Apontamentos

de

Matemática Numérica II

Ano Lectivo de 2006/2007

Luís Nunes Vicente

Departamento de Matemática da F.C.T.U.C.

A disciplina de Matemática Numérica II faz parte do tronco comum do plano de estudos da Licenciatura em Matemática do Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Esta disciplina, do primeiro semestre do terceiro ano, é antecedida da de Matemática Numérica I. Fazem parte do programa de Matemática Numérica I os seguintes tópicos: álgebra linear numérica (normas vectoriais e matriciais, sistemas de equações lineares, problemas de mínimos quadrados lineares, decomposições em valores próprios e em valores singulares); interpolação polinomial; resolução de equações não lineares de uma variável.

Estes apontamentos foram organizados em formato aula-a-aula, tipo *lecture notes*. Cada aula está descrita de forma o mais auto-contida possível. Evitaram-se, ao máximo, as referências dentro de cada aula e entre aulas. No final de cada lição, colocam-se exercícios sobre a matéria dada, para resolução nas aulas ou em trabalho-para-casa.

Os vários tópicos do programa da disciplina foram organizados da seguinte forma:

- Aulas 1–5: métodos numéricos para sistemas de equações não lineares.
- Aulas 6–8: métodos numéricos para optimização sem restrições e problemas de mínimos quadrados não lineares.
- Aula 9: diferenciação numérica.
- Aulas 10–12,16: integração numérica.
- Aulas 13–15,17–18: aproximação de funções.
- Aulas 19–23: métodos numéricos para problemas de condições de fronteira.
- Aulas 23–27: métodos numéricos para problemas de valor inicial.

Os exemplos numéricos foram corridos em MATLAB® (`http://www.mathworks.com`). As correspondentes `m-files` estão disponíveis a partir da página da disciplina, no endereço `http://www.mat.uc.pt/~lnv/mn2/mn2_05_06.html`.

Procurar-se-á melhorar a versão actual destes apontamentos no decorrer dos próximos anos. Reconhece-se a necessidade de cobrir outros tópicos, nomeadamente outras transformadas (Laplace e Ôndulas) e suas aplicações e discretizações. Tentar-se-á, também, incluir mais exemplos e ilustrações ao longo do texto.

Coimbra, 9 de Setembro de 2005, LNV.
(Data da última revisão: 19/02/2007.)

Aula 1: Método de Newton para Sistemas de Equações Não Lineares

A simulação computacional de modelos em ciência e engenharia, depara-se, frequentemente, com a necessidade de resolver sistemas de equações não lineares. Na maioria dos casos, estes sistemas têm origem na discretização de equações diferencias não lineares. Há situações, porém, em que os sistemas de equações não lineares são resultado directo da formulação dos problemas em causa, como é o caso do exemplo apresentado no final desta aula.

Existem diversas técnicas de resolução numérica para resolver sistemas de equações não lineares, sendo que as mais úteis e conhecidas assentam, directa ou indirectamente, no método de Newton e, consequentemente, na resolução de sistemas de equações lineares.

Há uma estreita ligação entre a resolução de sistemas de equações não lineares e a resolução de problemas de optimização não linear sem restrições, razão pela qual abordaremos estes dois tópicos em conjunto.

Considere, então, o sistema de equações não lineares

$$F(x) \;=\; 0,$$

em que F é uma função vectorial dada por

$$\begin{aligned} & F : D \subset \mathbb{R}^n \to \mathbb{R}^n, \\ & x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \to \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = F(x). \end{aligned}$$

As funções f_i são reais:

$$\begin{aligned} & f_i : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}, \\ & x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \longrightarrow f_i(x) = f_i(x_1, \ldots, x_n), \qquad i = 1, \ldots, n. \end{aligned}$$

A título de ilustração considere-se o seguinte exemplo:

$$\begin{aligned} & F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \\ & F(x) \;=\; \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} \;=\; \begin{bmatrix} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{bmatrix}. \end{aligned}$$

Para esta função F, o sistema $F(x) = 0$ tem as raízes $[3 \;\; 0]^\top$ e $[0 \;\; 3]^\top$.

Antes de introduzirmos o método de Newton para a resolução do sistema $F(x) = 0$, vamos rever algumas propriedades das funções de várias variáveis.

Uma função $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ diz-se continuamente diferenciável em x se as derivadas parciais de f existirem e forem contínuas em x. A função f diz-se continuamente diferenciável no aberto $D \subset \mathbb{R}^n$ se o for para todos os pontos de D. O gradiente de f em x é dado por

$$\nabla f(x) \;=\; \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix} \;\in\; \mathbb{R}^n.$$

Se f for continuamente diferenciável no conjunto aberto D de $\mathbb{R}^n$ então, para x e $x+p$ em D, com $[x, x+p] \subset D$, tem-se, pelo Teorema do Valor Médio na sua versão integral, que

$$f(x+p) - f(x) \;=\; \int_0^1 \nabla f(x+tp)^\top p \; dt.$$

Ser-nos-á útil, para motivar o método de Newton, a notação

$$\int_x^{x+p} \nabla f(z)^\top dz \;\stackrel{\text{def}}{=}\; \int_0^1 \nabla f(x+tp)^\top p \; dt.$$

Se agora tomarmos as n funções reais de n variáveis reais que definem o nosso sistema de equações não lineares e assumirmos que cada uma delas é continuamente diferenciável no conjunto aberto D de $\mathbb{R}^n$ então, para x e $x+p$ em D, com $[x, x+p] \subset D$, tem-se que

$$\begin{aligned} f_1(x+p) - f_1(x) &= \textstyle\int_x^{x+p} \nabla f_1(z)^\top dz \\ &\vdots \\ f_n(x+p) - f_n(x) &= \textstyle\int_x^{x+p} \nabla f_n(z)^\top dz. \end{aligned}$$

Em notação vectorial estas n igualdades são representadas por

$$F(x+p) - F(x) \;=\; \int_x^{x+p} J(z) dz,$$

em que $F(x)$ é o vector de componentes $f_1(x), \ldots, f_n(x)$ e

$$J(x) \;=\; \begin{bmatrix} \nabla f_1(x)^\top \\ \vdots \\ \nabla f_n(x)^\top \end{bmatrix}$$

é a matriz Jacobiana de F em x. Por extenso, a matriz Jacobiana toma a forma

$$J(x) \;=\; \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \cdots & \frac{\partial f_n}{\partial x_n}(x) \end{bmatrix} \;\in\; \mathbb{R}^{n \times n}.$$

No exemplo anterior, temos

$$\nabla f_1(x) \;=\; \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla f_2(x) \;=\; \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \quad \text{e} \quad J(x) \;=\; \begin{bmatrix} 1 & 1 \\ 2x_1 & 2x_2 \end{bmatrix}.$$

O método de Newton para a resolução do sistema $F(x) = 0$ é iterativo. Seja $x_0 \in \mathbb{R}^n$ uma aproximação inicial. Sabe-se que

$$F(x_0 + p) - F(x_0) \;=\; \int_{x_0}^{x_0+p} J(z)dz.$$

Idealmente, gostaríamos que $x_0 + p$ fosse uma solução do sistema, ou seja que

$$F(x_0 + p) \;=\; 0.$$

Logo, procuramos $p \in \mathbb{R}^n$ tal que

$$\int_{x_0}^{x_0+p} J(z)dz \;=\; -F(x_0).$$

Este integral é não linear em p, o que torna a resolução deste sistema tão difícil como a resolução do original.

Linearizar é a palavra de ordem do método de Newton. Ao aproximarmos este integral por

$$\int_{x_0}^{x_0+p} J(z)dz \;\simeq\; J(x_0)p$$

obtemos uma função linear em p (definida por $J(x_0)p$).

O passo p é determinado, assim, à custa da resolução do sistema de equações lineares

$$J(x_0)p_0 \;=\; -F(x_0).$$

A nova aproximação x_1 é dada por $x_1 = x_0 + p_0$.

O método de Newton, quando bem definido, gera a sucessão de pontos $\{x_k\}$ descrita pelo seguinte algoritmo.

Método de Newton para Sistemas de Equações não Lineares
Escolher $x_0 \in \mathbb{R}^n$.
Para $k = 0, 1, 2, \ldots$

1. Resolver o sistema de equações lineares $J(x_k)p_k = -F(x_k)$.
2. Fazer $x_{k+1} = x_k + p_k$.

A fórmula recursiva do método de Newton

$$x_{k+1} \;=\; x_k - J(x_k)^{-1}F(x_k)$$

reduz-se, no caso $n = 1$, usando a notação $f(x) = F(x)$, a

$$x_{k+1} \;=\; x_k - [f'(x_k)]^{-1}[f(x_k)] \;=\; x_k - \frac{f(x_k)}{f'(x_k)},$$

que identificamos como sendo a fórmula do método de Newton para a resolução de uma equação não linear com uma incógnita.

Para o exemplo desta aula tome-se $x_0 = [\,1\;\;5\,]^\top$. A primeira iteração do método de Newton calcula o passo

$$J(x_0)p_0 = -F(x_0) \iff \begin{bmatrix} 1 & 1 \\ 2 & 10 \end{bmatrix} p_0 = -\begin{bmatrix} 3 \\ 17 \end{bmatrix} \iff p_0 = \begin{bmatrix} -13/8 \\ -11/8 \end{bmatrix}.$$

Logo,

$$x_1 \;=\; x_0 + p_0 \;=\; \begin{bmatrix} -0.625 \\ 3.625 \end{bmatrix}.$$

A segunda iteração consiste em resolver

$$J(x_1)p_1 = -F(x_1) \iff \begin{bmatrix} 1 & 1 \\ -5/4 & 29/4 \end{bmatrix} p_1 = -\begin{bmatrix} 0 \\ 145/32 \end{bmatrix} \iff p_1 = \begin{bmatrix} -13/8 \\ -11/8 \end{bmatrix},$$

para, depois, calcular

$$x_2 \;=\; x_1 + p_1 \;\simeq\; \begin{bmatrix} -0.092 \\ 3.092 \end{bmatrix}.$$

Exercícios

1. Considere a função vectorial $\mathbb{R}^2 \longrightarrow \mathbb{R}^2$ definida por:

$$F(x_1, x_2) \;=\; \begin{bmatrix} x_1^2 + x_2 + 1 \\ x_1 x_2 \end{bmatrix}.$$

   (a) Calcule a matriz Jacobiana de F e as raízes de $F(x_1, x_2) = 0$.

   (b) Efectue, se possível, uma iteração do método de Newton partindo dos pontos $x_0 = [\,0\;\;0\,]^\top$ e $x_0 = [\,1\;\;1\,]^\top$.

2. Estude a aplicação do método de Newton ao problema: $\arctan(x) = 0$.

3. Considere a função vectorial $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ definida por:

$$F(x_1, x_2, \ldots, x_n) \;=\; \begin{bmatrix} x_1 x_2 \\ x_2 x_3 \\ x_3 x_4 \\ \vdots \\ x_{n-1} x_n \\ x_n \end{bmatrix},$$

em que n é um número inteiro ímpar maior que ou igual a 5.

(a) Calcule a matriz Jacobiana de F.

(b) Efectue, se possível, uma iteração do método de Newton partindo do ponto $x_0 = [\,1 \;\; 1 \;\; \cdots \;\; 1\,]^\top \in \mathbb{R}^n$.

(c) Efectue, se possível, uma iteração do método de Newton partindo do ponto $x_0 = [\,-1 \;\; -1 \;\; \cdots \;\; -1\,]^\top \in \mathbb{R}^n$.

(d) O que é que pode concluir sobre o comportamento do método de Newton a partir destas duas últimas alíneas?

Nota: Se não conseguir resolver este exercício na forma em que ele está colocado, resolva-o para $n = 5$.

4. Prove que o método de Newton converge numa iteração quando aplicado à resolução de um sistema de equações lineares (possível determinado).

5. Um exemplo simples e interessante de um sistema de equações não lineares descreve a intersecção de órbitas de planetas. Considere o seguinte sistema de equações não lineares:

$$F(t_1, t_2) \stackrel{\text{def}}{=} \begin{bmatrix} x_2(t_2) - x_1(t_1) \\ y_2(t_2) - y_1(t_1) \end{bmatrix} = 0.$$

Quando a função $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ se anula num ponto $[\,t_1 \;\; t_2\,]^\top$, as órbitas de dois planetas (1 e 2) intersectam-se. As órbitas são elípticas, com um dos focos na origem (*e.g.*, o sol). Por exemplo, a órbita do planeta 1 é definida por

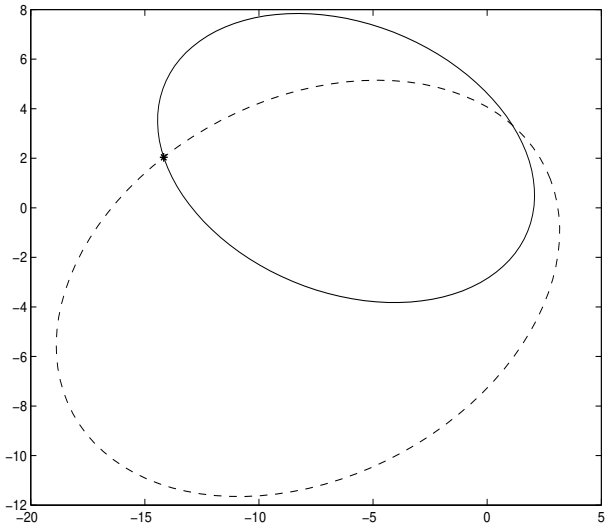
$$\begin{bmatrix} x_1(t_1) \\ y_1(t_1) \end{bmatrix} = \begin{bmatrix} \cos(\phi_1) & \text{sen}(\phi_1) \\ -\text{sen}(\phi_1) & \cos(\phi_1) \end{bmatrix} \begin{bmatrix} \frac{P_1 - A_1}{2} + \frac{P_1 + A_1}{2}\cos(t_1) \\ \sqrt{P_1 A_1}\,\text{sen}(t_1) \end{bmatrix},$$

em que ϕ_1 é o ângulo com que a órbita foi rodada e A_1 e P_1 são, respectivamente, a maior e a menor distância dos seus pontos ao sol. A órbita do planeta 2 é definida de modo análogo.

Corra, em MATLAB, as `m-files` disponíveis em

`ftp` : //`ftp.cs.cornell.edu/pub/cv`,

que aplicam o método de Newton para resolver, numericamente, uma instância deste sistema de equações não lineares. Mais informações sobre este problem são dadas no livro C. F. Van Loan, *An Introduction to Scientific Computing – A Matrix-Vector Approach Using* MATLAB, MATLAB Curriculum Series, Prentice Hall, Upper Saddle River, New Jersey, 1997.


8
6
4
2
0
−2
−4
−6
−8
−10
−12
−20
−15
−10
−5
0
5

Aula 2: Taxas de Convergência e Constantes de Lipschitz

O método de Newton apresenta uma taxa quadrática de convergência local. O erro absoluto entre as iteradas x_k e a solução x_* decresce quadraticamente, em norma, se x_0 estiver suficientemente próximo de x_*. A convergência diz-se local porque esta propriedade é garantida apenas para x_0 numa certa vizinhança de x_*.

Em MATLAB, corremos o método de Newton para o sistema de equações não lineares definido pela função vectorial

$$F(x) \;=\; \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ 5x_1^2 - x_2^2 - 2 \end{bmatrix},$$

com o ponto inicial $x_0 = [\,2 \;\; 2\,]^\top$. O método gera iteradas que se aproximam da solução $x_* = [\,\sqrt{2}/2 \;\; \sqrt{2}/2\,]^\top$. Reproduzimos, em baixo, o *output* do código.

```
>> Newton
------------------------------------------------------------------
                 Metodo de Newton para F(x) = 0
------------------------------------------------------------------
 Iteracao     |x(1)-x*(1)|          |x(2)-x*(2)|        || x-x* ||
     0     1.2928932188134525     1.2928932188134525    1.8284e+000
     1     0.4178932188134524     0.4178932188134524    5.9099e-001
     2     0.0776154410356746     0.0776154410356747    1.0976e-001
     3     0.0038384007210237     0.0038384007210238    5.4283e-003
     4     0.0000103617834891     0.0000103617834892    1.4654e-005
     5     0.0000000000759185     0.0000000000759186    1.0737e-010
     6     0.0000000000000000     0.0000000000000001    1.1102e-016
------------------------------------------------------------------
>>
```

O decréscimo quadrático do erro absoluto $\|x_k - x_*\|$ é evidente de iteração para iteração. Este erro comporta-se, a partir da segunda iteração, como

$$10^{-1} \quad 10^{-2} \quad 10^{-4} \quad 10^{-8} \quad 10^{-16}.$$

Seja $\{x_k\}$ uma sucessão de vectores de $\mathbb{R}^n$ a convergir para x_*. A convergência é quadrática, ou apresenta uma taxa quadrática, se existir uma constante positiva $M > 0$ tal que, para todo o k,

$$\|x_{k+1} - x_*\| \;\leq\; M\|x_k - x_*\|^2.$$

Os métodos de quasi-Newton apresentam uma taxa de convergência superlinear:

$$\lim_{k\to+\infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \;=\; 0.$$

Outros métodos convergem apenas linearmente. A sucessão $\{x_k\}$ converge para x_* linearmente, ou com taxa linear, se existir uma constante positiva $r \in (0,1)$ tal que, para todo o k,

$$\|x_{k+1} - x_*\| \ \leq \ r\|x_k - x_*\|.$$

A convergência quadrática implica a superlinear e esta, por sua vez, implica a linear. Estas taxas de convergência são designadas, por vários autores, usando a letra q (q-quadrática, q-superlinear e q-linear). Existe um outro tipo de taxas, as designadas pela letra r (ver exercício).

O estudo da convergência local do método de Newton requer um pouco mais do que a continuidade das funções f_i, $i = 1, \ldots, n$, que definem as componentes de F. É preciso que todas as suas derivadas parciais (coleccionadas na matriz Jacobiana J) sejam *contínuas à Lipschitz*. A matriz Jacobiana J diz-se contínua à Lipschitz em $D \subset \mathbb{R}^n$, com constante $\gamma > 0$, se

$$\|J(x) - J(y)\| \ \leq \ \gamma\|x - y\| \quad \text{para todos os } x \text{ e } y \text{ em } D.$$

A norma do lado esquerdo desta desigualdade é uma norma matricial, enquanto a norma do lado direito é uma norma vectorial. Vai ser necessário, assim, trabalhar com normas matriciais induzidas por normas vectoriais, como as normas ℓ_∞, ℓ_1 ou ℓ_2 (Euclideana).

Tentemos, primeiro, uma majoração do erro absoluto $\|x_1 - x_*\|$ em função do erro absoluto $\|x_0 - x_*\|$, assumindo, informalmente, que $J(x_0)$ é invertível. A função F é nula na solução x_*. A partir da forma do método de Newton, escrevemos

$$\begin{aligned} x_1 - x_* &= x_0 - x_* - J(x_0)^{-1}F(x_0) \\ &= J(x_0)^{-1}J(x_0)(x_0 - x_*) - J(x_0)^{-1}\left(F(x_0) - F(x_*)\right) \\ &= J(x_0)^{-1}\left[F(x_*) - F(x_0) - J(x_0)(x_* - x_0)\right]. \end{aligned}$$

Logo,

$$\|x_1 - x_*\| \ \leq \ \|J(x_0)^{-1}\| \ \|F(x_*) - [F(x_0) + J(x_0)(x_* - x_0)]\|.$$

A expressão $F(x_0) + J(x_0)(x_* - x_0)$ é o termo linear de uma expansão de Taylor de ordem um centrada em x_0. Assim sendo, é expectável que a diferença entre as normas de $F(x_*)$ e de $F(x_0) + J(x_0)(x_* - x_0)$ varie quadraticamente com $\|x_* - x_0\|$. Desta forma, obter-se-ia

$$\|x_1 - x_*\| \ \leq \ \|J(x_0)^{-1}\| \times \textit{constante} \times \|x_0 - x_*\|^2.$$

A proposição seguinte descreve o resultado que se utilizará para obter esta variação quadrática.

Proposição 1 *Seja $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ uma função vectorial definida num conjunto aberto D. Suponha-se que F é continuamente diferenciável em D e que a sua matriz Jacobiana J é contínua à Lipschitz em D com constante γ.*

Então, quaisquer que sejam x e $x + p$ em D, com $[x, x + p] \subset D$, é verdadeira a desigualdade

$$\|F(x + p) - F(x) - J(x)p\| \ \leq \ \frac{\gamma}{2}\|p\|^2.$$

Demonstração. O teorema do valor médio (em versão integral aplicada a todas as componentes de F) permite-nos escrever

$$F(x+p) - F(x) \;=\; \int_0^1 J(x+tp)p \; dt.$$

Por outro lado,

$$J(x)p \;=\; \int_0^1 J(x)p \; dt.$$

Assim, $F(x+p) - F(x) - J(x)p = \int_0^1 [J(x+tp) - J(x)]p \; dt$. Logo,

$$\|F(x+p) - F(x) - J(x)p\| \;\leq\; \int_0^1 \|J(x+tp) - J(x)\| \, \|p\| \; dt \;\leq\; \int_0^1 \gamma \|tp\| \, \|p\| \; dt.$$

A observação $\int_0^1 |t| \; dt = 1/2$ conclui a demonstração. ■

Aplicando esta proposição com $x = x_0$ e $p = x_* - x_0$, constatamos que a *constante* que procuramos é dada por $\gamma/2$.

Exercícios

1. Seja $\{x_k\}$ uma sucessão em $\mathbb{R}^n$ a convergir para x_*. A convergência diz-se r-quadrática se existir uma sucessão real $\{\alpha_k\}$, a convergir q-quadraticamente para zero, tal que, para todo o k,

   $$\|x_k - x_*\| \;\leq\; \alpha_k.$$

   Prove que se uma sucessão $\{y_k\}$ de $\mathbb{R}^n$ convergir q-quadraticamente para y_* então as sucessões $\{(y_k)_i\}$ convergem r-quadraticamente para $(y_*)_i$, para todas as componentes $i \in \{1, \ldots, n\}$.

   Nota: Este exercício aplica-se, também, aos casos linear e superlinear.

2. Mostre que a sucessão $\{1/k\}$ não converge para 0 q-linearmente.

3. Mostre que a sucessão $\{1 + 10^{-k}\}$ converge para 1 q-linearmente.

4. Prove que a sucessão $\{1 + (0.5)^{2^k}\}$ apresenta uma taxa de convergência q-quadrática para 1.

5. Estude a taxa de convergência da sucessão 3^{-k^2}.

6. Seja $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ uma função vectorial dada por

   $$F(x) \;=\; \begin{bmatrix} f_1(x_1, x_2, x_3) \\ f_2(x_1, x_2, x_3) \\ f_3(x_1, x_2, x_3) \end{bmatrix} \;=\; \begin{bmatrix} x_1 \\ x_2^2 + x_2 \\ e^{x_3} - 1 \end{bmatrix}.$$

(a) Indique valores para as constantes de Lipschitz das funções 1, $x+1$ e e^x em $[-a,a]$, $a>0$.

(b) Indique um valor para a constante de Lipschitz de $J(x)$ no conjunto $[-a,a]^3$:

$$\{x \in \mathbb{R}^3 : |x_i| \leq a\,,\ i=1,2,3\}.$$

(c) Em que região é que a sucessão gerada pelo método de Newton para a resolução de $F(x)=0$ é convergente se $(x_0)_3 = 0$? E se $(x_0)_2 = (x_0)_3 = 0$?

7. Considere a função vectorial $\mathbb{R}^2 \longrightarrow \mathbb{R}^2$ definida por:

$$F(x_1,x_2) \;=\; \begin{bmatrix} e^{x_1} - e^{x_2} \\ \frac{1}{3}x_2^3 \end{bmatrix}.$$

Indique um valor para constante de Lipschitz da matriz Jacobiana de F no conjunto $[-1,1]\times[-1,1]$.

8. Tente detectar com o método de Newton, em MATLAB, partindo de pontos iniciais diferentes, as quatro raízes de

$$F(x) \;=\; \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ 5x_1^2 - x_2^2 - 2 \end{bmatrix} \;=\; 0.$$

O que acontece se começar com o ponto $x_0 = [\,\sqrt{2}/2 \;\; 0\,]^\top$? Como poderia contornar o problema?

Aula 3: Taxa de Convergência Local do Método de Newton para Sistemas de Equações Não Lineares

Vamos provar que o erro absoluto entre as iteradas x_k geradas pelo método de Newton e a solução x_* decresce quadraticamente, em norma, se x_0 estiver suficientemente próximo de x_*. Demonstrar-se-á, também, que o método de Newton está bem definido localmente, ou seja, que, para um ponto inicial x_0 na vizinhança de uma solução x_* para a qual a matriz Jacobiana é não singular, as matrizes Jacobianas de todos os pontos x_k são, também, não singulares.

Vimos em aula anterior que

$$\|x_1 - x_*\| \;\leq\; \|J(x_0)^{-1}\|\,\|F(x_*) - [F(x_0) + J(x_0)(x_* - x_0)]\| \;\leq\; \frac{\gamma}{2}\|J(x_0)^{-1}\|\,\|x_0 - x_*\|^2,$$

em que γ é a constante de Lipschitz da matriz Jacobiana J da função vectorial F no domínio D.

As propriedades de convergência local do método de Newton são enunciadas e provadas no seguinte teorema.

Teorema 1 *Seja $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ uma função vectorial definida num conjunto aberto D. Suponha-se que F é continuamente diferenciável em D e que a sua matriz Jacobiana J é contínua à Lipschitz em D com constante γ.*

Seja $x_* \in D$ uma solução de $F(x) = 0$ para a qual $J(x_*)$ é não singular. Seja, ainda, β um escalar positivo tal que $\|J(x_*)^{-1}\| \leq \beta$.

Nestas condições, existe um escalar ϵ positivo tal que se

$$\|x_0 - x_*\| \;\leq\; \epsilon$$

então a sucessão $\{x_k\}$ gerada pelo método de Newton está bem definida, converge para x_* e satisfaz

$$\|x_{k+1} - x_*\| \;\leq\; \beta\gamma\|x_k - x_*\|^2.$$

Demonstração. Comecemos por colocar uma bola centrada em x_* dentro de D, o que é possível por este ser aberto. Seja $r > 0$ o raio dessa bola. Se $\epsilon \leq r$, então x_0 está nesta bola e, consequentemente, em D.

A demonstração é feita por indução. Vejamos, primeiro, o que acontece de x_0 para x_1. Da continuidade à Lipschitz de J em D vem que

$$\|J(x_*)^{-1}[J(x_0) - J(x_*)]\| \;\leq\; \|J(x_*)^{-1}\|\,\|J(x_0) - J(x_*)\| \;\leq\; \beta\gamma\|x_0 - x_*\|.$$

Desta forma, se escolhermos

$$\epsilon \;=\; \min\left\{r, \frac{1}{2\beta\gamma}\right\}$$

temos que

$$\|J(x_*)^{-1}[J(x_0) - J(x_*)]\| \;\leq\; \frac{1}{2} \;<\; 1.$$

Estamos, assim, em condições de aplicar o resultado do Exercício 2 (com $A = J(x_*)$ e $B = J(x_0)$) e afirmar que $J(x_0)$ é invertível. Para além disso, a desigualdade desse exercício diz-nos que

$$\|J(x_0)^{-1}\| \leq \frac{\|J(x_*)^{-1}\|}{1 - \|J(x_*)^{-1}[J(x_0) - J(x_*)]\|} \leq \frac{\|J(x_*)^{-1}\|}{1 - 1/2} \leq 2\beta.$$

Combinando a derivação feita antes do teorema com este limite superior para $\|J(x_0)^{-1}\|$, resulta em

$$\|x_1 - x_*\| \leq 2\beta\frac{\gamma}{2}\|x_0 - x_*\|^2 = \beta\gamma\|x_0 - x_*\|^2,$$

mostrando que estamos no caminho certo. Como, por hipótese, $\|x_0 - x_*\| \leq \epsilon$, esta desigualdade implica que

$$\|x_1 - x_*\| \leq \beta\gamma\|x_0 - x_*\|\frac{1}{2\beta\gamma} = \frac{1}{2}\|x_0 - x_*\| \leq \frac{\epsilon}{2} \leq \epsilon.$$

Logo, o ponto x_1 está nas mesmas condições das do ponto x_0. Seria possível, assim, provar os mesmos limites para o erro $\|x_2 - x_*\|$ em função do erro $\|x_1 - x_*\|$.

Raciocinando indutivamente, estabelece-se, para todo o k, que $J(x_k)$ é não singular, que

$$\|x_{k+1} - x_*\| \leq \beta\gamma\|x_k - x_*\|^2,$$

e que

$$\|x_{k+1} - x_*\| \leq \frac{1}{2}\|x_k - x_*\|.$$

A sucessão $\{x_k\}$ converge para x_* (porquê?) e a taxa de convergência das iteradas é quadrática. ■

Este teorema estuda a convergência local do método de Newton, ao assumir que o ponto inicial pertence a uma bola centrada numa solução. O raio desta bola é desconhecido aquando da aplicação do método, uma vez que depende dos valores F e de J nessa solução.

No entanto, é importante analisar o produto $\beta\gamma$ que aparece a multiplicar $\|x_k - x_*\|^2$, uma vez que, sendo grande, pode explicar uma certa deterioração da taxa quadrática de convergência para determinados problemas. Um produto $\beta\gamma$ grande sugere, também, uma região de convergência local mais pequena (pois ϵ é inversamente proporcional a $\beta\gamma$).

A constante β, que podemos considerar, nesta discussão, igual a $\|J(x_*)^{-1}\|$, indica a distância de $J(x_*)$ à singularidade matricial. No caso unidimensional, $J(x_*)^{-1} = 1/f'(x_*)$ é grande, em valor absoluto, quando o declive $f'(x_*)$ da recta tangente for pequeno em módulo. A condição $\|J(x_*)^{-1}\| \leq \beta$ reduz-se a

$$\left|\frac{1}{f'(x_*)}\right| \leq \beta.$$

A constante γ da continuidade à Lipschitz da matriz Jacobiana num vizinhança de x_* mede o grau de não linearidade de F. Quanto mais não linear é J, maior é a sua constante de Lipschitz γ.

O método de Newton pode ser *globalizado* de forma a que seja garantida a sua convergência a partir de pontos arbitrários. O estudo de *estratégias de globalização* e suas propriedades está fora do contexto desta disciplina. Porém, não resistimos a observar que o método da bisseção é uma estratégia de globalização para o caso unidimensional.

Exercícios

1. Seja $E \in \mathbb{R}^{n\times n}$. Prove que se $\|E\| < 1$ então $I - E$ é não singular e
$$\|(I-E)^{-1}\| \;\leq\; \frac{1}{1-\|E\|}.$$
Sugestão: Prove primeiro, por contradição, que $I - E$ é não singular. Depois, defina
$$S_k \;=\; \sum_{j=0}^{k} E^j$$
e mostre que
$$S_k - (I-E)^{-1} \;=\; -(I-E)^{-1}E^{k+1}$$
e
$$(I-E)^{-1} \;=\; \sum_{k=0}^{+\infty} E^k.$$

2. Sejam $A, B \in \mathbb{R}^{n\times n}$ duas matrizes tais que A é não singular e $\|-A^{-1}(B-A)\| < 1$. Demonstre, utilizando o exercício anterior, que B é não singular e que a norma da sua inversa satisfaz
$$\|B^{-1}\| \;\leq\; \frac{\|A^{-1}\|}{1-\|A^{-1}(B-A)\|}.$$

3. Prove, nas condições do Teorema 1 e utilizando os seus resultados, que a sucessão $\{F(x_k)\}$ converge quadraticamente para o vector nulo.

4. Neste exercício pretende-se analisar a taxa quadrática de convergência local do método de Newton *inexacto*. Suponha que em vez de ser calculado o passo de Newton exacto é determinado um passo p_k tal que
$$J(x_k)p_k \;=\; -F(x_k) + e_k,$$
em que $e_k \in \mathbb{R}^n$ representa o erro residual. Prove, nas condições do Teorema 1, que a sucessão $\{x_k\}$ converge quadraticamente para x_* se existir uma constante positiva c tal que
$$\|e_k\| \;\leq\; c\|F(x_k)\|^2 \qquad \text{para todo o } k.$$

Aula 4: Métodos de Quasi-Newton para Sistemas de Equações Não Lineares

Os valores das derivadas das funções-componente de F não se encontram disponíveis em diversas situações práticas. Uma das alternativas possíveis para resolver $F(x) = 0$ nestas situações é a aplicação de métodos de quasi-Newton (conhecidos, também, por métodos de secante). Outra alternativa passa pelo recurso a técnicas de diferenciação numérica.

O método de secante para a resolução de uma equação não linear a uma incógnita ($f(x) = 0$, $f : \mathbb{R} \longrightarrow \mathbb{R}$) tem uma interpretação geométrica simples. A equação da recta que passa pelos pontos $(x_0, f(x_0))$ e $(x_1, f(x_1))$ é dada por

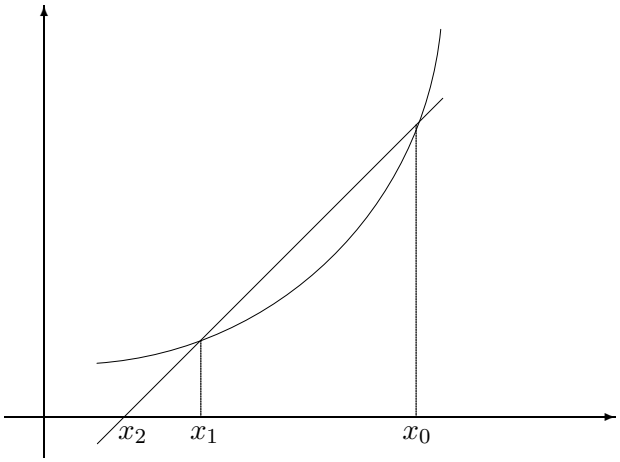
$$\frac{y - f(x_1)}{x - x_1} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

A nova iterada x_2 é a primeira coordenada do ponto de intersecção desta recta com o eixo das abcissas. Se fizermos $y = 0$ e $x = x_2$, obtemos

$$x_2 = x_1 - \frac{f(x_1)}{a_1} \quad \text{com} \quad a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Dados x_0 e x_1, o método de secante gera uma sucessão de pontos da forma

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



A fórmula do método de secante que determina a_{k+1} pode ser reescrita na forma

$$a_{k+1}\underbrace{(x_{k+1} - x_k)}_{\| \; s_k} = \underbrace{f(x_{k+1}) - f(x_k)}_{\| \; y_k}.$$

A equação linear em a

$$a\,(x_{k+1} - x_k) \;=\; f(x_{k+1}) - f(x_k)$$

é designada por equação de secante. No caso unidimensional, esta equação admite uma solução única (a_{k+1}).

Quando passamos a um sistema de n equações a n incógnitas, $F(x) = 0$, a equação de secante assume o aspecto

$$As_k \;=\; y_k$$

em que $A \in \mathbb{R}^{n\times n}$ é uma matriz de ordem n e s_k e y_k são dois vectores

$$s_k \;=\; x_{k+1} - x_k \;\in\; \mathbb{R}^n \qquad \text{e} \qquad y_k \;=\; F(x_{k+1}) - F(x_k) \;\in\; \mathbb{R}^n.$$

Esta equação de secante continua a ser linear (em A) mas passa a ter mais equações do que incógnitas. Existem n^2 componentes em A para apenas n igualdades.

Uma outra forma de chegar à equação de secante para sistemas de equações não lineares parte do *modelo* linear

$$m_{k+1}(s) \;\stackrel{\text{def}}{=}\; F(x_{k+1}) + A_{k+1}s,$$

enquanto aproximação para $F(x)$ numa vizinhança de x_{k+1},

$$F(x) \;\simeq\; m_{k+1}(x - x_{k+1}) \;=\; F(x_{k+1}) + A_{k+1}(x - x_{k+1}).$$

Se fizermos $s = 0$, vem que $m_{k+1}(0) = F(x_{k+1})$, o que traduz o facto do modelo ser exacto em x_{k+1}. Seria adequado que este mesmo modelo possuísse informação sobre o avanço de x_k para x_{k+1} (ou sobre o recuo de x_{k+1} para x_k). É imposta ao modelo, assim, a condição

$$m_{k+1}(-(x_{k+1} - x_k)) \;=\; F(x_k).$$

Ao fazermos as contas, verificamos que esta condição é equivalente a

$$A_{k+1}(x_{k+1} - x_k) \;=\; F(x_{k+1}) - F(x_k).$$

Resta-nos observar que A_{k+1} satisfaz a equação de secante $As_k = y_k$.

A equação de secante admite uma infinidade de soluções. Um tipo de soluções consiste em adicionar a A_k uma matriz de característica igual a 1 (a fim que A_k *mude* o menos possível). Se a diferença $A_{k+1} - A_k$ tiver característica um, então é porque existem vectores $u, v \in \mathbb{R}^n$ tais que

$$A_{k+1} - A_k \;=\; \underbrace{u}_{n\times 1}\,\underbrace{v^\top}_{1\times n}.$$

Se escolhermos $v = s_k$ então $(A_{k+1} - A_k)w = 0$ para todos os vectores w ortogonais a s_k, o que é atraente num cenário de *menor mudança* possível. Multiplicando ambos os membros desta equação por s_k, à esquerda, resulta em

$$\underbrace{A_{k+1}s_k}_{\shortparallel\; y_k} - A_k s_k \;=\; y_k - A_k s_k \;=\; (v^\top s_k)u,$$

o que nos informa que u é um múltiplo de $y_k - A_k s_k$. Com $v = s_k \neq 0$ vem que

$$u \;=\; \frac{1}{s_k^\top s_k}(y_k - A_k s_k).$$

Chegámos, assim, à seguinte expressão para A_{k+1}:

$$A_{k+1} \;=\; A_k + \frac{1}{s_k^\top s_k}(y_k - A_k s_k)s_k^\top,$$

conhecida por fórmula de actualização de Broyden.

O que vamos provar de seguida é que a matriz da actualização de Broyden é, de entre as que satisfazem a equação de secante, a que está mais perto, num certo sentido, de A_k.

Proposição 1 *Seja A_{k+1} a matriz da fórmula de actualização de Broyden (com $s_k \neq 0$) e $\|\cdot\|$ a norma ℓ_2.*

1. *A_{k+1} satisfaz a equação de secante $As_k = y_k$.*
2. $$\|A_{k+1} - A_k\| \;=\; \min\left\{\|A - A_k\| :\; As_k = y_k\right\}.$$

Demonstração. A Parte 1 resulta, trivialmente, da multiplicação de A_{k+1} por s_k:

$$A_{k+1}s_k \;=\; A_k s_k + \frac{1}{s_k^\top s_k}(y_k - A_k s_k)s_k^\top s_k \;=\; y_k.$$

Para provar a segunda parte, seja A uma matriz $n \times n$ a satisfazer $As_k = y_k$. Então

$$A_{k+1} - A_k \;=\; \frac{1}{s_k^\top s_k}\;\underbrace{(y_k - A_k s_k)}_{\begin{array}{c}\shortparallel\\ As_k - A_k s_k\end{array}}\; s_k^\top \;=\; \frac{1}{s_k^\top s_k}(A - A_k)s_k s_k^\top.$$

Aplicando propriedades da norma Euclideana, escrevemos

$$\|A_{k+1} - A_k\| \;\leq\; \|A - A_k\| \left\| \frac{1}{s_k^\top s_k} s_k s_k^\top \right\| \;=\; \|A - A_k\|,$$

o que conclui a demonstração. ■

É apresentado, de seguida, o método de quasi-Newton baseado na fórmula de actualização de Broyden.

Método de Broyden para Sistemas de Equações não Lineares
Escolher $x_0 \in \mathbb{R}^n$ e $A_0 \in \mathbb{R}^{n\times n}$ não singular.
Para $k = 0, 1, 2, \ldots$

1. Resolver o sistema de equações lineares $A_k s_k = -F(x_k)$.
2. Fazer $x_{k+1} = x_k + s_k$ e $y_k = F(x_{k+1}) - F(x_k)$.
3. Calcular
$$A_{k+1} \;=\; A_k + \frac{1}{s_k^\top s_k}(y_k - A_k s_k)s_k^\top.$$

Exercícios

1. Calcule as iteradas x_1 e x_2 do método de Broyden para a resolução do sistema de equações não lineares $F(x) = 0$, em que
$$F(x) \;=\; \begin{bmatrix} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{bmatrix},$$
começando com: (a) $x_0 = [\,1 \;\; 5\,]^\top$ e $A_0 = I$; (b) $x_0 = [\,1 \;\; 5\,]^\top$ e $A_0 = J(x_0)$.

2. Seja $v \in \mathbb{R}^n$ um vector não nulo. Prove que, na norma ℓ_2, é verdadeira a igualdade
$$\left\| \frac{1}{v^\top v} v v^\top \right\| \;=\; 1.$$

3. Seja $F(x) = Cx + c$, com $C \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}^n$. Considere dois pontos x_k e x_{k+1} distintos. Mostre que C satisfaz a equação de secante $As_k = y_k$.

Aula 5: Taxa de Convergência Local dos Métodos de Quasi-Newton para Sistemas de Equações Não Lineares

Os métodos de quasi-Newton ou de secante não apresentem a mesma taxa (quadrática) de convergência local do método de Newton. No entanto, convergem superlinearmente, o que, em termos práticos, é muito satisfatório. Em grande número de ocorrências, não se distinguem os comportamentos numéricos das taxas superlinear e quadrática.

Ilustramos o desempenho numérico do método de Broyden com o mesmo sistema de equações não lineares ao qual se aplicou o método de Newton e que foi definido pela função:

$$F(x) = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ 5x_1^2 - x_2^2 - 2 \end{bmatrix}.$$

Correu-se o método de Broyden, em MATLAB, começando com o mesmo ponto inicial $x_0 = [2 \quad 2]^\top$. A matriz A_0 escolhida foi $\|F(x_0)\|I$. O método gera iteradas que se aproximam da solução $x_* = [\sqrt{2}/2 \quad -\sqrt{2}/2]^\top$. Repare-se que, partindo do mesmo ponto inicial, os métodos de Broyden e de Newton convergiram para soluções diferentes. Reproduzimos, em baixo, o *output* do código.

```
>> Broyden
-----------------------------------------------------------------
               Metodo de Broyden para F(x) = 0
-----------------------------------------------------------------
 Iteracao     |x(1)-x*(1)|          |x(2)-x*(2)|        || x-x* ||
     0     1.2928932188134525     2.7071067811865475    3.0000e+000
     1     0.8456796233134946     1.8126795901866317    2.0002e+000
     2     0.4450899479441283     0.4689664890487301    6.4656e-001
     3     0.1888472653876047     0.7933170962329951    8.1548e-001
     4     0.0014854454921065     0.9329779661030725    9.3298e-001
     5     0.0135802307552687     0.7190994909274900    7.1923e-001
     6     0.0047682327329389     0.1462568630927651    1.4633e-001
     7     0.0002471415432675     0.0403884748676994    4.0389e-002
     8     0.0000998044959372     0.0067853796901328    6.7861e-003
     9     0.0000130697331487     0.0006553722147943    6.5550e-004
    10     0.0000003216540551     0.0000149397033994    1.4943e-005
    11     0.0000000020803608     0.0000000837210314    8.3747e-008
    12     0.0000000000239055     0.0000000008849602    8.8528e-010
    13     0.0000000000001673     0.0000000000061897    6.1920e-012
-----------------------------------------------------------------
>>
```

Observa-se que o erro não converge quadraticamente para zero mas, a partir da sexta iteração, o erro approxima-se de zero muito rapidamente. As primeiras cinco iterações, responsáveis por um desempenho mais lento em comparação com o registado para o

método de Newton, têm mais a ver com o comportamento global do método. O método de Broyden conduziu as iteradas para uma solução $x_*^B = [\,\sqrt{2}/2 \;\; -\sqrt{2}/2\,]^\top$, mais afastada do ponto inicial x_0 do que a solução detectada pelo método de Newton $x_*^N = [\,\sqrt{2}/2 \;\; \sqrt{2}/2\,]^\top$.

O teorema seguinte enuncia as condições sob as quais o método de Broyden converge localmente. A demonstração, demasiado longa e pormenorizada, é omitida.

Teorema 1 *Nas condições do teorema da convergência local do método de Newton, existem escalares ϵ e δ positivos tais que, se*

$$\|x_0 - x_*\| \;\leq\; \epsilon \quad e \quad \|A_0 - J(x_*)\| \;\leq\; \delta,$$

então a sucessão $\{x_k\}$ gerada pelo método de Broyden está bem definida e converge superlinearmente para x_*.

Para compreender melhor o método de Broyden torna-se indispensável estudar o comportamento assimptótico da sucessão $\{A_k\}$ das matrizes de quasi-Newton. Será que A_k tende para $J(x_*)$ quando x_k converge para x_*? No método de Newton, $J(x_k)$ converge para $J(x_*)$. Porém, A_k pode não convergir para $J(x_*)$, mesmo nas condições do teorema anterior. É simples provar, para o sistema de equações não lineares definido pela função

$$\begin{aligned} &F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \\ &F(x) \;=\; \left[\begin{array}{c} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{array}\right] \;=\; \left[\begin{array}{c} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{array}\right], \end{aligned}$$

que o método de Broyden, partindo de $x_0 = [\,1 \;\; 5\,]^\top$ e $A_0 = J(x_0)$, gera matrizes de secante A_k a satisfazer

$$\lim_{k \to +\infty} A_k \;=\; \left[\begin{array}{cc} 1 & 1 \\ 1.5 & 7.5 \end{array}\right] \;\neq\; \left[\begin{array}{cc} 1 & 1 \\ 0 & 6 \end{array}\right] \;=\; J(x_*),$$

com $x_* = [\,0 \;\; 3\,]^\top$.

Não se pense, porém, que as matrizes do método de Broyden não possuem as características assimptóticas adequadas. De facto, estas matrizes produzem passos $s_k = -A_k^{-1} F(x_k)$ que se aproximam assimptoticamente (sob escalonamento) dos passos de Newton $p_k = -J(x_k)^{-1} F(x_k)$. Ou seja, nas condições do teorema anterior, é verdadeiro o limite

$$\lim_{k \to +\infty} \frac{\|s_k - p_k\|}{\|s_k\|} \;=\; 0.$$

Corremos, novamente, o método de Broyden, mas desta vez para este segundo exemplo. Os resultados são apresentados de seguida. Imprimem-se as quantidades $\|s_k\|$, $\|s_k - p_k\|$ e $\|s_k - p_k\|/\|s_k\|$. Observa-se que todas, e em particular esta última, se aproximam de zero.

```
>> Broyden
-------------------------------------------------------------------
               Metodo de Broyden para F(x) = 0
-------------------------------------------------------------------
 Iteracao     ||s-p||          ||s||       ||s-p||/||s||     || x-x* ||
     1     0.0000e+000     2.1287e+000     0.0000e+000     8.8388e-001
     2     2.2845e-002     7.7675e-001     2.9412e-002     1.0714e-001
     3     1.5518e-002     8.9044e-002     1.7428e-001     1.8094e-002
     4     3.6730e-004     1.7650e-002     2.0810e-002     4.4381e-004
     5     1.8381e-006     4.4193e-004     4.1593e-003     1.8845e-006
     6     1.9628e-010     1.8843e-006     1.0416e-004     1.9712e-010
     7     6.9468e-017     1.9712e-010     3.5242e-007     2.0775e-016
-------------------------------------------------------------------
>>
```

Em termos gerais, os métodos de quasi-Newton apresentam a seguinte caracterização necessária e suficiente de convergência superlinear.

Teorema 2 *Seja $F : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ uma função vectorial definida num conjunto aberto D. Suponha-se que F é continuamente diferenciável em D e que a sua matriz Jacobiana J é contínua à Lipschitz em D com constante γ. Seja $x_* \in D$ uma solução de $F(x) = 0$ para a qual $J(x_*)$ é não singular.*

Considere um método a gerar uma sucessão de pontos da forma

$$x_{k+1} \;=\; x_k - A_k^{-1} F(x_k).$$

Suponha-se que as matrizes da sucessão $\{A_k\}$ são não singulares e que $\{x_k\}$ converge para x_* (com $x_k \neq x_*$ para todo o k).

Nestas condições, $\{x_k\}$ converge superlinearmente para x_* se e só se satisfaz a condição de Dennis-Moré

$$\lim_{k\to+\infty} \frac{\| \left(A_k - J(x_k)\right) s_k \|}{\|s_k\|} \;=\; 0.$$

O método de Broyden satisfaz a condição de Dennis-Moré e, por isso mesmo, apresenta uma taxa de convergência superlinear. É fácil provar, nas condições deste último teorema, que

$$\lim_{k\to+\infty} \frac{\| \left(A_k - J(x_k)\right) s_k \|}{\|s_k\|} \;=\; 0 \quad\Longleftrightarrow\quad \lim_{k\to+\infty} \frac{\|s_k - p_k\|}{\|s_k\|} \;=\; 0.$$

Para este efeito note-se que

$$\left(A_k - J(x_k)\right)\left(-A_k^{-1} F(x_k)\right) \;=\; -F(x_k) - J(x_k) s_k \;=\; J(x_k)\left(p_k - s_k\right).$$

Em algumas situações práticas, a possibilidade de contornar a utilização da matriz Jacobiana pode representar uma vantagem. Outra qualidade do método de Broyden que

o torna atraente, quando comparado com o método de Newton, é o seu baixo custo computacional por iteração.

O método de Newton requer a solução de um sistema de equações lineares $J(x_k)p = -F(x_k)$ em cada iteração, cujo o número de operações elementares é da ordem de n^3. Ora, os sistemas de equações lineares $A_k s = -F(x_k)$ do método de Broyden, dada a estrutura especial de A_k, podem ser resolvidos na ordem de n^2 operações.

De facto, utilizando a fórmula de Sherman-Morrison-Woodbury, escreve-se

$$A_{k+1}^{-1} \;=\; A_k^{-1} + \frac{1}{s_k^\top A_k^{-1} y_k} \left(s_k - A_k^{-1} y_k\right) s_k^\top A_k^{-1}.$$

Apresentamos, a seguir, o método de Broyden com base nesta actualização das inversas $B_k = A_k^{-1}$. Vê-se, assim, que cada iteração requer somente produtos internos e produtos matriz-vector, mantendo um número de operações elementares da ordem de n^2.

Método de Broyden para Sistemas de Equações não Lineares
Escolher $x_0 \in \mathbb{R}^n$ e $B_0 \in \mathbb{R}^{n\times n}$ não singular.
Para $k = 0, 1, 2, \ldots$

1. Fazer $s_k = -B_k F(x_k)$.
2. Fazer $x_{k+1} = x_k + s_k$ e $y_k = F(x_{k+1}) - F(x_k)$.
3. Calcular
$$B_{k+1} \;=\; B_k + \frac{1}{s_k^\top B_k y_k} \left(s_k - B_k y_k\right) s_k^\top B_k.$$

Exercícios

1. Mostre que se a sucessão $\{x_k\}$ convergir superlinearmente para x_* (com $x_k \neq x_*$ para todo o k) então
$$\lim_{k\to+\infty} \frac{\|x_{k+1} - x_k\|}{\|x_k - x_*\|} \;=\; 1.$$

2. Sejam u e v dois vectores de $\mathbb{R}^n$ e A uma matriz $n \times n$ não singular. Prove que $A + uv^\top$ é não singular se e só se
$$\sigma \;\stackrel{\text{def}}{=}\; 1 + v^\top A^{-1} u \;\neq\; 0.$$
Mostre que a fórmula (de Sherman-Morrison-Woodbury) para a inversa de $A + uv^\top$ quando $\sigma \neq 0$ é:
$$\left(A + uv^\top\right)^{-1} \;=\; A^{-1} + \frac{1}{\sigma} A^{-1} u v^\top A^{-1}.$$

3. **(Difícil.)** Nas condições do Teorema 2, prove que

$$\lim_{k\to+\infty} \frac{\|\left(A_k - J(x_k)\right) s_k\|}{\|s_k\|} \;=\; 0 \quad\Longleftrightarrow\quad \lim_{k\to+\infty} \frac{\|s_k - p_k\|}{\|s_k\|} \;=\; 0.$$

Aula 6: Conceitos Básicos sobre Optimização sem Restrições

Um ponto x_* diz-se um minimizante local (ou relativo) de $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ se existir um $\epsilon > 0$ a satisfazer

$$f(x_*) \leq f(x) \quad \text{para todo o } x \in D \text{ tal que } \|x - x_*\| \leq \epsilon.$$

O minimizante local x_* diz-se estrito ou forte se

$$f(x_*) < f(x) \quad \text{para todo o } x \in D, \ x \neq x_*, \ \text{tal que } \|x - x_*\| \leq \epsilon.$$

O minimizante local x_* diz-se isolado se for o único minimizante local de f na intersecção de D com a bola $\{x \in \mathbb{R}^n : \|x - x_*\| \leq \epsilon\}$. Todo o minimizante local isolado é estrito. Porém, no exemplo

$$f(x) = \begin{cases} x^4 \cos(1/x) + 2x^4 & x \neq 0, \\ 0 & x = 0, \end{cases}$$

o ponto $x_* = 0$ é um minimizante local estrito que não é isolado.

Está fora do âmbito desta disciplina a determinação de minimizantes globais (ou absolutos) para o problema

$$\min_{x \in D \subset \mathbb{R}^n} f(x).$$

A optimização diz-se sem restrições quando não é restrita a nenhum subconjunto de $\mathbb{R}^n$. Observa-se que o conjunto D representa o domínio da função e que a procura de minimizantes locais em D é irrestrita.

A determinação de minimizantes locais de uma função em $\mathbb{R}^n$ está relacionada com a resolução de sistemas de equações não lineares. O gradiente da função f é nulo num seu minimizante local. Logo, a minimização local de f passa, necessariamente, pela resolução do sistema de equações não lineares $\nabla f(x) = 0$.

Por outro lado, uma solução de um sistema de equações não lineares da forma $F(x) = 0$ é um minimizante (absoluto) da função $\|F(x)\|$, ou, se quisermos manter a suavidade de F, da função $\|F(x)\|^2$.

Seja $x_* \in D$ um minimizante local de uma função $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ continuamente diferenciável no aberto D. Consideremos a função real de variável real definida, para t suficientemente pequeno ($t \in [0, t_1)$), por

$$g(t) = f(x_* - t\nabla f(x_*)).$$

Tem-se que $g(0) = f(x_*)$ e $g'(0) = -\|\nabla f(x_*)\|^2$. Se ∇f não se anular em x_* então $g'(0) < 0$. Como, pelas hipóteses feitas sobre f, a derivada g' é contínua, então existe um $t_2 > 0$ para o qual $g'(s) < 0, s \in [0, t_2)$. Logo,

$$g(t) - g(0) = g'(s)(t - 0) < 0 \implies f(x_* - t\nabla f(x_*)) < f(x_*), \quad \forall t \in (0, \min\{t_1, t_2\}).$$

O que mostrámos contradiz a hipótese de x_* ser minimizante local. O gradiente de f ser nulo em minimizantes locais constitui a condição necessária de optimalidade de primeira ordem para funções continuamente diferenciáveis. Um ponto em que o gradiente de f se anule diz-se um ponto estacionário ou crítico de f.

De forma semelhante, provar-se-ia, para funções duas vezes continuamente diferenciáveis, que, em minimizantes locais, o gradiente ∇f é nulo e a matriz Hessiana $\nabla^2 f$ é semi-definida positiva (condições necessárias de optimalidade de segunda ordem).

As condições suficientes de optimalidade para funções duas vezes continuamente diferenciáveis são mais fortes do que as necessárias e exigem que a matriz Hessiana seja definida positiva.

Resumimos estes três factos na proposição seguinte.

Proposição 1 *Seja $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função definida num domínio D aberto.*

Se f for continuamente diferenciável em D e $x_* \in D$ for um minimizante local de f então $\nabla f(x_*) = 0$.

Se f for duas vezes continuamente diferenciável em D e $x_* \in D$ for um minimizante local de f então $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*)$ é semi-definida positiva.

Se f for duas vezes continuamente diferenciável em D, $\nabla f(x_*) = 0$ e $\nabla^2 f(x_*)$ for definida positiva, com $x_* \in D$, então x_* é um minimizante local (estrito) de f.

As condições suficientes de segunda ordem podem não ser necessárias. A função $f(x) = x^4$ admite um mínimo em $x_* = 0$ e, no entanto, $f''(x_*) = 0$.

O vector simétrico do vector gradiente $-\nabla f(x)$, utilizado na demonstração da condição necessária de optimalidade, desempenha um papel importante em optimização. Esta direcção é designada por *direcção de descida máxima.*

Uma direcção $p \in \mathbb{R}^n$ diz-se de descida em x se existir um $\bar{t} > 0$ para o qual

$$f(x + tp) \; < \; f(x), \quad \text{para todo o } \; t \in (0, \bar{t}).$$

Vimos, na demonstração da condição necessária de optimalidade, que $-\nabla f(x_*) \neq 0$ é uma direcção de descida. A explicação sobre o facto desta direcção ser de descida máxima fica para um exercício.

São de descida todas as direcções que fizerem um ângulo de amplitude inferior a $\pi/2$ com a direcção de descida máxima:

Proposição 2 *Seja $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função continuamente diferenciável no domínio D aberto e $x \in D$. A direcção $p \in \mathbb{R}^n$ é de descida em x se*

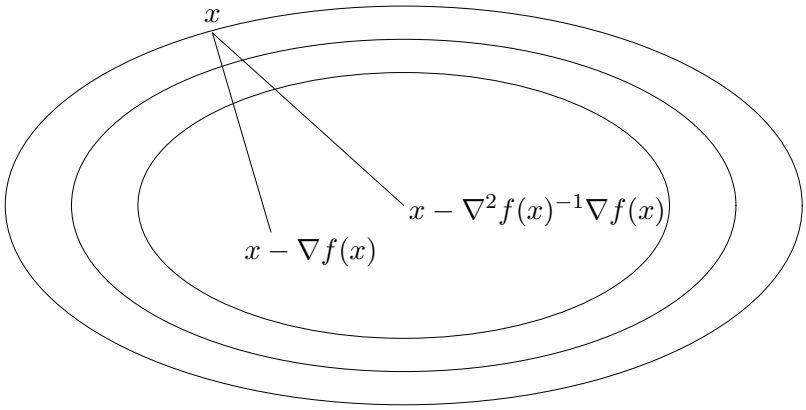
$$-\nabla f(x)^\top p \; > \; 0.$$

Outra direcção essencial em optimização é o passo de Newton. O passo de Newton para a resolução do sistema de equações não lineares $\nabla f(x) = 0$, quando bem definido, é dado por

$$-\nabla^2 f(x)^{-1} \nabla f(x)$$

(a matriz Jacobiana de ∇f é a matriz Hessiana de f). É fácil confirmar que o passo de Newton é uma direcção de descida se a matriz Hessiana for definida positiva num ponto em que o gradiente não se anule:

$$-\nabla f(x)^\top \left(-\nabla^2 f(x)^{-1}\nabla f(x)\right) \;>\; 0.$$



Exercícios

1. Demonstre as condições de segunda ordem (necessárias e suficientes) de optimalidade da Proposição 1.

2. Seja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função duas vezes continuamente diferenciável. Considere a função real de variável real $g(t) = f(x + tp)$, definida para $x, p \in \mathbb{R}^n$.

   (a) Escreva, para esta função g, a fórmula de Taylor de ordem um (com resto de Lagrange de ordem dois) centrada em 0. Identifique $p^\top \nabla f(x)$ como sendo a taxa de variação de f, a partir de x e ao longo de p.

   (b) Considere o seguinte problema em p

   $$\min_{p\in\mathbb{R}^n} \; p^\top \nabla f(x) \quad \text{sujeito a} \quad \|p\| = 1.$$

   Prove que a solução óptima deste problema é $-\frac{1}{\|\nabla f(x)\|}\nabla f(x)$. **Sugestão:** Utilize $p^\top \nabla f(x) = \cos(\text{ang}(p, \nabla f(x)))\|p\|\|\nabla f(x)\|$.

3. Seja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função duas vezes continuamente diferenciável. Dado um ponto $y \in \mathbb{R}^n$ em que a matriz Hessiana é não singular, considere a direcção $d(y)$ definida por

   $$d(y) \;=\; -\nabla f(y) - \nabla^2 f(y)^{-1}\nabla f(y)\,.$$

   (a) Mostre que $d(y)$ é uma direcção de descida quando $\nabla^2 f(y)$ é definida positiva e $\nabla f(y) \neq 0$.

(b) Mostre que $d(y)$ é uma direcção de descida se $\|\nabla^2 f(y)^{-1}\| < 1$ e $\nabla f(y) \neq 0$.

(c) Considere, agora, as funções reais de duas variáveis reais $f(x_1, x_2) = x_1^4/3 + 2x_2^3/3 + 8x_1x_2$ e $f(x_1, x_2) = x_1^4 + 6x_2^2$. Seja $y = [\,1 \;\; 1\,]^\top$. Calculando apenas $\nabla^2 f(y)$ e os seus valores próprios, mostre que $d(y)$ é, para ambas as funções, uma direcção de descida.

4. Seja $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função continuamente diferenciável e $x \in \mathbb{R}^n$. Considere a direcção $d \in \mathbb{R}^n$ dada por

$$d \;=\; -\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial f}{\partial x_i}(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

em que i é um índice para o qual $|\frac{\partial f}{\partial x_j}(x)|$ assume o maior valor em $j \in \{1, \ldots, n\}$.

Prove que d é uma direcção de descida se $\nabla f(x) \neq 0$.

5. Demonstre a Proposição 2.

6. Considere uma função $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ definida por

$$f(x) \;=\; \frac{\alpha}{2}\|x\|^2,$$

com α um número real positivo.

(a) Calcule a direcção de descida máxima, d, e o passo de Newton, p, para a função f. Conclua que as duas direcções são linearmente dependentes.

(b) Que valor escolheria para α de forma a que a direcção de descida máxima e o passo de Newton coincidissem?

(c) Faça, agora, $\alpha = 1$. Calcule $f(x + d)$ e $f(x + p)$. O que conclui?

Aula 7: Métodos de Newton e de Quasi-Newton para Optimização sem Restrições

O método de Newton para optimização sem restrições consiste na aplicação do método de Newton à resolução do sistema de equações não lineares $\nabla f(x) = 0$. Quando bem definido, este método gera a sucessão de pontos $\{x_k\}$ descrita pelo seguinte algoritmo.

Método de Newton para Optimização sem Restrições
Escolher $x_0 \in \mathbb{R}^n$.
Para $k = 0, 1, 2, \ldots$

1. Resolver o sistema de equações lineares $\nabla^2 f(x_k) p_k = -\nabla f(x_k)$.
2. Fazer $x_{k+1} = x_k + p_k$.

As propriedades de convergência local são semelhantes às descritas e provadas para os sistemas de equações não lineares.

Teorema 1 *Seja $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ uma função definida num conjunto aberto D. Suponha-se que f é duas vezes continuamente diferenciável em D e que a sua matriz Hessiana $\nabla^2 f$ é contínua à Lipschitz em D com constante γ.*

Seja $x_* \in D$ uma solução de $\nabla f(x) = 0$ para a qual $\nabla^2 f(x_*)$ é não singular. Seja, ainda, β um escalar positivo tal que $\|\nabla^2 f(x_*)^{-1}\| \leq \beta$.

Nestas condições, existe um escalar ϵ positivo tal que se

$$\|x_0 - x_*\| \ \leq \ \epsilon$$

então a sucessão $\{x_k\}$ gerada pelo método de Newton está bem definida, converge para x_* e satisfaz

$$\|x_{k+1} - x_*\| \ \leq \ \beta\gamma \|x_k - x_*\|^2.$$

O único comentário que importa fazer, neste contexto, diz respeito à hipótese de não singularidade da matriz Hessiana $\nabla^2 f(x_*)$. Seria mais lógico, por estarmos a minimizar f, impor que a matriz Hessiana $\nabla^2 f(x_*)$ fosse definida positiva. No entanto, não há nenhum elemento neste método de Newton que faça a distinção entre minimização e maximização. O objectivo deste método é anular o gradiente da função, uma condição que é necessária quer em minimizantes locais quer em maximizantes locais.

Tendo em vista o objectivo de minimização, o passo de Newton pode ser *modificado* perturbando a matriz Hessiana:

$$\left(\nabla^2 f(x_k) + E_k\right) p_k^m = -\nabla f(x_k).$$

O objectivo da matriz $E_k \in \mathbb{R}^{n \times n}$ é tornar $\nabla^2 f(x_k) + E_k$ definida positiva caso $\nabla^2 f(x_k)$ não goze desta propriedade. Desta forma, há a garantia de p_k^m ser uma direcção de descida.

Os objectivos dos métodos de quasi-Newton para optimização sem restrições incluem os destes métodos para os sistemas de equações não lineares: (i) usar uma ordem de derivadas inferior à do método de Newton; (ii) apresentar uma taxa superlinear de convergência local; (iii) baixar o custo da álgebra linear por iteração de $\mathcal{O}(n^3)$ (Newton) para $\mathcal{O}(n^2)$.

As matrizes de actualização de quasi-Newton ou de secante devem tomar em linha de conta a especificidade do sistema $\nabla f(x) = 0$, em particular o facto de a matriz Jacobiana de ∇f ser a Hessiana de f e, como tal, ser simétrica. Assim, exige-se que a equação de secante seja resolvida por matrizes simétricas:

$$Hs_k \;=\; y_k \quad \text{e} \quad H = H^\top,$$

com $s_k = x_{k+1} - x_k$ e $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

Como no caso dos sistemas de equações não lineares, existem, no contexto da optimização sem restrições, inúmeras actualizações de secante. Vamos apresentar a mais popular e eficiente numericamente: a actualização de BFGS (Broyden, Fletcher, Goldfarb e Shanno). A fórmula de actualização de BFGS é dada por

$$H_{k+1} \;=\; H_k - \frac{1}{s_k^\top H_k s_k} H_k s_k s_k^\top H_k + \frac{1}{y_k^\top s_k} y_k y_k^\top.$$

É fácil provar que H_{k+1} satisfaz a equação de secante, e que é simétrica se H_k o for (ver exercício).

A matriz H_{k+1} é obtida somando à matriz H_k uma matriz de característica dois. As actualizações de característica dois são típicas em optimização pois estão associadas à manutenção da positividade dos valores próprios das matrizes de secante. Prova-se que H_{k+1} é definida positiva se H_k for simétrica e definida positiva e $y_k^\top s_k > 0$ (ver exercício).

É possível aplicar a fórmula de Sherman-Morrison-Woodbury para calcular a inversa B_k de H_k. Descrevemos, de seguida, o método de BFGS, com recurso à actualização das inversas B_k.

Método de BFGS para Optimização sem Restrições
Escolher $x_0 \in \mathbb{R}^n$ e $B_0 \in \mathbb{R}^{n\times n}$ simétrica e definida positiva.
Para $k = 0, 1, 2, \ldots$

1. Fazer $s_k = -B_k \nabla f(x_k)$.
2. Fazer $x_{k+1} = x_k + s_k$ e $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
3. Calcular
$$B_{k+1} \;=\; \left(I - \rho_k s_k y_k^\top\right) B_k \left(I - \rho_k y_k s_k^\top\right) + \rho_k s_k s_k^\top,$$
com $\rho_k = 1/(y_k^\top s_k)$.

O número de operações elementares por iteração é, claramente, da ordem de n^2: todas as operações matriciais são produtos internos ou produtos matriz-vector.

Continuamos, enunciando a taxa de superlinearidade do método de BFGS. Quando comparamos as condições em que este resultado é obtido com as enunciadas para o método de Broyden, ou mesmo para o método de Newton para optimização, constatamos que a hipótese de $\nabla^2 f(x^*) = J_{\nabla f}(x_*)$ ser não singular é fortalecida, exigindo-se que esta matriz seja definida positiva. Esta alteração está em sintonia com a capacidade, da actualização de BFGS, de manter as matrizes de secante definidas positivas.

Teorema 2 *Considere as hipóteses do teorema da convergência local do método de Newton, mas na situação, mais restrita, de $\nabla^2 f(x_*)$ ser definida positiva.*

Nestas condições, existem escalares ϵ e δ positivos tais que, se

$$\|x_0 - x_*\| \ \leq \ \epsilon \quad e \quad \|H_0 - \nabla^2 f(x_*)\| \ \leq \ \epsilon,$$

então a sucessão $\{x_k\}$ gerada pelo método de BFGS está bem definida e converge superlinearmente para x_*.

A condição de Dennis-Moré, que o método de BFGS satisfaz, escreve-se, para optimização sem restrições, na forma

$$\lim_{k \to +\infty} \frac{\| \left(H_k - \nabla^2 f(x_k)\right) s_k \|}{\|s_k\|} \ = \ 0.$$

A actualização de Broyden faz com que a matriz B_{k+1} seja a que esteja mais perto de B_k, num certo sentido, de entre todas as matrizes simétricas a satisfazer a equação de secante. De facto, se B_k for simétrica e definida positiva e $y_k^\top s_k > 0$, então a matriz B_{k+1} é a solução óptima do problema

$$\min_{B \in \mathbb{R}^{n \times n}} \ \|B - B_k\|_W \quad \text{sujeito a} \quad \underbrace{s_k = B y_k}_{\Updownarrow \ Hs_k = y_k \ \ H = B^{-1}} \quad \text{e} \ \ B = B^\top,$$

em que $\|\cdot\|_W$ é a norma definida por $\|B\|_W = \|W^{\frac{1}{2}} B W^{\frac{1}{2}}\|_F$, com W uma matriz definida positiva a satisfazer $y_k = W s_k$.

Exercícios

1. Mostre que o método de Newton converge numa iteração quando aplicado à função

$$f(x) \ = \ \frac{1}{2} x^\top H x + g^\top x,$$

com $H \in \mathbb{R}^{n \times n}$ uma matriz não singular e simétrica e $g \in \mathbb{R}^n$.

2. Este exercício é para ser resolvido em MATLAB. As duas primeiras alíneas deverão ser justificadas matematicamente.

   (a) Escreva uma função que, dada a matriz H simétrica, devolva uma matriz simétrica E para a qual o menor valor próprio de $H + E$ não seja inferior a 10^{-4}.

   (b) Escreva uma função que, dada a matriz H simétrica, devolva uma matriz simétrica E para a qual $H + E$ seja definida positiva (sem recorrer ao cálculo de valores próprios de H).

   (c) Explique qual seria a utilidade destes procedimentos numa implementação do método de Newton modificado para optimização sem restrições.

3. Considere a actualização de BFGS para H_{k+1} em função de H_k.

   (a) Mostre que H_{k+1} satisfaz a equação de secante.

   (b) Mostre que se H_k for simétrica então H_{k+1} também o é.

   (c) Prove que, se H_k for simétrica e definida positiva e $y_k^\top s_k > 0$, então H_{k+1} é definida positiva.

4. **(Difícil.)** Seja H_k uma matriz simétrica e definida positiva e s_k e y_k vectores tais que $y_k^\top s_k > 0$. Considere a factorização de Cholesky de H_k, dada por $H_k = L_k L_k^\top$. Encare a equação de secante $Hs_k = y_k$ escrita na forma

$$Lv_k = y_k \quad \text{e} \quad L^\top s_k = v_k,$$

   em que v_k é um vector auxiliar.

   (a) Determine $L = L_{k+1}$, em função de L_k, y_k e v_k de tal forma que a equação $Lv_k = y_k$ seja satisfeita e que L_{k+1} difira de L_k numa matriz de característica um. Mais concretamente, faça

$$L_{k+1} \;=\; L_k + uv^\top,$$

   com $v = v_k/\|v_k\|^2$ e determine u.

   (b) Utilize a expressão calculada na alínea anterior e $L^\top s_k = v_k$ com $L = L_{k+1}$ para concluir que $v_k = \alpha_k L_k^\top s_k$, em que α_k é um escalar real.

   (c) Determine α_k.

   (d) Prove que L_{k+1} assim determinada é não singular.

   (e) Confirme que obteve a actualização de BFGS, ou seja, que $H_{k+1} = L_{k+1} L_{k+1}^\top$.

5. Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função duas vezes continuamente diferenciável e com matriz Hessiana definida positiva em $\mathbb{R}^n$. Considere a mudança de variáveis:

$$x \;=\; Ry\,, \quad \text{em que } R \in \mathbb{R}^{n\times n} \text{ é uma matriz não singular.}$$

   Considere a função $g : \mathbb{R}^n \to \mathbb{R}$ definida por $g(y) = f(Ry)$. Por derivação composta, tem-se que $\nabla g(y) = R^\top \nabla f(Ry)$ e $\nabla^2 g(y) = R^\top \nabla^2 f(Ry) R$.

(a) Prove que a matriz Hessiana de g também é definida positiva em $\mathbb{R}^n$.

(b) Mostre que o método de Newton é invariante ao escalonamento nas variáveis, ou seja, que as fórmulas $x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$ e $y_{k+1} = y_k - \nabla^2 g(y_k)^{-1} \nabla g(y_k)$ são equivalentes.

(c) Mostre que quando a matriz R é ortogonal, $x_{k+1} = x_k - \nabla f(x_k)$ e $y_{k+1} = y_k - \nabla g(y_k)$ são equivalentes.

Aula 8: Problemas de Mínimos Quadrados Não Lineares

Uma das classes de problemas de optimização sem restrições que mais frequentemente surgem em aplicações é a dos problemas de mínimos quadrados não lineares. Estes problemas colocam-se na forma

$$\min_{x\in\mathbb{R}^n} \; f(x) \stackrel{\text{def}}{=} \frac{1}{2}R(x)^\top R(x) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{m} r_i(x)^2,$$

com $m > n$. Pretende-se, nestes problemas, minimizar a norma Euclideana de uma função residual

$$R : D \subset \mathbb{R}^n \to \mathbb{R}^m,$$
$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \to \begin{bmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{bmatrix} = R(x).$$

Estes problemas aparecem no contexto de ajuste de dados. Suponha que a função $m(y,p)$, com $m : \mathbb{R}^{n_y+n_p} \longrightarrow \mathbb{R}$, descreve o comportamento de um determinado sistema, cujo estado é descrito pelas variáveis de estado $y \in \mathbb{R}^{n_y}$. A definição do modelo depende do valor das variáveis $p \in \mathbb{R}^{n_p}$, que podem representar parâmetros cujo valor é procurado ou controlos do sistema cujo valor importa igualmente conhecer. Com base em resultados experimentais, conhecem-se m respostas do sistema (designadas por $\bar{r}_1, \ldots, \bar{r}_m$) para m valores de estados $\bar{y}_1, \ldots, \bar{y}_m$. O objectivo é, então, conhecer os valores dos parâmetros ou controlos p que melhor ajustam o modelo, no sentido dos mínimos quadrados, às respostas do sistema conhecidas. Pretende-se, assim, resolver o problema

$$\min_{p\in\mathbb{R}^n} \; f(p) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{m}(m(\bar{y}_i,p) - \bar{r}_i)^2.$$

Neste caso, tem-se que $x = p$ e $r_i(x) = m(\bar{y}_i, x) - \bar{r}_i$, $i = 1, \ldots, m$.

Regressemos à formulação mais geral do problema de mínimos quadrados não lineares. A primeira coisa a fazer é escrever o gradiente da função objectivo f:

$$\nabla f(x) \;=\; \sum_{i=1}^{m} r_i(x)\nabla r_i(x) \;=\; J(x)^\top R(x),$$

em que $J(x) \in \mathbb{R}^{m\times n}$ representa a matriz Jacobiana, em x, da função vectorial R. Note-se que a matriz $J(x)$ tem mais linhas do que colunas ('verticalmente' rectangular) e que, portanto, $J(x)^\top$ tem mais colunas do que linhas ('horizontalmente' rectangular). A seguir, derivamos o gradiente para obter a matriz Hessiana de f:

$$\nabla^2 f(x) \;=\; \sum_{i=1}^{m} \left(\nabla r_i(x)\nabla r_i(x)^\top + r_i(x)\nabla^2 r_i(x)\right) \;=\; J(x)^\top J(x) + S(x),$$

com $S(x) = \sum_{i=1}^{m} r_i(x)\nabla^2 r_i(x)$. Note-se que

$$R(x_*) \;=\; 0 \qquad \Longrightarrow \qquad S(x_*) \;=\; 0 \quad \text{e} \quad \nabla^2 f(x_*) \;=\; J(x_*)^\top J(x_*).$$

O método de Gauss-Newton para problemas de mínimos quadrados não lineares consiste na aplicação do método de Newton à resolução do sistema de equações não lineares $\nabla f(x) = 0$, ignorando a contribuição do termo $S(x)$ da matriz Hessiana $\nabla^2 f(x)$. Quando bem definido, este método gera a sucessão de pontos $\{x_k\}$ descrita pelo seguinte algoritmo.

Método de Gauss-Newton para Mínimos Quadrados Não Lineares
Escolher $x_0 \in \mathbb{R}^n$.
Para $k = 0, 1, 2, \ldots$

1. Resolver o sistema de equações lineares $J(x_k)^\top J(x_k) p_k = -J(x_k)^\top R(x_k)$.
2. Fazer $x_{k+1} = x_k + p_k$.

Para estar bem definido, o método de Gauss-Newton tem de gerar iteradas x_k para as quais $J(x_k)^\top J(x_k)$ seja não singular, ou seja, para as quais a característica de $J(x_k)$ seja igual a n. Uma propriedade interessante deste método é a geração de passos que são direcções de descida para a função f (ver exercício).

O método de Gauss-Newton apresenta uma taxa quadrática de convergência local para pontos x_* tais que $R(x_*) = 0$.

Teorema 1 *Seja $R : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ uma função vectorial definida num conjunto aberto D, com $m > n$. Suponha-se que R é continuamente diferenciável em D e que a sua matriz Jacobiana J é contínua à Lipschitz em D com constante γ. Seja $\alpha > 0$ um limite superior para a norma de J em D.*

Seja $x_* \in D$ um ponto tal que $R(x_*) = 0$ e para o qual $J(x_*)$ tem característica n. Seja, ainda, β um escalar positivo tal que $\| \left(J(x_*)^\top J(x_*)\right)^{-1} \| \leq \beta$.

Nestas condições, existe um escalar ϵ positivo tal que se

$$\|x_0 - x_*\| \;\leq\; \epsilon$$

então a sucessão $\{x_k\}$ gerada pelo método de Gauss-Newton está bem definida, converge para x_* e satisfaz

$$\|x_{k+1} - x_*\| \;\leq\; \alpha\beta\gamma \|x_k - x_*\|^2.$$

Demonstração. O pouco que temos a fazer é observar que, se a característica de $J(x_0)$ for n, então

$$\begin{aligned} x_1 - x_* &= x_0 - x_* - \left(J(x_0)^\top J(x_0)\right)^{-1} J(x_0)^\top R(x_0) \\ &= \left(J(x_0)^\top J(x_0)\right)^{-1} J(x_0)^\top \left[R(x_*) - R(x_0) - J(x_0)(x_* - x_0)\right]. \end{aligned}$$

A demonstração é, praticamente, idêntica à do método de Newton para sistemas de equações não lineares. O valor de ϵ é dado, neste caso, por

$$\epsilon \;=\; \left\{ r, \frac{1}{2\alpha\beta\gamma} \right\}.$$

A única diferença entre as duas demonstrações é a presença de α ao lado de $\beta\gamma$. ■

Quando o resíduo $R(x_*)$ não for nulo, a taxa passa a ser linear, desde que a norma Euclideana de $R(x_*)$ seja inferior ao inverso do produto $\beta\gamma$. Esta hipótese é demasiado forte, por estar dependente das propriedades da função R na solução x_*.

Teorema 2 *Seja $R : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ uma função vectorial definida num conjunto aberto D, com $m > n$. Suponha-se que R é continuamente diferenciável em D e que a sua matriz Jacobiana J é contínua à Lipschitz em D com constante γ. Seja $\alpha > 0$ um limite superior para a norma de J em D.*

Seja $x_* \in D$ um ponto para o qual $J(x_*)$ tem característica n. Seja β um escalar positivo tal que $\| \left(J(x_*)^\top J(x_*)\right)^{-1} \| \leq \beta$. Suponhamos que

$$\|R(x_*)\| \;<\; \frac{1}{\beta\gamma}.$$

Nestas condições, existe um escalar ϵ positivo tal que se

$$\|x_0 - x_*\| \;\leq\; \epsilon$$

então a sucessão $\{x_k\}$ gerada pelo método de Gauss-Newton está bem definida, converge para x_* e satisfaz

$$\|x_{k+1} - x_*\| \;\leq\; r\|x_k - x_*\|,$$

com

$$r \;=\; \frac{1}{2} + \frac{c\beta\gamma\|R(x_*)\|}{2} \;\in\; (0,1) \quad e \quad c \in \left(1, \frac{1}{\beta\gamma\|R(x_*)\|}\right).$$

A demonstração é omitida por trazer pouco de novo. Repare-se que a constante r aproxima-se de 1 (o que é mau) quando $\|R(x_*)\|$ se aproxima do inverso de $\beta\gamma$.

Corremos, em MATLAB, os métodos de Newton e de Gauss-Newton para a função dada em baixo nos exercícios. Fez-se $m = 3$, $\bar{y}_i = i$, $\bar{r}_1 = 2$ e $\bar{r}_2 = 4$. Testaram-se três valores diferentes para $\bar{r}_3$, nomeadamente, -8, -1 e 8. Começou-se com $x_0 = 1$ em todos os casos. O número de iterações necessário para reduzir o gradiente de f para menos de 10^{-10} foi o seguinte:

$\bar{r}_3$	Gauss-Newton	Newton	$f(x_*)$
-8	5	7	0
-1	34	10	6.976
8	não convergiu	12	41.145

Foi possível observar, também, um decréscimo aproximadamente quadrático para o erro no método de Gauss-Newton para o primeiro caso, em que o resíduo foi nulo, e um decréscimo linear para o segundo caso, em que o resíduo ainda foi relativamente pequeno.

Exercícios

1. Escreva $R(x)$, $J(x)$, $\nabla f(x)$ e $\nabla^2 f(x)$ para a função f dada por

$$f(x) \;=\; \frac{1}{2}\sum_{i=1}^{m} \left(e^{\bar{y}_i x} - \bar{r}_i\right)^2 .$$

2. Seja R continuamente diferenciável numa vizinhança de x_k. Prove que se a característica de $J(x_k)$ for igual a n então o passo de Gauss-Newton

$$p_k = -\left(J(x_k)^\top J(x_k)\right)^{-1} J(x_k)^\top R(x_k)$$

é uma direcção de descida para a função $f = R^\top R/2$.

3. Considere o problema de mínimos quadrados lineares em que $R(x) = Ax - b$ e $A \in \mathbb{R}^{m\times n}$ tem característica n, com $m > n$.

   (a) Identifique a solução única x_*.

   (b) Escreva o gradiente e a Hessiana de $f = R^\top R/2$ num ponto x_k.

   (c) Prove que o método de Gauss-Newton precisa de apenas uma iteração para convergir para x_*.

4. Considere, agora, o contexto da resolução numérica de um sistema de equações não lineares $F(x) = 0$, com $F : \mathbb{R}^n \to \mathbb{R}^m$, n e m inteiros positivos a satisfazer $n > m$ e F continuamente diferenciável (e com matriz Jacobiana dada por $J(x)$). (O objectivo deste exercício é abordar os sistemas de equações não lineares *indeterminados*, que se relacionam, de uma certa forma, com os problemas de mínimos quadrados não lineares.)

   (a) Determine o conjunto das raízes do sistema definido por

$$F(x) \;=\; \begin{bmatrix} x_1^2 + x_2 + x_3 \\ x_1^2 + x_3 \end{bmatrix}.$$

   (b) Escreva a matriz Jacobiana de $F(x)$ para o exemplo da alínea anterior e mostre que tem sempre característica igual a 2.

(c) No caso geral, mostre que a direcção $d(x) = -J(x)^\top \left(J(x)J(x)^\top\right)^{-1} F(x)$ é uma direcção de descida para a função $(1/2)\|F(x)\|^2$, se $F(x) \neq 0$.

Tome, agora, $F(x) = Ax - b$, em que $A \in \mathbb{R}^{m\times n}$ é uma matriz com característica m e $b \in \mathbb{R}^m$.

(d) Mostre que $x_0 + d(x_0)$ é solução de $F(x) = 0$.

(e) **(Difícil.)** Prove que quando $x_0 = 0$ a solução encontrada é aquela que tem menor norma entre todas as soluções de $F(x) = 0$.

Aula 9: Diferenciação Numérica

Calcular valores para o gradiente de $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ não é possível em várias situações práticas. É fácil imaginar que é este o caso quando a função f resulta, por exemplo, de uma experiência física.

Uma aproximação para o valor do vector gradiente $\nabla f(x)$ pode ser determinada calculando a função em $n+1$ pontos. Esta aproximação dá pelo nome de diferenças (finitas) *progressivas* (de primeira ordem) e é definida por

$$\frac{f(x+\epsilon e_i) - f(x)}{\epsilon} \simeq \frac{\partial f}{\partial x_i}(x), \qquad i = 1, \ldots, n.$$

Os vectores e_i, $i = 1, \ldots, n$, formam as colunas da matriz identidade de ordem n. A questão essencial, nesta aproximação, é a escolha do número positivo ϵ. (A aproximação pode ser feita por diferenças *regressivas* de primeira ordem: $[f(x) - f(x - \epsilon e_i)]/\epsilon, i = 1, \ldots, n.$)

A aplicação de um resultado provado anteriormente para a análise local do método de Newton, diz-nos que, se uma função f for continuamente diferenciável em D (aberto) e se ∇f for contínua à Lipschitz (com constante $\gamma > 0$) em D, então

$$\left|f(x+p) - f(x) - \nabla f(x)^\top p\right| \leq \frac{\gamma}{2}\|p\|^2,$$

quaisquer que sejam x e $x+p$ em D, com $[x, x+p] \subset D$.

Desta forma, ao escolhermos $p = \epsilon e_i$, constatamos que

$$\frac{\partial f}{\partial x_i}(x) = \frac{f(x+\epsilon e_i) - f(x)}{\epsilon} + \delta_\epsilon, \quad \text{com} \quad |\delta_\epsilon| \leq (\gamma/2)\epsilon,$$

para $i = 1, \ldots, n$. O erro na fórmula das diferenças progressivas varia linearmente com ϵ, aproximando-se de zero, em aritmética exacta, para valores de ϵ cada vez mais pequenos.

No entanto, em aritmética de vírgula flutuante, este tipo de expressões aproximadas está sujeito a erros de arredondamento. Sempre que uma operação aritmética entre dois números, representados num sistema de vírgula flutuante, é calculada em aritmética de vírgula flutuante ocorre um erro. Este erro, quando medido de forma relativa, é limitado por $\mathbf{u}$, a unidade de arredondamento. Em dupla precisão tem-se que $\mathbf{u}$ é aproximadamente 10^{-16}. (Descubra o valor de `eps` em MATLAB...)

Simplificando a apresentação, suponhamos que o erro relativo nos valores de $f(x)$ e de $f(x+\epsilon e_i)$, quando calculados computacionalmente, é limitado por $\mathbf{u}$. Assim sendo,

$$\begin{aligned} |\text{calc}[f(x)] - f(x)| &\leq \alpha_f \mathbf{u}, \\ |\text{calc}[f(x+\epsilon e_i)] - f(x+\epsilon e_i)| &\leq \alpha_f \mathbf{u}, \end{aligned}$$

onde α_f representa um limite superior para f em x e em $x+\epsilon e_i$. Se, agora, considerarmos a expressão

$$\frac{\text{calc}[f(x+\epsilon e_i)] - \text{calc}[f(x)]}{\epsilon},$$

que corresponderia ao cálculo computacional da diferença progressiva, viria que o erro entre esta expressão e $\partial f/\partial x_i(x)$ seria limitado por

$$(\gamma/2)\epsilon + \frac{2\mathbf{u}\alpha_f}{\epsilon}.$$

Gostaríamos de escolher ϵ de forma a que este limite superior para o erro fosse o menor possível. É fácil verificar que o mínimo ocorre quando

$$\epsilon^2 \;=\; \frac{4\alpha_f \mathbf{u}}{\gamma}.$$

Vamos considerar que $\alpha_f/\gamma \simeq 1$, o que corresponderia a dizer que a função f é bem escalonada, no sentido em que o *ratio* entre os seus valores e os das suas derivadas é aproximadamente igual a um. Desta forma, chegamos a uma escolha optimal para ϵ:

$$\epsilon \;=\; \sqrt{\mathbf{u}},$$

para a qual o erro em cima fica da ordem de $\sqrt{\mathbf{u}}$.

Uma fórmula mais precisa para a aproximação das derivadas, designada por diferenças (finitas) *centrais* (de primeira ordem), é dada por

$$\frac{f(x+\epsilon e_i) - f(x-\epsilon e_i)}{2\epsilon} \;\simeq\; \frac{\partial f}{\partial x_i}(x), \qquad i = 1, \ldots, n.$$

Esta fórmula é mais dispendiosa do que a das diferenças progressivas pois requer $2n+1$ avaliações da função f (em comparação com as $n+1$ avaliações das diferenças progressivas). No entanto, é possível provar, com um grau de suavidade a mais (e aplicando o resultado do exercício em baixo para $p = \epsilon e_i$ e $p = -\epsilon e_i$), que

$$\frac{\partial f}{\partial x_i}(x) \;=\; \frac{f(x+\epsilon e_i) - f(x-\epsilon e_i)}{2\epsilon} + \delta_\epsilon, \quad \text{com} \quad |\delta_\epsilon| \;\leq\; (\gamma/4)\epsilon^2.$$

O erro é da ordem de ϵ^2. Mostra-se, igualmente, que a escolha optimal para ϵ, em diferenças centrais, é da ordem de $\mathbf{u}^{\frac{1}{3}}$ e que, portanto, o erro, quando expresso em termos da unidade de arredondamento, é da ordem de $\mathbf{u}^{\frac{2}{3}}$. A precisão das diferenças centrais, quando posta em termos de $\mathbf{u}$, não impressiona assim tanto.

diferenças	erro em ϵ	erro em $\mathbf{u}$
progressivas	$\mathcal{O}(\epsilon)$	$\mathcal{O}(\mathbf{u}^{\frac{1}{2}})$
centrais	$\mathcal{O}(\epsilon^2)$	$\mathcal{O}(\mathbf{u}^{\frac{2}{3}})$

Exemplificamos esta tabela quando $\epsilon = 10^{-8}$ e $\mathbf{u} = 10^{-16}$. Neste caso, o erro em ϵ das diferenças centrais é da ordem de 10^{-16} e a redução, relativamente às diferenças

progressivas, é significativa. No entanto, em termos de $\mathbf{u}$, o erro passa, das diferenças progressivas para as centrais, de 10^{-8} para $10^{-32/3} \simeq 10^{-10.67}$, o que já não constitui um progresso tão assinalável.

As derivadas parciais de segunda ordem podem ser aproximadas através das derivadas de primeira ordem. Se estas últimas não estiverem disponíveis, a aproximação pode ser feita recorrendo a avaliações da própria função. Vamos ver o caso das diferenças *centrais* de segunda ordem. Aplicando o resultado do mesmo exercício, mas desta vez com $p = \epsilon e_i$, $p = \epsilon e_j$ e $p = \epsilon(e_i + e_j)$, concluímos que

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \;=\; \frac{f(x + \epsilon e_i + \epsilon e_j) - f(x + \epsilon e_i) - f(x + \epsilon e_j) + f(x)}{\epsilon^2} + \delta_\epsilon,$$

com um erro δ_ϵ da ordem de ϵ. Esta aproximação é pouco atraente em optimização (pois tem precisão baixa e requer muitas avaliações de f). É preferível, por exemplo, aproximar o gradiente por diferenças centrais de primeira ordem e aplicar um método de quasi-Newton.

Em outros contextos numéricos em que n é reduzido (*e.g.*, $n = 1, 2, 3$), esta aproximação para as derivadas parciais de segunda ordem ocupa lugar mais relevante. No caso $n = 1$, a fórmula das diferenças centrais de segunda ordem para aproximação de $f''(y)$ aparece escrita, frequentemente, na forma

$$\frac{f(y - \epsilon) - 2f(y) + f(y + \epsilon)}{\epsilon^2}.$$

Exercícios

1. Mostre que se uma função f for duas vezes continuamente diferenciável em D (aberto) e se $\nabla^2 f$ for contínua à Lipschitz (com constante $\gamma > 0$) em D então

$$\left| f(x + p) - f(x) - \nabla f(x)^\top p - \frac{1}{2} p^\top \nabla^2 f(x) p \right| \;\leq\; \frac{\gamma}{4} \|p\|^3,$$

   quaisquer que sejam x e $x + p$ em D com $[x, x + p] \subset D$.

2. Mostre que a escolha optimal para ϵ em diferenças centrais de primeira ordem é da ordem de $\mathbf{u}^{\frac{1}{3}}$ e que, portanto, o erro desta aproximação, quando expresso em termos da unidade de arredondamento, é da ordem de $\mathbf{u}^{\frac{2}{3}}$.

3. A aproximação de uma matriz Jacobiana, $m \times n$, por diferenciação numérica pode ser feita linha a linha, aproximando cada gradiente por diferenças progressivas com uma precisão da ordem de ϵ (como foi explicado na aula). Se o seu objectivo fosse, porém, aproximar o produto de uma matriz Jacobiana $J(x)$ por um vector p, como poderia levar a cabo esta aproximação de forma mais económica e com igual precisão (ordem de ϵ)?

4. Deduza o limite superior para o erro (em termos de ϵ) na aproximação das diferenças centrais de segunda ordem dada para $\partial^2 f/\partial x_i \partial x_j(x)$.

5. Conte o número de avaliações de f necessárias para aproximar a matriz Hessiana (que se assumiria simétrica e de ordem n) por diferenças centrais de segunda ordem.

6. Suponha que pretende aproximar uma matriz Hessiana $\nabla^2 f(x)$, $n \times n$, podendo calcular o gradiente ∇f em pontos à sua escolha. Como aproximaria a matriz Hessiana? Qual a ordem de precisão? Tem garantia de que a aproximação calculada constituiria uma matriz simétrica? E se não, como poderia calcular uma aproximação simétrica?

7. Considere uma expansão em h dada por

$$\mathcal{A}(h) \;=\; \alpha_0 + \alpha_1 h + \alpha_2 h^2 + \mathcal{R}_3(h),$$

em que $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ e $|\mathcal{R}_3(h)| \leq Ch^3$ (com $C > 0$ independente de h). O propósito deste exercício é motivar a técnica de *extrapolação de Richardson* (que também pode ser utilizada em integração numérica).

   (a) Qual é a ordem com que $\mathcal{A}(h)$ aproxima α_0?

   (b) Mostre, especificando α_0, α_1, α_2 e $\mathcal{R}_3(h)$, que pode descrever, desta forma, a fórmula de Taylor de ordem 2 com resto de Lagrange, para uma função f em torno de um dado ponto x_0.

   (c) Multiplique a expansão dada em cima por $\delta \in (0,1)$. Substitua, também na expansão original, h por δh. Substraia as duas igualdades assim obtidas membro a membro. Conclua que obteve uma nova aproximação para α_0 da forma

$$\mathcal{B}(h) \;=\; \frac{\mathcal{A}(\delta h) - \delta \mathcal{A}(h)}{1-\delta}.$$

   (d) Qual é a ordem com que $\mathcal{B}(h)$ aproxima α_0?

Aula 10: Conceitos Básicos sobre Integração Numérica

Seja f uma função real de variável real, integrável no intervalo $[a, b]$. Em diversas aplicações, pode ser extremamente difícil, ou mesmo impossível, calcular, exactamente, o valor do integral

$$I(f) \stackrel{\text{def}}{=} \int_a^b f(x)\,dx.$$

Vamos estudar fórmulas de integração numérica para aproximar o valor de $I(f)$, conhecidas por *fórmulas (ou regras) de quadratura*.

A ideia básica das fórmulas de quadratura é simples. Considera-se uma aproximação f_n da função f, em que $n+1$ designa o número de pontos em $[a, b]$ nos quais a função f é avaliada. Calcula-se, seguidamente, o integral de f_n em $[a, b]$:

$$I_n(f) \stackrel{\text{def}}{=} I(f_n) = \int_a^b f_n(x)\,dx \simeq I(f).$$

Se $f \in C^0[a, b]$, então o erro de quadratura, $E_n(f) = I(f) - I_n(f)$, satisfaz

$$|E_n(f)| \leq \int_a^b |f(x) - f_n(x)|\,dx \leq (b-a) \underbrace{\max_{x\in[a,b]} |f(x) - f_n(x)|}_{\| f - f_n \|_\infty}.$$

Se soubéssemos uma estimativa do tipo $\|f - f_n\|_\infty \leq \epsilon$, poderíamos concluir um limite superior para o erro, da forma $|E_n(f)| \leq \epsilon(b-a)$.

Só faz sentido considerar fórmulas de quadratura para as quais f_n seja facilmente integrável. As *fórmulas de quadratura interpolatória* utilizam polinómios $f_n = \Pi_n f$, os quais interpolam f num conjunto $\{x_0, x_1, \ldots, x_n\}$ de pontos distintos de $[a, b]$, conhecidos por nós da quadratura. Quando os nós apresentam um espaçamento h uniforme,

$$x_{i+1} - x_i = h, \quad i = 0, 1, \ldots, n-1,$$

as fórmulas de quadratura interpolatória são designadas por *fórmulas de Newton-Cotes*.

O polinómio interpolador pode ser escrito como uma combinação linear dos polinómios de Lagrange $\ell_0, \ell_1, \ldots, \ell_n$:

$$\Pi_n f(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x),$$

em que os coeficientes da combinação linear são os valores de f nos pontos de interpolação. Logo, o integral $I_n(f)$, para $f_n = \Pi_n f$, pode ser expresso na forma

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i),$$

com

$$\alpha_i \;=\; \int_a^b \ell_i(x)\,dx, \quad i = 0, 1, \ldots, n.$$

As fórmulas de Newton-Cotes dividem-se em *abertas* e *fechadas*. Neste curso vamos considerar, essencialmente, as fórmulas *fechadas*, para as quais $x_0 = a$, $x_n = b$ e $h = (b-a)/n$. As fórmulas *abertas*, onde $x_0 = a + h$ e $x_n = b - h$, podem dar origem a coeficientes α_i negativos, o que as torna mais vulneráveis a erros de arredondamento. Nas fórmulas fechadas tem-se, obrigatoriamente, que $n \geq 1$. É possível, porém, conceber uma fórmula aberta com apenas um ponto ($n = 0$); ver exercício.

Uma propriedade relevante das fórmulas de Newton-Cotes é que os coeficientes α_i dependem de n e de i, mas não dependem do intervalo $[a, b]$, ou seja, não dependem de a e de b. Consequentemente, estes coeficientes podem ser tabelados *a priori*. Vejamos esta propriedade para o caso das fórmulas fechadas. O valor do polinómio de Lagrange ℓ_i em pontos x distintos dos nós é dado por

$$\ell_i(x) \;\stackrel{\text{def}}{=}\; \Pi_{k=0,k\neq i}^{n} \frac{x - x_k}{x_i - x_k} \;=\; \Pi_{k=0,k\neq i}^{n} \frac{(a+th) - (a+kh)}{(a+ih) - (a+kh)}, \quad i = 0, 1, \ldots, n,$$

onde, na última igualdade, fizemos a mudança de variável $x = a + th$, com $t \in [0, n]$. Constata-se que h é cancelável nesta fracção, e define-se

$$\ell_i(x) \;=\; \ell_i(a+th) \;=\; \Pi_{k=0,k\neq i}^{n} \frac{t-k}{i-k} \;\stackrel{\text{def}}{=}\; \phi_i(t), \quad i = 0, 1, \ldots, n.$$

Usando esta expressão para os polinómios de Lagrange, os coeficientes de integração são reescritos como

$$\alpha_i \;=\; \int_a^b \ell_i(x)\,dx \;=\; \int_0^n \left[\frac{d}{dt}(a+th)\right] \ell_i(a+th)\,dt \;=\; h \underbrace{\int_0^n \phi_i(t)\,dt}_{\| \; w_i^n}, \quad i = 0, 1, \ldots, n.$$

O integral de $\Pi_n f$ fica, então, igual a

$$I_n(f) \;=\; h \sum_{i=0}^{n} w_i^n f(x_i),$$

com

$$w_i^n \;=\; \int_0^n \phi_i(t)\,dt, \quad i = 0, 1, \ldots, n.$$

Os pesos $w_0^n, w_1^n, \ldots, w_n^n$ são independentes de a e de b. A sua dependência de n e de i é exemplificada na tabela seguinte.

$i\backslash n$	1	2	3	4	5	6
0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{3}{8}$	$\frac{14}{45}$	$\frac{95}{288}$	$\frac{41}{140}$
1	–	$\frac{4}{3}$	$\frac{9}{8}$	$\frac{64}{45}$	$\frac{375}{288}$	$\frac{216}{140}$
2	–	–	–	$\frac{24}{45}$	$\frac{250}{288}$	$\frac{27}{140}$
3	–	–	–	–	–	$\frac{272}{140}$

Nesta tabela listámos apenas os pesos até $[n/2]$, uma vez que os pesos são simétricos, ou seja,

$$w_i^n \;=\; w_{n-i}^n, \quad i = 0, 1, \ldots, n.$$

A simetria dos pesos w_i^n provém, obviamente, da simetria das funções ϕ_i.

Assim, no caso $n = 1$, temos que

$$w_0^1 \;=\; w_1^1 \;=\; \frac{1}{2}.$$

No caso $n = 2$, obtemos

$$w_0^2 \;=\; \frac{1}{3}, \quad w_1^2 \;=\; \frac{4}{3} \quad \text{e} \quad w_2^2 \;=\; \frac{1}{3}.$$

Nota final: Para aliviar a notação escrevemos $\ell_i(x)$, α_i e $\phi_i(t)$, quando, rigorosamente, deveríamos ter explicitado a dependência de n e ter escrito $\ell_i^n(x)$, α_i^n e $\phi_i^n(t)$.

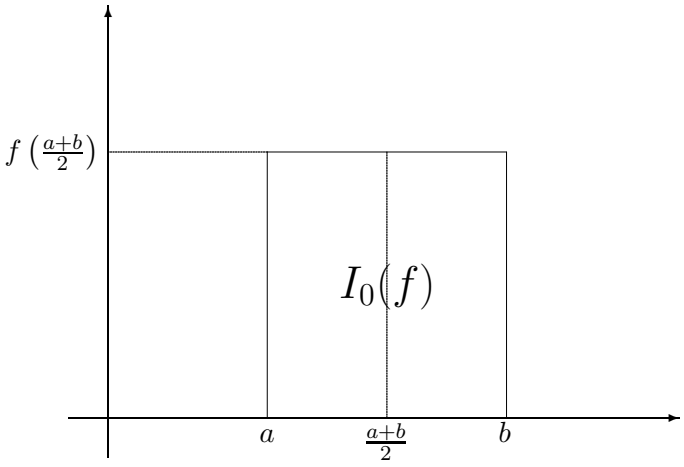
Exercícios

1. A fórmula aberta de Newton-Cotes para o caso $n = 0$ é designada por fórmula do ponto médio.

   (a) Indique os valores de h e de x_0.

   (b) Mostre que a fórmula do ponto médio é da forma

   $$I_0(f) \;=\; (b-a) f\left(\frac{a+b}{2}\right).$$



2. Sejam $a = -1$ e $b = 1$. Descubra qual a fórmula de Newton-Cotes fechada que tem a forma
$$\frac{1}{4}\left(f(-1) + 3f\left(-\frac{1}{3}\right) + 3f\left(\frac{1}{3}\right) + f(1)\right),$$
indicando o valor de n e identificando os respectivos pesos.

3. Considere o cálculo aproximado do integral $I(f) = \int_a^b f(x)\,dx$ através de uma aproximação para f em $[a, b]$ definida por $g(x) = f(y) + (x - y)f'(y)$, em que $y \in (a, b)$ e se considera que f tem derivada em y.

   (a) Deduza a fórmula para $I(g)$ dada por:
   $$I(g) \;=\; (b-a)f(y) + \frac{f'(y)}{2}(b-a)(a+b-2y).$$
   (De uma forma geral não se trata de uma fórmula de quadratura interpolatória.)

   (b) Que fórmula obtém quando $y = (a + b)/2$?

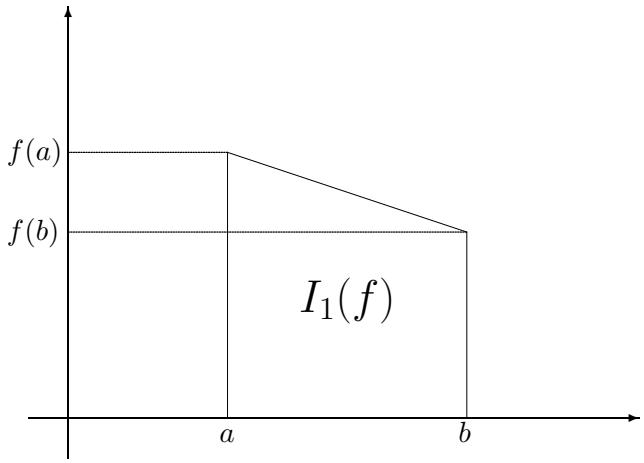
   (c) Deduza, na forma de um integral, uma expressão para o erro $I(f - g)$, assumindo que $f \in C^2[a, b]$.

   (d) Mostre que o erro é da ordem de $(b - a)^3$ e apresente um limite superior para a respectiva constante.

Aula 11: Integração Numérica – Fórmulas Trapezoidal e de Simpson

A fórmula de Newton-Cotes fechada mais simples que existe é a *fórmula trapezoidal (ou dos trapézios)*. Esta fórmula de quadratura, definida seleccionando $n = 1$, consiste em integrar o polinómio interpolador $\Pi_1 f$, de grau 1, nos nós de interpolação $x_0 = a$ e $x_1 = b$. Relembrando que $w_0^1 = w_1^1 = 1/2$ e $h = b - a$ quando $n = 1$, escrevemos a fórmula trapezoidal na forma

$$I_1(f) \;=\; \frac{b-a}{2}\left[f(a) + f(b)\right].$$



O erro desta fórmula é estimado integrando o erro do polinómio interpolador $\Pi_1 f$ enquanto aproximação para f. Se $f \in C^2[a,b]$ então

$$E_1(f) \;=\; I(f) - I_1(f) \;=\; \int_a^b [f(x) - \Pi_1 f(x)]\,dx \;=\; \int_a^b f[a,b,x]w_1(x)\,dx,$$

com $w_1(x) = (x-a)(x-b)$ e $f[a,b,x]$ a diferença dividida nos pontos a, b e x. Como $f[a,b,x]$ é uma função contínua em x e $w_1(x)$ é integrável em $[a,b]$ e não muda de sinal neste intervalo, sabe-se, por um teorema do valor médio integral, que

$$\int_a^b f[a,b,x]w_1(x)\,dx \;=\; f[a,b,\eta]\int_a^b w_1(x)\,dx \;=\; \frac{1}{2}f''(\xi)\int_a^b w_1(x)\,dx,$$

com $\eta, \xi \in (a,b)$. A última igualdade resulta das propriedades das diferenças divididas. Logo,

$$E_1(f) \;=\; \frac{1}{2}f''(\xi)\int_a^b (x-a)(x-b)\,dx \;=\; -\frac{1}{12}f''(\xi)(b-a)^3 \;=\; -\frac{f''(\xi)}{12}h^3.$$

A fórmula trapezoidal tem uma ordem de precisão 3. Diz-se que uma fórmula de quadratura de Newton-Cotes tem *ordem de precisão* p se $|I(f) - I_n(f)| \leq C\, h^p$, em que C é uma constante que não depende de n e de h. No caso da fórmula trapezoidal, esta constante é $C = \max_{x\in[a,b]} |f''(x)|/12$.

A fórmula de quadratura de Newton-Cotes fechada quando $n = 2$ é conhecida por *fórmula de Simpson ou de Cavalieri-Simpson* e resulta da integração do polinómio interpolador $\Pi_2 f$ de grau 2 nos nós de interpolação $x_0 = a$, $x_1 = (a+b)/2$ e $x_2 = b$. Os seus pesos são $w_0^2 = w_2^2 = 1/3$ e $w_1^2 = 4/3$. Neste caso, h é igual a $(b-a)/2$. Desta forma, a fórmula de Simpson é dada por

$$I_2(f) \;=\; \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Quando $f \in C^4[a,b]$, prova-se que o erro $E_2(f)$ pode ser expresso na forma

$$E_2(f) \;=\; -\frac{f^{(4)}(\xi)}{90}h^5,$$

com $\xi \in (a,b)$. (A demonstração é relegada para os exercícios.) A ordem de precisão da fórmula de Simpson é, desta forma, igual a 5. A constante C do erro é igual a $\max_{x\in[a,b]} |f^{(4)}(x)|/90$.

Caracterizámos estas duas fórmulas de acordo com a sua ordem de precisão. Uma outra propriedade das fórmulas de quadratura interpolatórias é o seu grau de exactidão. Diz-se que uma fórmula de quadratura interpolatória (e, consequentemente, uma fórmula de Newton-Cotes) tem um *grau de exactidão* $r \geq 0$ se r for o maior inteiro para o qual $I_n(f) \;=\; I(f)$ para todos os polinómios f de grau inferior ou igual a r.

É fácil constatar, pelas expressões dos erros, que a fórmula trapezoidal tem grau de exactidão 1 e que a fórmula de Simpson tem grau de exactidão 3.

As fórmulas de quadratura Gaussiana (a estudar mais adiante neste curso) atingem um grau de exactidão de $2n+1$. Este é, aliás, o grau máximo que uma fórmula de quadratura interpolatória pode alcançar (ver exercício).

Qualquer fórmula de quadratura interpolatória baseada em $n+1$ pontos tem um grau de exactidão nunca inferior a n. De facto, se f for um polinómio de grau inferior ou igual a n, então $\Pi_n f = f$ e, como é óbvio, $I_n(f) \stackrel{\text{def}}{=} I(\Pi_n f) = I(f)$.

A afirmação recíproca também é verdadeira, assumindo que os $n+1$ pontos são distintos: qualquer fórmula de quadratura que tenha grau de exactidão não inferior a n é, obrigatoriamente, do tipo interpolador. Seja

$$\bar{I}_n(f) \;=\; \sum_{i=0}^{n} \beta_i f(x_i)$$

uma fórmula de quadratura. Pretendemos provar que estes coeficientes β_i coincidem com os $\alpha_i = \int_a^b \ell_i(x)\,dx$, para $i = 0, 1, \ldots, n$, para, assim, concluir que a fórmula de quadratura é interpolatória. Sabe-se, por hipótese, que $\bar{I}_n(x^i) = I(x^i)$, $i = 0, 1, \ldots, n$, o que, na forma matricial, se escreve como

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} b-a \\ (b^2-a^2)/2 \\ \vdots \\ (b^{n+1}-a^{n+1})/(n+1) \end{bmatrix}.$$

A matriz (de Vandermonde) deste sistema é não singular pelo facto dos $n+1$ pontos serem distintos. Logo, este sistema só admite uma solução. Mas $[\,\alpha_0\ \alpha_1\ \cdots\ \alpha_n\,]^\top$ é solução do sistema, pois $I_n(x^i) = I(x^i)$, $i = 0, 1, \ldots, n$. Assim sendo, os α's e os β's coincidem. Resumimos estas duas implicações no enunciado seguinte.

Teorema 1 *Uma fórmula de quadratura, baseada em $n+1$ pontos distintos, tem grau de exactidão superior ou igual a n se e só se for do tipo interpolatório.*

Exercícios

1. Com este exercício pretende-se demonstrar a expressão dada para o erro na fórmula de Simpson. Seja $w_2(x) = (x-a)(x-(a+b)/2)(x-b)$.

   (a) Mostre que $w_2(x)$ muda de sinal em $[a, b]$. Conclua que não se pode aplicar ao erro
   $$E_2(f) = \int_a^b f[a, (a+b)/2, b, x] w_2(x)\,dx$$
   o teorema do valor médio integral que se aplicou no caso da fórmula trapezoidal.

   (b) Mostre que $\int_a^b w_2(x) = 0$. Utilizando este facto e
   $$f[a, (a+b)/2, b, x] = f[a, (a+b)/2, b, x_3] + f[a, (a+b)/2, b, x_3, x](x - x_3),$$
   conclua que
   $$E_2(f) = \int_a^b f[a, (a+b)/2, b, x_3, x] w_3(x)\,dx,$$
   com $w_3(x) = w_2(x)(x - x_3)$.

   (c) Faça, agora, $x_3 = x_1 = (a+b)/2$. Mostre que $w_3(x)$ assim definido não muda de sinal em $[a, b]$. Termine a demonstração aplicando o teorema do valor médio integral.

2. Considere a fórmula do ponto médio dada no exercício da aula anterior.

   (a) Prove que se $f \in C^2[a,b]$ então o erro da fórmula do ponto médio pode ser expresso na forma

   $$E_0(f) \;=\; \frac{h^3}{3} f''(\xi) \quad \text{com} \quad \xi \in (a,b).$$

   Sugestão: utilize um raciocínio semelhante ao aplicado na análise do erro da fórmula de Simpson.

   (b) Qual é a ordem de precisão da fórmula do ponto médio? E o seu grau de exactidão?

3. Considere $w_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$. Mostre que $I_n(w_n^2) = 0$ e que $\int_a^b w_n(x)^2\,dx > 0$. Retire, daqui, a conclusão de que $2n + 1$ é o grau máximo que uma fórmula de quadratura interpolatória pode atingir.

4. Suponha que o cálculo do integral $I(f) = \int_0^1 xf(x)\,dx$ é feito através da fórmula de quadratura

   $$I_2(f) \;=\; \alpha_0 f(0) + \alpha_1 f(1/2) + \alpha_2 f(1).$$

   (a) Determine α_0, α_1 e α_2 de forma a que a fórmula de quadratura tenha grau de exactidão não inferior a 2.

   (b) Seria possível encontrar valores para α_0, α_1 e α_2 tais que o grau de exactidão fosse igual a 3?

Aula 12: Integração Numérica – Fórmulas Compostas

O erro de uma fórmula de quadratura pode não ser pequeno se a amplitude do intervalo $[a, b]$ não for suficientemente pequena. Uma alternativa para esta situação passa pela integração do polinómio interpolador composto, construído em m subintervalos de $[a, b]$ de amplitude

$$H = \frac{b-a}{m} \qquad \text{com} \qquad m \geq 1.$$

As fórmulas compostas são construídas aplicando fórmulas de quadratura nos m subintervalos.

No caso da fórmula trapezoidal composta, os extremos dos subintervalos são os pontos

$$y_j = a + jH, \quad j = 0, \ldots, m.$$

Repare que $y_0 = a$ e $y_m = b$. Em cada subintervalo $[y_j, y_{j+1}]$, $j = 0, \ldots, m-1$, é aplicada uma fórmula trapezoidal. A fórmula trapezoidal composta é, então, definida pela expressão

$$I_{1,m}(f) = \frac{H}{2} \sum_{j=0}^{m-1} \left[f(y_j) + f(y_{j+1}) \right].$$

Observa-se, facilmente, que, com a excepção das extremidades $a = y_0$ e $b = y_m$, todos os outros extremos de subintervalos aparecem duas vezes nesta expressão. Deste modo,

$$I_{1,m}(f) = H \left[\frac{1}{2} f(y_0) + f(y_1) + \cdots + f(y_{m-1}) + \frac{1}{2} f(y_m) \right].$$

No que se segue, vamos partir do princípio de que $f \in C^2[a, b]$. O erro desta fórmula composta é a soma dos erros das m fórmulas trapezoidais:

$$E_{1,m}(f) = -\sum_{j=0}^{m-1} \frac{f''(\xi_j)}{12} H^3,$$

com $\xi_j \in (y_j, y_{j+1})$, $j = 0, \ldots, m-1$.

A aplicação do teorema do valor médio discreto com $g = f''$ (ver o final desta aula), permite-nos escrever

$$E_{1,m}(f) = -\left(\sum_{j=0}^{m-1} f''(\xi_j) \right) \frac{H^3}{12} = -\left(m f''(\eta) \right) \frac{H^3}{12} = -\frac{b-a}{12} f''(\eta) H^2,$$

com $\eta \in (a, b)$.

No caso da fórmula de Simpson composta, é preciso considerar dois pontos em cada um dos m subintervalos:

$$y_j = a + jH/2, \quad j = 0, \ldots, 2m.$$

Neste caso, temos que $y_0 = a$ e $y_{2m} = b$. É aplicada a fórmula de Simpson em cada um dos m subintervalos $[y_j, y_{j+2}]$, utilizando o ponto médio y_{j+1}, com j a variar de 0 até $2m-2$. Os pontos médios aparecerão, assim, apenas uma vez, enquanto que as extremidades dos intervalos contarão duas vezes. Somando as parcelas resultantes da aplicação das m fórmulas de Simpson e recorrendo a esta última observação, obtém-se a seguinte expressão para a fórmula de Simpson composta:

$$I_{2,m}(f) \;=\; \frac{H}{6}\left[f(y_0) + 2\sum_{r=1}^{m-1} f(y_{2r}) + 4\sum_{s=0}^{m-1} f(y_{2s+1}) + f(y_{2m})\right].$$

Se aplicássemos uma argumentação idêntica à utilizada para a fórmula trapezoidal composta, chegaríamos à conclusão de que o erro da fórmula de Simpson composta poderia ser expresso na forma

$$E_{2,m}(f) \;=\; -\frac{b-a}{2}\frac{f^{(4)}(\eta)(H/2)^4}{90} \;=\; -\frac{b-a}{2880}f^{(4)}(\eta)H^4.$$

Para o efeito seria preciso assumir que $f \in C^4[a,b]$. O escalar $\eta \in (a,b)$ resultaria da aplicação do teorema do valor médio discreto com $g = f^{(4)}$.

Para ilustrar o desempenho numérico destas duas fórmulas de quadratura compostas, escolhemos a função $f(x) = x^5$, que integrámos numericamente de 0 até 1. Os resultados, obtidos em MATLAB, são descritos em baixo em termos dos erros ocorridos.

m	$\lvert E_{1,m}(f)\rvert$	$\lvert E_{2,m}(f)\rvert$
10	4.16e-003	2.08e-006
100	4.17e-005	2.08e-010
1000	4.17e-007	2.10e-014
10000	4.17e-009	1.11e-016

Como seria de esperar, a fórmula de Simpson composta apresenta melhores resultados do que a fórmula trapezoidal composta.

Terminamos a aula com o enunciado e a demonstração do teorema do valor médio discreto.

Teorema 1 *Seja g uma função contínua em $[a,b]$ e seja $\{y_0, y_1, \ldots, y_q\}$ um conjunto de pontos em $[a,b]$.*

Seja, ainda, $\delta_0, \delta_1, \ldots, \delta_q$ uma sequência de escalares reais todos com o mesmo sinal.

Nestas condições, existe um escalar real $\eta \in [a,b]$ tal que

$$\sum_{j=0}^{q} \delta_j g(y_j) \;=\; g(\eta)\sum_{j=0}^{q} \delta_j.$$

Demonstração. Façamos a demonstração assumindo que todos os escalares δ_j são positivos. Considere-se a função auxiliar h definida por

$$h(x) \;=\; g(x)\sum_{j=0}^{q}\delta_j.$$

Garantida a existência de x_{min} e x_{max} em $[a,b]$ tais que

$$g(x_{min}) \;=\; \min_{x\in[a,b]} g(x) \quad \text{e} \quad g(x_{max}) \;=\; \max_{x\in[a,b]} g(x),$$

tem-se que

$$\underbrace{g(x_{min})\sum_{j=0}^{q}\delta_j}_{\substack{\shortparallel\\ h(x_{min})}} \;\leq\; \sum_{j=0}^{q}\delta_j g(y_j) \;\leq\; \underbrace{g(x_{max})\sum_{j=0}^{q}\delta_j}_{\substack{\shortparallel\\ h(x_{max})}}.$$

Como a função auxiliar h é contínua em $[a,b]$, o teorema de Bolzano-Cauchy aplicado a esta função prova-nos a existência de um $\eta \in [x_{min}, x_{max}] \subset [a,b]$ tal que

$$h(\eta) \;=\; \sum_{j=0}^{q}\delta_j g(y_j),$$

o que conclui a demonstração. ■

Exercícios

1. Escreva a fórmula do ponto médio composta e desenvolva uma expressão para o seu erro.

2. Calcule um valor para H para o qual exista a garantia de que o integral

$$\int_0^1 \operatorname{sen}(x^2)\,dx$$

é calculado com um erro inferior a 10^{-5} usando a:

   (a) Fórmula trapezoidal composta.

   (b) Fórmula de Simpson composta.

3. Demonstre o teorema do valor médio discreto no caso dos escalares δ_j serem todos negativos.

Aula 13: Conceitos Básicos sobre Aproximação de Funções

São frequentes as ocorrências em problemas de ciência e engenharia em que a informação sobre funções relevantes é parcial ou mesmo escassa. O valor de uma função pode ser conhecido, por exemplo, apenas num conjunto finito de pontos, conjunto de pontos esse que pode ser, ou não ser, especificado *a priori*.

Vários procedimentos em análise numérica ou em teoria da aproximação calculam, implicitamente ou explicitamente, uma função com características especiais, que pretende aproximar, de uma certa maneira, a função cuja expressão analítica, ou fórmula computacional, é desconhecida.

Suponha-se que é possível inferir, a partir do conhecimento de algumas propriedades estruturais do problema em causa, que a função f a aproximar está num determinado espaço vectorial real de funções E. Uma das técnicas mais populares em teoria da aproximação consiste em procurar uma função f_n, num subespaço de dimensão finita $S \subset E$, que se ajusta a f de acordo com algum critério predeterminado.

Seja $\{\psi_0, \psi_1, \ldots, \psi_n\}$ uma base de S. Na prática, o que se pretende calcular são escalares reais $\alpha_0, \alpha_1, \ldots, \alpha_n$ tais que

$$f_n \;=\; \alpha_0\psi_0 + \alpha_1\psi_1 + \cdots + \alpha_n\psi_n.$$

Vamos abordar, nesta aula, duas das técnicas mais utilizadas para calcular os escalares desta combinação linear. Outro método que se pode enquadrar desta forma é o da aproximação por funções *spline*.

Lembramos que nem sempre se aproxima f recorrendo a uma forma linear. Vimos, na aula sobre os problemas de mínimos quadrados não lineares, uma parameterização em que a dependência de f_n dos escalares $\alpha_0, \alpha_1, \ldots, \alpha_n$ era não linear.

Uma das abordagens pressupõe a existência de um produto interno $\langle \cdot, \cdot \rangle$ no espaço vectorial E. Este produto interno equipa o espaço E com uma norma $\|\cdot\|$, o que permite definir distância entre dois elementos de E. Deste modo, faz sentido escolher f_n como sendo a projecção ortogonal de f sobre S, denominada por f_n^*, por esta ser a que está menos distante de f de entre todas as funções de S (ver Teorema 1 desta aula). Por definição de projecção ortogonal, $f_n^* - f$ tem de ser ortogonal a todos os elementos de S e, em particular, aos $n+1$ elementos da base. Se tomarmos o produto interno entre $f_n - f$ e as $n+1$ funções $\psi_0, \psi_1, \ldots, \psi_n$, obtemos as igualdades

$$\begin{aligned}
\alpha_0\langle\psi_0,\psi_0\rangle + \alpha_1\langle\psi_1,\psi_0\rangle + \cdots + \alpha_n\langle\psi_n,\psi_0\rangle &= \langle f,\psi_0\rangle,\\
\alpha_0\langle\psi_0,\psi_1\rangle + \alpha_1\langle\psi_1,\psi_1\rangle + \cdots + \alpha_n\langle\psi_n,\psi_1\rangle &= \langle f,\psi_1\rangle,\\
&\vdots\\
\alpha_0\langle\psi_0,\psi_n\rangle + \alpha_1\langle\psi_1,\psi_n\rangle + \cdots + \alpha_n\langle\psi_n,\psi_n\rangle &= \langle f,\psi_n\rangle.
\end{aligned}$$

Este sistema de equações lineares admite uma solução única $\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*$. A matriz

$$\begin{bmatrix} \langle \psi_0, \psi_0 \rangle & \cdots & \langle \psi_n, \psi_0 \rangle \\ \vdots & \ddots & \vdots \\ \langle \psi_0, \psi_n \rangle & \cdots & \langle \psi_n, \psi_n \rangle \end{bmatrix},$$

conhecida por matriz de Gram, é não singular e simétrica (ver exercício).

A projecção ortogonal $f_n^* = \alpha_0^* \psi_0 + \alpha_1^* \psi_1 + \cdots + \alpha_n^* \psi_n$ de f sobre S é, de entre todas as funções em S, a que está mais perto de f. Enunciamos, seguidamente, este conhecido facto da Álgebra Linear.

Teorema 1 *Seja $\{\psi_0, \psi_1, \ldots, \psi_n\}$ uma base de S. Então, a função*

$$f_n^* = \sum_{k=0}^{n} \alpha_k^* \psi_k,$$

em que os coeficientes $\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*$ resolvem o sistema de equações lineares anterior, satisfaz

$$\|f_n^* - f\| = \min_{\psi \in S} \|\psi - f\|.$$

Observação: Este resultado permanece verdadeiro se o produto interno $\langle \cdot, \cdot \rangle$ for substituído por um semiproduto interno $\ll \cdot, \cdot \gg$ (com a respectiva norma $\| \cdot \|$ a dar lugar à seminorma $||| \cdot |||$). Um semiproduto interno obedece a todos os axiomas que definem um produto interno menos um: $\ll z, z \gg (= |||z|||^2)$ pode ser nulo sem que z seja zero.

A escolha da base de S é feita, geralmente, por forma a criar um padrão de esparsidade na matriz de Gram (diagonal, tridiagonal, *etc.*), reduzindo o custo computacional associado à resolução do sistema.

Os coeficientes $\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*$ são determinados com reduzido custo computacional se a base $\{\psi_0, \psi_1, \ldots, \psi_n\}$ for ortogonal. Neste caso, tem-se que

$$\alpha_k^* = \frac{\langle f, \psi_k \rangle}{\|\psi_k\|^2}, \quad k = 0, 1, \ldots, n.$$

Se a base for ortonormada, então $\alpha_k^* = \langle f, \psi_k \rangle$, $k = 0, 1, \ldots, n$.

A ortogonalidade das funções $\psi_0, \psi_1, \ldots, \psi_n$ é conveniente em diversas situações. Entre as funções ortogonais mais utilizadas encontram-se os polinómios ortogonais e os polinómios trigonométricos de Fourier, que estudaremos mais adiante.

No entanto, nem sempre esta ortogonalidade é desejável, como acontece quando se pretende descrever f' recorrendo a $\psi_0', \psi_1', \ldots, \psi_n'$ (assumindo-se bem definidas as respectivas derivadas). A ortogonalidade em $\{\psi_0, \psi_1, \ldots, \psi_n\}$ não implica, geralmente, a ortogonalidade em $\{\psi_0', \psi_1', \ldots, \psi_n'\}$. Observaremos uma situação deste tipo quando estudarmos o método dos elementos finitos para um problema de condição de fronteira.

A outra abordagem que aqui apresentamos aplica-se ao caso em que são conhecidos os valores de f em m pontos:

$$(x_1, f(x_1)), \ldots, (x_m, f(x_m)).$$

Nesta situação, é natural escolher um dado conjunto de pesos $w_1, \ldots, w_m$ positivos e, depois, procurar escalares $\alpha_0, \alpha_1, \ldots, \alpha_n$ de forma a minimizar

$$\sum_{i=1}^{m} w_i [f_n(x_i) - f(x_i)]^2 \;=\; \sum_{i=1}^{m} w_i \left[\left(\sum_{k=0}^{n} \alpha_k \psi_k(x_i) \right) - f(x_i) \right]^2 .$$

É fácil verificar que estamos na presença de um problema de mínimos quadrados lineares, que se pode escrever na forma

$$\min_{y \in \mathbb{R}^{n+1}} \left\| W^{\frac{1}{2}} B y - W^{\frac{1}{2}} b \right\|^2 ,$$

com W uma matriz diagonal $m \times m$ de elementos diagonais $w_1, \ldots, w_m$,

$$B \;=\; \begin{bmatrix} \psi_0(x_1) & \cdots & \psi_n(x_1) \\ \vdots & & \vdots \\ \psi_0(x_m) & \cdots & \psi_n(x_m) \end{bmatrix}, \quad y \;=\; \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{bmatrix} \quad \text{e} \quad b \;=\; \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}.$$

A norma $\|\cdot\|$, aqui utilizada, é a norma Euclideana em $\mathbb{R}^m$. Esta forma de aproximar f é conhecida por aproximação *discreta* no sentido dos mínimos quadrados.

Exercícios

1. Seja $\{\psi_0, \psi_1, \ldots, \psi_n\}$ uma base do subespaço S de um espaço vectorial real E. Prove que a matriz de Gram é simétrica e não singular.

2. Seja $S = \mathbb{P}_n$ o espaço dos polinómios de grau inferior ou igual a n. Dados os pontos $x_1, \ldots, x_m$ e os pesos positivos $w_1, \ldots, w_m$, considere

$$|||\, p \,||| \;=\; \left(\sum_{i=1}^{m} w_i [p(x_i)]^2 \right)^{\frac{1}{2}} .$$

   (a) Mostre, recorrendo ao produto interno usual em $\mathbb{R}^m$, que $|||\cdot|||$ é uma seminorma em $\mathbb{P}_n$, e que se trata de uma norma quando $m > n$.

   (b) Mostre que a seminorma $|||\cdot|||$ é essencialmente estrita, ou seja, que $|||\, p_1 + p_2 \,||| = |||\, p_1 \,||| + |||\, p_2 \,|||$ implica a existência de escalares β e γ tais que $\beta p_1(x_i) + \gamma p_2(x_i) = 0$, $i = 1, \ldots, m$ (assumindo $|||p_1|||$ e $|||p_2|||$ diferentes de zero).

(c) Faça, agora, $\psi_k(x) = x^k$, $k = 0, 1, \ldots, m-1$. Escreva, por extenso, a matriz B do problema da aproximação discreta no sentido dos mínimos quadrados. Observe que obteve a transposta da matriz de Vandermonde de ordem m.

3. Escreva as equações normais associadas ao problema de mínimos quadrados

$$\min_{y \in \mathbb{R}^{n+1}} \left\| W^{\frac{1}{2}} B y - W^{\frac{1}{2}} b \right\|^2 .$$

4. Seja f uma função contínua em $[0, 1]$. Considere a função real de $n+1$ variáveis reais dada por

$$g(\alpha_0, \alpha_1, \ldots, \alpha_n) \;=\; \int_0^1 \left(f(x) - \sum_{k=0}^{n} \alpha_k x^k \right)^2 dx,$$

que descreve uma forma de medir o erro da aproximação de f, em $[0, 1]$, por um polinómio de grau n.

(a) Determine escalares $\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*$ de forma a que $\nabla g(\alpha_0^*, \alpha_1^*, \ldots, \alpha_n^*) = 0$. (Estes escalares são a solução de um sistema de equações lineares. Não é necessário resolver o sistema.)

(b) O que teria de provar para afirmar que $[\, \alpha_0^* \; \alpha_1^* \; \ldots \; \alpha_n^* \,]^\top$ é o (único) minimizante da função g?

(c) A matriz deste sistema de equações lineares, conhecida por matriz de Hilbert, é extremamente mal-condicionada. Calcule, em MATLAB e *numa só instrução*, o número de condição desta matriz quando $n = 9$. Verifique, em MATLAB, que os seus valores próprios são todos positivos.

Aula 14: Polinómios Ortogonais – Legendre e Chebyshev

A escolha de bases ortogonais para S proporciona matrizes de Gram diagonais. Se as bases forem ortonormadas, a matriz de Gram coincide com a matriz identidade, sendo a mais bem condicionada de entre todas as escolhas possíveis. Entre as famílias de funções ortogonais de maior abrangência, encontram-se os polinómios ortogonais.

Para introduzir os polinómios ortogonais, vamos considerar o espaço de funções cujo quadrado é integrável, de forma ponderada, no intervalo $(-1, 1)$. Para isso assumimos a existência de uma função de ponderação (ou uma função-peso) $w(x)$, positiva e integrável em $(-1, 1)$. O nosso espaço de funções E vai ser, neste caso, definido por

$$L^2_w(-1,1) \;=\; \left\{ f : [-1,1] \longrightarrow \mathbb{R} : \int_{-1}^{1} f^2(x)\, w(x)\, dx \;<\; +\infty \right\}.$$

Neste espaço, está definido o produto interno[1]

$$\langle f, g \rangle_w \;=\; \int_{-1}^{1} f(x) g(x)\, w(x)\, dx$$

e a norma a ele associada

$$\|f\|_w \;=\; \left(\int_{-1}^{1} f(x)^2\, w(x)\, dx \right)^{\frac{1}{2}}.$$

Seja S o subespaço de E constituído por todos os polinómios de grau inferior ou igual a n. Este subespaço, denominado por $\mathbb{P}_n$, tem dimensão igual a $n+1$. Estamos interessados nas bases ortogonais ou ortonormadas $\{p_0, p_1, \ldots, p_n\}$ de $\mathbb{P}_n$ com a propriedade de

$$\text{grau}(p_k) = k, \quad k = 0, 1, \ldots, n.$$

Neste caso, diz-se que $p_0(x), p_1(x), \ldots, p_n(x)$ é uma sequência de *polinómios ortogonais* em $L^2_w(-1,1)$. Por exemplo, quando $n = 2$ e $w(x) = 1$, os polinómios

$$p_0(x) = 1, \quad p_1(x) = x \quad \text{e} \quad p_2(x) = 3x^2 - 1$$

são ortogonais:

$$\int_{-1}^{1} p_0(x) p_1(x)\, dx = 0, \quad \int_{-1}^{1} p_0(x) p_2(x)\, dx = 0 \quad \text{e} \quad \int_{-1}^{1} p_1(x) p_2(x)\, dx = 0.$$

Seja $p_0(x), p_1(x), \ldots, p_n(x)$ uma sequência de polinómios ortogonais em $L^2_w(-1,1)$. Os polinómios ortogonais gozam da seguinte propriedade, cuja demonstração é omitida por ser demasiado técnica.

[1] Trata-se de um semiproduto interno e de uma seminorma, a menos que as funções sejam contínuas em $[-1, 1]$.

Teorema 1 *Os polinómios ortogonais $p_0(x), p_1(x), \ldots, p_n(x)$ satisfazem uma fórmula de recorrência (de três termos) dada por*

$$p_{k+1}(x) \;=\; A_k(x - B_k)p_k(x) - C_k p_{k-1}(x), \quad k = 0, 1, \ldots, n-1,$$

com

$$A_k \;=\; \frac{\alpha_{k+1}}{\alpha_k}, \quad B_k \;=\; \frac{\langle x p_k(x), p_k(x)\rangle_w}{\|p_k\|_w^2} \quad e \quad C_k = \begin{cases} \textit{qualquer} & k = 0, \\ \frac{A_k \|p_k\|_w^2}{A_{k-1}\|p_{k-1}\|_w^2} & 1 \le k \le n-1, \end{cases}$$

em que α_k designa o coeficiente de x^k em $p_k(x)$ e, por convenção, $p_{-1}(x) = 0$.

O resultado deste teorema permite gerar sequências de polinómios ortogonais, concretizando o que não está especificado na fórmula de recorrência: a função peso w (que determina o produto interno), o valor para $p_0(x)$ e os coeficientes α_{k+1} (ou os escalares A_k). Se especificarmos o valor de p_{k+1} num dado ponto para todo o k, estamos, indirectamente, a especificar estes coeficientes (ver exercício). É isto que vamos fazer de seguida.

Por exemplo, se $w(x) = 1$, $p_0(x) = 1$ e $p_{k+1}(1) = 1$ para todo o $k \ge 0$, obtemos os chamados *polinómios de Legendre*. Os cinco primeiros polinómios de Legendre são os seguintes:

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= x, \\ L_2(x) &= (3x^2 - 1)/2, \\ L_3(x) &= (5x^3 - 3x)/2, \\ L_4(x) &= (35x^4 - 30x^2 + 3)/8. \end{aligned}$$

É possível mostrar que a fórmula de recorrência para os polinómios de Legendre é dada por

$$L_0(x) \;=\; 1, \quad L_1(x) \;=\; x, \quad L_{k+1}(x) \;=\; \frac{(2k+1)xL_k(x) - kL_{k-1}(x)}{k+1}, \quad k \ge 1.$$

Deduzimos, apenas, os três primeiros polinómios de Legendre. Por convenção, $p_{-1}(x) = 0$. Por escolha, $p_0(x) = 1$. A seguir vem

$$p_1(x) \;=\; A_0(x - B_0)p_0(x) - C_0 p_{-1}(x) \;=\; A_0(x - B_0) \;=\; A_0 x,$$

pois $B_0 = \langle x, 1\rangle_w / \|p_0\|_w = 0$. Escolha-se A_0 por forma a que $p_1(1) = 1$: $A_0 = 1$ (ou $\alpha_1 = 1$). Logo, $p_1(x) = x$. Depois, porque $B_1 = 0$ e $C_1 = A_1/3$,

$$p_2(x) \;=\; A_1(x - B_1)p_1(x) - C_1 p_0(x) \;=\; A_1 x^2 - \frac{1}{3}A_1.$$

Mas $p_2(1) = 1$ é equivalente a $A_1 = 3/2$ (ou $\alpha_2 = 3/2$). Desta forma, $p_2(x) = (3/2)x^2 - 1/2$.

Escolhendo $w(x) = (1-x^2)^{-\frac{1}{2}}$, $p_0(x) = 1$ e $p_{k+1}(1) = 1$ para todo o $k \geq 0$, obtemos os *polinómios de Chebyshev* (de primeira espécie). Os cinco primeiros polinómios de Chebyshev são os seguintes:

$$\begin{array}{rcl} T_0(x) & = & 1, \\ T_1(x) & = & x, \\ T_2(x) & = & 2x^2 - 1, \\ T_3(x) & = & 4x^3 - 3x, \\ T_4(x) & = & 8x^4 - 8x^2 + 1. \end{array}$$

Chega-se aos polinómios de Chebyshev de uma outra forma. De facto, considere-se a definição

$$T_k(x) \; = \; \cos(k \, \arccos(x)).$$

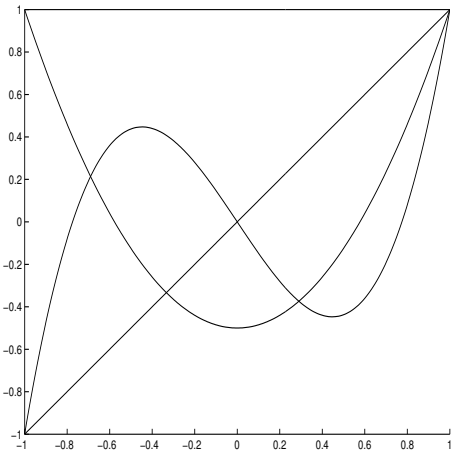
Vê-se, imediatamente, que $T_0(x) = 1$ e $T_k(1) = 1$ para todo o k. Mostra-se, utilizando as fórmulas aditivas das funções trigonométricas, que $T_k(x)$ é um polinómio em x para todo o k. Para além disso, prova-se que os polinómios são ortogonais:

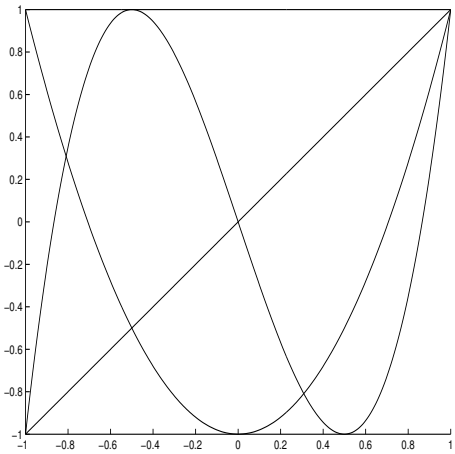
$$\begin{array}{rcl} \langle T_i, T_j \rangle_w & = & \int_{-1}^{1} (1-x^2)^{-\frac{1}{2}} T_i(x) T_j(x) \, dx \; = \; \int_0^{\pi} \cos(i\theta)\cos(j\theta) \, d\theta \\ & = & \frac{1}{2} \int_0^{\pi} \cos((i+j)\theta) + \cos((i-j)\theta) \, d\theta \\ & = & \left\{ \begin{array}{ll} \pi & i = j = 0, \\ \pi/2 & i = j \neq 0, \\ 0 & i \neq j. \end{array} \right. \end{array}$$

Assim sendo, estes polinómios são, necessariamente, os de Chebyshev. É possível mostrar, utilizando a sua definição trigonométrica, que a fórmula de recorrência para os polinómios de Chebyshev é dada por

$$T_0(x) \; = \; 1, \quad T_1(x) \; = \; x, \quad T_{k+1}(x) \; = \; 2xT_k(x) - T_{k-1}(x), \quad k \geq 1.$$

Terminamos a aula ilustrando os quatro primeiros polinómios de Legendre (figura à esquerda) e os quatro primeiros polinómios de Chebyshev (figura à direita).





Exercícios

1. Mostre que especificar A_k, α_{k+1} ou $p_{k+1}(1)$ determina, sem qualquer ambiguidade, o polinómio $p_{k+1}(x)$ na fórmula de recorrência do Teorema 1.

2. Prove as seguintes propriedades sobre os polinómios de Chebyshev:

   (a) $T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \; k \geq 1.$ (Use as fórmulas aditivas das funções trigonométricas.)

   (b) $\|T_k(x)\|_\infty = 1, \; k \geq 0.$

   (c) $T_k(x)$ atinge o valor absoluto de 1 nos pontos $x_j = \cos(j\pi/k), \; j = 0, \ldots, k.$ Mais precisamente: $T_k(x_j) = (-1)^j.$

   (d) As raízes de T_k são $\cos[(2j-1)\pi/(2k)]$, para $j = 1, \ldots, k.$

   (e) O polinómio $T_k(x)$ é uma função par se k for par e uma função ímpar se k for ímpar.

   (f) O coeficiente de x^k em $T_k(x)$ é dado por $\alpha_k = 2^{k-1}, \; k \geq 1.$

Aula 15: Polinómios Ortogonais – Propriedades

Consideremos o espaço vectorial real de funções $L^2_w(-1,1)$ e lembremos o produto interno definido neste espaço, $\langle f,g\rangle_w = \int_{-1}^{1} f(x)g(x)\,w(x)\,dx$, e a correspondente norma $\|\cdot\|_w$. Seja $\mathbb{P}_n$ o subespaço de $L^2_w(-1,1)$ constituído por todos os polinómios de grau inferior ou igual a n. Considere-se uma sequência $p_0(x), p_1(x), \ldots, p_n(x)$ de polinómios ortogonais em $\mathbb{P}_n$, para os quais grau$(p_k) = k$ para $k = 0, 1, \ldots, n$ e $\langle p_i, p_j\rangle_w = 0$ para $i \neq j$ com $i, j \in \{1, \ldots, n\}$.

Os polinómios ortogonais gozam de diversas propriedades e inúmeras aplicações. Vamos estudar algumas dessas propriedades (assumindo $k \leq n$).

Proposição 1 *Se $p(x)$ for um polinómio de grau inferior ou igual a $k \geq 0$ então existem escalares reais $c_0, c_1, \ldots, c_k$, unicamente determinados por $p(x)$, tais que*

$$p(x) \;=\; c_0p_0(x) + c_1p_1(x) + \cdots + c_kp_k(x).$$

Se a_k for o coeficiente de x^k em $p(x)$ então

$$c_k \;=\; \frac{a_k}{\alpha_k},$$

em que, como se sabe, α_k é o coeficiente de x^k em $p_k(x)$.

A demonstração da primeira parte desta propriedade assenta, exclusivamente, no facto de $\{p_0, p_1, \ldots, p_k\}$ ser uma base para $\mathbb{P}_k$. A demonstração da segunda parte é deixada como exercício. A partir desta propriedade conclui-se, imediatamente, a seguinte.

Proposição 2 *Se $p(x)$ for um polinómio de grau inferior a $k \geq 1$ então*

$$\langle p, p_k\rangle_w \;=\; 0.$$

Lembremos o exemplo da aula anterior em que $w(x) = 1$: $p_0(x) = 1$, $p_1(x) = x$ e $p_2(x) = 3x^2 - 1$. Tome-se $p(x) = x + 1$. Tem-se que

$$\langle p, p_2\rangle_w \;=\; \int_{-1}^{1} (1+x)(3x^2-1)\,dx \;=\; 0.$$

Esta última propriedade tem várias consequências importantes, entre as quais se encontra a seguinte.

Teorema 1 *O polinómio $p_k(x)$, $k \geq 1$, tem k raízes reais em $(-1,1)$, ou seja, $p_k(x)$ é da forma*

$$p_k(x) \;=\; \alpha_k(x-\xi_1)\cdots(x-\xi_k).$$

Continuando com o nosso exemplo,

$$p_0(x) \;=\; \alpha_0 \;=\; 1, \quad p_1(x) \;=\; \alpha_1(x-\xi_1) \;=\; 1(x-0)$$

e

$$p_2(x) \;=\; \alpha_2(x-\xi_1)(x-\xi_2) \;=\; 3(x+1/\sqrt{3})(x-1/\sqrt{3}).$$

Demonstração. Sejam $\xi_1, \ldots, \xi_r$ as r raízes reais de $p_k(x)$ em $(-1,1)$. Sabe-se que $r \leq k$. Pretende-se provar que $r = k$.

Tentemos chegar a uma contradição quando $r < k$. Escolha-se y entre a maior das raízes $\xi_1, \ldots, \xi_r$ e 1. O polinómio

$$p(x) \;=\; p_k(y)(x-\xi_1)\cdots(x-\xi_r)$$

tem, no intervalo $(-1,1)$, o mesmo sinal de $p_k(x)$. Logo, em todos os pontos x diferentes das ditas raízes,

$$p(x)p_k(x)\,w(x) \;>\; 0.$$

Porém, como o grau de $p(x)$ é igual a $r < k$, vem, pela propriedade anterior, que

$$\langle p, p_k\rangle_w \;=\; \int_{-1}^{1} p(x)p_k(x)\,w(x)\,dx \;=\; 0,$$

o que contradiz o facto de $p(x)p_k(x)w(x)$ ser positivo em $(-1,1)$ exceptuando num número finito de pontos (as raízes $\xi_1, \ldots, \xi_r$). ■

Exercícios

1. Sejam $p_0(x), p_1(x), \ldots, p_n(x)$ polinómios ortogonais em $\mathbb{P}_n$. Mostre que p_{k+1} (com $0 \leq k \leq n-1$) se pode escrever na forma

$$p_{k+1}(x) \;=\; \frac{\alpha_{k+1}}{\alpha_k} x p_k(x) + \sum_{i=0}^{k} d_i p_i(x),$$

para um determinado conjunto de escalares reais $d_0, d_1, \ldots, d_k$. (Este facto é o ponto de partida para a demonstração da fórmula de recorrência dada na aula anterior.)

Aula 16: Integração Gaussiana

Lembramos que qualquer fórmula de quadratura interpolatória tem um grau de exactidão que varia entre n e $2n+1$. Os polinómios ortogonais desempenham um papel fundamental no desenvolvimento de fórmulas de quadratura com um grau de exactidão elevado (por exemplo, $2n+1$).

Consideremos um conjunto de pontos $\{x_0, x_1, \ldots, x_n\}$, como nas aulas dadas sobre integração numérica, mas, desta vez, contidos no intervalo $[-1, 1]$ — para podermos, mais facilmente, aplicar o material dado sobre polinómios ortogonais. Fica relegada para mais tarde a integração Gaussiana sobre um intervalo $[a, b]$ arbitrário.

Pretendemos analisar fórmulas de quadratura interpolatórias para o cálculo de integrais do tipo

$$I_w(f) \;=\; \int_{-1}^{1} f(x)\, w(x)\, dx,$$

em que $w(x)$ é uma função de ponderação, positiva e integrável em $(-1, 1)$. A função f tem que definir bem este integral (*e.g.*, f contínua em $[-1, 1]$). Estas fórmulas de quadratura interpolatória são da forma:

$$I_{n,w}(f) \;=\; \int_{-1}^{1} \Pi_n f(x)\, w(x)\, dx \;=\; \sum_{i=0}^{n} \alpha_i f(x_i),$$

com $\alpha_i = \int_{-1}^{1} \ell_i(x)\, w(x)\, dx$, $i = 0, 1, \ldots, n$. A única novidade aqui é a presença da função-peso $w(x)$. Recordamos que $\Pi_n f(x)$ é o polinómio interpolador de f em $\{x_0, x_1, \ldots, x_n\}$ e que os $\ell_i's$ são os polinómios de Lagrange associados a este conjunto de pontos de interpolação.

O teorema seguinte vai permitir-nos desenvolver uma fórmula de quadratura com grau de exactidão igual a $2n+1$.

Teorema 1 *Seja m um número inteiro positivo. Uma fórmula de quadratura interpolatória Gaussiana tem grau de exactidão igual a $n+m$ se e só se o polinómio (dito nodal)*[2]

$$w_n(x) \;=\; (x - x_0)(x - x_1) \cdots (x - x_n)$$

for tal que

$$\int_{-1}^{1} w_n(x) p(x)\, w(x)\, dx \;=\; 0 \qquad \textit{para todo o} \quad p \;\in\; \mathbb{P}_{m-1}.$$

Demonstração. Pode ser facilmente confirmado que a condição é necessária.

Provemos a suficiência desta condição. Seja h um polinómio, qualquer, em $\mathbb{P}_{n+m}$. O nosso objectivo é provar que $I_{n,w}(h) = I_w(h)$, ou seja, que $E_{n,w}(h) = 0$. Como o grau

[2] O leitor vai aperceber-se que existe uma discrepância na notação utilizada nesta aula, uma vez que o polinómio nodal w_n tem grau $n+1$ e não grau n.

de h é igual a $n+m$, existe um polinómio-quociente $\pi_{m-1} \in \mathbb{P}_{m-1}$ e um polinómio-resto $q_n \in \mathbb{P}_n$ tais que

$$h(x) \;=\; w_n(x)\pi_{m-1}(x) + q_n(x).$$

Veja-se que $h(x_i) = q_n(x_i)$, $i = 0, 1, \ldots, n$, uma consequência da definição do polinómio nodal $w_n(x)$. A expressão para h e a condição suficiente implicam que

$$\int_{-1}^{1} q_n(x)\, w(x)\, dx \;=\; \int_{-1}^{1} h(x)\, w(x)\, dx - \int_{-1}^{1} w_n(x)\pi_{m-1}(x)\, w(x)\, dx \;=\; \int_{-1}^{1} h(x)\, w(x)\, dx.$$

Por outro lado, como a fórmula de quadratura interpolatória tem grau de exactidão pelo menos igual a n, vem que $I_{n,w}(q_n) = I_w(q_n)$. Logo,

$$\sum_{i=0}^{n} \alpha_i q_n(x_i) \;=\; \int_{-1}^{1} q_n(x)\, w(x)\, dx.$$

Combinando estas duas últimas igualdades e aplicando $h(x_i) = q_n(x_i)$, $i = 0, 1, \ldots, n$, resulta em

$$\int_{-1}^{1} h(x)\, w(x)\, dx \;=\; \sum_{i=0}^{n} \alpha_i q_n(x_i) \;=\; \sum_{i=0}^{n} \alpha_i h(x_i),$$

o que mostra que o erro $E_{n,w}(h)$ é nulo. ■

O valor máximo que podemos dar a m é $n+1$ (porquê?). Se fizermos $m = n+1$, a condição a satisfazer toma a forma

$$\int_{-1}^{1} w_n(x)p(x)\, w(x)\, dx \;=\; 0 \qquad \text{para todo o} \quad p \;\in\; \mathbb{P}_n.$$

Daqui concluímos que o polinómio nodal w_n, de grau igual a $n+1$, tem de ser ortogonal a todos os polinómios em $\mathbb{P}_n$. Logo, w_n deve ser um múltiplo de p_{n+1}, o polinómio ortogonal de grau $n+1$ associado à função-peso $w(x)$ (porquê?). Assim sendo, as raízes de w_n e de p_{n+1} têm de coincidir, ou seja, os pontos $x_0, x_1, \ldots, x_n$ têm de ser escolhidos de forma a satisfazer

$$p_{n+1}(x_i) \;=\; 0, \quad i = 0, 1, \ldots, n.$$

Concluímos, deste modo, que, para se atingir um grau de exactidão $2n+1$, dever-se-á escolher como pontos de interpolação $x_0, x_1, \ldots, x_n$ para a fórmula de quadratura as raízes do polinómio ortogonal p_{n+1} associado à função de ponderação $w(x)$.

Quando for dado um intervalo $[a, b]$ arbitrário, efectua-se uma mudança de variável e reduz-se os cálculos ao intervalo $[-1, 1]$. De facto, tem-se que

$$\int_{a}^{b} f(x)\, w(x)\, dx \;=\; \frac{b-a}{2} \int_{-1}^{1} f(\phi(\xi))\, w(\phi(\xi))\, d\xi,$$

com $\phi : [-1,1] \longrightarrow [a,b]$ a transformação afim definida por

$$\phi(\xi) \;=\; \frac{b-a}{2}\,\xi + \frac{a+b}{2}.$$

Se se aplicar ao cálculo do integral do membro do lado direito uma fórmula de quadratura Gaussiana do tipo

$$\sum_{i=0}^{n} \beta_i f(\phi(\xi_i)),$$

isso corresponde a aplicar, ao integral original, uma fórmula de quadratura da forma

$$\sum_{i=0}^{n} \alpha_i f(x_i),$$

com $x_i = \phi(\xi_i)$ e $\alpha_i = (b-a)\beta_i/2$, $i = 0, 1, \ldots, n$. Esta fórmula tem o mesmo grau de exactidão da fórmula aplicada no intervalo $[-1,1]$ (ver exercício).

Exercícios

1. Por que motivo é que a condição do Teorema 1 é necessária?

2. Considere o cálculo aproximado do integral

$$\int_a^b f(x)\, w(x)\, dx$$

por intermédio de uma fórmula de quadratura da forma

$$\sum_{i=0}^{n} \beta_i f(x_i)$$

com $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$. Calcule os x_i's e os β_i's de forma a que esta fórmula tenha grau de exactidão máximo $(2n+1)$ quando:

   (a) $w(x) = \sqrt{x}$, $a = 0$, $b = 1$ e $n = 1$.

   (b) $w(x) = 2x^2 + 1$, $a = -1$, $b = 1$ e $n = 0$.

   (c) $w(x) = 2$ se $0 < x \leq 1$ e $w(x) = 1$ se $-1 \leq x \leq 0$, $a = -1$, $b = 1$ e $n = 1$.

   Escreva somente a formulação, em cada alínea, dos sistemas que determinam os x_i's e os β_i's.

3. Mostre que, se a fórmula de quadratura Gaussiana $\sum_{i=0}^{n} \beta_i f(\phi(\xi_i))$ tem grau de exactidão igual a r em $[-1,1]$, então a fórmula de quadratura Gaussiana $\sum_{i=0}^{n} \alpha_i f(x_i)$, em $[a,b]$, tem, também, um grau de exactidão igual a r (com $x_i = \phi(\xi_i)$ e $\alpha_i = (b-a)\beta_i/2$, $i = 0, 1, \ldots, n$).

4. Considere uma fórmula de quadratura interpolatória, para aproximação do integral $I(f) = \int_{-1}^{1} f(x)\,dx$, da forma

$$I_1(f) \;=\; \alpha_0 f(x_0) + \alpha_1 f(\sqrt{3}/3).$$

   (a) Para que valores de α_0, α_1 e x_0 atinge esta fórmula o seu grau de exactidão máximo? (No caso de α_0 e α_1 só precisa de indicar as suas expressões.)

   (b) Considere $\alpha_1 = 0$. Para que valores de α_0 e x_0 atinge a fórmula $I_0(f) = \alpha_0 f(x_0)$ o seu grau de exactidão máximo?

   (c) Classifique as fórmulas que encontrou nas duas alíneas anteriores.

Aula 17: Introdução à Aproximação Trigonométrica

Muitos fenómenos físicos, como os que envolvem a luz ou o som, apresentam características periódicas. Uma função diz-se periódica se existir $\tau \in \mathbb{R}^+$ tal que $f(x+\tau) = f(x)$ para todo o x. Neste caso, o menor número real positivo τ para o qual esta propriedade se verifica chama-se período de f. Para aproximar funções deste tipo vamos considerar *polinómios trigonométricos*. Os polinómios algébricos são insatisfatórios para aproximar funções periódicas, uma vez que o único polinómio algébrico que é periódico é a função constante.

Um polinómio trigonométrico de ordem $2n+1$ é uma combinação linear de *polinómios de Fourier*. Os polinómios de Fourier são dados por

$$\phi_k(x) \;=\; e^{ikx} \;=\; \cos(kx) + i\,\mathrm{sen}(kx), \quad k = 0, \pm 1, \pm 2, \ldots$$

Os polinómios de Fourier (e, consequentemente, os trigométricos) são funções periódicas, com período 2π. Não ocorre perca de generalidade em particularizar o período da aproximação. Se a função original f tiver período τ, então $f(\tau x/2\pi)$ tem período 2π (ver exercício) e pode ser aproximada por polinómios trigonométricos (designe-se por $\bar{f}$ esta aproximação). A aproximação desejada é, então, dada por $\bar{f}(2\pi x/\tau)$ e tem período τ.

Observando que

$$\frac{1}{2\pi}\int_0^{2\pi} \phi_j(x)\overline{\phi_k(x)}\,dx \;=\; \frac{1}{2\pi}\int_0^{2\pi} e^{i(j-k)x}\,dx \;=\; \begin{cases} 1 & j = k, \\ 0 & j \neq k, \end{cases}$$

afirmamos que os polinómios de Fourier são ortonormados relativamente ao produto interno[3]

$$\langle f, g\rangle \;=\; \frac{1}{2\pi}\int_0^{2\pi} f(x)\overline{g(x)}\,dx.$$

No contexto geral da teoria de aproximação, consideramos o espaço vectorial complexo E de funções

$$L^2(0,2\pi) \;=\; \left\{ f : [0,2\pi] \longrightarrow \mathbb{C} : \int_0^{2\pi} |f(x)|^2\,dx \;<\; +\infty \right\},$$

onde este produto interno se encontra bem definido. A norma associada ao produto interno é dada por

$$\|f\|_{L^2(0,2\pi)} \;=\; \left(\frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx\right)^{\frac{1}{2}}.$$

Se tomarmos o subespaço de aproximação S como sendo o subespaço gerado pelos $2n+1$ polinómios de Fourier $\phi_k(x)$, $k = 0, \pm 1, \pm 2, \ldots, \pm n$, podemos aproximar f pela sua projecção ortogonal sobre S, dada por

$$f_{2n}^*(x) \;=\; \sum_{k=-n}^{n} \hat{f}_k \phi_k(x),$$

[3]Trata-se de um semiproduto interno e de uma seminorma, a menos, por exemplo, que as funções sejam contínuas em $[0, 2\pi]$.

em que os coeficientes $\hat{f}_k$, conhecidos por *coeficientes de Fourier*, são dados por

$$\hat{f}_k \;=\; \langle f, \phi_k \rangle_{L^2(0,2\pi)} \;=\; \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx.$$

Esta projecção ortogonal satisfaz $\|f_{2n}^* - f\|_{L^2(0,2\pi)} = \min_{\psi \in S} \|\psi - f\|_{L^2(0,2\pi)}$.

É frequente a função f a aproximar ser real, existindo interesse, nesta situação, em que f_{2n}^* seja, igualmente, uma função real. O polinómio trigonométrico f_{2n}^* é real se e só se $\overline{f_{2n}^*} = f_{2n}^*$. Como

$$\overline{\sum_{k=-n}^{n} \hat{f}_k \phi_k(x)} \;=\; \sum_{k=-n}^{n} \overline{\hat{f}_k} e^{-ikx} \;=\; \sum_{k=-n}^{n} \overline{\hat{f}_{-k}} e^{ikx}$$

e as funções ϕ_k são linearmente independentes, o polinómio trigonométrico f_{2n}^* é real se e só se

$$\hat{f}_k \;=\; \overline{\hat{f}_{-k}}, \quad k = 0, \pm 1, \pm 2, \ldots, \pm n.$$

Deste modo, se o polinómio trigonométrico f_{2n}^* for real, este pode ser escrito na forma

$$f_{2n}^*(x) \;=\; \hat{f}_0 + \sum_{k=1}^{n} \left[2\mathrm{Re}(\hat{f}_k) \cos(kx) - 2\mathrm{Im}(\hat{f}_k) \mathrm{sen}(kx) \right].$$

Vejamos o exemplo simples em que $f(x) = \mathrm{sen}(x)$. Tem-se que $f_2 = \mathrm{sen}$, uma vez que a função seno pertente ao subespaço gerado por ϕ_{-1}, ϕ_0 e ϕ_1. Se fizermos as contas vem que

$$\hat{f}_{-1} \;=\; i/2, \quad \hat{f}_0 \;=\; 0 \quad \text{e} \quad \hat{f}_1 \;=\; -i/2.$$

Logo,

$$f_2^*(x) \;=\; \hat{f}_0 + 2\mathrm{Re}(\hat{f}_1) \cos(x) - 2\mathrm{Im}(\hat{f}_1) \mathrm{sen}(x) \;=\; 0 + 0\, \cos(x) + 1\, \mathrm{sen}(x) \;=\; \mathrm{sen}(x).$$

Também se mostra, no caso real, que

$$f_{2n}^*(x) \;=\; \hat{f}_0 + \sum_{k=1}^{n} 2|\hat{f}_k| \cos(\theta_k + kx) \;\simeq\; f(x),$$

em que o ângulo θ_k desempenha o papel de um *deslocamento de fase*. A função periódica f é, assim, aproximada por uma soma de n oscilações harmónicas da forma

$$2|\hat{f}_k| \cos(\theta_k + kx).$$

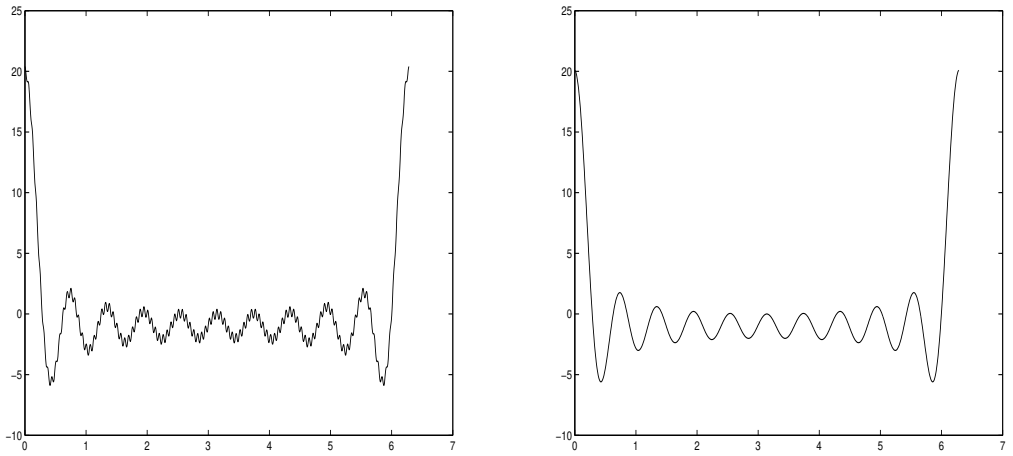
Cada uma destas oscilações tem

- *amplitude* $2|\hat{f}_k|$,
- *frequência* $\frac{k}{2\pi}$,
- período ou *comprimento de onda* $\frac{2\pi}{k}$,
- *ângulo de fase* θ_k.

A sequência $|\hat{f}_0|, |\hat{f}_1|, \ldots, |\hat{f}_n|$, ou a sequência dos seus quadrados, é conhecida por *espectro* (neste caso finito) de f.

Dependendo da forma como a *energia total* de f, $\|f\|^2_{L^2(0,2\pi)}$, se distribui sobre o seu espectro, a função f pode ser suave (quando o espectro decresce rapidamente) ou ruidosa (quando componentes do espectro não são pequenas para frequências k elevadas). A técnica de suavização de uma função f com ruído consiste, primeiro, em gerar os coeficientes de Fourier, ou uma sua aproximação, para, a seguir, *filtrar* estes coeficientes, suprimindo (anulando) os coeficientes correspondentes a frequências elevadas. Depois, reconstrói-se uma nova função (mais suave do que a original) através dos coeficientes filtrados.

Apresentamos, de seguida, o resultado de uma experiência de suavização, feita em MATLAB, onde foram filtrados perto de 85% dos coeficientes de Fourier (os de maior frequência).



Exercícios

1. Prove que se f tem período τ, então $f(\tau x/2\pi)$ tem período 2π.
2. Prove que se f for uma função real então

$$\hat{f}_k e^{ikx} + \hat{f}_{-k} e^{-ikx} \;=\; 2|\hat{f}_k| \cos(\theta_k + kx),$$

para um determinado θ_k. **Sugestão:** Escreva $\hat{f}_k$ na forma $|\hat{f}_k| e^{i\theta_k}$.

3. Se f tiver período 2π também o têm as funções $g_m(x) = f(mx)$ (com m um número inteiro) e $h_\alpha(x) = f(x - \alpha)$ (com α um número real). Mostre como se relacionam os coeficientes de Fourier de f com os de g_m e os de h_α.

Aula 18: Transformadas Discreta e Rápida de Fourier

A questão que se coloca com maior premência prática em aproximação de Fourier prende-se com o cálculo dos coeficientes de Fourier

$$\hat{f}_k \;=\; \langle f, \phi_k \rangle_{L^2(0,2\pi)} \;=\; \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx, \quad k = 0, \pm 1, \pm 2, \ldots, \pm n.$$

A expressão analítica da função f a aproximar é desconhecida, o que impossibilita esta integração de forma exacta. O que se conhece, em concreto, é o valor da função num conjunto finito de $N+1$ *pontos de amostragem* $\{x_0, x_1, \ldots, x_N\}$. Recorrendo a esta informação, é possível aplicar uma fórmula de quadratura para aproximar os integrais dos coeficientes de Fourier. Na sua notação original, a fórmula trapezoidal composta foi apresentada na forma

$$I_{1,m}(f) \;=\; H \left[\frac{1}{2} f(y_0) + f(y_1) + \cdots + f(y_{m-1}) + \frac{1}{2} f(y_m) \right].$$

Na situação que temos em mãos, $H = (2\pi - 0)/N$ e $x_j = 2\pi j/N$, $j = 0, 1, \ldots, N$. Reunindo esta informação, estamos em condições de aproximar os coeficientes de Fourier por

$$\hat{f}_k \;\simeq\; \tilde{f}_k \;=\; \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) e^{-ik\,x_j}.$$

Note que utilizámos o facto de f ter período 2π e, portanto, $f(0) = f(2\pi)$, ou seja, $f(x_0) = f(x_N)$. Assim sendo, o número de pontos de amostragem baixa para N.

Como veremos mais adiante, o número de pontos de amostragem deve ser igual ao número de coeficientes de Fourier a aproximar:

$$N \;=\; 2n + 1.$$

Os *coeficientes de Fourier discretos* $\tilde{f}_k$ definem uma aproximação para f^*_{2n}:

$$f^*_{2n}(x) \;\simeq\; \tilde{f}^*_{2n}(x) \;=\; \sum_{k=-n}^{n} \tilde{f}_k \phi_k(x).$$

Curiosamente, esta função $\tilde{f}^*_{2n}$ é a projecção ortogonal de f sobre o subespaço S gerado pelos $2n+1$ polinómios de Fourier $\phi_k(x)$, $k = 0, \pm 1, \pm 2, \ldots, \pm n$, mas, relativamente ao *semiproduto interno discreto*, definido por

$$\langle g, h \rangle_N \;=\; \frac{1}{N} \sum_{j=0}^{N-1} g(x_j) \overline{h(x_j)}$$

(que não é um produto interno em $L^2(0,2\pi)$ pois $\langle g,g\rangle_N$ pode ser nulo sem que g o seja). Confirma-se, no exercício em baixo, que os polinómios de Fourier são ortonormados neste semiproduto interno discreto. Além disso, constata-se, através de um tipo de cálculos semelhante ao deste exercício, que a aproximação discreta de Fourier $\tilde{f}^*_{2n}$ interpola f nos pontos x_ℓ, com $\ell = 0,1,\ldots,N-1$:

$$\begin{aligned}\tilde{f}^*_{2n}(x_\ell) \;=\; \textstyle\sum_{k=-n}^{n} \tilde{f}_k \phi_k(x_\ell) \;&=\; \textstyle\sum_{k=-n}^{n} \left(\frac{1}{N} \sum_{j=0}^{N-1} f(x_j) e^{-ik\,x_j} \right) e^{ik\,x_\ell} \\ &=\; \textstyle\frac{1}{N} \sum_{j=0}^{N-1} \left(\sum_{k=-n}^{n} e^{ik\,x_\ell} e^{-ik\,x_j} \right) f(x_j) \\ &=\; \cdots \\ &=\; f(x_\ell).\end{aligned}$$

A transformação linear que converte $f(x_0), f(x_1), \ldots, f(x_{N-1})$ em $\tilde{f}_{-n}, \ldots, \tilde{f}_{-1}, \tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_n$ chama-se *transformada discreta de Fourier*. Como $x_j = 2\pi j/N$, apresentamos os coeficientes de Fourier discretos na forma:

$$\tilde{f}_k \;=\; \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) w^{kj}, \quad k = -n, \ldots, -1, 0, 1, \ldots, n,$$

com

$$w \;(=\; w_N) \;=\; e^{-i2\pi/N}.$$

O número complexo w pertence ao círculo unitário $\{u \in \mathbb{C} : \; |u| = 1\}$ e satisfaz $w^N = 1 = w^0$. Antes de representarmos matricialmente os coeficientes de Fourier discretos, vamos fazer uma modificação nas linhas correspondentes aos índices $k = -n, \ldots, -1$, reescrevendo-as como

$$\tilde{f}_k \;=\; \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) w^{kj} \underbrace{w^{jN}}_{\shortparallel\; 1} \;=\; \frac{1}{N} \sum_{j=0}^{N-1} f(x_j) w^{(k+N)j}.$$

Quando o índice k varia de $-n$ até -1, o índice $k+N$ varia de $-n+N$ até $N-1$ ou seja, varia de $n+1$ até $N-1$:

k	-1	-2	$\cdots$	$-n$
$k+N$	$2n$	$2n-1$	$\cdots$	$n+1$

A forma matricial que procuramos é dada por

$$\begin{bmatrix} \tilde{f}_0 \\ \tilde{f}_1 \\ \vdots \\ \tilde{f}_n \\ \tilde{f}_{-n} \\ \vdots \\ \tilde{f}_{-1} \end{bmatrix} \;=\; \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w & w^2 & \cdots & w^{N-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^n & w^{n2} & \cdots & w^{n(N-1)} \\ 1 & w^{n+1} & w^{(n+1)2} & \cdots & w^{(n+1)(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{(N-1)2} & \cdots & w^{(N-1)^2} \end{bmatrix} \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \\ f(x_{n+1}) \\ \vdots \\ f(x_{N-1}) \end{bmatrix}.$$

A transformada discreta de Fourier representa-se, assim, por

$$\tilde{z} \;=\; \frac{1}{N} F z.$$

A matriz F é designada por *matriz da transformada discreta de Fourier*. Esta matriz é simétrica e não singular. A sua inversa é dada por $F^{-1} = \bar{F}/N$ (ver exercício). Quando $N = 5$, a matriz F toma a forma (lembre-se de que $w^5 = 1$):

$$F \;=\; \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 \\ 1 & w^2 & w^4 & w^6 & w^8 \\ 1 & w^3 & w^6 & w^9 & w^{12} \\ 1 & w^4 & w^8 & w^{12} & w^{16} \end{bmatrix} \;=\; \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 \\ 1 & w^2 & w^4 & w & w^3 \\ 1 & w^3 & w & w^4 & w^2 \\ 1 & w^4 & w^3 & w^2 & w \end{bmatrix}.$$

À transformação linear que, dados os coeficientes discretos de Fourier, reproduz os valores de f nos pontos de amostragem chama-se *transformada discreta de Fourier inversa*. Pelo que acabámos de ver, esta transformação linear opera da seguinte forma:

$$z \;=\; N F^{-1} \tilde{z} \;=\; \bar{F} \tilde{z}.$$

Na prática, para processar sinais, sobretudo quando sujeitos a ruído, é necessário tomar um número de amostragens elevado, o que se repercute em produtos matriz-vector (com as matrizes F e $\bar{F}$) de grandes dimensões.

Constata-se, facilmente, que o número de operações elementares (adições, subtracções, multiplicações e divisões) para efectuar o produto entre uma matriz em $\mathbb{C}^{N\times N}$ e um vector em $\mathbb{C}^N$ é um polinómio de grau 2 em N. É possível mostrar, aliás, que o primeiro termo deste polinómio é $8N^2$.

Os produtos complexos matriz-vector envolvendo as matrizes F ou $\bar{F}$ podem ser organizados de forma a que as operações elementares passem a ser, em número, dominadas por $5N \log_2 N$ (aplicando, recursivamente, a técnica do exercício em baixo). Este tipo de cálculos é conhecido por *transformada rápida de Fourier*, em inglês *fast Fourier transform (FFT)*. Os ganhos de $5N \log_2 N$ em relação a $8N^2$ são maiores do que aparentam (por conveniência da apresentação considera-se N par e potência de 2):

N	$\frac{8N^2}{5N \log_2 N}$
32	$\simeq 10$
1024	$\simeq 160$
32768	$\simeq 3500$
1048574	$\simeq 84000$

Exercícios

1. Mostre que $\langle \phi_k, \phi_\ell \rangle_N$ é igual a 1 se $k = \ell$ e igual a 0 se $k \neq \ell$, com k e ℓ a variar em $\{-n, \ldots, -1, 0, 1, \ldots, n\}$. (O caso $k = \ell$ é trivial. O caso $k \neq \ell$ resulta de uma soma geométrica.)

2. Prove que $\langle \phi_k, \phi_\ell \rangle_N$ é igual a 1 se $k = \ell (\text{mod } N)$ e igual a 0 se $k \neq \ell (\text{mod } N)$, quaisquer que sejam os inteiros k e ℓ.

3. Prove que ϕ_k e ϕ_ℓ coincidem em todos os pontos de amostragem $x_0, x_1, \ldots, x_{N-1}$ se $k = \ell (\text{mod } N)$. (O facto de não ser possível distinguir, no conjunto de amostragem, duas oscilações com estas características está na base do fenómeno conhecido por *distorção*.)

4. Mostre que a matriz F da transformada discreta de Fourier é simétrica e não singular (e que a sua inversa é dada por $\bar{F}/N$).

5. Mostre que, para N par,

$$\sum_{j=0}^{N-1} z_j x^j \;=\; p_{par}(x^2) + x\, p_{imp}(x^2)$$

   com

$$p_{par}(x) \;=\; z_0 + z_2 x + \cdots + z_{N-2} x^{\frac{N}{2}-1} \quad \text{e} \quad p_{imp}(x) \;=\; z_1 + z_3 x + \cdots + z_{N-1} x^{\frac{N}{2}-1}.$$

   (Esta técnica está na base da transformada rápida de Fourier.)

6. Considere a matriz da transformada discreta de Fourier de ordem 3:

$$F \;=\; \begin{bmatrix} 1 & 1 & 1 \\ 1 & w & w^2 \\ 1 & w^2 & w^4 \end{bmatrix}.$$

   (a) Escreva as componentes de F em função de, apenas, 1, w e w^2. Marque $1 = w^0$, $w = w^1$ e w^2 no círculo unitário do plano complexo.

   (b) Mostre que $F\bar{F} = 3I$. Identifique F^{-1}.

   (c) Mostre que $FC = DF$ em que

$$C \;=\; \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

   e D é a matriz diagonal de elementos diagonais 1, w e w^2. Quais são os valores próprios de C?

(d) Escreva a matriz (circulante)

$$H = \begin{bmatrix} h_0 & h_2 & h_1 \\ h_1 & h_0 & h_2 \\ h_2 & h_1 & h_0 \end{bmatrix}$$

em função das potências de C dadas por $I = C^0$, C e C^2.

(e) Multiplique esta expressão para H, à esquerda, por F e, à direita, por F^{-1}. Conclua que F também diagonaliza H. Quais são os valores próprios de H?

Aula 19: Formulação Variacional de um Problema de Condições de Fronteira

Dada uma função f, contínua em $[0,1]$, consideremos o seguinte problema de condições de fronteira, designado por (P):

$$\text{encontrar} \quad u \in C^2[0,1] \quad \text{tal que} \quad \begin{cases} -u''(x) \;=\; f(x) & \text{se } \; x \in (0,1), \\ u(0) \;=\; u(1) \;=\; 0. \end{cases} \tag{P}$$

É fácil constatar, ao integrar f duas vezes seguidas e ao utilizar as condições de fronteira $u(0) = u(1) = 0$ para determinar o valor das constantes de integração, que o problema (P) admite uma e uma só solução.

Se multiplicarmos a equação diferencial ordinária $-u''(x) = f(x)$ por uma *função teste* v definida em $[0,1]$ e integrarmos ambos os membros de 0 até 1 obtemos

$$-\int_0^1 u''(x)v(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx.$$

Ao integrarmos, por partes, o termo do membro do lado esquerdo vem que

$$\int_0^1 u'(x)v'(x)\,dx - \underbrace{[u'(x)v(x)]_0^1}_{\begin{array}{c}\shortparallel\\ u'(1)v(1) - u'(0)v(0)\end{array}} \;=\; \int_0^1 f(x)v(x)\,dx.$$

Se $v(0) = v(1) = 0$ então

$$\int_0^1 u'(x)v'(x)\,dx \;=\; \int_0^1 f(x)v(x)\,dx.$$

Estes dois integrais estão bem definidos se u', v e v' forem integráveis em $[0,1]$.

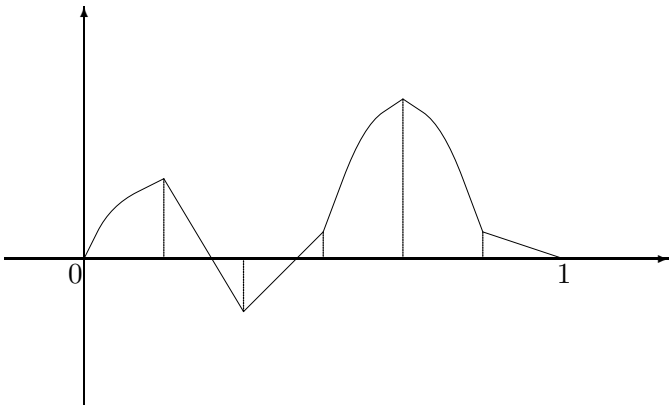
Esta integração *informal* sugere, à primeira vista, que consideremos o espaço vectorial

$$\{v:\; v \in C[0,1],\; v(0) = v(1) = 0,\; v' \in C[0,1]\}\,.$$

Os *vectores* deste espaço são, seguramente, funções integráveis e com derivadas integráveis em $[0,1]$. Para o tipo de esquemas numéricos que introduziremos mais tarde, este espaço mostrar-se-á insatisfatório. Será necessário considerar funções teste v cuja derivada é descontínua em determinados pontos do intervalo $[0,1]$. Assim, consideramos o espaço vectorial V definido por

$$V \;=\; \{v:\; v \in C[0,1],\; v(0) = v(1) = 0,\; v' \text{ contínua por troços e limitada em } [0,1]\}\,.$$

Ilustramos um exemplo de um elemento de V com seis troços na figura seguinte.



As funções derivadas das funções em V são, desta forma, integráveis em $[0,1]$. Note-se que estas funções derivadas não precisam de estar definidas nos pontos que delimitam os troços ou subintervalos. Assumimos que o número de troços é finito.

Estão bem definidos, em V, o produto interno

$$\langle v_1, v_2 \rangle \;=\; \int_0^1 v_1(x) v_2(x)\, dx,$$

e a norma

$$\|v\| \;=\; \sqrt{\langle v, v \rangle} \;=\; \sqrt{\int_0^1 v(x)^2\, dx}.$$

Estamos preparados para definir o problema variacional (V) associado ao problema de condições de fronteira (P):

$$\text{encontrar} \quad u \in V \quad \text{tal que} \quad \langle u', v' \rangle \;=\; \langle f, v \rangle \quad \text{para todo o } v \in V. \qquad \text{(V)}$$

Os problemas (P) e (V) são equivalentes, no sentido em que têm as mesmas soluções.

Teorema 1 *Se u for uma solução do problema de condições de fronteira (P) então u resolve o problema variacional (V).*

Se u for uma solução do problema variacional (V) e u'' existir e for contínua em $[0,1]$ então u resolve o problema de condições de fronteira (P).

Demonstração. A primeira implicação já foi provada informalmente. Se u for solução de (P) então u'' coincide com $-f$ e é, por hipótese, contínua em $[0,1]$. Assim sendo, toda a integração informal feita no início da aula está bem definida para funções teste v em V.

Provemos a implicação recíproca. Seja u uma solução do problema variacional (V). Então u satisfaz as condições de fronteira $u(0) = u(1) = 0$ e

$$\int_0^1 u'(x) v'(x)\, dx \;=\; \int_0^1 f(x) v(x)\, dx \quad \text{para todo o } v \in V.$$

Como u'' é contínua em $[0,1]$, é possível integrar, por partes, o integral do membro do lado esquerdo:

$$-\int_0^1 u''(x)v(x)\,dx + \underbrace{[u'(x)v(x)]_0^1}_{\begin{array}{c}\| \\ u'(1)v(1)-u'(0)v(0)=0\end{array}} = \int_0^1 f(x)v(x)\,dx \quad \text{para todo o } v \in V.$$

Desta forma,

$$-\int_0^1 (u''(x)+f(x))\,v(x)\,dx \;=\; 0 \quad \text{para todo o } v \in V,$$

o que, pela continuidade de $u''+f$, implica

$$u''(x)+f(x) \;=\; 0 \quad \text{para todo o } x \in (0,1),$$

demonstrando que u é solução da equação diferencial do problema (P). ■

Prova-se, ainda, que a solução do problema variacional (V) é única. Suponhamos que o problema (V) admitia duas soluções, u_1 e u_2. Para qualquer $v \in V$ viria que

$$\begin{aligned}\langle u_1', v'\rangle &= \langle f, v\rangle \\ \langle u_2', v'\rangle &= \langle f, v\rangle.\end{aligned}$$

Subtraindo estas igualdades membro a membro e escolhendo $v = u_1 - u_2 \in V$, resultaria em

$$\langle u_1' - u_2', u_1' - u_2'\rangle \;=\; 0.$$

Concluir-se-ia daqui (ver exercício) que $u_1 - u_2 = 0$ em $[0,1]$, o que demonstra a unicidade de solução do problema (V).

Exercícios

1. Considere a função
$$v(x) \;=\; \begin{cases} 2x & x \in [0,1/2], \\ 2-2x & x \in (1/2,1]. \end{cases}$$
Verifique que v está em V.

2. Prove que a aplicação que a cada par de elementos v_1, v_2 de V faz corresponder o número real $\langle v_1, v_2\rangle = \int_0^1 v_1(x)v_2(x)\,dx$ é um produto interno em V.

3. **(Difícil.)** Demonstre que se z é uma função contínua em $[0,1]$ e
$$\langle z, v\rangle \;=\; \int_0^1 z(x)v(x)\,dx \;=\; 0 \quad \text{para todo o } v \in V$$
então z é a função nula em $[0,1]$.

4. Seja v pertencente a V tal que $\langle v', v' \rangle = 0$.

   (a) Mostre, primeiro, que v' se anula em todos os pontos de $[0,1]$ que não são extremos de subintervalos a definir troços contínuos de v'.

   (b) Mostre que a função v é nula. Conclua, deste modo, que a função v' também é nula.

5. Dada uma função f, contínua em $[0,1]$, considere o seguinte problema de condições de fronteira:

$$\text{encontrar} \quad u \in C^2[0,1] \quad : \quad \begin{cases} -u''(x) + u(x) \;=\; f(x) & \text{se } \; x \in (0,1), \\ u(0) \;=\; u(1) \;=\; 0. \end{cases} \qquad \text{(P')}$$

   (a) Encontre a formulação variacional (V') deste problema, provando que se u for uma solução do problema de condições de fronteira (P') então u resolve o problema variacional (V').

   (b) Prove que se u for uma solução do problema variacional (V') e u'' existir e for contínua em $[0,1]$ então u resolve o problema de condições de fronteira (P').

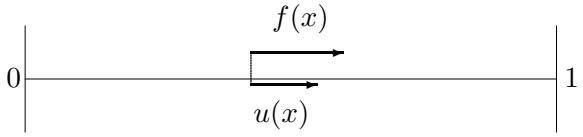
Aula 20: Princípio de Energia Potencial Mínima para um Problema de Condições de Fronteira

O problema de condições de fronteira,

$$\text{encontrar} \quad u \in C^2[0,1] \quad \text{tal que} \quad \begin{cases} -u''(x) \;=\; f(x) & \text{se } x \in (0,1), \\ u(0) \;=\; u(1) \;=\; 0, \end{cases} \tag{P}$$

com $f \in C[0,1]$, constitui um modelo elementar para várias situações em mecânica contínua.

Por exemplo, o deslocamento tangencial de uma barra elástica fixa em ambas as extremidades, quando sujeita a uma força tangencial de intensidade f, é traduzido pelo problema (P).

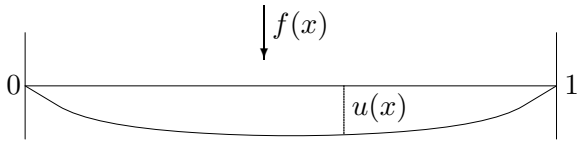


Designemos por $\sigma(x)$ a tracção e por $u(x)$ o deslocamento (ambos tangenciais). Sob a hipótese dos pequenos deslocamentos e supondo que o material é linearmente elástico, o deslocamento tangencial $u(x)$ obedece a

$$\begin{aligned} \sigma(x) \;&=\; Eu'(x) && \text{(lei de Hooke)}, \\ -\sigma'(x) \;&=\; f(x) && \text{(equação de equilíbrio)}, \\ u(0) \;&=\; u(1) \;=\; 0 && \text{(condições de fronteira)}. \end{aligned}$$

Assumindo que o módulo de elasticidade E é igual a 1 e substituindo $\sigma(x) = u'(x)$ na segunda equação, obtém-se o problema (P).

O deslocamento transversal $u(x)$ de uma corda elástica, com tensão igual a 1, fixa em ambas as extremidades e sujeita a uma força transversal de intensidade $f(x)$ obedece, também, à formulação (P).



Em ambos os problemas apresentados, as quantidades

$$\frac{1}{2}a(v,v) \stackrel{\text{def}}{=} \frac{1}{2}\langle v', v'\rangle = \frac{1}{2}\int_0^1 v'(x)^2\,dx \qquad \text{e} \qquad \langle f, v\rangle = \int_0^1 f(x)v(x)\,dx$$

representam, respectivamente, a energia elástica interna e a energia potencial associada ao deslocamento definido por $v \in V$. A funcional

$$J(v) \stackrel{\text{def}}{=} \frac{1}{2}a(v,v) - \langle f, v\rangle$$

mede a energia potencial total associada ao deslocamento $v \in V$.

Esta motivação leva-nos à colocação de uma segunda formulação do problema (P):

$$\text{encontrar} \quad u \in V \quad \text{tal que} \quad J(u) \leq J(v) \quad \text{para todo o } v \in V. \qquad \text{(M)}$$

Este problema está associado ao princípio da energia potencial mínima em mecânica. Matematicamente, este problema é equivalente ao problema de condições de fronteira (P), no sentido em que têm as mesmas soluções se u'' existir e for contínua. Vimos, na aula anterior que o problema (P) era equivalente, neste sentido, ao problema variacional (V). Vamos ver, agora, que os problemas (V) e (M) são equivalentes.

A notação $a(v,v) = \langle v', v'\rangle$ não foi introduzida ao acaso e corresponde à forma de designar as aplicações bilineares no contexto de formulações variacionais. De facto, a função que a cada par de vectores v_1 e v_2 em V faz corresponder o número real

$$a(v_1, v_2) \stackrel{\text{def}}{=} \langle v_1', v_2'\rangle = \int_0^1 v_1'(x)v_2'(x)\,dx$$

é uma *aplicação ou forma bilinear* (ver exercício). Com esta forma bilinear, reescrevemos o problema (V) na forma

$$\text{encontrar} \quad u \in V \quad \text{tal que} \quad a(u,v) = \langle f, v\rangle \quad \text{para todo o } v \in V. \qquad \text{(V)}$$

Teorema 1 *O problema variacional (V) e o problema de energia mínima (M) têm as mesmas soluções.*

Demonstração. Suponhamos que u é uma solução do problema (M). O facto de V ser um espaço vectorial implica que $u + tv$ está em V, quaisquer que sejam $v \in V$ e $t \in \mathbb{R}$. Escolha-se um v qualquer em V. Tem-se que $J(u) \leq J(u + tv)$ para todo o $t \in \mathbb{R}$. Defina-se a seguinte função g real de variável real (já nossa conhecida das aulas sobre optimização):

$$g(t) \stackrel{\text{def}}{=} J(u + tv).$$

Como $g(0) = J(u)$, é óbvio que g atinge um mínimo em zero.

Fazendo as contas, vemos que g é uma função quadrática em t:

$$g(t) \;=\; J(u + tv) \;=\; \frac{1}{2}a(u,u) + t\,a(u,v) + \frac{t^2}{2}a(v,v) - \langle f, u\rangle - t\langle f, v\rangle.$$

Sendo g uma função continuamente diferenciável em t, $g'(0) = 0$, ou seja,

$$a(u,v) - \langle f, v\rangle \;=\; 0,$$

o que mostra que u resolve (V).

Provemos, reciprocamente, que as soluções de (V) são minimizantes de J, por outras palavras, soluções de (M). Seja u uma solução do problema variacional (V). Para todo o v em V, considere $w = v - u \in V$. Veja que

$$J(v) \;=\; J(u + w) \;=\; \frac{a(u,u)}{2} + \underbrace{a(u,w)}_{\substack{\| \\ \langle f, w\rangle}} + \frac{a(w,w)}{2} - \langle f, u\rangle - \langle f, w\rangle \;=\; J(u) + \frac{a(w,w)}{2}.$$

Como $a(w,w) \geq 0$, tem-se que $J(v) \geq J(u)$, como queríamos demonstrar. ■

Exercícios

1. Considere a aplicação $a : V \times V \longrightarrow \mathbb{R}$, que a cada par de elementos v_1 e v_2 em V faz corresponder o número real

$$a(v_1, v_2) \stackrel{\text{def}}{=} \langle v_1', v_2'\rangle \;=\; \int_0^1 v_1'(x)v_2'(x)\,dx.$$

   (a) Prove que a é uma aplicação ou forma bilinear, ou seja, que satisfaz as igualdades

$$\begin{aligned} a(\alpha v_1 + \beta v_2, v_3) &= \alpha a(v_1, v_3) + \beta a(v_2, v_3), \\ a(v_1, \alpha v_2 + \beta v_3) &= \alpha a(v_1, v_2) + \beta a(v_1, v_3), \end{aligned}$$

   para todos os elementos v_1, v_2 e v_3 de V e para todos os escalares reais α e β.

   (b) Mostre que a é uma forma bilinear *simétrica*: $a(v_1, v_2) = a(v_2, v_1)$ para todos os $v_1, v_2 \in V$.

(c) Diga por que é que a é uma forma bilinear *definida positiva*: $a(v,v) > 0$ para todo o $0 \neq v \in V$.

2. Dada uma função f, contínua em $[0,1]$, considere, novamente, o problema de condições de fronteira:

$$\text{encontrar} \quad u \in C^2[0,1] \quad : \quad \begin{cases} -u''(x) + u(x) \;=\; f(x) & \text{se } \; x \in (0,1), \\ u(0) \;=\; u(1) \;=\; 0. \end{cases} \tag{P'}$$

Seja (V') o problema variacional encontrado no exercício da aula anterior.

(a) Considere, agora,

$$a(v_1, v_2) \stackrel{\text{def}}{=} \langle v_1', v_2' \rangle + \langle v_1, v_2 \rangle.$$

Prove que a é uma forma bilinear, simétrica e positiva definida.

(b) Reescreva o problema variacional (V') utilizando esta aplicação bilinear.

(c) Encontre o problema de energia mínima (M') associado a (V') e prove que são equivalentes (no sentido de terem as mesmas soluções).

Aula 21: Método de Elementos Finitos para um Problema de Condições de Fronteira

Relembremos a definição do espaço vectorial V:

$$V \;=\; \{v:\; v \in C[0,1],\; v(0) = v(1) = 0,\; v' \text{ contínua por troços e limitada em } [0,1]\}\,.$$

A ideia central do método de elementos finitos para a resolução do problema de condições de fronteira (P) consiste em considerar um subespaço de V com dimensão finita.

Este subespaço de dimensão finita é construído considerando uma partição do intervalo $[0,1]$, com $n+1$ subintervalos definidos pelos nós ou extremidades

$$0 = x_0 \;<\; x_1 \;<\; \cdots \;<\; x_n \;<\; x_{n+1} = 1.$$

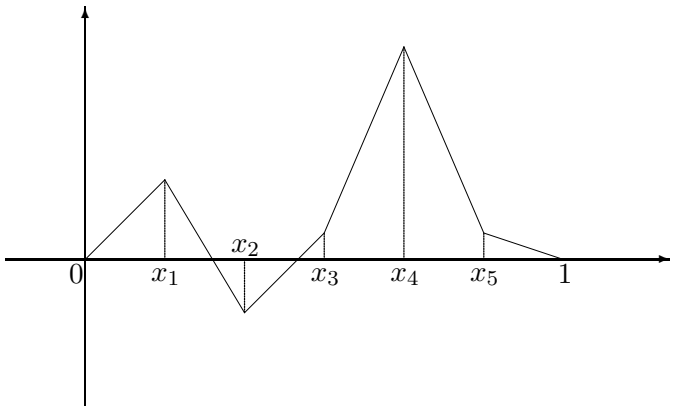
Para simplificar a apresentação, trabalharemos com subintervalos de amplitude uniforme:

$$x_{i+1} - x_i \;=\; h, \quad i = 0, \ldots, n, \qquad \text{em que} \qquad h \;=\; \frac{1}{n+1}.$$

Quanto mais pequeno h for (ou quanto maior n for) mais fina é a discretização do intervalo $[0,1]$.

Fixo um valor para h e dada a correspondente discretização do intervalo $[0,1]$, torna-se fácil identificar subespaços de V com dimensão finita. Para o efeito, basta escolher em cada subintervalo funções pertencentes a subespaços de dimensão finita, como, por exemplo, os polinómios de grau inferior ou igual a um dado número natural.

Uma escolha natural para um subespaço V_h de V, com dimensão finita, é dado pelas funções contínuas em $[0,1]$, lineares nos $n+1$ subintervalos (ou troços) de $[0,1]$.



Este subespaço V_h tem dimensão n. Em cada um dos $n+1$ troços há dois graus de liberdade, correspondentes à representação de polinómios de grau menor que ou igual a um. No entanto, é exigida a continuidade nos pontos interiores da discretização (em número igual a n) e valor zero nas extremidades 0 e 1. O resultado dá:

$$\dim(V_h) \;=\; 2(n+1) - n - 2 \;=\; n.$$

Uma base para V_h é dada pelas funções (lineares por troços) que tomam o valor 1 num nó interior e 0 nos restantes:

$$\psi_j(x_i) \;=\; \delta_{ij}$$

quaisquer que sejam $i, j \in \{1, \ldots, n\}$. É fácil verificar que estas n funções são linearmente independentes. Quando $n = 1$ e $h = 1/2$, V_h é gerado por uma única função.

Esta base de funções tem *suporte local*. A função ψ_j anula-se de 0 até x_{j-1} e de x_{j+1} até 1. O seu suporte (a região onde não se anula) vai de x_{j-1} a x_{j+1}. A derivada de ψ_j, considerada apenas no interior dos subintervalos, também se anula nos mesmos troços. As derivadas das funções da base têm, também, suporte local.

Consideremos um elemento v de V_h, escrito como combinação linear de $\psi_1, \ldots, \psi_n$

$$v(x) \;=\; \sum_{i=1}^{n} \eta_i \psi_i(x) \qquad x \in [0,1].$$

É fácil constatar que $\eta_i = v(x_i), i = 1, \ldots, n$ (o que mostra que esta combinação linear é única e que confirma o facto de $\{\psi_1, \ldots, \psi_n\}$ ser uma base de V_h).

O método de elementos finitos consiste em calcular uma aproximação para a solução u do problema de condições de fronteira (P) através da resolução do problema

$$\text{encontrar} \quad u_h \in V_h \quad \text{tal que} \quad a(u_h, v) \;=\; \langle f, v\rangle \quad \text{para todo o } v \in V_h \qquad (\mathrm{V}_h)$$

(método de Galerkin), ou do problema

$$\text{encontrar} \quad u_h \in V_h \quad \text{tal que} \quad J(u_h) \;\leq\; J(v) \quad \text{para todo o } v \in V_h \qquad (\mathrm{M}_h)$$

(método de Ritz). No contexto do problema (P), os problemas (V_h) e (M_h) são equivalentes.

Concentremo-nos na variante do método de Galerkin. Em primeiro lugar, observamos que se u_h resolve (V_h) então, em particular, u_h resolve

$$\begin{aligned} a(u_h, \psi_1) &= \langle f, \psi_1\rangle, \\ &\vdots \\ a(u_h, \psi_n) &= \langle f, \psi_n\rangle. \end{aligned}$$

Reciprocamente, suponhamos que u_h satisfaz estas n igualdades. Seja v um elemento de V_h. Escreva-se este elemento como combinação linear dos ψ's. Se multiplicarmos estas n

igualdades pelos coeficientes desta combinação linear e, depois, as somarmos, constatamos que u_h é solução de (V_h).

Agora, substituindo u_h pela sua combinação linear $u_h(x) = \sum_{i=1}^{n} \xi_i \psi_i(x)$, vem que

$$\begin{aligned} a(\psi_1,\psi_1)\xi_1 + \cdots + a(\psi_n,\psi_1)\xi_n &= \langle f, \psi_1\rangle, \\ &\vdots \\ a(\psi_1,\psi_n)\xi_1 + \cdots + a(\psi_n,\psi_n)\xi_n &= \langle f, \psi_n\rangle. \end{aligned}$$

Reescrevemos este sistema de equações lineares na forma matricial:

$$A\xi = b \iff \begin{bmatrix} a(\psi_1,\psi_1) & \cdots & a(\psi_n,\psi_1) \\ \vdots & \ddots & \vdots \\ a(\psi_1,\psi_n) & \cdots & a(\psi_n,\psi_n) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} \langle f,\psi_1\rangle \\ \vdots \\ \langle f,\psi_n\rangle \end{bmatrix}.$$

Provámos, assim, que resolver este sistema é equivalente a resolver o problema (V_h).

O suporte local das derivadas das funções da base está na génese do método de elementos finitos. A matriz A deste sistema é *esparsa*, ou seja, tem muitos elementos nulos:

$$a(\psi_i,\psi_j) = 0 \quad \text{se} \quad |i-j| > 1.$$

Os únicos elementos não nulos da matriz A são os das suas diagonal principal e subdiagonais principais:

$$a(\psi_i,\psi_i) = \int_{x_{i-1}}^{x_{i+1}} \psi_i'(x)^2\,dx = \int_{x_{i-1}}^{x_{i+1}} \frac{1}{h^2}\,dx = \frac{2}{h}$$

e

$$a(\psi_i,\psi_{i+1}) = \int_{x_i}^{x_{i+1}} \psi_i'(x)\psi_{i+1}'(x)\,dx = \int_{x_i}^{x_{i+1}} -\frac{1}{h^2}\,dx = -\frac{1}{h}.$$

A matriz $A \in \mathbb{R}^{n\times n}$ é, então, dada por

$$A = K = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & & 0 & 0 & 0 \\ 0 & -1 & 2 & & 0 & 0 & 0 \\ \vdots & & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & \ddots & 2 & -1 & 0 \\ 0 & 0 & 0 & & -1 & 2 & -1 \\ 0 & 0 & 0 & & 0 & -1 & 2 \end{bmatrix}.$$

Esta matriz é tridiagonal, simétrica e definida positiva.

A matriz K (que neste caso coincide com A) é designada por *matriz de rigidez* e o vector b por *vector de carga* — uma terminologia importada das aplicações originais do método de elementos finitos em mecânica.

Exercícios

1. Mostre que $\{\psi_1, \ldots, \psi_n\}$ é linearmente independente em V.

2. Prove que a matriz de rigidez K é definida positiva.

3. Considere o método de Ritz definido pela resolução do problema (M_h).

   (a) Mostre que este método consiste em encontrar $u_h \in V_h$ tal que o vector $[\, u_h(x_1) \ \cdots \ u_h(x_n)\,]^\top$ é solução do problema de optimização

   $$\min_{z \in \mathbb{R}^n} \ \frac{1}{2} z^\top A z - b^\top z.$$

   (b) Resolva o problema da alínea anterior e mostre que a solução (única) do método de Ritz coincide com a do método de Galerkin.

4. Considere o problema variacional (V') associado ao problema de condições de fronteira (P') mencionado nas duas últimas aulas.

   (a) Identifique o problema $(\mathrm{V'}_h)$, mostrando que a matriz A pode ser escrita na forma

   $$A \ = \ K + M,$$

   em que K é a matriz de rigidez e $M \in \mathbb{R}^{n \times n}$, designada por *matriz de massa*, é dada por

   $$M \ = \ \frac{h}{6} \begin{bmatrix} 4 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 4 & 1 & & 0 & 0 & 0 \\ 0 & 1 & 4 & & 0 & 0 & 0 \\ \vdots & & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & \ddots & 4 & 1 & 0 \\ 0 & 0 & 0 & & 1 & 4 & 1 \\ 0 & 0 & 0 & & 0 & 1 & 4 \end{bmatrix}.$$

   (b) Identifique o problema $(\mathrm{M'}_h)$ associado ao método de Ritz.

   (c) Prove (por dois processos diferentes) que os problemas $(\mathrm{V'}_h)$ e $(\mathrm{M'}_h)$ são equivalentes (no sentido de terem as mesmas soluções).

5. Prove que $\{\psi_1, \ldots, \psi_n\}$ é uma base de V_h (mostrando que é linearmente independente e que gera V_h).

6. Dada uma função f, contínua em $[0,1]$, considere o seguinte problema de condições de fronteira:

   $$\text{encontrar} \quad u \in C^4[0,1] \quad \text{tal que} \quad \begin{cases} u^{(4)}(x) \ = \ f(x) & \text{se } \ x \in (0,1), \\ u(0) \ = \ u(1) \ = \ 0, \\ u''(0) \ = \ u''(1) \ = \ 0. \end{cases}$$

(a) Mostre, através da mudança de variável $y = -u''$, que é possível reescrever a equação diferencial $u^{(4)}(x) = f(x)$ como um sistema formado pelas equações $-u''(x) = y(x)$ e $-y''(x) = f(x)$.

(b) Encontre, então, a seguinte formulação variacional para o problema de condições de fronteira (V é o espaço das funções teste usual): encontrar $u, y \in V$ tais que

$$\begin{aligned} \langle u', v' \rangle - \langle y, v \rangle &= 0 \quad \forall v \in V, \\ \langle y', v' \rangle &= \langle f, v \rangle \quad \forall v \in V. \end{aligned}$$

(c) Considere as funções teste $\psi_1, \ldots, \psi_n$ do método dos elementos finitos. Procedendo de forma análoga à deste método, encontre um sistema de equações lineares que permita calcular aproximações $u_h(x) = \sum_{i=1}^{n} \xi_i \psi_i(x)$ e $y_h(x) = \sum_{i=1}^{n} \pi_i \psi_i(x)$ para o problema variacional anterior.

Aula 22: Uma Estimativa para o Erro no Método de Elementos Finitos

Estamos interessados em estimar o erro absoluto entre a solução u_h obtida pelo método de elementos finitos e a solução u do problema de condições de fronteira (P).

Vimos que a solução u de (P) satisfaz a formulação variacional (V). Como $V_h \subset V$, vem que

$$a(u, v) \;=\; \langle f, v\rangle \quad \text{para todo o } v \in V_h.$$

Por outro lado, como também foi visto anteriormente, u_h satisfaz

$$a(u_h, v) \;=\; \langle f, v\rangle \quad \text{para todo o } v \in V_h.$$

Logo,

$$a(u - u_h, v) \;\stackrel{\text{def}}{=}\; \langle (u - u_h)', v'\rangle \;=\; 0 \quad \text{para todo o } v \in V_h.$$

Mostrámos, desta forma, que $u - u_h$ é ortogonal ao subespaço V_h para o produto interno da energia

$$\langle v_1, v_2\rangle_a \;\stackrel{\text{def}}{=}\; \langle v_1', v_2'\rangle \;=\; \int_0^1 v_1'(x) v_2'(x)\, dx.$$

Definindo a *norma da energia*

$$\|v\|_a \;\stackrel{\text{def}}{=}\; \|v'\| \;=\; \sqrt{\int_0^1 v'(x)^2\, dx},$$

temos, pelas propriedades da projecção ortogonal sobre um subespaço de dimensão finita, que

$$\|u - u_h\|_a \;\leq\; \|u - v\|_a \quad \text{para todo o } v \in V_h.$$

Formalizamos este resultado no seguinte enunciado.

Teorema 1 *Sejam u a solução de (P) e u_h a solução de (V_h). Então*

$$\|u - u_h\|_a \;\leq\; \|u - v\|_a \quad \textit{para todo o } v \in V_h.$$

A flexibilidade dada nesta estimativa permite-nos obter limites superiores para o erro na norma da energia $\|\cdot\|_a$, ao escolhermos diferentes elementos em V_h. Dentro destes, seleccionamos a função w_h em V_h que interpola a solução u nos nós $0, x_1, \ldots, x_n, 1$. É possível provar, recorrendo a argumentos interpolatórios, que

$$\|u - w_h\|_a \;\leq\; \frac{h}{\pi}\|u''\| \;=\; \frac{h}{\pi}\|f\|.$$

Aplicando o teorema anterior chegamos à nossa primeira estimativa para o erro.

Corolário 1 *Sejam u a solução de (P) e u_h a solução de (V_h). Então*

$$\|u - u_h\|_a \ \leq \ \frac{h}{\pi}\|f\|.$$

Vemos, assim, que o erro entre a solução u do problema de condições de fronteira (P) e a aproximação u_h do método de elementos finitos é da ordem de h, quando medido na norma da energia. Esta estimativa limita o erro nas derivadas $(u - u_h)'$ e é importante, em si mesma, no contexto dos problemas de aplicação. Por exemplo, nos dois exemplos que vimos anteriormente, u' representa uma deformação, desempenhando um papel mais relevante que o próprio deslocamento u.

A partir da estimativa para o erro $\|(u - u_h)'\|$ desenvolver-se-ia uma estimativa para o erro $\|u - u_h\|$ também linear em h (ver exercício). É possível, porém, derivar uma estimativa para o erro $\|u - u_h\|$ da ordem de h^2.

Teorema 2 *Sejam u a solução de (P) e u_h a solução de (V_h). Então*

$$\|u - u_h\| \ \leq \ \frac{h^2}{\pi^2}\|u''\| \ = \ \frac{h^2}{\pi^2}\|f\|.$$

Demonstração. Vamos utilizar um argumento de *dualidade*, também conhecido por *truque de Nitsche*. Considere-se o seguinte problema variacional:

$$\text{encontrar} \quad z \in V \quad \text{tal que} \quad a(z, v) \ = \ \langle u - u_h, v\rangle \quad \text{para todo o } v \in V.$$

Sendo z_h a aproximação calculada pelo método dos elementos finitos, tem-se, pelo que foi visto antes do teorema, que

$$\|z - z_h\|_a \ \leq \ \frac{h}{\pi}\|u - u_h\|.$$

Além disso, a escolha $v = u - u_h$ origina:

$$a(z, u - u_h) \ = \ \|u - u_h\|^2.$$

Vimos, no início da aula, que

$$a(v, u - u_h) \ = \ a(u - u_h, v) \ = \ 0 \quad \text{para todo o } v \in V_h.$$

Subtraindo membro a membro as duas últimas desigualdades, obtém-se

$$a(z - v, u - u_h) \ = \ \|u - u_h\|^2 \quad \text{para todo o } v \in V_h.$$

Logo, utilizando a desigualdade Cauchy-Schwartz

$$a(z - v, u - u_h) \ = \ \langle z - v, u - u_h\rangle_a \ \leq \ \|z - v\|_a\|u - u_h\|_a$$

e fazendo $v = z_h$, vem que

$$\|u - u_h\|^2 \;\leq\; \|z - z_h\|_a \|u - u_h\|_a.$$

Aplicamos, ao segundo membro, as estimativas já provadas para a aproximação do método de elementos finitos, escrevendo

$$\|u - u_h\|^2 \;\leq\; \left(\frac{h}{\pi}\|u - u_h\|\right)\left(\frac{h}{\pi}\|f\|\right).$$

■

Corremos, em MATLAB, o método dos elementos finitos para o problema de condições de fronteira (P) com $f(x) = x^2$. A solução deste problema é $u(x) = -(x^4 - x)/12$. Apresentamos, na tabela seguinte, o erro entre a aproximação u_h e a solução u, assim como o limite superior para o erro do Teorema 2 e o valor de $h^2 = 1/(n+1)^2$. O erro foi sempre inferior ao limite superior. O cálculo das normas em V foi aproximado pela fórmula trapezoidal composta $\|g\|^2 \simeq h\,(g(0)^2/2 + \sum_{i=1}^{n} g(x_i)^2 + g(1)^2/2)$.

n	$\|u - u_h\|$	$h^2\|f\|/\pi^2$	h^2
10	1.26e-004	3.77e-004	8.26e-003
100	1.49e-006	4.44e-006	9.80e-005
1000	1.52e-008	4.52e-008	9.98e-007
10000	1.52e-010	4.53e-010	1.00e-008

A matriz tridiagonal A foi armazenada de forma esparsa. Aplicou-se a fórmula trapezoidal composta para calcular cada componente do vector $b = [\,\langle f, \psi_1\rangle \;\cdots\; \langle f, \psi_n\rangle\,]^\top$.

Exercícios

1. Usando a relação, dada pelo teorema fundamental do cálculo integral,

   $$(u - u_h)(x) \;=\; \int_0^x (u - u_h)'(y)\, dy,$$

   prove que $\|(u - u_h)'\| \;\leq\; (h/\pi)\|f\|$ implica $\|u - u_h\| \;\leq\; (h/\pi)\|f\|$.

2. Calcule $\langle f, \psi_i\rangle$, $i = 1, \ldots, n$, recorrendo à fórmula trapezoidal composta.

3. Dada uma função f, contínua em $[0, 1]$, considere o problema de condições de fronteira (P), aqui descrito novamente:

   $$\text{encontrar} \quad u \in C^2[0,1] \quad \text{tal que} \quad \begin{cases} -u''(x) \;=\; f(x) & \text{se } x \in (0,1), \\ u(0) \;=\; u(1) \;=\; 0. \end{cases} \qquad \text{(P)}$$

O objectivo deste exercício é conhecer o *método das diferenças finitas* para a resolução do problema (P) e estabelecer a sua relação com o método dos elementos finitos. Considere o intervalo $[0,1]$ discretizado na forma

$$0 = x_0 \ < \ x_1 \ < \ \cdots \ < \ x_n \ < \ x_{n+1} = 1.$$

Seja u_k uma aproximação para $u(x_k)$, $k = 0, 1, \ldots, n, n+1$.

(a) Com k a variar de 1 até n, escreva uma aproximação para $u''(x_k)$ recorrendo à fórmula das diferenças centrais de segunda ordem.

(b) Tome o simétrico da aproximação obtida na alínea anterior e faça-o igual a $f(x_k)$. Reúna todas estas igualdades num sistema de equações lineares e escreva-o na sua forma matricial.

(c) Verifique que este sistema é equivalente ao que foi obtido para o método dos elementos finitos, quando se utiliza a fórmula trapezoidal composta para aproximar $\langle f, \psi_k \rangle$, $k = 1, \ldots, n$.

4. Dada uma função f, contínua em $[0,1]$, considere o seguinte problema de condições de fronteira (P):

$$\text{encontrar} \quad u \in C^2[0,1] \quad \text{tal que} \quad \begin{cases} -u''(x) \ = \ f(x) & \text{se} \ \ x \in (0,1), \\ u(0) \ = \ u(1) \ = \ 0. \end{cases} \tag{P}$$

Considere $\{\Psi_0, \Psi_1, \ldots, \Psi_n\} \subset C^2[0,1]$, com $n \in \mathbb{N}$. Tome $\bar{u}(x) = \sum_{i=0}^{n} \alpha_i \Psi_i(x)$ em que $\alpha_0, \alpha_1, \ldots, \alpha_n$ são coeficientes reais.

Este exercício pretende dar a conhecer os *métodos de colocação*.

(a) Substitua u por $\bar{u}$ na equação diferencial e diferencie.

(b) Que condições devem as funções $\Psi_0, \Psi_1, \ldots, \Psi_n$ e os coeficientes $\alpha_0, \alpha_1, \ldots, \alpha_n$ satisfazer para que $\bar{u}$ verifique as condições de fronteira?

(c) Considere, agora, $n-1$ pontos $x_1 < \cdots < x_{n-1}$ em $(0,1)$. Escreva um sistema de $n+1$ equações lineares (na forma matricial) que permita determinar $\bar{u}$ como aproximação da solução u.

(d) Faça, $\Psi_i(x) = x^i$, $i = 0, 1, \ldots, n$. Escreva a matriz do sistema da alínea anterior. Mostre que esta matriz é não singular quando $n = 3$.

Aula 23: Conceitos Básicos sobre Problemas de Valor Inicial

Um problema de valor inicial (ou problema de Cauchy) consiste em

$$\text{encontrar} \quad y \in C^1(I) \quad \text{tal que} \quad \begin{cases} y'(t) = f(t, y(t)) & \text{se } t \in I, \\ y(t_0) = y_0. \end{cases} \tag{I}$$

A função $f(t, y)$ é conhecida, bem como o instante inicial t_0 e o valor y_0 da função y em t_0. Tem-se que $y'(t) = dy/dt(t)$. É dado, também, um intervalo I, que contém o ponto t_0. Na prática, este intervalo nem sempre é especificado. A equação diferencial ordinária (de primeira ordem) $y' = f(t, y)$ descreve o declive da curva y no ponto t. Em muitos problemas de valor inicial, não é possível encontrar uma solução analítica, sendo necessário o recurso a um método numérico.

Como exemplo de um problema de valor inicial (I), tome-se

$$\begin{aligned} y' &= y \tan(t+3), \\ y(-3) &= -1, \end{aligned}$$

em que $f(t, y) = y \tan(t+3)$, $t_0 = -3$ e $y_0 = -1$. A solução analítica deste problema é $y(t) = \sec(t+3)$ para $-\pi/2 < t + 3 < \pi/2$.

As questões que se colocam, tradicionalmente, perante um problema deste tipo são saber se tem solução e se uma solução (quando existe) é única. É necessário impor condições a f para que um problema de valor inicial tenha solução e, mesmo assim, só é possível garantir existência numa vizinhança de t_0. Veja-se, por exemplo, o problema (I) em que

$$\begin{aligned} y' &= 1 + y^2, \\ y(0) &= 0. \end{aligned}$$

A solução y começará em $t = 0$ com $y'(0) = 1$. Quer a solução y quer a sua derivada y' são funções crescentes de $t > 0$. Logo, dada a expressão da equação diferencial, haverá, forçosamente, um valor finito para t a partir do qual não existirá solução. A solução para este problema é $y(t) = \tan(t)$ e esse valor de t é $\pi/2$. Em geral, só se consegue garantir a existência de solução numa vizinhança de t_0.

Teorema 1 *Seja f uma função contínua num rectângulo centrado em (t_0, y_0),*

$$R = \{(t, y) : |t - t_0| \leq \alpha, \ |y - y_0| \leq \beta\}$$

(com α e β números reais positivos). Então o problema de valor inicial (I) tem uma solução $y(t)$ para t a satisfazer

$$|t - t_0| \leq \min\left\{\alpha, \frac{\beta}{M}\right\},$$

em que M é o valor máximo que f atinge em R.

Um problema de valor inicial pode não ter solução única mesmo nas condições do teorema anterior. Um exemplo deste fenómeno é:

$$\begin{aligned} y' &= y^{\frac{2}{3}}, \\ y(0) &= 0. \end{aligned}$$

Verifica-se, facilmente, que existem duas soluções $y(t) = 0$ e $y(t) = t^3/27$. Para garantir a unicidade é necessário impor maior suavidade a f.

Teorema 2 *Sejam f e $\partial f/\partial y$ funções contínuas num aberto contendo o rectângulo R. Então o problema de valor inicial (I) tem uma solução única $y(t)$ no intervalo definido no teorema anterior.*

Uma alternativa a estes dois resultados é o seguinte teorema.

Teorema 3 *Se f for contínua para (t, y) no rectângulo R e se existir um escalar real $L > 0$ tal que*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|,$$

para quaisquer (t, y_1) e (t, y_2) em R, então o problema de valor inicial (I) tem uma única solução $y(t)$, com t a satisfazer

$$|t - t_0| \leq \min\left\{\alpha, \frac{\beta}{M}, \frac{1}{L}\right\},$$

em que M é o valor máximo que f atinge em R.

Este último teorema impõe à função contínua f apenas a continuidade à Lipschitz relativamente ao seu segundo argumento.

As soluções mencionadas nestes teoremas são restringidas a um domínio *local*. A existência de uma solução *global* única seria garantida alargando a continuidade à Lipschitz de f (relativamente a y) a todo o $I \times \mathbb{R}$.

Um problema de valor inicial também pode ser governado por um sistema de equações diferenciais ordinárias (de primeira ordem):

$$\text{encontrar} \quad y_1, \ldots, y_n \in C^1(I) \quad \text{tais que} \quad \begin{cases} y_1'(t) = f_1(t, y_1(t), \ldots, y_n(t)) & \text{se } t \in I, \\ \qquad \vdots \\ y_n'(t) = f_n(t, y_1(t), \ldots, y_n(t)) & \text{se } t \in I, \\ y_1(t_0) = y_0^1, \ldots, y_n(t_0) = y_0^n, \end{cases}$$

em que $f_1, \ldots, f_n$ são funções reais de $n+1$ variáveis reais. Neste caso, existem n funções $y_1, \ldots, y_n$ a determinar com base em n condições iniciais. Este problema pode ser escrito na forma vectorial:

$$\text{encontrar} \quad Y \in C^1(I) \quad \text{tal que} \quad \begin{cases} F'(t) = F(t, Y(t)) & \text{se } t \in I, \\ Y(t_0) = Y_0, \end{cases}$$

em que $F : \mathbb{R}^{n+1} \to \mathbb{R}^n$ tem componentes $f_1, \ldots, f_n$ e $Y_0 \in \mathbb{R}^n$. A título ilustrativo considere-se o problema de valor inicial vectorial definido por:

$$\begin{aligned} y_1'(t) &= y_1 + 4y_2 + e^t, \\ y_2'(t) &= y_1 + y_2 + 2e^t, \\ y_1(0) = 4 &\text{ e } y_2(0) = 5/4, \end{aligned}$$

em que $Y(t) = [\, y_1(t)\ y_2(t)\,]^\top$, $f_1(t, y_1, y_2) = y_1 + 4y_2 - e^t$, $f_2(t, y_1, y_2) = y_1 + y_2 + 2e^t$ e $Y_0 = [\, 4\ \ 5/4\,]^\top$. A solução deste problema é a seguinte:

$$y_1(t) = 4e^{3t} + 2e^{-t} - 2e^t \quad \text{e} \quad y_2(t) = 2e^{3t} - e^{-t} - e^t/4.$$

Os métodos numéricos analisados nas próximas aulas são descritos recorrendo ao problema de valor inicial escalar (I), mas a sua generalização à forma vectorial não apresenta qualquer obstáculo.

Um problema de valor inicial pode ser governado, também, por uma equação diferencial ordinária de ordem n, escrevendo-se como

$$\text{encontrar} \quad z \in C^n(I) \quad \text{tal que} \quad \begin{cases} z^{(n)}(t) = f(t, z(t), z'(t), \ldots, z^{(n-1)}(t)) & \text{se } t \in I, \\ z(t_0) = z_0, z'(t_0) = z_0^1, \ldots, z^{(n-1)}(t_0) = z_0^{n-1}. \end{cases}$$

Verifica-se, facilmente, que este problema é equivalente a um problema de valor inicial vectorial com

$$F(t, Y) = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ f(t, y_1, y_2, \ldots, y_n) \end{bmatrix} \quad \text{e} \quad Y_0 = \begin{bmatrix} z_0 \\ z_0^1 \\ \vdots \\ z_0^{n-1} \end{bmatrix},$$

através das mudanças de variável $y_1 = z, y_2 = z', \ldots, y_n = z^{(n-1)}$.

Exercícios

1. Mostre que $y(t) = -t^2/4$ e $y(t) = 1 - t$ são soluções do problema de valor inicial definido por:

$$\begin{aligned} y' &= \tfrac{1}{2}\left(\sqrt{t^2 + 4y} - t\right), \\ y(2) &= -1. \end{aligned}$$

   Por que é que este facto não contradiz os teoremas de unicidade dos problemas de valor inicial?

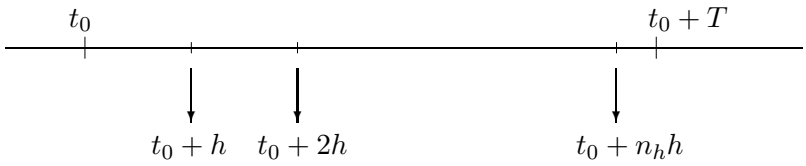
2. Escreva, na forma vectorial, o seguinte problema de valor inicial governado por uma equação diferencial ordinária de ordem 3:

$$\begin{aligned} &\operatorname{sen}(t)z^{(3)} + \cos(tz) + \operatorname{sen}(t^2 + z'') + (z')^3 \;=\; \log(t), \\ &z(2) = 7,\; z'(2) = 3,\; z''(2) = -4. \end{aligned}$$

Aula 24: Introdução aos Métodos Numéricos para Problemas de Valor Inicial

Um método numérico aplicado a um problema de valor inicial (I) gera uma sucessão de pares ordenados de números reais da forma $(t_0, u_0), (t_1, u_1), (t_2, u_2), \ldots$, em que o valor de u_k constitui uma aproximação para $y(t_k)$.

Vamos supor que o intervalo de integração é dado por $I = [t_0, t_0+T]$, onde $T \in (0, +\infty)$ mede o tempo de integração. Os nós da discretização são dados por $t_k = t_0 + kh$, em que h é o tamanho do passo da discretização. O índice k percorre os inteiros de 0 até n_h, com n_h o maior inteiro para o qual $t_0 + n_h h \leq t_0 + T$.



Estudaremos, neste curso, alguns dos *métodos de passo simples* mais conhecidos. Estes métodos caracterizam-se pelo facto de u_{k+1} depender apenas de u_k (de entre todos os valores $u_0, \ldots, u_k$). Quando u_{k+1} depende de u_k e de valores anteriores a u_k, dizemos que estamos na presença de *métodos de passos múltiplos* (ou *métodos multipasso*).

Uma das formas mais simples de desenvolver métodos de passo simples consiste em aplicar a fórmula de Taylor. Suponhamos, por exemplo, que $y(t)$ tem derivadas contínuas até à ordem 4. Assim,

$$y(t+h) \;=\; y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{3!}y^{(3)}(t) + \frac{h^4}{4!}y^{(4)}(t+\sigma h),$$

com $\sigma \in (0,1)$. Logo,

$$y(t+h) \;\simeq\; y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{3!}y^{(3)}(t).$$

Os valores para $y'(t), y''(t)$ e $y^{(3)}(t)$ podem ser obtidos através de $y'(t) = f(t, y(t))$. Por exemplo, quando

$$f(t, y(t)) \;=\; \cos(y(t)) + t^2$$

obter-se-ia

$$\begin{array}{lcll} y'(t) & = & f(t,y(t)) & = t^2 + \cos(y(t)), \\ y''(t) & = & f_t(t,y(t)) + f_y(t,y(t))y'(t) & = 2t - \mathrm{sen}(y(t))y'(t), \\ y^{(3)}(t) & = & \cdots & = 2 - y''(t)\mathrm{sen}(y(t)) - (y'(t))^2\cos(y(t)). \end{array}$$

Note-se que as derivadas parciais de f são tomadas em relação aos seus dois argumentos: $f(t,y) = \cos(y) + t^2$, $f_t(t,y) = 2t$, $f_y(t,y) = -\text{sen}(y)$.

Descrevemos, de seguida, o método de Taylor de ordem 3 para este problema. São dados do problema, para além da função f, os valores de t_0 e de y_0.

Método de Taylor de ordem 3 **para** $y'(t) = f(t, y(t)) = \cos(y(t)) + t^2,\ y(t_0) = y_0$
Fixar $T > 0$. Escolher $h > 0$. Determinar n_h. Fazer $u_0 = y_0$.
Para $k = 0, 1, \ldots, n_h - 1$

1. Fazer $u'_k = t_k^2 + \cos(u_k)$.
2. Fazer $u''_k = 2t_k - \text{sen}(u_k)u'_k$.
3. Fazer $u_k^{(3)} = 2 - u''_k\text{sen}(u_k) - (u'_k)^2\cos(u_k)$.
4. Fazer $u_{k+1} = u_k + h\left(u'_k + \frac{h}{2}\left(u''_k + \frac{h}{3}u_k^{(3)}\right)\right)$.
5. Fazer $t_{k+1} = t_k + h$.

Os métodos de Taylor são, conceptualmente, muito simples. É possível obter métodos de Taylor de *ordem*[4] muito elevada. No entanto, estes métodos requerem as derivadas parciais de $f(t,y)$ em ordem a t e a y. Ora, estas derivadas parciais podem não existir, ou podem existir e não estar disponíveis. No caso de não estarem disponíveis, é sempre possível recorrer à *diferenciação simbólica* ou à *diferenciação automática*, consoante se conheça, respectivamente, a expressão analítica de f ou um código fonte que a implemente. Existem, porém, outros métodos mais adequados a estas situações.

O método de Taylor de ordem 1 é conhecido por método de Euler (explícito ou progressivo).

Método de Euler Explícito para $y'(t) = f(t, y(t)),\ y(t_0) = y_0$
Fixar $T > 0$. Escolher $h > 0$. Determinar n_h. Fazer $u_0 = y_0$.
Para $k = 0, 1, \ldots, n_h - 1$

1. Fazer $u_{k+1} = u_k + hf(t_k, u_k)$.
2. Fazer $t_{k+1} = t_k + h$.

Todo o método de Taylor é *explícito*, uma vez que a expressão para u_{k+1} aparece dada *explicitamente* em função de $u_0, u_1, \ldots, u_k$. Um método de passo simples que seja explícito envolve, necessariamente, uma actualização do tipo

$$u_{k+1} \ = \ u_k + h\Phi(t_k, u_k, f_k; h).$$

[4] O conceito de ordem para um método numérico aplicado a (I) será definido, rigorosamente, mais à frente.

A função Φ é designada por *função incremental*. No caso do método de Euler, temos que $\Phi(t_k, u_k, f_k; h) = f_k$, com $f_k = f(t_k, u_k)$.

Um dos *métodos implícitos* mais simples é o método de Euler implícito ou regressivo. Neste método, o valor de u_{k+1} é solução da equação não linear

$$u_{k+1} = u_k + hf(t_{k+1}, u_{k+1}).$$

A determinação de u_{k+1} envolve a resolução desta equação e está dependente da forma como a função f é apresentada num dado problema. No exemplo anterior, esta equação toma o aspecto

$$u_{k+1} = u_k + h\left(\cos(u_{k+1}) + t_{k+1}^2\right),$$

em que $t_{k+1} = t_k + h$. Nesta equação, tudo é conhecido (u_k, h e t_{k+1}) menos u_{k+1}.

Nem sempre a resolução de $u_{k+1} = u_k + hf(t_{k+1}, u_{k+1})$ pode ser feita analiticamente, podendo ser necessária a aplicação de um método numérico (por exemplo, o método de Newton). A análise de métodos implícitos está fora do âmbito deste curso.

Exercícios

1. Escreva os passos do método de Taylor de ordem três aplicado ao problema de valor inicial dado por:
$$y' = y^2 + ye^t,$$
$$y(0) = 1.$$

2. Mostre que a aplicação de diferenças *progressivas* a $y'(t_k)$ permite escrever a fórmula de actualização do método de Euler explícito (ou *progressivo*).

3. Mostre que com a aplicação de diferenças *regressivas* a $y'(t_{k+1})$ se chega à fórmula de actualização do método de Euler implícito (ou *regressivo*).

Aula 25: Consistência, Estabilidade-Zero, Convergência

Consideremos, como base de trabalho, a fórmula de actualização de um método explícito e de passo simples, definida à custa da função incremental Φ e dada por

$$u_{k+1} \;=\; u_k + h\Phi(t_k, u_k, f(t_k, u_k); h), \quad 0 \leq k \leq n_h - 1, \quad u_0 \;=\; y_0.$$

Seja y a solução do problema de valor inicial (I). Faça-se $y_k = y(t_k)$ para todo o k. Se substituirmos u_{k+1} e u_k por, respectivamente, y_{k+1} e y_k, nas expressões anteriores, a fórmula deixa de ser necessariamente verdadeira. Seja $R_{k+1}(h)$ o erro residual ocorrido. Tem-se que

$$y_{k+1} \;=\; y_k + h\Phi(t_k, y_k, f(t_k, y_k); h) + R_{k+1}(h), \quad 0 \leq k \leq n_h - 1.$$

À quantidade $R_{k+1}(h)$ chama-se *erro de truncatura local* do método numérico. O *erro de truncatura global* é dado por

$$R(h) \;\stackrel{\text{def}}{=}\; \max_{0 \leq k \leq n_h - 1} |R_{k+1}(h)|.$$

Faz sentido exigir que $R(h)$ convirja para zero quando h tende para zero. Tomando limites e assumindo Φ limitada quando $h \to 0^+$, tal exigência traduziria a continuidade da solução y. Como veremos de seguida, é apropriado exigir mais dos erros de truncatura.

Note-se que a solução do problema (I) satisfaz, sob suavidade apropriada,

$$y(t_{k+1}) \;=\; y(t_k) + hy'(t_k) + \frac{h^2}{2} y''(t_k + \sigma_k h),$$

com $\sigma_k \in (0, 1)$. Numa notação diferente, tem-se que

$$y_{k+1} \;=\; y_k + hf(t_k, y_k) + \mathcal{O}(h^2),$$

com $\mathcal{O}(h^2)$ igual a uma constante (apenas dependente do valor máximo de y'' em I) vezes h^2. Comparando esta expressão para y_{k+1} com a expressão que envolve o erro de truncatura local, concluímos que faz sentido exigir que $R(h)/h$ convirja para zero quando h tende para zero. Dizemos, desta forma, que um método numérico (explícito e de passo simples) é *consistente com o problema de valor inicial (I)* se a sua solução y satisfizer

$$\lim_{h \to 0^+} \frac{R(h)}{h} \;=\; 0.$$

No método de Euler explícito, vimos que $\Phi(t_k, y_k, f(t_k, y_k); h) = f(t_k, y_k)$. É fácil mostrar que um método (explícito e de passo simples) é consistente se e só se

$$\lim_{h \to 0^+} \max_{0 \leq k \leq n_h - 1} |\Phi(t_k, y_k, f(t_k, y_k); h) - f(t_k, y_k)| \;=\; 0.$$

Um método (explícito e de passo simples) diz-se consistente de ordem p se a solução y de (I) satisfaz $R(h)/h \leq \mathcal{C}h^p$, com $\mathcal{C}$ uma constante real e independente de h.

Uma outra propriedade importante de um método numérico para problemas de valor inicial (I) é a estabilidade–zero. Diz-se que um método numérico (explícito e de passo simples) é *estável–zero* se, para qualquer $\epsilon > 0$, existir um $\bar{h} > 0$ e uma constante C (independente de h, n_h e ϵ) tais que, para todo o $h \in (0, \bar{h}]$,

$$|z_k - u_k| \ \leq \ C\epsilon, \quad 0 \leq k \leq n_h,$$

em que z_k se obtém perturbando a fórmula de actualização de u_k da forma

$$z_{k+1} \ = \ z_k + h\left[\Phi(t_k, z_k, f(t_k, z_k); h) + \delta_{k+1}\right], \quad 0 \leq k \leq n_h - 1, \quad z_0 \ = \ y_0 + \delta_0$$

e $\epsilon > 0$ é um limite superior para o tamanho da perturbação ($\delta_k \leq \epsilon$, $k = 0, 1, \ldots, n_h$). Um método que seja estável–zero é menos sensível à acumulação de erros de arredondamento na aplicação da sua fórmula. A designação estabilidade–*zero* provém da imposição de $|z_k - u_k| \leq C\epsilon$ para todo o h suficientemente perto de *zero*.

Finalmente, dizemos que um método numérico para problemas de valor inicial (I) é convergente se

$$|y_k - u_k| \ \leq \ i(h), \quad 0 \leq k \leq n_h,$$

em que $i(h)$ converge para zero quando h tende para 0 (diz-se que $i(h)$ é um infinitesimal em h). Um método diz-se convergente de ordem p se $i(h) = \mathcal{C}h^p$, com $\mathcal{C}$ uma constante real e independente de h (ou de n_h).

A teoria dos métodos numéricos para problemas de valor inicial estrutura-se em torno das duas seguintes propriedades (que apresentamos sem demonstração).

- Se a função Φ for contínua à Lipschitz em relação ao seu argumento u_k, então o método é estável-zero.
- Um método é convergente se e só se for consistente e estável–zero.

Deste modo, para provarmos que um método é convergente é suficiente mostrar que o método é consistente e que Φ é contínua à Lipschitz em relação ao seu argumento u_k.

Exemplifiquemos estas noções com o método de Euler (explícito ou progressivo). Sabemos que este método é consistente (pois satisfaz, trivialmente, a condição necessária e suficiente de consistência dada anteriormente). Além disso, tem-se que

$$|\Phi(t_k, u_k^1, f(t_k, u_k^1); h) - \Phi(t_k, u_k^2, f(t_k, u_k^2); h)| \ = \ |f(t_k, u_k^1) - f(t_k, u_k^2)|.$$

Logo, a continuidade à Lipschitz de f em relação ao seu segundo argumento (a tal condição para que o problema (I) tenha solução única) implica a continuidade à Lipschitz de Φ em relação ao seu argumento u_k. Assim sendo, o método de Euler explícito é estável–zero e, consequentemente, convergente.

Analisemos, agora, as ordens de consistência e convergência que ocorrem com o método de Euler explícito. Vimos que o erro de truncatura local pode ser expresso na forma

$$R_{k+1}(h) \;=\; \frac{h^2}{2} y''(t_k + \sigma_k h).$$

Logo,

$$\frac{R(h)}{h} \;\leq\; \left(\frac{1}{2} \max_{t\in I} |y''(t)|\right) h,$$

de onde se conclui que o método de Euler explícito tem ordem de consistência igual a 1.

Para estudarmos a ordem de convergência, começemos por notar que

$$y_{k+1} - u_{k+1} \;=\; \underbrace{(y_{k+1} - [y_k + hf(t_k, y_k)])}_{\shortparallel\; R_{k+1}(h)} - (u_{k+1} - [y_k + hf(t_k, y_k)]).$$

Mas $u_{k+1} - [y_k + hf(t_k, y_k)] = u_k - y_k + h(f(t_k, u_k) - f(t_k, y_k))$. Com o auxílio da notação $e_k \stackrel{\text{def}}{=} y_k - u_k$ e da continuidade à Lipschitz de f em relação ao seu segundo argumento, obtemos

$$|e_{k+1}| \;\leq\; R_{k+1}(h) + |e_k| + hL|e_k| \;\leq\; R(h) + (1 + hL)|e_k|.$$

Aplicando recursivamente esta desigualdade e somando a soma geométrica resultante dá origem a

$$|e_{k+1}| \;\leq\; \frac{(1 + Lh)^{k+1} - 1}{L} \frac{R(h)}{h}.$$

Daqui resulta que

$$|e_{k+1}| \;\leq\; \frac{(1 + Lh)^{k+1} - 1}{2L} \max_{t\in I} |y''(t)|\, h.$$

Como, $1 + hL \leq e^{hL}$ e $(k + 1)h = t_{k+1} - t_0$, chegamos a

$$|e_{k+1}| \;\leq\; \left(\frac{e^{L(t_{k+1} - t_0)} - 1}{2L} \max_{t\in I} |y''(t)|\right) h.$$

Finalmente, vem que

$$|e_k| \;\leq\; \left(\frac{e^{L\,T} - 1}{2L} \max_{t\in I} |y''(t)|\right) h, \quad 0 \leq k \leq n_h,$$

o que demonstra que o método de Euler explícito tem uma ordem de convergência igual a 1.

Exercícios

1. Mostre que um método da forma $u_{k+1} = u_k + h\Phi(t_k, u_k, f(t_k, u_k); h)$, $0 \leq k \leq n_h - 1$, $u_0 = y_0$ é consistente se e só se

$$\lim_{h \to 0^+} \max_{0 \leq k \leq n_h - 1} |\Phi(t_k, y_k, f(t_k, y_k); h) - f(t_k, y_k)| \ = \ 0.$$

2. Siga as sugestões dadas para mostrar que $|e_{k+1}| \ \leq \ \frac{(1+Lh)^{k+1}-1}{Lh} R(h)$.

3. Considere o problema de valor inicial definido por

$$\begin{aligned} y'(t) \ &= \ y(t) \qquad \text{se } \ t > 0, \\ y(0) \ &= \ 1, \qquad T \ = \ 1. \end{aligned}$$

   (a) Determine a constante que multiplica h na expressão do limite superior para o erro $|e_k|$ do método de Euler explícito.

   (b) Calcule uma expressão exacta para este erro. Faça $h = 0.1$ e veja quão realista foi o limite superior obtido anteriormente.

4. Considere o método de Euler modificado (a ser motivado mais à frente), definido por:

$$u_{k+1} \ = \ u_k + hf(t_k + h/2, u_k + hf(t_k, u_k)/2), \quad 0 \leq k \leq n_h - 1, \quad u_0 \ = \ y_0.$$

   Pode assumir que a função f é contínua à Lipschitz relativamente ao seu segundo argumento (com constante $L > 0$).

   (a) Identifique a função incremental $\Phi(t_k, u_k, f(t_k, u_k); h)$.

   (b) Mostre que o método é consistente com o problema de valor inicial.

   (c) Mostre que a função incremental é contínua à Lipschitz relativamente a u_k.

   (d) Prove que o método é convergente.

Aula 26: Métodos de Runge–Kutta para Problemas de Valor Inicial

Como vimos anteriormente, os métodos de Taylor requerem, para poderem ser aplicados, o cálculo de derivadas de f. Os métodos de Runge–Kutta evitam esta dificuldade, tentando reproduzir o efeito dos métodos de Taylor à custa de combinações sofisticadas de valores de f.

Na derivação que se faz de seguida omitem-se os argumentos t e y para simplificar a notação das expressões. Como $y' = f = f(t, y)$, vem que

$$\begin{aligned} y'' &= f_t + f_y y' = f_t + f_y f, \\ y^{(3)} &= f_{tt} + f_{ty} f + f_y (f_t + f_y f) + (f_{yt} + f_{yy} f) f. \end{aligned}$$

Um método de Taylor de ordem 2 baseia-se na expansão

$$\begin{aligned} y(t+h) &= y(t) + h y'(t) + \tfrac{h^2}{2} y''(t) + E_2^y(h) \\ &= y + hf + \tfrac{h^2}{2}(f_t + f_y f) + E_2^y(h) \\ &\simeq y + hf + \tfrac{h^2}{2}(f_t + f_y f). \end{aligned}$$

É possível reescrever esta expansão na forma

$$y(t+h) \simeq y + \frac{h}{2} f + \frac{h}{2}(f + [h] f_t + [hf] f_y).$$

A expressão entre parêntesis curvos corresponde a uma expansão de Taylor de f, centrada em (t, y) e ao longo de (h, hf), o que permitirá contornar a utilização das derivadas parciais f_t e f_y.

Assim sendo, vamos considerar uma segunda expansão de Taylor:

$$\begin{aligned} f(t+h, y+hf) &= f(t,y) + [h] f_t(t,y) + [hf] f_y(t,y) + E_1^f(h) \\ &= f + [h] f_t + [hf] f_y + E_1^f(h). \end{aligned}$$

Aplicou-se a fórmula de Taylor, de ordem 1, com resto de Lagrange, à função f (como função de duas variáveis), centrada em (t, y) e ao longo de (h, hf).

Se, agora, desprezarmos o erro $E_1^f(h)$, obtemos

$$\begin{aligned} y(t+h) &\simeq y + \tfrac{h}{2} f + \tfrac{h}{2} f(t+h, y+hf) \\ &= y(t) + \tfrac{h}{2} f(t, y(t)) + \tfrac{h}{2} f(t+h, y(t) + hf(t, y(t))). \end{aligned}$$

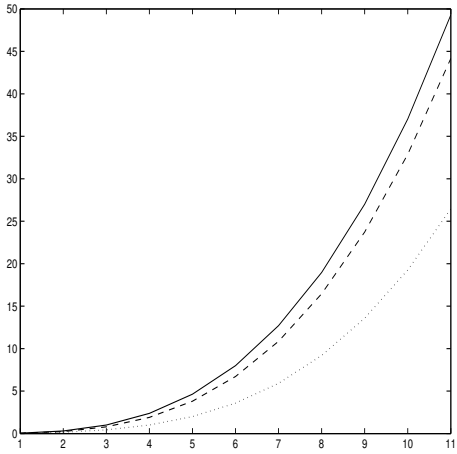
Surge, desta forma, um dos métodos de Runge–Kutta de ordem 2, conhecido por método de Heun, e que passamos a descrever de seguida.

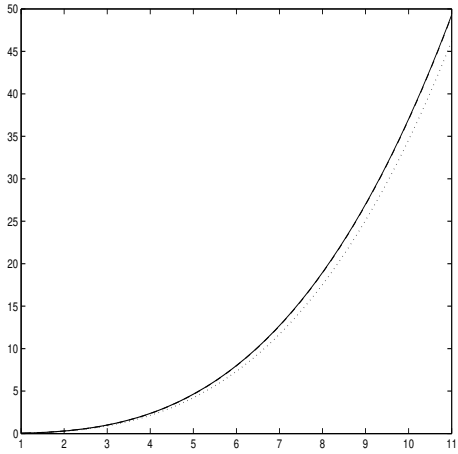
Método de Heun para $y'(t) = f(t, y(t)),\ y(t_0) = y_0$
Fixar $T > 0$. Escolher $h > 0$. Determinar n_h. Fazer $u_0 = y_0$.
Para $k = 0, 1, \ldots, n_h - 1$

1. Fazer $F_k^1 = hf(t_k, u_k)$ e $F_k^2 = hf(t_k + h, u_k + F_k^1)$.
2. Fazer $u_{k+1} = u_k + \frac{1}{2}(F_k^1 + F_k^2)$.

3. Fazer $t_{k+1} = t_k + h$.

Nos gráficos reproduzidos em baixo, ilustramos o desempenho numérico do método de Euler explícito (a ponteado) e do método de Heun (a tracejado), para o problema de valor inicial em que $f(t, y) = y^{2/3}$ e $t_0 = 1$. A integração numérica foi feita por um período de tempo $T = 10$. A solução é descrita a traço contínuo. A figura da esquerda corresponde a $h = 1$ e a da direita a $h = 0.1$. É bem patente o melhor comportamento numérico do método de ordem dois (Heun) em relação ao de ordem um (Euler explícito).





A família dos métodos de Runga–Kutta de ordem 2 pode ser obtida introduzindo parâmetros, na expressão

$$y + \frac{h}{2} f + \frac{h}{2}(f + h f_t + h f f_y),$$

da forma

$$y + w_1 h f + w_2 h(f + \alpha h f_t + \beta h f f_y).$$

Constata-se, imediatamente, que os parâmetros w_1, w_2, α e β têm de satisfazer o sistema de equações não lineares

$$w_1 + w_2 \;=\; 1, \quad w_2\alpha \;=\; \frac{1}{2} \quad \text{e} \quad w_2\beta \;=\; \frac{1}{2}.$$

A solução $w_1 = w_2 = 1/2$ e $\alpha = \beta = 1$ origina o método de Heun. Um outro conhecido método de Runge–Kutta de ordem 2 é o método de Euler modificado, que se obtém através da solução $w_1 = 0$, $w_2 = 1$ e $\alpha = \beta = 1/2$. Este método também é conhecido por método de Euler-Cauchy.

Método de Euler Modificado para $y'(t) = f(t, y(t)),\ y(t_0) = y_0$
Fixar $T > 0$. Escolher $h > 0$. Determinar n_h. Fazer $u_0 = y_0$.
Para $k = 0, 1, \ldots, n_h - 1$

1. Fazer $F_k^1 = hf(t_k, u_k)$ e $F_k^2 = hf(t_k + h/2, u_k + F_k^1/2)$.
2. Fazer $u_{k+1} = u_k + F_k^2$.
3. Fazer $t_{k+1} = t_k + h$.

Os cálculos associados à derivação de métodos de Runge–Kutta de ordem elevada são bastante complicados. As fórmulas finais, no entanto, são elegantes e simples de programar.

Exercícios

1. Identifique a função incremental $\Phi(t_k, u_k, f(t_k, u_k); h)$ nos métodos de Heun e de Euler modificado.

2. O método de Heun está associado ao método trapezoidal (também conhecido por método de Crank–Nicolson), cuja fórmula de actualização é dada por

$$u_{k+1} \;=\; u_k + \frac{h}{2}\left[f(t_k, u_k) + f(t_{k+1}, u_{k+1})\right].$$

   Mostre que esta fórmula resulta da aplicação da fórmula de quadratura trapezoidal ao integral

$$y(t) - y(t_k) \;=\; \int_{t_k}^{t} f(\tau, y(\tau))\, d\tau.$$

   Este método é explícito ou implícito?

3. Mostre que a fórmula de actualização do método de Heun pode ser obtida da do método trapezoidal, substituindo $f(t_{k+1}, u_{k+1})$ por $f(t_{k+1}, u_k + hf(t_k, u_k))$, ou seja, utilizando a fórmula de actualização do método de Euler explícito para aproximar u_{k+1}. (O método de Heun pode ser encarado como um processo de tornar explícito o método trapezoidal.)

4. Prove que o método de Heun tem um erro de truncatura local de ordem 2:

$$\frac{R_{k+1}(h)}{h} \;=\; \mathcal{O}(h^2).$$

   Decomponha, primeiro, o erro $R_{k+1}(h)$ na soma dos erros

$$S_1^k(h) \;=\; y_{k+1} - y_k - \frac{h}{2}\left[f(t_k, y_k) + f(t_{k+1}, y_{k+1})\right]$$

   e

$$S_2^k(h) \;=\; \frac{h}{2}\left[f(t_{k+1}, y_{k+1}) - f(t_k + h, y_k + hf(t_k, y_k))\right].$$

   Ao erro $S_1^k(h)$ aplique o que conhece sobre o erro da fórmula de quadratura trapezoidal. O segundo erro $S_2^k(h)$, analise-o à luz do que sabe sobre o erro de truncatura local do método de Euler explícito.

Aula 27: Estabilidade Absoluta de Métodos Numéricos para Problemas de Valor Inicial

Com a estabilidade absoluta pretende-se averiguar se uma aproximação u_k, para valores de h constantes, permanece limitada quando t_k tende para $+\infty$. A estabilidade absoluta de um método diz respeito ao seu comportamento assimptótico no tempo, o que contrasta com a estabilidade–zero, que mede, no intervalo de integração dado, o comportamento da aproximação u_k sob perturbação da sua fórmula de actualização.

A definição de estabilidade absoluta está associada ao *problema teste*:

$$\text{encontrar} \quad y \in C^1(I) \quad \text{tal que} \quad \begin{cases} y'(t) \;=\; \lambda y(t) & \text{se } \; t > 0, \\ y(0) \;=\; 1. \end{cases} \qquad (\mathrm{I}_\lambda)$$

A solução deste problema de Cauchy linear é $y(t) = e^{\lambda t}$. O parâmetro λ pode tomar valores complexos. Note-se que $\lim_{t\to+\infty} |y(t)| = 0$ se $\mathrm{Re}(\lambda) < 0$, ou seja, se λ está no semiplano

$$\mathbb{C}^- = \{z \in \mathbb{C} : \; \mathrm{Re}(z) < 0\}.$$

Escolhido um método numérico para a resolução aproximada de (I_λ), a aproximação u_k gerada vai depender de h e de λ. Diz-se que um método é absolutamente estável para um dado valor de h se

$$|u_k| \;\longrightarrow 0 \quad \text{quando} \quad t_k \;\longrightarrow +\infty.$$

A *região de estabilidade absoluta* de um método é o subconjunto do plano complexo

$$\mathcal{A} \;=\; \left\{\hat{h} = h\lambda \in \mathbb{C} : \quad |u_k| \;\longrightarrow 0 \quad \text{quando} \quad t_k \;\longrightarrow +\infty\right\}.$$

Vejamos, primeiro, o caso do método de Euler explícito. Como $u_{k+1} = u_k + h\lambda u_k$ e $u_0 = 1$, vem que

$$u_k \;=\; (1 + h\lambda)^k, \qquad k \geq 0.$$

Assim sendo, o método de Euler explícito é absolutamente estável se e só se $|1+h\lambda| < 1$, ou seja, se e só se $\hat{h} = h\lambda$ estiver no interior de um círculo do plano complexo de raio unitário e centro $(-1, 0)$. A sua região de estabilidade absoluta é $\mathcal{A} = \{\hat{h} \in \mathbb{C} : \; |1 + \hat{h}| < 1\}$.

É possível identificar o valor máximo de h a partir do qual o método de Euler explícito é absolutamente instável. De facto, mostra-se (ver exercício) que $|1 + h\lambda| < 1$, com h real positivo, é equivalente a

$$h\lambda \;\in\; \mathbb{C}^- \qquad \text{e} \qquad 0 \;<\; h \;<\; -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}.$$

O valor máximo de h procurado é dado por $-2\mathrm{Re}(\lambda)/|\lambda|^2$.

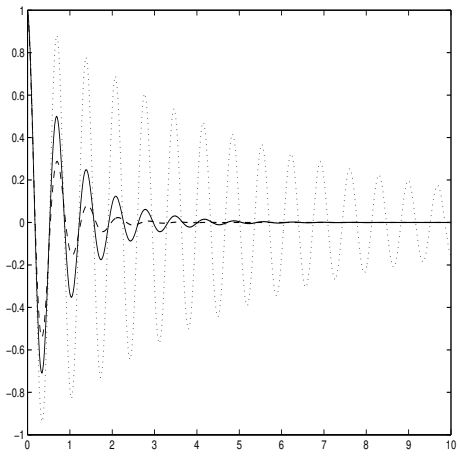
No caso do método de Euler implícito, tem-se que $u_{k+1} = u_k + h\lambda u_{k+1}$ e $u_0 = 1$, o que resulta em

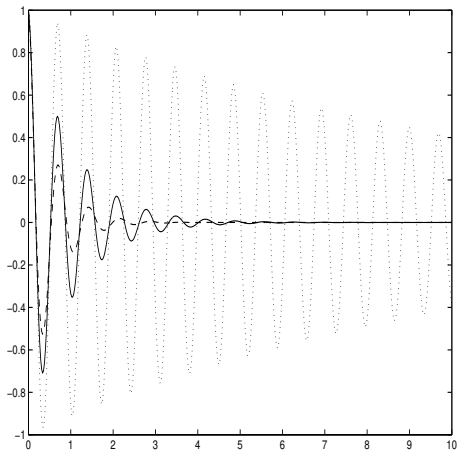
$$u_k = \frac{1}{(1-h\lambda)^k}, \qquad k \geq 0.$$

A região de estabilidade absoluta deste método é todo o plano complexo menos o círculo centrado em $(1, 0)$, de raio 1. A sua região de estabilidade absoluta contém $\mathbb{C}^-$.

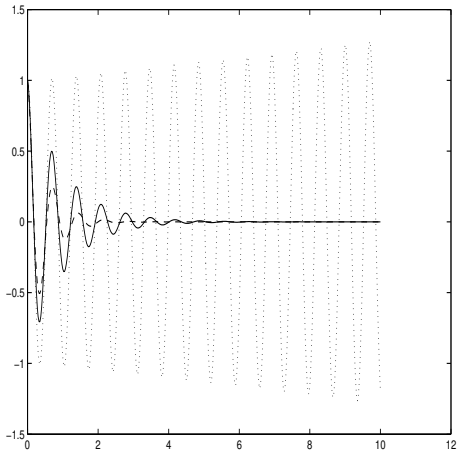
Aplicámos os métodos de Euler explícito e implícito ao problema teste (I_λ), escolhendo $\lambda = -1 + 9i$ e $T = 10$. Para este problema, o método de Euler explícito é absolutamente estável quando h toma valores até `0.0244`.

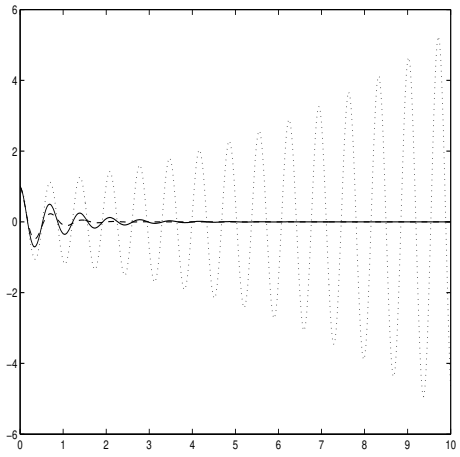
As duas primeiras figuras correspondem aos casos em que h tomou os valores `0.0200` e `0.0222`. Os testes foram feitos em MATLAB. As figuras ilustram o comportamento, no tempo, das partes reais da solução (a traço contínuo), da aproximação gerada pelo método de Euler explícito (a ponteado) e da aproximação gerada pelo método de Euler implícito (a tracejado). Verifica-se que o método implícito é melhor aproximação do que o explícito, mas constata-se que o explícito ainda exibe estabilidade (os seus gráficos aproximam-se do eixo das abcissas).





A seguir ilustramos o desempenho dos mesmos métodos quando h toma os valores `0.0250` e `0.0286`. Estes valores tornam o método de Euler explícito absolutamente instável, um fenómeno visível graficamente.





O método de Heun aplicado ao problema (I_λ) gera

$$u_k \;=\; \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right)^k, \qquad k \geq 0.$$

A sua região de estabilidade é um sobreconjunto da do método de Euler explícito.

Finalmente, com o método trapezoidal ou de Crank–Nicolson, obter-se-ia

$$u_k \;=\; \left(\frac{1 + h\lambda/2}{1 - h\lambda/2}\right)^k, \qquad k \geq 0.$$

A região de estabilidade deste método contém, claramente, o semiplano $\mathbb{C}^-$.

Um método numérico diz-se $\mathcal{A}$–*estável* se $\mathcal{A} \cap \mathbb{C}^- = \mathbb{C}^-$. Como as condições $\mathrm{Re}(h\lambda) < 0$ e $\mathrm{Re}(\lambda) < 0$ são equivalentes, um método é $\mathcal{A}$–*estável* se for absolutamente estável para todo o λ tal que $\mathrm{Re}(\lambda) < 0$.

Os métodos de Euler implícito e de Crank–Nicolson são $\mathcal{A}$–*estáveis*, mas os métodos de Euler explícito e de Heun não são $\mathcal{A}$–*estáveis*. Não existem métodos explícitos que sejam $\mathcal{A}$–*estáveis*. Esta propriedade de estabilidade é enfraquecida no estudo de métodos explícitos, considerando-se outras formas de estabilidade como a *estabilidade–rígida* ou a $\mathcal{A}_0$–*estabilidade*.

Exercícios

1. Mostre que $|1 + h\lambda| < 1$, com h real positivo, é equivalente a

$$h\lambda \;\in\; \mathbb{C}^- \qquad \text{e} \qquad 0 \;<\; h \;<\; -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2}.$$

2. Trace no plano complexo a região de estabilidade absoluta do método de Heun: $\mathcal{A} = \{\hat{h} \in \mathbb{C} : \ |1 + \hat{h} + \hat{h}^2/2| < 1\}$.

3. Confirme as expressões para u_k geradas pelos quatro métodos aplicados, nesta aula, ao problema teste (I_λ).

Lista das Aulas

1. Método de Newton para Sistemas de Equações Não Lineares
2. Taxas de Convergência e Constantes de Lipschitz
3. Taxa de Convergência Local do Método de Newton para Sistemas de Equações Não Lineares
4. Métodos de Quasi-Newton para Sistemas de Equações Não Lineares
5. Taxa de Convergência Local dos Métodos de Quasi-Newton para Sistemas de Equações Não Lineares
6. Conceitos Básicos sobre Optimização sem Restrições
7. Métodos de Newton e de Quasi-Newton para Optimização sem Restrições
8. Problemas de Mínimos Quadrados Não Lineares
9. Diferenciação Numérica
10. Conceitos Básicos sobre Integração Numérica
11. Integração Numérica – Fórmulas Trapezoidal e de Simpson
12. Integração Numérica – Fórmulas Compostas
13. Conceitos Básicos sobre Aproximação de Funções
14. Polinómios Ortogonais – Legendre e Chebyshev
15. Polinómios Ortogonais – Propriedades
16. Integração Gaussiana
17. Introdução à Aproximação Trigonométrica
18. Transformadas Discreta e Rápida de Fourier
19. Formulação Variacional de um Problema de Condições de Fronteira
20. Princípio de Energia Potencial Mínima para um Problema de Condições de Fronteira
21. Método de Elementos Finitos para um Problema de Condições de Fronteira
22. Uma Estimativa para o Erro no Método de Elementos Finitos
23. Conceitos Básicos sobre Problemas de Valor Inicial
24. Introdução aos Métodos Numéricos para Problemas de Valor Inicial

25. Consistência, Estabilidade-Zero, Convergência
26. Métodos de Runge–Kutta para Problemas de Valor Inicial
27. Estabilidade Absoluta de Métodos Numéricos para Problemas de Valor Inicial

Bibliografia

1. J. E. Dennis e R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

2. S. D. Conte e C. de Boor, *Elementary Numerical Analysis — An Algorithmic Approach*, terceira edição, Mc-Graw-Hill, Nova Iorque, 1980.

3. C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Lund, 1994.

4. M. Mori, *The Finite Element Method and Its Applications*, MacMillan Publishing Company, New York, 1983.

5. J. Nocedal e S. J. Wright, *Numerical Optimization*, Springer-Verlag, Nova Iorque, 1999.

6. H. Pina, *Métodos Numéricos*, McGraw-Hill, Lisboa 1995.

7. A. Quarteroni, R. Sacco e F. Saleri, *Numerical Mathematics*, Texts in Applied Mathematics, 37, Springer-Verlag, Berlim, 2000.

8. C. F. Van Loan, *Introduction to Scientific Computing — A Matrix-Vector Approach Using Matlab*, The Matlab Curriculum Series, Prentice-Hall, Upper Saddle River, New Jersey, 1997.