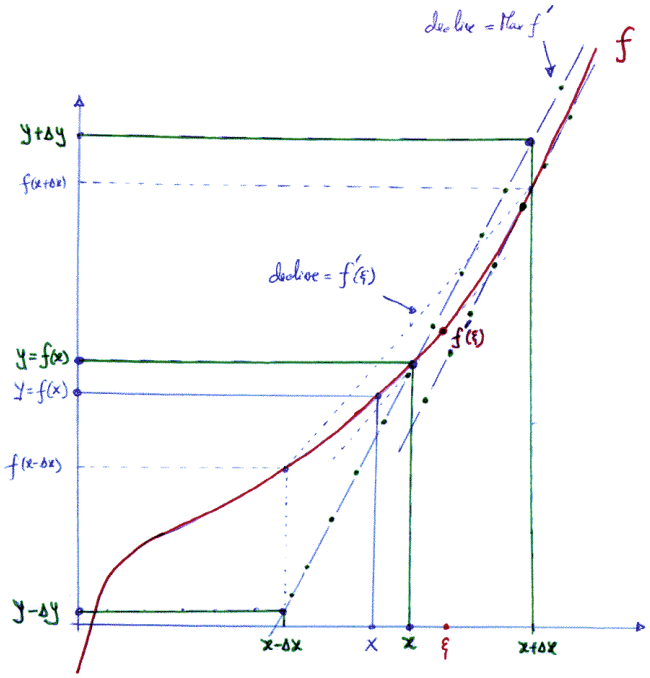


Figura 1.1: Representação do efeito de aproximar a derivada pelo seu valor máximo em I.

onde ξ_{xx} e $\xi_{yx} \in [x, x+\delta x] = I_x$, bem como ξ_{xy} e $\xi_{yy} \in [y, y+\delta y] = I_y$.

Daí que uma estimativa do erro com que z aproxima Z seja dada por

$$\begin{aligned} \Delta z &\leq \left|\frac{\partial f}{\partial x}(\xi_{xx}, \xi_{xy})\ \delta x\right| + \left|\frac{\partial f}{\partial y}(\xi_{yx}, \xi_{yy})\ \delta y\right| \\ &\leq \left|\frac{\partial f}{\partial x}(\xi_{xx}, \xi_{xy})\right|\ \Delta x + \left|\frac{\partial f}{\partial y}(\xi_{yx}, \xi_{yy})\right|\ \Delta y\,. \end{aligned} \tag{1.4.7}$$

Defenindo

$$\bar{f}_x \equiv \mathrm{Max}_{I_x, I_y}\left|\frac{\partial f}{\partial x}\right| \qquad \text{e} \qquad \bar{f}_y \equiv \mathrm{Max}_{I_x, I_y}\left|\frac{\partial f}{\partial y}\right|\,, \tag{1.4.8}$$

fica finalmente que

$$\Delta z \leq \bar{f}_x\ \Delta x + \bar{f}_y\ \Delta y\,. \tag{1.4.9}$$

Exemplo 1.4.2: Consideremos o produto de dois números: Z=XY. Se tivermos que $x-\Delta x \leq X \leq x+\Delta x$ e que $y-\Delta y \leq Y \leq y+\Delta y$, então podemos escrever que:

$$(x-\Delta x)(y-\Delta y) \leq XY \leq (x+\Delta x)(y+\Delta y)\,,$$

resultando que

$$xy-(y\Delta x+x\Delta y+\Delta x\Delta y) \leq XY \leq xy+(y\Delta x+x\Delta y-\Delta x\Delta y)\,.$$

Da expressão (1.4.9) temos que

$$\begin{aligned} \Delta z &= \mathrm{Max}_{I_y}|y|\cdot\Delta x+\mathrm{Max}_{I_x}|x|\cdot\Delta y=(y+\Delta y)\Delta x+(x+\Delta x)\Delta y \\ &= y\Delta x+x\Delta y+2\Delta x\Delta y \\ \Rightarrow \quad & y\Delta x+x\Delta y=\Delta z-2\Delta x\Delta y\,. \end{aligned}$$

Após substituir acima, temos que

$$xy-\Delta z+(3\Delta x\Delta y) \leq XY \leq xy+\Delta z-(\Delta x\Delta y)\,.$$

Desta forma fica mostrado que sendo $\Delta x\Delta y \geq 0$ temos necessariamente que

$$z-\Delta z \leq XY \leq z+\Delta z\,,$$

onde z=xy. Isto é, Δz é majorante do erro quando estimamos o valor de XY por xy.

Exemplo 1.4.3: Dispomos de valores aproximados para (X,Y), dados respectivamente por x=1.20 com um erro $\Delta x \leq 0.01$ e y=2.2 com um erro $\Delta y \leq 0.1$. Determinemos então o erro com que podemos conhecer um valor Z, quando este é obtido a partir de X e Y recorrendo-se à seguinte função;

$$Z=f(X,Y)\equiv Y\ \log X\ .$$

Usando a expressão (1.4.9), e já que

$$\frac{\partial f}{\partial x}=\frac{y}{x} \qquad \text{e} \qquad \frac{\partial f}{\partial y}=\log x\,,$$

temos que

$$\Delta z \leq \frac{2.3}{1.19}\,\Delta x+\log 1.21\ \Delta y \leq 0.04\ .$$

Temos então que z=2.2$\times\log 1.20$=0.40 representa Z com um erro $\Delta z \leq 0.04$.

1.4.3 Exemplos: soma e subtracção de valores

Seja $Z=aX+bY$, onde a e b são dois valores reais. Então, usando a Eq. (1.4.9), temos que

$$\Delta z \leq |a|\Delta x+|b|\Delta y\,. \tag{1.4.10}$$

Exemplo 1.4.4: No caso de x=0.012 (Δx=0.0005) e y=2.11 (Δy=0.005), e com a=b=1 temos que

$$z = 2.12 \pm 0.006 \qquad \text{e com} \qquad \frac{\Delta z}{|Z|} \leq 0.003 \, .$$

Mas no caso de x=1.193 (Δx=0.0005) e y=1.21 (Δy=0.005), e com a=$-b$=1 temos que

$$z = -0.017 \pm 0.006 \qquad \text{com} \qquad \frac{\Delta z}{|Z|} \leq 0.4 \, .$$

A razão de tal diferença entre os erros relativos, pode-se analisar escrevendo a expressão geral para o erro relativo;

$$\frac{\Delta z}{|Z|} \leq \frac{|a| \, \Delta x + |b| \, \Delta y}{|aX + bY|} \, . \tag{1.4.11}$$

Consideremos agora que

$$\varepsilon \geq \frac{\Delta x}{|X|} \qquad \text{e} \qquad \varepsilon \geq \frac{\Delta y}{|Y|} \, , \tag{1.4.12}$$

logo

$$\frac{\Delta z}{|Z|} \leq \frac{|a\,X| + |b\,Y|}{|aX + bY|} \, \varepsilon \, . \tag{1.4.13}$$

Se $(aX)\cdot(bY) > 0 \Rightarrow |aX|+|bY|=|aX+bY|$, temos que

$$\frac{\Delta z}{|Z|} \leq \varepsilon \, . \tag{1.4.14}$$

No entanto se $(aX)\cdot(bY) < 0 \Rightarrow |aX|+|bY| \geq |aX+bY|$, pode-se ter que

$$\frac{\Delta z}{|Z|} \gg \varepsilon \, , \tag{1.4.15}$$

no caso de $|Z|$ ser pequeno relativamente a $|aX|$ e $|bY|$.

1.5 Soluções de funções implícitas

Considermos agora o caso em que duas ou mais quantidades estão relacionadas por uma expressão não se podendo escrever qualquer delas com função explicita das restantes. Qualquer erro numa delas reflecte-se no valor que as outras terão.

Seja então F uma função de X e Y; a equação

$$F(X,Y)=0 \, , \tag{1.5.1}$$

define Y como função implícita de X, isto é $Y=Y(X)$.

Exemplo 1.5.1: Seja F a seguinte função;

$$F(X,Y) \equiv \mathrm{e}^{Y} - X - 1 \, .$$

Então $F(X,Y)=0$ define implicitamente a função $Y \equiv Y(X)$, que neste caso sabemos ser $Y(X)=\log(1+X)$ (para $X+1>1$).

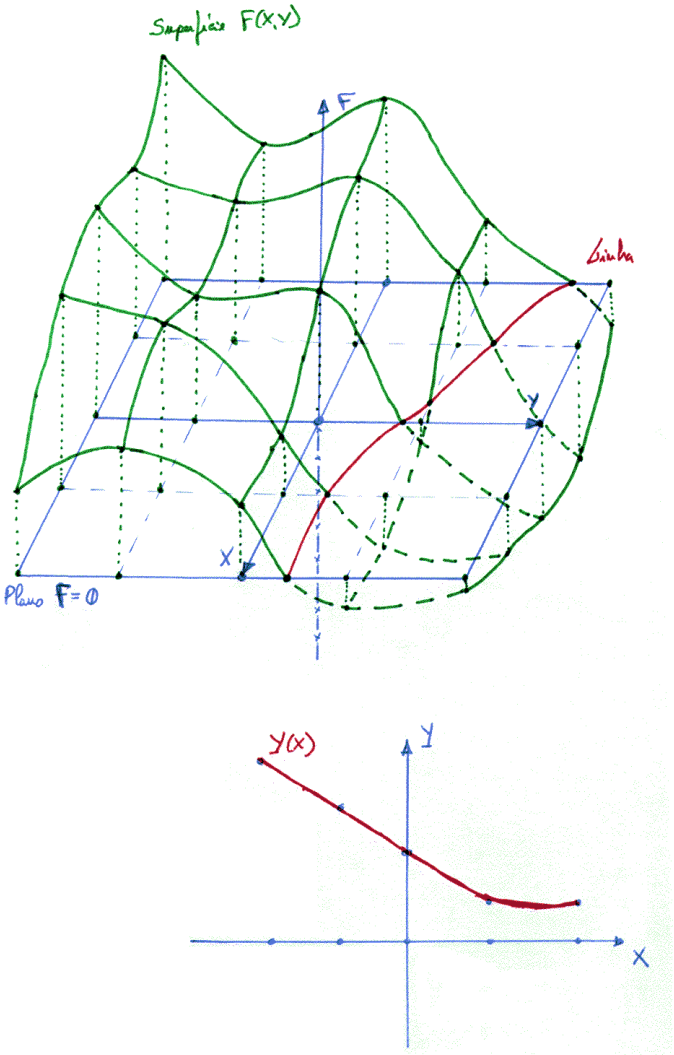


Figura 1.2: Representação da intersecção de uma superfície $F(X,Y)$ com o plano $F=0$, definindo-se desta forma pela curva de intersecção uma função implícita $Y=Y(X)$.

Exemplo 1.5.2: No entanto podem existir mais do que um função implícita, tal como definida por uma equação. Um exemplo é;

$$F(X,Y) = Y^2 - X^2 + X - Y = (Y-X)(Y+X-1) \,,$$

cujas funções implícitas são $Y(X)=X$ e $Y(X)=1-X$.

Consideremos agora que temos um valor x próximo de X, e determinemos o valor y associado a este, isto é, tal que $F(x,y)=0$. Sendo $\delta x=X-x$ e $\delta y=Y-y$, temos então que

$$\begin{aligned} F(X,Y) &= F(x+\delta x, y+\delta y) \\ &= F(x,y) + \frac{\partial F}{\partial x}(\xi_{xx},\xi_{xy})\,\delta x + \frac{\partial F}{\partial y}(\xi_{yx},\xi_{yy})\,\delta y \,. \end{aligned} \tag{1.5.2}$$

com ξ_{xx} e $\xi_{yx} \in [x, x+\delta x] = I_x$, e com ξ_{xy} e $\xi_{yy} \in [y, y+\delta y] = I_y$.

Daqui resulta que, pois $F(X,Y)=F(x,y)=0$,

$$\delta y = -\frac{\dfrac{\partial F}{\partial x}(\xi_{xx},\xi_{xy})}{\dfrac{\partial F}{\partial y}(\xi_{yx},\xi_{yy})}\,\delta x \,. \tag{1.5.3}$$

Dando que

$$y = Y + \frac{\dfrac{\partial F}{\partial x}(\xi_{xx},\xi_{xy})}{\dfrac{\partial F}{\partial y}(\xi_{yx},\xi_{yy})}\,(X-x) \,. \tag{1.5.4}$$

Da equação (1.5.3) temos também que

$$\Delta y = \left| \frac{\dfrac{\partial F}{\partial x}(\xi_{xx},\xi_{xy})}{\dfrac{\partial F}{\partial y}(\xi_{yx},\xi_{yy})} \right| \Delta x \,. \tag{1.5.5}$$

Se defenirmos

$$\bar{F}_x = \mathrm{Max}_{I_x,I_y} \left| \frac{\partial F}{\partial x} \right| \qquad \text{e} \qquad \bar{F}_y = \mathrm{Min}_{I_x,I_y} \left| \frac{\partial F}{\partial y} \right| \,, \tag{1.5.6}$$

então

$$\Delta y \leq \frac{\bar{F}_x}{\bar{F}_y}\,\Delta x \,. \tag{1.5.7}$$

Exemplo 1.5.3: Consideremos a seguinte equação, com $z \in [0, \pi/2]$, para a qual se conhece uma raiz;

$$1.2300^z - \tan z = 0 \,.$$

Determinemos então o que acontece à raiz da equação quando substituimos 1.2300 por 1,2345. Isto é, se

$$F(X,Y) = X^Y - \tan Y \,,$$

tem uma raiz $(X,Y)=(1.2300, Y_0)$, determinemos então a nova raiz que corresponde a $(x,y)=(1.2345, ?)$. Considerando que $F=0$ define $Y=Y(X)$ como função implícita, e depois de se usar a Eq. (1.5.4), ficamos com

$$y = Y_0 + \frac{1.2300^{Y_0-1} Y_0}{1.2300^{Y_0} \log 1.2300 - (1+\tan^2 Y_0)}\,(1.2300-1.2345) \,,$$

que nos dá uma estimativa, y, para a nova raiz.

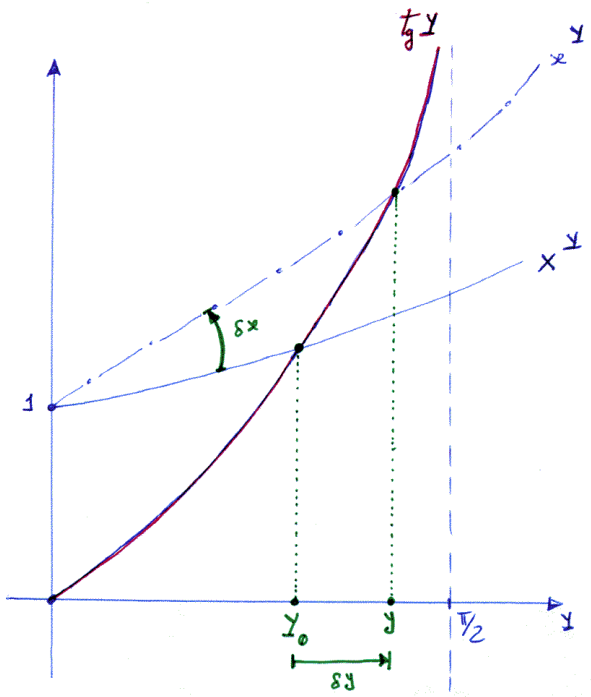


Figura 1.3: Representação do efeito da solução implícita do exemplo dado, quando se altera o valor de X=1.2300 para X=1.2345, a que corresponde uma nova raiz Y tal como pode ser calculada pela expressão obtida.

Para o caso mais geral de termos vários parâmetros, isto é, quando a função implícita é definida por

$$F(X_1, X_2, ..., X_n, Y) = 0\,, \tag{1.5.8}$$

então Eq. (1.5.2) corresponde a

$$F(X_1, X_2, ..., X_n, Y) = F(x_1, x_2, ..., x_n, y) + \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}\,\delta x_i + \frac{\partial F}{\partial y}\,\delta y\,, \tag{1.5.9}$$

de onde temos que

$$y = Y + \frac{\displaystyle\sum_{i=1}^{n} \frac{\partial F}{\partial x_i}\,(X - x_i)}{\dfrac{\partial F}{\partial y}}\,. \tag{1.5.10}$$

Daí que

$$\Delta y \leq \frac{1}{\bar{F}_y} \sum_{i=1}^{n} \bar{F}_{x_i}\,\Delta x_i\,, \tag{1.5.11}$$

onde

$$\bar{F}_y = \mathrm{Min}_{I_{x_1}, I_{x_2}, ..., I_{x_n}, I_y} \left|\frac{\partial F}{\partial y}\right| \qquad \text{e} \qquad \bar{F}_{x_i} = \mathrm{Max}_{I_{x_1}, I_{x_2}, ..., I_{x_n}, I_y} \left|\frac{\partial F}{\partial x_i}\right|\,. \tag{1.5.12}$$

1.6 Exercícios

E1.1) Qual o número de casa decimais e algarismos significativos dos seguintes valores: 0.012300, 123.0501, 3 e 3.000 ? Se foram escritos por arredondamento qual o valor do erro relativo com que cada um deles é dado?

E1.2) Escreva com 4 algarismos significativos, por arredondamento, os seguintes números:

$$29.63243, \quad 81.9773, \quad 4.4985001, \quad 11.63489, \quad 53908, \quad 0.0900038 \quad \text{e} \quad 2345234.$$

E1.3) Considere o seguinte número x=201.259023, e represente-o tal que:

a) $\Delta x \leq 0.01$

b) $\Delta x \leq 0.07$

c) $\Delta x \leq 1.5$

d) $\frac{\Delta x}{|X|} \leq 10^{-4}$

e) $\frac{\Delta x}{|X|} \leq 2\times 10^{-2}$

E1.4)* O número e tem a seguinte definição;

$$e \equiv \lim_{n\to\infty} \left(1+\frac{1}{10^n}\right)^{10^n} .$$

Tente obter este limite no seu computador usando valores de n sucessivamente mais elevados, comparando-os com o valor exacto. Explique o sucedido.

E1.5) Sabendo que o valor do número de Neper é dado por

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

determine o número de termos da série que necessita somar para obter o valor de e com quatro algarismos significativos. (Por definição 0!=1.)

E1.6) Calcule a soma da série ($n\geq 0$) de termo geral

$$a_n = \frac{x^{n+1}}{n!}$$

com erro inferior a 10^{-3}, para o caso de x=1/3.

E1.7)* Escreva um algoritmo que, a partir de um valor de x lido em graus, permita o cálculo de

$$y = \cos(x)$$

com erro inferior a ε através do seu desenvolvimento em série de Taylor. Implemente-o, e calcule o valor para ε=10^{-3} e 10^{-5}.

E1.8) Dadas as expressões equivalentes

$$\frac{1}{(2+\sqrt{3})^4} \qquad \text{e} \qquad 97-56\sqrt{3}$$

calcule os seus valores usando $\sqrt{3} = 1.732$. Interprete os resultados.

E1.9) Compare os resultados $f(500)$ e $g(500)$ usando seis algarismos significativos com arrendondamento, em que

$$f(x) = x\left(\sqrt{x+1} - \sqrt{x}\right) \qquad \text{e} \qquad g(x) = \frac{x}{\sqrt{x+1} + \sqrt{x}} .$$

Qual a razão para a diferença dos valores?

E1.10) Considere a seguinte equação (para $x \neq 0$);

$$\sum_{n=0}^{\infty} \frac{1}{n!} - log\, x^2 = 0 .$$

Determine as soluções da equação com um erro $\varepsilon \leq 10^{-2}$.

E1.11) Calcule o valor de $0.12^{3.1}$ e indique o erro absoluto do resultado sabendo que os dados foram aproximados por arredondamento.

E1.12) Calcule o valor de $\pi \times e$, com erro máximo absoluto inferior a 10^{-4}.

E1.13) Qual o número mínimo de casas decimais que deve considerar em valores aproximados de $\sqrt{3}$ e $\sqrt{5}$ para calcular $z = \sqrt{3}(\sqrt{5}-1)^2$ com erro não superior a 10^{-5}? Justifique.

E1.14) Seja

$$f(x) = \frac{3.55}{4.26\, x + 6.22} .$$

Determine os erros absoluto e relativo que se cometem ao calcular $f(x)$ para $x = 1.5$ (valor exacto) supondo os coeficientes obtidos por arredondamento.

E1.15) Dada a função $z = \log(\tan x) + 0.2^y$ com x expresso em radianos, determine os erros máximos absolutos que se podem admitir em x e y, para que se possa calcular z com erro absoluto não superior a 10^{-5}, sendo $x = 38^o 27' 5.3''$ e $y = 1.0759214$.

E1.16)* Calcule no seu computador favorito o valor da expressão

$$z = \frac{(x+y)^2 - x^2 - 2\, xy}{y^2} ,$$

com $x = 100.0$ e $y = 10^{-k}$, para $k = 0, 1, 2, 3, 4, \ldots$ Explique o sucedido e perdoe-lhe!

E1.17) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = r\, \sin x - e^x$, onde r é um número real. Se $X = -3.1696$ é raiz da equação para $r = 1.5$, então estime o valor da raiz a partir do valor dado no caso de $r = 1.501$.

E1.18) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = e^x - (x+2)$. Se o valor "2" na definição de $F(x)$ não é exacto mas sim escrito por arredondamento com um erro de 5×10^{-4}, calcule qual a precisão com que pode determinar a raiz pertencente ao intervalo $[1, 2]$.

E1.19) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = \cos(\pi x) - \frac{3}{2}\, \pi x e^{2x}$. Encontre a expressão que permite estimar a alteração do valor de cada uma das raízes da equação, se em vez dos valores exactos usarmos $e \sim 2.718$ e $\pi \sim 3.142$.

E1.20) Qual o valor aproximado da raíz da equação

$$1.24^{-x} - \sin(0.98x) = 0$$

sabendo que

$$1.24^{-x} - \sin(x) = 0$$

admite uma raíz cujo valor é aproximado por 0.95.

E1.21) Determine um valor aproximado da raíz da equação

$$x = e^{-x^2} \ ,$$

com $x>0$, sabendo que

$$1.03\, x = e^{-0.97\, x^2}$$

tem uma raíz de valor aproximado 0.65.

E1.22) Sendo 1.30 um valor aproximado da raíz da equação

$$\log_3 x - 3^{-x} = 0$$

determine um valor aproximado da raíz de

$$\log_{3.02} x - 3.01^{-x} = 0 \ .$$

Capítulo 2

Resolução numérica de equações

Nem sempre é possível obter uma solução analítica (exacta) de uma equação. Nesse caso necessitamos de recorrer a métodos numéricos para encontrar uma solução aproximada da nossa equação. Nesta Secção apresentamos alguns desses métodos, discutindo a forma como nos permitem calcular um valor aproximado da raiz, o erro com que tal é feito e as condições de aplicabilidade necessárias para a sua implementação.

2.1 Localização das raízes

De forma a tentarmos determinar numericamente uma raiz de uma equação precisamos primeiro de definir um intervalo que contenha a raiz, e apenas essa raiz. Para isso precisamos de encontrar métodos que nos permitam isolar as raízes de uma equação para implementarmos os métodos de cálculo numérico de raízes.

2.1.1 Números de Rolle

Seja $F(x)$ uma função real, de variável real, contínua e com derivada finita num domínio $D \subset \mathscr{R}$. Então:

- os pontos fronteira do domínio
- os zeros da derivada $F'(x)$ da função

constituem o conjunto dos números de Rolle da equação $F(x)=0$. Depois de ordenados, gozam das seguintes propriedades:

→ entre dois números consecutivos existe, quando muito, uma raiz da equação,

Pois se existissem duas (ou mais) raizes haveria, necessariamente, um zero da derivada entre elas já que a função é contínua e tem derivada finita em $\mathscr{D}$.

→ não terá nenhuma raiz se $F(x)$ tiver o mesmo sinal nesses dois pontos,

Se a função tem o mesmo sinal em dois pontos consecutivos, então terá obrigatoriamente - pois é continua - um número par de raizes entre estes dois pontos. Tal não é possível (pois, quando muito, tem uma) logo não tem nenhuma raiz.

→ existe uma raiz se o sinal for contrário.

Se a função tem sinais contrários em dois pontos consecutivos, então terá obrigatoriamente - pois é continua - um número impar de raizes entre estes dois pontos. Mas como não pode ter mais que uma, então tem necessariamente uma raiz.

Exemplo 2.1.1: Seja $F(x)=x\log x-1$ uma função definida em $x\in]0,+\infty[$, e queremos localizar todos os zeros da equação $F(x)=0$.

$$F'(x)=0 \quad \Rightarrow \quad \log x+1=0 \quad \Rightarrow \quad x=1/\mathrm{e}\,.$$

Assim, os números de Rolle desta equação são $\{0,1/\mathrm{e},+\infty\}$. Nestes pontos o sinal de $F(x)$ é;

$$\begin{aligned}\lim_{x\to 0^+} F(x) &= -1+\lim_{x\to 0^+}\frac{\log x}{1/x}=-1+\lim_{x\to 0^+}\frac{1/x}{-1/x^2}=-1\,,\\ F(1/\mathrm{e}) &= -(1+1/\mathrm{e})\,,\\ \lim_{x\to\infty} F(x) &= +\infty\,;\end{aligned}$$

isto é $\{-,-,+\}$. Logo temos que;
$\to$ não existem zeros no intervalo $]0,1/\mathrm{e}]$,
$\to$ a função tem um zero no intervalo $[1/\mathrm{e},+\infty[$.

Exemplo 2.1.2: Seja $F(x)=\mathrm{e}^x-(x+2)$ uma função definida em $x\in]-\infty,+\infty[$, e queremos localizar todos os zeros da equação $F(x)=0$.

$$F'(x)=0 \quad \Rightarrow \quad \mathrm{e}^x-1=0 \quad \Rightarrow \quad x=0\,.$$

Assim, os números de Rolle desta equação são $\{-\infty,0,+\infty\}$. Nestes pontos o sinal de $F(x)$ é;

$$\begin{aligned}\lim_{x\to-\infty} F(x) &= +\infty\,,\\ F(0) &= -1\,,\\ \lim_{x\to+\infty} F(x) &= +\infty\,;\end{aligned}$$

isto é $\{+,-,+\}$. Logo temos que;
$\to$ a função tem um zero no intervalo $]-\infty,0]$,
$\to$ a função tem um zero no intervalo $[0,+\infty[$.

2.1.2 Método gráfico

Por vezes determinar os zeros da derivada $F'(x)$ da função é tão difícil como encontrar os zeros de $F(x)$, pelo que é necessário encontrar outra forma de localizar os zeros da equação $F(x)=0$.

Como qualquer equação $F(x)=0$ pode ser escrita na forma

$$x=f(x) \qquad \text{ou} \qquad f_1(x)=f_2(x)\,, \tag{2.1.1}$$

recorremos a estas formas alternativas de escrever a equação, caso as funções $f(x)$ ou $f_{1,2}(x)$ sejam de facil representação gráfica. Os zeros da equação inicial correspondem às intersecções, no primeiro caso, da recta $y=x$ com a função $y=f(x)$, e no segundo, da intersecção das duas curvas $y=f_1(x)$ e $y=f_2(x)$.

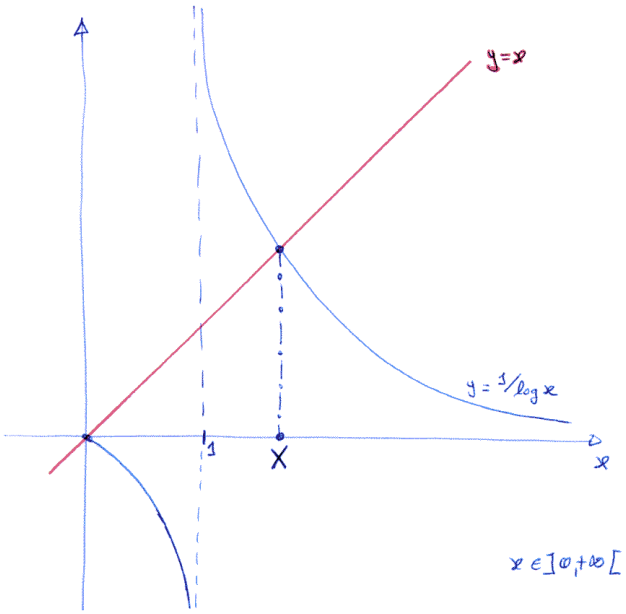


Figura 2.1: Representação das curvas $y{=}x$ e $y{=}f(x){=}1/\log x$. A intersecção entre elas corresponde aos zeros da equação $F(x){=}0$, que neste caso ocorre no intervalo $]1,+\infty[$.

Exemplo 2.1.3: Consideremos novamente a função $F(x){=}x\log x{-}1$, definida para a variável real $x{\in}]0,+\infty[$, e vamos tentar localizar todos os zeros da equação $F(x){=}0$ pelo método gráfico. A equação pode ser escrita na forma

$$x\log x = 1 \quad \Rightarrow \quad x = \frac{1}{\log x} \equiv f(x)\,,$$

se $x{\neq}1$. Representemos então a recta $y{=}x$ e a curva $y{=}1/\log x$ gráficamente na Fig. 2.1. É fácil verificar que as duas curvas apenas se cruzam uma vez num ponto que está necessariamente no intervalo $]1,+\infty[$. Encontramos assim a raiz pretendida e conseguimos localizá-la.
No entanto a equação também podia ser escrita na forma

$$f_1(x) \equiv \log x = \frac{1}{x} \equiv f_2(x)\,,$$

pelo que neste caso se representamos as duas curvas $y{=}\log x$ e $y{=}1/x$ (ver a Fig. 2.2) facilmente verificamos, mais uma vez, que estas apenas se cruzam uma única vez num ponto que está necessariamente no intervalo $]1,+\infty[$.

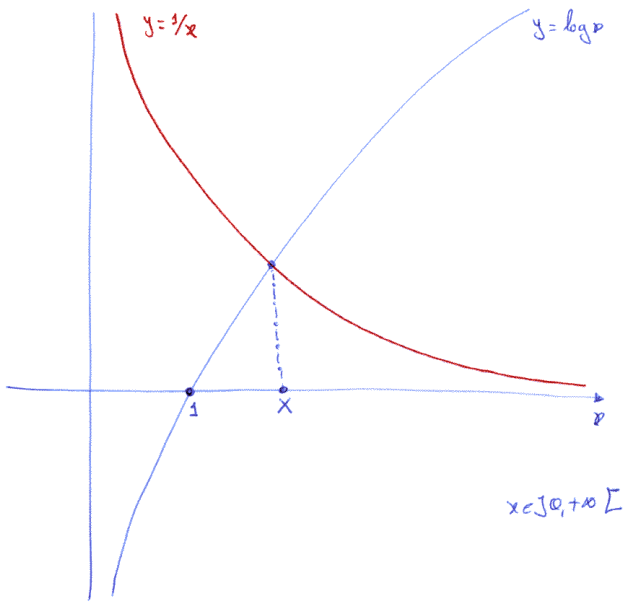


Figura 2.2: Representação das curvas $y{=}f_1(x){\equiv}\log x$ e $y{=}f_2(x){\equiv}1/x$. A intersecção entre elas corresponde aos zeros da equação $F(x){=}0$, e ocorre no intervalo $]1,+\infty[$.

Exemplo 2.1.4: Consideremos também a função $F(x){=}e^x{-}(x{+}2)$, definida para a variável real $x{\in}]-\infty,+\infty[$, e vamos tentar localizar todos os zeros da equação $F(x){=}0$ pelo método gráfico. A equação pode ser escrita na forma

$$e^x = x+2 \quad \Rightarrow \quad f_1(x) \equiv e^x = x+2 \equiv f_2(x)\,.$$

É fácil verificar que as duas curvas (uma exponencial e uma recta) apenas se cruzam duas vezes. Uma delas à esquerda do zero, e a outra à direita, pois para $x{=}0$ temos que

$$f_2(0) = 2 \quad > \quad 1 = f_1(0)\,,$$

enquanto que

$$\lim_{x\to-\infty} f_2(x) = -\infty \quad < \quad 0 = \lim_{x\to-\infty} f_1\,,$$

e

$$\lim_{x\to+\infty} \frac{f_1(x)}{f_2(x)} = +\infty \Rightarrow \qquad \lim_{x\to+\infty} f_2(x) \quad < \quad \lim_{x\to+\infty} f_1(x)\,.$$

2.2 Método das bissecções sucessivas

Este é um método iterativo que nos permite localizar a raiz de uma equação $F(x){=}0$ caso seja conhecido um intervalo onde esteja essa raiz (e apenas essa). Assim, seja;

- $F(x)$ uma função contínua em $[a,b]$,
- X um zero de $F(x)$ tal que $X{\in}[a,b]$.

Então podemos usar o seguinte algoritmo para encontrar a raiz, reduzindo o intervalo inicial a um intervalo de largura (2ε):

(*i*) $F_a{\equiv}F(a)$ e $F_b{=}F(b)$ tal que $F_a \cdot F_b \le 0$

(*ii*) para $c{=}\frac{a+b}{2}$ e $F_c{=}F(c)$:

→ se $F_a \cdot F_c \le 0$ então $b{=}c$ e $F_b{=}F_c$

→ se $F_a \cdot F_c > 0$ então $a{=}c$ e $F_a{=}F_c$

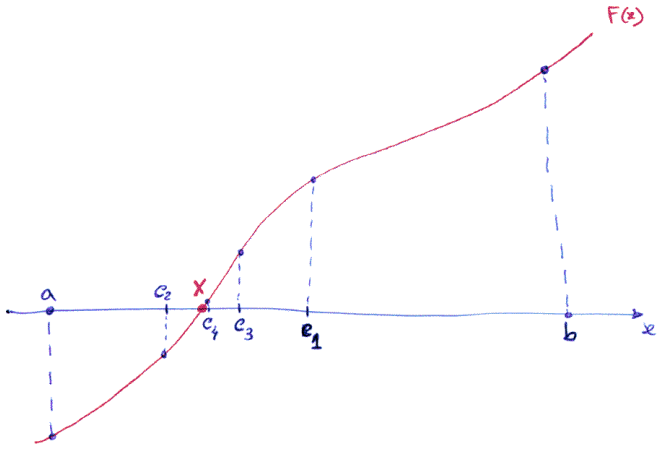


Figura 2.3: Sucessão de iterações feitas pelo método das bissecções sucessivas de forma a encontrar a raix X da equação $F(x)=0$, partindo-se do intervalo inicial $[a,b]$.

(*iii*) volta ao início (*i*) enquanto $\frac{b-a}{2} \geq \varepsilon$

O número de iterações necessárias para obter a precisão ε pretendida pode ser facilmente determinado, visto a dimensão do intervalo ser dado por:

$$\text{após } n \text{ iterações} \quad \Rightarrow \quad \text{amplitude do intervalo} = \frac{b-a}{2^n} \,. \tag{2.2.1}$$

Logo, temos que o erro (semi-amplitude do intervalo) é dado por

$$\frac{b-a}{2^{n+1}} \leq \varepsilon \quad \Rightarrow \quad n \geq \frac{\log(b-a) - \log\varepsilon}{\log 2} - 1 \,. \tag{2.2.2}$$

Exemplo 2.2.1: Consideremos a equação $F(x) \equiv x\log x - 1 = 0$, cuja raiz sabemos estar no intervalo $]1,\mathrm{e}[$, pois

$$\begin{aligned} a = 1 \quad &\Rightarrow \quad F_a \equiv F(1) = -1 < 0 \,, \\ b = \mathrm{e} \quad &\Rightarrow \quad F_b \equiv F(\mathrm{e}) = \mathrm{e} - 1 > 0 \,. \end{aligned}$$

Logo, $F_a \cdot F_b \leq 0$, podendo-se concluir que a raiz está no intervalo dado visto F ser contínua em $[1,\mathrm{e}]$. Vamos então calculá-la com uma precisão inferior a 0.1:

n	a	Sinal F_a	b	Sinal F_b	c	Sinal F_c	$\frac{b-a}{2}$
0	1.000	-	2.718	+	1.859	+	0.86
1	1.000	-	1.859	+	1.430	-	0.43
2	1.430	-	1.859	+	1.644	-	0.22
3	1.644	-	1.859	+	1.752	-	0.11
4	1.752	-	1.859	+	1.805		0.06

Temos assim que a raiz é dada por $X = 1.8 \pm 0.1$ Podemos ainda determinar se podiamos saber a priori o número mínimo de iterações que seria necessário calcular. Recorrendo à equação (2.2.2), encontramos que $n \geq 4$, valor este que é consistente com os cálculos indicados na tabela.

2.3 Método iterativo simples

Consideremos mais uma vez uma equação do tipo $F(x)=0$ para o qual queremos determinar uma raiz X num intervalo $[a,b]$.

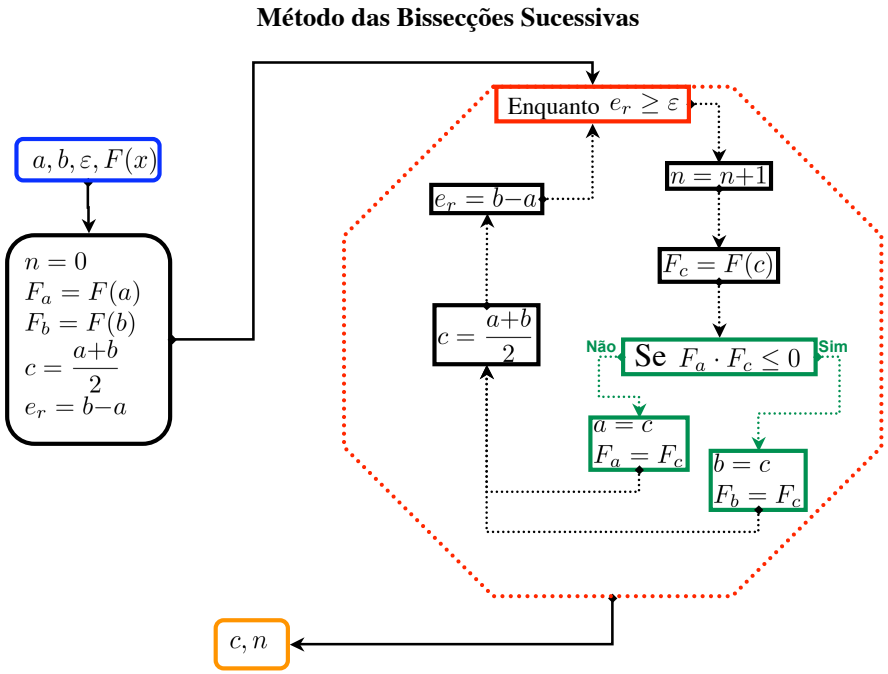


Figura 2.4: Algoritmo para implementação do método das bissecções sucessivas de forma a encontrar a raix X da equação $F(x)$=0, partindo-se do intervalo inicial $[a,b]$.

2.3.1 Condições de aplicabilidade

Comecemos por escrever a equação na forma

$$x = f(x)\ . \tag{2.3.1}$$

Temos então que a sucessão definida de acordo com,

$$\begin{cases} x_0 \in [a,b] \\ x_{n+1} = f(x_n) \qquad \text{para} \quad \text{n=0,1,2,...} \end{cases} \tag{2.3.2}$$

converge para a raiz X, se e só se,

$$\begin{array}{ll} (i) & f(x) \quad \text{é continua em} \quad [a,b] \\ (ii) & f([a,b]) \subseteq [a,b] \\ (iii) & |f(x_s)-f(x_t)| \leq L\,|x_s-x_t|, \quad \forall x_s, x_t \in [a,b] \quad \text{e com} \quad 0 \leq L < 1. \end{array} \tag{2.3.3}$$

Estas condições de aplicabilidade são tais que;

- (i) e (ii) garantem a existência da raiz no interval $[a,b]$
- (iii) garante a unicidade da raiz e a convergência de x_n para essa raiz X.

Suponhamos por exemplo que existem duas raízes α e β da equação $F(x)$=0 no intervalo $[a,b]$, com as condições acima $(i$-$iii)$ sendo válidas. Então

$$\begin{array}{rcll} \alpha & = & f(\alpha) & \\ \beta & = & f(\beta) & \Rightarrow |f(\alpha)-f(\beta)| = |\alpha-\beta|\ , \end{array} \tag{2.3.4}$$

pelo que usando a condição (iii) obtemos que

$$|\alpha-\beta| \leq L\,|\alpha-\beta| \quad \Rightarrow \quad L \geq 1\ , \tag{2.3.5}$$

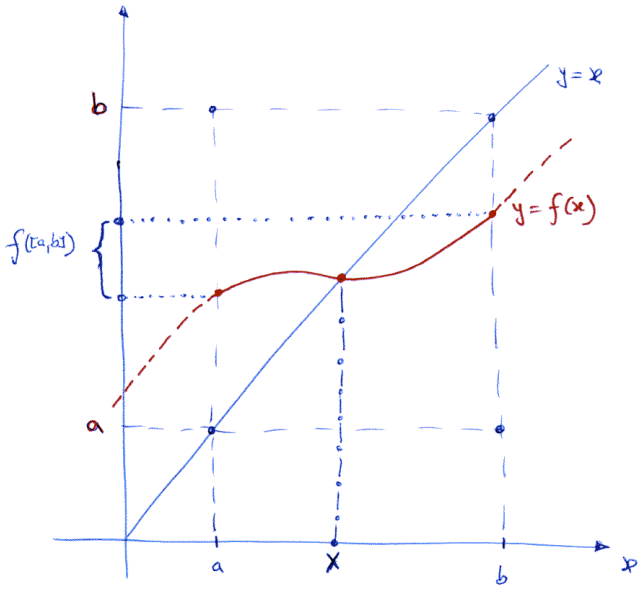


Figura 2.5: Representação gráfica da condição de aplicabilidade dada em (2.3.3ii). Pode ser verificado pelo gráfico que para $f([a,b])\subseteq[a,b]$, e sendo $f(x)$ contínua no intervalo, então a equação tem necessariamente uma raiz no intervalo $[a,b]$.

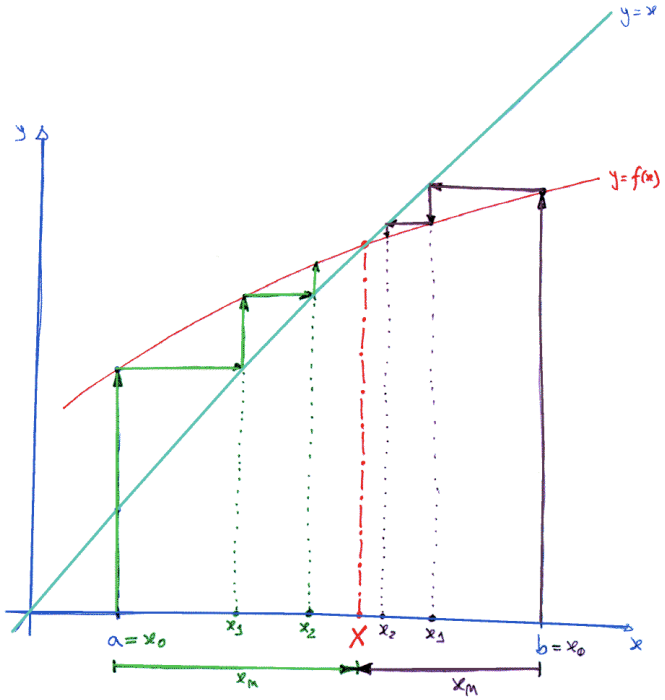


Figura 2.6: Sucessão de iterações feitas pelo método iterativo simples de forma a encontrar a raix X da equação $F(x)=0$ escrita, na forma $x=f(x)$, partindo-se do intervalo inicial $[a,b]$ e com $x_0=b$. Neste caso a derivada de $f(x)$ é positiva: $0<f'(x)<1$.

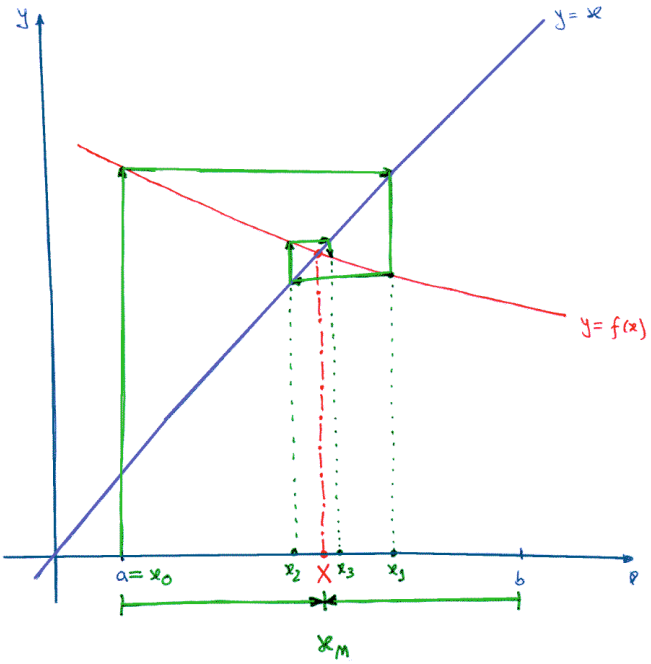


Figura 2.7: Sucessão de iterações feitas pelo método iterativo simples de forma a encontrar a raix X da equação $F(x)=0$, escrita na forma $x=f(x)$, partindo-se do intervalo inicial $[a,b]$ e com $x_0=b$. Neste caso a derivada de $f(x)$ é negativa: $-1<f'(x)<0$.

o que é impossível. Logo apenas pode existir uma raiz no intervalo $[a,b]$.

Se usarmos o teorema do valor médio podemos escrever que

$$\begin{aligned} f(x_s)-f(x_t) &= f(x_t+x_s-x_t)-f(x_t) \\ &= f(x_t)+f'(\xi)(x_s-x_t)-f(x_t) \\ &= f'(\xi)\cdot(x_s-x_t)\,, \end{aligned} \tag{2.3.6}$$

onde $\xi\in[a,b]$. Logo

$$|f(x_s)-f(x_t)| = |f'(\xi)|\cdot|x_s-x_t| \le L\,|x_s-x_t|\,, \tag{2.3.7}$$

se

$$L = Max_{\xi\in[a,b]}|f'(\xi)| < 1\,. \tag{2.3.8}$$

2.3.2 Convergência e expressão para o erro do termo de ordem n

Vejamos então se a sucessão de termos x_n converge para a raiz X. Usando (iii) temos que

$$\begin{aligned} |x_n-X| &= |f(x_{n-1})-f(X)| \le L\,|x_{n-1}-X| \\ &\le L\,|f(x_{n-2})-f(X)| \le L^2|x_{n-2}-X| \\ &\le \ldots \\ &\le L^n|x_0-X| \end{aligned} \tag{2.3.9}$$

$$\Rightarrow |x_n-X| \le L^n|b-a|\,. \tag{2.3.10}$$

Como $L<1$ temos então que $\lim_{n\to\infty}|x_n-X|=0$, pois $\lim_{n\to\infty}L^n=0$, logo fica demonstrada a convergência da sucessão de termos para a raiz; $\lim_{n\to\infty}x_n=X$.

Usando esta expressão podemos ainda obter a expressão que nos permite obter o número de termos que é necessário calcular para termos a raiz com uma precisão ε:

$$L^n|b-a|\le\varepsilon \qquad \Rightarrow \qquad n \ge \frac{\log\left(\frac{\varepsilon}{b-a}\right)}{\log L}\,. \tag{2.3.11}$$

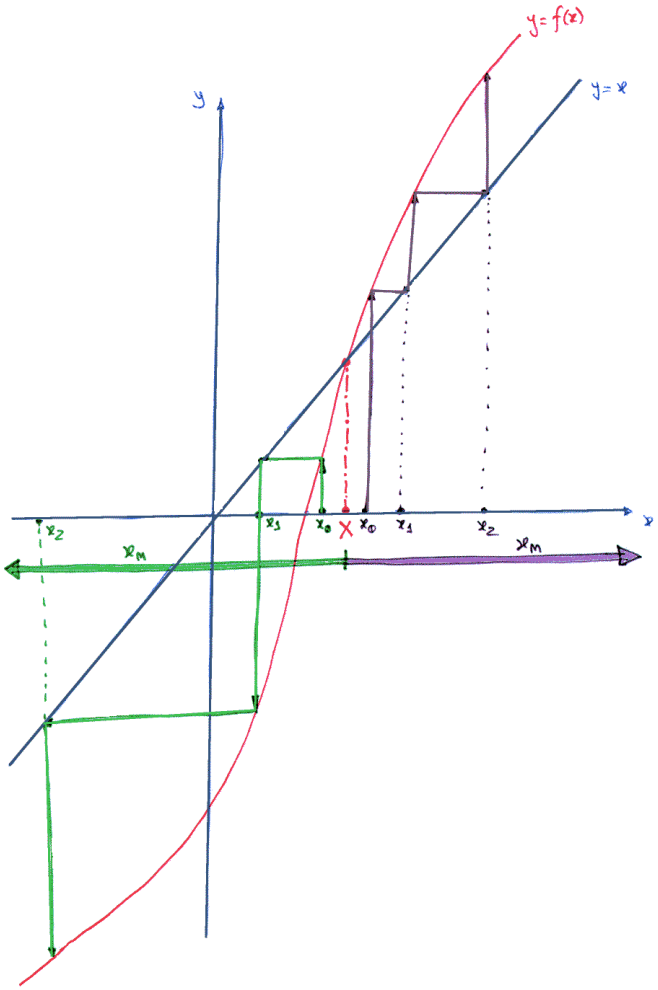


Figura 2.8: Sucessão de iterações feitas pelo método iterativo simples para o caso em que as condições de aplicabilidade não são verificadas, pois $f'(x)>1$, pelo que a sucessão de termos x_n diverge.

No caso de definirmos uma função $f(x)$ tal que $f'(x)>1$ no intervalo $[a,b]$, então a sucessão não converge para a raiz (ver Fig. 2.8). O mesmo acontece quando $f'(x)<-1$.

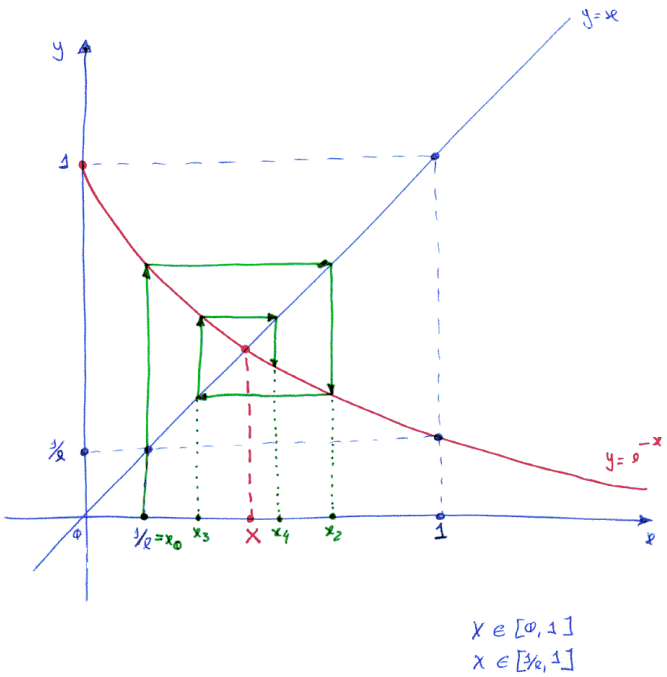


Figura 2.9: Localização da raiz da equação $e^{-x}-x=0$ no intervalo $[0,1]$ (ou ainda no intervalo $[1/e,1]$). Representa-se também os termos inciais da sucessão de iterações feitas pelo método iterativo simples.

Exemplo 2.3.1: Calculemos a raiz da equação $F(x)\equiv e^{-x}-x=0$. Começamos por escrever a equação na forma $x=e^{-x}\equiv f(x)$, pelo que queremos encontrar X tal que $X=f(X)$. Pelo método gráfico (ver Fig. 2.2) podemos ver que a raiz está no intervalo $[0,1]$. Vejamos então se a função $f(x)$ satisfaz as condições de aplicabilidade do método iterativo simples:

$$\begin{array}{ll} (i) & e^{-x} \quad \text{é contínua em} \quad [0,1]\,, \\ (ii) & f([0,1])=[1/e,1]\subseteq[0,1]\,, \\ (iii) & |f'(x)|=e^{-x} \quad \text{não é ``}<1\text{''}\,. \end{array}$$

A ultima condição não é verificada pois $|f'(0)|=1$, não sendo portanto estritamente menor que 1. A forma de contornar o problema é reduzir o intervalo, retirando o ponto 0 (onde o módulo da derivada é 1). Consideremos então o intervalo $[1/e,1]$ que ainda contém a raiz, pois $F(1/e)=e^{-1/e}-1/e>0$ e $F(1)=1/e-1<0$, e vejamos se as condições de aplicabilidade são verificadas;

$$\begin{array}{ll} (i) & e^{-x} \quad \text{é continua em} \quad [1/e,1]\,, \\ (ii) & f([1/e,1])=[1/e,e^{-1/e}]\subseteq[1/e,1]\,, \\ (iii) & |f'(x)|=e^{-x}\leq e^{-1/e}\equiv L<1\,. \end{array}$$

Então o método converge para a raiz. Seja $x_0=1/e$, o primeiro termo da sucessão:

$$\begin{aligned} x_1 &= f(x_0)=0.6922 \\ x_2 &= f(x_1)=0.5005 \\ x_3 &= f(x_2)=0.6062 \\ &\ldots \\ x_{18} &= f(x_{17})=0.5671 \end{aligned}$$

O erro associado ao termo da sucessão de ordem $n=18$, recorrendo a (2.3.10), é dado por

$$|x_{18}-X|\leq e^{-18/e}\,|1-1/e|\leq 9\times10^{-4}\,.$$

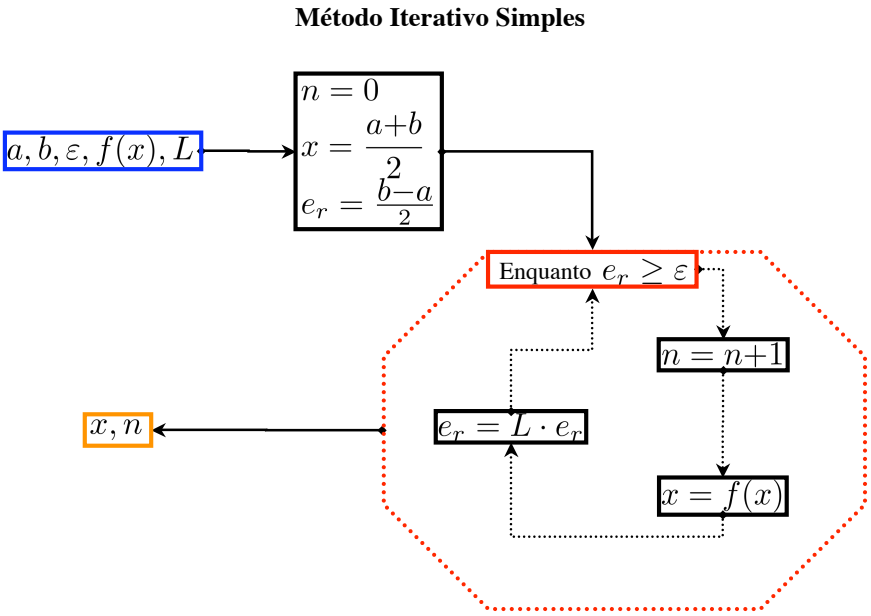


Figura 2.10: Algoritmo para implementação do método iterativo simples de forma a encontrar a raix X da equação $X=f(X)$ com precisão ε, partindo-se do intervalo inicial $[a,b]$.

2.3.3 Ordem de convergência

Diz-se que uma sucessão x_n converge linearmente para X, se existir um q, com $|q|<1$, tal que

$$q = \lim_{n\to\infty} \frac{X-x_{n+1}}{X-x_n} \,. \tag{2.3.12}$$

No caso do método iterativo simples, onde $x_n=f(x_{n-1})$ em que $f(x)$ obedece às condições de aplicabilidade do método, e com $\delta x_n=X-x_n$, temos que

$$\begin{aligned} X-x_n &= f(X)-f(x_{n-1}) = f(X)-f(X-\delta x_{n-1}) \\ &= f(X)-\left[f(X)-f'(\xi)\cdot\delta x_{n-1}\right] = f'(\xi)\cdot\delta x_{n-1} \,, \end{aligned} \tag{2.3.13}$$

com $\xi\in[a,b]$. Pelo que

$$q = \lim_{n\to\infty} \frac{X-x_n}{X-x_{n-1}} = f'(\xi) \,, \tag{2.3.14}$$

de onde resulta que

$$|q| = \left|f'(\xi)\right| \leq L < 1 \,, \tag{2.3.15}$$

pelo que o método iterativo simples é um método com ordem de convergência 1 (linear).

No entanto para uma função em que

$$f'(X) = f''(X) = \ldots = f^{(k-1)}(X) = 0 \,, \tag{2.3.16}$$

teremos uma ordem de convergência diferente, pois

$$\begin{aligned} X-x_n &= f(X)-f(x_{n-1}) = f(X)-f(X-\delta x_{n-1}) \\ &= f(X)-\left[f(X)+\sum_{i=1}^{k-1}\frac{f^{(i)}(X)}{i!}\left(-\delta x_{n-1}\right)^i+\frac{f^{(k)}(\xi)}{k!}\left(-\delta x_{n-1}\right)^k\right] \\ &= -\frac{f^{(k)}(\xi)}{k!}\left(-\delta x_{n-1}\right)^k \,. \end{aligned} \tag{2.3.17}$$

Neste caso, diz-se que a convergência é de ordem K se

$$\lim_{n\to\infty} \frac{X-x_{n+1}}{(X-x_n)^k} = q\,, \tag{2.3.18}$$

com $q\neq 0$. Ou seja, tal que

$$\lim_{n\to\infty} \frac{X-x_{n+1}}{(X-x_n)^k} = -\frac{(-1)^k}{k!} f^{(k)}(\xi) \neq 0\,. \tag{2.3.19}$$

Tal, sugere-nos que uma forma de melhorar a convergência de um método é por exemplo levar a que $f'(X)$ seja 0. Vejamos por exemplo o caso de definirmos a seguinte sucessão para uma equação do tipo $F(x)=0$;

$$x_{n+1} = x_n + \alpha\, F(x_n) \equiv f(x_n)\,. \tag{2.3.20}$$

Para ser convergente, já vimos que é necessário ter que

$$\left|f'(x)\right| < 1 \qquad \forall x \in [a,b]\,, \tag{2.3.21}$$

o que equivale a exigir que $|1+\alpha F'(x)|<1$. Mas se exigirmos que $f'(X)=0$ então o método converge quadráticamnete, pelo menos, pelo que melhoramos significativamente a forma de convergência da sucessão para a raiz X.

Para tal, basta considerar que

$$\alpha = -\frac{1}{F'(X)}\,. \tag{2.3.22}$$

Mas visto desconhecermos o valor de $F'(X)$, pois desconhecemos X, isto sugere que usemos então um sucessão do tipo

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)}\,. \tag{2.3.23}$$

Este é um novo método de calcular a raiz, cuja convergência é de ordem (quadrática), superior ao método iterativo simples.

2.4 Método iterativo de Newton

Vamos então calcular a raiz, contida no intervalo $[a,b]$, da equação $F(x)=0$ através de uma sucessão x_n que seja convergente para a raiz X.

Seja então $x_0 \in [a,b]$, e consideremos a seguinte fórmula de recorrência:

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} \qquad \text{para} \quad n = 0,1,2,... \tag{2.4.1}$$

Note que x_1 é o ponto de intersecção da recta tangente a $F(x)$ no ponto x_0 com o eixo dos x's (ver Fig. 2.11). É desta forma que a fórmula de recorrência tenta obter termos da sucessão cada vez mais perto da raiz X.

2.4.1 Condições de aplicabilidade

Seja então $F(x)=0$ uma equação com uma raiz $X \in [a,b]$. Se

- F, F' e F'' são contínuas em $[a,b]$,

e tal que

$$\begin{aligned}
&(i) \quad && F(a)\cdot F(b) \leq 0 \\
&(ii) \quad && F'(x) \neq 0, \quad \forall x \in [a,b] \\
&(iii) \quad && F''(x) \geq 0 \;\; \dot\vee \;\; F''(x) \leq 0, \quad \forall x \in [a,b] \\
&(iv) \quad && \text{e em alternativa;} \\
& && \rightarrow \quad x_0 \text{ é o extremo de } [a,b], \text{ onde } |F'| \text{ tem o menor valor, e tal que} \\
& && \qquad \left|\frac{F(x_0)}{F'(x_0)}\right| \leq b-a \\
& && \rightarrow \quad x_0 \text{ é o extremo de } [a,b], \text{ onde} \\
& && \qquad F(x_0)\cdot F''(x_0) > 0\,.
\end{aligned} \tag{2.4.2}$$

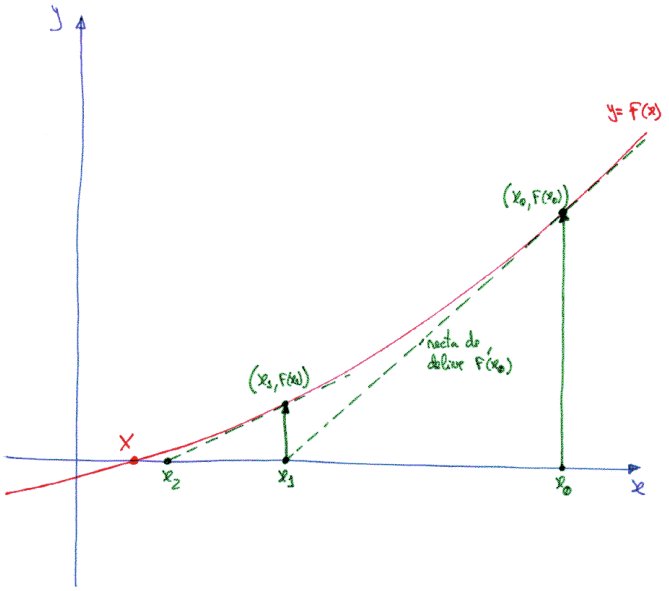


Figura 2.11: Representação de como o ponto x_1, obtido pela fórmula de recorrência dada pela expressão (2.4.1) a partir de x_0. Este valor, assim obtido, corresponde à intersecção da recta $y{=}F'(x_0)\cdot(x_1{-}x_0)+F(x_0)$, com o eixo dos x's ($y{=}0$).

Então a sucessão de termos

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} \qquad n = 0,1,2,\ldots \tag{2.4.3}$$

é tal que

$$\lim_{n\to\infty} x_n = X\ . \tag{2.4.4}$$

Pode ser facilmente verificado que;

(i)+ F contínua	$\Rightarrow$ existe uma raiz em $[a,b]$
(i)+(ii)+ F' contínua	$\Rightarrow$ existe uma só raiz em $[a,b]$
(iii)+ F'' contínua	$\Rightarrow$ não existem pontos de inflexão
(iv)	$\Rightarrow x_1 \in [a,b]$.

2.4.2 Convergência e expressão para o erro do termo de ordem n

Vejamos então a convergência da sucessão para a raiz X. Seja

- m - um minorante de $|F'(x)|$ para $x \in [a,b]$,
- M - um majorante de $|F''(x)|$ para $x \in [a,b]$,

então

$$X - x_{n+1} = X - x_n + \frac{F(x_n)}{F'(x_n)} \qquad \Rightarrow \qquad \delta x_{n+1} = \delta x_n + \frac{F(x_n)}{F'(x_n)}\ . \tag{2.4.5}$$

Como

$$F(X) = F(x_n+\delta x_n) = F(x_n) + F'(x_n)\cdot\delta x_n + \frac{F''(\xi_n)}{2}\,(\delta x_n)^2\ , \tag{2.4.6}$$

com $\xi_n{\in}[a,b]$, e relembrando que $F(X){=}0$, temos então, após substituir δx_n de (2.4.5) em (2.4.6), que

$$\delta x_{n+1} = -\frac{1}{2}\frac{F''(\xi_n)}{F'(x_n)}(\delta x_n)^2\ . \tag{2.4.7}$$

Daqui podemos tirar que

$$|\delta x_{n+1}| = \frac{|F''(\xi_n)|}{2|F'(x_n)|}\,|\delta x_n|^2 \le \frac{M}{2m}(\delta x_n)^2 \equiv \alpha\,|\delta x_n|^2\ , \tag{2.4.8}$$

onde definimos que

$$\alpha \equiv \frac{M}{2m}\,, \qquad \text{com} \qquad M \equiv \max_{\xi\in[a,b]} \left|F''(\xi)\right| \qquad \text{e} \qquad m \equiv \min_{\xi\in[a,b]} \left|F'(\xi)\right|\,. \tag{2.4.9}$$

Usando esta relação obtemos que

$$\begin{aligned} |\delta x_1| &\leq \alpha|\delta x_0|^2 \\ |\delta x_2| &\leq \alpha|\delta x_1|^2 \leq \alpha^3|\delta x_0|^4 \\ |\delta x_3| &\leq \alpha|\delta x_2|^2 \leq \alpha^7|\delta x_0|^8 \\ &\ldots \\ |\delta x_k| &\leq \alpha^{2^k-1}|\delta x_0|^{2^k} \end{aligned} \tag{2.4.10}$$

$$|\delta x_k| \leq \alpha^{2^k-1}(b-a)^{2^k}\,, \tag{2.4.11}$$

onde usamos o facto de $|\delta x_0| \leq b-a$. Temos assim, finalmente, que

$$\lim_{k\to\infty} |\delta x_k| \leq \frac{1}{\alpha} \lim_{k\to\infty} |\alpha\,(b-a)|^{2^k} = 0\,, \tag{2.4.12}$$

desde que $(b-a)<1/\alpha$. É sempre possivel encontrar $[a,b]$ tal que tenhamos $(b-a)<1/\alpha$, logo a sucessão converge para X.

Se quisermos determinar o número de iterações que é necessário fazer para obter o resultado com uma precisão ε, então basta usarmos a relação (2.4.11) para termos que

$$\Delta x_k = |\delta x_k| \leq \varepsilon \Rightarrow \alpha^{2^k-1}(b-a)^{2^k} \leq \varepsilon\,, \tag{2.4.13}$$

de onde tiramos que

$$k \geq \frac{1}{\log 2}\,\log\left\{\frac{\log(\alpha\varepsilon)}{\log[\alpha(b-a)]}\right\}\,, \tag{2.4.14}$$

que corresponde ao número mínimo de termos que é necessário calcular para obtermos a raiz com a precisão pedida (ε).

Exemplo 2.4.1: Pretende-se calcular o valor de $\sqrt{c}$ (para $c>0$) usando o método de Newton. Para tal temos que a equação a resolver é $F(x)\equiv x^2-c=0$. Neste caso a fórmula de recorrência é dada por

$$x_{n+1} = x_n - \frac{x_n^2-c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right) \qquad n = 0,1,2,\ldots$$

e com $x_0\neq 0$. Todas as condições de aplicabilidade são verificadas pois $F(x)$, $F'(x)$ e $F''(x)$ são contínuas em $\mathscr{R}$. Além disso a primeira e segunda derivadas são sempre positivas e x_0 pode ser escolhido como sendo c. Calculemos então o valor de $\sqrt{2}$:

n:	0	1	2	3	4	...
x_n:	2.0000	1.5000	1.4167	1.4142	1.4142	...
Δx_n:	1.0000	0.5000	0.1250	0.0078	0.00004	...

Para o cálculo de Δx_n usamos o facto de no intervalo $[1,2]$ termos que $m=M=2$, dando que $\alpha=0.5$.

Exemplo 2.4.2: O mesmo princípio pode ser também utilizado para calcular raízes de ordem m de um número real: $c^{1/m}$. Neste caso a sucessão é escrita para a função $F(x)=x^m-c$, sendo dada por

$$x_{n+1} = x_n - \frac{x_n^m-c}{m x_n^{m-1}} = \left(1-\frac{1}{m}\right)x_n + \frac{c}{m}\,x_n^{1-m} \qquad n = 0,1,2,\ldots$$

que é tal que $\lim_{n\to\infty} x_n = c^{1/m}$.

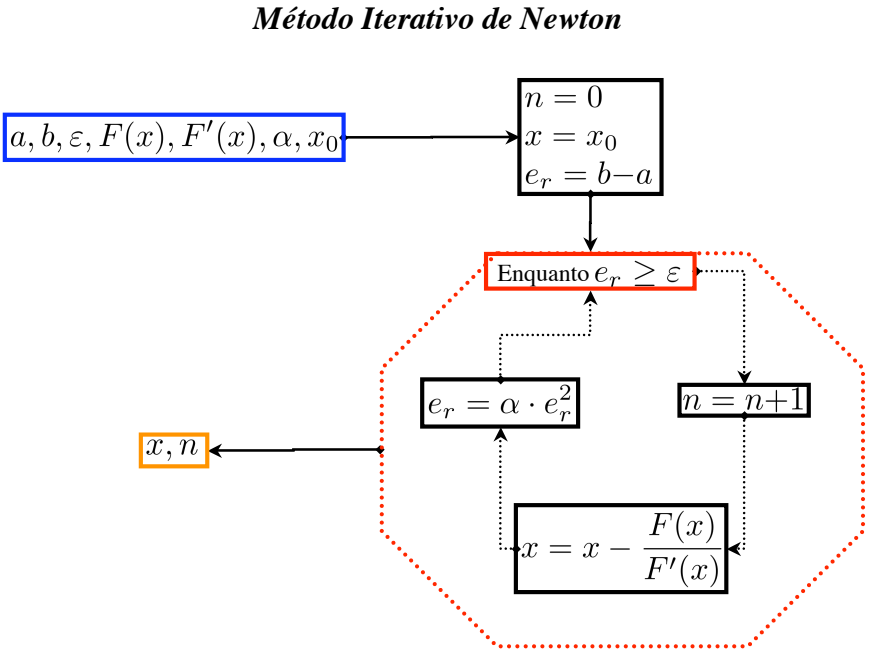


Figura 2.12: Algoritmo para implementação do método iterativo de Newton de forma a encontrar a raix X da equação $F(X)$=0 com precisão ε, partindo-se do intervalo inicial $[a,b]$.

2.4.3 Algoritmo de Horner

O algoritmo de Horner é uma forma de reduzir as operações necessárias para calcular o valor de um polinómio e das suas derivadas, num ponto. Este é construído de forma a evitar calcular as potências reduzindo o número de multiplicações feitas, ao mínimo.

Seja então $P(x)$ um polinómio de grau N;

$$P(x) = a_0 x^N + a_1 x^{N-1} + ... + a_{N-1} x + a_N \, . \tag{2.4.15}$$

Considermos a seguinte sucessão para calcular o valor de $P(\alpha)$:

$$\begin{aligned} b_0 &= a_0 \\ b_n &= \alpha \, b_{n-1} + a_n, \quad \text{para} \quad n = 1,2,...,N \, , \end{aligned} \tag{2.4.16}$$

então $P(\alpha)$=b_N. Para o cálculo da derivada $P'(\alpha)$ basta considerar que

$$\begin{aligned} c_0 &= b_0 \\ c_n &= \alpha \, c_{n-1} + b_n, \quad \text{para} \quad n = 1,2,...,N-1 \, , \end{aligned} \tag{2.4.17}$$

para se encontrar que $P'(\alpha)$=c_{N-1}.

Isto é, esquemáticamente, temos

$$\begin{array}{ccccccccccc} a_0 & & a_1 & & a_2 & & ... & & a_{N-2} & & a_{N-1} & & a_N \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ b_0 = a_0 & \rightarrow & b_1 & \rightarrow & b_2 & \rightarrow & ... & \rightarrow & b_{N-2} & \rightarrow & b_{n-1} & \rightarrow & b_N \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ c_0 = b_0 & \rightarrow & c_1 & \rightarrow & c_2 & \rightarrow & ... & \rightarrow & c_{N-2} & \rightarrow & c_{N-1} & & \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & & & \\ d_0 = c_0 & \rightarrow & d_1 & \rightarrow & d_2 & \rightarrow & ... & \rightarrow & d_{N-2} & & & & \end{array} \tag{2.4.18}$$

com

$$P(\alpha) = b_N\ , \qquad P'(\alpha) = c_{N-1} \qquad \text{e} \qquad \frac{P''(\alpha)}{2!} = d_{N-2}\ . \tag{2.4.19}$$

Sendo semelhante o cálculo das restantes derivadas do polinómio $P(x)$.

Exemplo 2.4.3: Calculemos o valor de $P(x)$, e suas derivadas, em x=1/2, onde

$$P(x) = 7x^4 + 5x^3 - 2x^2 + 8$$

Usando o esquema acima temos que

7	5	-2	0	8
7	8.5	2.25	1.125	8.5625
7	12	8.25	5.25	
7	15.5	16		
7	19 7			

Logo

$$\begin{aligned} P(1/2) &= 8.5625 \\ P^{(1)}(1/2) &= 5.25 \\ P^{(2)}(1/2) &= 2!\cdot 16 = 32 \\ P^{(3)}(1/2) &= 3!\cdot 19 = 114 \\ P^{(4)}(1/2) &= 4!\cdot 7 = 168 \end{aligned}$$

Ou seja,

$$P(x) = 7(x-1/2)^4 + 19(x-1/2)^3 + 16(x-1/2)^2 + 5.25(x-1/2) + 8.5625\ .$$

Exemplo 2.4.4: Consideremos então um exemplo de como o algoritmo de Horner pode ser usado para o cálculo de zeros de polinómios, através do uso de método de Newton. Seja $P(x)$ um polinómio dado por;

$$P(x) = x^3 - x^2 + 2x + 5\ ,$$

para o qual sabemos existir uma raiz perto de x_0=−1, tal que as condições de aplicabilidade do método iterativo de Newton são verificadas. Calculemos o valor da raiz;

n	x_n	Algoritmo de Horner: $\underline{P(x_n)}$ e $\underline{P'(x_n)}$				$\frac{P(x_n)}{P'(x_n)}$
0	−1	1	−1	2	5	
		1	−2	4	$\underline{1}$	$\frac{1}{7}$
		1	−3	$\underline{7}$		
1	−1.14286	1	−1	2	5	
		1	−2.14286	4.44898	$\underline{-0.08455}$	$-\frac{0.08455}{8.20408}$
		1	−3.28571	$\underline{8.20408}$		
2	−1.13255	1	−1	2	5	
		1	−2.13255	4.41522	$\underline{-0.00047}$	$-\frac{0.00047}{8.11312}$
		1	−3.26510	$\underline{8.11312}$		
3	−1.13249					

Obtemos assim que $X \simeq -1.1325$.

2.4.4 Variantes: declive fixo, secante e falsa posição

Por vezes há vantagem em simplificar o método de Newton, evitando por exemplo ter de calcular $F'(x_n)$ para todos os termos. Surgem assim as variantes ao método de Newton. Entre elas temos;

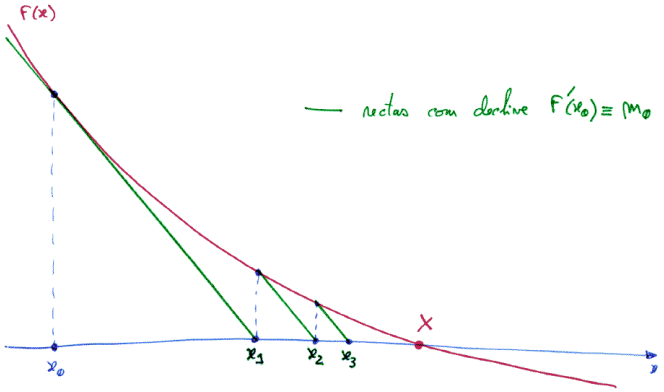


Figura 2.13: Representação do método de declive fixo em que se considera a fórmula de recorrência dada em (2.4.20).

Declive Fixo: neste caso aproximamos o valor de $F'(x_n)$ pelo seu valor para o primeiro termo da sucessão, isto é, por $F'(x_0)$. Temos então a seguinte fórmula de recorrência:

$$x_{n+1} = x_n - \frac{F(x_n)}{m_0}\,, \qquad \text{onde} \quad m_0 \equiv F'(x_0)\,. \tag{2.4.20}$$

Exemplo 2.4.5: Determinemos o valor de $\sqrt{2}$ utilizando o método do declive fixo. Como $F(x)=x^2-2$, temos que;

$$m_0 = F'(2) = 4\,.$$

Assim, a série de termos é dada por,

$$x_{n+1} = x_n - \frac{x_n^2-2}{4}\,,$$

de onde obtemos que

n:	0	1	2	3	4	...
x_n:	2.0000	1.5000	1.4375	1.4209	1.4146	...

Método da Secante: neste caso tenta-se ser mais eficiente, substituindo a derivada não por uma constante, mas por uma aproximação mais próxima do valor real. Para tal escreve-se que

$$F'(x_n) \simeq \frac{F(x_n)-F(x_{n-1})}{x_n-x_{n-1}}\,, \tag{2.4.21}$$

obtendo-se a seguinte fórmula de recorrência;

$$x_{n+1} = x_n - \frac{F(x_n)}{F(x_n)-F(x_{n-1})}\,(x_n-x_{n-1})\,. \tag{2.4.22}$$

Note-se que neste caso é necessário ter x_0 e x_1 para iniciar o cálculo dos termos da sucessão. Mas x_1 pode ser facilmente obtido, por uma das fórmulas de recorrência dadas anteriormente.

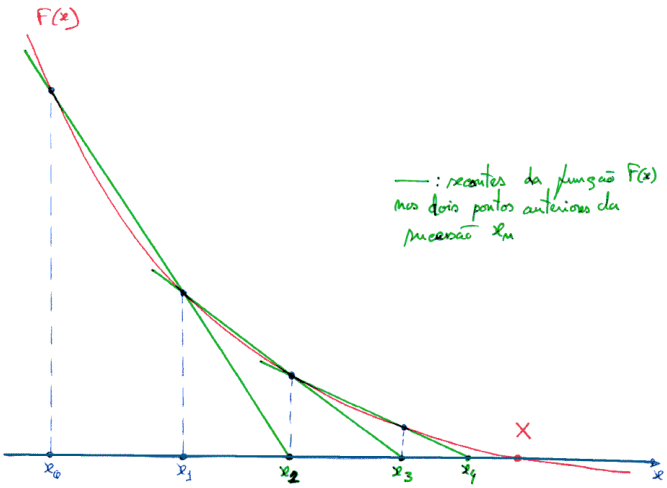


Figura 2.14: Representação do método da secante em que se considera a fórmula de recorrência dada em (2.4.22).

Exemplo 2.4.6: Determinemos o valor de $\sqrt{2}$ utilizando o método da secante. Como $F(x)=x^2-2$, temos que a série de termos é dada por,

$$x_{n+1} = x_n - \frac{x_n^2-2}{x_n+x_{n-1}} ,$$

de onde obtemos, considerando que $x_0=2$ e $x_1=1.5$, que

n:	0	1	2	3	4	...
x_n:	2.0000	1.5000	1.4286	1.4146	1.4142	...

Método da Falsa Posição: neste caso obtamos por substituir a tangente à função $F(x)$ no ponto x_n pela secante relativa a um ponto de referência x_r. Assim, a fórmula de recorrência é dada por

$$x_{n+1} = x_n - \frac{F(x_n)}{F(x_n)-F(x_r)}\,(x_n-x_r) . \tag{2.4.23}$$

Exemplo 2.4.7: Determinemos o valor de $\sqrt{2}$ utilizando o método da false posição. Como $F(x)=x^2-2$, e considerando que $x_r=2$ com $F(x_r)=2$, a série de termos é dada por,

$$x_{n+1} = x_n - \frac{x_n^2-2}{x_n+2} ,$$

de onde obtemos que

n:	0	1	2	3	4	...
x_n:	2.0000	1.5000	1.4286	1.4167	1.4146	...

2.4.5 Resolução de equações dadas por funções implícitas

Consideremos o caso de queremos resolver a seguinte equação implícita;

$$F(x,y) = 0 \qquad \text{que nos dá} \quad y=y(x) . \tag{2.4.24}$$

Usemos o método iterativo de Newton para obter o valor de $y=Y$ para $x=X$. Como,

$$F(X,y) = F[X,y_n+(y-y_n)] = F(X,y_n) + \frac{\partial F}{\partial y}(X,\xi_n)\cdot(y-y_n) , \tag{2.4.25}$$

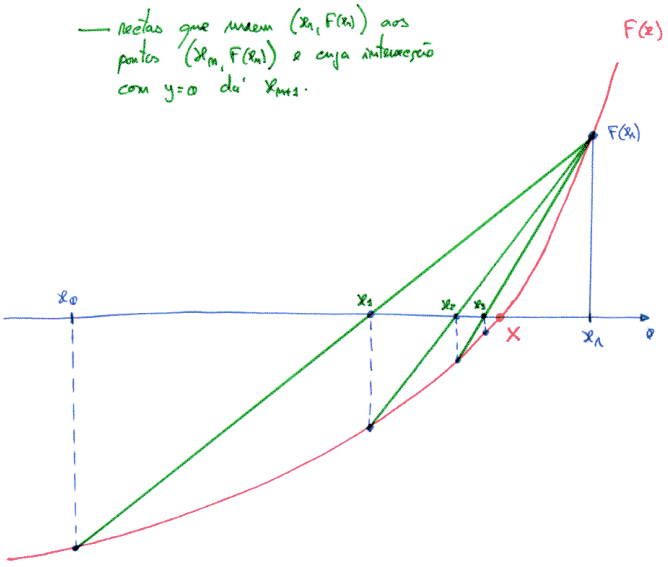


Figura 2.15: Representação do método da falsa posição em que se considera a fórmula de recorrência dada em (2.4.23), com o ponto de referência x_r sendo o outro extremo do intervalo $[a,b]$, que não o termo inicial x_0 da sucessão.

com $\xi_n \in [y_n, y]$, então temos que

$$y \simeq y_n - \frac{F(X,y_n)}{\dfrac{\partial F}{\partial y}(X,y_n)} \,. \tag{2.4.26}$$

Logo podemos definir uma fórmula de recorrência, que converge para $Y = y(X)$, como sendo

$$y_{n+1} = y_n - \frac{F(X,y_n)}{F_y(X,y_n)} \,, \tag{2.4.27}$$

onde

$$F_y(X,y_n) \equiv \frac{\partial F}{\partial y}(X,y_n) \,, \tag{2.4.28}$$

tendo-se então que $\lim\limits_{n\to\infty} y_n = Y$.

2.5 Exercícios

E2.1) Dada a equação $F(x)$=0, tal que F é continua em $[a,b]$ e $F(a)\cdot F(b)\leq 0$, proponha um algoritmo que traduza o cálculo da raíz da equação, com erro não superior a uma tolerância ε, pelo método das bissecções sucessivas. Incluindo também;

a) A determinação do número de iterações a efectuar.

b) Estimando o erro absoluto e relativo em cada etapa.

E2.2) Implemente o algoritmo desenvolvido, e aplique-o na resolução dos seguintes casos de equações:

a) $F(x) \equiv 0.123^x - x = 0$ com $\varepsilon = 10^{-4}$

b) $F(x) \equiv x^3 - 2\,\mathrm{e}^{-x} = 0$ com $\varepsilon = 10^{-5}$

E2.3) Supondo que as condições de aplicabilidade do método iterativo simples para a resolução da equação $F(X)$=0, escrita na forma X=$f(X)$, são satisfeitas no intervalo $[a,b]$, proponha um algoritmo que traduza o cálculo dessa raíz com erro não superior a uma tolerância ε. Incluindo também;

a) A determinação do número de iterações a efectuar.

b) Estimando o erro absoluto e relativo em cada etapa.

E2.4) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = \cos(\pi x) - \frac{3}{2}\pi x e^{2x}$.

a) Encontre as duas maiores raízes negativas da equação $F(X)$=0 e indique intervalos, de amplitude não superior a 0.1, contendo cada uma delas.

b) Escrevendo a equação $F(X)$=0 na forma X=$f(X)$, encontre uma expressão para a função $f(x)$ que lhe permita usar o método iterativo simples para a determinação da raíz $X_o \in [0.1, 0.2]$. Indique os três primeiros termos que resultam da aplicação deste método iterativo ao cálculo desta raiz.

E2.5) Dada a função

$$F(x) = |\log\, x| - \frac{\sin\, x}{5}$$

para x>0, separe as raízes da equação $F(X)$=0. Determine um intervalo de amplitude 0.01 que contenha a menor raíz e usando o algoritmo desenvolvido acima para a implementação do método iterativo simples, encontre o seu valor com um erro inferior a 10^{-7}. Compare o número de iterações feitas com o número que seria necessário fazer caso se usasse o método das bisseções sucessivas.

E2.6) Considere a seguinte equação (para $x\neq 0$); $g(x) - log\, x^2 = 0$, onde $g(x)$ é uma função real.

a) Verifique que para $g(x)$=$x-3/2$ existem três raízes da equação dada. Indique um intervalo de amplitude não superior a 0.1 para cada uma delas.

b) Verifique que para $g(x)$=$x-9/8$ pode usar o método iterativo simples para determinar a raíz contida no intervalo $[3.7, 3.8]$ e calcule-a com um erro inferior a 10^{-2}.

E2.7) Supondo que as condições de aplicabilidade do método de Newton para a resolução da equação $F(X)$=0, são satisfeitas no intervalo $[a,b]$, proponha um algoritmo que traduza o cálculo dessa raíz com erro não superior a uma tolerância ε. Incluindo também;

a) A determinação do número de iterações a efectuar.

b) Estimando o erro absoluto e relativo em cada etapa.

E2.8) Considere a equação

$$2\,\log\, x = x-2$$

a) Separe as raízes reais e encontre um intervalo de amplitude 0.1 que contenha a maior delas.

b) Verifique que é possivel aplicar o método de Newton para calcular essa raíz, e efectue três iterações indicando o erro para cada uma delas.

c) Compare com a precisão obtida pelos métodos das bissecções sucessivas e iterações simples, se apenas três iterações forem igualmente feitas.

E2.9) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = r\ \sin x - e^x$, onde r é um número real.

a) Verifique que para $r \geq 5$ existem pelo menos duas raizes positivas da equação $F(X)=0$. Indique intervalos no caso de $r=6$, de amplitude não superior a 0.1, contendo cada uma delas.

b) Verifique que pode usar o método de Newton para encontrar a raiz contida no intervalo $[-3\pi/2, -\pi]$ para qualquer $r \geq 1$. Indique o primeiro termo da sucessão e escreva a fórmula de recorrência que permitirá obter os restantes termos. Cálcule os três primeiros termos da sucessão se utilizar a fórmula de recorrência para o caso de declive fixo, com $r=1.2$.

E2.10) Caso seja possível calcule as duas maiores raízes da seguinte equação pelo método de Newton;

$$\sin x = \frac{x+1}{x-1}$$

com uma precisão de 2×10^{-6}.

E2.11) Usando o método de Newton, juntamente com o método de Horner para o cálculo de um polinómio e a sua derivada, obtenha a raíz positiva do seguinte polinómio;

$$p(x) = x^3 - x^2 - x - 1$$

com precisão superior a 10^{-3}.

E2.12) Considere a seguinte função definida em $\mathbb{R}$; $F(x) = e^x - (x+2)$.

a) Determine quantas raizes têm a equação $F(X)=0$, e indique intervalos contendo cada uma delas.

b) Mostre que se escrever a equação na forma $X = f(X) = e^X - 2$ não pode utilizar o método iterativo simples para determinar a maior das raízes. Descreva gráficamente a razão deste comportamento.

c) Verifique que pode no entanto usar o método de Newton para encontrar a raiz. Indique o primeiro termo da sucessão e escreva a fórmula de recorrência que permitirá obter os restantes termos. Determine o número de iterações que teria que calcular para obter a raiz com um erro inferior a 10^{-5}.

Capítulo 3

Interpolação numérica

Por vezes apenas conhecemos uma quantidade muito limitada de informação sobre uma função. Para podermos estimar valores aproximados que essa função toma em pontos onde o valor da função não é conhecido, torna-se necessário recorrer a um método de interpolação numérica. Este permite-nos reconstruir localmente a função usando a informação conhecida para simular, através do uso de *funções interpoladoras*, o comportamento da função tabelada para valores onde não é conhecida. Neste capítulo vamos considerar métodos de "reconstrução" da função recorrendo a diferentes expressões e formas de construir a função interpoladora.

3.1 Função interpoladora

Para construir uma função que "interpole" uma tabela de valores, e nos permita estimar outros valores em pontos que não estão tabelados, é necessário recorrer a uma função cujas propriedades são bem conhecidas e para a qual seja fácil calcular o valor em qualquer ponto. No entanto tal função tem claramente que incluir como seus os pontos da tabela inicial, pois é necessário que pelo menos nesses pontos a função original e a função por nós construída coincidam. Quando tal acontece dizemos que estamos a usar uma função interpoladora, relativamente à tabela de pontos dados.

Vamos então supor que temos uma tabela de $n+1$ valores conhecidos, de uma função $f(x)$ cuja expressão é desconhecida: Tabela **??**. Pretende-se então, usando esta tabela de valores, determinar $f(x_r)$, para $x_r \in [a,b]$. O intervalo é definido por a e b, tal que para $\forall i \in \{0,1,2,...,n\}$ se tenha que $x_i \in [a,b]$. É então necessário construir uma função $y(x)$, tal que

$$y(x_i) \equiv f_i \qquad \forall i=0,1,...,n\ , \tag{3.1.1}$$

de forma a aproximarmos o valor de $f(x_r)$ por $y(x_r)$ (ver Fig. 3.1). Ou seja, para considerarmos que

$$f(x_r) \simeq y(x_r)\ . \tag{3.1.2}$$

Tabela 3.1: Informação disponível sobre uma função $f(x)$ que não conhecemos, mas para a qual dispomos de $n+1$ pontos de abcissas x_i.

i	x_i	$f(x_i)$
0	x_0	f_0
1	x_1	f_1
2	x_2	f_2
...	...	...
n	x_n	f_n

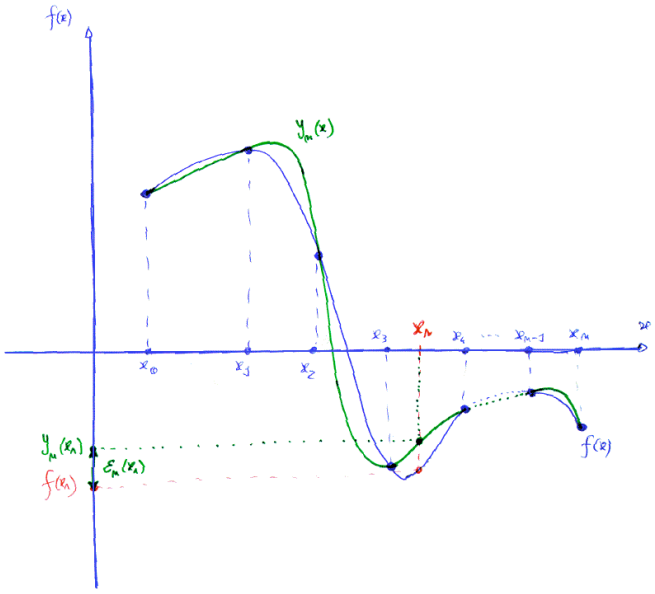


Figura 3.1: Representação de uma função $f(x)$, desconhecida, e para a qual apenas temos os valores que toma nos pontos x_i, com i=0, 1, ..., n (ver Tabela 3.1). Precisamos então de definir uma função interpoladora $y(x)$ que nos permite usar estes pontos para estimar o valor que a função f toma em x_r.

Nestas condições, dizemos que $y(x)$ é a função interpoladora de $f(x)$, relativamente à tabela de n+1 pontos $\{(x_i, f_i)\}_{i=0}^{n}$.

A forma como construimos $y(x)$ pode ser qualquer, desde que possamos garantir que as condições impostas em (3.1.1) são verificadas. Uma das formas mais simples de definir uma função interpoladora que nos permita verificar as condições de interpolação é simplesmente defini-la como sendo uma combinação linear de funções, isto é

$$y(x) \equiv \sum_{j=0}^{n} a_j \, \varphi_j(x) \, , \tag{3.1.3}$$

onde as funções $\varphi(x)$ são linearmente independentes de forma a garantir que a solução é única. Desta forma reduzimos o problema de encontrar a função interpoladora ao cálculo da solução de um sistema linear de equações, que corresponde a encontrar os valores dos coeficientes a_i. Esse sistema é simplesmente obtido impondo as condições (3.1.1);

$$\sum_{j=0}^{n} a_j \, \varphi_j(x_i) = f_i \qquad \text{para} \qquad i = 0, 1, ..., n \, . \tag{3.1.4}$$

As funções de referência $\varphi(x)$ podem ser quaisquer, dependendo das características do problema que estamos a tratar, ou seja, das propriedades da função $f(x)$. Um exemplo que consideramos a seguir corresponde a escolher para $\varphi(x)$ os monómios. Desta forma é possível construir uma função interpoladora que corresponde a um polinómio (combinação linear de monómios de diferente grau).

3.2 Interpolação polinomial

Uma das forma mais fáceis de construir uma função interpoladora, é recorrendo a funções simples como os polinómios, cujas propriedades são bem conhecidas. Assim, para que $P(x)$ seja um polinómio interpolador de $f(x)$ nos n+1 pontos da tabela $\{(x_i, f_i)\}_{i=0}^{n}$ de abcissas distintas, é necessário que este tenha grau n, sendo da forma, usando $\varphi_j(x) \equiv x^j$,

$$P(x) = \sum_{j=0}^{n} a_j \, x^j = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \, , \tag{3.2.1}$$

com os coeficientes a_j ($J = 0,1,...,n$) a serem definidos pelas $n+1$ condições;

$$P(x_i) = f_i \quad ; i = 0,1,...,n \; . \tag{3.2.2}$$

Este conjunto de condições corresponde a $n+1$ equações que nos permitem calcular os coeficientes a_j;

$$\begin{cases} a_0 + x_0 a_1 + x_0^2 a_2 + ... + x_0^n a_n = f_0 \\ a_0 + x_1 a_1 + x_1^2 a_2 + ... + x_1^n a_n = f_1 \\ \qquad\qquad ... \\ a_0 + x_n a_1 + x_n^2 a_2 + ... + x_n^n a_n = f_n \; . \end{cases} \tag{3.2.3}$$

Ou seja,

$$\begin{pmatrix} 1 & x_0 & x_0^2 & ... & x_0^n \\ 1 & x_1 & x_1^2 & ... & x_1^n \\ ... & ... & ... & & ... \\ 1 & x_n & x_n^2 & ... & x_n^n \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ ... \\ f_n \end{pmatrix} \quad \Leftrightarrow \quad \mathscr{D} \cdot \vec{a} = \vec{f} \; . \tag{3.2.4}$$

O sistema tem solução, se e só se, o determinante da matriz $\mathscr{D}$ é não nulo. Tal é verdade, se e só se, todos os pontos $\{x_i\}_{i=0}^n$ são estritamente diferentes, como considerado inicialmente.

O polinómio (3.2.1) pode ser construido de variadas formas, mas existe apenas um polinómio que é solução do sistema de equações dado em (3.2.4).

3.2.1 Polinómio interpolador na fórmula de Lagrange

Na construção do polinómio interpolador pela fórmula de Lagrange usa-se um princípio base que nos permite obter o polinómio como uma combinação linear de polinómios mais simples que tem o mesmo grau que aquele que queremos construir, e tal que os coeficientes da combinação linear sejam os valores tabelados da função.

Comecemos por considerar o caso de uma recta; isto é, queremos construir uma recta que interpole a função $f(x)$ nos pelos pontos (x_0, f_0) e (x_1, f_1). Para tal, consideremos que o polinómio interpolador $y_{0,1}(x)$ (que é uma recta - polinómio de grau 1) é escrito como a combinação linear de duas rectas $\ell_0(x)$ e $\ell_1(x)$ de acordo com

$$y_{0,1}(x) \equiv \ell_0(x) \cdot f_0 + \ell_1(x) \cdot f_1 \; . \tag{3.2.5}$$

Como a recta deve passar pelos pontos dados, teremos que

$$y_{0,1}(x_0) = f_0 \qquad \text{e} \qquad y_{0,1}(x) = f_1 \; , \tag{3.2.6}$$

logo segue que

$$\begin{array}{rcll} \ell_0(x_0) & = & 1 & \ell_0(x_1) = 0 \\ \ell_1(x_0) & = & 0 & \ell_1(x_1) = 1 \; . \end{array} \tag{3.2.7}$$

Com tais condições podemos facilmente obter que

$$\begin{array}{rcll} \ell_0(x_1) & = & 0 & \Rightarrow \quad \ell_0(x) = C_0(x - x_1) \\ \ell_1(x_0) & = & 0 & \Rightarrow \quad \ell_1(x) = C_1(x - x_0) \; , \end{array} \tag{3.2.8}$$

bem como que

$$\begin{array}{rcll} \ell_0(x_0) & = & 1 & \Rightarrow \quad C_0 = \dfrac{1}{x_0 - x_1} \\ \ell_1(x_1) & = & 1 & \Rightarrow \quad C_1 = \dfrac{1}{x_1 - x_0} \; . \end{array} \tag{3.2.9}$$

Daqui resulta que

$$y_{0,1}(x) = f_0 \, \frac{x - x_1}{x_0 - x_1} + f_1 \, \frac{x - x_0}{x_1 - x_0} \; . \tag{3.2.10}$$

Para encontrarmos a parábola $y_{0,1,2}(x)$ que interpola três pontos; (x_0, f_0), (x_1, f_1) e (x_2, f_2), basta escrever que

$$y_{0,1,2}(x) = \ell_0(x) f_0 + \ell_1(x) f_1 + \ell_2(x) f_2 \; , \tag{3.2.11}$$

onde neste caso $\ell_i(x)$ (i=0,1,2) são parábolas. Impondo as condições de interpolação temos que

$$\begin{array}{rcllll} \ell_0(x_0) & = & 1 & \ell_0(x_1) = 0 & \ell_0(x_2) = 0 \\ \ell_1(x_0) & = & 0 & \ell_1(x_1) = 1 & \ell_1(x_2) = 0 \\ \ell_2(x_0) & = & 0 & \ell_2(x_1) = 0 & \ell_2(x_2) = 1 \; , \end{array} \tag{3.2.12}$$

que de forma análoga nos permitem calcular a seguinte expressão para o polinómio interpolador

$$y_{0,1,2}(x) = f_0 \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} . \tag{3.2.13}$$

Em geral, se temos os $n+1$ pontos da tabela $\{(x_i, f_i)\}_{i=0}^{n}$, e queremos o polinómio interpolador $y_{0,1,\ldots,n}(x)$ desta tabela, isto é, tal que

$$y_{0,1,\ldots,n}(x_i) \equiv f_i \qquad \forall i \in \{0,1,2,\ldots,n\} , \tag{3.2.14}$$

então escrevemos que

$$y_{0,1,\ldots,n}(x) = \ell_0(x)\cdot f_0 + \ell_1(x)\cdot f_1 + \ldots + \ell_n(x)\cdot f_n , \tag{3.2.15}$$

com

$$\ell_j(x_i) = \delta_{ij} \qquad i,j \in \{0,1,2,\ldots,n\} , \tag{3.2.16}$$

onde $\delta_{ij}=1$ se $i=j$, e $\delta_{ij}=0$ se $i\neq j$. Tal como fizemos nos casos mais simples, também agora obtemos, ao impôr as condições de anulação de (3.2.15) na expressão (3.2.16), que

$$\ell_k(x_i) = 0 \quad \text{para } i \neq j \qquad \Rightarrow \qquad \ell_j(x) = C_j \prod_{k=0;k\neq j}^{n} (x-x_k) . \tag{3.2.17}$$

Finalmente, impondo o restante conjunto de condições, temos que

$$\ell_j(x_j) = 1 \quad \text{para } j = 0,1,\ldots,n \qquad \Rightarrow \qquad C_j = \frac{1}{\prod_{k=0;k\neq j}^{n} (x_j-x_k)} , \tag{3.2.18}$$

de onde vem que

$$\ell_j(x) = \frac{\prod_{k=0;k\neq j}^{n} (x-x_k)}{\prod_{k=0;k\neq j}^{n} (x_j-x_k)} . \tag{3.2.19}$$

Pelo que podemos então escrever o polinómio interpolador, na denominada fórmula de Lagrange;

$$y_{0,1,\ldots,n}(x) = \sum_{j=0}^{n} \left(f_j \cdot \frac{\prod_{k=0;k\neq j}^{n} (x-x_k)}{\prod_{k=0;k\neq j}^{n} (x_j-x_k)} \right) . \tag{3.2.20}$$

Exemplo 3.2.1: Consideremos uma função $f(x)$, da qual conhecemos três pontos: $\{(0.0,-1.0),(1.0,0.1),(1.5,2.0)\}$. Calculemos o valor da função em x=0.93 por interpolação polinomial. Recorrendo à expressão de Lagrange para o polinómio interpolador temos que (ver 3.2.13);

$$\begin{aligned} y_{0,1,2}(x) &= -1.0 \frac{(x-1.0)(x-1.5)}{(0.0-1.0)(0.0-1.5)} + 0.1 \frac{(x-0.0)(x-1.5)}{(1.0-0.0)(1.0-1.5)} \\ &\quad +2.0 \frac{(x-0.0)(x-1.0)}{(1.5-0.0)(1.5-1.0)} \\ &= -\frac{2}{3}(x-1.0)(x-1.5) - \frac{1}{5}x(x-1.5) + \frac{8}{3}x(x-1.0) . \end{aligned}$$

Pelo que obtemos finalmente que $f(0.93)\simeq -0.094$, recorrendo ao polinómio interpolador obtido.

3.2.2 Erro de aproximação usando interpolação polinomial

Como apenas usamos a tabela de pontos para inferir o valor da função, supondo para isso que um polinómio descreve aproximadamente o comportamento local da função, temos aquilo que designamos por erro da aproximação

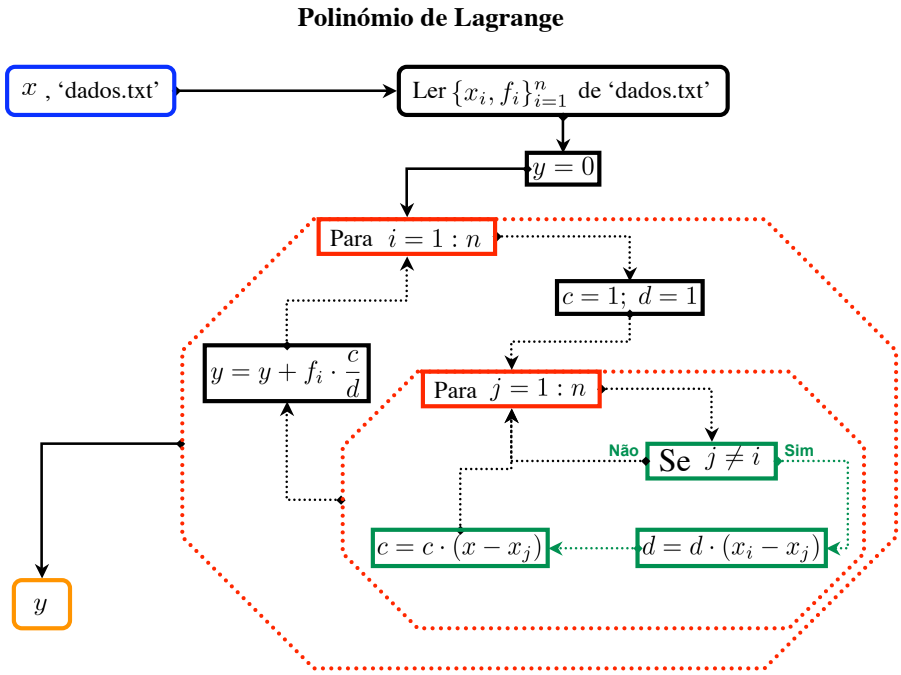


Figura 3.2: Algoritmo para implementação do polinómio de Lagrange com o objectivo de determinar o valor y, do polinómio interpolador nos pontos $\{x_i, f_i\}_{i=1}^n$, em x.

do valor da função num ponto por interpolação polinomial. Isto é, se $f(x_r)$ é o valor exacto de $f(x)$ em x_r, então $y_{0,1,\ldots,n}(x_r)$ é apenas uma aproximação desse valor com um erro formalmente definido por

$$\varepsilon_{0,1,\ldots,n}(x_r) \equiv f(x_r) - y_{0,1,\ldots,n}(x_r)\ . \tag{3.2.21}$$

Precisamos agora de encontrar uma forma de majorar este erro para todos os valores possíveis de $x_r \in [a,b]$.

Consideremos então uma função $f(x)$ e uma tabela $\{(x_i, f_i)\}_{i=0}^n$ de $n+1$ pontos desta função, que são conhecidos. Se:

- $[a,b]$ é tal que contém todos os x_i, $\forall i \in \{0,1,2,\ldots,n\}$,
- $f(x)$, $f^{(1)}(x)$, $f^{(2)}(x)$, ..., $f^{(n)}(x)$ existem e são contínuas em $[a,b]$,
- $f^{(n+1)}(x)$ existe e é contínua em $[a,b]$,

então

$$\exists \xi \in]m,M[\quad \text{com} \quad \begin{cases} m \equiv Min\{x_0, x_1, \ldots, x_n, x_r\} \\ M \equiv Max\{x_0, x_1, \ldots, x_n, x_r\}\ , \end{cases} \tag{3.2.22}$$

tal que

$$\varepsilon_{0,1,\ldots,n}(x_r) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x_r - x_i)\ . \tag{3.2.23}$$

Vejamos que assim é:

a) se $x_r = x_i$, $\forall i \in \{0,1,2,\ldots,n\}$, então

$$\prod_{i=0}^{n} (x_r - x_i) = 0 \quad \Rightarrow \quad \varepsilon_{0,1,\ldots,n}(x_r) = 0 \quad \Rightarrow \quad y_{0,1,\ldots,n}(x_i) = f(x_i)\ , \tag{3.2.24}$$

tal como usado para construir o polinómio $y_n(x)$.

b) se $x \neq x_i$, $\forall i \in \{0,1,2,...,n\}$; define-se

$$F(t) \equiv f(t) - y_{0,1,...,n}(t) - C\prod_{i=0}^{n}(t-x_i)\ , \tag{3.2.25}$$

com C (constante em t) sendo dado por, após usarmos a condição de que $F(x_r)=0$,

$$C = \frac{f(x_r) - y_{0,1,...,n}(x_r)}{\prod_{i=0}^{n}(x_r-x_i)}\ . \tag{3.2.26}$$

Temos assim que $F(t)=0$ para $t \in \{x_0, x_1, ..., x_n, x_r\}$. Logo, podemos afirmar que

$F(t)$ tem $n+2$ zeros, pelo menos, em $[a,b]$

$F^{(1)}(t)$ tem $n+1$ zeros, pelo menos, em $[a,b]$

$F^{(2)}(t)$ tem n zeros, pelo menos, em $[a,b]$

...

$F^{(n)}(t)$ tem 2 zeros, pelo menos, em $[a,b]$

$F^{(n+1)}(t)$ tem 1 zero, pelo menos, em $[a,b]$

Chamemos ξ ao zero de $F^{(n+1)}(t)$ no intervalo $[a,b]$. Então, depois de derivar (3.2.25), temos ainda que

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - C\left[\prod_{i=0}^{n}(\xi-x_i)\right]^{(n+1)} \equiv 0\ , \tag{3.2.27}$$

pois a derivada de ordem $n+1$ de um polinómio de grau n é zero. Usando o resultado

$$\left[\prod_{i=0}^{n}(\xi-x_i)\right]^{(n+1)} = (n+1)!\ , \tag{3.2.28}$$

temos finalmente que

$$C = \frac{f^{(n+1)}(\xi)}{(n+1)!}\ . \tag{3.2.29}$$

De onde se tira, depois de usar a definição (3.2.26) para C, que

$$\varepsilon_{0,1,...,n}(x_r) = f(x_r) - y_{0,1,...,n}(x_r) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x_r-x_i)\ , \tag{3.2.30}$$

tal como queriamos provar!

Desta forma mostramos que ε, tal como dado em (3.2.23), é o erro cometido em aproximar $f(x_r)$ pelo valor do polinómio interpolador de grau n da tabela $\{(x_i,f_i)\}_{i=0}^{n}$, no ponto x_r. Em termos de módulos (erro absoluto) temos que

$$\begin{aligned} |f(x_r) - y_{0,1,...,n}(x_r)| &= \left|\frac{f^{(n+1)}(\xi)}{(n+1)!}\quad\prod_{i=0}^{n}(x_r-x_i)\right| \\ &\leq \frac{\max\limits_{\xi\in[a,b]}\left|f^{(n+1)}(\xi)\right|}{(n+1)!}\quad\left|\prod_{i=0}^{n}(x_r-x_i)\right|\ . \end{aligned} \tag{3.2.31}$$

Exemplo 3.2.2: Consideremos a função $f(x)\equiv e^x$, e suponhamos que vamos usar apenas três pontos desta: $\{(0,1),(1,e),(2,e^2)\}$ para estimar o seu valor em $x=0.5$. Recorrendo à expressão de Lagrance para o polionómio interpolador temos que (ver 3.2.15);

$$\begin{aligned} y_{0,1,2}(x) &= 1\,\frac{(x-1)(x-2)}{(0-1)(0-2)} + e\,\frac{(x-0)(x-2)}{(1-0)(1-2)} + e^2\,\frac{(x-0)(x-1)}{(2-0)(2-1)} \\ &= \frac{1}{2}(x-1)(x-2) - e\,x(x-2) + \frac{e^2}{2}\,x(x-1)\,. \end{aligned}$$

Pelo que $f(0.5)\simeq 1.49$, recorrendo ao polinómio interpolador obtido. Logo o erro obtido (porque conhecemos a função) é

$$\varepsilon_{0,1,2}(0.5) = e^{1/2} - 1.49 = 0.16\,.$$

Verifiquemos agora que este valor é compatível com a expressão obtida em (3.2.31);

$$|f(x) - y_{0,1,2}(x)| \leq \frac{e^2}{3!} \quad |(x-x_0)(x-x_1)(x-x_2)| \equiv \varepsilon\,.$$

Para $x=0.5$, obtemos que $\varepsilon=0.47$. Tal como esperado este valor é um majorante do valor real do erro de interpolação obtido acima. Se considerarmos que $0.5\in[0,1]$, então podemos majorar a terceira derivada de $f(x)$ usando o seu valor em $x=1$, obtendo-se dessa forma que $\varepsilon=0.18$, sendo um valor bastante mais próximo do valor real do erro (0.16).

3.2.3 Polinómio interpolador por recorrência: fórmula de Aitken-Neville

Neste caso, tenta-se evitar a incoveniência que se tem na construcção do polinómio pela fórmula de Lagrange que está associada ao facto de ser necessário reconstruir o polinómio caso se queira adicionar mais um ponto á tabela. Neste método, que usa um processo de construção do polinómio interpolador por recorrência, pode-se facilmente adicionar a informação correspondente a um ponto, usando a expressão já calculada para a tabela inicial.

Seja então $\{(x_i,f_i)\}_{i=0}^n$ uma tabela de $n+1$ pontos (de abcissas distintas) que pretendemos interpolar construindo um polinómio $y_{\mathscr{A}}(x)$. Comecemos por definir o conjunto de abcissas;

$$\mathscr{A} \equiv \{x_i\}_{i=0}^n\,, \tag{3.2.32}$$

e dois subconjuntos $\mathscr{S}$ e $\mathscr{T}$ deste, com n pontos cada, de acordo com

$$\mathscr{S} = \mathscr{A}_0 \cup \{x_s\} \qquad \text{e} \qquad \mathscr{T} = \mathscr{A}_0 \cup \{x_t\}\,, \tag{3.2.33}$$

e tal que

$$\mathscr{A}_0 = \mathscr{A}/\{x_s,x_t\} \qquad \text{sendo} \qquad x_s \neq x_t\,. \tag{3.2.34}$$

Isto é,

$$\mathscr{A}_0 \cup \{x_s,x_t\} = \mathscr{A}\,. \tag{3.2.35}$$

Nestas condições, se

- $y_{\mathscr{S}}(x)$ é o polinómio interpolador de $f(x)$ nos pontos de $\mathscr{S}$,
- $y_{\mathscr{T}}(x)$ é o polinómio interpolador de $f(x)$ nos pontos de $\mathscr{T}$,

então o polinómio $y_{\mathscr{S}\cup\mathscr{T}}(x)\equiv y_{\mathscr{A}}(x)$, que interpola os pontos de $\mathscr{S}\cup\mathscr{T}\equiv\mathscr{A}$, é dado por

$$y_{\mathscr{A}}(x) = \alpha_s(x)\,y_{\mathscr{S}}(x) + \alpha_t(x)\,y_{\mathscr{T}}(x)\,, \tag{3.2.36}$$

onde $\alpha_s(x)$ e $\alpha_t(x)$ são polinómios de grau 1, definidos de acordo com as seguintes condições;

$$\begin{aligned} y_{\mathscr{A}}(x_s) &= y_{\mathscr{S}}(x_s) \\ y_{\mathscr{A}}(x_t) &= y_{\mathscr{T}}(x_t)\,. \end{aligned} \tag{3.2.37}$$

Daí que

$$\begin{aligned} \alpha_s(x_s) = 1 \quad \alpha_t(x_s) = 0 \\ \alpha_s(x_t) = 0 \quad \alpha_t(x_t) = 1\,, \end{aligned} \tag{3.2.38}$$

Polinómio de Aitken-Neville

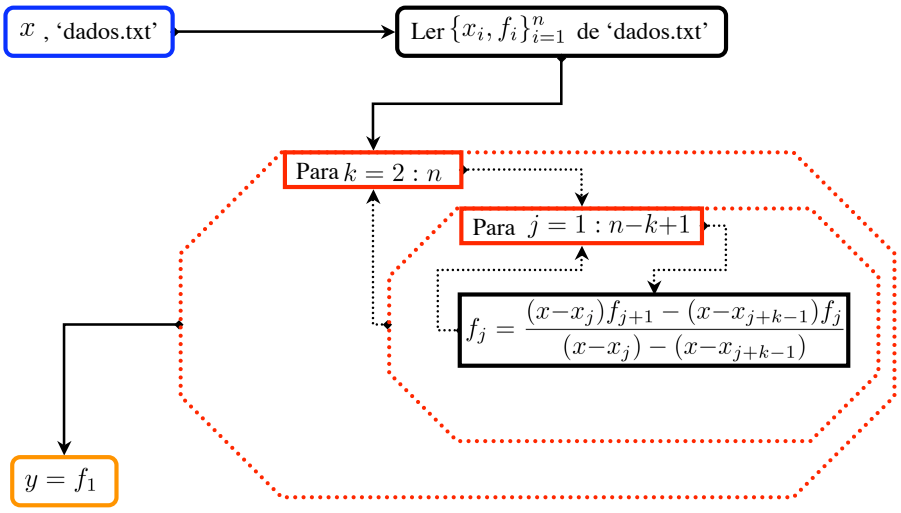


Figura 3.3: Algoritmo para implementação do polinómio de Aitken-Neville com o objectivo de determinar o valor y, do polinómio interpolador nos pontos $\{x_i, f_i\}_{i=1}^{n}$, em x.

nos dê a seguinte expressão para o polinómio interpolador;

$$y_{\mathscr{A}}(x) = \frac{x-x_t}{x_s-x_t}\, y_{\mathscr{S}}(x) + \frac{x-x_s}{x_t-x_s}\, y_{\mathscr{T}}(x) = \frac{(x-x_t)\, y_{\mathscr{S}}(x) - (x-x_s)\, y_{\mathscr{T}}(x)}{x_s-x_t} \,. \tag{3.2.39}$$

Vejamos como podemos usar esta expressão para construir a parábola que interpola uma função $f(x)$ nos seguintes pontos $\{x_0, x_1, x_2\}$.

- Primeiro contruimos a recta que interpola $f(x)$ em $\{x_0, x_1\}$;

$$\begin{array}{ll} x-x_0 & f_0 \\ x-x_1 & f_1 \end{array} \quad \Rightarrow \quad y_{0,1}(x) = \frac{(x-x_0)\, f_1 - (x-x_1)\, f_0}{x_1-x_0} \,, \tag{3.2.40}$$

- Depois contruimos a recta que interpola $f(x)$ em $\{x_1, x_2\}$;

$$\begin{array}{ll} x-x_1 & f_1 \\ x-x_2 & f_2 \end{array} \quad \Rightarrow \quad y_{1,2}(x) = \frac{(x-x_1)\, f_2 - (x-x_2)\, f_1}{x_2-x_1} \,, \tag{3.2.41}$$

- Finalmente contruimos a parábola que interpola $f(x)$ em $\{x_0, x_1, x_2\}$;

$$\begin{array}{ll} x-x_0 & f_0 \\ x-x_1 & f_1 \\ x-x_2 & f_2 \end{array} \Rightarrow \begin{array}{l} y_{0,1}(x) \\ y_{1,2}(x) \end{array} \Rightarrow \quad y_{0,1,2}(x) = \frac{(x-x_0)\, y_{1,2}(x) - (x-x_2)\, y_{0,1}(x)}{x_2-x_0} \,. \tag{3.2.42}$$

Claramente que esta expressão tem de ser equivalente (representa o mesmo polinómio) que a obtida anteriormente, pela fórmula de Lagrange. Tal é verdade pois como vimos inicialmente existe apenas um polinómio de grau 2

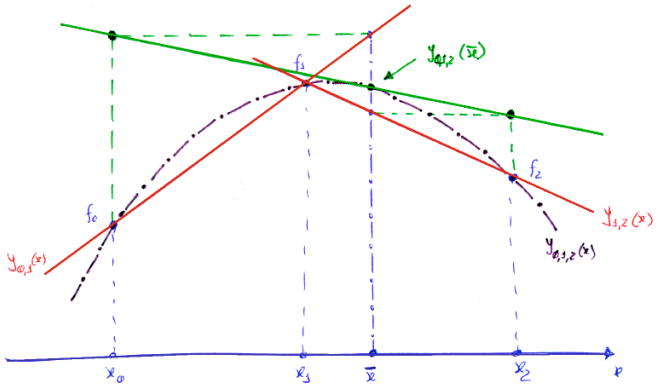


Figura 3.4: Representação da construção da parábola $y_{0,1,2}(x)$ interpoladora de uma função $f(x)$ nos pontos $\{x_0, x_1, x_2\}$, de forma a podermos estimar o valor que a função toma em $\bar{x}$. São indicadas as duas rectas $y_{0,1}(x)$ e $y_{1,2}(x)$ que nos permitem calcular por recorrência a parábola.

que interpola três pontos distintos dados. Podemos assim verificar facilmente que de facto ambos os polinómios (3.2.42) e (3.2.15) são a mesma parábola;

$$\begin{aligned} y_{0,1,2}(x) &= \frac{(x-x_0)\, y_{1,2}(x) - (x-x_2)\, y_{0,1}(x)}{x_2-x_0} \\ &= \frac{(x-x_0)}{x_2-x_0}\,\frac{(x-x_1)\, f_2 - (x-x_2)\, f_1}{x_2-x_1} - \frac{(x-x_2)}{x_2-x_0}\,\frac{(x-x_0)\, f_1 - (x-x_1)\, f_0}{x_1-x_0} \\ &= f_0\,\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f_1\,\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + f_2\,\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\,. \end{aligned} \tag{3.2.43}$$

Exemplo 3.2.3: Consideremos mais uma vez a função $f(x)$, da qual conhecemos três pontos: $\{(0.0,-1.0),(1.0,0.1),(1.5,2.0)\}$. Calculemos o valor da função em x=0.93 por interpolação polinomial, mas desta vez recorrendo ao método de Aitken-Neville para construir o polionómio interpolador (ver 3.2.40 a 3.2.42). Temos então de construir a seguinte tabela;

$$\begin{array}{ll} x-x_0 & f_0 \\ x-x_1 & f_1 \\ x-x_2 & f_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} y_{0,1}(x) = \dfrac{(x-x_0)\, f_1 - (x-x_1)\, f_0}{x_1-x_0} \\ y_{1,2}(x) = \dfrac{(x-x_1)\, f_2 - (x-x_2)\, f_1}{x_2-x_1}\,, \end{array}$$

de onde se segue que

$$y_{0,1,2}(x) = \frac{(x-x_0)\, y_{1,2}(x) - (x-x_2)\, y_{0,1}(x)}{x_2-x_0}\,.$$

Substituindo os valores, temos então que;

$$\begin{array}{ll} 0.93 & -1.0 \\ -0.07 & 0.1 \\ -0.57 & 2.0 \end{array} \quad \Rightarrow \quad \begin{array}{l} y_{0,1}(0.93) = 0.023 \\ y_{1,2}(0.93) = -0.166 \end{array} \quad \Rightarrow \quad y_{0,1,2}(0.93) = -0.094\,.$$

Pelo que temos assim que $f(0.93) \simeq -0.094$. Note que o valor terá de ser exactamente o mesmo que foi obtido no Exemplo 3.2.1, pois o poliómio é o mesmo.

Em geral, se quisermos construir a expressão do polinómio de ordem n que interpola $n+1$ pontos conhecidos de

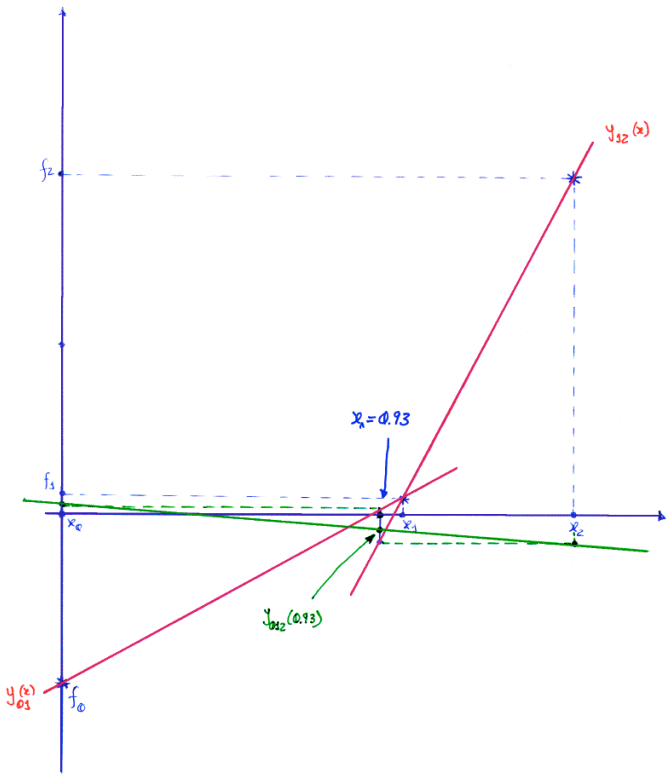


Figura 3.5: Representação do cálculo do valor da parábola $y_{0,1,2}(x)$ no ponto x=0.93 tal como calculado no Exemplo 3.2.3. São indicadas as duas rectas $y_{0,1}(x)$ e $y_{1,2}(x)$, e os valores que tomam em x=0.93, e que nos permitem calcular por recorrência a parábola interpoladora.

$\mathscr{A}=\{x_i\}_{i=0}^n$, teremos de construir todos os polinómios de menor grau que interpolam subconjuntos deste. Isto é;

$$\begin{array}{llllllll}
x-x_0 & f_0 & & & & & & \\
 & & y_{0,1} & & & & & \\
x-x_1 & f_1 & & y_{0,1,2} & & & & \\
 & & y_{1,2} & & y_{0,1,2,3} & & & \\
x-x_2 & f_2 & & y_{1,2,3} & & y_{0,1,2,3,4} & & \\
 & & & & & & & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & y_{0,1,\ldots,n-1,n} \\
 & & & & & & & \\
x-x_{n-2} & f_{n-2} & & y_{n-3,n-2,n-1} & & y_{n-3,n-2,n-1,n} & & \\
 & & y_{n-2,n-1} & & y_{n-3,n-2,n-1,n} & & & \\
x-x_{n-1} & f_{n-1} & & y_{n-2,n-1,n} & & & & \\
 & & y_{n-1,n} & & & & & \\
x-x_n & f_n & & & & & &
\end{array}$$

Note-se que caso nos seja dado um ponto adicional da tabela (x_{n+1}, f_{n+1}), então bastará calcular no esquema acima os valores de;

$$x-x_{n+1} \quad f_{n+1} \quad \rightarrow \quad y_{n,n+1} \quad \rightarrow \quad y_{n-1,n,n+1} \quad \rightarrow \quad \ldots \quad \rightarrow \quad y_{0,1,\ldots,n,n+1}\ , \tag{3.2.44}$$

sem ser necessário recalcular o resto da tabela. Esta é uma das grandes vantagens de um método que funcione por recorrência, tal como faz o método de Aitken-Neville para a construção do polinómio interpolador. Outra vantagem, é ser fácil também usar outra informação da função que não valores que esta toma.

Por exemplo, suponhamos que o valor de derivada da função é conhecido num dos pontos da tabela, então torna-se simples incluir esta informação no esquema dado acima, para o cálculo do polinómio interpolador. Seja f'_k o valor da derivada de $f(x)$ no ponto (x_k, f_k); então podemos definir uma recta que passe por este ponto e cujo declive é dado por f'_k;

$$y_{k,k}(x) = f_k + (x-x_k)\, f'_k\ . \tag{3.2.45}$$

Se escrevermos esta relação no formato adequado para o esquema de Aitken-Neville, temos que inserir uma linha com a informação da derivada da função no ponto x_k de forma a podermos determinar o polinómio interpolador que tem a mesma derivada que a função nesse ponto. Isto é feito alterando a tabela dada previamente da seguinte forma;

$$
\begin{array}{llllll}
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
 & & \ldots & & \ldots & \\
x-x_{k-1} & f_{k-1} & & y_{k-2,k-1,k} & & \ldots \\
 & & y_{k-1,k} & & y_{k-2,k-1,k,k} & \\
x-x_k & f_k & & y_{k-1,k,k} & & \ldots \\
 & & y_{k,k} & & y_{k-1,k,k,k+1} & \\
x-x_k & f'_k & & y_{k,k,k+1} & & \ldots \\
 & & y_{k,k+1} & & y_{k,k,k+1,k+2} & \\
x-x_{k+1} & f_{k+1} & & y_{k,k+1,k+2} & & \ldots \\
 & & \ldots & & \ldots & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{array}
$$

Podemos assim calcular o polinómio $y_{0,1,..,k-1,k,k,k+1,...,n}(x)$ que interpola não só a função nos pontos, mas também que tem a derivada no ponto x_k igual ao valor da derivada da função $f(x)$ nesse ponto.

Exemplo 3.2.4: Vejamos agora o caso de termos dois pontos de uma função $f(x)$: $\{(0,1),(1,2)\}$, bem como o valor de $f'(x)$ em x=0; f'_0=0. Calculemos então o valor que o polinómio interpolador toma em x=0.5, usando toda a informação disponível sobre a função;

$$
\begin{array}{ll}
x-x_0 & f_0 \\
 & \\
x-x_0 & f'_0 \\
 & \\
x-x_1 & f_1
\end{array}
\quad \Rightarrow \quad
\begin{array}{l}
y_{0,0}(x) = f'_0(x-x_0) + f_0 \\
 \\
y_{0,1}(x) = \dfrac{(x-x_0)\, f_1 - (x-x_1)\, f_0}{x_1-x_0};,
\end{array}
$$

de onde se segue que

$$
y_{0,0,1}(x) = \frac{(x-x_0)\, y_{0,1}(x) - (x-x_1)\, y_{0,0}(x)}{x_1-x_0} \, .
$$

Substituindo os valores, temos então que;

$$
\begin{array}{ll}
0.5 & 1 \\
 & \\
0.5 & 0 \\
 & \\
-0.5 & 2
\end{array}
\quad \Rightarrow \quad
\begin{array}{l}
y_{0,0}(0.5) = 1.0 \\
 \\
y_{0,1}(0.5) = 1.5
\end{array}
\quad \Rightarrow \quad y_{0,0,1}(0.5) = 1.25 \, .
$$

Temos assim que $f(0.5) \simeq 1.25$ por interpolação polinomial, em que recorremos a uma parábola usando o valor de derivada da função num ponto.

Exemplo 3.2.5: Considermos uma função $f(x)$, da qual conhecemos três pontos: $\{(0.0,-1.0),(1.0,0.1),(1.5,2.0)\}$. Esta tem um zero no intervalo $[0,2]$, e pretendemos estimar a localização desse zero recorrendo a interpolação polinomial. Para tal construimos a seguinte tabela;

$$\begin{array}{ll} 0-f_0 & x_0 \\ 0-f_1 & x_1 \\ 0-f_2 & x_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} y_{0,1}(0) = \dfrac{(0-f_0)\,x_1-(0-f_1)\,x_0}{f_1-f_0} \\ y_{1,2}(0) = \dfrac{(0-f_1)\,x_2-(0-f_2)\,x_1}{f_2-f_1} \end{array}$$

de onde se segue que

$$y_{0,1,2}(0) = \frac{(0-f_0)\,y_{1,2}(0)-(0-f_2)\,y_{0,1}(0)}{f_2-f_0}\,.$$

Substituindo os valores, temos então que;

$$\begin{array}{ll} 1.0 & 0.0 \\ -0.1 & 1.0 \\ -2.0 & 1.5 \end{array} \quad \Rightarrow \quad \begin{array}{l} y_{0,1}(0) = 0.9091 \\ y_{1,2}(0) = 0.9737 \end{array} \quad \Rightarrow \quad y_{0,1,2}(0) = 0.9306\,.$$

Temos assim que $x \simeq 0.9306$, é uma estimativa para a localização do zero de $f(x)$. A esta forma de calcular um zero de um função chama-se *interpolação polinomial inversa*.

3.2.4 Polinómio interpolador por recorrência: fórmula de Newton

Consideremos uma outra forma de construirmos por recorrência o polinómio interpolador. A fórmula de Newton para um polinómio de grau n é;

$$y_{0,1,\ldots,n}(x) = \bar{a}_0 + \bar{a}_1(x-x_0) + \bar{a}_2(x-x_0)(x-x_1) + \ldots + \bar{a}_n(x-x_0)(x-x_1)\ldots(x-x_{n-1})\,. \tag{3.2.46}$$

Se agora usarmos este polinómio como função de interpolação da tabela $\{(x_i,f_i)\}_{i=0}^n$, as condições de interpolação requerem que

$$y_{0,1,\ldots,n}(x_i) = f_i \qquad \forall i \in \{0,1,\ldots,n\}\,, \tag{3.2.47}$$

de onde resulta que;

$$\begin{array}{lll}
y_{0,1,\ldots,n}(x_0) = f_0 & \Rightarrow & \bar{a}_0 = f_0 \\
y_{0,1,\ldots,n}(x_1) = f_1 & \Rightarrow & \bar{a}_0 + \bar{a}_1(x_1-x_0) = f_1 \\
 & \Rightarrow & \bar{a}_1 = \frac{f_1-f_0}{x_1-x_0} \\
y_{0,1,\ldots,n}(x_2) = f_2 & \Rightarrow & \bar{a}_0 + \bar{a}_1(x_2-x_0) + \bar{a}_2(x_2-x_0)(x_2-x_1) = f_2 \\
 & \Rightarrow & \bar{a}_2 = \dfrac{\dfrac{f_2-f_1}{x_2-x_1} - \dfrac{f_1-f_0}{x_1-x_0}}{x_2-x_0} \\
\ldots\ etc\ \ldots & & \\
y_{0,1,\ldots,n}(x_k) = f_k & \Rightarrow & \bar{a}_k = f[x_0,x_1,\ldots,x_k] \\
\ldots\ etc\ \ldots & & \\
y_{0,1,\ldots,n}(x_n) = f_n & \Rightarrow & \bar{a}_n = f[x_0,x_1,\ldots,x_n]\,.
\end{array} \tag{3.2.48}$$

Temos assim de introduzir uma fórmula de recorrência para calcular cada um dos $\bar{a}_k$, usando para isso os valores anteriores $(\bar{a}_{k-1},\bar{a}_{k-2},\ldots,\bar{a}_0)$. Esta fórmula de recorrência é

$$f[x_j,x_{j+1},\ldots,x_{k-1},x_k] = \frac{f[x_{j+1},\ldots,x_{k-1},x_k]-f[x_j,x_{j+1},\ldots,x_{k-1}]}{x_k-x_j}\,. \tag{3.2.49}$$

Como

$$\bar{a}_k = f[x_0,x_1,\ldots,x_k]\,, \tag{3.2.50}$$

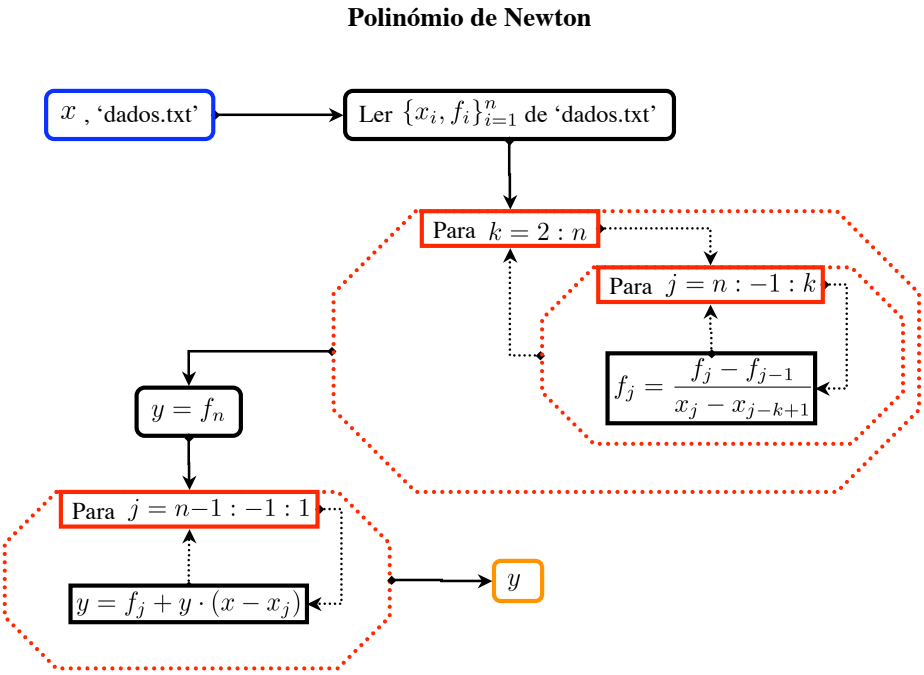


Figura 3.6: Algoritmo para implementação do polinómio de Newton com o objectivo de determinar o valor y, do polinómio interpolador nos pontos $\{x_i, f_i\}_{i=1}^n$, em x.

então

$$y_{0,1,\ldots,n}(x) = \sum_{k=0}^{n} \left\{ f[x_0, x_1, \ldots, x_k] \prod_{j=0}^{k-1} (x - x_j) \right\} , \tag{3.2.51}$$

pressupondo que

$$\prod_{j=0}^{-1} (x - x_j) = 1 \qquad \text{e} \qquad f[x_0] \equiv f_0 \, . \tag{3.2.52}$$

De uma forma análoga ao método de Aitken-Neville, podemos estabelecer um esquema simples para se implementar o cálculo por recorrência dos diferentes coeficientes $\bar{a}_k$. Este é;

$$\begin{array}{llllll}
x_0 & f_0 = \bar{a}_0 & & & & \\
 & & f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \bar{a}_1 & & & \\
x_1 & f_1 & & f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \bar{a}_2 & & \\
 & & f[x_1, x_2] = \frac{f_2 - f_1}{x_2 - x_1} & & \ldots & \\
x_2 & f_2 & & f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1} & & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \\
 & & f[x_{n-1}, x_n] = \frac{f_n - f_{n-1}}{x_n - x_{n-1}} & & & \\
x_n & f_n & & & &
\end{array} \tag{3.2.53}$$

Vejamos mais uma vez qual a expressão do parábola interpoladora de uma função $f(x)$, nos três pontos $\{(x_0, f_0),$

$(x_1,f_1),(x_2,f_2)\}$. Comecemos por calcular os coeficientes $\bar{a}_0$, $\bar{a}_1$ e $\bar{a}_2$;

$$\begin{array}{lll} x_0 & f_0=\bar{a}_0 & \\ & f[x_0,x_1]=\frac{f_1-f_0}{x_1-x_0}=\bar{a}_1 & \\ x_1 & f_1 & f[x_0,x_1,x_2]=\frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}=\bar{a}_2\ . \\ & f[x_1,x_2]=\frac{f_2-f_1}{x_2-x_1} & \\ x_2 & f_2 & \end{array} \tag{3.2.54}$$

Logo, a expressão do polinómio interpolador é;

$$\begin{aligned} y_{0,1,2}(x) &= \bar{a}_0+\bar{a}_1(x-x_0)+\bar{a}_2(x-x_0)(x-x_1) \\ \Rightarrow &= f_0+\frac{f_1-f_0}{x_1-x_0}(x-x_0)+\frac{\frac{f_2-f_1}{x_2-x_1}-\frac{f_1-f_0}{x_1-x_0}}{x_2-x_0}(x-x_0)(x-x_1)\ . \end{aligned} \tag{3.2.55}$$

Mais uma vez este polinómio é exactamente o mesmo que encontramos em (3.2.15) e (3.2.42) pelas fórmulas anteriormente apresentadas.

Na fórmula de Newton para construir o polinómio interpolador, como já acontecia com a fórmula de Aitken-Neville, quando temos de adicionar mais um ponto e reconstruir o polinómio interpolador podemos faze-lo facilmente usando o já trabalho feito. Vejamos então como tal pode ser feito; seja $y_{0,1,\ldots,n}(x)$ um polinómio de grau n que interpola os $n+1$ pontos $\{(x_i,f_i)\}_{i=0}^n$, calculado pelo método de Newton. Se adicionarmos um ponto extra (x_{n+1},f_{n+1}), vejamos como podemos obter o novo polinómio interpolador $y_{0,1,\ldots,n,n+1}(x)$ de grau $n+1$ que interpola os $n+2$ pontos $\{(x_i,f_i)\}_{i=0}^{n+1}$;

$$\begin{aligned} y_{0,1,\ldots,n,n+1}(x) &= \bar{a}_0+\bar{a}_1(x-x_0)+\bar{a}_2(x-x_0)(x-x_1)+\ldots+\bar{a}_n(x-x_0)(x-x_1)\ldots(x-x_{n-1})+ \\ &\quad +\bar{a}_{n+1}(x-x_0)(x-x_1)\ldots(x-x_{n-1})(x-x_n) \\ &= y_{0,1,\ldots,n}(x) \quad + \quad \bar{a}_{n+1}(x-x_0)(x-x_1)\ldots(x-x_{n-1})(x-x_n)\ . \end{aligned} \tag{3.2.56}$$

Ou seja, basta calcular o valor de $\bar{a}_{n+1}=f[x_0,x_1,\ldots,x_n,x_{n+1}]$, o que pode ser facilmente feito estendendo a tabela dada em (3.2.53). Feito isto temos a expressão do novo polinómio interpolador, tal como dado em (3.2.56).

Decorre naturalmente da forma como construimos o polinómio interpolador, que caso $y_{0,1,\ldots,n}(x)$ seja o polinómio interpolador de $f(x)$ na tabela $\{(x_i,f_i)\}_{i=0}^n$, então o polinómio que interpola o conjunto de $n+2$ pontos

$$\{(x_i,f_i)\}_{i=0}^n \cup \{(x_r,f_r)\}\ , \tag{3.2.57}$$

é dado por;

$$y_{0,1,\ldots,n,n+1}(x)=y_{0,1,\ldots,n}(x)+f[x_0,x_1,\ldots,x_n,x_r]\ (x-x_0)(x-x_1)\ldots(x-x_n)\ . \tag{3.2.58}$$

Desta expressão também decorre naturalmente que o erro com que $y_{0,1,\ldots,n}(x_r)$ estima $f(x_r)$ é dado por

$$\begin{aligned} f(x_r)-y_{0,1,\ldots,n}(x_r) &= y_{0,1,\ldots,n,n+1}(x_r)-y_{0,1,\ldots,n}(x_r) \\ &= f[x_0,x_1,\ldots,x_n,x_r]\ (x_r-x_0)(x_r-x_1)\ldots(x_r-x_n)\ . \end{aligned} \tag{3.2.59}$$

Se definirmos a função

$$F(t)\equiv f(t)-y_{0,1,\ldots,n}(t)\ , \tag{3.2.60}$$

então sabemos que esta tem necessariamente $n+1$ zeros, e que estes são $t\in\{x_0,x_1,\ldots,x_n\}$. Consequentemente, podemos afirmar que $F^{(n)}(t)$ tem necessariamente pelo menos um zero. Seja $x=\xi$ esse zero, de onde resulta que

$$F^{(n)}(\xi)=0 \qquad \Rightarrow \qquad f^{(n)}(\xi)-y^{(n)}_{0,1,\ldots,n}(\xi)=0\ . \tag{3.2.61}$$

Mas, sendo

$$y^{(n)}_{0,1,\ldots,n}(\xi)=\bar{a}_n\ n!=f[x_0,x_1,\ldots,x_n]\ n!\ , \tag{3.2.62}$$

significa que

$$f[x_0,x_1,\ldots,x_n]=\frac{f^{(n)}(\xi)}{n!}\ . \tag{3.2.63}$$

Assim, usando a Eq. (3.2.59), podemos finalmente escrever que

$$\begin{aligned} f(x_r) - y_{0,1,\ldots,n}(x_r) &= f[x_0, x_1, \ldots, x_n, x_r]\,(x_r - x_0)(x_r - x_1)\ldots(x_r - x_n) \\ &= \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x_r - x_i)\,, \end{aligned} \tag{3.2.64}$$

ou

$$|f(x_r) - y_{0,1,\ldots,n}(x_r)| \leq \frac{\max \left| f^{(n+1)}(\xi) \right|}{(n+1)!} \left| \prod_{i=0}^{n} (x_r - x_i) \right|\,, \tag{3.2.65}$$

como já tinhamos visto em (3.2.31).

Exemplo 3.2.6: Consideremos mais uma vez a função $f(x)$, da qual conhecemos três pontos: $\{(0.0, -1.0), (1.0, 0.1), (1.5, 2.0)\}$. Calculemos o valor da função em x=0.93 por interpolação polinomial, mas desta vez recorrendo à fórmula de Newton para construir o polionómio interpolador (ver 3.2.55). Temos então de construir a seguinte tabela;

$$\begin{array}{ll} x_0 & f_0 = \bar{a}_0 \\ & \\ x_1 & f_1 \\ & \\ x_2 & f_2 \end{array} \quad \Rightarrow \quad \begin{array}{l} f[x_0, x_1] = \dfrac{f_1 - f_0}{x_1 - x_0} = \bar{a}_1 \\ \\ f[x_1, x_2] = \dfrac{f_2 - f_1}{x_2 - x_1}\,, \end{array}$$

de onde se segue que

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \bar{a}_2\,.$$

Substituindo os valores, temos então que;

$$\begin{array}{lll} 0.0 & \bar{a}_0 = -1.0 & \\ & & \bar{a}_1 = f[x_0, x_1] = 1.1 \\ 1.0 & 0.1 & \Rightarrow \\ & & f[x_1, x_2] = 3.8 \\ 1.5 & 2.0 & \end{array} \quad \Rightarrow \quad \bar{a}_2 = f[x_0, x_1, x_2] = 1.8\,.$$

Pelo que temos que a expressão do polinómio é

$$y_{0,1,2}(x) = -1.0 + 1.1x + 1.8x(x-1)\,,$$

de onde resulta que $f(0.93) \simeq y_{0,1,2}(0.93) = -0.094$. Note que o valor terá de ser exactamente o mesmo que foi obtido nos Exemplos 3.2.1 e 3.2.3, pois o polinómio é o mesmo.

3.3 Interpolação por splines polinomiais

Nesta seccção consideramos outro tipo de função interpoladora. Se tivermos uma tabela extensa de pontos, claramente teremos de utilizar um polinómio interpolador de grau elevado. Tal pode originar problemas pois polinómios de elevado grau são extremamente "corrugados", isto é, podem ter localmente elevados valores da sua derivada. Em termos gráficos isto pode ser descrito como sendo uma função interpoladora que eventualmente oscila de cima para baixo rapidamente, entre pontos de interpolação.

Assim, de forma a reduzir este problema mas mantendo a vantagem de ser fácil construir polinómios, surgem as *splines polinomiais*. Estas são funções que apenas localmente coincidem com polinómios de baixo grau, mas que no entanto podem interpolar um número elevado de pontos, mantendo-se assim uma função interpoladora que é regular, não apresentando variações rápidas entre pontos que interpola.

Uma *spline S* de *grau* m_S ($m_S \geq 0$), com *n nodos*:

$$x_0 < x_1 < x_2 < \ldots < x_n\,, \tag{3.3.1}$$

no intervalo $[a,b]$ (em que $a{=}x_0$ e $b{=}x_n$), é definida de acordo com;

$$\begin{cases} \rightarrow S \quad \text{coincide em cada subintervalo} \quad \Omega_i{=}[x_{i-1},x_i], \quad \text{para} \quad i \in \{1,2,...,n\}, \\ \qquad\quad \text{com um polinómio de grau menor ou igual a } m_S \\ \\ \rightarrow S,\, S',\, ...,\, S^{(m_S-1)} \text{ são contínuas em } [a,b]. \end{cases} \tag{3.3.2}$$

Define-se ainda a *característica da malha*, como sendo

$$h \equiv \max_{1\leq i\leq n} h_i \qquad \text{onde} \qquad h_i \equiv x_i{-}x_{i-1}\ , \tag{3.3.3}$$

e os *parâmetros* M_i da spline, de acordo com

$$M_i \equiv S''(x_i) \qquad i=0,1,...,n\ . \tag{3.3.4}$$

Isto é, os parâmetros correspondem aos valores da segunda derivada da spline nos nodos. Finalmente, por *splines parciais* $S_i(x)$ de uma spline polinomial $S(x)$ de grau m_S, referimo-nos às funções que coincidem com a spline apenas em cada intervalo Ω_i. Ou seja, por definição, temos que a spline parcial $S_i(x)$, cujo domínio é o interval Ω_i, é dada por

$$S_i(x) = S(x) \quad \text{para} \quad x \in [x_{i-1},x_i]\ . \tag{3.3.5}$$

De notar que a função $S_i(x)$ só está definida para $x{\in}[x_{i-1},x_i]$, sendo necessáriamente um polinómio de grau igual ou inferior a m_S. Logo, para construir a spline basta-nos encontrar os diferentes polinómios de grau m_S que correspondem aos diferentes intervalos Ω_i, e que verificam as condições estabelecidas na definição de splines polinomiais.

Assim, para interpolar uma tabela $\{(x_i,f_i)\}_{i=0}^n$ de $n{+}1$ pontos, bastará construir a spline $S(x)$ cujos nodos são os pontos x_i ($i{=}0,1,..,n$) e tal que esta tenha os valores f_i nesses nodos, isto é, $S(x_i){=}f_i$.

3.3.1 Splines de grau 0, 1 e 2

A título de exemplo consideremos primeiro três casos simples de construção de splines polinomiais interpoladoras de uma função $f(x)$ em $n{+}1$ pontos.

<u>*Splines de grau* $m_S{=}0$</u>: neste caso, e por definição, as splines parciais em cada intervalo Ω_i são necessariamente constantes (polinómio de grau 0), pelo que basta construir $S(x)$ usando que

$$S_i(x) = C_i \qquad \text{para} \qquad x \in \Omega_i\ . \tag{3.3.6}$$

Como, queremos que a spline interpole os pontos $\{(x_i,f_i)\}_{i=0}^n$, então basta impor que

$$S_i(x_i) = f_i \qquad \Rightarrow \qquad C_i = f_i\ . \tag{3.3.7}$$

A função que resulta daqui é representada na Fig. 3.7.

Note que em vez de (3.3.7), podíamos ter decidido usar a condição à esquerda, dizendo que $S_i(x_i){=}f_{i-1}$, e de onde resultaria que $C_i{=}f_{i-1}$. Em qualquer uma das opções há sempre um extremo que não é usado na definição das constantes C_i.

O erro de interpolação pode ser facilmente calculado pois $S_i(x)$ é um polinómio em Ω_i, logo

$$S_i(x) = f(x) + f'(\xi)\,(x{-}x_i) \qquad \text{para} \quad \xi \in [x_{i-1},x_i]\ , \tag{3.3.8}$$

de onde resulta que (ver 3.2.31);

$$\begin{aligned} |e_i(x)| &\equiv |S_i(x){-}f(x)| = \left|f'(\xi)\right|\ |x{-}x_i| \\ &\leq \max_{\xi\in[x_{i-1},x_i]} \left|f'(\xi)\right| \cdot h_i\ . \end{aligned} \tag{3.3.9}$$

<u>*Splines de grau* $m_S{=}1$</u>: neste caso, as funções $S_i(x)$ são rectas nos intervalos Ω_i. Mas de acordo com a definição, precisamos ainda de exigir a continuidade da função spline $S(x)$.

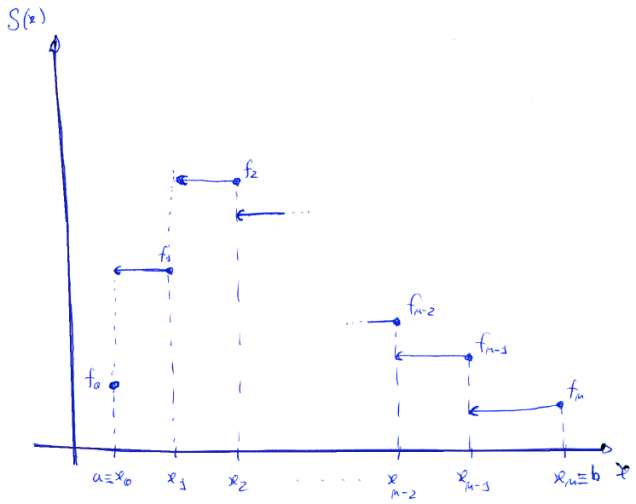


Figura 3.7: Representação de uma spline $S(x)$, de grau m_S=0, que interpola a função $f(x)$ nos pontos $\{x_0, x_1, ..., x_n\}$.

Se $S(x)$ interpola $f(x)$ nos pontos $\{(x_i, f_i)\}_{i=0}^{n}$, e é contínua aí, então teremos de ter que

$$S_i(x_{i-1}) = f_{i-1} \qquad \text{e} \qquad S_i(x_i) = f_i \qquad i = 1, ..., n \; . \tag{3.3.10}$$

Comecemos por construir a primeira spline parcial, $S_1(x)$. Sendo um recta será dada por

$$S_1(x) = a + bx \qquad \text{para} \qquad x_0 \leq x \leq x_1 \; , \tag{3.3.11}$$

Impondo as duas condição de interpolação para os pontos x_0 e x_1, temos que

$$\begin{cases} S_1(x_0) = f_0 \\ S_1(x_1) = f_1 \; , \end{cases} \tag{3.3.12}$$

de onde resulta que (ver a Secção 3.2 sobre interpolação polinomial),

$$S_1(x) = f_0 \, \frac{x_1 - x}{h_1} + f_1 \, \frac{x - x_0}{h_1} \; . \tag{3.3.13}$$

Podemos então ver que de forma análoga é possível construir todas as outras splines parciais

$$S_i(x) = f_{i-1} \, \frac{x_i - x}{h_i} + f_i \, \frac{x - x_{i-1}}{h_i} \qquad \text{com} \quad x_{i-1} \leq x \leq x_i \quad \text{e para} \quad i = 1, 2, ..., n \; . \tag{3.3.14}$$

Assim, temos que a spline $S(x)$ que interpola a tabela é construída usando segmentos de recta que se ligam nos nodos da spline (ver Fig. 3.8), garantindo assim que $S(x)$ é contínua.

Neste caso, o erro de interpolação é mais uma vez obtido a partir do erro de interpolação polinomial para uma recta, em cada intervalo Ω_i. Isto é,

$$\begin{aligned} |e_i(x)| \quad &= \quad \left| \frac{f^{(2)}(\xi)}{2!} \right| \cdot |(x - x_{i-1})(x - x_i)| \\ &\leq \quad \max_{\xi \in [x_{i-1}, x_i]} |f^{(2)}(\xi)| \cdot \frac{h_i^2}{8} \quad \text{para} \quad i = 1, 2, ..., n \; , \end{aligned} \tag{3.3.15}$$

pois $|(x - x_{i-1})(x - x_i)| \leq h_i^2/4$.

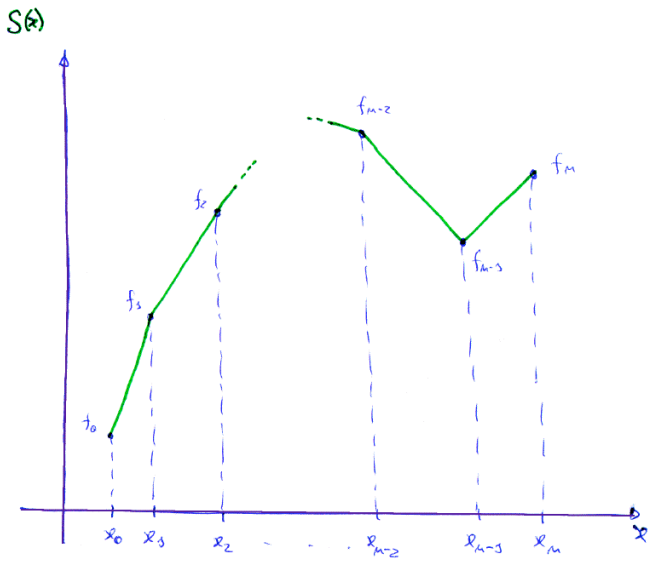


Figura 3.8: Representação de uma spline $S(x)$, de grau m_S=1, que interpola a função $f(x)$ nos pontos $\{x_0, x_1, ..., x_n\}$. Esta é constituída por segmentos de recta que se ligam nos nodos de forma a garantir que a função $S(x)$ é contínua, tal como exigido pela definição de spline.

Exemplo 3.3.1: Consideremos que temos a seguinte tabela de três pontos de uma função real, $\{(0,1),(1,0),(4,2)\}$, que pretendemos interpolar usando uma spline de grau m=1. Basta-nos então construir os segmentos de recta entre cada ponto, obtendo que

$$S(x) = \begin{cases} S_1(x) \equiv 1-x & ; 0 \leq x \leq 1 \\ S_2(x) \equiv \frac{2}{3}(x-1) & ; 1 \leq x \leq 4 \, . \end{cases}$$

Podemos assim estimar o valor da função em x=2, pois $f(2) \simeq S(2) = S_2(2) = 2/3$.

Splines de grau m_S=2: mais uma vez temos que as splines parciais são desta vez parábolas. Assim a spline é construída de acordo com;

$$\begin{cases} S_i(x) & \text{são parábolas em cada } \Omega_i \text{ respectivo,} \\ S(x) \quad \text{e} \quad S'(x) & \text{são contínuas.} \end{cases} \tag{3.3.16}$$

Consideremos então primeiro a continuidade de $S(x)$:

$$\forall i \in \{0,1,...,n\} \qquad S(x_i) = f_i \, , \tag{3.3.17}$$

pelo que $S_i(x_{i-1}) = f_{i-1}$, de onde segue, visto $S_i(x)$ ser uma parábola, que

$$S_i(x) = f_{i-1} + m_{i-1}(x - x_{i-1}) + \frac{M_i}{2}(x - x_{i-1})^2 \, . \tag{3.3.18}$$

Aqui introduzimos os valores da primeira derivada nos nodos, como sendo

$$m_i \equiv S'(x_i) \qquad \text{para} \qquad i = 0,1,...,n \, , \tag{3.3.19}$$

e usamos os parâmetros da spline, definidos em (3.3.4). Se derivarmos a expressão (3.3.18), temos que

$$S_i'(x) = m_{i-1} + M_i(x - x_{i-1}) \, , \tag{3.3.20}$$

que quando avaliado em x=x_i requer, pois $S'(x)$ tem de ser contínua, que

$$m_i \equiv S'(x_i) = m_{i-1} + M_i \, h_i \, , \tag{3.3.21}$$

de onde resulta que

$$M_i = \frac{m_i - m_{i-1}}{h_i} \ . \tag{3.3.22}$$

Finalmente, a continuidade de $S(x)$ em x_i também exige que se tenha

$$\begin{aligned} S_i(x_i) = f_i \qquad &\Rightarrow \qquad f_i = f_{i-1} + m_{i-1}(x_i - x_{i-1}) + \frac{m_i - m_{i-1}}{2h_i}(x_i - x_{i-1})^2 \\ &\Rightarrow \quad f_i - f_{i-1} = m_{i-1}h_i + \frac{1}{2}(m_i - m_{i-1})h_i \ , \end{aligned} \tag{3.3.23}$$

para $i=1,2,...,n$; correspondendo a n condições. Estas podem ser escritas na forma

$$m_i + m_{i-1} = 2\,\frac{f_i - f_{i-1}}{h_i} \qquad i=1,2,...,n \ . \tag{3.3.24}$$

Como a continuidade da derivada apenas pode ser imposta nos nodos internos, falta-nos uma condição para podermos definir a spline $S(x)$. Isto é, precisamos, por exemplo, do valor de m_0, já que as n relações dadas em (3.3.24) envolvem $n+1$ variáveis $\{m_0, m_1, ..., m_n\}$. Dada a condição extra, sob a forma da derivada num dos nodos, por exemplo, é possível calcular o valor de todos os m_i. A partir dos quais se pode então construir a função interpoladora sob a forma de uma spline polinomial de grau 2:

$$x \in [x_{i-1}, x_i] \quad \Rightarrow \quad S(x) = S_i(x) \equiv f_{i-1} + m_{i-1}(x - x_{i-1}) + \frac{m_i - m_{i-1}}{2h_i}(x - x_{i-1})^2 \ , \tag{3.3.25}$$

ou seja, usamos a spline parcial para um dos valores de $i \in \{1,2,...,n\}$, dependendo do valor x em que queremos avaliar $S(x)$.

Exemplo 3.3.2: Consideremos novamente que temos a seguinte tabela de três pontos de uma função; $\{(0,1),(1,0),(4,2)\}$, que pretendemos interpolar usando agora uma spline de grau $m_S=2$. Para tal precisamos primeiro de especificar um dos valores de m_i; seja então o valor de $f'(1)=0$. Logo, temos que $m_1=0$, e usando (3.3.24) temos mais duas equações;

$$m_1 + m_0 = 2\,\frac{0-1}{1-0} \qquad \text{e} \qquad m_2 + m_1 = 2\,\frac{2-0}{4-1} \ ,$$

Destas resulta que $m_0=-2$, $m_1=0$ e $m_2=4/3$. Podemos então construir os segmentos de parábola entre cada par de pontos (ver 3.3.25), obtendo que

$$S(x) = \begin{cases} S_1(x) \equiv 1 - 2x + x^2 & ; 0 \le x \le 1 \\ \\ S_2(x) \equiv \frac{2}{9}\,(x-1)^2 & ; 1 \le x \le 4 \ . \end{cases}$$

Para estimar o valor da função em $x=2$, basta calcular $f(2) \simeq S(2) = S_2(2) = 2/9$.

3.3.2 Splines cúbicas (grau 3)

Tal como definido em (3.3.2), uma spline $S(x)$ cúbica (com grau $m_S=3$) é tal que

$$\begin{cases} S_i(x) & \text{são polinómios de grau } m_S \le 3 \text{ em cada } \Omega_i \text{ respectivo,} \\ \\ S(x)\ , S'(x) \quad \text{e} \quad S''(x) & \text{são contínuas.} \end{cases} \tag{3.3.26}$$

Daí que a segunda derivada das splines parciais sejam rectas, pelo que podemos escrever que

$$S_i''(x) = M_{i-1}\,\frac{x_i - x}{h_i} + M_i\,\frac{x - x_{i-1}}{h_i} \qquad \text{para} \quad x \in [x_{i-1}, x_i] \ , \tag{3.3.27}$$

onde, mais uma vez, $M_i \equiv S''(x_i)$ para $i=0,1,..,n$. Ao escrevermos as segundas derivadas das splines parciais nesta forma, estamos automaticamente a assegurar que $S''(x)$ é uma função contínua. Agora precisamos de exigir a continuidade da primeira derivada nos nodos, bem como que a spline tome os valores f_i nos nodos. Para tal, integramos duas vezes a expressão dada em (3.3.27) obtendo que

$$S_i(x) = M_{i-1}\,\frac{(x_i - x)^3}{6h_i} + M_i\,\frac{(x - x_{i-1})^3}{6h_i} + C_i\,\frac{x_i - x}{h_i} + D_i\,\frac{x - x_{i-1}}{h_i} \ , \tag{3.3.28}$$

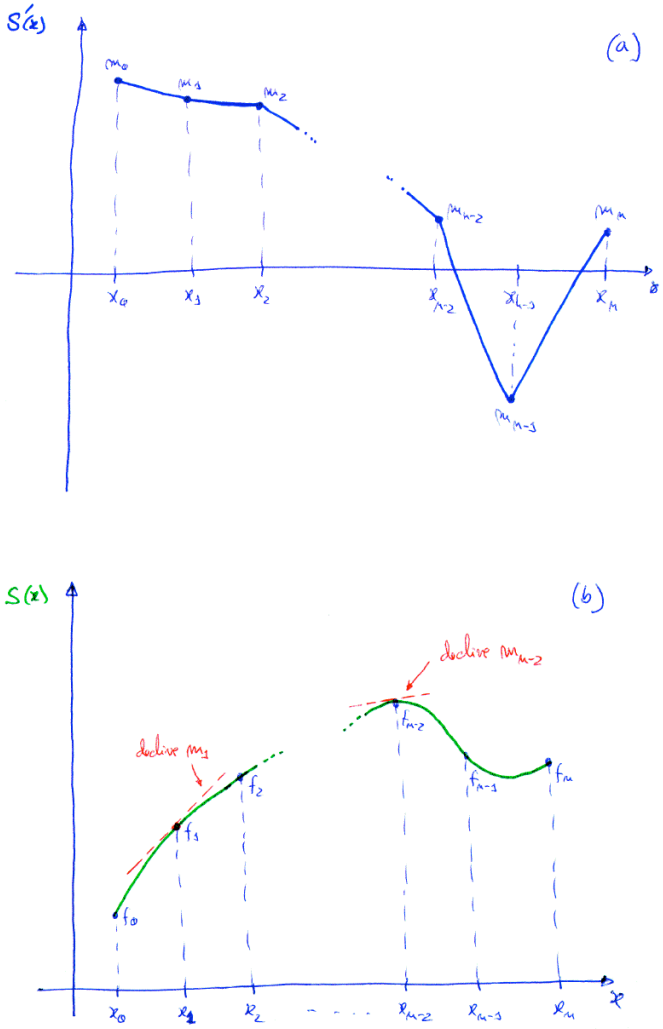


Figura 3.9: (a) Representação da derivada de uma spline $S(x)$, de grau m_S=2, que interpola a função $f(x)$ nos pontos $\{x_0, x_1, ..., x_n\}$. Esta é constituída por segmentos de recta que se ligam nos nodos de forma a garantir que a função $S'(x)$ é contínua, tal como exigido pela definição de spline. (b) Representação da spline $S(x)$, correspondendo a parábolas para cada intervalo Ω_i, e tal que em cada nodo a tangente tem declive m_i.

onde C_i e D_i são constantes de integração, para o intervalo $x\in[x_{i-1},x_i]$. Comecemos então por usar as condições de interpolação;

$$S_i(x_{i-1}) = f_{i-1} \qquad \text{e} \qquad S_i(x_i) = f_i \,, \tag{3.3.29}$$

obtendo-se que

$$\begin{aligned} f_{i-1} &= M_{i-1}\,\frac{(x_i-x_{i-1})^3}{6h_i} + C_i\,\frac{x_i-x_{i-1}}{h_i} = M_{i-1}\,\frac{h_i^2}{6} + C_i \\ &\Rightarrow C_i = f_{i-1} - M_{i-1}\,\frac{h_i^2}{6} \end{aligned} \tag{3.3.30}$$

$$\begin{aligned} f_i &= M_i\,\frac{(x_i-x_{i-1})^3}{6h_i} + D_i\,\frac{x_i-x_{i-1}}{h_i} = M_i\,\frac{h_i^2}{6} + D_i \\ &\Rightarrow D_i = f_i - M_i\,\frac{h_i^2}{6}\,, \end{aligned} \tag{3.3.31}$$

para i=1,2,...,n. Finalmente, falta-nos impor a condição que garanta a continuidade da primeira derivada. Para tal basta considerar que

$$S_i'(x_i) = S_{i+1}'(x_i) \qquad \text{para} \quad i = 1,2,...,n-1\,. \tag{3.3.32}$$

Derivando (3.3.28), e depois de substituir as expressões obtidas em (3.3.30) e (3.3.31), ficamos com

$$\begin{aligned} S_i'(x) &= -M_{i-1}\,\frac{(x_i-x)^2}{2h_i} + M_i\,\frac{(x-x_{i-1})^2}{2h_i} - \frac{f_{i-1}}{h_i} + M_{i-1}\,\frac{h_i}{6} + \frac{f_i}{h_i} - M_i\,\frac{h_i}{6} \\ &= -M_{i-1}\,\frac{(x_i-x)^2}{2h_i} + M_i\,\frac{(x-x_{i-1})^2}{2h_i} + \frac{f_i-f_{i-1}}{h_i} - (M_i-M_{i-1})\,\frac{h_i}{6}\,, \end{aligned} \tag{3.3.33}$$

de onde se tem, usando (3.3.31), que

$$M_i\,\frac{h_i}{2} + \frac{f_i-f_{i-1}}{h_i} - (M_i-M_{i-1})\,\frac{h_i}{6} = -M_i\,\frac{h_{i+1}}{2} + \frac{f_{i+1}-f_i}{h_{i+1}} - (M_{i+1}-M_i)\,\frac{h_{i+1}}{6}\,. \tag{3.3.34}$$

Isto é, temos $n-1$ condições, para as $n+1$ incógnitas M_i, que são escritas na forma

$$M_{i-1}\frac{h_i}{6} + M_i\,\frac{h_i+h_{i+1}}{3} + M_{i+1}\,\frac{h_{i+1}}{6} = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i} \qquad i = 1,2,...,n-1\,. \tag{3.3.35}$$

Temos assim o seguinte sistema de $n-1$ equações

$$\begin{cases} M_0\dfrac{h_1}{6} + M_1\,\dfrac{h_1+h_2}{3} + M_2\,\dfrac{h_2}{6} = \dfrac{f_2-f_1}{h_2} - \dfrac{f_1-f_0}{h_1} & ;\, i{=}1 \\ M_1\dfrac{h_2}{6} + M_2\,\dfrac{h_2+h_3}{3} + M_3\,\dfrac{h_3}{6} = \dfrac{f_3-f_2}{h_3} - \dfrac{f_2-f_1}{h_2} & ;\, i{=}2 \\ .... & \\ M_{n-2}\dfrac{h_{n-1}}{6} + M_{n-1}\,\dfrac{h_{n-1}+h_n}{3} + M_n\,\dfrac{h_n}{6} = \dfrac{f_n-f_{n-1}}{h_n} - \dfrac{f_{n-1}-f_{n-2}}{h_{n-1}} & ;\, i{=}n{-}1\,, \end{cases} \tag{3.3.36}$$

que pode ser escrito na forma matricial como,

$$\begin{bmatrix} \frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \dots & & & & \\ 0 & \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \dots & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \\ & & & \dots & \frac{h_{n-2}}{6} & \frac{h_{n-2}+h_{n-1}}{3} & \frac{h_{n-1}}{6} & 0 & \\ & & & \dots & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} & \frac{h_n}{6} & \end{bmatrix}_{(n-1)\times(n+1)} \times \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ \dots \\ M_{n-2} \\ M_{n-1} \\ M_n \end{bmatrix}_{(n+1)} =$$

$$= \begin{bmatrix} \frac{f_2-f_1}{h_2} - \frac{f_1-f_0}{h_1} \\ \frac{f_3-f_2}{h_3} - \frac{f_2-f_1}{h_2} \\ \dots \\ \frac{f_{n-1}-f_{n-2}}{h_{n-1}} - \frac{f_{n-2}-f_{n-3}}{h_{n-2}} \\ \frac{f_n-f_{n-1}}{h_n} - \frac{f_{n-1}-f_{n-2}}{h_{n-1}} \end{bmatrix}_{(n-1)}\,. \tag{3.3.37}$$

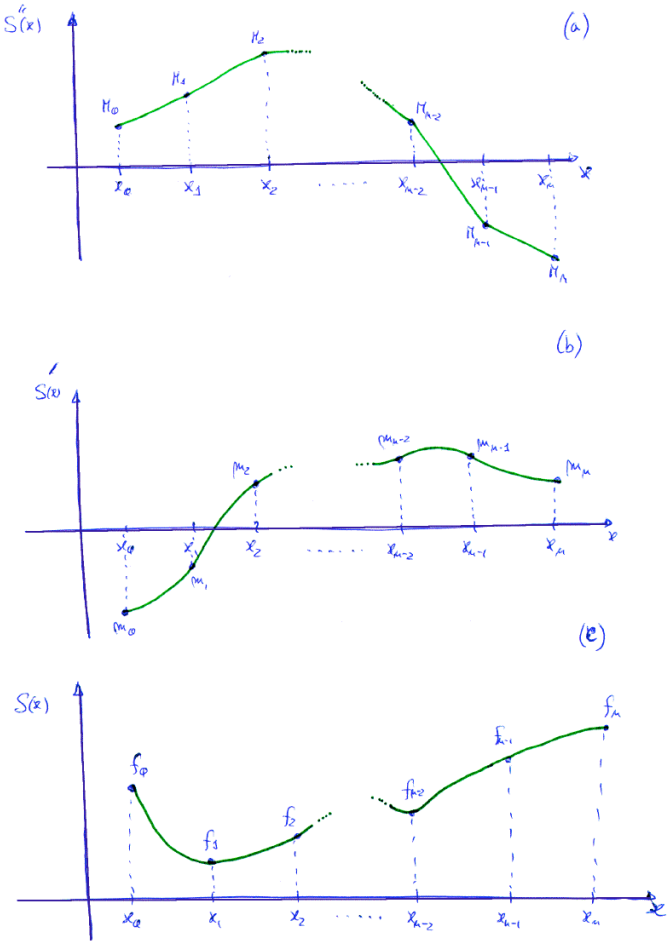


Figura 3.10: (a) Representação da segunda derivada de uma spline $S(x)$, de grau m_S=3, que interpola a função $f(x)$ nos pontos $\{x_0, x_1, ..., x_n\}$. Esta é constituída por segmentos de recta que se ligam nos nodos de forma a garantir que a função $S''(x)$ é contínua, tal como exigido pela definição de spline. (b) Representação da derivada da spline $S'(x)$, correspondendo a parábolas para cada intervalo Ω_i. (c) Representação da spline $S(x)$, correspondendo a cúbicas em cada intervalo entre nodos da spline.

Para podermos construir a spline cúbica ainda nos faltam duas condições, pois só então poderemos calcular todos os valores M_i. Existem várias opções que são mais comumente usadas para especificar as duas condições em falta, sendo algumas delas dadas em baixo.

Após termos calculado todos os valores dos parâmetros M_i podemos escrever a expressão para as splines parciais de forma a calcular $S(x)$ em qualquer ponto. Pois a spline parcial de ordem i é dada por

$$\begin{aligned} S_i(x) &= M_{i-1}\frac{(x_i-x)^3}{6h_i} + M_i\frac{(x-x_{i-1})^3}{6h_i} + \\ &+ \left(f_{i-1} - M_{i-1}\frac{h_i^2}{6}\right)\frac{x_i-x}{h_i} + \left(f_i - M_i\frac{h_i^2}{6}\right)\frac{x-x_{i-1}}{h_i}\,, \end{aligned} \tag{3.3.38}$$

para $x \in [x_{i-1}, x_i]$.

Spline completa: uma das formas de poder fechar o sistema de equações que nos permite calcular os parâmetros da spline é conhecendo o valor que a derivada da função toma nos extremos. Ou seja, sabendo que

$$\begin{cases} S_1'(x_0) = f_0' \\ S_n'(x_n) = f_n'\,, \end{cases} \tag{3.3.39}$$

Após usarmos a expressão (3.3.32) aqui, podemos adicionar às $n-1$ equações dadas em (3.3.35) as seguintes duas

condições;

$$\begin{aligned} M_0\,\frac{h_1}{3}+M_1\,\frac{h_1}{6} &= \frac{f_1-f_0}{h_1}-f_0' \\ M_{n-1}\,\frac{h_n}{6}+M_n\,\frac{h_n}{3} &= f_n'-\frac{f_n-f_{n-1}}{h_n}\,. \end{aligned} \tag{3.3.40}$$

A função spline que resulta é usualmente denominada *spline completa*.

Note-se que os dois valores da derivada da função $f(x)$ podem ser especificados em qualquer um dos nodos da Spline, ou mesmo em qualquer ponto do intervalo $[a,b]$.

Exemplo 3.3.3: Consideremos mais uma vez a seguinte tabela de três pontos de uma função real, $\{(0,1),(1,0),(4,2)\}$, que pretendemos interpolar usando agora uma spline de grau m_S=3 *completa*. Para tal precisamos primeiro de especificar as duas condições dadas em (3.3.39) usando o facto de $f'(0)$=1 e $f'(4)$=0;

$$\begin{aligned} M_0\,\frac{1}{3}+M_1\,\frac{1}{6} &= \frac{0-1}{1}-1 \\ M_1\,\frac{3}{6}+M_2\,\frac{3}{3} &= 0-\frac{2-0}{3}\,. \end{aligned}$$

Precisamos agora adicionar a condições que falta recorrendo a (3.3.35) que é (i=1)

$$M_0\frac{1}{6}+M_1\,\frac{1+3}{3}+M_2\,\frac{3}{6}=\frac{2-0}{3}-\frac{0-1}{1}\,.$$

Temos assim que

$$\begin{cases} 2M_0+M_1=-12 \\ 3M_1+6M_2=-4 \\ M_0+8M_1+3M_2=10\,. \end{cases}$$

Resolvendo este sistema resulta que M_0=$-15/2$, M_1=3 e M_2=$-13/6$. Podemos então construir os segmentos de cúbicas entre cada par de pontos (ver 3.3.37), obtendo que

$$S(x)=\begin{cases} S_1(x)\equiv -\frac{5}{4}(1-x)^3+\frac{1}{2}x^3+\frac{9}{4}(1-x)-\frac{1}{2}x & ;\,0\leq x\leq 1 \\ S_2(x)\equiv \frac{1}{6}(4-x)^3-\frac{13}{108}(x-1)^3-\frac{3}{2}(4-x)+\frac{7}{4}(x-1) & ;\,1\leq x\leq 4\,. \end{cases}$$

Para estimar o valor da função em x=2, basta calcular que $f(2)\simeq S(2)=S_2(2)$, logo tem-se que $f(2)\simeq -1/27$.

<u>*Spline natural*</u>: caso não se disponha de mais nenhuma informação sobre a função $f(x)$ então a opcção normalmente usada é pressupor que a segunda derivada da spline nos extremos é nula. Isto é;

$$\begin{cases} M_0\equiv S_0''(x_0)=0 \\ M_n\equiv S_n''(x_n)=0\,. \end{cases} \tag{3.3.41}$$

Temos então apenas que resolver as $n-1$ equações dadas em (3.3.35) para obter os valores dos parâmetros $\{M_1,M_2,\ldots,;M_{n-1}\}$.

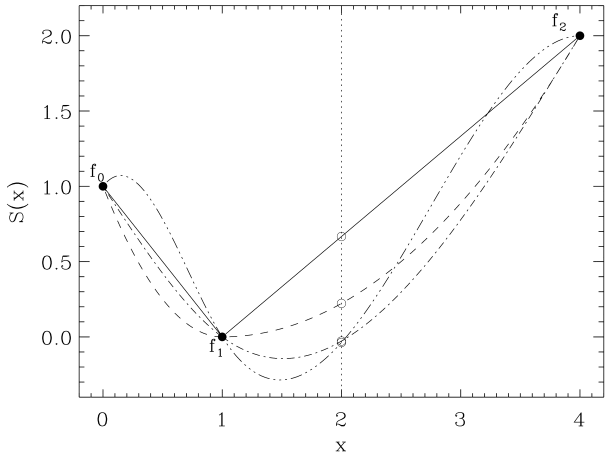


Figura 3.11: Representação de splines de grau m_S=1 (linha contínua), m_S=2 (linha tracejada) e m_S=3 (*completa* - linha tracejada-3ponteada; *natural* - linha tracejada-ponteada), que correspondem aos Exemplos 3.3.1-4. Todas estas splines interpolam a função $f(x)$ nos pontos $\{(0,1),(1,0),(4,2)\}$, permitindo-nos estimar o valor de $f(2)$.

Exemplo 3.3.4: Consideremos mais uma vez a tabela de três pontos para uma função real, $\{(0,1),(1,0),(4,2)\}$, que pretendemos interpolar usando agora uma spline de grau m=3 *natural*. Para tal precisamos primeiro de especificar as duas condições dadas em (3.3.38) que correspondem a M_0=0 e M_2=0. Precisamos ainda de adicionar a condição que falta recorrendo a (3.3.35) que é (i=1)

$$M_0\frac{1}{6}+M_1\,\frac{1+3}{3}+M_2\,\frac{3}{6}=\frac{2-0}{3}-\frac{0-1}{1}\,.$$

Resolvendo esta equação obtém-se que M_0=0, M_1=5/4 e M_2=0. Podemos então construir os segmentos de cúbicas entre cada par de pontos (ver 3.3.37), obtendo que

$$S(x)=\begin{cases} S_1(x)\equiv \frac{5}{24}x^3+(1-x)-\frac{5}{24}x & ;\,0\leq x\leq 1\\ S_2(x)\equiv \frac{5}{72}(4-x)^3-\frac{5}{8}(4-x)+\frac{2}{3}(x-1) & ;\,1\leq x\leq 4\end{cases}$$

Para estimar o valor da função em x=2, basta calcular que $f(2)\simeq S(2)=S_2(2)$, logo tem-se que $f(2)\simeq -1/36$.

Spline com continuidade da terceira derivada: outra alternativa, é usar para condições extras a continuidade da terceira derivada em dois nodos; por exemplo em x_1 e x_{n-1}. Desta forma não precisamos de adicionar informação extra explicitamente sendo as condições impostas de uma forma "interna". A terceira derivada é dada por;

$$S_i'''(x)=\frac{M_i-M_{i-1}}{h_i}\,. \tag{3.3.42}$$

Logo, as duas condições que precisamos correspondem a

$$S_1'''(x_1)=S_2'''(x_1)\qquad\text{e}\qquad S_{n-1}'''(x_{n-1})=S_n'''(x_{n-1})\,, \tag{3.3.43}$$

de onde resulta que

$$\begin{aligned} -\frac{M_0}{h_1}+\left(\frac{1}{h_1}+\frac{1}{h_2}\right)M_1-\frac{M_2}{h_2}&=0\\ -\frac{M_{n-2}}{h_{n-1}}+\left(\frac{1}{h_{n-1}}+\frac{1}{h_n}\right)M_{n-1}-\frac{M_n}{h_n}&=0\,. \end{aligned} \tag{3.3.44}$$

3.3.3 Resolução de sistemas de equações lineares

Embora não esteja normalmente incluído nos assuntos abordados em Métodos Numéricos, vamos aqui brevemente descrever um das técnicas mais básicas para resolver sistemas de equações algébricas, pois o cálculo de Splines envolve a resolução de sistemas lineares com um elevado número de equações.

O único método que se apresenta aqui é o *método de eliminação Gaussiana com pivotagem*. Existem outras formas de encontrar a solução de um sistema de equações, mas que podem também ser facilmente encontradas na literatura.

Comecemos então por considerar o seguinte sistema de n equações lineares para as n variáveis $\{x_1, x_2, ..., x_n\}$;

$$\begin{cases} a_{11}x_1 & + a_{12}x_2 & + ... & + a_{1n}x_n & = f_1 \\ a_{21}x_1 & + a_{22}x_2 & + ... & + a_{2n}x_n & = f_2 \\ ... & ... & ... & ... & ... \\ a_{n1}x_1 & + a_{n2}x_2 & + ... & + a_{nn}x_n & = f_n \, . \end{cases} \tag{3.3.45}$$

Este pode ser escrito na forma

$$\mathscr{A} \cdot \vec{x} = \vec{f} \, , \tag{3.3.46}$$

onde a matriz é dada por

$$\mathscr{A} \equiv \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix} \, , \tag{3.3.47}$$

e os vectores por

$$\vec{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} \qquad \text{e} \qquad \vec{f} \equiv \begin{bmatrix} f_1 \\ f_2 \\ ... \\ f_n \end{bmatrix} \, . \tag{3.3.48}$$

Para a resolução de sistemas de equações lineares é essencial utilizar os seguintes resultados;

→ Dois sistemas de equações lineares dizem-se equivalentes se possuirem o mesmo conjunto de soluções. Ou seja, se as soluções de ambos forem iguais.

→ É condição suficiente para dois sistemas

$$\mathscr{A} \cdot \vec{x} = \vec{f} \qquad \text{e} \qquad \mathscr{B} \cdot \vec{y} = \vec{g} \, , \tag{3.3.49}$$

serem equivalentes, que exista uma matriz $\mathscr{C}$ invertível, tal que

$$\mathscr{B} = \mathscr{C} \otimes \mathscr{A} \qquad \text{e} \qquad \vec{g} = \mathscr{C} \cdot \vec{f} \, . \tag{3.3.50}$$

Algumas das operações que verificam esta condição, são por exemplo;

→ Permutações de duas linhas. Se queremos, por exemplo, permutar as linhas 1 e 2, então;

$$\mathscr{C} \equiv \begin{bmatrix} 0 & 1 & 0 & 0 & ... \\ 1 & 0 & 0 & 0 & ... \\ 0 & 0 & 1 & 0 & ... \\ 0 & 0 & 0 & 1 & ... \\ ... & ... & ... & ... & ... \end{bmatrix} \, . \tag{3.3.51}$$

→ Multiplicação de uma linha por uma constante α (não nula). Se queremos multiplicar a linha 2 por α, então;

$$\mathscr{C} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & ... \\ 0 & \alpha & 0 & 0 & ... \\ 0 & 0 & 1 & 0 & ... \\ 0 & 0 & 0 & 1 & ... \\ ... & ... & ... & ... & ... \end{bmatrix} \, . \tag{3.3.52}$$

→ Soma de uma linha com o produto de outra por uma constante β. Se queremos substituir a linha 2 pela soma desta com o produto da linha 1 pela constante β, então;

$$\mathscr{C} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & ... \\ \beta & 1 & 0 & 0 & ... \\ 0 & 0 & 1 & 0 & ... \\ 0 & 0 & 0 & 1 & ... \\ ... & ... & ... & ... & ... \end{bmatrix} . \tag{3.3.53}$$

→ Etc ...

<u>*Método de Gauss*</u>: vamos então usar uma combinação destes tipos de operações para resolver o sistema pelo método de Gauss. Este consiste em reduzir o sistema de equações inicial (3.3.45) a um sistema equivalente que seja *triangular*. Isto é, partimos de

$$\mathscr{A} \cdot \vec{x} = \vec{f} \qquad \text{com} \qquad \mathscr{A} \equiv \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} \\ ... & ... & ... & ... & ... \\ a_{n1} & a_{n2} & a_{12} & ... & a_{nn} \end{bmatrix} , \tag{3.3.54}$$

para chegar a

$$\mathscr{B} \cdot \vec{x} = \vec{g} \qquad \text{com} \qquad \mathscr{B} \equiv \begin{bmatrix} b_{11} & b_{12} & b_{13} & ... & b_{1n} \\ 0 & b_{22} & b_{23} & ... & b_{2n} \\ 0 & 0 & b_{33} & ... & b_{3n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & b_{nn} \end{bmatrix} , \tag{3.3.55}$$

de forma que estes dois sistemas sejam equivalentes.

(*i*) o primeiro passo é naturalmente adicionar a cada uma das equações a partir da segunda e até à última (linha n), uma equação que resulta de multiplicar a primeira linha por uma constante, de forma a que se torne todos os a_{i1} (i=2,...,n) iguais a zero. Ou seja, temos de executar as seguintes operações:

$$i \in \{2,3,...,n\} : \quad \begin{cases} \beta_i = -\frac{a_{i1}}{a_{11}} \\ k \in \{1,2,...,n\} : \quad a_{ik} = a_{ik} + \beta_i \cdot a_{1k} \\ f_i = f_i + \beta_i \cdot f_1 \ . \end{cases} \tag{3.3.56}$$

Por exemplo, a nossa matrix $\mathscr{C}_2$ (no caso de i=2), é dada por

$$\mathscr{C} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & ... \\ -a_{21}/a_{11} & 1 & 0 & 0 & ... \\ 0 & 0 & 1 & 0 & ... \\ 0 & 0 & 0 & 1 & ... \\ ... & ... & ... & ... & ... \end{bmatrix} . \tag{3.3.57}$$

(*ii*) o passo seguinte será agora no sentido de eliminar (tornar zero) todos os elementos da coluna 2 a partir da linha 3; a_{2i} para $i \in \{3,...,n\}$. Para tal basta pegar no resultado da transformação feita em (*i*) e executar o mesmo tipo de operação que se defeniu em (3.3.56) mas agora a partir da linha 3 para toda a coluna 2. Mantendo os mesmos nomes, mas com os novos valores obtidos no passo anterior, temos que

$$i \in \{2,3,...,n\} : \quad \begin{cases} \beta_i = -\frac{a_{i1}}{a_{11}} \\ k \in \{1,2,...,n\} : \quad a_{ik} = a_{ik} + \beta_i \cdot a_{1k} \\ f_i = f_i + \beta_i \cdot f_1 \ . \end{cases} \tag{3.3.58}$$

(*iii*) os restantes passos consistem em aplicar o anterior para eliminar todos os elementos da matriz abaixo da diagonal, repetindo, com as necessárias adaptações, o procedimento dado em (3.3.58).

Todo o procedimento atrás descrito pode ser transcrito no seguinte algoritmo:

$$k{=}1,...,n{-}1 \quad \rightarrow \begin{cases} i{=}k{+}1,...,n \quad \rightarrow \begin{cases} \beta_{ik} = -\frac{a_{ik}}{a_{kk}} \\ f_i = f_i + \beta_{ik} f_k \\ a_{ik} = 0 \\ j{=}k{+}1,...,n \quad \rightarrow \begin{cases} a_{ij} = a_{ij} + \beta_{ik} a_{kj} \ , \end{cases} \end{cases} \end{cases} \tag{3.3.59}$$

que nos dá o sistema triangular de equações lineares equivalente ao inicial. Este sistema de equações (do tipo dado em 3.3.54) pode ser facilmente usado para encontrar a solução $\vec{x}$. Escrevendo o novo sistema de equações calculado (notar que embora se usem os mesmos simbolos, os valores foram entretanto alterados pelo procedimento 3.3.58);

$$\begin{array}{rrrrr} a_{11}x_1 & + a_{12}x_2 & + \ldots & + a_{1n}x_n & = f_1 \\ & a_{22}x_2 & + \ldots & + a_{2n}x_n & = f_2 \\ & & \ldots & \ldots & \ldots \\ & & & a_{nn}x_n & = f_n \ . \end{array} \tag{3.3.60}$$

Logo

$$\begin{aligned} x_n &= \frac{f_n}{a_{nn}} \\ x_{n-1} &= \frac{1}{a_{n-1,n-1}} \ (f_{n-1} - b_{n-1,n}x_n) \\ &\ldots \\ x_1 &= \frac{1}{a_{11}} \ (f_1 - a_{1n}x_n - a_{1n-1}x_{n-1} - \ldots - a_{12}x_2) \ . \end{aligned} \tag{3.3.61}$$

Este conjunto de operações corresponde ao seguinte procedimento;

$$\begin{cases} x_n = \frac{f_n}{a_{nn}} \\ k{=}n{-}1,...,1 \quad \rightarrow \begin{cases} x_k = \frac{1}{a_{kk}} \left(f_k - \sum_{j=k+1}^{n} a_{kj} x_j \right) \ . \end{cases} \end{cases} \tag{3.3.62}$$

Exemplo 3.3.5: Consideremos o seguinte sistema de equações lineares

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 9 \\ 8 \end{bmatrix} \ .$$

Usando o procedimento descrito em (3.3.56), temos de multiplicar a linha 1 por (-2) e somá-la à linha 2, bem como multiplicar a linha 1 por (-3) e somá-la à linha 3, de onde resulta que o sistema passa a ser dado por;

$$\begin{bmatrix} 1 & 2 & 3 \\ & -1 & -2 \\ & -2 & -8 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ -10 \end{bmatrix} \ .$$

Agora, resta multiplicar a linha 2 por (-2) e somá-la à linha 3, resultando finalmente uma sistema triangular dado por

$$\begin{bmatrix} 1 & 2 & 3 \\ & -1 & -2 \\ & & -4 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ -4 \end{bmatrix} \ .$$

Daqui obtemos então que

$$x_3 = 1 \quad \Rightarrow \quad x_2 = 1 \quad \Rightarrow \quad x_1 = 1 \ .$$

Foi assim encontrada a solução deste sistema de três equações lineares.

Exemplo 3.3.6: Vejamos agora o caso do seguinte sistema de equações lineares

$$\begin{bmatrix} 0.0003 & 1.246 \\ 0.4370 & -2.402 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.249 \\ 1.968 \end{bmatrix} ,$$

cuja solução sabemos ser exactamente x_1=10 e x_2=1. Vamos usar o procedimento descrito em (3.3.56), mantendo quatro algarismos significativos nos cálculos. Para tal temos de multiplicar a linha inicial por

$$\beta_{21} = -\frac{0.4370}{0.0003} = -1457.$$

e adicionar o resultado à segunda linha. Esta operação resulta no seguinte sistema triangular;

$$\begin{bmatrix} 0.0003 & 1.246 \\ & -1817. \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.249 \\ -1818. \end{bmatrix} .$$

Deste sistema obtemos agora que

$$x_2 = \frac{1818}{1817} = 1.001 ,$$

e, partindo deste valor, que

$$x_1 = \frac{1}{0.0003} (1.249 - 1.256 \times 1.001) = 6.667 .$$

Isto é, o resultado está *incorrecto* devido ao facto de os erros de arredondamento terem reduzido o número de algarismos significativos, no resultado para x_1, a zero.

Pivotagem parcial: por vezes se os elementos de uma coluna são valores que diferem em muitas ordens de grandeza é necessário ter o cuidado de controlar o efeito dos erros de arredondamento no resultado quando se tenta obter a solução numericamente. Isto pode ser feito recorrendo-se à pivotagem parcial; esta técnica consiste em cada fase da iteração descrita em (3.3.56) - isto é, para cada valor de k - devemos permutar com a linha k (linha de referência) a linha j_r, em que j_r é o valor de j=$k,\ldots,n$) para o qual o valor de $|a_{jk}|$ é máximo. Desta forma garantimos que ao eliminar todos os elementos da matriz nessa coluna minoramos o efeito dos erros de arredondamento, escolhendo o melhor (isto é, o menor) valor para $|\beta_{ik}|$.

O procedimento para reduzir o sistema de equações a um sistema triangular, passa assim a ser

$$k{=}1,\ldots,n{-}1 \quad \rightarrow \begin{cases} a)\ i{=}k,\ldots,n \quad \rightarrow \begin{cases} j_r = \text{valor de } j \text{ onde } |a_{jk}| \text{ é máximo} \end{cases} \\ b)\ \text{Trocar a linha } k \text{ do sistema de equações, pela linha } j_r \\ c)\ i{=}k{+}1,\ldots,n \quad \rightarrow \begin{cases} \beta_{ik} = -\frac{a_{ik}}{a_{kk}} \\ f_i = f_i + \beta_{ik} f_k \\ a_{ik} = 0 \\ j{=}k{+}1,\ldots,n \quad \rightarrow \begin{cases} a_{ij} = a_{ij} + \beta_{ik} a_{kj} . \end{cases} \end{cases} \end{cases} \tag{3.3.63}$$

Exemplo 3.3.7: Vejamos novamente o caso do seguinte sistema de equações lineares

$$\begin{bmatrix} 0.0003 & 1.246 \\ 0.4370 & -2.402 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.249 \\ 1.968 \end{bmatrix} ,$$

cuja solução sabemos ser exactamente x_1=10 e x_2=1. Se usarmos o procedimento descrito em (3.3.63), onde se inclui pivotagem parcial, e mais uma vez mantendo-se quatro algarismos significativos nos cálculos, o sistema triangular é agora;

$$\begin{bmatrix} 0.4370 & -2.402 \\ & 1.248 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.968 \\ 1.248 \end{bmatrix} ,$$

pois, tal como requerido na pivotagem, tivemos que permutar a linha 1 com a linha 2, visto ser esta a que tem o maior valor para $|a_{k1}|$. Para construir este novo sistema de equações usamos o valor (comparar com o Exemplo 3.3.6)

$$\beta_{12} = -\frac{0.0003}{0.4370} = -0.0006865 .$$

A solução que agora obtemos é

$$x_2 = \frac{1.248}{1.248} = 1.000 ,$$

e, usando este valor,

$$x_1 = \frac{1}{0.4370} \left(1.968 + 2.402 \times 1.000\right) = 10.00 .$$

Isto é, no cálculo da solução não se perdeu algarismos significativos pois usamos a linha com o maior coeficiente como linha de referência.

Sistemas tri-diagonais de equações lineares: este tipo de sistemas de equações lineares é obviamente um subconjunto daqueles que podem ser resolvidos pelo método de Gauss discutido acima. No entanto visto não ser necessário efectuar a maioria das operações descritas por um algoritmo do tipo dado em (3.3.59) ou (3.3.63), e sendo um tipo de sistemas que aparece frequentemente (como por exemplo no cálculo de splines), vamos aqui considerar a forma como se pode simplificar a determinação da solução.

Um sistema tri-diagonal pode então ser escrito como

$$\begin{array}{cccccccc} a_1x_1 & +b_1x_2 & & & & & & = f_1 \\ c_2x_1 & +a_2x_2 & +b_2x_3 & & & & & = f_2 \\ & c_3x_2 & +a_3x_3 & +b_3x_4 & & & & = f_3 \\ & & \ldots & \ldots & \ldots & & & \ldots \\ & & & c_{n-1}x_{n-2} & +a_{n-1}x_{n-1} & +b_{n-1}x_n & & = f_{n-1} \\ & & & & c_nx_{n-1} & +a_nx_n & & = f_n , \end{array} \tag{3.3.64}$$

cuja resolução é agora simplificada para um procedimento do tipo

$$k{=}2,\ldots,n \quad \rightarrow \begin{cases} \beta_k = -\frac{c_k}{a_{k-1}} \\ c_k = 0 \\ a_k = a_k + \beta_k b_{k-1} \\ f_k = f_k + \beta_k f_{k-1} . \end{cases} \tag{3.3.65}$$

Após o qual, o sistema de equações passa a ser triangular;

$$\begin{array}{cccccccc} a_1x_1 & +b_1x_2 & & & & & & = f_1 \\ & +a_2x_2 & +b_2x_3 & & & & & = f_2 \\ & & a_3x_3 & +b_3x_4 & & & & = f_3 \\ & & & \ldots & \ldots & & & \ldots \\ & & & & a_{n-1}x_{n-1} & +b_{n-1}x_n & & = f_{n-1} \\ & & & & & a_nx_n & & = f_n , \end{array} \tag{3.3.66}$$

Sendo a solução calculada simplesmente de acordo com

$$\begin{cases} x_n = \frac{f_n}{a_n} \\ k{=}n{-}1,...,1 \end{cases} \rightarrow \left\{ x_k = \frac{f_k - b_k x_{k+1}}{a_k} \right. . \tag{3.3.67}$$

Exemplo 3.3.8: Consideremos o seguinte sistema tri-diagonal de equações lineares

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 3 & 4 \\ 0 & 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 9 \\ 5 \end{bmatrix} .$$

De forma a usar o procedimento descrito em (3.3.64), temos de definir

$$\vec{a} = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 2 \\ 4 \\ - \end{bmatrix} \qquad \vec{c} = \begin{bmatrix} - \\ 2 \\ 4 \end{bmatrix} .$$

Após aplicarmos o procedimento descrito em (3.3.64) ficamos com

$$\vec{a} = \begin{bmatrix} 1 \\ -1 \\ 17 \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 2 \\ 4 \\ - \end{bmatrix} \qquad \vec{f} = \begin{bmatrix} 3 \\ 3 \\ 17 \end{bmatrix} ,$$

de onde resulta que

$$x_3 = \frac{17}{17} = 1 \quad \Rightarrow \quad x_2 = \frac{3-1\times 4}{-1} = 1 \quad \Rightarrow \quad x_1 = \frac{3-1\times 2}{1} = 1 .$$

Foi assim encontrada a solução deste sistema tri-diagonal de três equações lineares.

3.4 Outras funções interpoladoras

Claramente nem sempre será adequado para descrever a função $f(x)$ tabela o uso de polinómios, ou funções que recorrem a polinómios. Assim podemos definir de uma forma mais geral a função interpoladora de uma tabela $\{(x_i, f_i)\}_{i=0}^n$, cujas abcissas estão num intervalo $[a,b]$, como sendo uma combinação linear de funções padrão $\phi_i(x)$;

$$y(x) = \sum_{i=0}^{n} a_i \, \phi_i(x) . \tag{3.4.1}$$

O caso discutido na Secção 3.2 corresponde naturalmente a considerar que $\phi_i(x) \equiv x^i$, definindo-se assim um polinómio de grau n. No entanto há várias outras opções que podem ser feitas.

Após definirmos o conjunto de $n{+}1$ funções padrão, basta emtão impôr as condições que exigem que $y(x)$ seja uma função interpoladora para determinar os coeficientes da combinação linear, ou seja, que

$$f_j = \sum_{i=0}^{n} a_i \, \phi_i(x_j) \qquad \text{para} \quad j = 0,1,2,...,n . \tag{3.4.2}$$

Temos assim um sistema de $n{+}1$ equações para as incógnitas a_i. O sistema será bem definido, tendo uma solução única, se as funções $\phi_i(x)$ são linearmente independendentes para a malha $\{x_i\}_{i=0}^n$.

Duas funções $\phi_i(x)$ e $\phi_k(x)$, dizem-se que *não são linearmente independentes* relativamente aos pontos $\{x_i\}_{i=0}^n$, se

$$\exists C \in \mathscr{R} \; ; \; \phi_i(x_j) = C \, \phi_k(x_j) \quad \text{para todo o } j{=}0,1,...,n . \tag{3.4.3}$$

Sendo as funções padrão linearmente independentes sabemos então que o sistema (3.4.2) têm solução única que pode ser determinada resolvendo o sistema de equações. Este pode ser escrito na forma matricial como sendo

$$\begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & ... & \phi_n(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \phi_2(x_1) & ... & \phi_n(x_1) \\ ... & ... & ... & ... & ... \\ \phi_0(x_n) & \phi_1(x_n) & \phi_2(x_n) & ... & \phi_n(x_n) \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ ... \\ f_n \end{bmatrix} . \tag{3.4.4}$$

Cuja solução (a_i) nos permite construir a função interpoladora (3.4.1).

Exemplo 3.4.1: Consideremos a seguinte tabela de três pontos de uma função $f(x)$ que sabemos ser periódica; $\{(0,0),(\pi/2,1),(\pi,3)\}$. Pretende-se estimar por interpolação o valor que a função toma em $x{=}\pi/4$. Face às propriedades de $f(x)$, podemos escrever a função interpoladora, recorrendo a (3.4.1) como sendo

$$y(x) = a_0\cos(x) + a_1\sin(x) + a_3\cos(2x)\ .$$

Introduzimos assim funções padrão $\phi_j(x)$ periódicas. Podemos então impôr as condições de interpolação obtendo o seguinte sistema (ver 3.4.4) de três equações, a três incógnitas,

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}\ ,$$

cuja solução é $a_0{=}-3/2$, $a_1{=}5/2$ e $a_2{=}3/2$. Temos então a expressão da função interpoladora que podemos usar para estimar $f(\pi/4){\simeq}y(\pi/4){=}\sqrt{2}/2$.

Exemplo 3.4.2: Consideremos agora uma tabela de pontos $\{(0,0),(1,1),(2,4)\}$ que sabemos pertencer a uma função $f(x)$ cujo comportamento é esperado ser próximo do tipo exponencial. Mais uma vez precisa-se de estimar por interpolação o valor que a função toma em $x{=}1/2$. Face às propriedades de $f(x)$, podemos escrever a função interpoladora, recorrendo a (3.4.1), como sendo

$$y(x) = a_0\ \mathrm{e}^{x} + a_1\ \mathrm{e}^{-x} + a_3\ \mathrm{e}^{2x}\ .$$

Introduzimos assim funções padrão $\phi_j(x)$ que são exponenciais. Podemos então impôr as condições de interpolação obtendo o seguinte sistema (ver 3.4.4) de três equações, a três incógnitas,

$$\begin{bmatrix} 1 & 1 & 1 \\ \mathrm{e} & 1/\mathrm{e} & \mathrm{e}^2 \\ \mathrm{e}^2 & 1/\mathrm{e}^2 & \mathrm{e}^4 \end{bmatrix} \times \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \end{bmatrix}\ ,$$

cuja solução é $a_0{=}0.3422$, $a_1{=}{-}0.3701$ e $a_2{=}0.0279$. Temos então a expressão da função interpoladora que usamos para estimar $f(1/2){\simeq}y(1/2){=}0.4156$.

No entanto não existem apenas funções interpoladores que são combinações lineares (tal como dado em 3.4.1) de funções padrão. Podemos também considerar funções de $n{+}1$ parâmetros, mas cuja dependência nestes é não linear. Claramente este tipo de funções interpoladoras são mais difíceis de determinar, sendo na maior parte dos casos (quando temos muitos pontos a interpolar) inviável faze-lo.

Exemplo 3.4.3: Consideremos agora a seguinte tabela de pontos $\{(0,1),(1,2)\}$ que sabemos pertencer a uma função $f(x)$ do tipo exponencial. Calculemos então a função interpoladora $y(x)$ da forma

$$y(x) = a_0\ \mathrm{e}^{a_1\,x}\ .$$

Impondo as condições de interpolação temos o seguinte sistema de equações (não lineares);

$$\begin{cases} a_0 = 1 \\ a_0\ \mathrm{e}^{a_1} = 2\ . \end{cases}$$

Cuja solução corresponde a $a_0{=}1$ e $a_1{=}\log 2$. Assim, a função interpoladora encontrada é dada por

$$y(x) = 2^x\ .$$

Podemos então estimar o valor de $f(1/2)$, calculando $f(1/2){\simeq}y(1/2){=}\sqrt{2}$.

3.5 Exercícios

E3.1) Dada a tabela calcule o valor de $\tan(\pi/5)$ usando interpolação linear e interpolação parabólica, indicando o correspondente erro de interpolação, pelos métodos de,

a) Lagrange.

b) Aitken-Neville.

x	$\tan x$
0	0
$\pi/6$	0.57735
$\pi/4$	1

E3.2)* Proponha um algoritmo que permita o cálculo, num ponto x, do valor do polinómio de grau n que interpola a tabela $\{x_i, f_i\}_{i=0,\ldots,n}$, pela fórmula de Lagrange.

E3.3) Considere a seguinte tabela da função $f(x)=1/x$. Determine o polinómio de menor grau que permite calcular o valor de $f(2.10)$ com 4 casas decimais significativas e calcule o correspondente valor aproximado de $f(2.10)$.

x	$f(x)$
2.09	0.478469
2.11	0.473934
2.13	0.469484
2.15	0.465116

E3.4)* Proponha um algoritmo que permita o cálculo, num ponto x, do valor do polinómio de grau n que interpola a tabela $\{x_i, f_i\}_{i=0,\ldots,n}$, pelo método de Aitken-Neville.

E3.5)* Implementando o algoritmo desenvolvido na pergunta anterior, utilize a informação toda da seguinte tabela de forma a calcular o valor de $f(1.11)$. Indique um majorante de $|f^{(7)}(\xi)|$, com $\xi \in [1.00, 1.30]$, de forma que o erro no cálculo do valor em 1.11 seja dominado apenas pelo erro de arredondamento com que os valores de $f(x)$ são dados na tabela.

x	$f(x)$
1.00	1.017452
1.05	0.971622
1.10	0.930208
1.15	0.892646
1.20	0.858464
1.25	0.827269
1.30	0.798724

E3.6) Sabe-se que a função $f(x)$ tem um só zero no intervalo $]0,1[$. A partir da tabela dada, determine por interpolação inversa um valor aproximado da raíz com erro inferior a 10^{-3}, sabendo que para esse intervalo se têm $\left|\{f^{-1}\}^{(n)}(x)\right| < 0.1$.

x	$f(x)$
0.08	-0.53213
0.10	-0.31259
0.12	-0.13466
0.14	$+0.01429$
0.16	$+0.14182$

E3.7)* Proponha um algoritmo que permita o cálculo, num ponto x, do valor do polinómio de grau n que interpola a tabela $\{x_i, f_i\}_{i=0,\ldots,n}$, pela fórmula de Newton.

E3.8)* A partir da tabela ao lado estime $\sqrt{2.15}$ usando o método de Newton e determine um limite superior do erro.

x	$\sqrt{x}$
2.0	1.414214
2.1	1.449138
2.2	1.483240
2.3	1.516675
2.4	1.549193

E3.9) Conhecem-se os seguintes valores de uma função real $f(x)$ em três pontos: $f(x_0{=}-1){=}0.5$, $f(x_1{=}0){=}0.0$

e $f(x_2{=}1){=}-0.1$. Estime por interpolação polinomial (usando a fórmula de Newton) o valor da função $f(x)$ em $x{=}0.4$.

E3.10) Conhecem-se os seguintes valores de uma função real $f(x)$ em três pontos; $f(0.0){=}-0.5$, $f(0.8){=}0.0$ e $f(2.0){=}1.0$, bem como um valor da sua derivada; $f'(2.0){=}0.0$. Estime, recorrendo a uma spline quadrática, e usando toda a informação dada, o valor de $f(1)$. Esboce o gráfico da derivada da spline interpoladora calculada.

E3.11) Conhecem-se os seguintes valores tabelados de uma função $f(x)$ nos seguintes pontos: $f(x_0{=}0){=}0.0$, $f(x_1{=}1){=}1.5$, $f(x_2{=}2){=}0.5$ e $f(x_3{=}3){=}0.0$.

a) Estime por interpolação, usando uma spline cúbica natural, o valor da função em $x{=}1.5$ usando todos os pontos fornecidos.

b) Que valor de $f'(1.8)$ se obtém se usarmos a spline da alinea anterior para o cálculo da derivada?

E3.12) Construa a Spline quadrática $S(x)$ que interpola os pontos da tabela ao lado, tal que $m_0{-}m_3{=}1$. Obtenha o valor de $S(1.5)$. Justifique graficamente a diferença para $S(1.5)$ no caso de usar $m_0{=}0$.

x_i	f_i
0	1.1
1	1.5
2	0.5
3	1.0

E3.13) Construa a Spline cúbica natural que interpola os pontos da tabela dada na pergunta anterior. Qual o valor que esta tem em $x{=}1.5$? Se em vez de considerar o caso de esta ser natural, usar as condições $M_0{=}1$ e $M_1{=}M_2$ que valor se obtém para o mesmo ponto?

E3.14)* Recorrendo a uma Spline cúbica natural que interpola os pontos da tabela dada ao lado, estime o valor da função f em $x{=}1.5$?

x	f
0.0	1.41
1.1	3.44
1.6	2.48
2.2	3.51
3.4	0.54

E3.15) Conhecem-se os seguintes valores tabelados de uma função $f(x)$ nos seguintes pontos: $f(x_0{=}1){=}0$, $f(x_1{=}2){=}2$, $f(x_2{=}3){=}0$ e $f(x_3{=}4){=}4$. Sabe-se ainda que para esta função se têm que $f''(x_1){=}1$ e $f''(x_2){=}0$. Estime por interpolação (usando toda a informação fornecida sobre a função), através de uma spline cúbica, o valor da função $f(x)$ em $x{=}{}^7\!/_2$.

E3.16) Conhecem-se os dois valores de uma função $f(x)$ nos seguintes pontos: $f(x_0{=}1){=}0$, $f(x_1{=}2){=}2$. Estime por interpolação, o valor da função $f(x)$ em $x{=}{}^3\!/_2$ considerando que a função interpoladora é representada por;

a) $y(x) \equiv a_0 \cos x + a_1 e^x$;

b) $y(x) \equiv \dfrac{2 + a_0 \cos x}{1 + a_1 x}$.

Capítulo 4

Aproximação numérica

Por vezes temos informação sobre uma função na forma de vários pontos, que são apenas indicativos do comportamento da função. Isto é, porque têm um erro de medida associado por exemplo, não os podemos considerar como sendo pontos da função mas apenas indicativos do seu comportamento. Daí que num caso deste tipo não faz sentido recorrer a interpolação para estimar o valor da função num ponto. Em vez disso considera-se um *função aproximadora* que tenta, através de um critério pré estabelecido, representar da *melhor forma* o comportamento dos pontos tabelados. Nesta secção vamos considerar algumas formas de calcular essa função aproximadora de forma a estimar o comportamento da função para a qual apenas dispomos de uma tabela de valores, e que não são necessáriamente pontos dela.

4.1 Função aproximadora

Por função aproximadora vamos representar a função que segundo um critério de escolha bem definido seja a *melhor* função que representa uma tabela de pontos $\{(x_i, f_i)\}_{i=0}^{n}$. Para tal temos primeiro que definir o tipo de função (um polinómio de grau m, por exemplo, com $m \ll n$), e depois o critério que nos permita escolher entre as várias funções de um tipo qual é delas a que melhor representa a tabela de valores.

Seja $y(x)$ a função aproximadora. Então um critério possível de escolha seria dizer que a melhor função $y(x)$ é aquela que minimiza o valor de

$$\mathscr{R} = \max_{i=0,1,\ldots,n} |y(x_i) - f_i| \, . \tag{4.1.1}$$

Se $y(x)$ é por exemplo um polinómio de grau 1, então a condição (4.1.1) permite-nos determinar os dois parâmetros que definem a recta, de forma que esta seja a melhor recta que representa os pontos de acordo com o critério estabelecido. Mas através de um método de construção de $y(x)$ como descrito aqui, a função aproximadora não passa necessariamente pelos pontos da tabela (ver Fig. 4.1).

Outro critério possível, e mais frequentemente usado, corresponde a encontrar a função $y(x)$ que minimiza

$$\mathscr{R} = \sum_{i=0}^{n} |y(x_i) - f_i|^2 \, . \tag{4.1.2}$$

Tal deve ser usado, por exemplo quando os valores f_i têm associado um erro de medida que obedece a uma distribuição gaussiana em torno do valor exacto. Sendo por isso o critério mais comumentemente usado em Astronomia, Física ou outras ciências.

Vejamos agora de que forma este critério pode ser usado para encontrar a melhor recta, por exemplo. Temos então que a função aproximadora é do tipo

$$y(x) = a_0 + a_1 x \, . \tag{4.1.3}$$

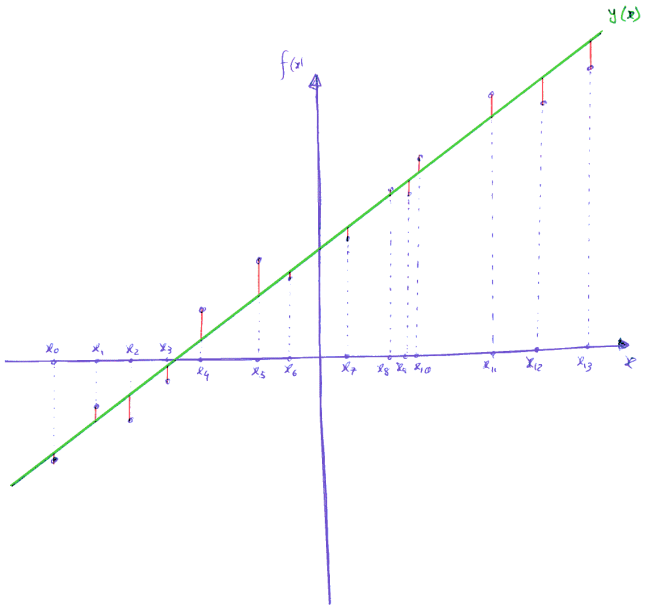


Figura 4.1: Representação de uma recta como função aproximadora de uma tabela de valores para uma função $f(x)$. Esta recta foi determinada usando um critério do tipo dado em (5.1.2).

Logo formalmente a condição (4.1.2) corresponde a encontrar os valores de a_0 e a_1 que são um mínimo de

$$\mathscr{R} \equiv \mathscr{R}(a_0, a_1) = \sum_{i=0}^{n} |(a_0 + a_1 x_i) - f_i|^2 \,. \tag{4.1.4}$$

Os valores são então encontrados obtendo as soluções de

$$\frac{\partial \mathscr{R}}{\partial a_0} = 0 \qquad \text{e} \qquad \frac{\partial \mathscr{R}}{\partial a_1} = 0 \,, \tag{4.1.5}$$

pois queremos o mínimo de $\mathscr{R}$. Assim, temos duas condições que nos permitem calcular os parâmetros;

$$\begin{cases} \displaystyle\sum_{i=0}^{n} [(a_0 + a_1 x_i) - f_i] = 0 \\ \displaystyle\sum_{i=0}^{n} x_i \, [(a_0 + a_1 x_i) - f_i] = 0 \,, \end{cases} \tag{4.1.6}$$

que pode ser escrito como

$$\begin{cases} a_0 \displaystyle\sum_{i=0}^{n} 1 + a_1 \sum_{i=0}^{n} x_i = \sum_{i=0}^{n} f_i \\ a_0 \displaystyle\sum_{i=0}^{n} x_i + a_1 \sum_{i=0}^{n} x_i^2 = \sum_{i=0}^{n} x_i f_i \,. \end{cases} \tag{4.1.7}$$

A solução destas duas equações lineares permite-nos definir a recta que melhor aproxima os pontos. A solução é única se

$$\begin{vmatrix} \displaystyle\sum_{i=0}^{n} 1 & \displaystyle\sum_{i=0}^{n} x_i \\ \displaystyle\sum_{i=0}^{n} x_i & \displaystyle\sum_{i=0}^{n} x_i^2 \end{vmatrix} \neq 0 \,. \tag{4.1.8}$$

Temos toda a liberdade de escolher a função aproximadora, pelo que deve ser tomado em conta o tipo de comportamento esperado para os pontos. Vamos considerar, por exemplo, uma função aproximadora do género

$$y(x) \equiv a_0 \, x^5 \,. \tag{4.1.9}$$

Então neste caso temos de minimizar

$$\mathscr{R}(a_0) = \sum_{i=0}^{n} \left(a_0 x_i^5 - f_i \right)^2 \,. \tag{4.1.10}$$

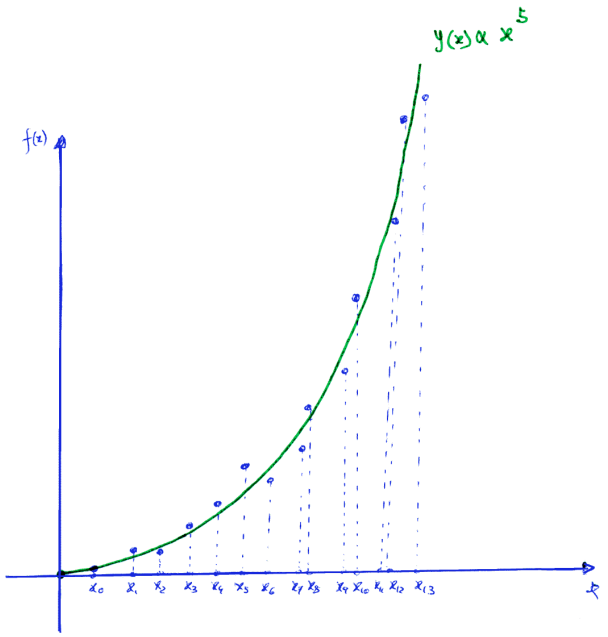


Figura 4.2: Representação da melhor função aproximadora do tipo $y(x) \propto x^5$ para um conjunto de pontos $\{(x_i, f_i)\}_{i=0}^{n}$.

O mínimo de $\mathscr{R}$ corresponde a

$$\frac{\partial \mathscr{R}}{\partial a_0} = 2\sum_{i=0}^{n} x_i^5 \left(a_0 x_i^5 - f_i\right) = 0\,, \tag{4.1.11}$$

cuja solução é

$$a_0 = \frac{\sum_{i=0}^{n} x_i^5 f_i}{\sum_{i=0}^{n} x_i^{10}}\,. \tag{4.1.12}$$

Dada a tabela, podemos calcular a_0, logo temos a função aproximadora (ver Fig. 4.2).

Consideremos agora o caso geral de termos um função aproximadora que é escrita como uma combinação linear de funções de referência $\phi_j(x)$;

$$y(x) = \sum_{j=0}^{m} a_j\, \phi_j(x)\,, \tag{4.1.13}$$

com que pretendemos aproximar $n{+}1$ valores da tabela $\{x_i, f_i\}_{i=0}^{n}$, usando um dos critérios de escolha da melhor aproximação. Esse critério vai permitir-nos calcular os m valores dos coeficientes a_j construindo assim a função aproximadora. Para o fazer precisamos primeiro de arranjar bases de funções referência $\phi_j(x)$ que sejam linearmente independentes bem como de definir os critérios de selecção da melhor aproximação.

Note que se o número de pontos $n{+}1$ é muito superior ao número de parâmetros $m{+}1$ que definem a função aproximadora, esta não passará na maioria deles, isto é, os pontos da tabela não serão pontos da função aproximadora.

Mas no caso limite de termos $n{=}m$ então a função interpoladora (que para $n{+}1$ pontos é definida por $n{+}1$ parâmetros) terá $\mathscr{R}{=}0$. Logo, a função interpoladora é a melhor função aproximadora neste limite.

4.2 Método dos Mínimos Quadrados

O critério de selecção da melhor função aproximadora é neste caso dado pela escolha da função que minimiza a soma dos quadrados dos desvios, isto é, que corresponde ao mínimo de

$$\mathscr{R} \equiv \sum_{i=0}^{n} \Big[y(x_i) - f_i \Big]^2 . \tag{4.2.1}$$

Sendo a função dada em geral por uma expressão do tipo (4.1.13), temos que

$$\mathscr{R}(a_0, a_1, a_2, ..., a_m) = \sum_{i=0}^{n} \left[\sum_{j=0}^{m} a_j \, \phi_j(x_i) - f_i \right]^2 . \tag{4.2.2}$$

Assim, o mínimo desta função corresponde aos valores de a_0, a_1, ..., a_m que são determinados por

$$\frac{\partial \mathscr{R}}{\partial a_k} = 0 \qquad \text{para} \quad k = 0, 1, ..., m . \tag{4.2.3}$$

Estas $m+1$ condições correspondem a

$$\begin{aligned} & \frac{\partial \mathscr{R}}{\partial a_k} = 2 \sum_{i=0}^{n} \left\{ \phi_k(x_i) \Big[\sum_{j=0}^{m} a_j \, \phi_j(x_i) - f_i \Big] \right\} = 0 \\ \Rightarrow \quad & \sum_{j=0}^{m} \left\{ a_j \sum_{i=0}^{n} \Big[\phi_k(x_i) \, \phi_j(x_i) \Big] \right\} = \sum_{i=0}^{n} \Big[\phi_k(x_i) \, f_i \Big] , \end{aligned} \tag{4.2.4}$$

para k=0, 1, ..., m. Temos assim que resolver o seguinte sistema de $m+1$ equações, a $m+1$ incógnitas (a_j), dado por

$$\mathscr{B} \cdot \vec{a} = \vec{g} , \tag{4.2.5}$$

com

$$b_{kj} = b_{jk} = \sum_{i=0}^{n} \Big[\phi_k(x_i) \, \phi_j(x_i) \Big] ; \quad \vec{a} = \begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{bmatrix} ; \quad \text{e} \quad g_k = \sum_{i=0}^{n} \Big[\phi_k(x_i) \, f_i \Big] . \tag{4.2.6}$$

Precisamos agora definir quais as funções referência $\phi_k(x)$ que vamos usar, de forma a podermos usar este sistema de equações para determinar a melhor função aproximadora pelo método dos mínimos quadrados.

4.2.1 Aproximação por monómios

Como já vem sendo usual comecemos pelo mais fácil, que corresponde a usar polinómios como funções de referência. Consideremos então que

$$\phi_k(x) \equiv x^k , \tag{4.2.7}$$

correspondendo à base de monómios geradora de polinómios. A nossa função aproximadora é então um polinómio de grau m, dado por

$$y(x) = \sum_{j=0}^{m} a_j x^j . \tag{4.2.8}$$

Neste caso temos que (ver 4.2.6),

$$g_k = \sum_{i=0}^{n} \Big(x_i^k \, f_i \Big) \qquad \text{e} \qquad b_{kj} = \sum_{i=0}^{n} x_i^{k+j} . \tag{4.2.9}$$

É fácil concluir que mesmo para valores baixos de M os termos b_{kj} podem ter valores de ordem de grandeza muito diversa. Surge-nos assim um problema similar que nos levou a introduzir *pivotagem* quando discutimos a resolução numérica de sistemas de equações na Secção 3.3.3. Embora aqui isso não baste pois caso o número n de pontos seja elevado a precisão no cálculo dos parâmetros a_j torna-se rapidamente muito baixa.

Exemplo 4.2.1: Consideremos a seguinte tabela de pontos

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
f_i :	2.0	2.2	2.3	4.1	5.5	7.0

que queremos aproximar por uma parábola, logo temos que usar $\phi_0(x)$=1, $\phi_1(x)$=x e $\phi_2(x)$=x^2;

$$y(x) = a_0 + a_1 x + a_2 x^2 \, .$$

O sistema de equações dado em (4.2.5) é então,

$$\begin{bmatrix} 6 & 17 & 56.5 \\ 17 & 56.5 & 206.75 \\ 56.5 & 206.75 & 803.125 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 23.1 \\ 77.95 \\ 291.825 \end{bmatrix} \, .$$

Notar a diferença de ordem de grandeza dos elementos destas matrizes. Resolvendo o sistema, obtemos que

$$y(x) = 2.4446 - 0.9697x + 0.4410x^2 \, ,$$

é a parábola que melhor aproxima a tabela de pontos.

4.2.2 Aproximação por bases ortonormais de polinómios

A solução é considerar bases geradores de polinómios que sejam ortogonais e normalizadas de forma que a matrix $\mathscr{B}$ gerada a partir dessas bases de polinómios, tenha elementos da mesma ordem de grandeza.

Para tal podemos usar uma fórmula de recorrência que nos permita gerar polinómios nas condições necessárias. Primeiro definimos produto interno de duas funções, u e v, num intervalo Ω=$[a,b]$ relativamente à função peso $\omega(x)$ (definida positiva nesse intervalo):

$$(u,v) \equiv \int_a^b u(x) \cdot v(x) \quad \omega(x) \, \mathrm{d}x \, . \tag{4.2.10}$$

Diz-se então que duas funções, $u_1(x)$ e $u_2(x)$, são ortogonais no intervalo Ω, relativamente à função peso $\omega(x)$, se

$$(u_1, u_2) = 0 \, . \tag{4.2.11}$$

Podemos ainda introduzir o conceito de norma de uma função como sendo o valor de

$$||u|| \equiv \sqrt{(u,u)} \, . \tag{4.2.12}$$

Fórmula de recorrência para gerar polinómios ortogonais: se $p_k(x)$, para k=$0,1,...$, é uma família de polinómios ortogonais em que o grau de $p_k(x)$ é k, e se α_k é o coeficiente do termo de ordem k (isto é, de x^k) de $p_k(x)$ então os polinómios verificam a seguinte relação de recorrência:

$$p_{k+1}(x) = A_k(x - B_k) \cdot p_k(x) - C_k \cdot p_{k-1}(x) \qquad \text{para} \quad k = 0,1,2,... \, , \tag{4.2.13}$$

onde $p_{-1}(x)$≡0 e

$$\begin{aligned} A_k &= \frac{\alpha_{k+1}}{\alpha_k} \\ B_k &= \frac{(x\, p_k, p_k)}{||p_k||^2} \\ C_k &= \frac{A_k \, ||p_k||^2}{A_{k-1} \, ||p_{k-1}||^2} \, . \end{aligned} \tag{4.2.14}$$

Assim, para definirmos uma família de polinómios ortogonais basta-nos definir o intervalo Ω, a função $\omega(x)$ e os valores de α_k (normalização).

Polinómios de Legendre: um exemplo são os frequentemente usados polinómios de Legendre que são definidos escolhendo-se;

$$\begin{cases} \omega(x) = 1 \\ \Omega = [-1,1] \\ \alpha_k \quad \text{é tal que} \quad p_k(1) = 1 \, . \end{cases} \tag{4.2.15}$$

Como $p_0(x)$ é constante (grau "0") então temos que

$$p_0(x) = 1 \,. \tag{4.2.16}$$

Logo segue (de 4.2.13) que

$$p_1(x) = A_0(x - B_0)p_0(x) \,. \tag{4.2.17}$$

Recorrendo a (4.2.14) obtemos que

$$||p_0||^2 = \int_{-1}^{1} \mathrm{d}x = 2 \qquad \text{e} \qquad (x\, p_0, p_0) = \int_{-1}^{1} x p_o^2(x)\mathrm{d}x = \int_{-1}^{1} x\mathrm{d}x = 0 \,, \tag{4.2.18}$$

pelo que B_0=0. Finalmente temos que A_0=α_1, dando que

$$p_1(x) = \alpha_1 x \,. \tag{4.2.19}$$

Impondo a condição de normalização, $p_1(1)$=1, chegamos a

$$p_1(x) = x \,. \tag{4.2.20}$$

Usando mais uma vez (4.2.13), segue-se que

$$p_2(x) = A_1(x - B_1)\, p_1(x) - C_1\, p_0(x) \,. \tag{4.2.21}$$

Como

$$A_1 = \alpha_2 \tag{4.2.22}$$

$$||p_1||^2 = \int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3} \qquad \text{e} \qquad (xp_1, p_1) = \int_{-1}^{1} x^3\mathrm{d}x = 0 \Rightarrow B_1 = 0$$

$$||p_0||^2 = 2$$

$$C_1 = \frac{A_1||p_1||^2}{A_0||p_0||^2} = \frac{\alpha_2}{3} \,, \tag{4.2.23}$$

obtemos que

$$p_2(x) = \alpha_2 \left(x^2 - \frac{1}{3}\right) \,. \tag{4.2.24}$$

De $p_2(1)$=1 temos que α_2=3/2, pelo que encontramos finalmente o polinómio de Legendre de grau 2;

$$p_2(x) = \frac{3x^2 - 1}{2} \,. \tag{4.2.25}$$

Podemos assim gerar por recorrência todos os polinómios de Legendre, de forma a construir uma base de funções geradora de qualquer polinómio de grau m.

<u>*Fórmula de Rodrigues*</u>: esta é outra forma de gerar os polinómios de Legendre. É escrita na forma,

$$p_k(x) = \frac{1}{2^k\, k!}\, \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(x^2 - 1\right)^k \qquad k \geq 1 \,, \tag{4.2.26}$$

com $p_0(x)$=1. Claramente esta fórmula pode ser relacionada com (4.2.13), notando que desta expressão chegamos a

$$p_k(x) = \frac{(2k+1)\, x\, p_k(x) - k\, p_{k-1}(x)}{k+1} \qquad k = 0, 1, \ldots \,, \tag{4.2.27}$$

logo,

$$A_k = \frac{2k+1}{k+1} \,, \qquad B_k = 0 \,, \qquad \text{e} \qquad C_k = \frac{k}{k+1} \,, \tag{4.2.28}$$

pois

$$||p_k||^2 = \frac{2}{2k+1} \qquad \text{e} \qquad \alpha_k = \frac{(2k)!}{2^k\, (k!)^2} \,. \tag{4.2.29}$$

<u>*Polinómios de Chebyshev (primeira espécie)*</u>: outra família de polinómios ortogonais usada em diferentes situações é gerada considerando que

$$\begin{cases} \omega(x) = \dfrac{1}{\sqrt{1-x^2}} \\ \Omega = [-1, 1] \\ T_k(1) = 1 \,. \end{cases} \tag{4.2.30}$$

Logo o produto interno está definido por

$$(T_k, T_j) = \int_{-1}^{1} \frac{T_k(x) T_j(x)}{\sqrt{1-x^2}} \, \mathrm{d}x \,, \tag{4.2.31}$$

dando que

$$\begin{aligned} &T_0(x) = 1 \\ &T_1(x) = x \\ &T_2(x) = 2x^2 - 1 \\ &T_3(x) = 4x^3 - 3x \\ &\ldots \end{aligned} \tag{4.2.32}$$

A fórmula de recorrência (4.2.13) corresponde neste caso a

$$T_{k+1}(x) = 2x \, T_k(x) - T_{k-1}(x) \qquad k = 1, 2, \ldots \,, \tag{4.2.33}$$

tendo-se que $|T_k(x)| \leq 1$ $(\forall_{k=0,1,2,\ldots})$. Os polinómios podem ser alternativamente escritos como

$$T_k(x) = \cos\Big[k \cdot \arccos(x)\Big] \,. \tag{4.2.34}$$

Logo temos também que

$$(T_k, T_j) = \int_0^{\pi} \cos(k\theta) \cdot \cos(j\theta) \, \mathrm{d}\theta \,, \tag{4.2.35}$$

como sendo o produto interno entre dois polinómios de Chebyshev.

Tal como já aconteceu com os polinómios de Legendre, também aqui temos que as funções geradoras da função aproximadora estão definidas para um intervalo $\Omega = [-1, 1]$. Logo, caso a nossa tabela seja dada para pontos x_i num intervalo $[a, b]$, é necessário converter os pontos ao intervalo Ω de forma a podermos usar uma função aproximadora do tipo

$$y(x) = \sum_{j=0}^{m} a_j \, T_j(x) \,, \tag{4.2.36}$$

determinando os coeficientes a_j. A renormalização da tabela é feita considerando que o método dos Mínimos Quadrados deve ser aplicado aos pontos $\{(t_i, f_i)\}_{i=0}^{n}$, tais que

$$t_i = \frac{2x_i - (b+a)}{b-a} \qquad \text{bem como} \qquad x_i = \frac{(b-a)t_i + (b+a)}{2} \,. \tag{4.2.37}$$

Estas expressões permitem-nos converter a tabela, bem como a função aproximadora encontrada para t_i na função para os x_i, tal como pretendido.

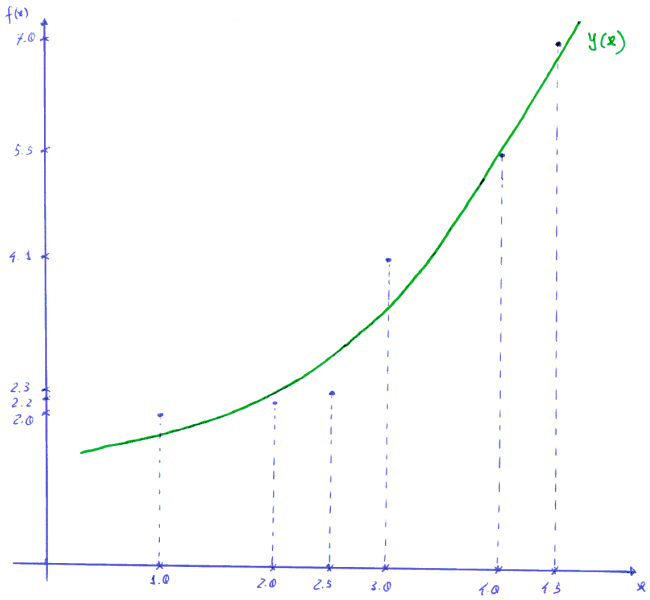


Figura 4.3: Representação da melhor parábola aproximadora, tal como calculada pelo método dos Mínimos Quadrados, de uma tabela de valores para uma função $f(x)$ (ver Exemplo 4.2.2).

Exemplo 4.2.2: Consideremos a seguinte tabela de pontos

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
f_i :	2.0	2.2	2.3	4.1	5.5	7.0

que queremos aproximar por uma parábola, logo temos que usar $T_0(t)=1$, $T_1(t)=t$ e $T_2(t)=2t^2-1$. Mas como os polinómios de Chebyshev estão definidos para o intervalo $[-1,1]$, é necessário normalizar a tabela de valores. Como $x_i \in [1,4.5]$, basta-nos escrever $x=(3.5t+5.5)/2$ pelo que $t=(2x-5.5)/3.5$, passando a nossa tabela a ser

t_i :	-1.0	-0.4286	-0.1429	0.1429	0.7143	1.0
f_i :	2.0	2.2	2.3	4.1	5.5	7.0

para uma função aproximadora do tipo

$$y(t) = a_0 + a_1\, t + a_2\, (2t^2-1)\ .$$

O sistema de equações que nos permite calcular os coeficientes é agora;

$$\begin{bmatrix} 6 & 0.2857 & -0.5305 \\ 0.2857 & 2.7348 & 0.2857 \\ -0.5305 & 0.2857 & 4.2404 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 23.1 \\ 8.2430 \\ -1.5821 \end{bmatrix}\ .$$

Resolvendo-o, obtemos que

$$y(x) = 3.7884 + 2.5478 \left(\frac{2x-5.5.}{3.5}\right) + 0.6754 \left[2\left(\frac{2x-5.5.}{3.5}\right)^2 - 1\right]\ ,$$

ou ainda

$$y(t) = 3.1130 + 2.5478\, t + 1.3508\, t^2\ ,$$

correspondendo à parábola que melhor aproxima a tabela de pontos (ver a Fig. 4.3).

4.2.3 Aproximação por outras funções

Mais uma vez, e voltando à expressão geral (4.1.13), podemos notar que as funções padrão $\phi_j(x)$ podem ser quaisquer e não apenas funções geradoras de polinómios, tal como usamos até agora.

Podemos definir uma função aproximadora que é gerada como combinação linear de funções trignométricas ou exponencias, por exemplo, à semelhança do que fizemos no caso da interpolação (Secção 3.4). A definição da base de funções padrão que escolhermos dependerá do comportamento esperado para a função $f(x)$ que estamos a tentar aproximar.

Exemplo 4.2.3: Consideremos novamente a seguinte tabela de pontos

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
f_i :	2.0	2.2	2.3	4.1	5.5	7.0

que queremos aproximar por uma função do tipo

$$y(x) = a_0 \cos x + a_1 \mathrm{e}^x \, .$$

O sistema de equações que nos permite calcular os coeficientes é agora (ver 4.2.5 e 4.2.6);

$$\begin{bmatrix} 2.5587 & -85.914 \\ -85.914 & 11698. \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} -10.807 \\ 1062.5 \end{bmatrix} \, .$$

Resolvendo-o, obtemos que

$$y(x) = -1.5583 \, \cos x + 0.79382 \, \mathrm{e}^x \, ,$$

é a função que melhor aproxima a tabela de pontos.

Outra possibilidade que não consideramos ainda é a de escrevermos uma função aproximadora que não corresponde a uma combinação linear de funções padrão. Em tal caso a determinação da função aproximadora é mais difícil pois não podemos reduzir o problema a um sistema de equações lineares para os parâmetros que definem a função aproximadora. Um exemplo seria uma função do tipo,

$$y(x) = a_0 \, \mathrm{e}^{a_1 x} \, . \tag{4.2.38}$$

Podemos ainda usar o método dos Mínimos Quadrados, mas o cálculo dos parâmetros a_j terá de ser feito encontrando o mínimo de $\mathscr{R}$ por um método não linear de cálculo numérico de mínimos de funções.

No entanto em alguns casos é possível reduzir um problema não linear a um que o seja. Vejamos por exemplo a expressão (4.2.38); após usar logaritmos temos que

$$g(x) \equiv \log[y(x)] = \log a_0 + a_1 x \equiv b_0 + b_1 x \, . \tag{4.2.39}$$

Ou seja, a tabela de pontos $\{(x_i, \log f_i)\}_{i=0}^n$ pode ser aproximada por uma função linear $g(x)$ que corresponde a uma recta, obtendo-se tal como anteriormente feito os valores de b_0 e b_1. A partir destes valores temos então a função aproximadora de $\{(x_i, f_i)\}_{i=0}^n$ como sendo dada por

$$y(x) = \mathrm{e}^{b_0 + b_1 x} \, . \tag{4.2.40}$$

Exemplo 4.2.4: Consideremos novamente a seguinte tabela de pontos

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
f_i :	2.0	2.2	2.3	4.1	5.5	7.0

que queremos aproximar por uma função do tipo

$$y(x) = a_0 \, \mathrm{e}^{a_1 x} \, .$$

Usando logaritmos (ver 4.2.39) trasnformamos esta tabela em

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
$\log f_i$:	0.6931	0.7885	0.8329	1.4110	1.7047	1.9459

cuja função aproximadora será agora $g(x)=\log a_0+a_1 x \equiv b_0+b_1 x$. O sistema de equações que nos permite calcular os coeficientes é agora (ver 4.2.5 e 4.2.6);

$$\begin{bmatrix} 6 & 17 \\ 17 & 56.5 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 7.376 \\ 24.16 \end{bmatrix}$$

Resolvendo-o, obtemos que b_0=0.12048 e b_1=0.39136. Temos então que a função aproximadora $y(x)$ é

$$y(x) = \mathrm{e}^{0.12048} \cdot \mathrm{e}^{0.39136\, x} = 1.128 \, \mathrm{e}^{0.39136 x} \, ,$$

pois $a_0=\mathrm{e}^{b_0}$=1.128 e $a_1=b_1$=0.39136.

4.2.4 Mínimos Quadrados ponderados

A aproximação linear é usada frequentemente para casos nos quais apenas conhecemos valores aproximados (que tem um erro associado) da função $f(x)$. Tal acontece quando estes são por exemplo obtidos por experimentação, a qual tem sempre um error de medida/experimental associado. A tabela de valores passa agora a ser $\{(x_i, f_i, \sigma_i)\}_{i=0}^{n}$, onde σ_i é a incerteza associada ao valor f_i, sendo portanto uma forma de classificar a confiança que temos nos valores tabelados de $f(x)$.

Desta forma, caso alguns dos pontos tabelados tenham um erro associado superior aos restantes é importante garantir que a função aproximadora tenha em conta o erro de forma a não ser dominada por aqueles pontos em que confiamos *menos* (ver Fig. 4.4).

Vejamos então como podemos introduzir no processo de escolha da melhor função pelo método dos Mínimos Quadrados uma avaliação da qualidade dos pontos, de forma a que os melhores pontos (menor erro) tenham um contribuição maior na definição da função aproximadora. Introduzimos assim o peso p_i (valor positivo) do ponto (x_i, f_i) como sendo a quantidade que é proporcional à qualidade do ponto. Isto é, quando menor é a incerteza no valor de f_i maior será o valor do peso p_i.

Aquilo que é relevante, não é o valor absoluto que de facto os pesos tomam, mas sim o valor relativo entre si. Logo, um ponto será *de confiança* se o valor do seu peso é superior à média dos pesos para todos os pontos (ver Fig. 4.5). Quanto mais acima estiver do valor médio melhor é a qualidade do ponto (menor o erro associado). O mesmo se podendo dizer para o contrário, em que quanto menor é o peso relativamente ao valor médio mais baixa é a qualidade do ponto, daí que deva ser reduzida a sua influência na definição da função aproximadora. Este definição de peso leva a que baste redefinir o conceito de resíduos, escrevendo que

$$\mathscr{R} = \sum_{i=0}^{n} p_i \, [f_i - y(x_i)]^2 \, , \tag{4.2.41}$$

em que, tal como antes,

$$y(x) = \sum_{j=0}^{m} a_j \, \phi_j(x) \, . \tag{4.2.42}$$

Em tal caso o sistema anterior (4.2.5), escrito como

$$\mathscr{B} \cdot \vec{a} = \vec{g} \, , \tag{4.2.43}$$

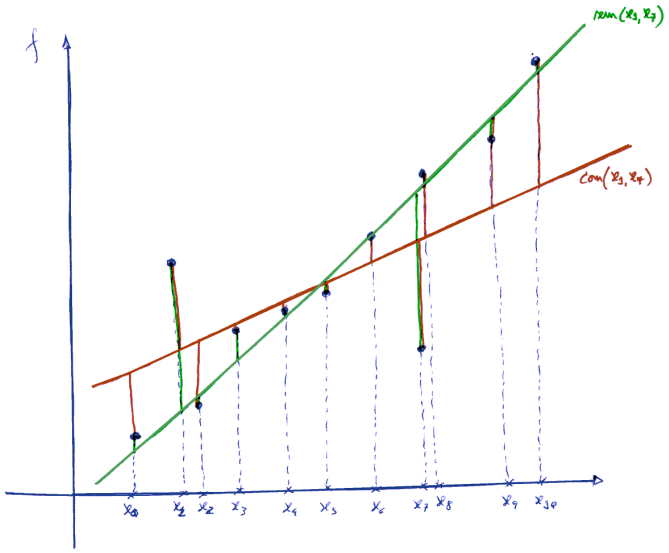


Figura 4.4: Representação da melhor recta aproximadora, tal como calculada pelo método dos Mínimos Quadrados, de uma tabela de valores para uma função $f(x)$ no caso de dois pontos *maus*; x_1 e x_7, serem usados (linha a tracejado) ou ignorados (linha contínua).

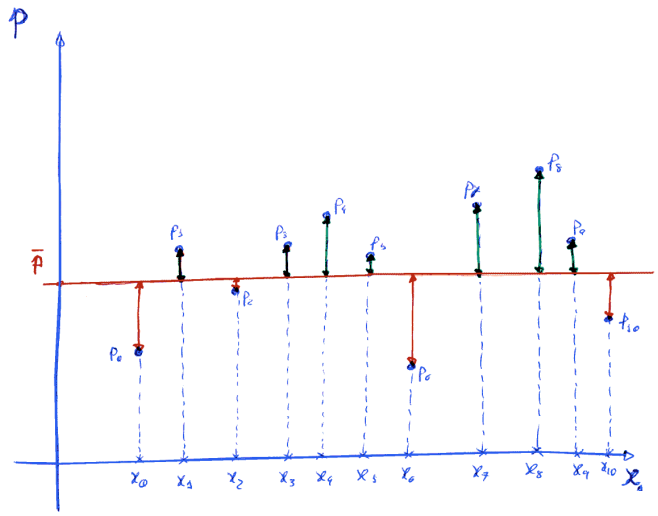


Figura 4.5: Representação da grandeza relativa dos pesos de uma tabela de valores. A quantidade mais relevante a classificar a *qualidade* de um ponto é a grandeza relativa (e não a grandeza absoluta) do peso, sendo a normalização dada por exemplo pelo valor médio $\bar{p}$ dos pesos.

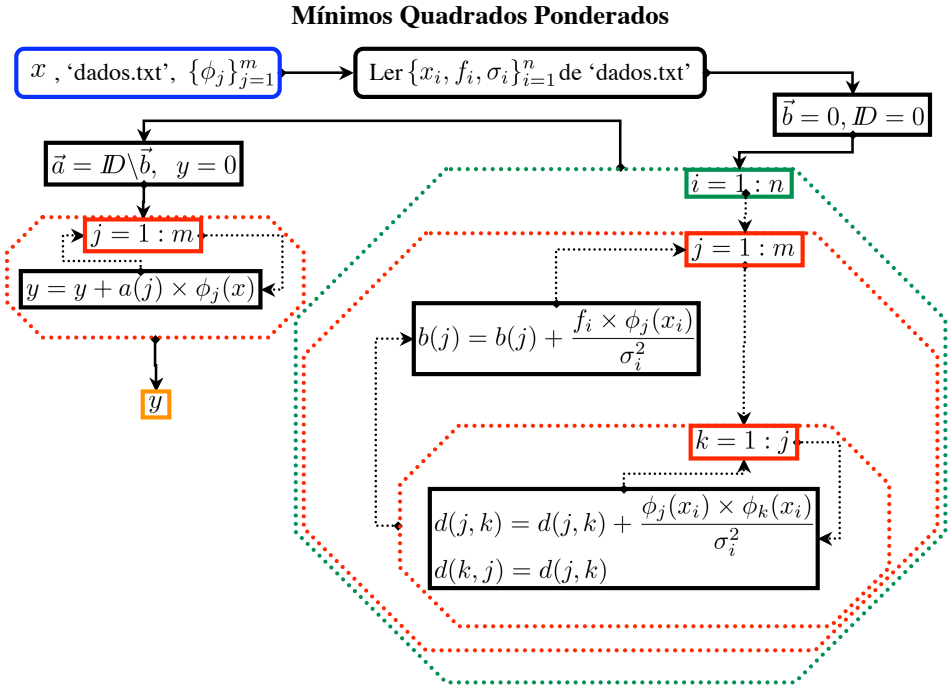


Figura 4.6: Algoritmo para implementação do Método de aproximação χ-Quadrado, que permite determinar o valor $y = \sum_{j=1}^{m} a_j \phi_j(x)$ que melhor aproxima a função $f(x)$ nos pontos $\{x_i, f_i, \sigma_i\}_{i=1}^{n}$, e em que σ_i é a incerteza associada ao ponto i.

passa agora a ser dado por

$$b_{kj} = b_{jk} = \sum_{i=0}^{n} \left[p_i \, \phi_k(x_i) \, \phi_j(x_i) \right] \qquad \text{e} \qquad g_k = \sum_{i=0}^{n} \left[p_i \, \phi_k(x_i) \, f_i \right] . \tag{4.2.44}$$

Aproximação χ-quadrado: no caso particular de a qualidade ser estimada pelo erro experimental σ_i, define-se o peso como sendo

$$p_i = \frac{1}{\sigma_i^2} , \tag{4.2.45}$$

a que corresponde uma aproximação pelos Mínimos Quadrados ponderados do tipo χ-quadrado, sendo frequentemente usada em ciências experimentais.

Neste caso os resíduos são dados por

$$\chi^2 = \sum_{i=0}^{n} \left[\frac{f_i - y(x_i)}{\sigma_i} \right]^2 . \tag{4.2.46}$$

Logo, temos também que

$$b_{kj} = b_{jk} = \sum_{i=0}^{n} \left[\frac{\phi_k(x_i) \, \phi_j(x_i)}{\sigma_i^2} \right] \qquad \text{e} \qquad g_k = \sum_{i=0}^{n} \left[\frac{\phi_k(x_i) \, f_i}{\sigma_i^2} \right] . \tag{4.2.47}$$

Exemplo 4.2.5: Consideremos a seguinte tabela de pontos

x_i :	1.0	2.0	2.5	3.0	4.0	4.5
f_i :	2.0	2.2	2.3	4.1	5.5	7.0
σ_i :	0.1	0.2	0.1	0.3	0.4	0.4

de forma a que, usando a incerteza σ na medida de cada valor, obtenha a melhor aproximação por uma função do tipo

$$y(x) = a_0 \cos x + a_1 \mathrm{e}^x \, .$$

O sistema de equações que nos permite calcular os coeficientes é agora (ver 4.2.47);

$$\begin{bmatrix} 111.543 & -1468.58 \\ -1468.58 & 90803.0 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} -175.882 \\ 10482.1 \end{bmatrix} \, .$$

Resolvendo-o, obtemos que

$$y(x) = -0.070246 \, \cos x + 0.11443 \, \mathrm{e}^x \, ,$$

é a função que melhor aproxima a tabela de pontos.

4.3 Aproximação de funções

Podemos também substituir uma função $f(x)$ definida num intervalo $[a,b]$, por uma função aproximadora escrita na forma

$$y(x) = \sum_{j=0}^{m} a_j \, \phi_j(x) \, . \tag{4.3.1}$$

Neste caso a expressão dos resíduos deixa de ser um somatório, pois já não temos apenas uma tabela de pontos, passando a ser um integral;

$$\mathscr{R} = \int_a^b [f(x) - y(x)]^2 \, \omega(x) \, \mathrm{d}x \, , \tag{4.3.2}$$

onde $\omega(x)$ é a função peso definida no intervalo $[a,b]$, sendo positiva em todos os pontos. Podemos então calcular os coeficientes a_j da combinação linear determinando o mínimo da quantidade $\mathscr{R}$. Tal corresponde a encontrar os zeros das derivadas,

$$\frac{\partial \mathscr{R}}{\partial a_j} = 0 \qquad \text{para} \qquad j = 0, 1, ..., m \, . \tag{4.3.3}$$

Daqui resulta que

$$\begin{aligned} & \sum_{j=0}^{m} a_j \left[\int_a^b \phi_k(x) \phi_j(x) \omega(x) \, \mathrm{d}x \right] = \int_a^b \phi_k(x) f(x) \omega(x) \, \mathrm{d}x \\ \Rightarrow \quad & \sum_{j=0}^{m} (\phi_k, \phi_j) \, a_j = (\phi_k, f) \, , \end{aligned} \tag{4.3.4}$$

para k=0, 1, ..., m e tal que

$$(u,v) \equiv \int_a^b u(x) v(x) \omega(x) \, \mathrm{d}x \, . \tag{4.3.5}$$

Temos então o seguinte sistemas de m+1 equações

$$\mathscr{B} \cdot \vec{a} = \vec{g} \, , \tag{4.3.6}$$

onde

$$b_{kj} = b_{jk} = (\phi_k, \phi_j) \qquad \text{e} \qquad g_k = (\phi_k, f) \, . \tag{4.3.7}$$

Exemplo 4.3.1: Determinemos pelo método dos Mínimos Quadrados a função aproximadora de

$$f(x) = \pi^2 - x^2 \,,$$

no intervalo $[0,\pi]$, que seja da forma

$$y(x) = a_0 + a_1 \, \cos x \,,$$

para $\omega(x)$=1. Usando (4.3.2) e (4.3.4) temos que

$$b_{11} = \int_0^\pi \mathrm{d}x = \pi \qquad b_{12} = \int_0^\pi \cos x \, \mathrm{d}x = 0$$
$$b_{21} = b_{12} = 0 \qquad b_{22} = \int_0^\pi \cos^2 x \, \mathrm{d}x = \pi/2$$

$$g_1 = \int_0^\pi (\pi^2 - x^2)\mathrm{d}x = 2\pi^3/3 \qquad g_2 = \int_0^\pi (\pi^2 - x^2)\cos x \, \mathrm{d}x = 2\pi$$

Resolvendo o sistema de equações (4.3.1) para a_0 e a_1, obtemos que

$$y(x) = \frac{2}{3}\pi^2 + 4\cos x \,,$$

como sendo a função aproximadora de $f(x)$ no intervalo especificado.

Exemplo 4.3.2: Consideremos agora a função

$$f(x) = \sin(2x-1) \,,$$

definida no intervalo $[0,1]$, e determinemos qual a parábola que melhor aproxima esta função, com um peso $\omega(x)$=1. Se recorrermos aos polinómios de Legendre, temos que a função aproximadora é

$$y(t) = a_0 + a_1 t + a_2 \left(\frac{3t^2-1}{2} \right) \,,$$

com t=$2x-1$, de forma a termos que $t \in [-1,1]$. Assim, o sistema de equações (4.3.1), para a variável t, é dado por

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2/3 & 0 \\ 0 & 0 & 2/5 \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2(\sin 1 - \cos 1) \\ 0 \end{bmatrix} \,,$$

de onde resulta que a_0=0, a_1=$3(\sin 1 - \cos 1)$ e a_2=0. A função aproximadora em x, é então dada por

$$y(x) = 3(\sin 1 - \cos 1)(2x-1) \,.$$

Esta é a parábola que melhor aproxima $f(x)$ no intervalo considerado.

4.4 Exercícios

E4.1) Prove que a recta obtida por aproximação linear de um conjunto de pontos passa pelo ponto $(\bar{x},\bar{y})$, onde a barra representa a média aritmética dos valores para todos os pontos.

E4.2) Obtenha pelo método dos mínimos quadrados a melhor combinação linear das funções $\cos(x)$ e e^x que aproxima os pontos da tabela ao lado.

x	$f(x)$
-2	2.14
-1	2.36
0	3.15
1	4.72
2	9.35

E4.3) Pretende-se construir uma estrada parabólica entre Lisboa e o Porto, cidades estas cujas coordenadas (l,d) (l é medido na direcção sul-norte enquanto que d é na direcção oeste-este) são respectivamente (0,0) e (350,50). Determine a curvatura que a estrada terá de ter de forma a minimizar a soma dos quadrados das distâncias (medidas na direcção oeste-este) que ligam as cidades, incluídas na tabela, á estrada planeada. Como se altera a curvatura se a ligação a Coimbra for mais importante (conta a triplicar) do que as restantes cidades?

Cidade	(l,d)
Aveiro	$(300,50)$
Coimbra	$(200,60)$
Leiria	$(100,20)$
Santarem	$(50,50)$

E4.4) Conhecem-se os seguintes pontos de uma função real $f(x)$ em três pontos: $f(x_0{=}-1){=}0.5$, $f(x_1{=}0){=}0.0$ e $f(x_2{=}1){=}-0.1$. Calcule pelo método dos mínimos quadrados a recta que, passando necessáriamente pelo ponto $(2,-0.4)$, melhor aproxima os valores dados de f em (x_0,x_1,x_2). Represente num gráfico os pontos dados e a recta aproximadora encontrada pelo método dos mínimos quadrados.

E4.5) Em Sismologia Estelar usa-se o valor determinado por observação de Δ, para obter o valor da massa M da estrela. Para tal usa-se um modelo de referência, que neste caso corresponde a $M_0/M_\odot{=}1.10$ e $\Delta_0{=}55.47\ \mu$Hz, e recorre-se à expressão

$$\frac{M}{M_0} = D_0 \times \left(\frac{\Delta}{\Delta_0}\right)^2 .$$

Usando os valores da tabela ao lado determine pelo método dos mínimos quadrados ponderados o valor de $M_{\rm obs}$ que corresponde a $\Delta_{\rm obs}{=}55.2\ \mu$Hz.

$M/M_\odot$	$\Delta\ (\mu\text{Hz})$	erro (μHz)
1.05	54.27	0.32
1.07	54.69	0.54
1.10	55.50	0.12
1.10	55.47	0.17
1.10	55.75	0.57
1.15	56.74	0.23

E4.6) Indique para as seguintes funções $y(x)$, aproximadoras de uma tabela de pontos $\{x_i,f_i\}$, como pode linearizar o problema para aplicação do método dos mínimos quadrados;

a) $y(x) = \dfrac{\alpha}{\beta+x}$

b) $y(x) = \alpha\, e^{\beta x^2}$

E4.7) Seja $f(x)$ uma função real que descreve o comportamento de um sistema físico. Por experimentação mediram-se os seguintes valores: $f(-1){=}0.736$, $f(0){=}1.99$ e $f(+1){=}5.44$. Sabendo que se espera um comportamento para $f(x)$ do tipo $\alpha\, e^{\beta x}$ determine a melhor aproximação recorrendo ao método dos mínimos quadrados para calcular os valores de α e β.

Capítulo 5

Cálculo numérico de derivadas e integrais

Por vezes precisamos de calcular propriedades de uma função, tal como a sua derivada num ponto ou o seu integral num intervalo, conhecendo apenas alguns pontos da função. Para tal temos mais uma vez de recorrer a métodos numéricos que nos permitam estimar o valor dessas quantidades usando apenas a informação disponível. Nesta Secção consideramos alguns desses métodos para o cálculo de derivadas e integrais de funções tabeladas.

5.1 Cálculo numérico da derivada de uma função

Comecemos por considerar a definição de derivada de uma função real $f(x)$ num ponto x:

$$f'(x) \equiv \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \,. \tag{5.1.1}$$

Se considerarmos a expansão da função em torno de x temos que

$$f(x+h) = f(x) + f'(x)\, h + \frac{1}{2} f''(\xi)\, h^2 \,, \tag{5.1.2}$$

em que $\xi \in [x, x+h]$. Daqui, tiramos que

$$f'(x) = \frac{f(x+h)-f(x)}{h} - \frac{1}{2} f''(\xi)\, h \,. \tag{5.1.3}$$

De onde resulta que

$$f'(x) \simeq \frac{f(x+h)-f(x)}{h} \,, \tag{5.1.4}$$

é uma aproximação do valor da derivada, com um erro

$$\varepsilon = \left| \frac{1}{2} f''(\xi)\, h \right| \propto h \,. \tag{5.1.5}$$

Isto é, podemos usar a expressão (5.1.4) para estimar a derivada tal como definida em (5.1.1), tendo então um erro ε que é proporcional ao valor de h, usado.

Exemplo 5.1.1: Conhecemos o valor de uma função $f(x)$ nos seguintes pontos: $\{(0,0),(1,2)\}$, e queremos determinar o valor da sua derivada em x=0. Recorrendo à expressão (5.1.4), temos que (para h=1−0=1)

$$f'(0) \simeq \frac{f(0+h)-f(0)}{h} = \frac{2-0}{1} = 2 \,,$$

é uma estimativa do valor da derivada, cujo erro é $\varepsilon \propto h$.

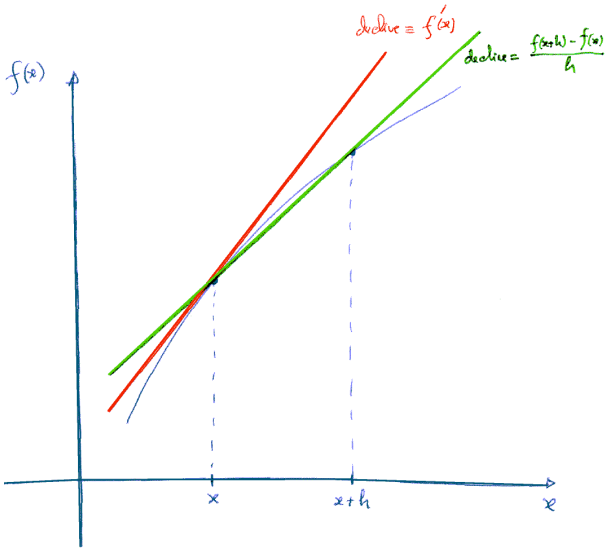


Figura 5.1: Representação da derivada da função $f(x)$ em x, estimada através do valor da função em x e em $x+h$. Isto é, recorremos à secante da função em x e $x+h$ para estimar o declive da tangente em x.

5.1.1 Fórmula das diferenças centrais de segunda ordem

No entanto podemos tentar reduzir o erro na estimativa da derivada caso disponhamos de mais pontos da função. Para tal comecemos por considerar a expansão dada em (5.1.2), mas agora até ao termo na terceira derivada;

$$\begin{aligned} f(x+h) &= f(x) + f'(x)\,h + \frac{1}{2}f''(x)\,h^2 + \frac{1}{6}f'''(\xi_+)\,h^3 \\ f(x-h) &= f(x) - f'(x)\,h + \frac{1}{2}f''(x)\,h^2 - \frac{1}{6}f'''(\xi_-)\,h^3\,, \end{aligned} \tag{5.1.6}$$

em que $\xi_+, \xi_- \in [x-h, x+h]$. Por subtracção das duas expressões encontramos que

$$\begin{aligned} f(x+h) - f(x-h) &= 2f'(x)\,h + \frac{1}{6}\left[f'''(\xi_+) + f'''(\xi_-)\right]h^3 \\ &= 2f'(x)\,h + \frac{1}{3}f'''(\xi)h^3\,, \end{aligned} \tag{5.1.7}$$

com $\xi \in [x-h, x+h]$, de acordo com o teorema do valor médio.

Temos assim, a partir desta expressão, uma nova forma de estimar a derivada;

$$f'(x) \simeq \frac{f(x+h) - f(x-h)}{2h}\,, \tag{5.1.8}$$

cujo erro é agora dado por

$$\varepsilon = \left|\frac{1}{6}f'''(\xi)h^2\right| \propto h^2\,. \tag{5.1.9}$$

Como a dependência em h é agora quadrática, temos que o erro pode ser bastante menor, do que no caso anterior (5.1.4), se h é pequeno.

Exemplo 5.1.2: Conhecemos o valor de uma função $f(x)$ nos seguintes pontos: $\{(-1,-1),(1,2)\}$, e queremos determinar o valor da sua derivada em x=0. Recorrendo à expressão (5.1.8), temos que (para h=1−0=0−(−1)=1)

$$f'(0) \simeq \frac{f(0+h) - f(0-h)}{2h} = \frac{2-(-1)}{2} = \frac{3}{2}\,,$$

é uma estimativa do valor da derivada, cujo erro é $\varepsilon \propto h^2$.

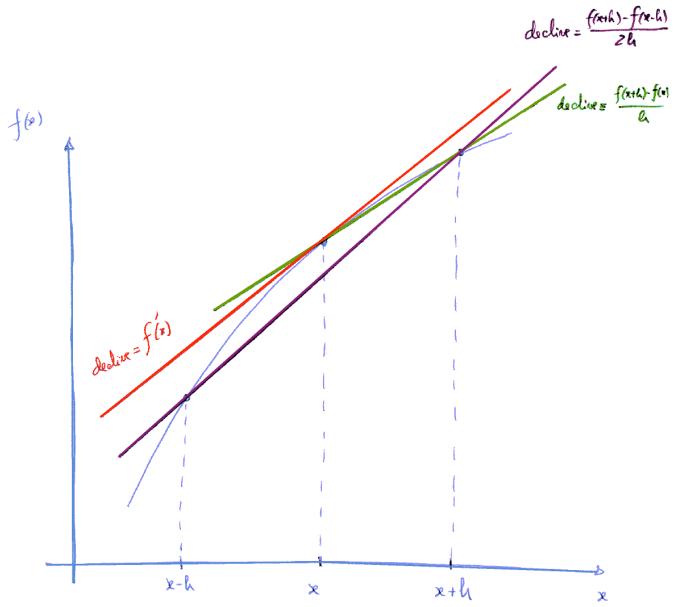


Figura 5.2: Representação da derivada da função $f(x)$ em x, estimada através do valor da função em $x-h$ e em $x+h$. Isto é, recorremos à secante da função em $x-h$ e $x+h$ para estimar o declive da tangente em x. Também é representada a aproximação discutida na Fig. 5.1.

Exemplo 5.1.3: De forma a ilustrar a importância do erro consideremos a seguinte função,

$$f(x) = \mathrm{e}^x \,,$$

e calculemos a sua derivada em x=0. Sabemos que $f'(0)$=1; vamos então agora comparar as estimativas que se encontram, para diferentes valores de h, usando a expressão (5.1.8);

$$f'(0) \simeq D_h \equiv \frac{f(0+h)-f(0-h)}{2h} = \frac{\mathrm{e}^{+h}-\mathrm{e}^{-h}}{2h} \,,$$

obtendo-se que

$h:$	10^0	10^{-1}	10^{-2}
$D_h:$	1.175201	1.001668	1.000017

Isto é, como $\varepsilon \propto h^2$, para valores de h menores temos um erro menor, obtendo-se uma estimativa para a derivada mais próxima do valor exacto.

5.1.2 Fórmula das diferenças centrais de quarta ordem

Consideremos uma vez mais que adicionamos mais informação sobre a função $f(x)$. Neste caso, passamos a ter quatro valores desta nos pontos $\{x-2h, x-h, x+h, x+2h\}$, e queremos determinar a derivada em x. Para tal, escrevemos mais uma vez as expansões da função, em torno de x, para estes quatro pontos

$$\begin{aligned}
f(x+2h) &= f(x) + f^{(1)}(x)\, 2h + \frac{1}{2} f^{(2)}(x)\, 4h^2 + \frac{1}{6} f^{(3)}(x)\, 8h^3 + \frac{1}{24} f^{(4)}(\xi_{2+})\, 16h^4 \\
f(x+h) &= f(x) + f^{(1)}(x)\, h + \frac{1}{2} f^{(2)}(x)\, h^2 + \frac{1}{6} f^{(3)}(x)\, h^3 + \frac{1}{24} f^{(4)}(\xi_{1+})\, h^4 \\
f(x-h) &= f(x) - f^{(1)}(x)\, h + \frac{1}{2} f^{(2)}(x)\, h^2 - \frac{1}{6} f^{(3)}(x)\, h^3 + \frac{1}{24} f^{(4)}(\xi_{1-})\, h^4 \\
f(x-2h) &= f(x) - f^{(1)}(x)\, 2h + \frac{1}{2} f^{(2)}(x)\, 4h^2 - \frac{1}{6} f^{(3)}(x)\, 8h^3 + \frac{1}{24} f^{(4)}(\xi_{2-})\, 16h^4 \,,
\end{aligned} \tag{5.1.10}$$

em que $\xi_{2+}, \xi_{1+}, \xi_{1-}, \xi_{2-} \in [x-2h, x+2h]$. Temos agora de combinar estas quatro equações de forma a eliminar $f(x)$, $f^{(2)}(x)$ e $f^{(3)}(x)$, mantendo o termo em $f^{(1)}(x)$. Para tal basta considerar que

$$\begin{aligned}
&a_1 f(x+2h) + a_2 f(x+h) + a_3 f(x-h) + a_4 f(x-2h) = \\
&\quad = 0 \cdot f(x) + 1 \cdot f^{(1)}(x)\, h + 0 \cdot f^{(2)}(x) + 0 \cdot f^{(3)}(x) + ? \cdot f^{(4)}(x)\, h^4 \,.
\end{aligned} \tag{5.1.11}$$

Que corresponde a impor as seguintes condições;

$$\begin{aligned} a_1+a_2+a_3+a_4 &= 0 \\ 2a_1+a_2-a_3-2a_4 &= 1 \\ 4a_1+a_2+a_3+4a_4 &= 0 \\ 8a_1+a_2-a_3-8a_4 &= 0\,. \end{aligned} \tag{5.1.12}$$

A solução deste sistema é

$$a_1=-\frac{1}{12} \qquad a_2=\frac{8}{12} \qquad a_3=-\frac{8}{12} \qquad a_4=\frac{1}{12}\,. \tag{5.1.13}$$

Depois de substituir estes valores, ficamos com

$$\begin{aligned} \frac{-f(x+2h)+8f(x+h)-8f(x-h)+f(x-2h)}{12} &= f^{(1)}(x)\,h+ \\ &+\frac{h^4}{36}\left[-2f^{(4)}(\xi_{2+})+f^{(4)}(\xi_{1+})-f^{(4)}(\xi_{1-})+2f^{(4)}(\xi_{2-})\right]\,. \end{aligned} \tag{5.1.14}$$

Usando o teorema do valor médio, e a expansão para $f^{(4)}(\xi_b)$, temos ainda que

$$\begin{aligned} -2f^{(4)}(\xi_{2+})+f^{(4)}(\xi_{1+})-f^{(4)}(\xi_{1-})+2f^{(4)}(\xi_{2-}) &= -3f^{(4)}(\xi_a)+3f^{(4)}(\xi_b) \\ &= 3f^{(5)}(\xi)(\xi_b-\xi_a)\,, \end{aligned} \tag{5.1.15}$$

onde $\xi_a,\xi_b,\xi\in[x-2h,x+2h]$. Temos então que a expressão para a diferença central de quarta ordem é

$$f'(x)\simeq\frac{-f(x+2h)+8f(x+h)-8f(x-h)+f(x-2h)}{12h}\,, \tag{5.1.16}$$

com um erro dado por

$$\varepsilon=\left|\frac{h^3}{36}\cdot 3f^{(5)}(\xi)(\xi_b-\xi_a)\right|\leq\left|\frac{h^3}{12}f^{(5)}(\xi)(4h)\right|=\left|\frac{1}{3}f^{(5)}(\xi)\,h^4\right|\,, \tag{5.1.17}$$

em que $\xi\in[x-2h,x+2h]$.

Exemplo 5.1.4: Conhecemos o valor de uma função $f(x)$ nos seguintes pontos: $\{(-2,0);(-1,-1);(1,2);(2,4)\}$, e queremos determinar o valor da sua derivada em x=0. Recorrendo à expressão (5.1.16), temos que (visto h=1)

$$\begin{aligned} f'(0) &\simeq \frac{-f(0+2h)+8f(0+h)-8f(0-h)+f(0-2h)}{12h} \\ &\simeq \frac{-4+8\cdot 2-8\cdot(-1)+0}{12}=\frac{5}{3}\,, \end{aligned}$$

é uma estimativa do valor da derivada, cujo erro é $\varepsilon\propto h^4$.

Exemplo 5.1.5: De forma a ilustrar a importância do erro consideremos, mais uma vez, a seguinte função,

$$f(x)=\mathrm{e}^x\,,$$

e calculemos a sua derivada em x=0. Sabemos que $f'(0)$=1; pelo que vejamos qual a estimativa que se encontra, para diferentes valores de h, usando a expressão (5.1.16);

$$\begin{aligned} f'(0)\simeq D_h &\equiv \frac{-f(0+2h)+8f(0+h)-8f(0-h)+f(0-2h)}{12h} \\ &= \frac{-\mathrm{e}^{+2h}+8\mathrm{e}^{+h}-8\mathrm{e}^{-h}+\mathrm{e}^{-2h}}{12h}\,, \end{aligned}$$

obtendo-se que

h:	10^0	10^{-1}	10^{-2}
D_h:	0.962458	0.999997	1.000000

Isto é, como $\varepsilon\propto h^4$ temos que para valores de h menores temos um erro menor, obtendo-se uma estimativa para a derivada mais próxima do valor exacto, e melhor do que obtido usando a fórmula de segunda ordem (ver Exemplo 5.1.3).

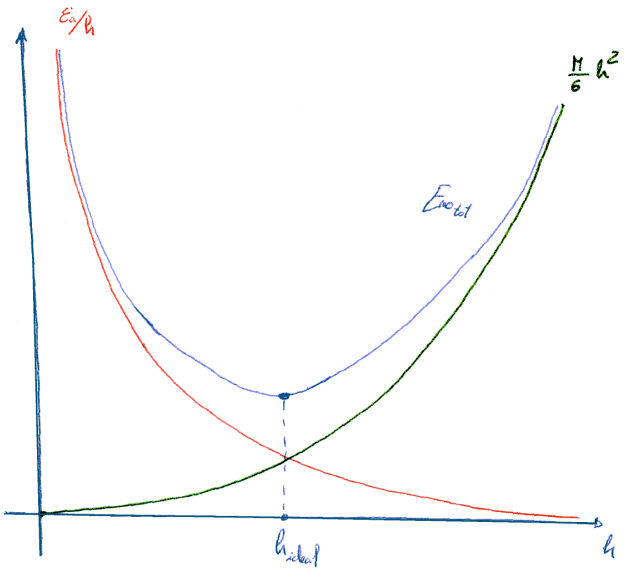


Figura 5.3: Representação das contribuições do erro de truncatura e de arredondamento para o erro total no cálculo da derivada pela expressão (5.1.8). O valor ideal de h corresponde ao mínimo desta função; para valores pequenos de h o erro de arredondamento domina, enquanto que para valores maiores de h o erro de truncatura da fórmula é a contribuição dominante.

5.1.3 Efeito dos erros de arrendondamento no cálculo da derivada

Como no cálculo numérico temos sempre de ter em conta o efeito dos erros de arrendondamento, é importante notar que não basta reduzir o valor de h para melhorar a estimativa encontrada para a derivada, pois eventualmente teremos um resultado que é dominado pelo erro de arredondamento cometido no cálculo.

Vejamos então qual será o valor óptimo para h, no caso da *fórmula das diferenças centrais de segunda ordem*. Se

$$f'(x) = \frac{f(x+h)-f(x-h)}{2h} + \text{Erro}_{\text{tru}} \,, \tag{5.1.18}$$

temos que

$$\text{Erro}_{\text{tot}} = \text{Erro}_{\text{arr}} + \text{Erro}_{\text{tru}} \,. \tag{5.1.19}$$

Já vimos em (5.1.9) que,

$$\text{Erro}_{\text{tru}} \leq \frac{h^2}{6} M \,, \qquad \text{onde} \qquad M = \max_{\xi\in[x-h,x+h]} \left|f'''(\xi)\right| \,. \tag{5.1.20}$$

Se os valores de $f(x)$ tém um erro de arredondamento de ε_a, então do uso da expressão (5.1.8) resulta um erro;

$$\text{Erro}_{\text{arr}} \leq \frac{\varepsilon_a}{h} \,. \tag{5.1.21}$$

Substituindo, temos que

$$\text{Erro}_{\text{tot}} \leq \frac{\varepsilon_a}{h} + \frac{h^2}{6} M \,. \tag{5.1.22}$$

O mínimo desta expressão corresponde a

$$\frac{\text{dErro}_{\text{tot}}}{\text{d}h} = 0 \quad \Rightarrow \quad -\frac{\varepsilon_a}{h^2} + \frac{2h}{6} M = 0 \,, \tag{5.1.23}$$

cuja solução é o valor

$$h_{\text{ideal}} = \left(\frac{3\varepsilon_a}{M}\right)^{1/3} \,. \tag{5.1.24}$$

As duas contribuições para o erro total na estimativa da derivada estão representadas na Fig. 5.3, onde também se indica o valor de h a que corresponde o mínimo.

Como exemplo, temos que para $\varepsilon_a{=}10^{-11}$ e $M{=}1$, o valor ideal é $h_{\text{ideal}}{\simeq}3{\times}10^{-4}$.

Enquanto que no caso da *fórmula das diferenças centrais de quarta ordem* o valor ideal de h, se tivermos em consideração os erros de arredondamento, será diferente. Pois, como

$$f'(x) = \frac{-f(x{+}2h) + 8f(x{+}h) - 8f(x{-}h) + f(x{-}2h)}{12h} + \text{Erro}_{\text{tru}} \,, \tag{5.1.25}$$

temos então que

$$\text{Erro}_{\text{tot}} = \text{Erro}_{\text{arr}} + \text{Erro}_{\text{tru}} \,. \tag{5.1.26}$$

Mas agora, usando (5.1.17),

$$\text{Erro}_{\text{tru}} \leq \frac{h^4}{3}\, M \qquad \text{onde} \qquad M = \max_{\xi \in [x-2h, x+2h]} \left| f^{(5)}(\xi) \right| \,, \tag{5.1.27}$$

enquanto que para o erro de arredondamento da expressão (5.1.25), temos que

$$\text{Erro}_{\text{arr}} \leq \frac{3\varepsilon_a}{2h} \,. \tag{5.1.28}$$

Substituindo, obtém-se que

$$\text{Erro}_{\text{tot}} \leq \frac{3\varepsilon_a}{2h} + \frac{h^4}{3}\, M \,, \tag{5.1.29}$$

cujo mínimo é dado por

$$\frac{\text{dErro}_{\text{tot}}}{\text{d}h} = 0 \quad \Rightarrow \quad -\frac{3\varepsilon_a}{2h^2} + \frac{4h^3}{3}\, M = 0 \,, \tag{5.1.30}$$

a que corresponde o valor

$$h_{\text{ideal}} = \left(\frac{9\varepsilon_a}{8M} \right)^{1/5} \,. \tag{5.1.31}$$

De forma a compararmos com a expressão anterior, vejamos mais uma vez, como exemplo, o caso de $\varepsilon_a{=}10^{-11}$ e $M{=}1$. Para estes dados temos que o valor ideal é agora $h_{\text{ideal}}{\simeq}6{\times}10^{-3}$.

5.1.4 Cálculo da derivada recorrendo a interpolação

Podemos facilmente generalizar as expressões obtidas em (5.1.8) ou (5.1.16) recorrendo a interpolação polinomial, para estimar o valor de derivada de uma função $f(x)$ num ponto usando uma tabela de valores que não estão necessariamente igualmente espaçados.

Seja $\{(x_i, f_i)\}_{i=0}^{n}$ uma tabela de $n{+}1$ pontos cujo polinómio interpolador é $y(x)$. Então

$$f'(x) \simeq y'(x) \,. \tag{5.1.32}$$

O erro da estimativa para a derivada é agora dado àcusta da derivada do erro de interpolação, tal como encontrado em (3.2.31).

No caso de termos três pontos, o polinómio interpolador (parábola) é dado por

$$y(x) = f_0 \frac{(x{-}x_1)(x{-}x_2)}{(x_0{-}x_1)(x_0{-}x_2)} + f_1 \frac{(x{-}x_0)(x{-}x_2)}{(x_1{-}x_0)(x_1{-}x_2)} + f_2 \frac{(x{-}x_0)(x{-}x_1)}{(x_2{-}x_0)(x_2{-}x_1)} \,, \tag{5.1.33}$$

cuja derivada nos dá que

$$f'(x) \simeq y'(x) = f_0 \frac{2x{-}x_1{-}x_2}{(x_0{-}x_1)(x_0{-}x_2)} + f_1 \frac{2x{-}x_0{-}x_2}{(x_1{-}x_0)(x_1{-}x_2)} + f_2 \frac{2x{-}x_0{-}x_1}{(x_2{-}x_0)(x_2{-}x_1)} \,. \tag{5.1.34}$$

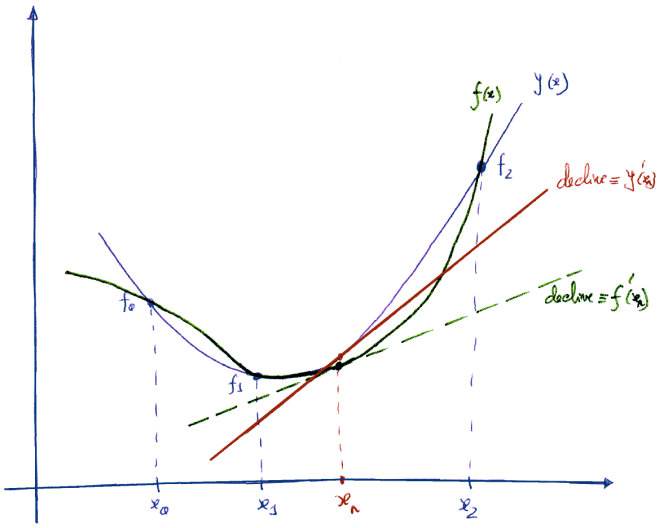


Figura 5.4: Representação da derivada da função $f(x)$ no ponto x_r, e da estimativa dada pela parábola interpoladora da função nos pontos $\{x_0, x_1, x_2\}$.

Exemplo 5.1.6: Mostremos que no caso de $x_1-x_0=x_2-x_1=h$, a expressão (5.1.34) dá uma estimativa para a derivada em x_1 que é equivalente à obtida em (5.1.8). Substituindo, temos que para (5.1.34),

$$f'(x_1) \simeq f_0 \frac{-h}{(-h)(-2h)} + f_1 \frac{h-h}{(h)(-h)} + f_2 \frac{h}{(2h)(h)} = \frac{f_2-f_0}{2h} .$$

Enquanto que para a expressão (5.1.8) temos que

$$f'(x_1) \simeq \frac{f(x_1+h)-f(x_1-h)}{2h} = \frac{f_2-f_0}{2h} ,$$

que corresponde à expressão obtida acima, como queriamos provar.

Exemplo 5.1.7: Consideremos o caso de termos os pontos $\{(x_0,f_0);(x_1,f_1)\}$, de uma função cujo comportamento esperamos que seja da forma;

$$y(x) = a_0 + a_1 e^x .$$

Pretende-se estimar o valor de $f'(x_0)$. Comecemos por determinar a função inteproladora nos pontos conhecidos: $y(x_0)=f_0$ e $y(x_1)=f_1$. Obtendo-se que,

$$a_1 = \frac{f_1-f_0}{e^{x_1}-e^{x_0}} .$$

Assim, como

$$y'(x) = a_1 e^x , \quad \text{temos que} \quad f'(x_0) \simeq y'(x_0) = \frac{e^{x_0}}{e^{x_1}-e^{x_0}} (f_1-f_0) .$$

Esta expressão permite-nos assim estimar o valor da derivada recorrendo à função interpoladora de $f(x)$.

Este método pode também ser usado para estimar as outras derivadas da função. Por exemplo se pretendemos estimar a segunda derivada basta usar a expressão (5.1.34) para escrever que

$$f''(x) \simeq y''(x) = \frac{2f_0}{(x_0-x_1)(x_0-x_2)} + \frac{2f_1}{(x_1-x_0)(x_1-x_2)} + \frac{2f_2}{(x_2-x_0)(x_2-x_1)} . \tag{5.1.35}$$

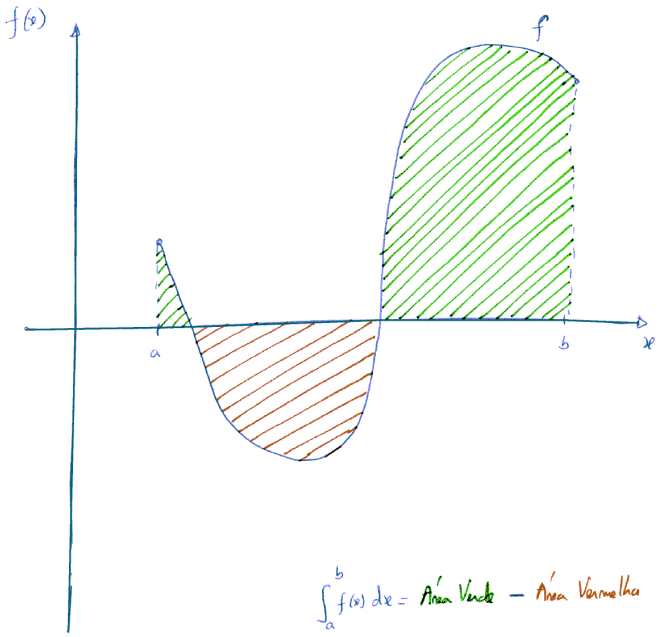


Figura 5.5: Representação da área definida pela função $f(x)$ e o eixo dos x's. O integral da função no intervalo $[a,b]$ corresponde ao valor da área indicada, tal que áreas abaixo do eixo tem uma contribuição negativa.

Exemplo 5.1.8: Voltemos a considerar o caso de $x_1-x_0=x_2-x_1=h$, e usemos a expressão (5.1.35) para estimar o valor da segunda derivada da função em $x=x_1$. Substituindo, temos que

$$f''(x_1) \simeq \frac{2f_0}{(-h)(-2h)} + \frac{2f_1}{(h)(-h)} + \frac{2f_2}{(2h)(h)} = \frac{f_2 - 2f_1 + f_0}{h^2} \,,$$

nos dá a estimativa pedida.

5.1.5 Cálculo da derivada recorrendo a splines

Outra opção é obviamente a utilização de splines como função interpoladora. Se usarmos uma spline de grau $m=3$, então por construção vimos que os parâmetros M_i da spline nos dá a segunda derivada nos nodos, enquanto que os valores de m_i correspondem ao valor da primeira derivada nos nodos. Para qualquer outro valor de x podemos ainda calcular os valores das derivadas recorrendo à spline parcial $S_i(x)$ que corresponde ao valor de x. As derivadas são então estimadas, recorrendo à expressão (3.3.38), calculando-se os valores da primeira derivada de acordo com,

$$f'(x) \equiv S_i'(x) = -M_{i-1}\,\frac{(x_i-x)^2}{2h_i} + M_i\,\frac{(x-x_{i-1})^2}{2h_i} + \frac{f_i-f_{i-1}}{h_i} - (M_i-M_{i-1})\,\frac{h_i}{6} \,, \tag{5.1.36}$$

e da segunda derivada de acordo com,

$$f''(x) \equiv S_i''(x) = M_{i-1}\,\frac{x_i-x}{h_i} + M_i\,\frac{x-x_{i-1}}{h_i} \,. \tag{5.1.37}$$

5.2 Cálculo numérico de integrais

Por vezes precisamos determinar o valor do integral de uma função num intervalo, conhecendo apenas alguns pontos da função nesse intervalo. Para tal temos de desenvolver uma forma de estimar o valor do integral usando, da melhor forma, a informação disponível sobre a função.

O integral I de uma função $f(x)$, num intervalo $[a,b]$, é escrito como

$$I \equiv \int_a^b f(x)\,\mathrm{d}x \,. \tag{5.2.1}$$

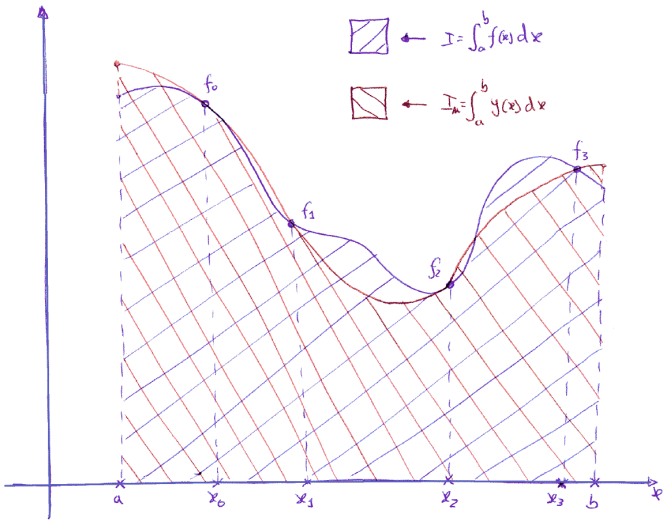


Figura 5.6: Representação da integral da função $f(x)$ e do integral do polinómio interpolador nos pontos (x_i, f_i). Desta forma é possível estimar o valor de I, calculando o valor de I_n, se apenas conhecemos os pontos x_i da função.

Este corresponde a calcular a área definida pela função e o eixo do x's, tal que zonas abaixo do eixo têm uma contribuição negativa.

No entanto se da função, no intervalo $[a,b]$, apenas conhecemos os pontos $\{(x_i, f_i)\}_{i=0}^n$, então para calcular o valor do integral precisamos substituir a função por outra que seja uma representação de $f(x)$, e que seja analiticamente integrável. Tal pode ser feito recorrendo-se a uma função interpoladora de $f(x)$ na tabela de pontos dada. Assim, se $y(x)$ é, por exemplo, o polinómio interpolador de $f(x)$ nos pontos x_i, podemos usar a seguinte expressão para estimar o valor do integral

$$I \simeq I_n \equiv \int_a^b y(x)\,\mathrm{d}x\,, \tag{5.2.2}$$

pois um polinómio é sempre integrável.

Usando a fórmula de Lagrange (Eq. 3.2.20) para o polinómio, temos que

$$y(x) = \sum_{i=0}^{n} f_i\, l_i(x)\,, \tag{5.2.3}$$

onde os $l_i(x)$ são dados por (3.2.19). Substituindo, encontramos que

$$I_n = \int_a^b \left[\sum_{i=0}^{n} f_i l_i(x)\right] \mathrm{d}x = \sum_{i=0}^{n} \left[f_i \int_a^b l_i(x)\,\mathrm{d}x\right] = \sum_{i=0}^{n} f_i A_i\,, \tag{5.2.4}$$

onde definimos

$$A_i \equiv \int_a^b l_i(x)\,\mathrm{d}x\,. \tag{5.2.5}$$

Temos assim uma forma de estimar o integral I da função a partir do valor de I_n, calculado à custa do polinómio interpolador $y(x)$ que usa a informação disponível sobre a função (ver Fig. 5.6).

O erro cometido ao substituir $f(x)$ pelo polinómio interpolador no cálculo do integral é obtido obviamente à custa do erro de interpolação, sendo dado por (ver por exemplo 3.2.59)

$$\varepsilon_n = \left| \int_a^b f[x_0, x_1, ..., x_n, x] \prod_{i=0}^{n} (x - x_i)\mathrm{d}x \right|\,. \tag{5.2.6}$$

Se $\prod_{i=0}^{n}(x-x_i)$ não muda de sinal em $[a,b]$, então pelo teorema do valor médio podemos escrever que

$$\varepsilon_n = \left| f[x_0, x_1, ..., x_n, \eta] \int_a^b \prod_{i=0}^{n} (x - x_i)\mathrm{d}x \right|\,, \tag{5.2.7}$$

onde $\eta \in [a,b]$. Mas, como vimos em (3.2.63), existe um $\xi \in [a,b]$, tal que

$$f[x_0,x_1,...,x_n,\eta] = \frac{f^{(n+1)}(\xi)}{(n+1)!} , \tag{5.2.8}$$

de onde resulta que

$$\varepsilon_n \leq \frac{M_{n+1}}{(n+1)!} \left| \int_a^b \prod_{i=0}^{n} (x-x_i) \mathrm{d}x \right| , \tag{5.2.9}$$

onde

$$M_{n+1} = \max_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right| . \tag{5.2.10}$$

Temos assim uma forma de estimar o erro cometido, se conhecermos M_{n+1}.

No caso de

$$\int_a^b \prod_{i=0}^{n} (x-x_i) \mathrm{d}x = 0 , \tag{5.2.11}$$

então é necessário considerar um ponto $x_{n+1} \in [a,b]$, tal que $\prod_{i=0}^{n+1}(x-x_i)$ não mude de sinal em $[a,b]$, escrevendo nesse caso que

$$f[x_0,x_1,...,x_n,x] = f[x_0,x_1,...,x_n,x_{n+1}] + f[x_0,x_1,...,x_n,x_{n+1},x](x-x_{n+1}) , \tag{5.2.12}$$

de onde segue que

$$\begin{aligned} \varepsilon_n \quad = \quad & \left| \int_a^b f[x_0,x_1,...,x_n,x_{n+1}] \prod_{i=0}^{n} (x-x_i) \mathrm{d}x + \right. \\ & \left. + \int_a^b f[x_0,x_1,...,x_n,x_{n+1},x] \prod_{i=0}^{n+1} (x-x_i) \mathrm{d}x \right| \end{aligned} \tag{5.2.13}$$

Como o primeiro termo é nulo (como decorre de 5.2.11), temos que

$$\varepsilon_n = \left| \int_a^b f[x_0,x_1,...,x_n,x_{n+1},x] \prod_{i=0}^{n+1} (x-x_i) \mathrm{d}x \right| , \tag{5.2.14}$$

de onde resulta que

$$\varepsilon_n \leq \frac{M_{n+2}}{(n+2)!} \left| \int_a^b \prod_{i=0}^{n+1} (x-x_i) \mathrm{d}x \right| , \tag{5.2.15}$$

neste caso particular. Isto é, quando se verifica (5.2.11).

5.2.1 Regras simples de integração

Usemos então a expressão geral dada em (5.2.4) para definir algumas regras simples de integração numérica de funções.

<u>*Regra do rectângulo*</u>: suponhamos que apenas temos um ponto x_0 da função no intervalo $[a,b]$. Nesse caso

$$I_\mathrm{r} = f_0 A_0 , \tag{5.2.16}$$

com

$$A_0 = \int_a^b l_0(x) \, \mathrm{d}x = \int_a^b 1 \, \mathrm{d}x = b-a . \tag{5.2.17}$$

Logo a regra do rectângulo corresponde a

$$I \simeq I_\mathrm{r} = f_0 (b-a) . \tag{5.2.18}$$

Neste caso o erro de integração (ver 5.2.9) é dado por

$$\varepsilon_\mathrm{r} \leq M_1 \left| \int_a^b (x-x_0) \mathrm{d}x \right| = \frac{M_1}{2} \left| (b-x_0)^2 - (x_0-a)^2 \right| \leq \frac{M_1}{2} (b-a)^2 . \tag{5.2.19}$$

No caso particular de $x_0{=}a$ temos o rectângulo à esquerda

$$I \simeq I_{ra} = f_a(b{-}a)\,, \tag{5.2.20}$$

enquanto que, se $x_0{=}b$, temos o rectângulo à direita

$$I \simeq I_{rb} = f_b(b{-}a)\,. \tag{5.2.21}$$

Exemplo 5.2.1: Integremos a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I{=}\sqrt{e}{-}1{=}0.6487$), usando primeiro a regra do rectângulo à esquerda. Temos então, de (5.2.20), que

$$I_{\rm ra} = e^0\,(1/2{-}0) = 0.5000\,.$$

Enquanto que o resultado para a regra do rectângulo à direita, usando (5.2.21), é

$$I_{\rm rb} = e^{1/2}\,(1/2{-}0) = 0.8244\,.$$

Para o erro de integração, em ambos os casos, temos que

$$\varepsilon_{\rm ra} \leq \frac{\sqrt{e}}{2}\,(1/2{-}0)^2 = 0.21\,.$$

<u>*Regra do ponto médio*</u>: neste caso temos a regra do rectângulo mas aplicada ao ponto médio do intervalo $[a,b]$, que é dado por

$$x_0 = \frac{a{+}b}{2}\,. \tag{5.2.22}$$

Logo,

$$I \simeq I_{\rm pm} = f\left(\frac{a{+}b}{2}\right)\,(b{-}a)\,. \tag{5.2.23}$$

Para este caso temos que

$$\int_a^b \left(x - \frac{a{+}b}{2}\right) {\rm d}x = 0\,, \tag{5.2.24}$$

pelo que temos de recorrer à expressão encontrada em (5.2.15). De forma a garantir que $\prod_{i=0}^{n+1}(x{-}x_i)$ não muda de sinal em $[a,b]$, usamos como ponto adicional $x{=}(a{+}b)/2$. Resultando que

$$\varepsilon_{\rm pm} \leq \frac{M_2}{24}(b{-}a)^3\,, \tag{5.2.25}$$

o que significa que esta é uma melhor aproximação para o cálculo do integral do que a defenida pela regra do rectângulo.

Exemplo 5.2.2: Integremos novamente a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I{=}\sqrt{e}{-}1{=}0.6487$), usando desta vez a regra do ponto médio. Temos então, de (5.2.23), que

$$I_{\rm pm} = e^{1/4}\,(1/2{-}0) = 0.6420\,.$$

Para o erro de integração temos que

$$\varepsilon_{\rm pm} \leq \frac{\sqrt{e}}{24}\,(1/2{-}0)^3 = 0.009\,.$$

<u>*Regra do trapézio*</u>: vejamos agora o caso de dispormos de dois pontos da função; $\{(a,f_a),(b,f_b)\}$. O polinómio interpolador é agora

$$y(x) = f_a \cdot l_a(x) + f_b \cdot l_b(x)\,, \tag{5.2.26}$$

de onde se obtém que

$$I_{\rm t} = f_a A_a + f_b A_b\,. \tag{5.2.27}$$

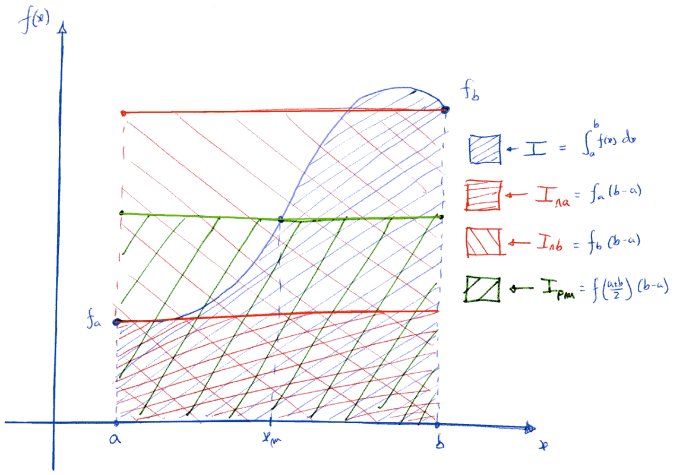


Figura 5.7: Representação da estimativa do integral da função $f(x)$ no intervalo $[a,b]$ obtida pelas regras do rectângulo - à direita e à esquerda (5.2.20-5.2.21), e do ponto médio (5.2.23).

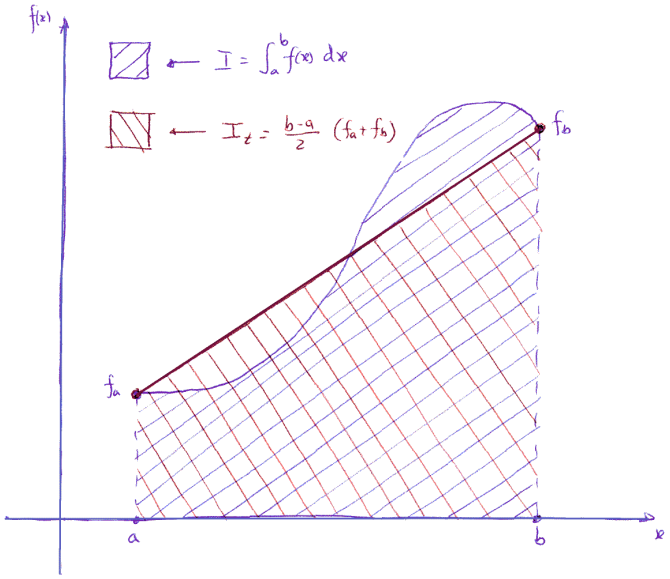


Figura 5.8: Representação da estimativa do integral da função $f(x)$ no intervalo $[a,b]$ obtida pela regra do trapézio (5.2.29).

Das expressões para $l_i(x)$ e de (5.2.5) temos que

$$A_a = \int_a^b \frac{x-b}{a-b}\,\mathrm{d}x = \frac{b-a}{2} \qquad \text{e} \qquad A_b = \int_a^b \frac{x-a}{b-a}\,\mathrm{d}x = \frac{b-a}{2}\,. \tag{5.2.28}$$

Substituindo, temos a regra do trapézio na seguinte forma;

$$I \simeq I_{\mathrm{t}} = \frac{b-a}{2}\,(f_a+f_b)\,. \tag{5.2.29}$$

Neste caso o erro da estimativa é dado por (ver 5.2.9)

$$\varepsilon_{\mathrm{t}} \leq \frac{M_2}{12}\,(b-a)^3\,, \tag{5.2.30}$$

pois

$$\int_a^b (x-a)(x-b)\mathrm{d}x = -\frac{(b-a)^3}{6}\,. \tag{5.2.31}$$

Exemplo 5.2.3: Integremos novamente a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I=\sqrt{e}-1=0.6487$), usando desta vez a regra do trapézio. Temos então, de (5.2.29), que

$$I_t = \frac{1/2-0}{2}\left(e^0+e^{1/2}\right) = 0.6622\,.$$

Para o erro de integração, usando (5.2.30), temos que

$$\varepsilon_t \leq \frac{\sqrt{e}}{12}\,(1/2-0)^3 = 0.018\,.$$

Regra de Simpson: consideremos agora que temos o valor da função em três pontos $\{(a,f_a),(x_m,f_m),(b,f_b)\}$, em que x_m é o ponto médio do intervalo $[a,b]$,

$$x_m \equiv \frac{a+b}{2} \qquad \text{e} \qquad f_m \equiv f(x_m)\,. \tag{5.2.32}$$

O polinómio interpolador é agora

$$y(x) = f_a\cdot l_a(x) + f_m\cdot l_m(x) + f_b\cdot l_b(x)\,, \tag{5.2.33}$$

de onde resulta que

$$I_s = f_a A_a + f_m A_m + f_b A_b\,. \tag{5.2.34}$$

Os coeficientes são agora dados por

$$\begin{aligned} A_a &= \int_a^b \frac{(x-x_m)(x-b)}{(a-x_m)(a-b)}\,dx = \frac{b-a}{6} \\ A_m &= \int_a^b \frac{(x-a)(x-b)}{(x_m-a)(x_m-b)}\,dx = 2\,\frac{b-a}{3} \\ A_b &= \int_a^b \frac{(x-a)(x-x_m)}{(b-a)(b-x_m)}\,dx = \frac{b-a}{6}\,. \end{aligned} \tag{5.2.35}$$

Temos então que a estimativa para o valor do integral é, neste caso, dada por

$$I \simeq I_s = \frac{b-a}{6}\left(f_a + 4\,f_m + f_b\right). \tag{5.2.36}$$

Mais uma vez temos que recorrer a (5.2.15), pois

$$\int_a^b (x-a)(x-x_m)(x-b)dx = 0\,. \tag{5.2.37}$$

Usando como nodo adicional $x=x_m$, de forma a garantir que $(x-a)(x-x_m)^2(x-b)$ tem sempre o mesmo sinal para $x\in[a,b]$, temos então que

$$\varepsilon_s \leq \left|\frac{M_4}{2880}\,(b-a)^5\right|\,. \tag{5.2.38}$$

Exemplo 5.2.4: Integremos novamente a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I=\sqrt{e}-1=0.6487$), usando desta vez a regra de Simpson. Temos então, de (5.2.36), que

$$I_s = \frac{1/2-0}{6}\left(e^0+4\,e^{1/4}+e^{1/2}\right) = 0.6487\,.$$

Para o erro de integração, usando (5.2.38), temos que

$$\varepsilon_s \leq \frac{\sqrt{e}}{2880}\,(1/2-0)^5 = 0.00002\,.$$

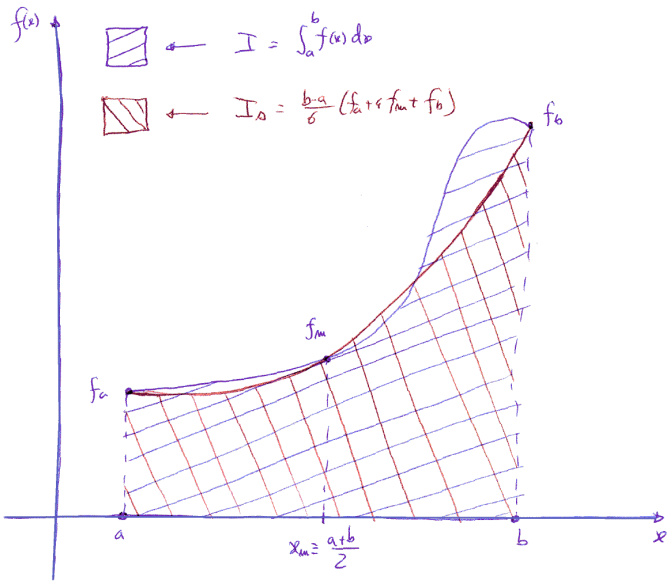


Figura 5.9: Representação da estimativa do integral da função $f(x)$ no intervalo $[a,b]$ obtida pela regra de Simpson (5.2.36).

5.2.2 Regras compostas de integração

O facto do erro de estimação do integral pelas regras que obtivemos na Secção anterior ser proporcional a uma potência de $(b-a)$, sugere-nos que tentemos reduzir o erro final da estimativa através da redução desta largura do intervalo. Isto é, usamos as regras atrás descritas mas para intervalos de muito menor amplitude de forma a reduzir tanto quanto possível o erro com que estimamos o valor do integral. Obviamente que estamos sempre limitados pela informação que dispomos/podemos dispôr da função $f(x)$.

Consideremos então que introduzimos a seguinte tabela de pontos (ordenados) da função $f(x)$ no intervalo $[a,b]$

$$\Big\{(x_i,f_i)\Big\}_{i=0}^{n} \qquad \text{com} \qquad h_i \equiv x_i - x_{i-1} \qquad \text{para} \quad i=1,2,...,n\,, \tag{5.2.39}$$

com $x_0{=}a$ e $x_n{=}b$. Então

$$I \equiv \int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x\,. \tag{5.2.40}$$

Usemos agora as regras de integração apresentadas antes para estimar o valor do integral em cada um dos intervalos $[x_{i-1},x_i]$, tal como representado por

$$I_i \equiv \int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x\,. \tag{5.2.41}$$

Nesse caso, o erro de integração, passa a ser dado por

$$\varepsilon \equiv \sum_{i=1}^{n} \varepsilon_i\,, \tag{5.2.42}$$

em que ε_i é o erro cometido para cada um dos intervalos, cuja largura é agora h_i, podendo ser portanto muito menor do que o erro inicial que era proporcional a $b-a$.

<u>*Regra do ponto médio composta*</u>: comecemos então por considerar que usamos a regra do ponto médio (5.2.23) para estimar cada um dos integrais I_i;

$$I_i \simeq h_i\, f\left(\frac{x_{i-1}+x_i}{2}\right)\,, \tag{5.2.43}$$

com um erro (ver 5.2.25)

$$\varepsilon_i \leq \frac{M_{2i}}{24}\, h_i^3\,. \tag{5.2.44}$$

Assim, obtemos que

$$I \simeq I_{\mathrm{pmc}} = \sum_{i=1}^{n} \left[h_i\, f\left(\frac{x_{i-1}+x_i}{2}\right)\right]\,, \tag{5.2.45}$$

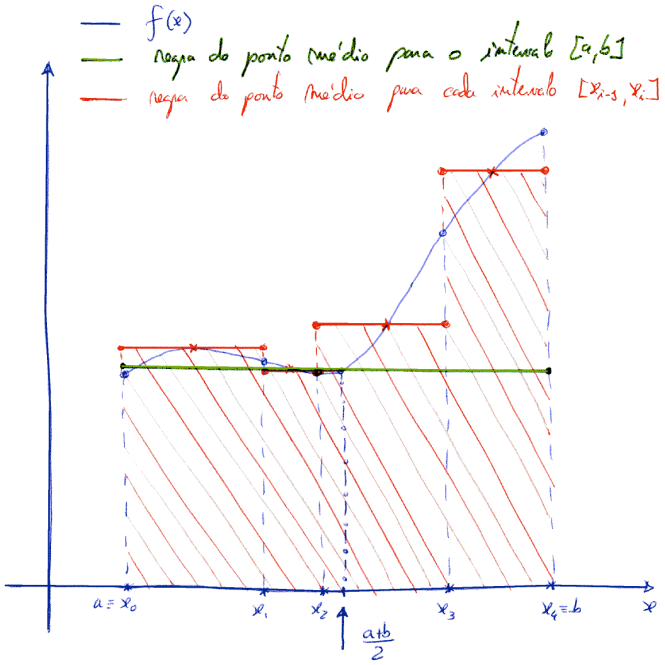


Figura 5.10: Representação do efeito de aplicar a regra do ponto médio composta (5.2.45) para estimar o integral da função $f(x)$ no intervalo $[a,b]$. Para comparação, também se indica a regra simples do ponto médio para o mesmo intervalo.

com um erro

$$\varepsilon_{\rm pmc} \leq \sum_{i=1}^{n} \frac{M_{2i}}{24}\, h_i^3 \,. \tag{5.2.46}$$

No caso particular de $h_i{=}h$ (para todo o $i{=}1,\ldots,n$), então

$$I_{\rm pmc} = h \sum_{i=1}^{n} f\left(\frac{x_{i-1}{+}x_i}{2}\right) , \tag{5.2.47}$$

com um erro

$$\varepsilon_{\rm pmc} \leq \frac{h^3}{24} \sum_{i=1}^{n} M_{2i} = \frac{n\,h^3}{24} M_2 = \frac{M_2}{24}\, h^2 (b{-}a) \,, \tag{5.2.48}$$

já que $b{-}a{=}nh$.

Exemplo 5.2.5: Integremos a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I{=}\sqrt{\mathrm{e}}{-}1{=}0.6487$), usando desta vez a regra do ponto médio composta para quatro intervalos igualmente espaçados; $h{=}1/8$. Temos então, de (5.2.47), que

$$I_{\rm pmc} = \frac{1}{8}\left(\mathrm{e}^{1/16} + \mathrm{e}^{3/16} + \mathrm{e}^{5/16} + \mathrm{e}^{7/16}\right) = 0.6483 \,.$$

Para o erro de integração, usando (5.2.48), temos que

$$\varepsilon_{\rm pmc} \leq \frac{\sqrt{\mathrm{e}}}{24}\,(1/8)^2(1/2{-}0) = 0.0006 \,.$$

Regra do trapézio composta: se recorrermos agora à regra do trapézio (5.2.29) para calcular o valor de I_i temos que

$$I \simeq I_{\rm tc} = \sum_{i=1}^{n} \frac{h_i}{2}\,[f(x_{i-1}){+}f(x_i)] \,. \tag{5.2.49}$$

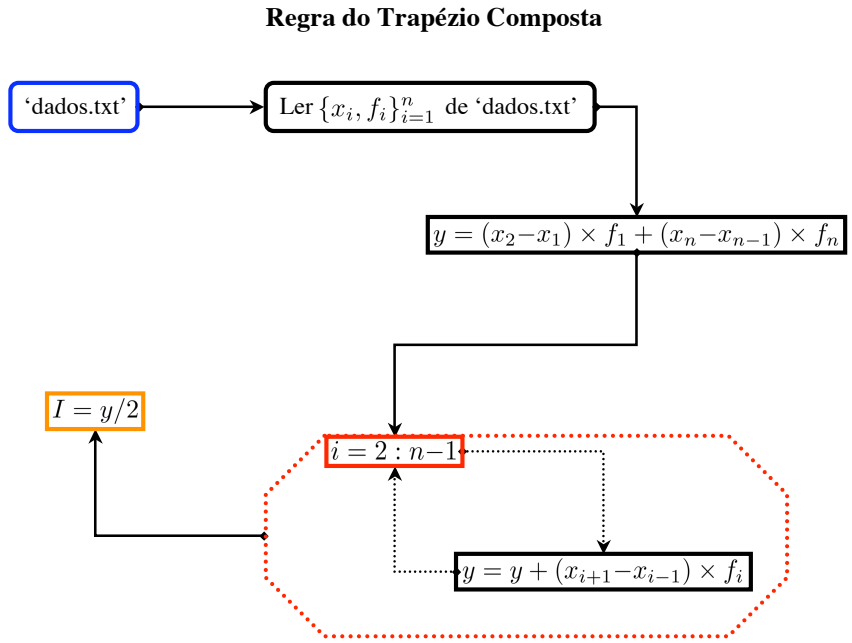


Figura 5.11: Algoritmo para implementação do regra do traézio composta que permite calcular o integral $I = \int_{x_1}^{x_n} f(x)\,\mathrm{d}x$, a partir da tabela $\{x_i, f_i\}_{i=1}^n$.

Logo, temos que

$$I_{\text{tc}} = \frac{1}{2}\left[h_1 f(x_0) + \sum_{i=1}^{n-1}\left(h_i + h_{i+1}\right) f(x_i) + h_n f(x_n)\right] , \tag{5.2.50}$$

cujo erro é majorado por

$$\varepsilon_{\text{tc}} = \sum_{i=1}^{n} \varepsilon_i \leq \sum_{i=1}^{n} \frac{M_{2i}}{12}\, h_i^3 . \tag{5.2.51}$$

Mais uma vez, para pontos igualmente espaçados ($h_i = h$), temos que

$$I_{\text{tc}} = \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right] , \tag{5.2.52}$$

e

$$\varepsilon_{\text{tc}} \leq \frac{h^3}{12}\sum_{i=1}^{n} M_{2i} \leq \frac{h^3}{12}\, nM_2 = \frac{M_2}{12}(b-a)h^2 , \tag{5.2.53}$$

sendo que $M_2 \equiv \max_{\xi\in[x_0,x_n]} |f''(\xi)| = \max_{i\in[1,n]} M_{2i}$.

Exemplo 5.2.6: Integremos novamente a função e^x no intervalo $[0, 1/2]$ (cujo resultado sabemos ser $I = \sqrt{\mathrm{e}} - 1 = 0.6487$), usando desta vez a regra do trapézio composta para quatro intervalos igualmente espaçados; $h = 1/8$. Temos então, de (5.2.52), que

$$I_{\text{tc}} = \frac{1}{16}\left(\mathrm{e}^0 + 2\mathrm{e}^{1/8} + 2\mathrm{e}^{1/4} + 2\mathrm{e}^{3/8} + \mathrm{e}^{1/2}\right) = 0.6496 .$$

Para o erro de integração, usando (5.2.53), temos que

$$\varepsilon_{rmtc} \leq \frac{\sqrt{\mathrm{e}}}{12}\,(1/8)^2(1/2 - 0) = 0.0012 .$$

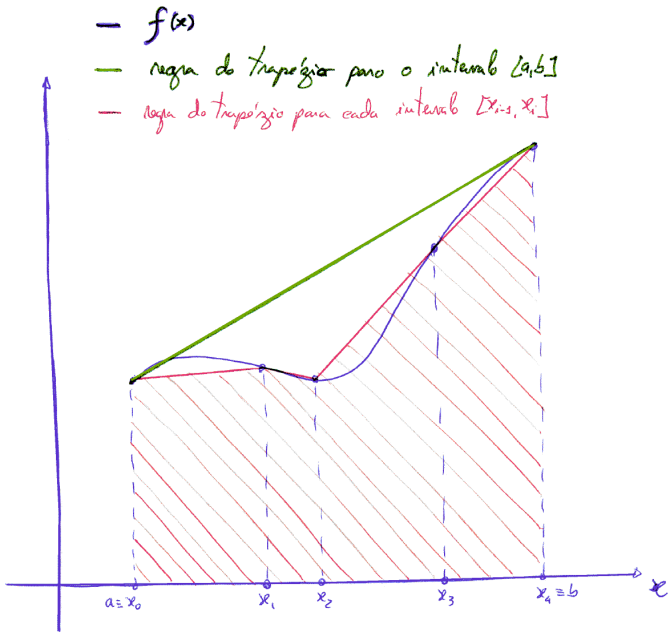


Figura 5.12: Representação do efeito de aplicar a regra do trapézio composta (5.2.50) para estimar o integral da função $f(x)$ no intervalo $[a,b]$. Para comparação, também se indica a regra simples do trapézio para o mesmo intervalo.

Regra de Simpson composta: se recorrermos agora à regra de Simpson (5.2.36) para calcular o valor de I_i temos que

$$I \simeq I_{\rm sc} = \sum_{i=1}^{n} \frac{h_i}{6}\left[f(x_{i-1})+4f\left(\frac{x_{i-1}+x_i}{2}\right)+f(x_i)\right] . \tag{5.2.54}$$

Logo, resulta que

$$I_{\rm sc} = \frac{1}{6}\left[h_1 f(x_0)+\sum_{i=1}^{n-1}\left(h_i+h_{i+1}\right)f(x_i)+4\sum_{i=1}^{n} h_i f\left(\frac{x_{i-1}+x_i}{2}\right)+h_n f(x_n)\right] , \tag{5.2.55}$$

cujo erro é majorado por

$$\varepsilon_{\rm sc} = \sum_{i=1}^{n}\varepsilon_i \leq \sum_{i=1}^{n}\frac{M_{4i}}{2880}\,h_i^5 . \tag{5.2.56}$$

Mais uma vez, para pontos igualmente espaçados (h_i=h), temos que

$$I_{\rm sc} = \frac{h}{6}\left[f(x_0)+2\sum_{i=1}^{n-1} f(x_i)+4\sum_{i=1}^{n} f\left(\frac{x_{i-1}+x_i}{2}\right)+f(x_n)\right] , \tag{5.2.57}$$

com o erro dado por

$$\varepsilon_{\rm sc} \leq \frac{h^5}{2880}\sum_{i=1}^{n} M_{4i} = \frac{h^5}{2880}\,nM_4 = \frac{M_4}{2880}(b-a)h^4 , \tag{5.2.58}$$

sendo que $M_4 \equiv \max_{\xi\in[x_0,x_n]}\left|f^{(4)}(\xi)\right| = \max_{i\in[1,n]} M_{4i}$.

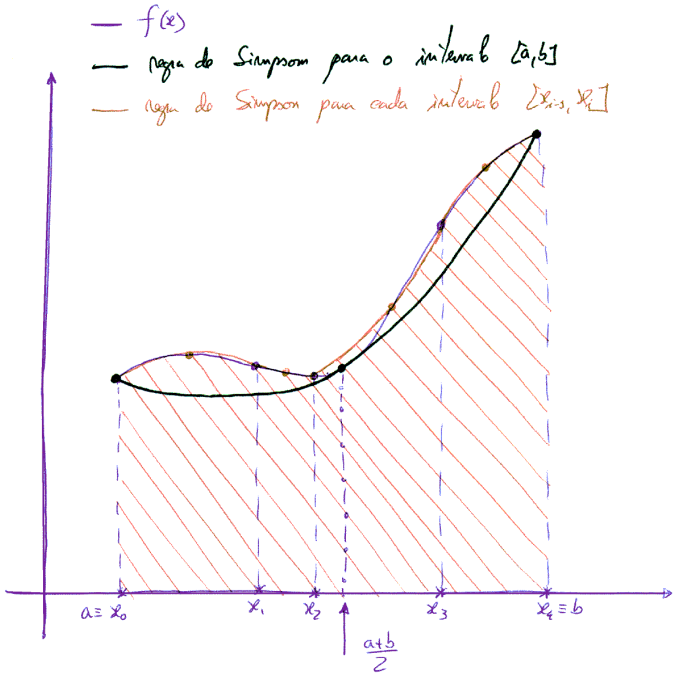


Figura 5.13: Representação do efeito de aplicar a regra de Simpson composta (5.2.55) para estimar o integral da função $f(x)$ no intervalo $[a,b]$. Para comparação, também se indica a regra simples de Simpson para o mesmo intervalo.

Exemplo 5.2.7: Integremos novamente a função e^x no intervalo $[0,1/2]$ (cujo resultado sabemos ser $I=\sqrt{e}-1=0.6487$), usando desta vez a regra de Simpson composta para quatro intervalos igualmente espaçados; $h=1/8$. Temos então, de (5.2.57), que

$$\begin{aligned} I_{sc} &= \frac{1}{48}\left[e^0+2\left(e^{1/8}+e^{1/4}+e^{3/8}\right)+4\left(e^{1/16}+e^{3/16}+e^{5/16}+e^{7/16}\right)+e^{1/2}\right] \\ &= 0.6487\,. \end{aligned}$$

Para o erro de integração, usando (5.2.58), temos que

$$\varepsilon_{sc} \leq \frac{\sqrt{e}}{2880}\,(1/8)^4(1/2-0) = 0.00000007\,.$$

5.2.3 Cálculo do integral recorrendo a splines cúbicas

Não precisamos recorrer unicamente a polinómios interpoladores de forma a substituir a função no integral. Tal pode ser também feito por outras funções (e em particular por splines) que também sejam integráveis.

Seja então $\{(x_i,f_i)\}_{i=0}^n$ uma tabela de pontos de uma função $f(x)$ (com $a=x_0$ e $b=x_n$). Se a spline cúbica interpoladora de $f(x)$ nos pontos desta tabela é $S(x)$, com as splines parciais dadas por

$$\begin{aligned} S_i(x) &= M_{i-1}\frac{(x_i-x)^3}{6h_i}+M_i\frac{(x-x_{i-1})^3}{6h_i}+ \\ &\quad +\left(f_{i-1}-M_{i-1}\frac{h_i^2}{6}\right)\frac{x_i-x}{h_i}+\left(f_i-M_i\frac{h_i^2}{6}\right)\frac{x-x_{i-1}}{h_i}\,, \end{aligned} \tag{5.2.59}$$

para $x\in[x_{i-1},x_i]$ (com $i=1,2,\ldots,n$), então podemos usar esta função interpoladora para estimar o valor do integral de $f(x)$;

$$I \equiv \int_a^b f(x)\,dx \simeq \int_a^b S(x)\,dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} S_i(x)\,dx \equiv \sum_{i=1}^n I_i\,. \tag{5.2.60}$$

Substituimos a expressão (5.2.59), de forma a obter que

$$I_i \equiv \int_{x_{i-1}}^{x_i} S_i(x)\,\mathrm{d}x = \frac{h_i}{2}\,(f_{i-1}+f_i) - \frac{h_i^3}{24}\,(M_{i-1}+M_i)\,. \tag{5.2.61}$$

De onde concluímos que

$$I \simeq \sum_{i=1}^{n} \frac{h_i}{2}\,\left[(f_{i-1}+f_i) - \frac{h_i^2}{12}\,(M_{i-1}+M_i)\right]\,, \tag{5.2.62}$$

é a estimativa que temos para o integral de $f(x)$ no intervalo $[a,b]$, quando usamos splines.

Exemplo 5.2.8: Consideremos a tabela de três pontos para uma função real, $\{(0,1),(1,0),(4,2)\}$, cujo integral em $[0,4]$ pretendemos estimar usando agora uma spline cúbica natural. Como já calculado no Exemplo 3.3.4, os parâmetros da spline cúbica natural que interpola a função nestes pontos são M_0=0, M_1=5/4 e M_2=0. Podemos então calcular o valor de integral recorrendo a (5.2.62),

$$\begin{aligned}\int_0^4 f(x)\,\mathrm{d}x \quad &\simeq \quad \frac{1-0}{2}\left[(1+0) - \frac{(1-0)^2}{12}(0+5/4)\right] + \\ &\quad + \frac{4-1}{2}\left[(0+2) - \frac{(4-1)^2}{12}(5/4+0)\right] \simeq 2.5729\,.\end{aligned}$$

5.2.4 Cálculo do integral recorrendo a outras funções interpoladoras

Qualquer função interpoladora da tabela de valores para os quais queremos determinar o integral pode ser usada de forma a estimar o integral. Desde que seja uma função interpoladora integrável analiticamente temos vantagem em a usar, pois esta descreve o comportamento esperado da função $f(x)$ pelo que a estimativa do integral da função será mais aproximado do valor real.

5.3 Exercícios

E5.1) Conhecem-se os seguintes valores tabelados de uma função real $f(x)$ em três pontos; $f(0.0)=-0.5$, $f(0.8)=0.0$ e $f(2.0)=1.0$, bem como um valor da sua derivada; $f'(2.0)=0.0$. Estime o valor de $f'(1)$ tal que o valor encontrado seja exacto se $f(x)$ é um polinómio de grau 2 ou inferior.

E5.2) Conhecem-se os seguintes valores tabelados de uma função real $f(x)$ em três pontos: $f(x_0=-1)=0.5$, $f(x_1=0)=0.0$ e $f(x_2=1)=-0.1$.

a) Considerando que o valor aproximado da derivada de f num ponto x pode ser dado por uma relação do tipo

$$f'(x) \sim a(h_1,h_2)\cdot f(x_1) + b(h_1,h_2)\cdot f(x_2)$$

onde $h_i=x-x_i$ (com $i=1,2$), calcule as expressões para os coeficientes (a,b) que tornam esta expressão exacta pelo menos para polinómios de grau um ou inferior.

b) Utilize a expressão obtida para determinar uma estimativa para o valor de $f'(0.4)$.

E5.3) Conhecem-se os seguintes valores de uma função $f(x)$ em quatro pontos: $f(x_0=0)=0.0$, $f(x_1=1)=1.5$, $f(x_2=2)=0.5$ e $f(x_3=3)=0.0$. Escrevendo a derivada de f como;

$$f'(x) \sim a(x)\, f(x_0) + b(x)\, f(x_1) + c(x)\, f(x_2)\,,$$

calcule as expressões dos polinómios (a,b,c) de forma a que esta relação seja exacta para funções polinomiais do maior grau possível. Utilize-a para calcular o valor de $f'(1.8)$.

E5.4) Pretende-se aproximar $f''(x)$ por uma fórmula do tipo;

$$f''(x) \sim a_1 f(x+h) + b_1 f'(x+h) + a_2 f(x-h) + b_2 f'(x-h)\,.$$

Quais os valores dos coeficientes para que a fórmula seja exacta para polinómios do maior grau possível?

E5.5) Utilizando a tabela de valores dados ao lado para a função $f(x)=\sinh(x)$ calcule $f'(0.4)$ recorrendo à fórmula de diferenças centrais de 2^a ordem para $h=0.001$ e $h=0.002$. Compare os resultados.

x	$f(x)$
0.398	0.40859
0.399	0.40967
0.400	0.41075
0.401	0.41183
0.402	0.41292

E5.6) Deduza a seguinte fórmula de aproximação para a segunda derivada obtida a partir do polinómio interpolador nos pontos x_0, x_1 e x_2, em que $x_1-x_0=h$ e $x_2-x_1=\alpha h$;

$$f''(x) \sim \frac{2}{h^2}\left[\frac{f(x_0)}{1+\alpha} - \frac{f(x_1)}{\alpha} + \frac{f(x_2)}{\alpha(\alpha+1)}\right]\,.$$

Mostre que para $\alpha=1$ se tem a fórmula das diferenças centrais.

E5.7) Seja $g(x)$ uma função real definida em $\mathbb{R}$.

a) Construa o polinómio interpolador de $g(x)$ nos seguintes pontos: $g(-1)=-0.1$, $g(0)=0.0$ e $g(3)=-0.2$, usando a expressão de Newton.

b) Cálcule o erro na determinação de $\int_a^b g(x)\,\mathrm{d}x$ se for usada a regra do ponto médio, no caso de $g(x)$ representar uma parábola. Em que condições o valor dado pela regra do ponto médio coincide com o valor da regra do trapézio?

c) Mostre que no cálculo aproximado da derivada através da fórmula de diferença central de segunda ordem o resultado é exacto se a função for uma parábola. Represente gráficamente a aproximação á derivada que tal fórmula fornece no caso de uma função geral.

E5.8) Empregue a regra do trapézio para calcular o integral

$$I = \int_0^1 x\, e^{-x^2}\, \mathrm{d}x\,.$$

Compare o resultado obtido para I com a soma resultante da aplicação da regra composta a N intervalos; $\sum_{n=1}^{N} I_n$, nos casos de N=2 e N=4.

E5.9) Pretende-se calcular pela regra do trapézio o integral

$$I = \int_0^1 e^{e^x} \mathrm{d}x\,,$$

com erro inferior a 10^{-4}. Quantos pontos se devem usar?

E5.10) Pretende-se calcular o valor de $\int_a^b f(x)\,\mathrm{d}x$ sabendo-se apenas os seguintes valores: $f(a) = f_1$, $f'(a)$=0 e $f(b)$=f_2.

a) Construa o polinómio interpolador de $f(x)$, usando toda esta informação, pelo método de Aitken-Neville. Mostre igualmente que este pode ser escrito como,

$$y(x) = f_1\,\frac{b-x}{b-a}\,\left(1+\frac{x-a}{b-a}\right) + f_2\,\left(\frac{x-a}{b-a}\right)^2\,.$$

b) Aproximando a função $f(x)$ por $y(x)$ no integral, mostre que se obtem a seguinte estimativa para o seu valor,

$$\int_a^b f(x)\,\mathrm{d}x \sim \frac{2f_1+f_2}{3}\,(b-a)\,.$$

c) Calcule o valor no caso de ter (a,b)=$(0,1)$ e (f_1,f_2)=$(1.0000, 1.7183)$. Compare com o valor obtido se usar a regra do trapézio (ignorando neste caso o facto de saber a derivada da função em a), e comente a diferença.

E5.11) Seja $I(x)$=$\int_0^x e^{-t^2}\mathrm{d}t$, uma função real definida em $[0,+\infty]$.

a) Dividindo o intervalo $[0,x]$ em N intervalos de dimensão h, use a regra do trapézio para calcular $I(t_i+h)-I(t_i)$ (com i=$0,..,N-1$ e t_i=$i\cdot h$) e mostre que a regra do trapézio composta para N=$(b-0)/h$ intervalos dá a seguinte aproximação para $I(b)$;

$$I(b) \equiv \int_0^b e^{-t^2}\mathrm{d}t \sim \frac{h}{2}\left[1+2\sum_{i=1}^{N-1} e^{-(ih)^2} + e^{-b^2}\right]$$

b) Calcule o valor de $\int_0^\infty e^{-t^2}\mathrm{d}t$, com um erro inferior a 0.05, sabendo que o erro ao estimar o valor de $\lim_{x\to\infty} I(x)$ pela regra do trapézio composta (com h=0.4) é majorado por 0.03.

E5.12) Seja $g(x)$ uma função real definida em $\mathbb{R}$.

a) Construa o polinómio interpolador de $g(x)$ usando os seguintes pontos: $g(-1)$=-1.1, $g(0)$=0.2 e $g(2)$=0.0. Indique para o caso geral de se ter n pontos, em que condições o polinómio aproximador (no sentido dos mínimos quadrados) coincide com o polinómio interpolador nesses pontos.

b) Mostre que no cálculo de $\int_a^b g(x)\,\mathrm{d}x$ a regra de Simpson é exacta se $g(x)$ representar uma parábola. Porquê?

c) Mostre que o mesmo também acontece para o cálculo aproximado da derivada através da fórmula de diferença central de primeira ordem. Represente gráficamente a aproximação á derivada que esta fórmula fornece no caso geral.

E5.13) Seja $f(x)$ uma função real que descreve o comportamento de um sistema físico. Por experimentação mediram-se os seguintes valores: $f(-1)$=0.736, $f(0)$=1.99 e $f(+1)$=5.44.

a) Calcule o valor aproximado de $I = \int_{-1}^{+1} x^2 f(x)\,\mathrm{d}x$ recorrendo á regra de Simpson para o cálculo de integrais. Represente gráficamente o valor calculado para I.

b) Sabendo que o erro absoluto no cálculo do integral pela regra de Simpson é majorado por 0.004 e que os valores medidos para $f(x)$ têm um erro relativo associado de 10^{-3}, determine o erro relativo com que determinou o valor de I na alinea anterior.

Capítulo 6

Resolução numérica de equações diferenciais

Nem sempre é possível determinar a solução analítica de um sistema de equações diferenciais. Nesse caso torna-se necessário recorrer ao cálculo numérico dessas soluções impondo condições iniciais ou condições fronteira. A título de exemplo, nesta secção, vamos considerar um dos métodos mais simples de resolver sistemas de equações diferenciais sujeitos a condições iniciais ou a condições fronteira.

6.1 Problemas de valor inicial

Consideremos o caso simples de conhecermos um ponto da função a determinar (valor inicial) a partir do qual podemos construir a função usando uma equação diferencial que nos descreve o modo como esta varia.

6.1.1 Uma equação diferencial de primeira ordem

Consideremos uma equação diferencial para uma função $u(t)$;

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t,u(t)) \,, \tag{6.1.1}$$

sujeita à condição inicial: $u(t_0)=u_0$.

Resolver a equação significa determinar o valor da função u para $t\neq t_0$. Isto é, calcular

$$u(t) = u(t_0) + \int_{t_0}^{t} \frac{\mathrm{d}u}{\mathrm{d}\xi}\,\mathrm{d}\xi = u_0 + \int_{t_0}^{t} f(\xi,u(\xi))\,\mathrm{d}\xi \,. \tag{6.1.2}$$

Logo é necessário estimar numericamente o valor de

$$\Delta u(t) \equiv u(t) - u(t_0) = \int_{t_0}^{t} f(\xi,u(\xi))\,\mathrm{d}\xi \,. \tag{6.1.3}$$

Recorrendo por exemplo à regra do rectângulo à esquerda podemos escrever que

$$\Delta u(t) \simeq f(t_0,u(t_0)) \cdot (t-t_0) \,. \tag{6.1.4}$$

6.1.2 Método de Runge-Kutta de segunda ordem

Tentemos então usar uma melhor forma de estimar $\Delta u(t)$. Consideremos a regra do trapézio para obter que

$$\int_{t_0}^{t} f(\xi,u(\xi))\,\mathrm{d}\xi \simeq \frac{t-t_0}{2}\left[f(t_0,u_0) + f(t,u(t))\right] \,, \tag{6.1.5}$$

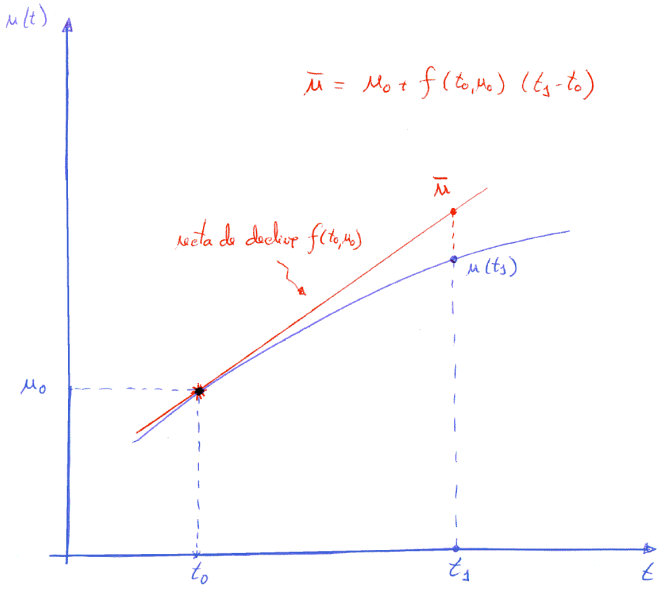


Figura 6.1: Exemplo do significado de usar a derivada da função $u(t)$ em t_0 para estimar o valor que esta toma em $u(t)$. Isto corresponde a resolver a equação diferencial (6.1.1) recorrendo a estimativa dada por (6.1.4).

de onde segue, definindo-se $h \equiv t - t_0$, que

$$u(t) \simeq u_0 + \frac{h}{2}\left[f(t_0,u_0) + f(t,u(t))\right] . \tag{6.1.6}$$

No entanto não podemos usar esta expressão pois necessitamos de $u(t)$ no lado direito para obter o valor pretendido. Assim, é necessário substituir $u(t)$ na expressão da direita por uma estimativa. Tal é feito recorrendo a (6.1.4), de forma que

$$\begin{aligned} u_h &= u_0 + h \cdot f(t_0,u_0) \\ u(t) &\simeq u_0 + \frac{h}{2}\left[f(t_0,u_0) + f(t,u_h)\right] . \end{aligned} \tag{6.1.7}$$

Temos assim uma expressão para estimar o valor de $u(t)$ que é melhor do que a obtida em (6.1.4), pois aí o erro (no cálculo do integral) era proporcional a $(t-t_0)^2$ enquanto que para a (6.1.7) o erro é proporcional a $(t-t_0)^3$.

A implementação do método de Runge-Kutta de segunda ordem pode ser resumida da seguinte forma; seja $u(t)$ uma função que obedece à seguinte equação diferencial,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t,u(t)) . \tag{6.1.8}$$

Se temos que $u(t_i)=u_i$, então para determinar $u_{i+1}=u(t_{i+1})$, com $h \equiv t_{i+1} - t_i$, procedemos da seguinte forma; calcula-se,

$$\begin{cases} F_1 = f(t_i,u_i) \\ F_2 = f(t_i+h, u_i+hF_1) , \end{cases} \tag{6.1.9}$$

de onde se obtém que

$$u_{i+1} = u_i + \frac{h}{2}(F_1+F_2) . \tag{6.1.10}$$

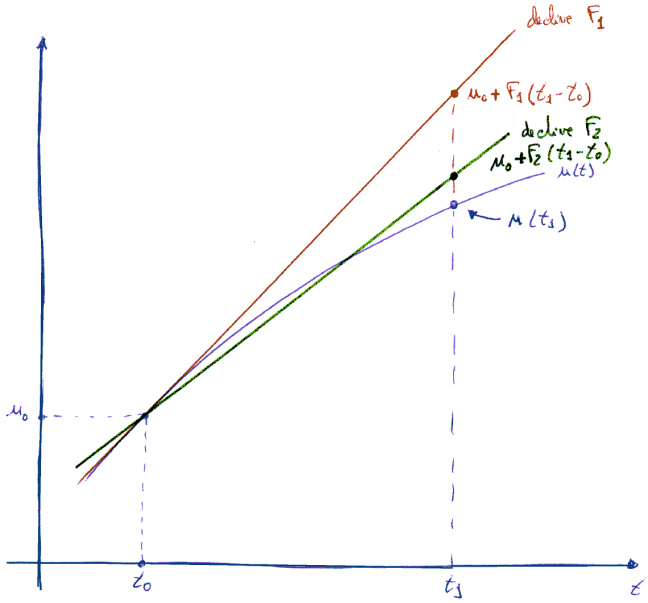


Figura 6.2: Representação gráfica do método de Runge-Kutta de segunda ordem que permite estimar o valor da função em t a partir do valor desta em t_0 e da sua derivada tal como dada por $f(t,u)$.

Exemplo 6.1.1: Consideremos a seguinte equação diferencial,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = t \cdot u(t) \,,$$

com a seguinte condição inicial: $u(0)=1$, cuja solução sabemos ser $u(t)=\mathrm{e}^{t^2/2}$. Usemos o método de Runge-Kutta de segunda ordem para estimar o valor de $u(0.1)$. Como $f(t,u)=t\cdot u$, recorrendo a (6.1.96.1.9) temos que (com $h=0.1$)

$$\begin{aligned} F_1 &= f(0,1) = 0 \\ F_2 &= f(0+h, 1+h\cdot 0) = h \,, \end{aligned}$$

de onde se obtém que

$$u(0.1) \simeq u(0) + \frac{h}{2}(F_1+F_2) = 1 + \frac{h}{2}(0+h) = 1.005 \,.$$

O valor exacto é $u(0.1)=\mathrm{e}^{0.005}=1.0050125$.

6.1.3 Método de Runge-Kutta de quarta ordem

Podemos no entanto reduzir o erro envolvido na resolução da equação recorrendo a uma melhor forma de estimar o integral dado em (6.1.3). Tal corresponde, por exemplo, a usar a regra de Simpson - isto é

$$\int_{t_0}^{t} f(\xi, u(\xi))\, \mathrm{d}\xi \simeq \frac{t-t_0}{6} \left[f(t_0,u_0) + 4f(t_m, u(t_m)) + f(t, u(t)) \right] \,, \tag{6.1.11}$$

onde $t_m \equiv (t+t_0)/2$ coresponde ao valor médio do intervalo $[t_0,t]$. Mais uma vez temos o problema de usar uma expressão deste tipo pois desconhecemos o valor de $u(t_m)$ e de $u(t)$ para substituir na expressão obtida para o

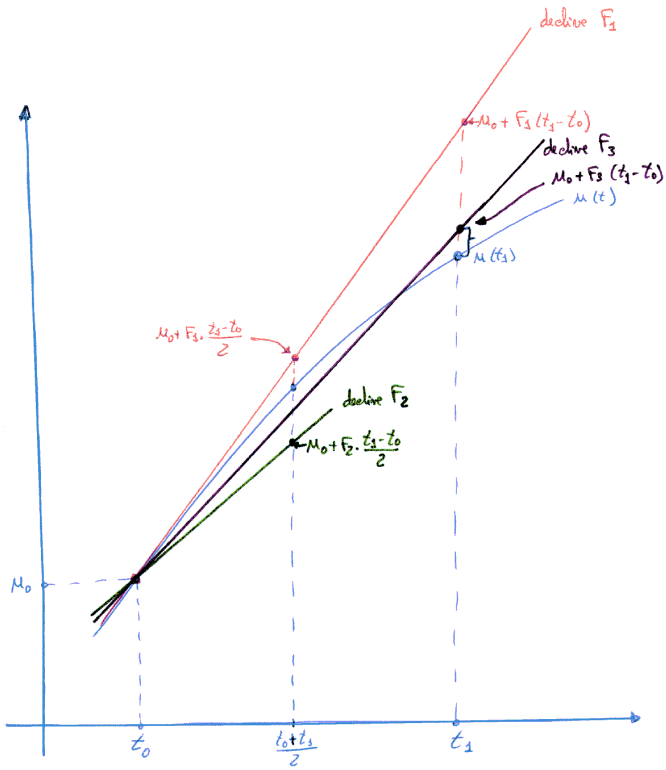


Figura 6.3: Representação gráfica do método de Runge-Kutta de quarta ordem que permite estimar o valor da função em t a partir do valor desta em t_0 e da sua derivada tal como dada por $f(t,u)$.

integral. Assim recorremos a estimativas destes valores escrevendo, para $h \equiv t - t_0$, que

$$\begin{cases} f_0 \equiv f(t_0, u_0) \\ f(t_m, u(t_m)) \simeq f_{m1} \equiv f\left(t_m, u_0 + \dfrac{h}{2} \cdot f_0\right) \\ f(t_m, u(t_m)) \simeq f_{m2} \equiv f\left(t_m, u_0 + \dfrac{h}{2} \cdot f_{m1}\right) \\ f(t, u(t)) \equiv f_t \simeq f\left(t, u_0 + h \cdot f_{m2}\right) \ . \end{cases} \tag{6.1.12}$$

Temos assim estimativas para usar em (6.1.11) obtendo que

$$u(t) \simeq u_0 + \frac{t - t_0}{6} \left[f_0 + (2 f_{m1} + 2 f_{m2}) + f_t \right] \ . \tag{6.1.13}$$

Como estamos a usar a fórmula de Simpson para estimar o valor do integral, temos que o erro associado a este método de integração da equação diferencial é proporcional a $(t - t_0)^5$.

Mais uma vez, a implementação deste método (Runge-Kutta de quarta ordem) é simples, podendo ser resumida da seguinte forma; seja $u(t)$ uma função que obedece à seguinte equação diferencial,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u(t)) \ . \tag{6.1.14}$$

Se temos que $u(t_i) = u_i$, então para determinar $u_{i+1} = u(t_{i+1})$, com $h \equiv t_{i+1} - t_i$, procedemos da seguinte forma: calcula-se,

$$\begin{cases} F_1 = f(t_i, u_i) \\ F_2 = f\left(t_i + \dfrac{h}{2}, u_i + \dfrac{h}{2} F_1\right) \\ F_3 = f\left(t_i + \dfrac{h}{2}, u_i + \dfrac{h}{2} F_2\right) \\ F_4 = f(t_i + h, u_i + h F_3) \ , \end{cases} \tag{6.1.15}$$

de onde se obtém que

$$u_{i+1} = u_i + \frac{h}{6}(F_1+2F_2+2F_3+F_4)\ . \tag{6.1.16}$$

Exemplo 6.1.2: Consideremos mais uma vez a seguinte equação diferencial,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = t\cdot u(t)\ ,$$

com a condição inicial: $u(0)$=1, e cuja solução já sabemos ser $u(t)$=$\mathrm{e}^{t^2/2}$. Usemos agora o método de Runge-Kutta de quarta ordem para estimar o valor de $u(0.1)$. Como $f(t,u)$=$t\cdot u$, recorrendo a (6.1.15) temos que (com h=0.1)

$$\begin{aligned}
F_1 &= f(0,1) = 0\\
F_2 &= f\left(0+\frac{h}{2}, 1+\frac{h}{2}\cdot 0\right) = \frac{h}{2}\\
F_3 &= f\left(0+\frac{h}{2}, 1+\frac{h}{2}\cdot\frac{h}{2}\right) = \frac{h}{2}\left(1+\frac{h^4}{4}\right)\\
F_4 &= f\left(0+h, 1+h\cdot\frac{h}{2}\left(1+\frac{h^4}{4}\right)\right) = h\cdot\left[1+\frac{h^2}{2}\left(1+\frac{h^4}{4}\right)\right]\ ,
\end{aligned}$$

de onde se obtém que

$$\begin{aligned}
u(0.1) &\simeq u(0)+\frac{h}{6}(F_1+2F_2+2F_3+F_4)\\
&= 1+\frac{h^2}{6}\left[2+\left(1+\frac{h^4}{4}\right)\left(1+\frac{h^2}{2}\right)\right] = 1.0050084\ .
\end{aligned}$$

O valor exacto é $u(0.1)$=$\mathrm{e}^{0.005}$=1.0050125.

Uma das formas de diminuir o erro na estimativa do valor de $u(t)$ é recorrermos à integração da equação diferencial por passos, pois o erro na utilização da regra de Runge-Kutta é proporcional a uma potência de h. Isto é, a integração entre t_0 e t pela regra the Runge-Kutta de quarta ordem pode ser substituida pela integraçã com a Regra de Runge-Kutta de segunda ordem de t_0 a $(t+t_0)/2$, seguida da integração de $(t+t_0)/2$ a t.

Exemplo 6.1.3: Consideremos novamente a mesma equação diferencial,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = t \cdot u(t) \qquad \text{com} \qquad u(0)=1 \,,$$

e estimemos o valor de $u(0.1)$ recorrendo ao método de Runge-Kutta de segunda ordem, mas para os intervalos $[0,0.05]$ e $[0.05,0.1]$. Recorrendo a (6.1.9) para calcular $u(0.05)$ temos que (com h=0.05):

$$\begin{aligned} F_1 &= f(0,1) = 0 \\ F_2 &= f\,(0{+}h, 1{+}h{\cdot}F_1) = h{\cdot}(1{+}h{\cdot}0) = h \,, \end{aligned}$$

de onde se obtém que

$$u(0.05) \simeq u(0) + \frac{h}{2}(F_1{+}F_2) = 1 + \frac{h}{2}(0{+}h) = 1 + \frac{h^2}{2} = 1.00125 \,.$$

Voltemos agora a integrar novamente para obter $u(0.1)$ a partir de $u(0.05)$. Temos que h=0.1−0.05=0.05, com

$$\begin{aligned} F_1 &= f\left(0.05, 1{+}\frac{h^2}{2}\right) = 0.05\left(1+\frac{h^2}{2}\right) \\ F_2 &= f\left(0.05{+}h, 1{+}\frac{h^2}{2}{+}h{\cdot}F_1\right) = (0.05{+}h)\left[1{+}\frac{h^2}{2}{+}0.05h\left(1+\frac{h^2}{2}\right)\right] \\ &= (0.05{+}h)(1{+}0.05{\cdot}h)\left(1{+}\frac{h^2}{2}\right) \,, \end{aligned}$$

de onde se obtém que

$$\begin{aligned} u(0.1) &\simeq u(0.05) + \frac{h}{2}(F_1{+}F_2) \\ &= \left(1+\frac{h^2}{2}\right) + \frac{h}{2}\left[0.05\left(1+\frac{h^2}{2}\right) + (0.05{+}h)(1{+}0.05h)\left(1+\frac{h^2}{2}\right)\right] \\ &= \left(1+\frac{h^2}{2}\right)\left[1 + 0.05\frac{h}{2} + (0.05{+}h)(1{+}0.05h)\frac{h}{2}\right] = 1.0050109 \,. \end{aligned}$$

O valor exacto é $u(0.1)$=e$^{0.005}$=1.0050125.

6.1.4 Duas equações diferenciais de primeira ordem

Em geral podemos ter várias equações diferenciais que envolvem diferentes funções a determinar, conhecendo-se apenas os seus valores num ponto. Neste caso podemos mais uma vez recorrer ao método de Runge-Kutta adaptando-o ao facto de que temos agora várias equações diferenciais.

Vejamos a título de exmplo o caso de termos duas equações diferenciais para duas funções $u_1(t)$ e $u_2(t)$;

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= f_1(t, u_1(t), u_2(t)) \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= f_2(t, u_1(t), u_2(t)) \,, \end{aligned} \tag{6.1.17}$$

que pretendemos resolver, calculando $u_1(t)$ e $u_2(t)$ a partir das seguintes condições iniciais para t=t_0,

$$u_1(t_0) = u_{10} \qquad \text{e} \qquad u_2(t_0) = u_{20} \,. \tag{6.1.18}$$

Aplicando o método de Runge-Kutta de segunda ordem (ver Secção 6.1.2) para resolver este sistema de equações diferenciais temos, para h=$t-t_0$, que

$$\begin{cases} F_{11} = f_1(t_0, u_{10}, u_{20}) \\ F_{12} = f_2(t_0, u_{10}, u_{20}) \end{cases} \tag{6.1.19}$$

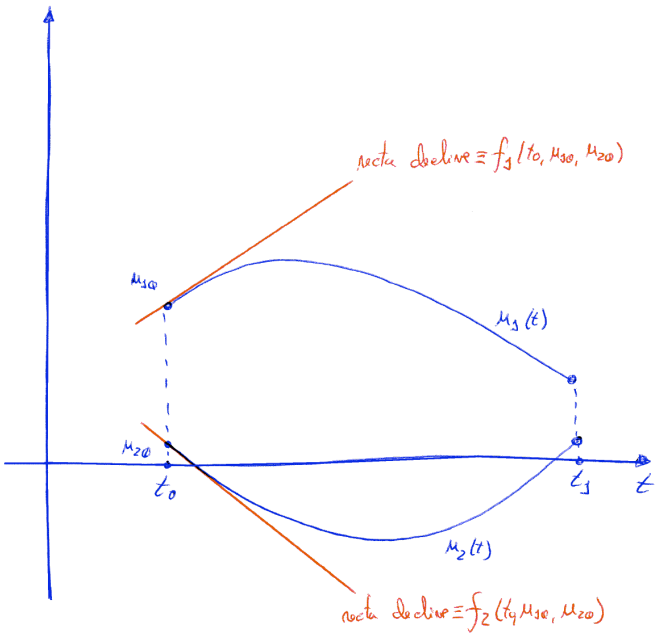


Figura 6.4: Representação gráfica de como usamos duas equações diferenciais do tipo (6.1.17) para estimar o valor das funções em t a partir do valor destas em t_0 e das suas derivadas tal como dadas por $f_1(t,u_1,u_2)$ e $f_2(t,u_1,u_2)$.

$$\begin{cases} F_{21} = f_1\left(t_0{+}h, u_{10}{+}hF_{11}, u_{20}{+}hF_{12}\right) \\ F_{22} = f_2\left(t_0{+}h, u_{10}{+}hF_{11}, u_{20}{+}hF_{12}\right) \end{cases}$$

de onde se obtém que;

$$\begin{cases} u_1(t_0{+}h) = u_{10} + \dfrac{h}{2}(F_{11}{+}F_{21}) \\ u_2(t_0{+}h) = u_{20} + \dfrac{h}{2}(F_{12}{+}F_{22}) \end{cases} \tag{6.1.20}$$

Temos assim uma estimativa do valor de ambas as funções no ponto t.

Exemplo 6.1.4: Consideremos o seguinte sistema de equações diferenciais de primeira ordem,

$$\begin{aligned} \frac{\mathrm{d}u_1}{\mathrm{d}t} &= u_2 \equiv f_1(t,u_1,u_2) \\ \frac{\mathrm{d}u_2}{\mathrm{d}t} &= -\frac{\alpha}{u_1^2} \equiv f_2(t,u_1,u_2) \end{aligned}$$

em que α=0.1. As duas condições fronteira de que dispomos são: $u_1(0)$=1 e $u_2(0)$=0. Usando o método de Runge-Kutta de segunda ordem (dado em 6.1.19 e 6.1.20), determinemos o valor das duas funções para t=0.1 (sendo portanto h=0.1);

$$\begin{cases} F_{11} = f_1(0,1,0) = 0 \\ F_{12} = f_2(0,1,0) = -0.1 \end{cases}$$

$$\begin{cases} F_{21} = f_1\left(0{+}h, 1{+}hF_{11}, 0{+}hF_{12}\right) = -0.1h \\ F_{22} = f_2\left(0{+}h, 1{+}hF_{11}, 0{+}hF_{12}\right) = -0.1 \end{cases}$$

Temos então que

$$\begin{cases} u_1(0.1) = 1 + \dfrac{0.1}{2}(0{-}0.1h) = 0.995 \\ u_2(0.1) = 0 + \dfrac{0.1}{2}(-0.1{-}0.1) = -0.01 \end{cases}$$

De igual modo, se usarmos desta vez o método de Runge-Kutta de quarta ordem (ver Secção 6.1.3), com $h{=}t{-}t_0$, temos que

$$\begin{cases} F_{11} = f_1(t_0, u_{10}, u_{20}) \\ \\ F_{12} = f_2(t_0, u_{10}, u_{20}) \end{cases}$$

$$\begin{cases} F_{21} = f_1\left(t_0{+}\dfrac{h}{2}, u_{10}{+}\dfrac{h}{2}F_{11}, u_{20}{+}\dfrac{h}{2}F_{12}\right) \\ F_{22} = f_2\left(t_0{+}\dfrac{h}{2}, u_{10}{+}\dfrac{h}{2}F_{11}, u_{20}{+}\dfrac{h}{2}F_{12}\right) \end{cases} \tag{6.1.21}$$

$$\begin{cases} F_{31} = f_1\left(t_0{+}\dfrac{h}{2}, u_{10}{+}\dfrac{h}{2}F_{21}, u_{20}{+}\dfrac{h}{2}F_{22}\right) \\ F_{32} = f_2\left(t_0{+}\dfrac{h}{2}, u_{10}{+}\dfrac{h}{2}F_{21}, u_{20}{+}\dfrac{h}{2}F_{22}\right) \end{cases}$$

$$\begin{cases} F_{41} = f_1(t_0{+}h, u_{10}{+}hF_{31}, u_{20}{+}hF_{32}) \\ \\ F_{42} = f_2(t_0{+}h, u_{10}{+}hF_{31}, u_{20}{+}hF_{32}) \end{cases}$$

de onde se obtém que

$$\begin{cases} u_1(t_0{+}h) = u_{10} + \dfrac{h}{6}(F_{11}{+}2F_{21}{+}2F_{31}{+}F_{41}) \\ u_2(t_0{+}h) = u_{20} + \dfrac{h}{6}(F_{12}{+}2F_{22}{+}2F_{32}{+}F_{42}) \end{cases} \tag{6.1.22}$$

Mais uma vez temos as expressões que permitem estimar o valor das duas funções em t.

6.1.5 Uma equação diferencial de segunda ordem

No caso de um equação diferencial de segunda ordem, podemos usar o método descrIto acima, transformando a equação num sistema de duas equações de primeira ordem. Isto é, consideremos a equação diferencial (2^a ordem);

$$\frac{\mathrm{d}^2 v}{\mathrm{d}t^2} = g\left(t, v, \frac{\mathrm{d}v}{\mathrm{d}t}\right) , \tag{6.1.23}$$

com as seguintes condições iniciais,

$$v(t_0) = v_0 \qquad \text{e} \qquad \frac{\mathrm{d}v}{\mathrm{d}t}(t_0) = v_0' \, . \tag{6.1.24}$$

Definindo as seguintes funções,

$$u_1(t) \equiv v(t) \qquad \text{e} \qquad u_2(t) \equiv \frac{\mathrm{d}v}{\mathrm{d}t} \, , \tag{6.1.25}$$

temos que a Eq. (6.1.23) pode ser escrita como,

$$\frac{\mathrm{d}u_2}{\mathrm{d}t} = g(t, u_1, u_2) \, . \tag{6.1.26}$$

Temos assim um sistema de duas equações diferenciais para u_1 e u_2, dado por,

$$\begin{cases} \dfrac{\mathrm{d}u_1}{\mathrm{d}t} = u_2(t) \\ \dfrac{\mathrm{d}u_2}{\mathrm{d}t} = g(t, u_1, u_2) \, , \end{cases} \tag{6.1.27}$$

cujas condições iniciais, para $t{=}t_0$, são $u_1(t_0){=}v_0$ e $u_2(t_0){=}v_0'$. Este sistema é equivalente ao definido em (6.1.17), pelo que podemos usar novamente o método de Runge-Kutta descrito na Secção anterior para calcular a função $v(t)$.

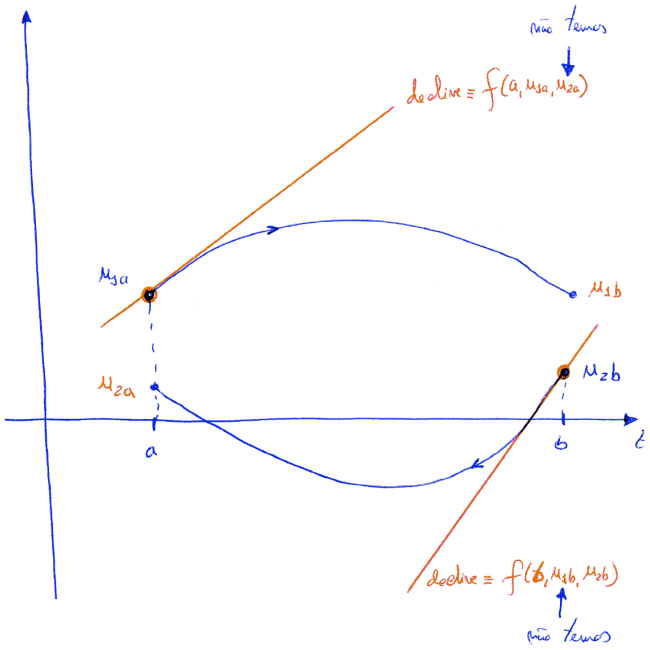


Figura 6.5: Representação gráfica de como dispondo de condições fronteira não nos é possível usar directamente um método do tipo Runge-Kutta para calcular as soluções a partir das equações diferenciais.

Exemplo 6.1.5: Vejamos o caso da trajectória de um corpo no campo gravitacional gerado por um corpo central de massa M. Se $r(t)$ é a distância a que o corpo se encontra no instante t do objecto central, então a aceleração a que está sujeito é dada por:

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = -\frac{GM}{r^2} \, .$$

Sabendo a posição (r_0) e a velocidade (v_0) que o coropo tem no instante $t{=}t_0$, podemos usar esta equação diferencial para calcular a trajectória deste usando o método de Runge-Kutta. Para tal basta-nos definir a função velocidade, como sendo

$$v(t) \equiv \frac{\mathrm{d}r}{\mathrm{d}t} \, .$$

Pelo que passamos a ter um sistema de duas equações diferenciais de primeira ordem:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = v(t) \qquad \text{e} \qquad \frac{\mathrm{d}v}{\mathrm{d}t} = -\frac{GM}{r^2(t)} \, ,$$

com condições iniciais: $r(t_0){=}r_0$ e $v(t_0){=}v_0$.

6.2 Problemas com condições fronteira

Neste caso não dispomos de todas as condições iniciais que nos permitem definir a solução num ponto t_0. Temos sim condições que têm de ser impostas em pontos distintos. Isso significa que não podemos de uma forma directa usar um método que parta de um ponto e use as expressões das derivadas para chegar a outro. É necessário encontrar um método que nos permita usar simultaneamente todas as condições impostas (condições fronteira) mesmo sendo estas válidas em pontos diferentes.

Vejamos o caso de termos duas equações;

$$\begin{cases} \dfrac{\mathrm{d}u_1}{\mathrm{d}t} = f_1(t,u_1(t),u_2(t)) \\ \dfrac{\mathrm{d}u_2}{\mathrm{d}t} = f_2(t,u_1(t),u_2(t)) \,, \end{cases} \tag{6.2.1}$$

cuja solução é especificada pelas seguintes condições;

$$u_1(a) = u_{1a} \qquad \text{e} \qquad u_2(b) = u_{2b} \,, \tag{6.2.2}$$

onde $a \neq b$. Neste caso temos o problema de que as condições que especificam a solução são dadas em pontos diferentes. Pelo que não temos toda a informação necessária num ponto - a partir do qual podemos estimar o valor das funções noutro ponto, sendo agora necessário utilizar simultaneamente as duas condições fronteira nesse cálculo.

Exemplo 6.2.1: Consideremos a função $f(x)$, tal que

$$f''(x) = \frac{f'(x)}{1+x} + f(x) \,,$$

e para a qual sabemos que $f'(0)=1$ e $f(1)=0$. Usemos a regra de Runge-Kutta de segunda ordem para estimar o valor de $f(0)$. Primeiro convertemos a equação de segunda ordem num sistema de equações de primeira ordem, definindo $v(x) \equiv f'(x)$. Neste caso as equações são:

$$\frac{\mathrm{d}v}{\mathrm{d}x} = \frac{v(x)}{1+x} + f(x) \qquad \text{e} \qquad \frac{\mathrm{d}f}{\mathrm{d}x} = v(x) \,,$$

cujas condições fronteira são $f(1)=0$ e $v(0)=1$. Usando o método de Runge-Kutta de segunda ordem temos que ($h=1$):

$$\begin{cases} F_{11} = \dfrac{v(0)}{1+0} + f(0) = 1 + f(0) \\ F_{12} = v(0) = 1 \end{cases}$$

$$\begin{cases} F_{21} = \dfrac{v(0)+hF_{11}}{1+0+h} + f(0)+hF_{12} = 2+2f(0) \\ F_{22} = v(0)+hF_{11} = 1+1 = 2 \end{cases}$$

de onde obtemos que

$$v(1) \simeq v(0) + \frac{1}{2}\Big[1+f(0)+2+2f(0)\Big] = 1 + \frac{3}{2}\Big[1+f(0)\Big] \qquad \text{e} \qquad f(1) = 0 \simeq f(0) + \frac{1}{2}(1+2) \,.$$

Da segunda equação temos que $f(0) \simeq -3/2$, como queriamos determinar.

6.2.1 Método "shooting"

Um exemplo simples de como podemos encontrar a solução é o método "shooting", que nos permite recorrer aos métodos anteriores, através do uso de técnicas que permitem calcular zeros de funções de forma a encontrar a solução procurada.

Para aplicarmos este método começamos por introduzir dois parâmetros

$$u_1(b) = C_b \qquad \text{e} \qquad u_2(a) = C_a \,, \tag{6.2.3}$$

que correspondem aos valores desconhecidos das duas funções $u_1(t)$ e $u_2(t)$ que não são fornecidos pelas condições iniciais.

Caso disponhamos de valores para C_a e C_b é possível usar um dos métodos de Runge-Kutta para calcular a soluções, quer a partir de $t=a$ quer a partir de $t=b$. Pois em ambos os casos dispomos agora dos valores das funções nesses pontos.

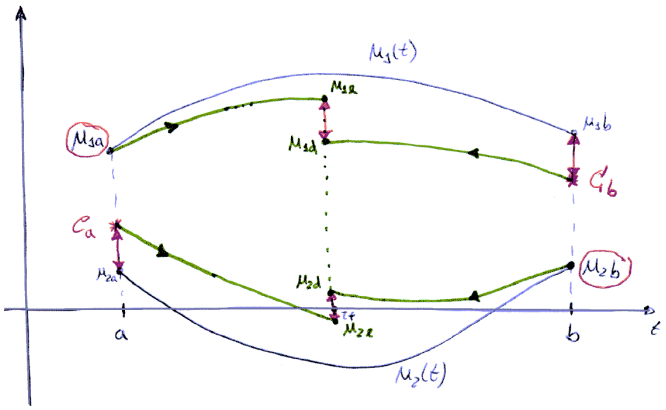


Figura 6.6: Representação gráfica de como dispondo de condições fronteira em $\{a,b\}$ e estimativas C_a e C_b então já podemos usar o método de Runge-Kutta para calcular a solução, partindo de a e/ou de b.

Seja $t_f \in [a,b]$, e integremos as equações (recorrendo a C_a) de a até t_f, onde obtemos

$$u_{1e} = u_1(t_f) \qquad \text{e} \qquad u_{2e} = u_2(t_f)\,. \tag{6.2.4}$$

Façamos o mesmo, agora a partir de b (recorrendo a C_b) até t_f, onde obtemos desta vez os valores

$$u_{1d} = u_1(t_f) \qquad \text{e} \qquad u_{2d} = u_2(t_f)\,. \tag{6.2.5}$$

Se os valores C_a e C_b correspondem à solução procurada, então teremos necessáriamente que $u_{1e}{=}u_{1d}$ e $u_{2e}{=}u_{2d}$. No entanto como estes valores são desconhecidos à partida, apenas podemos ter "palpites" do seu valor, pelo que os valores à esquerda não serão iguais aos valores à direita (ver Fig. 6.6). Isto é, o vector

$$\vec{\triangle} \equiv \begin{bmatrix} u_{1d}-u_{1e} \\ u_{2d}-u_{2e} \end{bmatrix}\,, \tag{6.2.6}$$

é não nulo. Tal como foi calculado, temos que $\vec{\triangle}{=}\vec{\triangle}(C_a,C_b)$, pois foi à custa destes valores que calculamos $\vec{\triangle}$. Por construção, temos ainda que a solução do sistema de equações diferenciais corresponde a ter $\vec{\triangle}{=}\vec{0}$.

Temos então um problema simples, que nos permite determinar a solução. Basta para isso recorrer a um dos métodos para o cálculo de raizes de forma a determinar os valores de C_a e C_b que correspondem ao zero de $\vec{\triangle}$. Usemos por exemplo o método iterativo de Newton: seja $\vec{C}{\equiv}(C_a,C_b)$, então pela expressão de Newton temos que se $\vec{C}_0$ é uma estimativa do zero, uma melhor aproximação será dada por

$$\vec{\triangle}(\vec{C}{+}\delta\vec{C}) = \vec{\triangle}(\vec{C}) + \frac{\mathrm{d}\vec{\triangle}}{\mathrm{d}\vec{C}} \cdot \delta\vec{C}\,, \tag{6.2.7}$$

onde $\delta\vec{C}$ é a correcção dada pelo método de Newton ao valor da raiz, e

$$\frac{\mathrm{d}\vec{\triangle}}{\mathrm{d}\vec{C}} \equiv \begin{bmatrix} \dfrac{\partial \triangle_1}{\partial C_a} & \dfrac{\partial \triangle_1}{\partial C_b} \\ \dfrac{\partial \triangle_2}{\partial C_a} & \dfrac{\partial \triangle_2}{\partial C_b} \end{bmatrix}\,, \tag{6.2.8}$$

é a matriz que dá a derivada de $\vec{\triangle}$ em função dos parâmetros C_a e C_b.

Como queremos que $\vec{\triangle}(\vec{C}{+}\delta\vec{C}){=}\vec{0}$, então podemos usar o sistema de equações dado em (6.2.7) para calcular a correcção $\delta\vec{C}$, bastando para isso resolver

$$\frac{\mathrm{d}\vec{\triangle}}{\mathrm{d}\vec{C}} \cdot \delta\vec{C} = -\vec{\triangle}(\vec{C})\,. \tag{6.2.9}$$

De forma a termos a matriz (6.2.8) necessitamos de recorrer a um dos métodos, discutidos na Secção 5, sobre o cálculo numérico de derivadas.

Um processo destes para a resolução de um sistema de equações diferenciais com condições fronteira pressupõe que os "palpites" iniciais de C_a e C_b estão suficientemente próximos da solução de forma a garantir a convergência do método iterativo usado para calcular o zero de $\vec{\triangle}$.

6.3 Exercícios

E6.1) Pretende-se calcular o valor de $\int_0^1 f(x)\,\mathrm{d}x$ sabendo apenas que

$$f'(x) = (1-4x)\,f(x) + (1+x)e^{2x(1-x)}$$

e $f(x{=}0){=}1$. Obtenha o valor aproximado do integral usando a fórmula de Simpson e a regra de Runge-Kutta de segunda ordem.

E6.2) Sabendo que uma função $f(x)$ é tal que

$$f'(x) = \cos\left[\pi f(x)\right] + e^{2x\,f(x)}$$

com $f(x{=}0){=}0$, recorra á regra de Runge-Kutta de segunda ordem para a integração de equações diferenciais e á regra do trapézio para o cálculo aproximado de integrais, de forma a estimar o valor de

$$\int_0^1 x\,f'(x)\,\mathrm{d}x.$$

E6.3) Seja $f(x)$ uma função real definida para $x \in \mathbb{R}$.

a) Sabendo que $f(x)$ é tal que

$$f'(x) = \cos\left[\pi f(x)\right] + e^{2x\,f(x)}$$

com $f(x{=}0){=}0$, recorra á regra de Runge-Kutta de segunda ordem para a integração de equações diferenciais e á regra da diferença central de segunda ordem para o cálculo aproximado de derivadas, de forma a estimar o valor de $f''(0.2)$.

b) Mostre que a regra de Runge-Kutta de segunda ordem quando usada para calcular o $\int_a^b f(x)\,\mathrm{d}x$, para qualquer função $f(x)$, fornece o mesmo resultado que a regra do trapézio.

E6.4) Seja $f(x)$ uma função real definida para $x \in \mathbb{R}$. Sabendo que $f(x)$ é tal que

$$f''(x) = xf'(x) + f(x)$$

com $f(x{=}0){=}0$ e $f'(x{=}0){=}1$, recorra á regra de Runge-Kutta de segunda ordem para a integração de equações diferenciais para estimar o valor de $f(1)$.

E6.5) Seja $f(x)$ uma função real da qual se conhecem os pontos $(x_i, f_i) \in \{(0.0, 1.0); (0.1, 0.5)\}$. Sabendo que a função obedece à seguinte equação diferencial;

$$f'(x) = \frac{\alpha}{x+1} - f(x)\,e^x\,,$$

recorra ao método de Runge-Kutta de $2^{\underline{a}}$ ordem para estimar o valor de α que reproduz os pontos conhecidos da função.

E6.6) Seja

$$f'(x) = \frac{f(x)}{x+1} + e^x$$

uma função real definida para $x \in [0, +\infty]$. Sabendo que $f(1){=}2$, estime o valor de $f(0)$ recorrendo à regra de Runge-Kutta de segunda ordem.

Bibliografia

Conte S. D., de Boor C., 1980, *Elementary Numerical Analysis: an Algorithmic Approach*, McGraw-Hill

Mathews J. H., 1992, *Numerical Methods for Mathematics, Science and Engineering* (2^a edição), Prentice-Hall International Editions

Pina H., 1995, *Métodos Numéricos*, McGraw-Hill de Portugal

Valença M. R., 1988, *Métodos Numéricos*, INIC

Apêndice A

Trabalhos práticos

Neste anexo apresentam-se diversos trabalhos práticos que requerem a utilização do computador para a resolução de problemas específicos de ciências e/ou engenharia.

A.1 Análise de um modelo da estrutura interna do Sol

Pretende-se estudar a estrutura interna de uma estrela recorrendo a uma modelo estelar tabelado. Este foi obtido a partir de um sistema de equações diferenciais não lineares que descrevem o conjunto de relações física que determinam o funcionamento de uma estrela, cujas características são:

R - raio da estrela (distância da superfície ao centro da estrela),
M - massa da estrela (massa contida na esfera de raio R),
L - luminosidade da estrela (energia total emitida por unidade de tempo na superfície).

Da tabela que descreve a estrutura (em ficheiro disponível no servidor das "Aulas na Web" da FCUP) constam os seguintes valores:

r/R	distância relativa ao centro da estrela,
$\log_{10} P$	logaritmo da pressão P (em unidades CGS),
m/M	massa relativa contida numa esfera de raio r,
L_r/L	energia relativa que por unidade de tempo atravessa a esfera de raio r,
$\nabla \equiv \frac{\mathrm{d}\log_{10} T}{\mathrm{d}\log_{10} P}$	gradiente logaritmico da temperatura T em ordem à pressão P,
$\log_{10} k$	logaritmo da opacidade k do gás (unidades CGS).

Utilize a tabela para obter os seguintes dados sobre esta estrela: o Sol!

a) Uma das quantidades fundamentais que descrevem a distribuição do gás no interior da estrela é a densidade ρ. Esta está relacionada com a massa "m" através da seguinte equação diferencial:

$$\frac{\mathrm{d}m}{\mathrm{d}r} = 4\pi r^2 \rho \,.$$

Obtenha o comportamento de $\rho \equiv \rho(r/R)$, e represente gráficamente $\log_{10}(\rho/\bar{\rho})$ como função de r/R, sendo $\bar{\rho} \equiv \frac{3M}{4\pi R^3}$.

b) Uma estrela apenas consegue estar em equilíbrio porque produz a energia que é continuamente perdida à superfície na forma de luminosidade. O processo de produção de energia recorre a reacções de fusão sendo

a emissividade ε (energia produzida por unidade de massa e por unidade de tempo) dada por:

$$\frac{\mathrm{d}L_r}{\mathrm{d}m} = \varepsilon \, .$$

Determine em que locais da estrela se produz energia, calculando a partir da tabela a função $\varepsilon \equiv \varepsilon(r/R)$, e representando-a graficamente.

c) Uma das quantidades que é fundamental para se perceber a forma como uma estrela produz e transporta energia no seu interior é a temperatura T. Recorra à tabela dada para determinar o comportamento da temperatura, representando graficamente $\log_{10} T$ em função de $\log_{10} P$. Para tal considere que à superfície (r=R) se tem que $T(R) \equiv T_{\mathrm{eff}}$=5777.538 K.

d) Em interiores estelares a interacção entre a radiação e o gás leva-nos a esperar que a opacidade se comporte de acordo com,

$$k \propto T^{-\beta} \, .$$

Encontre, por aproximação numérica "local", os valores de β que são representativos das diferentes regiões no interior da estrela.

e) Partindo da tabela original construa uma tabela (por interpolação) de $[\log_{10}\rho \ , \ \log_{10} k]$ com entradas espaçadas de 0.5, e que vão de $\log_{10}\rho = -10$ até $\log_{10}\rho = 1$.

f) Devido à dificuldade em modelar a forma de transporte de energia que a estrela usa perto da superfície, existe uma incerteza de $\pm 5\%$ na função ∇ para $1.0 \geq r/R \geq 0.95$. Estime a incerteza com que, nestas condições, podemos estimar o valor da temperatura no centro da estrela, $T_c \equiv T(r=0)$, recorrendo à tabela dada.

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

A.2 Monitorização das populações de coelhos e raposas

Pretende-se introduzir numa ilha onde existem coelhos uma população de raposas de forma a obter um equilíbrio nas populações de ambas as espécies. Pelas condições disponíveis os coelhos não tem limite de alimento (a ilha é húmida, logo rica em erva), no entanto as raposas dependem quase exclusivamente da população de coelhos para se alimentarem. Através de estudos inciais em populações controladas de coelhos e/ou raposas procurou-se medir os parâmetros que definem o comportamento destas duas espécies de forma a se poder determinar as melhores condições de equilíbrio na ilha, assegurando assim o sucesso da introdução das duas espécies.

a) Começou-se por estimar a capacidade dos coelhos se reproduzirem quando sujeitos à caça por um grupo de raposas. Para uma população inicial de coelhos $y_c(0)$=100, determinou-se qual seria a sua evolução na presença de duas populações distintas de raposas (y_r) por um período de doze meses:

Número de raposas y_r **:**	40	60
Número de coelhos y_c **para** t=12 **:**	1100	10
Incerteza :	95	2

Considere que a variação do número de coelhos é descrita por uma equação diferencial do tipo;

$$\frac{\mathrm{d}y_c}{\mathrm{d}t} = y_c \cdot (C_a - C_c \, y_r) \ ,$$

onde "C_a" descreve a velocidade de reprodução dos coelhos, e "C_c" o efeito que a caça das raposas tem na população de coelhos. Nesta descrição o valor de y_r é considerado constante no tempo.

Recorrendo aos dados registados determine os parâmetros (C_a, C_c) que descrevem a evolução do número de coelhos, estimando a incerteza dos valores obtidos.

b) Com objectivo de se estimar a forma como a variação da população de raposas depende da população de coelhos, fizeram-se vários testes em que partindo de um mesmo número de raposas e fixando o número de coelhos disponíveis por cada raposa, se determinou de que forma a população de raposas variava. Os dados obtidos foram os seguintes:

Nº de coelhos/raposa:	5	10	15	20	25	30	35	40
Variação das raposas (%):	−65	−22	−9	−1	5	11	15	21

Construa um código que permita estimar o valor de uma função por interpolação de uma tabela de $n+1$ valores, usando uma função da forma:

$$f_b(x) = \sum_{j=0}^{n} a_j \, \psi_j(x) \, .$$

Recorra a este código para estimar o número "C_d" de coelhos por raposa, que corresponde a ter um população constante de raposas. Para tal recorra a interpolação polinomial e ainda a interpolação por funções do tipo $\psi_k(x)=x^{-k}$, comparando os resultados obtidos.

c) De forma a se determinar a capacidade das raposas de aumentar a sua população fez-se um estudo do seu ciclo reprodutivo ao longo de 12 meses. Este estudo consistiu em, garantindo a alimentação suficiente, medir mensalmente a variação da população tal como dado pelo número de novos indivíduos adultos que surgem na população. Os valores obtidos foram;

Mês:	1	2	3	4	5	6	7	8	9	10	11	12
Taxa de variação (%):	2	1	1	1	5	10	15	20	20	20	10	5
Erro:	1	1	1	1	2	2	5	2	3	5	4	3

Construa um programa que a partir de uma tabela de $n+1$ pontos $\{x_i, f_i\}_{i=0}^{n}$ permita encontrar, usando o Método dos Mínimos Quadrados Ponderados, a função que melhor aproxima essa tabela;

$$f_c(x) = \sum_{j=0}^{k} a_j \, \psi_j(x) \, .$$

Recorrendo à tabela dada para a população de raposas, e considerando que a taxa de variação da população pode ser descrita por uma função do tipo;

$$R_r(t) = C_b \left[1 + \cos\left(\omega \, t - \phi\right)\right] \, ,$$

determine os melhores valores de "C_b" e "ϕ" que descrevem o comportamento registado. Para tal considere que a frequência ω corresponde ao período de um ano: $\omega = 2\pi/12$.

d) Para modelar a interacção de duas espécies podemos usar um modelo matemático simples que consiste no seguinte sistemas de equações diferenciais para as populações de coelhos (y_c) e raposas (y_r) em função do tempo (t é medido em meses):

$$\begin{aligned} \frac{\mathrm{d}y_c}{\mathrm{d}t} &= y_c \left(C_a - C_c \, y_r\right) \\ \frac{\mathrm{d}y_r}{\mathrm{d}t} &= y_r \left\{ -C_b \left[1 + \cos\left(2\pi \, \frac{t-9}{12} \right) \right] + C_c \left(y_c - C_d \, y_r \right) \right\} \, . \end{aligned}$$

Considere ainda que no caso em questão C_a=1.0 e C_b=10.0 (estes valores medem a dependência da variação da população no número de indivíduos dessa população), enquanto que C_c=0.02 (medindo o impacto que tem na populacao de cada espécie a interacção com a outra). Considere ainda que o parâmetro C_d=20 (indicando o número mínimo de coelhos que é necessário existirem por raposa para que estas não passem fome).

Para um grupo inicial de 20 raposas encontre o numero mínimo de coelhos que deve adicionar para garantir que ambas as populações não se extinguem (a extinção de uma espécie ocorre quando o número de indivíduos é menor que 2).

e) Obtenha os gráficos que mostram a evolução (por três anos) das duas populações para os seguintes valores iniciais das populações:

Número inicial de coelhos:	100	100	100	200	400	1000
Número inicial de raposas:	100	50	20	40	20	100

f) Sabendo que as populações de coelhos sofrem, frequentemente, surtos de uma doença que reduz em 80% a população existente, determine após que período de tempo t_d tal pode acontecer sem comprometer a sobrevivência das duas espécies - para tal considere o caso em que partimos de populações iniciais de 20 raposas e 300 coelhos. Obtenha os gráficos com a evolução das populações por dois anos, nos casos de t menor e t maior que esse valor limite.

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

A.3 Lançamento de projécteis

O projéctil disparado por um canhão descreve, quando sujeito á força de gravidade e ao atrito do ar, a seguinte trajectória $\vec{r}=(d,h)$ (onde d é a distância na horizontal e h a altitude), ao longo do tempo t:

$$\begin{aligned} h(t;v_0,\theta_0,C) &= C\,(v_h+gC)\left(1-\mathrm{e}^{-t/C}\right)-gCt \\ d(t;v_0,\theta_0,C) &= C\,v_d\left(1-\mathrm{e}^{-t/C}\right)\,. \end{aligned}$$

A velocidade inicial dada ao projéctil é $\vec{v}\equiv(v_d,v_h)=\left(v_0\cos\theta_0, v_0\sin\theta_0\right)$, onde θ_0 é o ângulo com a horizontal ($20^o\leq\theta_0\leq80^o$), sendo a aceleração da gravidade g=9.75$\times10^{-3}$ km s^{-2}. A "transparência" do ar é medida por C, e o canhão está localizado na origem $(0,0)$, disparando em t=0. O tipo de munições disponíveis são:

Tipo	v_0 (km/s)	Alcance máximo (km)
A	0.5	~ 5
B	1.0	~ 10
C	2.0	~ 21

Pretende-se construir um código que indique as condições que se devem usar para se poder atingir um ponto no terreno. Este código será desenvolvido para um canhão colocado num planalto que controla um vale onde o inimigo de move. A região em causa é descrita por uma malha (x,y) de pontos onde se conhece a altitude relativa ao ponto onde está instalado o canhão: $(0,0)$. Estes dados estão disponíveis por download.

a. De forma a determinar qual o atrito do ar na zona de disparo, efectuaram-se vários testes no planalto escolhendo a velocidade v_0 e o ângulo θ_0. Para diferentes tipos de munições mediu-se os seguintes valores da distância na horizontal d_a, a que o projéctil atingiu o solo.

v_0 (km/s)	θ_0 (^{o})	d_a (km)
0.5	40	4.86
0.5	50	4.14
0.5	60	3.24
0.5	70	2.22
0.5	80	1.13
1.0	70	4.49
1.0	80	2.28

Construa um programa que a partir dos valores de (v_0,θ_0) e a distância a que o projéctil atingiu o solo d_a, determine o tempo t_a que demorou a atingir o solo e o valor C que caracteriza a transparência do ar no local. Use os valores da tabela para estimar o valor médio de C para o local de disparo.

Note que no instante em que o projéctil atinge o solo no planalto se tem que $h(t_a)$=0 (com t_a>0), e ainda $d(t_a)$=d_a.

b. Construa um programa que a partir da malha que descreve o terreno determine por interpolação (a duas dimensões - em x e y) o valor da altitude $z\equiv z(x,y)$ no terreno para qualquer ponto (x,y) da região. Determine o perfil do terreno, representando-o num gráfico, quando nos deslocamos em linha recta entre o ponto em que se encontra o canhão e a posição x=13 km e y=9.0 km.

c. Construa um programa que determine, após escolha da velocidade do disparo (tipo de munição), qual o ângulo θ_0 de disparo que se deve usar para atingir um alvo cuja localização no terreno é: (x_a,y_a). Use esse código para determinar o valor de θ_0 para os seguintes alvos (especificando qual o tipo de munição que deve usar):

Alvo	x_a (km)	y_a (km)
1	1.03	3.17
2	3.04	5.21
3	5.52	0.51
4	6.11	9.52
5	10.11	2.78
6	17.03	2.04
7	12.15	8.09
8	19.10	8.99

Represente graficamente o trajecto do projéctil para quatro dos alvos indicados na tabela, mostrando simultaneamente o perfil do terreno ao longo desses trajectos.

Dada a localização do alvo (x_a,y_a), pode definir o valor de d_a=$\sqrt{x_a^2+y_a^2}$ e h_a=$z(x_a,y_a)$ (note que nas tabelas fornecidas z é dado em metros enquanto que x e y são dados em quilómetros).

Escolhido o valor de v_0 (tipo de munição), pode usar as equações do movimento para determinar qual o valor de θ_0 necessário para atingir esse ponto. Para tal deve usar o valor de C estimado na alinea *a)*.

d. **[Pergunta de valorização - facultativa!]** Face à incerteza em estimar o valor de C, e na precisão com que pode calcular a altitude do alvo, encontre uma forma de estimar o erro com que é possível apontar o canhão nestas condições.

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

A.4 Sismologia do Sol

No estudo da estrutura interna do Sol, recorre-se aquilo que são conhecidas como as oscilações solares. Para todos os efeitos são o equivalente a sismos na Terra, e tal como cá, também no Sol podem ser usados para inferir a estrutura nas camadas internas da nossa estrela. Esta informação vem sob a forma de frequências de oscilação (ω) cujo valor está associado a modos próprios caracterizados por dois (pressupondo simetria esférica) números quânticos - isto é, duas dimensões espaciais. Estes são o grau (ℓ) do modo e a sua ordem (n). Daí que cada valor de frequência seja da forma $\omega_\ell(n)$.

Uma das informações que se pode extrair diz respeito à base da zona exterior convectiva (zona de "ebulição" do gás). Em particular onde esta se localiza e suas características principais. Tal pode ser feito através do estudo do sinal periódico, originado por esta região do Sol, presente nas frequências de oscilação que é facilmente extraído através da seguinte função:

$$S_d(\omega_\ell) \equiv \frac{\mathrm{d}^2\omega_\ell}{\mathrm{d}n_\ell^2} \, .$$

Esta função pode ser calculada para todos os conjuntos de valores que têm o mesmo grau ℓ.

→ Recorra ao cálculo numérico para obter os seguintes resultados;

a. Cálcule os parâmetros M_i da spline cúbica natural que interpola a seguinte tabela de valores;

x:	0.0	1.5	2.0	3.5	4.0	5.0	6.1	7.0	8.2	9.0
y:	2.0	1.0	0.0	0.5	0.8	2.0	0.0	−1.0	0.0	0.0

De forma a poder utilizar o programa nas alineas seguintes garanta que este é suficientemente geral para poder ser utilizado para qualquer tabela de $n+1$ pontos.

b. Obtenha os valores de $S_d(\omega_\ell)$, recorrendo a interpolação por splines polinomiais cúbicas naturais, para cada subconjunto de valores de $\omega_\ell(n)$ (dados nas tabelas anexas) com o mesmo valor de ℓ. Represente graficamente todos os valores de $S_d(\omega_\ell)$ obtidos para as duas tabelas dadas (SM e OM).

c. Através da descrição teórica da origem deste sinal detectado na alinea **b)** é possível determinar que este tem uma forma do tipo

$$\begin{aligned} S_d(\omega_\ell) = \quad & a_0 + a_1 \frac{\omega_\ell}{3000} + a_2 \cos\left(\alpha_a \frac{\omega_\ell}{3000}\right) + a_3 \sin\left(\alpha_a \frac{\omega_\ell}{3000}\right) + \\ & + \frac{a_4}{\left(\frac{\omega_\ell}{2000}\right)^2} \cos\left(\alpha_b \frac{\omega_\ell}{2000}\right) + \frac{a_5}{\left(\frac{\omega_\ell}{2000}\right)^2} \sin\left(\alpha_b \frac{\omega_\ell}{2000}\right), \end{aligned}$$

onde $\alpha_a \simeq 85.2$ e $\alpha_b \simeq 18.35$. Recorrendo ao método dos Mínimos Quadrados obtenha a melhor função aproximadora, calculando os coeficientes a_j, nos dois casos em estudo; SM e OM. Represente graficamente os pontos usados e as funções aproximadoras obtidas para ambos os casos. Conclua se os dois modelos do Sol em estudo são iguais usando os valores obtidos para os parâmetrso a_j nos dois casos.

Todas as Tabelas estão disponíveis em formato electrónico.

Apresente um relatório descrevendo o método de resolução usado, e inclua também a listagem do(s) programa(s) construído(s) bem como os resultados obtidos (não esquecer os gráficos pedidos).

A.5 Planeamento de uma pista de ski

Na estação de ski de que se junta o plano de pistas pretende-se construir uma nova pista passando nos pontos $A-H$ assinalados.

Para projectar o trajecto desta pista, e devido ás restrições na curvatura que é necessário impor na sua construção, deve-se recorrer ao uso de *splines cúbicas* como função interpoladora.

a. Construa um programa que dada uma tabela de pontos $\{x_i, f_i\}_{i=0}^n$ calcule, num ponto x, o valor da *spline cúbica completa* que interpola os pontos da tabela.

b. As cotas dos pontos indicados no mapa são;

Ponto	d	h
A	0.0	0.0
B	2.9	1.1
C	7.0	1.4
D	10.4	2.0
E	13.8	2.4
F	15.0	3.8
G	16.3	5.6
H	18.0	6.2

Recorra ao uso de uma *spline cúbica* para determinar o trajecto que necessáriamente passa nos pontos indicados e exigindo que a pista seja horizontal tanto na zona de partida como de chegada. Obtenha as cotas do trajecto assim calculado para os seguintes valores de d: $\{1.0, 8.0, 14.0, 17.0\}$.

c. Esboce o gráfico da nova pista, indicando também os pontos calculados na alinea **b).**

Apresente um pequeno relatório descrevendo as propriedades da função interpoladora usada e método de cálculo desta. Inclua também a listagem do programa construído bem como os resultados obtidos/pedidos.

A.6 Trajectória de um cometa

De acordo com a primeira lei de Kepler um cometa deve ter uma órbita elíptica, parabólica ou hiperbólica (ignorando as forças de atracção gravitacional dos planetas).

A posição do cometa é dada, em coordenadas polares, pela equação

$$r = \beta - e\,[r\,\cos(\vartheta)]\,,$$

onde β é uma constante e e é a excentricidade da órbita, sendo $e<1$ para uma órbita elíptica, $e=1$ para uma parabólica e $e>1$ para uma hiperbólica.

Pretende-se estimar a posição de um cometa a partir de dados obtidos por observação.

a. Escreva um programa que calcule os coeficientes da combinação linear de um conjunto de funções dadas $\{\Phi_j(x)\}_{j=0}^{m}$ que melhor representa, no sentido dos mínimos quadrados, uma tabela de $N+1$ pontos $\{x_i, y_i, \Delta y_i\}_{i=0}^{N}$, onde Δy_i são os erros cometidos na medição dos valores y_i.

O programa deve calcular o valor dessa combinação num ponto x dado, o resíduo nos pontos tabelados e o erro médio quadrático.

b. Suponha que se fizeram as seguintes observações de um cometa desconhecido:

ϑ :	1.00	1.13	1.32	1.37	1.54	1.72	1.89	2.04	2.19	2.26
r :	1.01	1.08	1.20	1.25	1.42	1.65	1.95	2.30	2.75	3.00
Δr :	0.04	0.03	0.05	0.02	0.02	0.01	0.01	0.05	0.03	0.04

Usando o programa escrito determine o tipo de órbita do cometa e calcule aproximadamente a sua posição quando ϑ=4.6 (radianos). Esboce um gráfico com os pontos medidos, as barras de erro e a função aproximadora encontrada.

Apresente um pequeno relatório descrevendo o método de resolução usado. Inclua também a listagem do programa construído bem como os resultados obtidos.

A.7 Alinhamento de astros

A posição da projecção de corpos celestes no céu é definida por duas coordenadas angulares denominadas **declinação** (δ - medida em graus: 0^o-360^o) e **ascenção recta** (α - medida em horas: 0^h-24^h) que cobrem assim toda a esfera celeste. Definindo a origem, qualquer corpo celeste pode ser encontrado se forem fornecidas as duas coordenadas (α, δ).

Três astros estão alinhados (isto é, definem um circulo maior) quando as suas coordenadas (α_a, δ_a), (α_b, δ_b) e (α_c, δ_c) satisfazem a seguinte equação:

$$\tan(\delta_a)\,\sin(\alpha_b-\alpha_c)+\tan(\delta_b)\,\sin(\alpha_c-\alpha_a)+\tan(\delta_c)\,\sin(\alpha_a-\alpha_b)=0\,.$$

Consideremos então duas estrelas com as seguintes coordenadas:

$$\begin{aligned} \text{Castor } (\alpha \text{ Gem}): \quad & \alpha_a = 7^h\ 34^m\ 16.40^s \\ & \delta_a = 31^o\ 53'\ 51.2" \\ \text{Pollux } (\beta \text{ Gem}): \quad & \alpha_b = 7^h\ 45^m\ 00.10^s \\ & \delta_b = 28^o\ 02'\ 12.5" \ . \end{aligned}$$

Construa os seguintes programas;

a. Que cálcule o valor de $\sin x$ usando o facto de

$$\sin x \equiv \sum_{i=1}^{\infty} (-1)^{i-1}\ \frac{x^{2i-1}}{(2i-1)!}\ ,$$

com erro inferior a ε. Obtenha os valores nos casos de x=$\{\pi/4, \pi/2, \pi\}$ usando ε=10^{-6}.

b. Que determine a hora, com erro inferior a ε=0.05^s, em que Marte esteve em alinhamento com as estrelas dadas acima, no dia 1 de Outubro de 1994 ($t \in [1,2]$). Para tal considere que nesse dia o deslocamento de Marte na esfera celeste é descrito por:

$$\begin{aligned} \alpha_c(t) \quad &= (2^m\ 25.67^s)\ t + 7^h\ 58^m\ 23.32^s \\ \delta_c(t) \quad &= -(5'\ 50.7")\ t + 21^o\ 35'\ 28.9" \end{aligned}$$

onde t é o tempo em unidades de dias, contados a partir do início de Outubro.

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

A.8 Construção de um oleoduto

Pretende-se construir uma conduta de petróleo entre um planalto e a planicie (ver figura). Para tal é necessário definir os pontos do trajecto que esta deve seguir. Duas propostas foram feitas tendo em consideração a acessibilidade dos nodos da conduta e custos de construção.

Notando que a curvatura da conduta é determinante na capacidade de transporte de petróleo bem como na resistência desta, estabeleceu-se como critério de selecção entre as duas propostas, que fosse escolhida aquela que apresentar um menor máximo da curvatura em todo o trajecto.

$\rightarrow$ Construa os seguintes programas;

a. Que resolva um sistema tri-diagonal de N equações lineares. A resolução numérica do sistema não deve recorrer ao uso de matrizes $N \times N$, aproveitando por isso, o facto da matriz operadora ser tridiagonal. Utilize o programa no cálculo de $\vec{x}$, tal que

$$\begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 3 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 2 & 3 \end{bmatrix} \times \vec{x} = \begin{pmatrix} -1 \\ 1/2 \\ 1/2 \\ -1/2 \end{pmatrix} .$$

b. Que dada uma tabela de n pontos cálcule os parâmetros da spline cúbica natural que os interpola. Considere a seguinte tabela, e usando o programa desenvolvido cálcule os valores dos parâmetros M_i da spline interpoladora nos pontos da tabela;

x:	0	10	20	30	40	50	60	70	80	90	100
y:	2.0	1.0	0.0	1.5	1.8	2.0	1.0	−1.0	0.0	0.5	0.0

c. Que dado um trajecto determine o máximo da curvatura em todo o percurso, usando como modelo para a conduta uma spline cúbica. Use o programa para determinar qual o trajecto a escolher, entre os dois fornecidos, face ao critério definido.

Trajecto 1

x (km)	h (m)
0.0	0.0
5.0	5.0
26.0	25.0
35.5	80.0
44.0	105.0
50.0	109.0
60.0	100.0

Trajecto 2

x (km)	h (m)
0.0	0.0
9.0	8.0
20.0	30.0
30.0	35.0
36.0	80.0
41.0	116.0
53.0	106.0
60.0	100.0

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

A.9 Control da dosagem de um fármaco

O atracúrio é um fármaco que produz bloqueio neuromuscular, sendo correntemente utilizado em Anestesia para induzir um relaxamento muscular em pacientes durante intervenções cirúrgicas.

O modelo que descreve a relação (dinâmica) entre a quantidade administrada $u(t)$ (por via endovenosa na situação presente) e a concentração de efeito induzida $C_e(t)$, é referido por "Modelo Farmacocinético". Para o atracúrio, e nas dosagens habitualmente utilizadas na prática clínica, o modelo é razoavelmente bem descrito por um sistema linear de 3^o ordem, com uma resposta impulsional do tipo,

$$C_e(t) = a_1\, e^{-\lambda_1 \cdot t} + a_2\, e^{-\lambda_2 \cdot t} + a_3\, e^{-\lambda_3 \cdot t}\ ; \quad a, b > 0\ .$$

O efeito induzido pelo atracúrio $r(t)$ (o nível de bloqueio neuromuscular) está relacionado com a concentração de efeito. O modelo que descreve esta relação é designado por "Modelo Farmacodinâmico". Para o atracúrio, este modelo pode ser descrito pela relação não linear (equação de Hill):

$$r(t) = \frac{100}{1 + \left[\dfrac{C_e(t)}{C_{p50}}\right]^{\gamma}}\ ; \quad C_{p50}, \gamma > 0\ .$$

Nestas equações os parâmetros $(a_1, a_2, a_3, \lambda_1, \lambda_2, \lambda_3, C_{50}, \gamma)$ são (muito) diferentes de indivíduo para indivíduo, sendo ainda de admitir que possam variar de forma significativa ao longo de uma intervenção cirúrgica. Reparar que numa situação de ausência total de relaxamento, $C_e(t)$=0 e $r(t)$=100. Enquanto que numa situação de grande relaxamento, $C_e(t) >> 1$ e $r(t) \sim 0$. Na prática clínica é corrente induzir um relaxamento alto ($r(t) \sim 0$) no início de uma intervenção cirúrgica para uma intubação fácil. Este efeito é obtido através da administração rápida (por via endovenosa) do fármaco (dose típica para o atracúrio = 500 μ g/kg).

$\rightarrow$ Usando os resultados de um teste ao doente, fornecidos em anexo, proceda aos seguintes cálculos;

a. Construa um programa geral que dado $(n+1)$ pontos $\{x_i, f_i, e_i\}_{i=0}^{n}$, onde e_i são os erros na medida f_i, e $(m+1)$ funções $[\varphi_0, \varphi_1, ..., \varphi_m]$, encontre a melhor combinação linear destas no sentido dos minimos quadrados pesados que aproxima os pontos dados. Construa o programa de forma que os coeficientes sejam calculados para valores normalizados dos x_i, isto é, a partir destes que seja a tabela $\{x'_i, f_i, e_i\}_{i=0}^{n}$ que é usada no cálculo dos coeficientes, tal que $x'_i \in [-1, 1]$.

Aplique ao caso de: $\varphi_0(x)$=1, $\varphi_1(x)$=x e $\varphi_2(x)$=$(3x^2-1)/2$ e os pontos dados na tabela;

i	x_i	f_i	e_i
0	0.0	0.0	0.1
1	5.0	5.0	0.2
2	26.0	25.0	0.4
3	35.5	80.0	0.2
4	44.0	105.0	1.1
5	50.0	109.0	0.8
6	60.0	100.0	0.5

b. Obtenha no caso do doente cujos valores são fornecidos, a função $C_e(t)$, sabendo que

$$(\lambda_1, \lambda_2, \lambda_3, C_{p50}, \gamma) = (0.315, 0.035, 0.100, 0.652, 4.25) ,$$

no sentido da aproximação definida na alinea anterior.

Apresente um relatório descrevendo o método de resolução usado nas várias alineas. Inclua também a listagem dos programas construídos bem como os resultados obtidos.

• Tabela de valores disponível electrónicamente.

©2015 Universidade do Porto, Portugal