Análise Numérica (Teoria)

Carlos J. S. Alves

Instituto Superior Técnico

Versão 0.5

(Dezembro de 2012 - compilação)

LMAC, MMA, MEIC

Conteúdo

Capítulo 1

Aproximação de funções

1.1 Interpolação de Lagrange

Consideramos um subespaço finito $G = \langle \mathbf{g} \rangle$, gerado por $\mathbf{g} = \{g_0, ..., g_n\}$, uma lista de funções. Dada uma lista de nós distintos $\mathbf{x} = \{x_0, ..., x_n\}$ e uma lista de valores $\mathbf{y} = \{y_0, ..., y_n\}$, pretende-se encontrar $\phi \in G : \phi(\mathbf{x}) = \mathbf{y}$.

• O problema tem uma solução imediata que consiste na resolução do sistema

$$\mathbf{g}(\mathbf{x})\mathbf{a} = \mathbf{y} \Leftrightarrow \begin{bmatrix} g_0(x_0) & \cdots & g_n(x_0) \\ \vdots & \ddots & \vdots \\ g_0(x_n) & \cdots & g_n(x_n) \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix} \tag{1.1.1}$$

em que $\mathbf{g}(\mathbf{x}) = (\mathbf{g}(x_0), \cdots, \mathbf{g}(x_n))$ é uma matriz quadrada, em que cada linha k é o vector $\mathbf{g}(x_k) = (g_0(x_k), ..., g_n(x_k))$. Da mesma forma usaremos $\mathbf{g}(t)$ para identificar o vector num qualquer ponto t.

A matriz $\mathbf{g}(\mathbf{x})$ é invertível se as funções g_k forem linearmente independentes em $\mathbf{x}$, formando uma base.

Através do vector $\mathbf{a}$ obtemos imediatamente $\phi(t) = \mathbf{g}(t) \cdot \mathbf{a}$, porque

$$\phi(\mathbf{x}) = \mathbf{g}(\mathbf{x})\mathbf{a} = \mathbf{y}. \tag{1.1.2}$$

Observação: Para verificar a invertibilidade da matriz podemos usar o seguinte resultado de Álgebra Linear:

- *Se* $\mathbf{A}$ *é matriz quadrada, temos* $\mathbf{Av} = 0 \Rightarrow \mathbf{v} = 0$ *sse* $\mathbf{A}$ *é invertível.*

Trata-se de outra maneira de dizer que a independência das colunas de uma matriz quadrada é equivalente à invertibilidade. De facto, escrevendo $\mathbf{A} = (\mathbf{A}_1, ..., \mathbf{A}_n)$ em que $\mathbf{A}_k$ é a coluna k, ficamos com $0 = \mathbf{Av} = v_1\mathbf{A}_1 + ... + v_n\mathbf{A}_n \Rightarrow \mathbf{v} = (v_1, ..., v_n) = 0$, o que traduz a independência linear dos vectores coluna. □

• O facto de haver uma solução imediata, não significa que seja este o melhor caminho. A matriz $\mathbf{g}(\mathbf{x})$ pode ser mal condicionada, o que pode representar um problema para a resolução do sistema, e por outro lado, também se poderá tentar reduzir o número de operações.

Exemplo 1. Pretende-se determinar uma função que interpole os pontos (x_k, y_k) que são

$$(-1, 0),\ (0, 1),\ (1, 0),$$

mas que tenda para zero no infinito. Se usarmos polinómios sabemos que a condição no infinito não será verificada, por isso consideramos outras funções base que tenham esse comportamento, $g_0(x) = \frac{2}{1+x^2}, g_1(x) = \frac{3}{2+x^2}, g_2(x) = \frac{x}{1+x^2}$. À partida não sabemos se estas funções são linearmente independentes no conjunto de nós $\mathbf{x} = \{-1, 0, 1\}$, verificamos isso construindo o sistema:

$$\begin{bmatrix} g_0(-1) & g_1(-1) & g_2(-1) \\ g_0(0) & g_1(0) & g_2(0) \\ g_0(1) & g_1(1) & g_2(1) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 1 & 1 & -\frac{1}{2} \\ 2 & \frac{3}{2} & 0 \\ 1 & 1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

obtendo $a_0 = 2, a_1 = -2, a_2 = 0$, ou seja $\phi(x) = \frac{4}{1+x^2} - \frac{6}{2+x^2}$. Notamos que se tivéssemos escolhido todas as funções g_k pares, então $g_k(-1) = g_k(1)$ pelo que a primeira e última linha coincidiriam, havendo dependência linear.

1.2 Interpolação de Lagrange Polinomial

Trata-se do caso de funções de variável real (ou complexa) em que $\mathbf{v}(t) = \{1, t, t^2, ..., t^n\}$, ou seja as funções base são monómios, e o subespaço G consiste nos polinómios de grau menor ou igual que n, normalmente designado $\mathcal{P}_n$. Como é claro, poderá escolher-se outra base de polinómios, mas tendo escolhido a base canónica, designaremos por $\mathbf{v}$ em vez de $\mathbf{g}$, por coerência com o nome da matriz $\mathbf{v}(\mathbf{x})$ que é designada matriz de Vandermonde (ou Van der Monde)

$$\mathbf{v}(\mathbf{x}) = \begin{bmatrix} 1 & x_0 & \cdots & (x_0)^n \\ \vdots & & \ddots & \vdots \\ 1 & x_n & \cdots & (x_n)^n \end{bmatrix}. \tag{1.2.1}$$

Neste caso, para verificarmos que a matriz quadrada $\mathbf{v}(\mathbf{x})$ é invertível (ou seja, que os monómios são linearmente independentes no conjunto de nós $\mathbf{x}$) basta usar o teorema fundamental da álgebra, pois $\mathbf{v}(\mathbf{x})\mathbf{a} = 0$ significa que o polinómio de grau menor ou igual a n

$$p_n(t) = \mathbf{v}(t) \cdot \mathbf{a} = a_0 + a_1 t + ... + a_n t^n \tag{1.2.2}$$

tem $n + 1$ raízes, em $\mathbf{x} = \{x_0, ..., x_n\}$, o que implica que seja o polinómio nulo, logo $\mathbf{a} = 0$.

1.2.1 Fórmula de Lagrange.

Um outro processo de explicitar a existência consiste em considerar uma base adaptada a esse conjunto de pontos, constituída por polinómios $L_j \in \mathcal{P}_n$ que verifiquem $L_j(x_i) = \delta_{ij}$, pois dessa forma obtemos a matriz identidade, ou seja, $\mathbf{L}(\mathbf{x}) = \mathbf{I}$, e não é necessário resolver qualquer sistema, ficamos com $\mathbf{a} = \mathbf{y}$, e por isso

$$p_n(t) = \mathbf{L}(t) \cdot \mathbf{y}$$

esta fórmula será designada por *fórmula de Lagrange*, e explicitando o cálculo das componentes L_i de $\mathbf{L}$,

$$L_j(t) = \prod_{i=0, i \neq j}^{n} \frac{t - x_i}{x_j - x_i} \tag{1.2.3}$$

verificam $L_i(x_j) = \delta_{ij}$ e tratam-se de polinómios de grau n, podendo ser escritos na base canónica (ou seja, passamos para a solução do sistema de Vandermonde).

1.2.2 Fórmula de Newton.

Uma outra escolha de base,

$$\begin{aligned} w_0(t) &= 1, \\ w_1(t) &= (t-x_0), \\ &\vdots \quad \vdots \quad \vdots \\ w_n(t) &= (t-x_0)\cdots(t-x_{n-1}) \end{aligned}$$

permite a possibilidade de simplificar o sistema, mas não ao ponto da diagonalização, a matriz $\mathbf{w}(\mathbf{x})$ será apenas triangular inferior.

$$\mathbf{w}(\mathbf{x}) = \begin{bmatrix} w_0(x_0) & \cdots & w_n(x_0) \\ \vdots & \ddots & \vdots \\ w_0(x_n) & \cdots & w_n(x_n) \end{bmatrix} = \begin{bmatrix} w_0(x_0) & 0\cdots & 0 \\ \vdots & \ddots & \vdots \\ w_0(x_n) & \cdots & w_n(x_n) \end{bmatrix}$$

De facto, é imediato ver que $w_j(x_i) = (x_i - x_0)\cdots(x_i - x_{j-1}) = 0,$ se $i < j$.

Por outro lado, os elementos da diagonal não são nulos, $w_i(x_i) = (x_i - x_0)\cdots(x_i - x_{i-1}) \neq 0$, pois os nós de interpolação são distintos.

A invertibilidade é assim imediata, e de

$$\sum_{j=0}^{k} a_j w_j(x_k) = y_k$$

podemos explicitar a solução do sistema $\mathbf{w}(\mathbf{x})\mathbf{a} = \mathbf{y}$ de forma recursiva

$$\begin{aligned} a_0 &= y_0 \\ a_k &= \frac{1}{w_k(x_k)} \left(y_k - \sum_{j=0}^{k-1} a_j w_j(x_k) \right), \text{ para } k = 1, ..., n, \end{aligned} \tag{1.2.4}$$

estes valores a_k são normalmente designados por *diferenças divididas*, escrevendo-se

$$a_k = y_{[x_0,\ldots,x_k]}.$$

Obtemos assim a denominada *fórmula de Newton*:

$$\phi(t) = \mathbf{a} \cdot \mathbf{w}(t) = \sum_{k=0}^{n} y_{[x_0,\ldots,x_k]}(t-x_0)\cdots(t-x_{k-1}). \tag{1.2.5}$$

• Uma vantagem adicional desta fórmula é que a adição de um ponto de interpolação x_{n+1} com valor y_{n+1} não implica mudar todas as funções base, como acontece com a fórmula de Lagrange, apenas adicionamos $w_{n+1}(t) = (t-x_0)\cdots(t-x_{n-1})(t-x_n)$. O cálculo recursivo dos coeficientes mantém-se, apenas necessitamos de considerar um novo $a_{n+1} = y_{[x_0,\ldots,x_{n+1}]}$. Assim, sendo $p_n(t) = \mathbf{a}^{[n]} \cdot \mathbf{w}^{[n]}(t)$ o polinómio interpolador nos nós $(x_0, ..., x_n)$, obtemos

$$\begin{aligned} p_{n+1}(t) &= \mathbf{a}^{[n+1]} \cdot \mathbf{w}^{[n+1]}(t) = \mathbf{a}^{[n]} \cdot \mathbf{w}^{[n]}(t) + a_{n+1} w_{n+1}(t) \\ &= p_n(t) + y_{[x_0,\ldots,x_{n+1}]}(t-x_0)\cdots(t-x_n), \end{aligned}$$

o polinómio interpolador nos nós $(x_0, ..., x_n, x_{n+1})$. Concluímos ainda que o coeficiente de maior grau é a diferença dividida $y_{[x_0,\ldots,x_{n+1}]}$.

Exercício 1. Usando as funções base da Fórmula de Newton, encontre o polinómio interpolador que verifica

$$p(-1) = p(0) = p(1) = 1,\ p(2) = 7.$$

Resolução: Basta calcular com $w_0(t) = 1, w_1(t) = t + 1, w_2(t) = (t + 1)t, w_3(t) = (t + 1)t(t - 1)$:

$$\mathbf{w}(\mathbf{x})\mathbf{a} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 1 & 3 & 6 & 6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 7 \end{bmatrix}$$

e assim obtemos $\mathbf{a} = (1, 0, 0, 1)$, ou seja $p(t) = \mathbf{a} \cdot \mathbf{w}(t) = 1 + (t+1)t(-1) = 1 - t + t^3$, um polinómio de terceiro grau, conforme esperado.

Exercício 2. Aplique o resultado do exercício anterior para encontrar uma função interpoladora ϕ que verifique $\phi(-1) = \frac{1}{2}, \phi(0) = 1, \phi(1) = \frac{1}{2}, \phi(2) = \frac{7}{17}$, mas que tenda para zero no infinito.

Resolução: Podemos escolher $\phi(t) = \frac{1}{1+t^4}(a_0 + a_1 t + a_2 t^2 + a_3 t^3)$ que tende para zero no infinito. Querendo que $\phi(x_i) = y_i$ obtemos $a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 = y_i(1 + x_i^4)$.
Reduzimos assim a um problema de interpolação polinomial, alterando o valor da imagem $\hat{y}_0 = y_0(1 + x_0^2) = \frac{1}{2}(1 + (-1)^4) = 1$, e de forma semelhante $\hat{y}_1 = y_1 = 1, \hat{y}_2 = 2y_2 = 1, \hat{y}_3 = y_3(1 + 2^4) = 7$. Obtivémos os valores do exercício anterior, para evitar novos cálculos, e por isso a solução é

$$\phi(t) = \frac{1}{1 + t^4}(1 - t + t^3).$$

Observação 1. *(Diferenças divididas)* A razão do nome "diferenças divididas" para $a_n = y_{[x_0,...,x_n]}$ está relacionada com outra propriedade interessante:

$$y_{[x_0,...,x_n]} = \frac{y_{[x_1,...,x_n]} - y_{[x_0,...,x_{n-1}]}}{x_n - x_0}$$

que é a normalmente usada para obter de forma recursiva $a_n = y_{[x_0,...,x_n]}$.

Observação 2. *(Número de operações)* Através deste estudo podemos concluir que o número de operações elementares através das Fórmulas de Lagrange ou Newton em $O(n^2)$ compensa face à resolução do sistema com a matriz de Vandermonde que envolve $O(n^3)$ operações.

1.2.3 Erro de interpolação polinomial.

Até aqui as imagens $y_0, ..., y_n$ atribuídas aos nós $x_0, ..., x_n$ são completamente arbitrárias, havendo possibilidades infinitas para funções que tomem esses valores. No entanto, se associarmos os valores y_k aos valores $f_k = f(x_k)$, para uma função f com alguma regularidade, é possível obter estimativas que indicam em que medida o polinómio interpolador constitui uma aproximação razoável da função f fora dos nós de interpolação. É claro que quanto mais próximo dos nós, melhor será a aproximação. Quando consideramos o cálculo do polinómio fora do intervalo $[x_0; ...; x_n]$ (que contém todos os pontos), é habitual falar denominar a aproximação por *extrapolação*.

Consideremos p_n o polinómio interpolador nos nós $x_0, ..., x_n$, e f uma função qualquer.

Definimos o erro num ponto $z \notin \{x_0, ..., x_n\}$, como sendo $E(z) = f(z) - p_n(z)$.

Começamos por reparar que podemos considerar z como um nó de interpolação adicional, logo

$$p_{n+1}(t) = p_n(t) + f_{[x_0,\ldots,x_n,z]}(t-x_0)\cdots(t-x_n)$$

e como se trata de um nó de interpolação $p_{n+1}(z) = f(z)$, portanto $E(z) = f(z) - p_n(z) = p_{n+1}(z) - p_n(z)$, ou seja

$$E(z) = f_{[x_0,\ldots,x_n,z]}(z-x_0)\cdots(z-x_n).$$

Esta fórmula tem utilidade prática limitada, porque não podemos calcular $f_{[x_0,\ldots,x_n,z]}$ sem conhecer $f(z)$, mas tem utilidade teórica.

Diferenças divididas e diferenciação. Podemos obter um teorema que relaciona a diferenciação com as diferenças divididas.

Teorema 1. *Assumindo que $f \in C^m[x_0;\ldots;x_m]$, então*

$$\exists \xi \in [x_0;\ldots;x_m] : f_{[x_0,\ldots,x_m]} = \frac{1}{m!} f^{(m)}(\xi). \tag{1.2.6}$$

Demonstração. Consideremos p_n o polinómio interpolador em $x_0,\ldots,x_m$. A função $E = f - p_m$ tem pelo menos $m+1$ zeros em $[x_0;\ldots;x_m]$, e é diferenciável continuamente, logo pelo Teorema de Rolle E' tem pelo menos m zeros em $[x_0;\ldots;x_m]$. Da mesma forma E'' terá pelo menos $m-1$ zeros em $[x_0;\ldots;x_m]$, e assim sucessivamente até que concluímos que $E^{(m)}$ tem pelo menos um zero ξ em $[x_0;\ldots;x_m]$. Agora, basta reparar que

$$0 = E^{(m)}(\xi) = f^{(m)}(\xi) - p_m^{(m)}(\xi) = f^{(m)}(\xi) - f_{[x_0,\ldots,x_m]} m!,$$

porque o coeficiente de grau m de p_m é exactamente $f_{[x_0,\ldots,x_m]}$. □

Teorema 2. *Assumindo que $f \in C^{n+1}[x_0;\ldots;x_n]$, então*

$$\exists \xi \in [x_0;\ldots;x_n;z] : E(z) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(z-x_0)\cdots(z-x_n). \tag{1.2.7}$$

Demonstração. Resulta do teorema anterior considerando $m = n+1$ com $x_{n+1} = z$ e de

$$E(z) = f_{[x_0,\ldots,x_n,z]}(z-x_0)\cdots(z-x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(z-x_0)\cdots(z-x_n).$$

□

Através desta fórmula de erro podemos ainda escrever uma igualdade, semelhante à expansão em série de Taylor,

$$f(z) = p_n(z) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(z-x_0)\cdots(z-x_n), \text{ com } \xi \in [x_0;\ldots;x_n;z],$$

onde p_n é o polinómio interpolador, e usando a fórmula de Newton ficamos com

$$\begin{aligned} f(z) &= f(x_0) + f_{[x_0,x_1]}(z-x_0) + \cdots + f_{[x_0,\ldots,x_n]}(z-x_0)\cdots(z-x_{n-1}) \\ &\quad + \frac{f^{(n+1)}(\xi)}{(n+1)!}(z-x_0)\cdots(z-x_n), \end{aligned}$$

assim a expansão em série de Taylor surge como caso limite quando $x_1, ..., x_n \to x_0$ porque

$$f_{[x_0,\ldots,x_m]} = \frac{1}{m!}f^{(m)}(\xi) \to \frac{1}{m!}f^{(m)}(x_0).$$

Observação 3. Ou seja, quando repetirmos $n+1$ vezes um nó x, isso leva à identificação:

$$f_{[x,\ldots,x]} = \frac{1}{m!}f^{(m)}(x) \tag{1.2.8}$$

Estimativa de Erro: Para efeitos práticos, como o ponto ξ será desconhecido, majoramos o erro:

$$|E(z)| \leq \frac{||f^{(n+1)}||_\infty}{(n+1)!}|z-x_0|\cdots|z-x_n|,$$

em que $||.||_\infty$ representa a norma do máximo no intervalo considerado, ou seja $[x_0; ...; x_n; z]$. Recordamos que a norma uniforme num intervalo $[a, b]$ é dada por

$$||u||_\infty = \max_{t\in[a,b]} |u(t)|. \tag{1.2.9}$$

1.3 Aplicação à regularização de dados. Filtros.

Podemos admitir que os dados que queremos interpolar são inexactos e resultam de valores experimentais sujeitos a ruído aleatório. Ou seja, que os valores correctos seriam $f_0, ..., f_N$, mas devido a imprecisão, ou ruído, obtivémos $\tilde{f}_0, ..., \tilde{f}_N$. Uma maneira de contornar o problema desse ruído é usar filtros, que permitem integrar o ruído, regularizando os dados.

1.3.1 Formulação contínua

Seja f a função original, e seja $\tilde{f}$ a função após uma perturbação "ruído" ρ, tal que

$$\tilde{f}(x) = f(x) + \rho(x),$$

onde a distribuição de "ruído" é tal que $\int_{x-\varepsilon}^{x+\varepsilon} \rho(t)dt = 0$.

Neste caso, a integração permite minorar o ruído, pois

$$\frac{1}{2\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon} \tilde{f}(t)dt = \frac{1}{2\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon} f(t)dt + \frac{1}{2\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon} \rho(t)dt = f(\tilde{x}),$$

com $\tilde{x} \in [x-\varepsilon, x+\varepsilon]$, aplicando o teorema do valor intermédio para integrais. Quanto $\varepsilon \to 0$, temos $f(\tilde{x}) \to f(x)$, o que justifica a aproximação para funções contínuas. Este é um caso particular, em que consideramos como filtro de regularização a integração com uma função descontínua.

Definição 1. Baseados nesta ideia, consideramos outro tipo de regularizações, definindos *filtros regularizadores* enquanto funções μ_ε verificando:

$$||\mu_\varepsilon||_{L^1(\mathbb{R})} = 1, \text{ em que } \mu \text{ é positiva e par, com suporte em } [-\varepsilon, \varepsilon].$$

Proposição 1. *Para $f \in C[-\varepsilon, \varepsilon]$, os filtros regularizadores verificam*

$$\langle \mu_\varepsilon, f \rangle_{L^2(\mathbb{R})} \underset{\varepsilon \to 0}{\longrightarrow} f(0).$$

Demonstração. O suporte de μ_ε implica que a função é nula fora do intervalo $]-\varepsilon, \varepsilon[$,

$$\langle \mu_\varepsilon, f \rangle_{L^2(\mathbb{R})} = \int_{\mathbb{R}} \mu_\varepsilon(t) f(t) dt = \int_{-\varepsilon}^{\varepsilon} \mu_\varepsilon(t) f(t) dt$$

como $\mu_\varepsilon \geq 0$ aplicando o teorema do valor intermédio para integrais, $\xi \in [-\varepsilon, \varepsilon]$,

$$\langle \mu_\varepsilon, f \rangle_{L^2(\mathbb{R})} = f(\xi) \int_{-\varepsilon}^{\varepsilon} \mu_\varepsilon(t) dt = f(\xi) ||\mu_\varepsilon||_{L^1(\mathbb{R})} = f(\xi) \underset{\varepsilon \to 0}{\longrightarrow} f(0).$$

□

Estes filtros estão centrados em zero, mas podem ser deslocados, considerando uma translação do centro para um z qualquer, fazendo $\mu_{\varepsilon,z}(x) = \mu_\varepsilon(x - z)$.

1.3.2 Exemplos de filtros.

A proposição generaliza a propriedade apresentada inicialmente em que se considerava um filtro descontínuo

$$\mu_\varepsilon^{[0]}(x) = \begin{cases} \frac{1}{2\varepsilon}, & (|x| < \varepsilon) \\ 0, & (|x| \geq \varepsilon) \end{cases} \tag{1.3.1}$$

mas podemos ainda considerar filtros contínuos,$\mu_\varepsilon^{[1]} \in C(\mathbb{R})$,

$$\mu_\varepsilon^{[1]}(x) = \begin{cases} \frac{\varepsilon - |x|}{\varepsilon^2}, & (|x| < \varepsilon) \\ 0, & (|x| \geq \varepsilon) \end{cases} \tag{1.3.2}$$

ou ainda diferenciáveis, $\mu_\varepsilon^{[2]} \in C^1(\mathbb{R})$,

$$\mu_\varepsilon^{[2]}(x) = \begin{cases} \frac{15(\varepsilon+x)^2(\varepsilon-x)^2}{16\varepsilon^5}, & (|x| < \varepsilon) \\ 0, & (|x| \geq \varepsilon) \end{cases}$$

e de um modo geral podemos definir ainda filtros mais regulares, $\mu_\varepsilon^{[p]} \in C^{p-1}(\mathbb{R})$,

$$\mu_\varepsilon^{[p]}(x) = \begin{cases} M_p \frac{(\varepsilon+x)^p(\varepsilon-x)^p}{\varepsilon^{2p+1}}, & (|x| < \varepsilon) \\ 0, & (|x| \geq \varepsilon) \end{cases} \tag{1.3.3}$$

em que M_p é uma constante tal que $||\mu_\varepsilon^{[p]}||_{L^1(\mathbb{R})} = 1$ (esta constante não tem fórmula explícita, sendo $M_1 = \frac{3}{4}, M_2 = \frac{15}{16}, M_3 = \frac{35}{32}, M_4 = \frac{315}{256}, ...$).

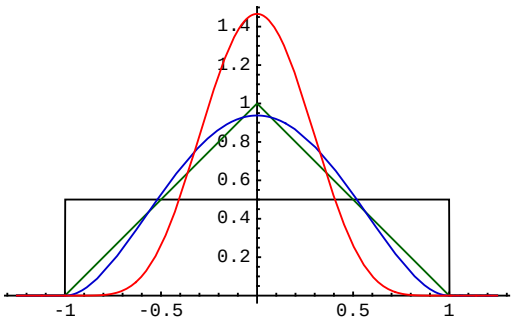

Figura 1.3.1: Diversos filtros:$\mu_1^{[0]}$ (preto),$\mu_1^{[1]}$ (verde), $\mu_1^{[2]}$ (azul), $\mu_1^{[6]}$ (vermelho).

1.3.3 Delta de Dirac

Apesar de termos visto que $\langle \mu_\varepsilon, f \rangle_{L^2(\mathbb{R})}$ converge para $f(0)$, não há nenhuma função no limite de μ_ε quando $\varepsilon \to 0$, porque o suporte da função é reduzido a $[-\varepsilon, \varepsilon]$, ao mesmo tempo que o integral deve ser 1, e não zero. Por isso é definido um símbolo (uma distribuição), denominado *Delta de Dirac* δ que representa este limite. Assim definimos, para funções f contínuas,

$$\langle \delta, f \rangle_{L^2(\mathbb{R})} = f(0).$$

Quando mudamos o centro, por translação para um ponto y definimos o delta de Dirac δ_y

$$\langle \delta_y, f \rangle_{L^2(\mathbb{R})} = f(y). \tag{1.3.4}$$

Desta forma, o valor de uma função num ponto y pode ser aproximado considerando a translação de um filtro para y, ou seja, $\mu_{\varepsilon,y}$ porque

$$f(y) = \langle \delta_y, f \rangle_{L^2(\mathbb{R})} \approx \langle \mu_{\varepsilon,y}, f \rangle_{L^2(\mathbb{R})} \text{ (quando } \varepsilon \to 0).$$

1.3.4 Produto de Convolução

Sejam $f, g \in L^2(\mathbb{R})$, com base na translação define-se o produto de convolução

$$(f * g)(y) = \int_{\mathbb{R}} f(x) g(y - x) dx \tag{1.3.5}$$

que é comutativo e verifica as propriedades habituais do produto, cujo elemento neutro é o delta de Dirac, pois

$$(f * \delta)(y) = (\delta * f)(y) = \int_{\mathbb{R}} \delta(x) f(y - x) dx = f(y - 0) = f(y).$$

Desta forma podemos definir a função regularizada, que resulta da aplicação de um filtro por translação

$$(f * \mu_\varepsilon)(y) = \int_{\mathbb{R}} f(x) \mu_\varepsilon(y - x) dx = \langle \mu_{\varepsilon,y}, f \rangle_{L^2(\mathbb{R})} = f(\tilde{y}) \approx f(y), \text{ com } \tilde{y} \in [y - \varepsilon, y + \varepsilon] \tag{1.3.6}$$

1.3.5 Derivadas generalizadas

Esta noção de regularização permite ainda estender a noção de derivada.

Proposição 2. *Se $f, \mu_\varepsilon \in C^1(\mathbb{R})$, em que μ_ε é um filtro, então*

$$\langle \mu_\varepsilon, f' \rangle_{L^2(\mathbb{R})} = - \langle \mu'_\varepsilon, f \rangle_{L^2(\mathbb{R})} . \tag{1.3.7}$$

Demonstração. Como $\mu(\pm\varepsilon) = 0$, integrando por partes, obtém-se

$$\langle \mu_\varepsilon, f' \rangle_{L^2(\mathbb{R})} = \int_{-\varepsilon}^{\varepsilon} \mu_\varepsilon(t) f'(t) dt = [\mu_\varepsilon(t) f'(t)]_{t=-\varepsilon}^{t=\varepsilon} - \int_{-\varepsilon}^{\varepsilon} \mu_\varepsilon(t) f'(t) dt = 0 - \langle \mu'_\varepsilon, f \rangle_{L^2(\mathbb{R})} .$$

□

Este resultado mostra que podemos definir uma aproximação da derivada, mesmo quando ela não tem sentido clássico, passando a derivada para o filtro regularizador. Apesar de demonstrarmos o resultado exigindo que $\mu_\varepsilon \in C^1(\mathbb{R})$, a integração de Lebesgue permite mesmo considerar a derivada no caso em que não há descontinuidades.

Por exemplo, podemos derivar $\mu_\varepsilon^{[1]}$ seccionalmente

$$\mu_\varepsilon^{[1]\prime}(x) = \begin{cases} \frac{1}{\varepsilon^2}, & (-\varepsilon < x < 0) \\ -\frac{1}{\varepsilon^2}, & (0 < x < \varepsilon) \\ 0, & (|x| \geq \varepsilon) \end{cases}$$

e considerar a aproximação da derivada

$$\begin{aligned} f'(y) &= \langle \delta_y, f' \rangle_{L^2(\mathbb{R})} \approx \langle \mu_{\varepsilon,y}, f' \rangle_{L^2(\mathbb{R})} = & (1.3.8) \\ &= - \left\langle \mu'_{\varepsilon,y}, f \right\rangle_{L^2(\mathbb{R})} = -\frac{1}{\varepsilon^2} \left(\int_{y-\varepsilon}^{y} f(t) dt - \int_{y}^{y+\varepsilon} f(t) dt \right) & (1.3.9) \end{aligned}$$

e este procedimento pode ser aplicado aos outros filtros, e ainda a derivadas de maior ordem, por aplicação sucessiva. No entanto, convém notar que $\mu_\varepsilon^{[1]}$ não deve ser derivado segunda vez no sentido clássico. Com efeito, podemos ver que a derivação de funções descontínuas leva à noção de delta de Dirac.

Observação 4. Se considerarmos a denominada *função de Heaviside:*

$$H(x) = \begin{cases} 1, & (0 < x) \\ 0, & (x \leq 0) \end{cases}$$

obtemos para qualquer função f diferenciável com suporte limitado (tal que $f(x) = 0$, para $x > R$),

$$\langle H', f \rangle_{L^2(\mathbb{R})} = - \langle H, f' \rangle_{L^2(\mathbb{R})} = - \int_0^R f'(t) dt = -f(R) + f(0) = f(0) = \langle \delta, f \rangle_{L^2(\mathbb{R})} ,$$

o que leva à identificação $H' = \delta$, ou seja, do delta de Dirac com a derivada da função descontínua de Heaviside. Isto mostra ainda por que $\mu_\varepsilon^{[0]\prime} = \frac{1}{2\varepsilon}(\delta_{-\varepsilon} - \delta_\varepsilon)$ ou $\mu_\varepsilon^{[1]\prime\prime} = \frac{1}{\varepsilon^2}(\delta_{-\varepsilon} - 2\delta_0 + \delta_\varepsilon)$ sendo derivadas de funções descontínuas, expressas através de deltas de Dirac, não têm correspondente no sentido clássico, mas fazem sentido enquanto fórmulas de diferenças finitas.

1.3.6 Formulação discreta

Na maioria das aplicações não temos dados contínuos para uma função f, mas apenas dados nalguns nós $f(x_0) = f_0, \cdots, f(x_N) = f_N$. Ainda assim, esses dados podem estar perturbados por ruído aleatório, medindo-se as perturbações $\tilde{f}_0, ..., \tilde{f}_N$. Nesse caso, não faz sentido considerar as integrações, ou os filtros regularizadores definidos em todos os pontos.

Vamos considerar que esses dados resultam de nós igualmente espaçados, de forma a nos concentrarmos apenas nos valores da função.

Assumimos implicitamente que a função é periódica, de forma que $f_N = f_0$, e apenas consideramos os índices de 0 até $N-1$.

O produto interno em L^2 é substituído pelo seu equivalente discreto l^2, sendo necessário ter especial atenção ao produto de convolução discreto, $\mathbf{v} * \mathbf{w}$, que é um vector, definido pelas componentes

$$[\mathbf{v} * \mathbf{w}]_k = \sum_{j=0}^{N-1} v_j w_{k-j} \tag{1.3.10}$$

subentendendo-se que os valores de índices negativos são módulo N, ou seja $-j = N - j \,(\text{mod } N)$. É ainda claro que o elemento neutro será o delta de Kronecker centrado no indíce 0, ou seja o vector $\delta_{0j} = (1, 0, \cdots, 0)$.

Também os filtros de regularização, passam a vectores, que designaremos por $\mathbf{w}$, e são centrados no índice 0, com as propriedades discretas correspondentes (para um $E < \frac{N}{2}$):

- soma unitária $||\mathbf{w}||_1 = 1$, não negativos$w_j \geq 0$,
- com suporte limitado $w_j = 0$ para $E < j < N - E$, e que são simétricos $w_{-j} = w_j$.

Assim, no caso mais simples, correspondente a $\mu_\varepsilon^{[0]}$, com $\varepsilon = \frac{E}{N}$, tomamos o vector

$$\mathbf{w}^{[0]} = \frac{1}{2E+1}(1, \overbrace{1, \cdots, 1}^{E\,vezes}, 0, \cdots, 0, \overbrace{1, \cdots, 1}^{E\,vezes}) \tag{1.3.11}$$

e a regularização do vector de dados $\tilde{\mathbf{f}}$ será dada pela convolução $\tilde{\mathbf{f}} * \mathbf{w}^{[0]}$, sendo fácil verificar neste caso que

$$[\tilde{\mathbf{f}} * \mathbf{w}^{[0]}]_k = \sum_{j=0}^{N-1} \tilde{f}_j w_{k-j} = \frac{1}{2\epsilon+1} \sum_{j=k-\epsilon}^{k+\epsilon} \tilde{f}_j$$

o que corresponde a uma média de $2E+1$ valores adjacentes e por isso regulariza $\tilde{f}_k$ aproximando-o de f_k.

Mesmo em casos simples, a convolução pode revelar-se dispendiosa computacionalmente, sendo preferível usar um cálculo através da Transformada de Fourier Rápida, que iremos estudar em seguida.

1.3.7 Exercícios

Exercício 3. Considere uma função f e a sua perturbação $\tilde{f}(x) = f(x) + a\sin(Mx)$. Determine a convolução $\tilde{f} * \mu$ aplicando o filtro $\mu = \mu_{\varepsilon=\pi/M}^{[0]}$, e comente o resultado.

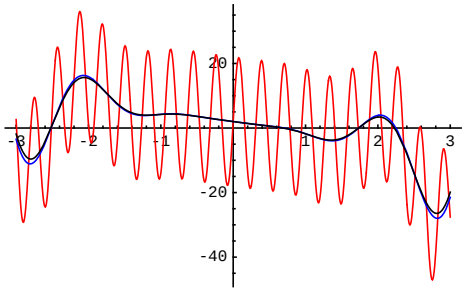

Figura 1.3.2: Exemplo com $f = x^3\sin(4x) - 3x + 2, a = M = 20$, apresentando $\tilde{f}$ (vermelho), f (azul), $\tilde{f} * \mu$(preto).

Resolução: Como μ é par, $\mu(y-x) = \mu(x-y)$, tem suporte em $[y-\varepsilon, y+\varepsilon]$,

$$(\tilde{f} * \mu)(y) = \int_{\mathbb{R}} \tilde{f}(x)\mu(y-x)dx = \int_{y-\frac{\pi}{M}}^{y+\frac{\pi}{M}} \tilde{f}(x)\mu(x-y)\,dx = \int_{y-\frac{\pi}{M}}^{y+\frac{\pi}{M}} (f(x) + a\sin(Mx))\,\frac{M}{2\pi}\,dx$$

e usando o teorema do valor intermédio para integrais, com $\xi \in [y-\frac{\pi}{M}, y+\frac{\pi}{M}]$,

$$= \frac{M}{2\pi}\int_{y-\frac{\pi}{M}}^{y+\frac{\pi}{M}} f(x)dx + \frac{aM}{2\pi}\int_{y-\frac{\pi}{M}}^{y+\frac{\pi}{M}} \sin(Mx)\,dx = \frac{M}{2\pi}f(\xi)\frac{2\pi}{M} + \frac{aM}{2\pi}\left[\frac{-1}{M}\cos(Mx)\right]_{x=y-\frac{\pi}{M}}^{x=y+\frac{\pi}{M}} = f(\xi),$$

porque $\cos(My+\pi) = \cos(My-\pi)$, e a 2ª parcela dá zero.

Neste caso aplicando a regularização recuperamos um valor aproximado da função original, e o erro será $f(y) - (\tilde{f} * \mu)(y) = f(y) - f(\xi) \to 0$ quando $M \to \infty$. Reparamos ainda que o resultado é independente da amplitude a.

Exercício 4. Mostre que se $f \in C^3$ então $(f * \mu_\varepsilon^{[1]\prime})(y) = f'(y) - \frac{f^{(3)}(\xi)}{12}\varepsilon^2$, e também $|(f * \mu_\varepsilon^{[1]\prime})(y)| \leq |f'(\xi)|$, para certo $\xi \in [y-\varepsilon, y+\varepsilon]$.

Resolução: Seja F a primitiva de f. De (1.3.8) $(f*\mu_\varepsilon^{[1]\prime})(y) = \frac{1}{\varepsilon^2}\left(\int_y^{y+\varepsilon} f(t)dt - \int_{y-\varepsilon}^{y} f(t)dt\right) = \frac{F(y+\varepsilon)-F(y)}{\varepsilon^2} - \frac{F(y+\varepsilon)-F(y)}{\varepsilon^2} = \frac{F(y+\varepsilon)-2F(y)+F(y-\varepsilon)}{\varepsilon^2} = F''(y) - \frac{F^{(4)}(\xi)}{12}\varepsilon^2$, notando que $f' = F''$.

Por outro lado, $|(f*\mu_\varepsilon^{[1]\prime})(y)| = \frac{1}{\varepsilon^2}\left|\int_y^{y+\varepsilon} f(t)dt - \int_{y-\varepsilon}^{y} f(t)dt\right| = \frac{1}{\varepsilon^2}\left|\varepsilon f(\xi^+) - \varepsilon f(\xi^-)\right| = \frac{|\xi^+-\xi^-|}{\varepsilon}\left|\frac{f(\xi^+)-f(\xi^-)}{\xi^+-\xi^-}\right| = \frac{|\xi^+-\xi^-|}{\varepsilon}|f'(\xi)| \leq |f'(\xi)|$ porque $\xi^+ \in [y, y+\varepsilon]$, $\xi^- \in [y-\varepsilon, y]$, logo $|\xi^+ - \xi^-| \leq \varepsilon$.

Exercício 5. Considere $\tilde{\mathbf{f}} = \mathbf{f} + \mathbf{r}$ tal que $r_{k-\epsilon} + \ldots + r_{k+\epsilon} = 0$, então $[\tilde{\mathbf{f}} * \mathbf{w}^{[0]}]_k = \frac{f_{k-\epsilon}+\ldots+f_{k+\epsilon}}{2\epsilon+1}$ onde $\mathbf{w}^{[0]}$ é definido por (1.3.11).

Resolução: Seja $[\tilde{\mathbf{f}} * \mathbf{w}]_k = \sum_{j=0}^{N-1}(f_j + r_j)w_{k-j} = \sum_{j=k-\epsilon}^{k+\epsilon} f_j w_{j-k} + \sum_{j=k-\epsilon}^{k+\epsilon} r_j w_{j-k} = \sum_{j=k-\epsilon}^{k+\epsilon} f_j \frac{1}{2\epsilon+1} + \sum_{j=k-\epsilon}^{k+\epsilon} r_j \frac{1}{2\epsilon+1} = \frac{1}{2\epsilon+1}\sum_{j=k-\epsilon}^{k+\epsilon} f_j$ e obtemos a média.

1.4 Interpolação Trigonométrica e TFD

1.4.1 Caso Geral

As fórmulas para o cálculo do polinómio interpolador são ainda válidas quando consideramos funções complexas, apenas serão diferentes as fórmulas para a relação do erro de interpolação com as derivadas, já que utilizámos teoremas de análise real, como o Teorema de Rolle, que não são válidos em análise complexa.

Estamos interessados em estudar um caso particular, em que os pontos de interpolação são da forma

$$x_k = e^{it_k} = \cos(t) + i\sin(t), \text{ com } t_k \in [0, 2\pi[, \text{ para } k = 0, ..., 2n,$$

que é idêntico a um caso de interpolação trigonométrica. Na interpolação trigonométrica consideramos como subespaço finito $G = \langle \mathbf{g} \rangle$, gerado por

$$\mathbf{g} = \{g_0, ..., g_{2n}\} = \{1, \cos(t), \sin(t), ..., \cos(nt), \sin(nt)\},$$

mas utilizando a exponencial complexa, podemos considerar $G = \langle \mathbf{u} \rangle$

$$\mathbf{u} = \{u_0, ..., u_{2n}\} = \{e^{-itn}, ..., e^{-it}, 1, e^{it}, ..., e^{itn}\},$$

em que a separação entre senos e co-senos é feita pela representação nos complexos.

Para além disso, reparamos que $\mathbf{u}$ tem relação directa com $\mathbf{v} = \{x^{-n}, ..., x, 1, x, ..., x^n\}$, efectuando a mudança de variável $x = e^{it}$. Assim, dados valores $\mathbf{y} = \{y_0, ..., y_{2n}\}$ associados aos nós $\mathbf{t} = \{t_0, ..., t_{2n}\}$, a resolução do sistema

$$\mathbf{u}(\mathbf{t})\mathbf{a} = \mathbf{y}$$

é equivalente a resolver $\mathbf{v}(\mathbf{x})\mathbf{a} = \mathbf{y}$ considerando $\mathbf{x} = \{e^{it_0}, ..., e^{it_{2n}}\}$, em que $\mathbf{v}(\mathbf{x})$ é uma matriz de Vandermonde (dividida por x^n).

1.4.2 Aplicação das fórmulas de Lagrange e Newton

Tendo obtido os coeficientes a_k, e a representação na forma complexa

$$\phi(t) = \sum_{k=-n}^{n} a_k e^{itk},$$

para funções reais a passagem para a forma trigonométrica

$$\phi(t) = c_0 + \sum_{k=1}^{n} (c_k \cos(kt) + b_k \sin(kt))$$

é efectuada considerando que $a_{-k} = \bar{a}_k$, e assim $c_k = 2Re(a_k), b_k = -2Im(a_k)$.

Com as devidas transformações, as fórmulas de Lagrange e Newton obtidas para a interpolação polinomial são ainda válidas neste caso.

Por exemplo, da fórmula de Lagrange, obtém-se $\phi = \mathbf{y} \cdot \mathbf{L}$, com $z = e^{i\tau}$,

$$L_j(\tau) = \prod_{k=0, k\neq j}^{2n} \frac{z - x_k}{x_j - x_k} = \prod_{k=0, k\neq j}^{2n} \frac{e^{i\tau} - e^{it_k}}{e^{it_j} - e^{it_k}}$$

e da fórmula de Newton $\phi = \mathbf{y}_{[x_0,...,x_n]} \cdot \mathbf{w}$,

$$\phi(\tau) = \sum_{k=0}^{2n} y_{[x_0,...,x_k]}(e^{i\tau} - e^{it_0}) \cdots (e^{i\tau} - e^{it_{k-1}})$$

com

$$y_{[x_0,...,x_k]} = \frac{y_{[x_1,...,x_k]} - y_{[x_0,...,x_{k-1}]}}{e^{it_k} - e^{it_0}}.$$

1.4.3 Nós igualmente espaçados

Num caso simples, de N nós igualmente espaçados em $[0, 2\pi[$,

$$t_k = \frac{2\pi k}{N}, \text{ com } k = 0, ..., N-1,$$

observamos que $(e^{it_k})^m = e^{2\pi i \frac{k}{N} m} = (e^{it_m})^k$ e considerando $\mathbf{w} = \{1, e^{it}, ..., e^{it(N-1)}\}$, a matriz $\mathbf{W} = \mathbf{w}(\mathbf{t})$ tem uma forma simplificada, simétrica,

$$\mathbf{W} = \begin{bmatrix} 1 & e^{it_0} & \cdots & (e^{it_0})^{N-1} \\ 1 & e^{it_1} & \cdots & (e^{it_1})^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{it_{N-1}} & \cdots & (e^{it_{N-1}})^{N-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{2\pi i \frac{1}{N}} & \cdots & e^{2\pi i \frac{(N-1)}{N}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{2\pi i \frac{(N-1)}{N}} & \cdots & e^{2\pi i \frac{(N-1)^2}{N}} \end{bmatrix}.$$

Podemos ainda verificar que $\mathbf{W}^* = \overline{\mathbf{W}}^\top$ está directamente relacionada com a matriz inversa, pois

$$\mathbf{W}^*\mathbf{W} = N\,\mathbf{I} = \mathbf{W}\mathbf{W}^*.$$

Esta igualdade resulta de $e^{it_k m} = e^{it_m k}$, verificando que $(k \neq m)$:

$$\begin{aligned} [\mathbf{W}\mathbf{W}^*]_{km} &= (1, e^{it_k}, ..., e^{it_k(N-1)}) \cdot \overline{(1, e^{it_1 m}, ..., e^{it_{N-1} m})} \\ &= (1, e^{it_k}, ..., e^{it_k(N-1)}) \cdot (1, e^{-it_m}, ..., e^{-it_m(N-1)}) \\ &= 1 + e^{i(t_k - t_m)} + \cdots + e^{i(t_k - t_m)(N-1)}, \text{ e se } k \neq m, \\ &= \frac{1 - e^{i(t_k - t_m)N}}{1 - e^{i(t_k - t_m)}} = \frac{1 - e^{2\pi i(k-m)}}{1 - e^{i(t_k - t_m)}} = \frac{1-1}{1 - e^{i(t_k - t_m)}} = 0 \end{aligned}$$

e no caso $k = m$, os elementos da diagonal, obtemos obviamente $[\mathbf{W}\mathbf{W}^*]_{kk} = 1 + e^0 + \cdots + e^0 = N$.

Assim, a solução do sistema $\mathbf{w}(\mathbf{t})\mathbf{a} = \mathbf{y}$ é simplesmente

$$\mathbf{a} = \frac{1}{N}\mathbf{w}(\mathbf{t})^*\mathbf{y} \Leftrightarrow a_k = \frac{1}{N}\sum_{m=0}^{N-1} y_m e^{-it_m k}$$

e o problema de interpolação trigonométrica tem uma solução imediata na forma complexa,

$$\phi(\tau) = \mathbf{u}(\tau) \cdot \mathbf{a} = \frac{1}{N}(\mathbf{W}^*\mathbf{y}) \cdot \mathbf{u}(\tau), \text{ para } \tau \in [0, 2\pi[.$$

• Para obter a expressão na forma real, quando todos os dados são reais, basta considerar a parte real da função.

Exercício 6. Calcular o interpolador trigonométrico que verifique
$\phi(0) = 1, \phi(\frac{\pi}{2}) = 0, \phi(\pi) = -1, \phi(\frac{3\pi}{2}) = 1$.

Resolução: Neste caso $N = 4$, e as funções base são $u_0(t) = 1, u_1(t) = e^{it}, u_2(t) = e^{2it}, u_3(t) = e^{3it}$. O sistema de interpolação é

$$\mathbf{Wa} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix} \mathbf{a} = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}$$

cuja solução sai assim de forma simples,$\mathbf{a} = \frac{1}{N}\mathbf{W}^*\mathbf{y}$,

$$\mathbf{a} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}\begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 1 \\ 2+i \\ -1 \\ 2-i \end{bmatrix}$$

por isso, $\phi(t) = \Re[\frac{1}{4}(1, 2+i, -1, 2-i)\cdot(1, e^{it}, e^{2it}, e^{3it})] = \Re[\frac{1}{4}(1 + (2+i)e^{it} - e^{2it} + (2-i)e^{3it})] =$
$= \frac{1}{4}(1 + 2\cos(t) - \sin(t) - \cos(2t) + 2\cos(3t) + \sin(3t)).\square$

Observação 5. Há outra possibilidade para considerar apenas a parte real. Considerando $N = 2n+1$, e usando a periodicidade, podemos substituir a parte de indíces $\{n+1, ..., 2n\}$ por $\{-n, ..., -1\}$ sem afectar os cálculos, pois $e^{2\pi i\frac{n+k}{2n+1}} = e^{2\pi i\frac{2n+1+k-1-n}{2n+1}} = e^{2\pi i\frac{k-1-n}{2n+1}}$, e basta considerar a mudança dos índices $n+k$ para $k-1-n$, com $k = 1, ..., n$. Consequentemente a relação

$$a_k = \frac{1}{2n+1}\sum_{m=-n}^{n} y_{m+n}e^{-it_m k}$$

é ainda válida, e obtém-se para $k = 0, ..., n$,

$$\begin{aligned} c_k &= 2Re(a_k) = \frac{2}{2n+1}\sum_{m=0}^{2n} y_m\cos(kt_m) \\ b_k &= -2Im(a_k) = \frac{2}{2n+1}\sum_{m=0}^{2n} y_m\sin(kt_m). \end{aligned}$$

1.4.4 Transformação de Fourier Discreta

Definição 2. Dada uma lista $\mathbf{y}$ de valores (reais ou complexos), associados a nós igualmente espaçados $\mathbf{t} = \frac{2\pi}{N}(0, 1, \cdots, N-1)$, e à lista de funções $\mathbf{u}(t) = \{1, e^{it}, ..., e^{it(N-1)}\}$, designa-se $\mathbf{u(t)}^*\mathbf{y}$ a sua transformada de Fourier Discreta,

$$\begin{array}{rccl} \mathcal{F}: & \mathbb{C}^N & \to & \mathbb{C}^N \\ & \mathbf{y} & \longmapsto \mathcal{F}\mathbf{y} = & \mathbf{u(t)}^*\mathbf{y} = \left[\sum_{j=0}^{N-1} y_j e^{-\frac{2\pi i}{N}kj}\right]_k \end{array}$$

Desta forma, a solução do problema de interpolação trigonométrica pode ser escrita na forma $\phi(t) = \frac{1}{N}(\mathcal{F}\mathbf{y})\cdot\mathbf{u}(t)$.

Como $\mathbf{u(t)}^*\mathbf{u(t)} = N\,\mathbf{I}$, é claro que se $\mathcal{F}\mathbf{y} = \mathbf{u(t)}^*\mathbf{y} = \mathbf{z}$, então $\mathbf{y} = \frac{1}{N}\mathbf{u(t)z}$, e a transformada de Fourier inversa é dada por

$$\mathcal{F}^{-1}\mathbf{z} = \frac{1}{N}\mathbf{u(t)z}.$$

Observação 6. Há ainda uma forma alternativa, normalizada, de apresentar a transformada de Fourier discreta e a sua inversa (e que é usada no *Mathematica* com a rotina Fourier[*lista*]):

$$\hat{\mathcal{F}}\mathbf{y} = \frac{1}{\sqrt{N}}\mathbf{u(t)}^*\mathbf{y} \quad , \quad \hat{\mathcal{F}}^{-1}\mathbf{z} = \frac{1}{\sqrt{N}}\mathbf{u(t)z}.$$

Notamos que desta forma, a matriz $\mathbf{U} = \frac{1}{\sqrt{N}}\mathbf{u(t)}$ fica unitária, pois $\mathbf{U}^*\mathbf{U} = \mathbf{I}$.

A transformação de Fourier discreta tem propriedades semelhantes à transformação de Fourier usual.

Proposição 3. *Temos as igualdades de Plancherel e Parseval (discretas):*

$$\frac{1}{N}\langle(\mathcal{F}\mathbf{y}),(\mathcal{F}\mathbf{z})\rangle = \langle\mathbf{y},\mathbf{z}\rangle\,; \qquad \frac{1}{\sqrt{N}}||\mathcal{F}\mathbf{y}||_2 = ||\mathbf{y}||_2.$$

Demonstração. (Exercício). Relembramos que o produto interno complexo é definido com o conjugado $\langle\mathbf{y},\mathbf{z}\rangle = \bar{\mathbf{y}}\cdot\mathbf{z} = (\bar{\mathbf{y}})^\top\mathbf{z} = \mathbf{y}^*\mathbf{z}$,logo

$$\begin{aligned}\frac{1}{N}\langle(\mathcal{F}\mathbf{y}),(\mathcal{F}\mathbf{z})\rangle &= \frac{1}{N}(\mathbf{u}(\mathbf{t})^*\mathbf{y})^*(\mathbf{u}(\mathbf{t})^*\mathbf{z}) = \frac{1}{N}(\mathbf{y}^*(\mathbf{u}(\mathbf{t})^*)^*\mathbf{u}(\mathbf{t})^*\mathbf{z})\\ &= \frac{1}{N}\mathbf{y}^*(\mathbf{u}(\mathbf{t})\mathbf{u}(\mathbf{t})^*)\mathbf{z} = \frac{1}{N}\mathbf{y}^*(N\mathbf{I})\mathbf{z} = \mathbf{y}^*\mathbf{z} = \langle\mathbf{y},\mathbf{z}\rangle\,.\end{aligned}$$

A igualdade de Parseval é agora imediata, $\frac{1}{N}||\mathcal{F}\mathbf{y}||_2^2 = \frac{1}{N}\langle(\mathcal{F}\mathbf{y}),(\mathcal{F}\mathbf{y})\rangle = \langle\mathbf{y},\mathbf{y}\rangle = ||\mathbf{y}||_2^2$.

□

1.4.5 Transformação de Fourier Rápida (FFT)

O número de operações necessário para o cálculo da Transformada de Fourier Discreta (TFD) consiste em N multiplicações e $N-1$ somas, para cada uma das N componentes, o que implica $O(N^2)$ operações. Esse valor pode ser reduzido significativamente através de algoritmos mais eficazes exigindo apenas $O(N\log_2 N)$ operações, e que são designados normalmente por Transformações de Fourier Rápidas (ou FFT, do inglês *Fast Fourier Transform*). Desta forma, o cálculo de uma TFD para $N = 1000$ pontos, pode reduzir-se de 1 milhão para 10 mil, aproximadamente.

A ideia poderá remontar a Gauss (séc. XIX), mas foi reintroduzida no contexto actual por Cooley e Tukey na década de 1960, e consiste na utilização de um processo recursivo. Consideremos o caso em que $N = 2^M$, o cálculo da TFD pode ser sucessivamente decomposto no cálculo de duas TFD com metade dos pontos. De facto, sendo $\mathbf{Y} = \mathcal{F}\mathbf{y}$, temos

$$\begin{aligned}Y_{2k} &= \sum_{m=0}^{\frac{N}{2}-1}(y_m + y_{m+N/2})e^{-2\pi i\frac{2km}{N}},\ \text{para } k = 0,...,\frac{N}{2}-1\\ Y_{2k+1} &= \sum_{m=0}^{\frac{N}{2}-1}(y_m - y_{m+N/2})e^{-2\pi i\frac{(2k+1)m}{N}},\ \text{para } k = 0,...,\frac{N}{2}-1\end{aligned}$$

o que resulta de

$$\begin{aligned}Y_k &= \sum_{j=0}^{N-1} y_j e^{-2\pi i\frac{kj}{N}} = \sum_{j=0}^{\frac{N}{2}-1} y_j e^{-2\pi i\frac{kj}{N}} + \sum_{j=\frac{N}{2}}^{N-1} y_j e^{-2\pi i\frac{kj}{N}}\\ &= \sum_{m=0}^{\frac{N}{2}-1} y_m e^{-2\pi i\frac{km}{N}} + \sum_{m=0}^{\frac{N}{2}-1} y_{m+N/2} e^{-2\pi i\frac{k(m+N/2)}{N}}\\ &= \sum_{m=0}^{\frac{N}{2}-1}(y_m + y_{m+N/2}e^{-k\pi i})e^{-2\pi i\frac{km}{N}}\end{aligned}$$

e como $e^{-k\pi i} = (-1)^k$, a igualdade é diferente para termos pares e ímpares. Agora basta reparar que

$$Y_{2k} = \sum_{m=0}^{\frac{N}{2}-1} (y_m + y_{m+N/2}) e^{-2\pi i \frac{km}{N/2}} = \mathcal{F}\mathbf{y}^{[N/2,0]}$$

em que $\mathbf{y}_m^{[N/2,0]} = y_m + y_{m+N/2}$ dá um vector com metade da dimensão original, e da mesma forma

$$Y_{2k+1} = \sum_{m=0}^{\frac{N}{2}-1} (y_m - y_{m+N/2}) e^{-2\pi i \frac{m}{N}} e^{-2\pi i \frac{km}{N/2}} = \mathcal{F}\mathbf{y}^{[N/2,1]}$$

em que $\mathbf{y}_m^{[N/2,1]} = (y_m - y_{m+N/2}) e^{-\pi i \frac{m}{N/2}}$ dá igualmente um vector com metade da dimensão. Portanto $\mathcal{F}\mathbf{y}$ pode ser obtido através do cálculo de $\mathcal{F}\mathbf{y}^{[N/2,0]}$ e de $\mathcal{F}\mathbf{y}^{[N/2,1]}$, recorrendo a algumas operações extra, $\frac{N}{2}$ somas para o cálculo dos termos pares, $\frac{N}{2}$ subtracções e multiplicações para o cálculo dos termos ímpares. De forma sucessiva, $\mathcal{F}\mathbf{y}^{[N/2,0]}$ poderá ser calculado recorrendo a $\mathcal{F}\mathbf{y}^{[N/4,00]}, \mathcal{F}\mathbf{y}^{[N/4,01]}$, e $\mathcal{F}\mathbf{y}^{[N/2,1]}$ poderá ser calculado recorrendo a $\mathcal{F}\mathbf{y}^{[N/4,10]}$ e $\mathcal{F}\mathbf{y}^{[N/4,11]}$. Após $M-1$ passos chegamos ao cálculo de $\mathbf{y}^{[N/2^{M-1},\cdots]}$ e $N/2^{M-1} = 2$. Concluímos que após $M-1$ passos com N somas e $N/2$ multiplicações intermédias, bastará calcular os valores de $\mathcal{F}\mathbf{y}^{[2;0\ldots0]}, \ldots, \mathcal{F}\mathbf{y}^{[2;1\ldots1]}$, em que cada um exige apenas 1 soma e 1 multiplicação (num total de N). Logo o número de operações total envolvido é inferior a $2NM$, ou seja é $O(N \log_2 N)$. Para além disso, a notação $\mathbf{y}^{[2;0101101]}$ permite obter imediatamente o índice respectivo, escrevendo o índice pretendido k na notação binária 0101101. Há múltiplas variantes, considerando outras bases, e outras ordenações.

1.4.6 Exemplos de TFD

Exercício 7. Mostre algumas propriedades da TFD:
i) $[\mathcal{F}\left(\binom{N-1}{n}\right)]_k = \left(1 + \exp(\frac{-2\pi i}{N}k)\right)^{N-1}$
ii) Se $\mathbf{v} \in \mathbb{R}^N$ então $[\overline{\mathcal{F}(\mathbf{v})}]_k = [\mathcal{F}(\mathbf{v})]_{N-k}$
iii) Fixo m, $[\mathcal{F}\left(v_n e^{\frac{2\pi i}{N}nm}\right)]_k = [\mathcal{F}(\mathbf{v})]_{k-m}$

Resolução:
(i) Temos $[\mathcal{F}\binom{N-1}{n}]_k = \sum_{j=0}^{N-1} \binom{N-1}{j} e^{-\frac{2\pi i}{N}kj} = \sum_{j=0}^{N-1} \binom{N-1}{j} \left(e^{-\frac{2\pi i}{N}k}\right)^j = \left(1 + e^{-\frac{2\pi i}{N}k}\right)^{N-1}$ usando a expressão do binómio de Newton $(1+c)^{N-1} = \sum_{j=0}^{N-1} \binom{N-1}{j} c^j$.
(ii) Como $v_j \in \mathbb{R}$,

$$[\mathcal{F}(\mathbf{v})]_{N-k} = \sum_{j=0}^{N-1} v_j e^{-\frac{2\pi i}{N}(N-k)j} = \sum_{j=0}^{N-1} v_j \underbrace{e^{-\frac{2\pi i}{N}Nj}}_{=1} e^{\frac{2\pi i}{N}kj} = \sum_{j=0}^{N-1} \overline{v_j e^{-\frac{2\pi i}{N}kj}} = \overline{[\mathcal{F}(\mathbf{v})]_k}.$$

(iii) $[\mathcal{F}(v_n e^{\frac{2\pi i}{N}nm})]_k = \sum_{j=0}^{N-1} (v_j e^{\frac{2\pi i}{N}jm}) e^{-\frac{2\pi i}{N}kj} = \sum_{j=0}^{N-1} v_j e^{-\frac{2\pi i}{N}(k-m)j} = [\mathcal{F}(v_n)]_{k-m}$

Exercício 8. Seja $\Delta v_n = v_{n+1} - v_n$. Mostre que:
(i) Fixo m, $[\mathcal{F}(v_{n+m})]_k = e^{\frac{2\pi i}{N}km}[\mathcal{F}(\mathbf{v})]_k$
(ii) Fixo p, $[\mathcal{F}(\Delta^p v_n)]_k = (e^{\frac{2\pi i}{N}k} - 1)^p [\mathcal{F}(\mathbf{v})]_k$

Resolução:
(i) $[\mathcal{F}(v_{n+m})]_k = \sum_{j=0}^{N-1} v_{j+m} e^{-\frac{2\pi i}{N}kj} = \sum_{j=m}^{N+m-1} v_j e^{-\frac{2\pi i}{N}k(j-m)}$

$= e^{\frac{2\pi i}{N}km} \sum_{j=m}^{N+m-1} v_j e^{-\frac{2\pi i}{N}kj} = e^{\frac{2\pi i}{N}km}[\mathcal{F}(v_n)]_k$
(ii) Usamos (i) com $m = 1$, e para $p = 1$ obtemos:

$$[\mathcal{F}(\Delta v_n)]_k = [\mathcal{F}(v_{n+1})]_k - [\mathcal{F}(v_n)]_k = (e^{\frac{2\pi i}{N}k} - 1)[\mathcal{F}(v_n)]_k$$

.
Basta agora reparar que $\Delta^{p+1}v_n = \Delta(\Delta^p v_n)$ e por indução em p :
$[\mathcal{F}\left(\Delta^{p+1}v_n\right)]_k = [\mathcal{F}(\Delta(\Delta^p v_n))]_k = (e^{\frac{2\pi i}{N}k} - 1)[\mathcal{F}(\Delta^p v_n)]_k = (e^{\frac{2\pi i}{N}k} - 1)\left((e^{\frac{2\pi i}{N}k} - 1)^p[\mathcal{F}(v_n)]_k\right) = (e^{\frac{2\pi i}{N}k} - 1)^{p+1}[\mathcal{F}(v_n)]_k$.
Esta propriedade é o correspondente discreto para a fórmula de derivação.

1.4.7 Propriedades da convolução vectorial com a TFD

A transformada de Fourier discreta, através da FFT, permite calcular rapidamente um produto de convolução, devido à seguinte propriedade.

Teorema 3. *Seja* $\mathbf{v} \bullet \mathbf{w}$ *um produto de vectores definido por componentes* $[\mathbf{v} \bullet \mathbf{w}]_k = v_k w_k$. *Então verifica-se:*
(i) $\mathcal{F}(\mathbf{v} * \mathbf{w}) = \mathcal{F}(\mathbf{v}) \bullet \mathcal{F}(\mathbf{w})$, *o que implica* $\mathbf{v} * \mathbf{w} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{v}).\mathcal{F}(\mathbf{w}))$
(ii) $\mathcal{F}(\mathbf{v} \bullet \mathbf{w}) = \frac{1}{N}\mathcal{F}(\mathbf{v}) * \mathcal{F}(\mathbf{w})$

Demonstração. Sendo $\mathbf{y} = \mathbf{v} * \mathbf{w}$

$$[\mathcal{F}\mathbf{y}]_k = \sum_{j=0}^{N-1} y_j e^{-2\pi i\frac{kj}{N}} = \sum_{j=0}^{N-1}\sum_{m=0}^{N-1} v_m w_{j-m} e^{-2\pi i\frac{kj}{N}} = \sum_{m=0}^{N-1} v_m \sum_{j=0}^{N-1} w_{j-m} e^{-2\pi i\frac{kj}{N}}$$

$$= \sum_{m=0}^{N-1} v_m \sum_{r=-m}^{N-1-m} w_r e^{-2\pi i\frac{k(r+m)}{N}} = \sum_{m=0}^{N-1} v_m e^{-2\pi i\frac{km}{N}} \sum_{r=-m}^{N-1-m} w_r e^{-2\pi i\frac{kr}{N}} = [\mathcal{F}\mathbf{v}]_k[\mathcal{F}\mathbf{w}]_k.$$

A segunda igualdade é semelhante (Exercício). □

1.5 Operador de Interpolação Polinomial

Definição: Consideremos $\{x_0, ..., x_n\} \subset [a, b]$, então define-se $\mathcal{L}$ o operador de interpolação polinomial de Lagrange associado a esses pontos,

$$\begin{array}{rccc} \mathcal{L}: & C[a,b] & \to & \mathcal{P}_n \subset C[a,b] \\ & f & \longmapsto & p_n \end{array}$$

Notamos que $\mathcal{L}$ é um operador linear, porque sendo p_n o polinómio interpolador para f e q_n o polinómio interpolador para g, então

$$(\alpha f + \beta g)(x_i) = \alpha f(x_i) + \beta g(x_i) = \alpha p_n(x_i) + \beta q_n(x_i) = (\alpha p_n + \beta q_n)(x_i)$$

e portanto $\alpha p_n + \beta q_n$ é o (único) polinómio interpolador de grau n, logo $\mathcal{L}(\alpha f + \beta g) = \alpha p_n + \beta q_n = \alpha\mathcal{L}(f) + \beta\mathcal{L}(g)$.

Para além disso, notamos que se trata de uma projecção, pois $\mathcal{L}^2 = \mathcal{L}$, já que

$$\mathcal{L}^2(f) = \mathcal{L}\left(\mathcal{L}(f)\right) = \mathcal{L}(p_n) = p_n = \mathcal{L}(f), \forall f \in C[a, b],$$

notando que é imediato que $\mathcal{L}(p_m) = p_m$, se o grau m for menor ou igual a n.

Introduzimos também o funcional *delta de Dirac,* para $z \in [a,b]$,

$$\begin{array}{rccc} \delta_z : & C[a,b] & \to & \mathbb{R} \\ & f & \longmapsto & f(z) \end{array}$$

e observamos que, usando a fórmula de Lagrange, temos

$$\mathcal{L}f = \sum_{k=0}^{n} f(x_k) l_k = \sum_{k=0}^{n} l_k \delta_{x_k} f$$

podendo escrever-se abreviadamente $\mathcal{L} = \mathbf{l} \cdot \delta_{\mathbf{x}}$, com $\mathbf{l} = (l_0, ..., l_n)$ e $\delta_{\mathbf{x}} = (\delta_{x_0}, ..., \delta_{x_n})$.

O operador $\mathcal{L}$ é um operador contínuo, pertencendo a $L(C[a,b])$, podendo mesmo ser determinada a sua norma.

Proposição 4. $||\mathcal{L}||_{L(C[a,b])} = \Lambda_n$, *em que*

$$\Lambda_n = \left\| \sum_{k=0}^{n} |l_k| \right\|_\infty \quad \textit{é a constante de Lebesgue.}$$

Demonstração. Recordamos que $||\mathcal{L}||_{L(C[a,b])} = \sup_{f \neq 0} \frac{||\mathcal{L}f||_\infty}{||f||_\infty}$, e como

$$\begin{aligned} ||\mathcal{L}f||_\infty &= \max_{t\in[a,b]} |\sum_{k=0}^{n} f(x_k) l_k(t)| \leq \max_{t\in[a,b]} \sum_{k=0}^{n} |f(x_k)||l_k(t)| \\ &\leq \max_{t\in[a,b]} \sum_{k=0}^{n} ||f||_\infty |l_k(t)| = ||f||_\infty \max_{t\in[a,b]} \sum_{k=0}^{n} |l_k(t)| = \Lambda_n ||f||_\infty \end{aligned}$$

concluímos que $||\mathcal{L}||_{L(C[a,b])} \leq \Lambda_n$.
Por outro lado, escolhendo um ponto $t^* : \sum_{k=0}^{n} |l_k(t^*)| = \Lambda_n$, podemos considerar uma função $s \in C[a,b]$: $s(x_k) = sign(l_k(t^*))$, e $||s||_\infty = 1$ (por exemplo, s toma valores em $[-1,1]$, podendo considerar uma função seccionalmente $\mathcal{P}_1$, unindo por uma linha os pontos $(x_k, s(x_k))$, ou seja, um spline linear). Dessa forma $s(x_k) l_k(t^*) = sign(l_k(t^*)) l_k(t^*) = |l_k(t^*)| \geq 0$, logo

$$\begin{aligned} ||\mathcal{L}s||_\infty &= \max_{t\in[a,b]} |\sum_{k=0}^{n} s(x_k) l_k(t)| \geq |\sum_{k=0}^{n} s(x_k) l_k(t^*)| \\ &= \sum_{k=0}^{n} s(x_k) l_k(t^*) = \sum_{k=0}^{n} |l_k(t^*)| = \Lambda_n \end{aligned}$$

e concluímos que $||\mathcal{L}||_{L(C[a,b])} = \sup_{f\neq 0} \frac{||\mathcal{L}f||_\infty}{||f||_\infty} \geq \frac{||\mathcal{L}s||_\infty}{||s||_\infty} = \Lambda_n$.

□

Observação 7. *Estabilidade da interpolação de Lagrange.*
Consideremos duas funções f e $\tilde{f}$, em que normalmente $\tilde{f}$ é considerada uma perturbação ou aproximação de f. O polinómio interpolador p_n associado a f será diferente de $\tilde{p}_n$ associado a $\tilde{f}$. A constante de Lebesgue permite controlar a influência que um erro em f tem no cálculo do polinómio interpolador,

$$\begin{aligned} ||p_n - \tilde{p}_n||_\infty &= ||\mathcal{L}f - \mathcal{L}\tilde{f}||_\infty = ||\mathcal{L}(f - \tilde{f})||_\infty \\ &\leq ||\mathcal{L}||_{L(C[a,b])} ||f - \tilde{f}||_\infty = \Lambda_n ||f - \tilde{f}||_\infty. \end{aligned}$$

Exercício 9. Mostre que sendo $\tilde{f} = f * \mu_\varepsilon$, em que μ_ε é um filtro, e sendo p_n (resp. $\tilde{p}_n$) o polinómio interpolador de f (resp. $\tilde{f}$) nos nós $x_0, ..., x_n \in [a+\varepsilon, b-\varepsilon]$, temos para $f \in C^1[a,b]$:

$$||p_n - \tilde{p}_n||_\infty \leq \Lambda_n \left\|f'\right\|_\infty \varepsilon.$$

Resolução: Pela observação anterior, basta mostrar que $||f - \tilde{f}||_\infty \leq \left\|f'\right\|_\infty \varepsilon$. Como vimos em (1.3.6) que $\tilde{f}(y) = f * \mu_\varepsilon(y) = f(\tilde{y})$, com $\tilde{y} \in [y-\varepsilon, y+\varepsilon]$, temos pelo teorema de Lagrange

$$\left|f(y) - \tilde{f}(y)\right| = |f(y) - f(\tilde{y})| = |f'(\xi)|\,|y - \tilde{y}| \leq \left\|f'\right\|_\infty \varepsilon.$$

1.6 Interpolação com Splines

A interpolação polinomial clássica pode levar a problemas de instabilidade, pois ao aumentar o número de nós aumentamos o grau do polinómio interpolador. Para evitar isso, podemos fazer uma partição do intervalo e considerar funções seccionalmente polinomiais. Se estas funções colarem com regularidade, então somos levados à noção de spline:

Definição 3. Dada uma partição do intervalo $[a,b] = \cup_{k=1}^{N} [x_{k-1}, x_k]$ com $X = \{x_0, ..., x_N\} \in [a,b]$, designamos *spline* de ordem $r \geq 1$ uma função $s \in \mathcal{S}_r(X)$:
(i) $s \in C^{r-1}[a,b]$,
(ii) $s_{|[x_{k-1},x_k]} \in \mathbb{P}_r$, para $k = 1, ..., N$ (ou seja, s é um polinómio de grau r em cada sub-intervalo $[x_{k-1}, x_k]$.

Como um spline é apenas uma função seccionalmente polinomial, a interpolação por splines não obriga a aumentar o grau do polinómio interpolador cada vez que aumentamos o número de pontos de interpolação. Iremos ver os dois casos mais habituais: interpolação por splines lineares ($r = 1$) e por splines cúbicos ($r = 3$).

1.6.1 Splines Lineares $\mathcal{S}_1$

Este é o caso mais simples.

Procurar a função $s \in \mathcal{S}_1(X) : s(X) = f(X)$, resume-se a considerar $s_{|[x_{k-1},x_k]}(x) = f_{k-1} + f[x_{k-1}, x_k](x - x_{k-1})$, para $x \in [x_{k-1}, x_k]$.

Uma base para os splines lineares consiste nas funções

$$g_k(x) = \begin{cases} \frac{x-x_{k-1}}{x_k - x_{k-1}}, \text{ se } x \in [x_{k-1}, x_k] \\ \frac{x-x_{k+1}}{x_k - x_{k+1}}, \text{ se } x \in [x_k, x_{k+1}] \\ 0, \text{ se } x \notin [x_{k-1}, x_{k+1}] \end{cases} \quad (k = 1, \ldots, N-1)$$

$$g_0(x) = \begin{cases} \frac{x-x_1}{x_0 - x_1}, \text{ se } x \in [x_0, x_1] \\ 0, \text{ se } x \notin [x_0, x_1] \end{cases} \quad g_N(x) = \begin{cases} \frac{x-x_{N-1}}{x_N - x_{N-1}}, \text{ se } x \in [x_{N-1}, x_N] \\ 0, \text{ se } x \notin [x_{N-1}, x_N] \end{cases}$$

que tornam a matriz de Vandermonde a identidade, escrevendo-se

$$s(x) = f_0 g_0(x) + \ldots + f_N g_N(x).$$

Alternativamente, podemos escrever directamente

$$x \in [x_{k-1}, x_k] \implies s(x) = f_{k-1} + f[x_{k-1}, x_k](x - x_{k-1}).$$

Erro de Interpolação por Splines Lineares
Seja $f \in C^2[a,b]$. Em cada subintervalo, $x \in [x_{k-1}, x_k]$ temos

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

O máximo valor de $|w(x)| = |x - x_k||x - x_{k-1}|$ é atingido no ponto médio $x = \frac{x_{k-1}+x_k}{2}$, logo designando $h_{k-1} = x_k - x_{k-1}$
obtemos $\max_{x\in[a,b]} |w(x)| = \frac{h_{k-1}^2}{4}$ e a estimativa de erro para a interpolação por splines lineares é (considerando $h = \max_{k=0}^{N-1} |h_k|$)

$$||f - s||_\infty \leq \frac{||f''||_\infty}{8} h^2. \tag{1.6.1}$$

Teorema 4. *O conjunto $\mathcal{S}_r(X)$ dos splines de grau r é um espaço vectorial de dimensão $N+r$.*

Demonstração. Através da derivada de ordem $r-1$ obtemos $s^{(r-1)}$ como função contínua, que é seccionalmente $\mathbb{P}_1$, ou seja pode ser definida por um spline linear, que tem $N+1$ funções base. A primitivação de $s^{(r-1)}$ até s acrescenta $r-1$ incógnitas que definem a dimensão adicional do espaço, e no total a dimensão é $N + 1 + r - 1 = N + r$. □

Observação 8. Se aumentarmos o grau para 2, podemos definir splines de grau 2, que omitimos a dedução, por ser semelhante à de grau 3, que apresentaremos pois é mais utilizada por minimizar a curvatura da função interpoladora.

Notamos que se forem apenas consideradas funções seccionalmente polinomiais de grau 2 (sem serem splines), é normal fazer uma partição com N par, definindo o polinómio interpolador de grau 2 usando os subintervalos $[x_{2k-2}, x_{2k}]$ com $k = 1, ..., N/2$. Por exemplo, usando a fórmula de Newton, temos para $x \in [x_{2k-2}, x_{2k}]$

$$\phi(x) = f_{2k-2} + f[x_{2k-2}, x_{2k-1}](x - x_{2k-2}) + f[x_{2k-2}, x_{2k-1}, x_{2k}](x - x_{2k-2})(x - x_{2k-1}),$$

mas esta expressão não dá um spline de grau 2, pois apesar da função ser contínua, não é exigido que a derivada o seja. É esta aproximação seccionalmente $\mathbb{P}_2$ que é usada na regra de integração de Simpson.

1.6.2 Splines Cúbicos $\mathcal{S}_3$

Começamos por explicitar a dedução do sistema que permite calcular os splines cúbicos. Começamos por recordar que a segunda derivada será um spline linear. Assim, para $x \in [x_k, x_{k+1}]$ temos $s''(x) = s''_k + (x - x_k)s''_{[x_k,x_{k+1}]}$, e primitivando

$$s'(x) = s'_k + (x - x_k)s''_k + \frac{1}{2}(x - x_k)^2 s''_{[x_k,x_{k+1}]} \tag{1.6.2}$$

e daqui, em x_{k+1}, temos $s'_{k+1} = s'(x_{k+1})$ dado por

$$s'_{k+1} = s'_k + s''_k h_k + \frac{h_k^2}{2} s''_{[x_k,x_{k+1}]} = s'_k + s''_k h_k + \frac{h_k}{2}(s''_{k+1} - s''_k) \tag{1.6.3}$$

e primitivando (1.6.2), temos

$$s(x) = s_k + (x - x_k)s'_k + \frac{1}{2}(x - x_k)^2 s''_k + \frac{1}{6}(x - x_k)^3 s''_{[x_k, x_{k+1}]} \tag{1.6.4}$$

Logo $s_{k+1} = s_k + h_k s'_k + \frac{h_k^2}{2} s''_k + \frac{h_k^3}{6} s''_{[x_k, x_{k+1}]}$, e por s interpolar f temos $f_{[x_k, x_{k+1}]} = \frac{s_{k+1} - s_k}{h_k}$, assim

$$f_{[x_k, x_{k+1}]} = s'_k + \frac{h_k}{2} s''_k + \frac{h_k}{6}(s''_{k+1} - s''_k) = s'_k + \frac{h_k}{3} s''_k + \frac{h_k}{6} s''_{k+1} \tag{1.6.5}$$

$$f_{[x_k, x_{k+1}]} - f_{[x_{k-1}, x_k]} = s'_k - s'_{k-1} + \frac{h_k}{3} s''_k + \frac{h_k}{6} s''_{k+1} - \frac{h_{k-1}}{3} s''_{k-1} - \frac{h_{k-1}}{6} s''_k$$

De (1.6.3) temos $s'_k - s'_{k-1} = s''_{k-1} h_{k-1} + \frac{h_{k-1}}{2}(s''_k - s''_{k-1})$, e agrupando em s''_j

$$f_{[x_k, x_{k+1}]} - f_{[x_{k-1}, x_k]} = \frac{h_{k-1}}{6} s''_{k-1} + \frac{h_{k-1}}{3} s''_k + \frac{h_k}{3} s''_k + \frac{h_k}{6} s''_{k+1}. \tag{1.6.6}$$

Obtemos um sistema tridiagonal para calcular os valores $s''_k = s''(x_k)$.

– **Condições na derivada:** $s'_0 = s'(a) = f'_0$, $s'_N = s'(b) = f'_N$

$$\begin{bmatrix} \frac{h_0}{3} & \frac{h_0}{6} & 0 & \cdots & 0 \\ \frac{h_0}{6} & \frac{h_0+h_1}{3} & \frac{h_1}{6} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \frac{h_{N-2}+h_{N-1}}{3} & \frac{h_{N-1}}{6} \\ 0 & \cdots & 0 & \frac{h_{N-1}}{6} & \frac{h_{N-1}}{3} \end{bmatrix} \begin{bmatrix} s''_0 \\ s''_1 \\ \vdots \\ s''_{N-1} \\ s''_N \end{bmatrix} = \begin{bmatrix} f_{[x_0,x_1]} - f'_0 \\ f_{[x_1,x_2]} - f_{[x_0,x_1]} \\ \vdots \\ f_{[x_{N-1},x_N]} - f_{[x_{N-2},x_{N-1}]} \\ f'_N - f_{[x_{N-1},x_N]} \end{bmatrix}$$

Notando que (1.6.6) se aplica para $k = 1, .., N-1$, enquanto para a primeira equação ($k = 0$) aplicamos (1.6.5) para obter $f_{[x_0,x_1]} = s'_0 + \frac{h_k}{3} s''_0 + \frac{h_k}{6} s''_1$.
De forma semelhante, seria possível obter $f_{[x_k,x_{k+1}]} = s'_{k+1} - \frac{h_k}{3} s''_{k+1} - \frac{h_k}{6} s''_k$, aplicando-se à última equação ($k + 1 = N$).

– **Condições naturais:** $s''_0 = s''(a) = 0$, $s''_N = s''(b) = 0$

$$\begin{bmatrix} \frac{h_0+h_1}{3} & \frac{h_1}{6} & 0 & \cdots & 0 \\ \frac{h_1}{6} & \frac{h_1+h_2}{3} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{h_{N-2}}{6} \\ 0 & \cdots & 0 & \frac{h_{N-2}}{6} & \frac{h_{N-2}+h_{N-1}}{3} \end{bmatrix} \begin{bmatrix} s''_1 \\ \vdots \\ \vdots \\ s''_{N-1} \end{bmatrix} = \begin{bmatrix} f_{[x_1,x_2]} - f_{[x_0,x_1]} \\ \vdots \\ \vdots \\ f_{[x_{N-1},x_N]} - f_{[x_{N-2},x_{N-1}]} \end{bmatrix}$$

Partindo destes s''_k é possível obter, de (1.6.5),

$$s'_k = f[x_k, x_{k+1}] - \frac{h_k}{6}(2s''_k + s''_{k+1})$$

e de (1.6.4) a expressão de $s_{|[x_k,x_{k+1}]} \in \mathcal{P}_3$, para $x \in [x_k, x_{k+1}]$ $_{(k=0,\ldots,N-1)}$:

$$s(x) = f_k + (x - x_k)s'_k + (x - x_k)^2 \frac{s''_k}{2} + (x - x_k)^3 \frac{s''[x_k, x_{k+1}]}{6}.$$

Exercício 10. Verifique que o spline cúbico natural, que interpola os pontos $\begin{bmatrix} x_k \\ f_k \end{bmatrix} = \{\begin{bmatrix} -2 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 7 \end{bmatrix}\}$, é dado por

$$s(x) = 1 + x + x^2 + \begin{cases} \frac{1}{4}x^2(2+x), \text{ se } x \in [-2,0] \\ \frac{1}{4}x^2(2-x), \text{ se } x \in [0,+2] \end{cases}$$

Resolução: Basta reparar que em ambos os troços é um polinómio cúbico tendo-se

$$s'(x) = 1 + 2x + \begin{cases} x + \frac{3}{4}x^2, \text{ se } x \in [-2,0] \\ x - \frac{3}{4}x^2, \text{ se } x \in [0,+2] \end{cases}, \; s''(x) = 2 + \begin{cases} 1 + \frac{3}{2}x, \text{ se } x \in [-2,0] \\ 1 - \frac{3}{2}x, \text{ se } x \in [0,+2] \end{cases}$$

pelo que se verifica a continuidade C^2 em $x = 0$ (o único ponto de ligação)

$$s(0^-) = 1 = s(0^+),\; s'(0^-) = 1 = s'(0^+),\; s''(0^-) = 3 = s''(0^+).$$

tendo-se ainda $s''(-2) = 0 = s''(2)$. Neste caso reparamos que o sistema seria unidimensional, confirmando-se $s_1'' = 3$:

$$\frac{h_0 + h_1}{3} s_1'' = f_{[x_1,x_2]} - f_{[x_0,x_1]} \Leftrightarrow \frac{2+2}{3} s_1'' = 3 - (-1) \Leftrightarrow s_1'' = 3$$

e daqui obtemos $s_0^{'} = -1 - \frac{2}{6}(2s_0^{''} + s_1^{''}) = -2$, o que dá $s(x) = 3 - 2(x+2) + 0(x+2)^2 + \frac{3-0}{6\times 2}(x+2)^3$ que é a expressão em $[-2,0]$.

Exercício 11.

(i) Determinar o spline natural que verifica $s(\pm 2) = s(\pm 1) = 0, s(0) = 1$.

(ii) Analogamente, determine o spline com condições nulas sobre as derivadas.

Resolução: (i) Sendo $h_k = 1$, obtemos o sistema:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} s_1'' \\ s_2'' \\ s_3'' \end{bmatrix} = \begin{bmatrix} s_{[-1,0]} - s_{[-2,-1]} \\ s_{[0,1]} - s_{[-1,0]} \\ s_{[1,2]} - s_{[0,1]} \end{bmatrix} = \begin{bmatrix} 1 - 0 = 1 \\ -1 - 1 = -2 \\ 0 - (-1) = 1 \end{bmatrix}$$

cuja solução é $\frac{6}{7}(3,-5,3)$, obtendo-se $\mathbf{s''} = \frac{6}{7}(0,3,-5,3,0)$ e de $\mathbf{s'} = \frac{3}{7}(-1,2,0,-2,...)$ após cálculos, retiramos

$$s(x) = \frac{1}{7}\left\{ \underbrace{3(x+1)(x+2)(x+3)}_{x\in[-2,-1]}, \underbrace{7 - 15x^2 - 8x^3}_{x\in[-1,0]}, \underbrace{7 - 15x^2 + 8x^3}_{x\in[0,1]}, \underbrace{-3(x-1)(x-2)(x-3)}_{x\in[1,2]} \right\}.$$

(ii) Obtemos o sistema:

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} s_0'' \\ \vdots \\ s_4'' \end{bmatrix} = \begin{bmatrix} s_{[-2,-1]} - s_0' \\ s_{[-1,0]} - s_{[-2,-1]} \\ s_{[0,1]} - s_{[-1,0]} \\ s_{[1,2]} - s_{[0,1]} \\ s_4' - s_{[1,2]} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

cuja solução é $\mathbf{s''} = \frac{3}{2}(-1,2,-3,2,-1)$, obtendo-se

$$s(x) = \frac{1}{4}\left\{ 3\underbrace{(x+1)(x+2)^2}_{x\in[-2,-1]}, \underbrace{4 - 9x^2 - 5x^3}_{x\in[-1,0]}, \underbrace{4 - 9x^2 + 5x^3}_{x\in[0,1]}, \underbrace{-3(x-1)(x-2)^2}_{x\in[1,2]} \right\}$$

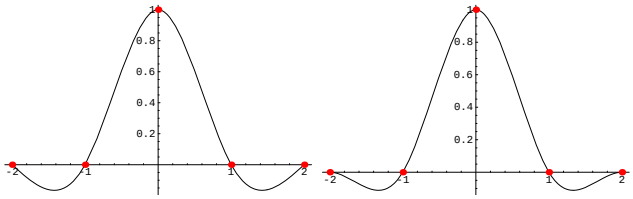

Figura 1.6.1: Spline cúbico natural, e spline com derivada nula, soluções do Exercício.

1.6.3 Estimativas sobre splines cúbicos

Os splines cúbicos verificam uma importante propriedade.

Teorema 5. *Considere qualquer $g \in C^2[a,b] : g(x_k) = f_k$. O spline cúbico interpolador $s(x_k) = f_k$ tal que $s''(a) = s''(b) = 0$, spline natural, (ou com condições nas derivadas $s'(a) = g'(a)$, $s'(b) = s'(b)$), verifica a propriedade de minimização*

$$\int_a^b |s''(x)|^2 dx \leq \int_a^b |g''(x)|^2 dx, \tag{1.6.7}$$

que resulta da igualdade

$$||g'' - s''||^2_{L^2[a,b]} = ||g''||^2_{L^2[a,b]} - ||s''||^2_{L^2[a,b]}. \tag{1.6.8}$$

Demonstração.

$$\int_a^b (g'' - s'')^2 dt = \int_a^b (g'')^2 dt - \int_a^b (s'')^2 dt + \int_a^b (2s''g'' - 2(s'')^2) dt$$

$$2\int_a^b (g'' - s'')s'' dt = 2[(g' - s')s'']_a^b - 2\int_a^b (g'' - s'')s''' dt = -2\sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} (g' - s')s''' dt$$

porque $(g' - s')s''$ é nulo em a e b, devido às condições naturais ou sobre as derivadas. Finalmente,

$$\int_{x_k}^{x_{k+1}} (g' - s')s''' dt = [(g - s)s''']_{x_k}^{x_{k+1}} - \int_{x_k}^{x_{k+1}} (g - s)s'''' dt = 0,$$

porque em cada subintervalo$[x_k, x_{k+1}]$ temos $s \in \mathbb{P}_3$ logo $s'''' = 0$,
e ainda temos $(g - s)(x_k) = f_k - f_k = 0$, por ser interpolador. □

Teorema 6. *Dada uma função $f \in C^2[a,b]$, sendo $h = \max_{k=0}^{N-1} h_k$, obtemos*

$$||f - s||_\infty \leq \frac{h^{3/2}}{2} ||f''||_{L^2[a,b]}, \quad \textit{e ainda } ||f' - s'||_\infty \leq h^{1/2} ||f''||_{L^2[a,b]}.$$

Quando $f \in C^4[a,b]$, obtemos estimativas de erro mais precisas:

$$||f - s||_\infty \leq \frac{h^4}{16} ||f^{(4)}||_\infty. \ \square$$

sendo ainda possível obter $||f - s||_\infty \leq \frac{5h^4}{384} ||f^{(4)}||_\infty$.

Demonstração. Uma vez que s interpola f, a função $r = f - s$ tem pelo menos $n+1$ zeros que são $\{x_0, ..., x_n\}$. Logo pelo teorema de Rolle a derivada r' tem pelo menos n zeros, que designamos $z_1, ..., z_n$ verificando-se $z_k \in [x_{k-1}, x_k]$. Por outro lado, a função $|r'|$ tem ponto de máximo em $z \in [x_{m-1}, x_m]$ (para certo m), verificando-se $|z - z_m| \leq h$.
Como $r'(z_m) = 0$, podemos escrever

$$r'(x) = \int_{z_m}^{x} r''(t)dt \implies ||r'||_\infty = \max_{t\in[a,b]} |r'(t)| = |r'(z)| = \left| \int_{z_m}^{z} r''(t)dt \right|$$

e pela desigualdade de Schwarz ($|\langle f, g\rangle| \leq ||f||\,||g||$, aplicada em $L^2[a,b]$)

$$||r'||_\infty^2 = \left| \int_{z_m}^{z} r''(t)dt \right|^2 \leq ||1||^2_{L^2[z_m,z]} ||r''||^2_{L^2[z_m,z]} \leq |z - z_m|^2 \left| \int_a^b r''(t)^2 dt \right| \leq h||r''||^2_{L^2[a,b]}$$

finalmente como vimos no teorema anterior $||r''||^2_{L^2[a,b]} = ||f'' - s''||^2_{L^2[a,b]} \leq ||f''||^2_{L^2[a,b]}$ concluindo-se $||r'||^2_\infty \leq h||f''||^2_{L^2[a,b]}$ e a estimativa no erro da derivada.
A estimativa no erro da função, é semelhante *(Exercício)*:
Agora a função $|r|$ tem ponto de máximo em $w \in [x_{p-1}, x_p]$ (para certo p), verificando-se $|w - w_p| \leq h/2$, com $w_p = x_p$ ou com $w_p = x_{p-1}$. Assim,

$$r(x) = \int_{w_p}^{x} r'(t)dt \implies ||r||_\infty = \max_{t\in[a,b]} |r(t)| = |r(w)| = \left| \int_{w_p}^{w} r'(t)dt \right| \leq |w - w_p|\,||r'||_\infty$$

e pela estimativa da derivada, $||r||_\infty \leq \frac{h}{2}||r'||_\infty \leq \frac{h}{2}h^{1/2}||f''||_{L^2[a,b]}$.
Finalmente, a última e melhor estimativa em $O(h^4)$, resulta de considerar que 0 é o spline linear por interpolação de r, e pela estimativa de erro para splines lineares, isso implica $||r - 0||_\infty \leq \frac{1}{8}h^2||r''||_\infty$, e de forma semelhante, sendo σ o spline linear para f'' (notar que não é s'') temos $||f'' - \sigma||_\infty \leq \frac{1}{8}h^2||f^{''(4)}||_\infty$, quando $f \in C^4[a,b]$. Sendo possível mostrar que $||s'' - \sigma||_\infty \leq 3||f'' - \sigma||_\infty$, o resultado surge pela desigualdade triangular.

□

Exercício 12. Considere $f(x) = \sin(x)$ e uma partição de $[0, \frac{\pi}{2}]$ em N subintervalos, determine o número de valores de seno que precisam de ficar armazenados para calcularmos o seno em qualquer ponto, com erro inferior a 10^{-8} (precisão simples), usando funções seccionalmente $\mathbb{P}_3$.

Resolução: Neste caso$h_k = h = \frac{\pi}{2N}$ e pela melhor estimativa (como $f \in C^4, f^{(4)} = f = \sin$) temos para o spline cúbico interpolador

$$||f - s||_\infty \leq \frac{5h^4}{384}||f^{(4)}||_\infty = \frac{5}{384}(\frac{\pi}{2N})^4 < 10^{-8} \implies N > 10^2\frac{\pi}{2}(\frac{5}{384})^{1/4} = 53.06$$

bastando por isso armazenar $N + 1 = 55$ valores. As propriedades periódicas do seno mostram que bastaria tabelar aprox. 50 valores de seno com $h = \frac{\pi}{2N} \approx 0.03$ para ter precisão simples com a expressão do spline cúbico (os cinquenta valores s''_k deveriam estar previamente guardados, para evitar a resolução do sistema). No caso do seno isto não é justificado, mas serve como processo geral para outras funções em que o cálculo é moroso.

1.6.4 B-splines

No caso de nós igualmente espaçados, $h = \frac{b-a}{N}$, o cálculo do spline cúbico pode ser simplificado, usando funções base que são denominadas B-splines cúbicos:

$$B_3(x) = \begin{cases} \frac{1}{6}(2 - |x|)^3 - \frac{2}{3}(1 - |x|)^3 & \text{se } |x| \leq 1 \\ \frac{1}{6}(2 - |x|)^3 & \text{se } 1 \leq |x| \leq 2 \\ 0 & |x| \geq 2 \end{cases}$$

que é um spline cúbico (Exercício) natural com derivada nulas nos extremos de $[-2,2]$. Assim, com translações para $x_k = a + kh$, é possível definir uma aproximação

$$s(x) = \sum_{k=-1}^{N+1} a_k B_3\left(\frac{x - x_k}{h}\right)$$

em que os $N+3$ coeficientes $a_{-1}, a_0, ..., a_{N+1}$ são determinados resolvendo um sistema simples, quase tridiagonal

$$\begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ 0 & \cdots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} a_{-1} \\ a_0 \\ \vdots \\ \vdots \\ a_N \\ a_{N+1} \end{bmatrix} = \begin{bmatrix} hf'_0 \\ f_0 \\ \vdots \\ f_N \\ hf'_N \end{bmatrix}$$

resultando de para $m = 0, ..., N$ termos

$$\begin{aligned} f_m &= s(x_m) = \sum_{k=m-1}^{m+1} a_k B_3\left(\frac{x_m - x_k}{h}\right) = a_{m-1} B_3\left(\frac{x_m - x_{m-1}}{h}\right) + a_m B_3(0) + a_{m+1} B_3\left(\frac{x_m - x_{m+1}}{h}\right) \\ &= a_{m-1} B_3(1) + a_m B_3(0) + a_{m+1} B_3(-1) = a_{m-1}\frac{1}{6} + a_m \frac{2}{3} + a_{m+1}\frac{1}{6} \end{aligned}$$

o que se justifica por $B_3\left(\frac{x_m - x_k}{h}\right)$ ser nulo se $\frac{|x_m - x_k|}{h} \geq 2$. De forma semelhante, derivando, obtemos a primeira, $s'(x_0) = f'_0$, e a última, $s'(x_N) = f'_N$, equações.

Exercício 13. Aplique este método para determinar o spline cúbico $s(\pm 2) = s(\pm 1) = 0, s(0) = 1$, com condições nulas nas derivadas (exercício anterior) .

Exercício 14. Deduza o sistema a resolver no caso de splines quadráticos usando

$$B_2(x) = \begin{cases} 1 - \frac{1}{2}\left((\frac{1}{2} - |x|)^2 + (\frac{1}{2} + |x|)^2\right) & \text{se } |x| \leq \frac{1}{2} \\ \frac{1}{2}(\frac{3}{2} - |x|)^2 & \text{se } \frac{1}{2} \leq |x| \leq \frac{3}{2} \\ 0 & |x| \geq \frac{3}{2} \end{cases}$$

adicionando a equação $s'(x_0) = f'_0$ para determinar as incógnitas $a_{-1}, a_0, ..., a_N$.

1.7 Interpolação de Hermite

Objectivo: Consideramos um subespaço finito $G = \langle \mathbf{g} \rangle$, gerado por $\mathbf{g} = \{g_0, ..., g_N\}$, uma lista de funções. Dada uma lista de nós distintos $\mathbf{x} = \{x_0, ..., x_m\}$ e uma lista de listas $\mathbf{y} = \{\{y_0^{(0)}, ..., y_0^{(\alpha_0)}\}, ..., \{y_m^{(0)}, ..., y_m^{(\alpha_m)}\}\}$, em que a cada nó x_k está associada uma lista $\{y_k^{(0)}, ..., y_k^{(\alpha_k)}\}$ correspondente aos valores que a função (e as suas derivadas até ordem α_k) devem tomar nesse

nó. Pode-se formar uma tabela geral de interpolação

$$\begin{array}{llll} x_0 & x_1 & \cdots & x_m \\ y_0^{(0)} & y_1^{(0)} & \cdots & y_m^{(0)} \\ \vdots & \vdots & & \vdots \\ y_0^{(\alpha_0)} & \vdots & \cdots & y_m^{(\alpha_m)} \\ & y_1^{(\alpha_1)} & & \end{array}$$

Pretende-se encontrar $\phi \in G : \{\phi, \phi', ...\phi^{(\alpha_k)}\}(x_k) = \{y_k^{(0)}, ..., y_k^{(\alpha_k)}\}$.

Mais abreviadamente, designando $\Phi = [\{\phi, \phi', ...\phi^{(\alpha_k)}\}]_k$, poderíamos também escrever $\Phi(\mathbf{x}) = \mathbf{y}$.

- A solução do problema consiste na resolução do sistema alargado

$$\mathbf{G}(\mathbf{x})\mathbf{a} = \mathbf{y} \Leftrightarrow \begin{bmatrix} \{g_0, ..., g_0^{(\alpha_0)}\}(x_0) & \cdots & \{g_N, ..., g_N^{(\alpha_0)}\}(x_0) \\ \vdots & \ddots & \vdots \\ \{g_0, ..., g_0^{(\alpha_m)}\}(x_m) & \cdots & \{g_N, ..., g_N^{(\alpha_m)}\}(x_m) \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} \{y_0^{(0)}, ..., y_0^{(\alpha_0)}\} \\ \vdots \\ \{y_m^{(0)}, ..., y_m^{(\alpha_m)}\} \end{bmatrix}$$

mas como não é possível um cálculo eficiente com listas, passamos as listas para colunas, inserindo novas linhas na matriz.

$$\begin{bmatrix} g_0(x_0) & & g_N(x_0) \\ \vdots & \cdots & \vdots \\ g_0^{(\alpha_0)}(x_0) & & g_N^{(\alpha_0)}(x_0) \\ \vdots & \ddots & \vdots \\ g_0(x_m) & & g_N(x_m) \\ \vdots & \cdots & \vdots \\ g_0^{(\alpha_m)}(x_m) & & g_N^{(\alpha_m)}(x_m) \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} y_0^{(0)} \\ \vdots \\ y_0^{(\alpha_0)} \\ \vdots \\ y_m^{(0)} \\ \vdots \\ y_m^{(\alpha_m)} \end{bmatrix}$$

e para que haja um número de incógnitas igual ao número de equações, temos que ter

$$N + 1 = (\alpha_0 + 1) + \cdots + (\alpha_m + 1) = |\alpha| + m + 1$$

ou seja, $N = |\alpha| + m$.

A matriz $\mathbf{G}(\mathbf{x})$ é invertível se as funções $g_k^{(\alpha_j)}$ forem linearmente independentes em $\mathbf{x}$, formando uma base.

Através da solução $\mathbf{a}$ obtemos $\phi(t) = \mathbf{g}(t) \cdot \mathbf{a}$, ou $\Phi(t) = \mathbf{G}(t) \cdot \mathbf{a}$, verificando

$$\Phi(\mathbf{x}) = \mathbf{G}(\mathbf{x})\mathbf{a} = \mathbf{y}.$$

1.7.1 Interpolação polinomial de Hermite

Trata-se mais uma vez do caso de funções de variável real (ou complexa) em que $\mathbf{v}(t) = \{1, t, t^2, ..., t^N\}$, ou seja as funções base são monómios, e o subespaço $G = \mathcal{P}_N$. Agora a matriz de Vandermonde generalizada passa a ser

$$\mathbf{V}(\mathbf{x}) = \begin{bmatrix} \{1, 0, ..., 0\} & \{x_0, 1, 0, ..., 0\} & \cdots & \{x_0^N, ..., N \cdots (N - \alpha_0 + 1)x_0^{N-\alpha_0}\} \\ \vdots & \ddots & & \vdots \\ \{1, 0, ..., 0\} & \{x_N, 1, 0, ..., 0\} & \cdots & \{x_0^N, ..., N \cdots (N - \alpha_n + 1)x_m^{N-\alpha_m}\} \end{bmatrix}.$$

Mais uma vez, como asseguramos que o número de incógnitas é igual ao número de equações, com $N = |\alpha|+m$, para verificar que a matriz quadrada $\mathbf{V}(\mathbf{x})$ é invertível, basta ver que a solução do problema homogéneo terá que ser nula, ou seja, $\mathbf{V}(\mathbf{x})\mathbf{a} = 0 \Rightarrow \mathbf{a} = 0$.

Ora $\mathbf{V}(\mathbf{x})\mathbf{a} = \mathbf{0}$ significa que o polinómio

$$p_N(t) = \mathbf{v}(t)\mathbf{a} = a_0 + a_1 t + ... + a_N t^N$$

tem raízes em $\mathbf{x} = \{x_0, ..., x_m\}$, incluindo as múltiplas, sendo definida essa multiplicidade da raiz x_k pelo valor α_k+1. Assim, contando com as multiplicidades, há $\alpha_0+1+...+\alpha_m+1 = |\alpha|+m+1$ raízes, ou seja o polinómio de grau N tem $N+1$ raízes, o que implica que seja o polinómio nulo, logo $\mathbf{a} = 0$.

1.7.2 Aplicação da Fórmula de Newton

Podemos usar a fórmula de Newton já conhecida para calcular o polinómio interpolador de Hermite, ao invés de resolvermos o sistema definido pela matriz de Vandermonde generalizada.

Para esse efeito, consideramos uma repetição dos nós apropriada

$$\begin{array}{ccccccc} z_0 & \cdots & z_{\alpha_0} & \cdots & z_{N-\alpha_m} & \cdots & z_N \\ \{x_0 & \cdots & x_0\} & \cdots & \{x_m & \cdots & x_m\} \end{array}$$

em que cada x_k é repetido $\alpha_k + 1$ vezes. Tendo feito isto, a fórmula de Newton mantém-se

$$p_N(t) = \sum_{k=0}^{N} f_{[z_0,...,z_k]}(t - z_0)\cdots(t - z_{k-1})$$

entendendo o significado de $f_{[z_0,...,z_k]}$ com a repetição de nós, como um limite, por exemplo:

$$f_{[\underbrace{x_0, ..., x_0}_{\alpha_0+1 \, vezes}]} = \frac{1}{\alpha_0!} f^{(\alpha_0)}(x_0),$$

e mantendo-se o cálculo nos restantes, com nós diferentes, por exemplo:

$$f_{[\underbrace{x_0, ..., x_0}_{\alpha_0+1 \, vezes}, x_1]} = \frac{f_{[\underbrace{x_0, ..., x_0}_{\alpha_0 \, vezes}, x_1]} - f_{[\underbrace{x_0, ..., x_0}_{\alpha_0+1 \, vezes}]}}{x_1 - x_0}.$$

De novo, usando uma tabela de diferença divididas isto leva a um cálculo fácil.

Observação 9. Uma outra possibilidade será considerar a base de polinómios de Newton, definindo

$$\mathbf{w}(t) = \{1, (t - x_0), \cdots, (t - x_0)^{\alpha_0+1}, (t - x_0)^{\alpha_0+1}(t - x_1), \cdots\cdots, (t - x_0)^{\alpha_0+1}\cdots(t - x_m)^{\alpha_m}\}$$

o que permite obter o sistema na forma triangular inferior.

Exercício 15. Considere as tabelas de interpolação, e determine os polinómios interpoladores de Hermite:

(a)

x	0	1
$f(x)$	1	-1
$f'(x)$	-3	0

(b)

x	0	1
$f(x)$	1	-1
$f'(x)$	-3	
$f''(x)$	6	

Resolução: (a) Havendo 4 condições será polinómio de grau ≤ 3. Usamos a Fórmula de Newton, notando que $f[0,0] = f'(0) = -3$, $f[1,1] = f'(1) = 0$, para substituir na tabela de diferenças generalizada:

$$\begin{array}{cccccc} x & : & 0 & 0 & 1 & 1 \\ \hline f(x) & : & \mathbf{1} & 1 & -1 & -1 \\ & & f_{[0,0]} = \mathbf{-3} & \frac{-1-1}{1-0} = -2 & f_{[1,1]} = 0 & \\ & & & \frac{-2+3}{1-0} = \mathbf{1} & \frac{0-(-2)}{1-0} = 2 & \\ & & & & \frac{2-1}{1-0} = \mathbf{1} & \end{array}$$

e assim $p_3(x) = 1 - 3x + x^2 + x^2(x-1) = 1 - 3x + x^3$.

(b) De forma análoga, agora notamos que $f''(0) = 6 \implies f_{[0,0,0]} = \frac{f''(0)}{2!} = 3$

$$\begin{array}{cccccc} x & : & 0 & 0 & 0 & 1 \\ \hline f(x) & : & \mathbf{1} & 1 & 1 & -1 \\ & & f_{[0,0]} = \mathbf{-3} & f_{[0,0]} = -3 & \frac{-1-1}{1-0} = -2 & \\ & & & f_{[0,0,0]} = \mathbf{3} & \frac{-2-(-3)}{1-0} = 1 & \\ & & & & \frac{1-3}{1-0} = -2 & \end{array}$$

e assim $p_3(x) = 1 - 3x + 3x^2 - 2x^3$.

1.7.3 Fórmula com polinómios base de Hermite (1ª derivada)

Tal como no caso da interpolação de Lagrange, é também possível encontrar polinómios base de Hermite, que transformem a matriz de Vandermonde generalizada na matriz identidade. No entanto, essa expressão não é simples no caso geral, pelo que nos restringimos a apresentar o caso em que há uma tabela com os valores da função f_k e da sua derivada f'_k. Relembramos que neste caso há $2m+2$ condições, o que leva a polinómios de grau menor ou igual a $2m+1$.

Sendo L_k os polinómios base de Lagrange, definimos os polinómios base de Hermite:

$$\begin{aligned} H_k^0(x) &= (1 - 2L_k'(x_k)(x - x_k))\, L_k(x)^2 \\ H_k^1(x) &= (x - x_k)\, L_k(x)^2 \end{aligned} \tag{1.7.1}$$

Não é difícil verificar que (Exercício):

(i) $H_k^0(x_j) = \delta_{kj}$, $H_k^1(x_j) = 0$,

(ii) $(H_k^0)'(x_j) = 0$, $(H_k^1)'(x_j) = \delta_{kj}$.

Desta forma obtemos directamente a expressão para o polinómio interpolador

$$p_{2m+1}(x) = \sum_{k=0}^{m} f_k H_k^0(x) + \sum_{k=0}^{m} f'_k H_k^1(x) \tag{1.7.2}$$

1.7.4 Expressão do Erro

Da fórmula de Newton generalizada, com a repetição dos nós, e pela expressão já conhecida do erro de interpolação, aplicamos facilmente ao caso da interpolação de Hermite:

$$\begin{aligned} E(x) = f(x) - p_{m+|\alpha|}(x) &= f_{[z_0,..,z_0,...,z_m,...,z_m,x]} \prod_{k=0}^{m} (x - x_k)^{\alpha_k+1} \\ &= \frac{f^{(|\alpha|+m+1)}(\xi_x)}{(|\alpha|+m+1)!} \prod_{k=0}^{m} (x - x_k)^{\alpha_k+1} \end{aligned} \tag{1.7.3}$$

com $\xi_x \in [x_0; \ldots; x_m; x]$, desde que $f \in C^{|\alpha|+m+1}$.

Observação 10. Considerando a interpolação de Hermite de uma função $f \in C^{r+1}$ num ponto y usando derivadas até grau r nesse ponto, pela fórmula de Newton, obtemos o polinómio de Taylor em que o erro é o resto de Lagrange

$$\begin{aligned} p_r(x) &= \sum_{k=0}^{r} f_{[\underbrace{y,\cdots,y}_{(k+1)}]}(x-y)^k = \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}(x-y)^k \\ E(x) &= f(x) - p_r(x) = \frac{f^{(r+1)}(\xi_x)}{(r+1)!}(x-y)^{r+1} \end{aligned}$$

Exercício 16. Mostre que a expressão do polinómio interpolador de uma função $f \in C^\infty$ num nó y, nas derivadas até grau r, e num outro nó $z = y + h$, é dada por

$$p_{r+1}(x) = f(z)\left(\frac{x-y}{h}\right)^{r+1} + \sum_{k=0}^{r} h^k \frac{f^{(k)}(y)}{k!}\left(\left(\frac{x-y}{h}\right)^k - \left(\frac{x-y}{h}\right)^{r+1}\right).$$

Apresente ainda uma majoração do erro.

Resolução: Usando a fórmula de Newton temos

$$p_{r+1}(x) = \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}(x-y)^k + f_{[y,\cdots,y,z]}(x-y)^{r+1}$$

e substituindo $x = z$, devemos ter $f(z) = \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}h^k + f_{[y,\cdots,y,z]}h^{r+1}$, de onde obtemos

$$f_{[y,\cdots,y,z]} = \left(f(z) - \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}h^k\right)h^{-r-1}$$

$$p_{r+1}(x) = \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}(x-y)^k + \left(f(z) - \sum_{k=0}^{r} \frac{f^{(k)}(y)}{k!}h^k\right)\left(\frac{x-y}{h}\right)^{r+1}$$

de onde sai o resultado agrupando no somatório.

1.8 Diferenciação Numérica

1.8.1 Aproximação por interpolação de Lagrange

Começamos por recordar a fórmula de interpolação de Lagrange em que usando pontos $x_0, ..., x_n$ uma função f é aproximada pelo polinómio interpolador e assim podemos tentar aproximar

$f'(x)$ por $p'_n(x) = \sum_{k=0}^{n} f(x_k)L'_k(x)$. O erro cometido nesta aproximação é obtido pela fórmula do erro com diferenças divididas:

$$E_n(x) = f(x) - p_n(x) = f_{[x_0,...,x_n,x]} \overbrace{(x-x_0)...(x-x_n)}^{W_{n+1}(x)}$$

Notando que

$$\frac{d}{dx} f_{[x_0,...,x_n,x]} = \lim_{\varepsilon \to 0} \frac{f_{[x_0,...,x_n,x+\varepsilon]} - f_{[x_0,...,x_n,x]}}{\varepsilon} = \lim_{\varepsilon \to 0} f_{[x_0,...,x_n,x,x+\varepsilon]} = f_{[x_0,...,x_n,x,x]}$$

a derivada $E'_n(x)$ dá:

$$E'_n(x) = f'(x) - p'_n(x) = f[x_0,...,x_n,x,x]W_{n+1}(x) + f[x_0,...,x_n,x]W'_{n+1}(x).$$

Admitindo que $f \in C^{n+2}([x_0,x_n])$, sabemos que para $x \in [x_0,x_n]$,

$$E'_n(x) = f'(x) - p'_n(x) = \frac{f^{(n+2)}(\xi_2)}{(n+2)!}W_{n+1}(x) + \frac{f^{(n+1)}(\xi_1)}{(n+1)!}W'_{n+1}(x), \tag{1.8.1}$$

com $\xi_1, \xi_2 \in]x_0, x_n[$.

Supondo que os pontos estão igualmente espaçados, $x_k = x_0 + kh$, e para h suficientemente pequeno, temos

$$W_{n+1}(x) = O(h^{n+1}), W'_{n+1}(x) = O(h^n).$$

• Sempre que escolhermos $x = x_j$ (um dos nós), temos $W_{n+1}(x) = 0$ e o erro dependerá do valor $W'_{n+1}(x)$. Não tendo erro nulo, será então da ordem $O(h^n)$.

• Para obtermos um erro da ordem $O(h^{n+1})$ convém assim escolher um ponto x tal que $W'_{n+1}(x) = 0$.

Aproximação da 1ª derivada

Vamos analisar alguns casos particulares.

- **Caso** $n = 1$. Consideramos apenas dois pontos x_0, x_1, e a aproximação de f' será p'_1

$$\begin{aligned} p_1(x) &= f(x_0) + f_{[x_0,x_1]}(x - x_0) \\ p'_1(x) &= f_{[x_0,x_1]} \end{aligned}$$

Quanto ao erro, como $W_2(x) = (x-x_0)(x-x_1)$ obtemos

$$W'_2(x) = \frac{d}{dx}((x-x_0)(x-x_1)) = 2x - (x_0+x_1)$$

e um primeiro objectivo será escolher $z : W'_2(z) = 0$.

Diferença centrada.

Sendo $z = \frac{x_0+x_1}{2}$, temos $W'_2(z) = 0$, e com $x_0 = z - h$, $x_1 = z + h$, obtemos de (1.8.1)

$$f'(z) - f_{[z-h,\, z+h]} = \frac{f^{(3)}(\xi)}{3!}W_2(z) = -\frac{f^{(3)}(\xi_2)}{6}h^2$$

Ou seja, temos a fórmula da diferença centrada ($\xi \in [z-h, z+h]$):

$$f'(z) = \frac{f(z+h) - f(z-h)}{2h} - \frac{f'''(\xi)}{6}h^2.$$

Diferença progressiva
No caso de escolhermos $x_0 = z$, e $x_1 = z + h$, temos$W_2(z) = 0,\ W_2'(z) = -h$. Portanto,

$$f'(z) = f_{[z,\, z+h]} - \frac{f''(\xi_1)}{2}h,$$

o que dá exactamente o resto de Lagrange da série de Taylor, e justifica o resto em $O(h)$.
Diferença regressiva
De forma semelhante, escolhendo $x_0 = z - h$, e $x_1 = z$, temos$W_2(z) = 0,\ W_2'(z) = h$.

$$f'(z) = f_{[z-h,\, z]} + \frac{f''(\xi_1)}{2}h$$

- **Caso** $n = 2$. Considerando agora três pontos x_0, x_1, x_2 obtendo

$$p_2(x) = f(x_0) + f_{[x_0,x_1]}(x - x_0) + f_{[x_0,x_1,x_2]}(x - x_0)(x - x_1)$$

o que implica

$$p_2'(x) = f_{[x_0,x_1]} + f_{[x_0,x_1,x_2]}(2x - x_0 - x_1). \tag{1.8.2}$$

Diferença centrada (n=2)
Neste caso corresponde a considerar $z = x_1$, com $x_0 = z - h$, $x_2 = z + h$, ficando

$$p_2'(z) = f_{[x_0,x_1]} + f_{[x_0,x_1,x_2]}h = f_{[x_0,x_1]} + \frac{1}{2}(f_{[x_1,x_2]} - f_{[x_0,x_1]}) = f_{[x_0,x_2]},$$

e obtemos a expressão anterior.
Notando que aqui $W_3'(x) = 3(x-z)^2 - h^2$, temos $W_3(z) = 0$, mas $W_2'(z) = -h^2 \neq 0$, confirmando-se a fórmula em $O(h^2)$, mas[1] não em $O(h^3)$.

Diferença progressiva (n=2)
Considerando $x_0 = z$, $x_1 = z + h$, $x_2 = z + 2h$ obtemos de (1.8.2)

$$f'(z) \approx p_2'(z) = f_{[x_0,x_1]} - 3hf_{[x_0,x_1,x_2]} = \frac{5}{2}f_{[x_0,x_1]} - \frac{3}{2}f_{[x_1,x_2]},$$

e como $W_3'(z) = W_3'(x_0) = 3h^2 - h^2 = 2h^2$, temos

$$f'(z) = \frac{4f(z+h) - 3f(z) - f(z+2h)}{2h} + \frac{h^2 f'''(\xi)}{3}.$$

Exercício: determinar a expressão para a diferença regressiva com $n = 2$.

- **Caso** $n > 2$. Para valores de n superiores o processo será semelhante. No entanto, no que se segue e na prática, as aproximações mais frequentes não utilizam n maior que 2.

[1] – Para obter essa ordem superior deveríamos considerar $W_3'(x) = 3(x-z)^2 - h^2 = 0$, ou seja $x = z \pm \frac{1}{\sqrt{3}}h$. Este valor não é normalmente considerado pois quebra o espaçamento uniforme. De qualquer forma, podemos obter

$$f'(x) = f_{[z-h,z]} - f_{[z-h,z,z+h]}\tfrac{3-\sqrt{3}}{3}h + O(h^3)$$

Aproximação da segunda derivada

Usando ainda a interpolação de Lagrange, somos levados a considerar

$$f''(x) \approx p_n''(x).$$

Em termos de cálculo do erro, de $E_n'(x) = f'(x) - p_n'(x) = f_{[x_0,\ldots,x_n,x,x]} W_{n+1}(x) + f_{[x_0,\ldots,x_n,x]} W_{n+1}'(x)$. obtemos

$$\begin{aligned} E_n''(x) &= f''(x) - p_n''(x) = \\ &= f[x_0,\ldots,x_n,x,x,x]W_{n+1}(x) + 2f[x_0,\ldots,x_n,x,x]W_{n+1}'(x) + f[x_0,\ldots,x_n,x]W_{n+1}''(x), \end{aligned}$$

e usando a relação com as derivadas, para $f \in C^{n+3}$, temos

$$E_n''(x) = \frac{f^{(n+3)}(\xi_3)}{(n+3)!} W_{n+1}(x) + 2\frac{f^{(n+2)}(\xi_2)}{(n+2)!} W_{n+1}'(x) + \frac{f^{(n+1)}(\xi_1)}{(n+1)!} W_{n+1}''(x),$$

Mais uma vez notamos que $W_{n+1}(x) = O(h^{n+1}), W_{n+1}'(x) = O(h^n), W_{n+1}''(x) = O(h^{n-1})$, no entanto apenas podemos esperar encontrar x tal que $W_{n+1}''(x) = 0$, para manter o erro em $O(h^n)$.

- **Caso** $n = 2$. Notamos que no caso $n = 1$ teríamos $p_1'' = 0$, e como $W_2''(x) = O(1)$, o erro não tenderia para zero, e não constituiria uma aproximação credível, pelo que começamos com $n = 2$.
  A partir da fórmula de Newton,

  $$p_2''(x) = 2f_{[x_0,x_1,x_2]}$$

  e em termos do erro, notamos que

  $$W_3'(x) = \frac{d}{dx}(x-x_0)(x-x_1)(x-x_2) = (x-x_0)(x-x_1) + (x-x_1)(x-x_2) + (x-x_0)(x-x_2)$$

  logo

  $$W_3''(x) = 6x - 2(x_0 + x_1 + x_2),$$

  portanto $f''(z)$ será aproximado por $p_2''(z) = 2f_{[x_0,x_1,x_2]}$ com erro $O(h^2)$, se $W_3''(z) = 0$, ou seja:

  $$z = \frac{1}{3}(x_0 + x_1 + x_2)$$

  Quando $f \in C^4$, isso pode ser obtido com diferenças centradas, escolhendo $x_0 = z - h$, $x_1 = z$, $x_2 = z + h$:

  $$\begin{aligned} f''(z) &= 2f_{[x_0,x_1,x_2]} + \frac{f^{(4)}(\xi_1)}{4!} W_2'(z) \\ &= \frac{f(z+h) - 2f(z) + f(z-h)}{h^2} - \frac{f^{(4)}(\xi_1)}{12} h^2 \end{aligned} \tag{1.8.3}$$

  porque $W_3'(z) = 0 + 0 + h(-2h) = -2h^2$.

1.8.2 Método dos coeficientes indeterminados

Vamos agora ver um processo diferente para obter a aproximação das derivadas. Para melhor compreensão, veremos como obter a fórmula (1.8.3) para a segunda derivada com diferenças centradas, por este outro método. Queremos ainda utilizar os três pontos $x_{-1} = x + h, x_0 = x, x_1 = x + h$,

$$f''(x) \approx Af(x+h) + Bf(x) + Cf(x-h),$$

em que os valores A, B e C são desconhecidos e serão utilizados para obter a aproximação mais conveniente.

Usando o desenvolvimento em série de Taylor para $f(x+h)$ e para $f(x-h)$, obtemos

$$Af(x+h)+Bf(x)+Cf(x-h) = A\left(f(x)+hf'(x)+\frac{h^2}{2}f''(x)+...\right)+Bf(x)+C\left(f(x)-hf'(x)+\frac{h^2}{2}f''(x)-...\right)$$

se usarmos o resto de Lagrange para $f^{(iv)}$ obtém-se $Af(x+h) + Bf(x) + Cf(x-h) =$

$$\begin{aligned} = \quad & f(x)(A+B+C) + f'(x)(A-C) + \frac{h^2}{2}f''(x)(A+C) + \frac{h^3}{6}f''(x)(A-C) + \\ & + A\frac{h^4}{24}f^{(4)}(\xi_1) + C\frac{h^4}{24}f^{(4)}(\xi_2) \end{aligned}$$

O objectivo é agora anular todas as expressões, excepto a que tem a segunda derivada, que é a que pretendemos aproximar, e as expressões das quartas derivadas, que constituirão o resto. Para além disso, para obtermos apenas a segunda derivada, devemos exigir que o seu coeficiente seja unitário, ou seja, $\frac{h^2}{2}(A+C) = 1$. Ficamos assim com o sistema

$$\begin{cases} A+B+C=0 \\ A-C=0 \\ \frac{h^2}{2}(A+C)=1 \end{cases} \qquad (A,B,C) = \frac{1}{h^2}(1,-2,1),$$

e podemos concluir a mesma fórmula (1.8.3), pois

$$\frac{f(x+h)-2f(x)+f(x-h)}{h^2} = f''(x) + \frac{h^2}{24}f^{(4)}(\xi_1) + \frac{h^2}{24}f^{(4)}(\xi_2).$$

Exercício 17. Obter a fórmula de diferenças centradas para a primeira derivada, que já foi apresentada, mas usando agora o método dos coeficientes indeterminados.

Exercício 18. (erros de arredondamento). Considere que os valores de $f_{-1} = f(x-h), f_0 = f(x)$ e $f_1 = f(x+h)$ estavam afectados de erros, e apenas dispunhamos dos valores aproximados respectivos $\tilde{f}_{-1}, \tilde{f}_0, \tilde{f}_1$. Comente o efeito dos erros $\varepsilon = \max_{k\in\{-1,0,1\}} |f_k - \tilde{f}_k|$ no cálculo da aproximação de $f''(x)$ usando a fórmula com diferenças centradas.

Resolução:

$$f''(x) = \frac{\tilde{f}_1 - 2\tilde{f}_0 + \tilde{f}_{-1}}{h^2} - \frac{h^2}{12}f^{(iv)}(\xi) + \frac{\varepsilon_1 - 2\varepsilon_0 + \varepsilon_{-1}}{h^2},$$

em que $\varepsilon_k = f_k - \tilde{f}_k$ são os erros. Portanto, como $|\varepsilon_k| \leq \varepsilon$, quando $h \to 0$ temos

$$\left| f''(x) - \frac{\tilde{f}_1 - 2\tilde{f}_0 + \tilde{f}_{-1}}{h^2} \right| \leq \underbrace{\frac{h^2}{12} \left| f^{(iv)}(\xi) \right|}_{\longrightarrow 0} + \underbrace{\frac{4\varepsilon}{h^2}}_{\not\longrightarrow 0}.$$

Reparamos assim que a parcela que contém os erros de arredondamento não irá decrescer *a priori* para zero, a menos que o ε acompanhe o decrescimento do h, ou seja, deveremos ter $\varepsilon = o(h^2)$. Isto significa que ao diminuir o h devemos ter o cuidado de que os erros acompanhem o decréscimo, ou doutra forma a aproximação perderá a eficácia, ou mesmo o significado.

1.8.3 Introdução à teoria das diferenças

Iremos agora definir algumas noções básicas que estão relacionadas com a diferenciação numérica. Dada uma sucessão (u_n) analogamente definimos os seguintes operadores de diferenças

- Diferenças progressivas : $\Delta u_n = u_{n+1} - u_n$.
- Diferenças regressivas : $\nabla u_n = u_n - u_{n-1}$.
- Diferenças centradas : $\mathrm{d}u_n = \frac{1}{2}(u_{n+1} - u_{n-1})$ ou ainda[2] $\mathrm{d}u_n = u_{n+1/2} - u_{n-1/2}$

Vamos concentrar-nos nas diferenças progressivas enunciando algumas propriedades imediatas.

Teorema 7. *No caso de usarmos nós igualmente espaçados $x_k = x_0 + kh$ temos ($k \in \mathbb{N}_0$)*

$$f_{[x_0,\ldots,x_k]} = \frac{\Delta^k f_0}{k! h^k}$$

e por isso a fórmula de interpolação de Newton fica

$$p_n(x) = \sum_{k=0}^{n} \frac{\Delta^k f_0}{k! h^k} \prod_{j=0}^{k-1} (x - x_j).$$

Demonstração. Provamos por indução, sendo óbvio para $k = 1$ (ou mesmo $k = 0$), pois $f_{[x_0,x_1]} = \frac{f_1 - f_0}{h} = \frac{\Delta f_0}{h}$. Sendo válida para quaisquer $\{x_0, ..., x_k\}$, temos para $\{x_0, ..., x_{k+1}\}$:

$$f_{[x_0,\ldots,x_{k+1}]} = \frac{f_{[x_1,\ldots,x_{k+1}]} - f_{[x_0,\ldots,x_k]}}{x_{k+1} - x_0} = \frac{\frac{\Delta^k f_1}{k! h^k} - \frac{\Delta^k f_0}{k! h^k}}{(k+1)h} = \frac{\Delta^k(\Delta f_0)}{(k+1)k! h h^k} = \frac{\Delta^{k+1} f_0}{(k+1)! h^{k+1}}.$$

□

Proposição 5. *As seguintes propriedades são evidentes:*
i) Se $u_n = C$ (constante) então $\Delta u_n = 0$.
ii) $\Delta(\alpha u_n + \beta v_n) = \alpha \Delta u_n + \beta \Delta v_n$
iii) $\Delta(u_n v_n) = v_{n+1}\Delta u_n + u_n \Delta v_n = u_{n+1}\Delta v_n + v_n \Delta u_n$
iv) $\Delta(C^n) = (C-1)C^n$, e em particular[3] $\Delta(2^n) = 2^n$.

[2] Neste caso, supomos que u_n se trata da imagem de uma certa função u calculada num ponto x_n, em que os pontos x_n verificam $x_{n+1} = x_n + h$, com $h > 0$ fixo. Assim, entendemos $u_{n+1/2}$ como sendo o valor de u calculado num ponto $x_{n+1/2} = x_n + h/2$, e consequentemente $u_{n-1/2}$ será o valor no ponto $x_{n-1/2} = x_n - h/2$.

[3]

Proposição 6. *A sucessão $u_n = 2^n$ é um ponto fixo do operador de diferenças progressivas, tal como e^x é um ponto fixo para o operador de derivação habitual*

Demonstração. Imediata. (*Exercício*: obtenha resultados semelhantes para o operador de diferenças centradas $\mathrm{d}u_n$). □

Proposição 7. *Temos a propriedade telescópica e outra propriedade elementar*
i) $\sum_{k=0}^{n-1} \Delta u_k = u_n - u_0$
ii) $\Delta\left(\sum_{k=0}^{n-1} u_k\right) = u_n$

Demonstração. Imediata. (*Exercício*: obtenha resultados semelhantes para o operador de diferenças centradas $\mathrm{d}u_n$). □

Exercício 19. Verifique que

$$\sum_{k=0}^{n-1} u_k \Delta v_k = [u_k v_k]_{k=0}^{k=n} - \sum_{k=0}^{n-1} v_{k+1} \Delta u_k$$

e explicite $\sum_{k=0}^{n-1} kC^k$.

Resolução: Pela proposição anterior (iii), temos $\sum_{k=0}^{n-1} \Delta(u_k v_k) = \sum_{k=0}^{n-1}(u_k \Delta v_k + v_{k+1}\Delta u_k)$ e pela propriedade telescópica,

$$u_n v_n - u_0 v_0 = \sum_{k=0}^{n-1} u_k \Delta v_k + \sum_{k=0}^{n-1} v_{k+1} \Delta u_k.$$

Aplicado com $u_k = k, v_k = \frac{C^k}{C-1}$ porque $\Delta v_k = C^k$, e como $\Delta u_k = 1$, obtemos

$$\sum_{k=0}^{n-1} kC^k = \left[k\frac{C^k}{C-1}\right]_{k=0}^{k=n} - \sum_{k=0}^{n-1} \frac{C^{k+1}}{C-1} = \frac{nC^n}{C-1} - \frac{C}{C-1}\sum_{k=0}^{n-1} C^k = \frac{nC^n}{C-1} - C\frac{C^n-1}{(C-1)^2}.$$

Proposição 8. *Definindo* $n^{[p]} = \frac{n!}{(n-p)!} = n(n-1)\cdots(n-p+1)$, *obtemos*

$$\Delta n^{[p]} = pn^{[p-1]}, \text{ e por isso } \sum_{k=0}^{n-1} k^{[p]} = \frac{n^{[p+1]}}{p+1},$$

em particular $\sum_{k=0}^{n-1} k = \frac{1}{2}n(n-1)$, $\sum_{k=0}^{n-1} k(k-1) = \frac{1}{3}n(n-1)(n-2)$.

Demonstração. Obtemos directamente

$$\begin{aligned} \Delta n^{[p]} = (n+1)^{[p]} - n^{[p]} &= (n+1)n\cdots(n-p+2) - n\cdots(n-p+1) \\ &= (n+1-n+p-1)n\cdots(n-p+2) = pn\cdots(n-p+2) = pn^{[p-1]}. \end{aligned}$$

e depois basta notar que $n^{[p+1]} = \sum_{k=0}^{n-1} \Delta k^{[p+1]} = (p+1)\sum_{k=0}^{n-1} k^{[p]}$. □

Exercício 20. Mostre que $\Delta\frac{1}{n^{[p]}} = \frac{-p}{(n+1)^{[p+1]}}$ e conclua que $\sum_{k=0}^{n-1} \frac{1}{(k+1)(k+2)(k+3)} = \frac{1}{4} - \frac{1}{2(n+2)(n+1)}$, ou de modo geral

$$\sum_{k=0}^{n} \frac{1}{(k+p+1)^{[p+1]}} = \frac{1}{p}\left(\frac{1}{p!} - \frac{1}{(n+p)^{[p]}}\right) \underset{n\to\infty}{\longrightarrow} \frac{1}{p!p}.$$

Resolução: Se $u_n \neq 0$, como

$$\Delta\frac{1}{u_n} = \frac{1}{u_{n+1}} - \frac{1}{u_n} = \frac{u_n - u_{n+1}}{u_{n+1}u_n} = \frac{-\Delta u_n}{u_{n+1}u_n}$$

obtemos para $u_n = n^{[p]}$, com $n \geq p$

$$\Delta \frac{1}{n^{[p]}} = \frac{-pn^{[p-1]}}{(n+1)^{[p]}n^{[p]}} = \frac{-p}{(n+1)^{[p]}(n-p+1)} = \frac{-p}{(n+1)^{[p+1]}}.$$

notando que $n^{[p]} = n^{[p-1]}(n-p+1)$. Em particular,

$$\sum_{k=0}^{n-1} \frac{-2}{(k+3)^{[3]}} = \sum_{k=0}^{n-1} \Delta \frac{1}{(k+2)^{[2]}} = \frac{1}{(n+2)^{[2]}} - \frac{1}{2^{[2]}}.$$

De modo geral,

$$\sum_{k=0}^{\infty} \frac{-p}{(k+p+1)^{[p+1]}} = \lim_n \sum_{k=0}^{n} \Delta \frac{1}{(k+p)^{[p]}} = \lim_n \frac{1}{(n+p)^{[p]}} - \frac{1}{p^{[p]}}$$

Observação 11. A teoria das diferenças pode ser ligada ainda por analogia a resultados de equações às diferenças que veremos mais tarde.

1.8.4 Aplicação da teoria das diferenças à aproximação de derivadas

Para motivar a obtenção de fórmulas de aproximação da derivação por diferenças, começamos por obter o análogo da fórmula de Taylor para diferenças. Com efeito, introduzindo o *operador de sucessão*

$$Su_n = u_{n+1},$$

temos $\Delta u_n = Su_n - u_n = (S-I)u_n$, logo $\Delta = S - I \Leftrightarrow S = \Delta + I$, e assim podemos estabelecer o resultado.

Proposição 9. *Temos para* $u_{n+m} = S^m u_n$

$$u_{n+m} = \sum_{k=0}^{m} \frac{m^{[k]}}{k!} \Delta^k u_n.$$

Demonstração. Basta reparar que $S^m = (\Delta + I)^m = \sum_{k=0}^m \binom{m}{k} \Delta^k = \sum_{k=0}^m \frac{m^{[k]}}{k!} \Delta^k$. □

Seja u uma função analítica e $u_n = u(x_n)$, em que $x_n = a + nh$, temos pelo desenvolvimento em série de Taylor,

$$u_{n+1} = u(x_n + h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} (\partial^k u)(x_n)$$

e designando por ∂ o operador de derivação, podemos escrever pela expansão da exponencial,

$$Su_n = \sum_{k=0}^{\infty} \frac{(h\partial)^k}{k!} u_n = (e^{h\partial}) u_n$$

em que abreviadamente escrevemos $(e^{h\partial})u_n$ para $(e^{h\partial})u$ calculado em x_n. Desta forma estabelecemos $S = e^{h\partial}$, e portanto, por exemplo,

$$\Delta + I = S = e^{h\partial} \Leftrightarrow \partial = \frac{1}{h} \log(\Delta + I).$$

Fica assim estabelecida uma relação formal entre derivadas e diferenças progressivas. Tendo obtido esta fórmula, podemos utilizar a expansão do logaritmo,

$$\log(x+1) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k$$

e desta forma, formalmente,

$$\partial = \frac{1}{h}\log(\Delta + I) = \frac{1}{h}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}\Delta^k.$$

Truncando a série, por exemplo com um termo, $\partial \approx \frac{1}{h}\sum_{k=1}^{1} \frac{(-1)^{k+1}}{k}\Delta^k = \frac{\Delta}{h}$, ou seja $\partial u(x_n) \approx \frac{u_{n+1}-u_n}{h}$. Se usarmos dois termos,

$$\partial \approx \frac{1}{h}\sum_{k=1}^{2} \frac{(-1)^{k+1}}{k}\Delta^k = \frac{\Delta}{h} - \frac{\Delta^2}{2h},$$

esta última fórmula leva à aproximação,

$$\partial u(x_n) \approx \frac{2\Delta u_n - \Delta^2 u_n}{2h} = \frac{2u_{n+1} - 2u_n - u_{n+2} + 2u_{n+1} - u_n}{2h} = \frac{4u_{n+1} - 3u_n - u_{n+2}}{2h},$$

que já tinha sido encontrada no caso $n \geq 2$ pela interpolação de Lagrange. Para obter fórmulas de graus superiores, bastará considerar mais termos da série.

Exercício 21. Estabelecer relações semelhantes para as outras diferenças, por exemplo, $\mathrm{d} = 2\sinh(\frac{h}{2}\partial)$.

1.9 Aproximação de Funcionais Lineares

Um espaço vectorial com produto interno $\langle .,. \rangle$ é denominado pré-hilbertiano (ou euclidiano), notando que no caso em que o corpo de escalares é complexo temos

$$\langle \alpha u, v\rangle = \bar{\alpha}\,\langle u, v\rangle \text{ e também } \langle v, u\rangle = \overline{\langle u, v\rangle},$$

por isso convencionamos que a conjugação se efectua no primeiro termo. Associa-se a norma definida por $||u||^2 = \langle u, u\rangle$, relembramos que a existência de produto interno num espaço normado pode ser avaliada pela igualdade do paralelogramo:

$$||\frac{u+v}{2}||^2 + ||\frac{u-v}{2}||^2 = \frac{1}{2}(||u||^2 + ||v||^2).$$

No caso complexo, é imediato que

$$||u+v||^2 = ||u|| + 2\ \mathrm{Re}\,\langle u, v\rangle + ||v||^2$$

e se $\langle u, v\rangle = 0$ obtemos o teorema de Pitágoras $||u+v||^2 = ||u||^2 + ||v||^2$. Outro resultado conhecido é a desigualdade de Cauchy-Schwarz

$$|\langle u, v\rangle| \leq ||u||\,||v||.$$

Quando o espaço pré-hilbertiano é completo (as sucessões de Cauchy convergem), designa-se espaço de Hilbert H. Relembramos que este é o caso de todos os espaços de dimensão finita (isomorfos a $\mathbb{R}^n$), ou das funções em $L^2(a,b)$. No entanto, se considerarmos$C[a,b]$, e apesar do produto interno L^2 estar bem definido, as sucessões de Cauchy nessa norma $L^2(a,b)$ podem convergir para uma função $L^2(a,b)$ que não é contínua... Por isso, o espaço de funções contínuas $C[a,b]$ não é habitualmente considerado no âmbito dos espaços de Hilbert, sendo apenas completo com a norma do máximo, a que não está associado produto interno. Iremos analisar o caso de espaços métricos completos, isto é espaços de Banach, noutro capítulo.

Definição 4. Dado um funcional linear $F : H \to \mathbb{R}$, definimos uma fórmula de aproximação linear $\tilde{F}$ usando nós $z_0, \ldots, z_m$ e coeficientes $\alpha_0, \cdots, \alpha_m$

$$\tilde{F}(g) = \alpha_0 g(z_0) + \cdots + \alpha_m g(z_m).$$

A fórmula é exacta num subespaço $S = \langle g_0, \cdots, g_n \rangle \subseteq H$ se verificar $\tilde{F}(g) = F(g), \forall g \in S$.

Observação 12. Por exemplo, um subespaço polinomial $S = \mathcal{P}_n = \langle 1, t, \cdots, t^n \rangle$. Se essa aproximação for exacta para polinómios de grau $k \leq n$, ou seja $\tilde{F}(t^k) = F(t^k)$, dizemos que a fórmula tem pelo menos grau n (e diz-se ter exactamente grau n se não tiver grau $n+1$).

Método dos Coeficientes Indeterminados.

Dado um conjunto de nós $z_0, \ldots, z_m$ podemos determinar os coeficientes α_j de maneira a que a fórmula seja exacta em S :

$$\begin{cases} \alpha_0 g_0(z_0) + \cdots + \alpha_m g_0(z_m) = F(g_0) \\ \vdots \\ \alpha_0 g_n(z_0) + \cdots + \alpha_m g_n(z_m) = F(g_n) \end{cases} \Leftrightarrow \begin{bmatrix} g_0(z_0) & \cdots & g_0(z_m) \\ \vdots & \ddots & \vdots \\ g_n(z_0) & \cdots & g_n(z_m) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} F(g_0) \\ \vdots \\ F(g_n) \end{bmatrix}$$

sistema onde encontramos a transposta da matriz de Vandermonde.

Exemplo 2. Um caso de aplicação deste *método de coeficientes indeterminados* consiste em encontrar regras de integração, com $F(g) = I(g) = \int_a^b g(t)dt$, exactas para polinómios $g_k(t) = t^k$, com $k = 0, ..., n$, obtemos o sistema:

$$\begin{bmatrix} 1 & \cdots & 1 \\ z_0 & \cdots & z_m \\ \vdots & \ddots & \vdots \\ z_0^n & \cdots & z_m^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \int_a^b 1dt = b-a \\ \int_a^b tdt = \frac{1}{2}(b^2 - a^2) \\ \vdots \\ \int_a^b t^n dt = \frac{1}{n+1}(b^{n+1} - a^{n+1}) \end{bmatrix}.$$

São assim determinadas as regras de quadratura simples:

- Regra do Ponto Médio, com $m = 0$, $z_0 = \frac{a+b}{2} \implies \alpha_0 = b - a$, logo

$$\tilde{I}(g) = \alpha_0 g(z_0) = (b-a) g(\frac{a+b}{2})$$

- Regra dos Trapézios, com $m = 1$, $z_0 = a$, $z_1 = b \implies$

$$\begin{bmatrix} 1 & 1 \\ z_0 & z_1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} b-a \\ \frac{1}{2}(b^2 - a^2) \end{bmatrix} \implies \alpha_0 = \alpha_1 = \frac{b-a}{2}$$

$$\tilde{I}(g) = \alpha_0 g(z_0) + \alpha_1 g(z_1) = \frac{b-a}{2}(g(a) + g(b))$$

- Regra de Simpson, com $m = 2$, $z_0 = a$, $z_1 = \frac{a+b}{2}$, $z_2 = b \implies$

$$\begin{bmatrix} 1 & 1 & 1 \\ z_0 & z_1 & z_2 \\ z_0^2 & z_1^2 & z_2^2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} b-a \\ \frac{1}{2}(b^2-a^2) \\ \frac{1}{3}(b^3-a^3) \end{bmatrix} \implies \alpha_0 = \alpha_2 = \frac{b-a}{6},\ \alpha_1 = 4\alpha_0$$

$$\tilde{I}(g) = \frac{b-a}{6}(g(a) + 4g(\frac{a+b}{2}) + g(b))$$

Exemplo 3. De forma semelhante podemos obter fórmulas de diferenciação numérica. Por exemplo, com $F(g) = g''(y)$ e com $z_0 = y - h$, $z_1 = y$, $z_2 = y + h$, obtemos

$$\begin{bmatrix} 1 & 1 & 1 \\ z_0 & z_1 & z_2 \\ z_0^2 & z_1^2 & z_2^2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \implies \alpha_0 = \alpha_2 = \frac{1}{h^2}, \alpha_1 = -2\alpha_0$$

$$\tilde{F}(g) = \frac{1}{h^2}(g(y-h) - 2g(y) + g(y+h)).$$

Observação 13. A própria interpolação pode ser vista neste contexto. Para cada ponto x, $F(g) = g(x)$, ou ainda $F = \delta_x$ (o delta de Dirac centrado em x). Assim, no caso da interpolação polinomial, os valores α_j são dados pelos polinómios de Lagrange $\alpha_k = L_k(y)$.

Teorema 8. *Seja $\tilde{F}$ uma aproximação do funcional F com pelo menos grau $n \geq m$. O erro da aproximação é dado por*

$$F(g) - \tilde{F}(g) = F(\varepsilon_n),$$

onde $\varepsilon_n(x) = g[z_0, \ldots, z_n, x](x - z_0) \cdots (x - z_n)$ é o erro de interpolação nos nós $z_0, \ldots, z_m$, definidos em $\tilde{F}$, acrescentando nós adicionais $z_{m+1}, \ldots, z_n$ (eventualmente repetidos).

Demonstração. Sendo p_n o polinómio interpolador, o erro de interpolação é $\varepsilon_n(x) = g(x) - p_n(x) = g[z_0, \ldots, z_n, x](x - z_0) \cdots (x - z_n)$, portanto como $\tilde{F}$ é exacta para grau n, obtemos

$$F(g) - \tilde{F}(g) = F(p_n + \varepsilon_n) - \tilde{F}(p_n + \varepsilon_n) = \underbrace{F(p_n) - \tilde{F}(p_n)}_{=0} + F(\varepsilon_n) - \tilde{F}(\varepsilon_n) = F(\varepsilon_n) - \tilde{F}(\varepsilon_n).$$

Finalmente, $\tilde{F}(\varepsilon_n) = 0$ porque o erro é nulo nos nós de interpolação $\varepsilon_n(z_j) = 0$, e assim

$$\tilde{F}(\varepsilon_n) = \sum_{j=0}^{m} \alpha_j \varepsilon_n(z_j) = 0.$$

□

Exemplo 4. Erro da Regra dos Trapézios - obtemos a conhecida expressão

$$I(g) - \tilde{I}(g) = I(\varepsilon_1) = \int_a^b g[a, b, x](x-a)(x-b)dx = g[a, b, \xi] \int_a^b (x-a)(x-b)dx = \frac{g''(\xi)}{2} \frac{(a-b)^3}{6}$$

(aplicando o Teorema do Valor Intermédio para integrais).

Observação 14. De forma semelhante, podemos estabelecer o erro para a fórmula da segunda derivada, fazendo notar que ela é válida mesmo para polinómios de grau 3.

$$\begin{aligned} \partial^2(g) - \tilde{\partial^2}(g) &= \partial^2(\varepsilon_3) = \partial^2(g[z-h, z, z, z+h, x](x - z + h)(x - z)^2(x - z - h))_{x=z} = \\ &= 0 + 2g[z-h, z, z, z+h, x](x - z + h)(x - z - h)_{x=z} = -2g[z-h, z, z, z+h, z]h^2 = -\frac{9}{} \end{aligned}$$

1.10 Sistema Normal e Mínimos Quadrados

Consideramos a aproximação num subespaço vectorial de H gerado por funções base, $S = \langle g_1, \cdots, g_n \rangle$, que tem dimensão finita n, onde está definido um produto interno. Relativamente à distância definida nesse espaço de Hilbert, pelo produto interno $dist(u,v) = ||u-v|| = \langle u-v, u-v\rangle^{1/2}$, a melhor aproximação que é neste caso única, é dada pela resolução do sistema normal. Relembramos que dado $f \in H$ isso corresponde a encontrar g tal que

$$||f-g|| = \inf_{\varphi \in S} ||f-\varphi||$$

como S tem dimensão finita existe um mínimo, que resulta de encontrar equivalentemente[4] o único $g \in S$ tal que

$$\langle \varphi, f-g\rangle = 0, \forall \varphi \in S$$

De facto, escrevendo $g = a_1 g_1 + ... + a_n g_n$ basta verificar a condição para as funções base e assim $\langle g_k, f-g\rangle = 0$ leva ao sistema normal

$$\langle g_k, g\rangle = \langle g_k, f\rangle \Leftrightarrow \begin{bmatrix} \langle g_1,g_1\rangle & \cdots & \langle g_1,g_n\rangle \\ \vdots & \ddots & \vdots \\ \langle g_n,g_1\rangle & \cdots & \langle g_n,g_n\rangle \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \langle g_1,f\rangle \\ \vdots \\ \langle g_n,f\rangle \end{bmatrix}$$

Observação 15. A matriz $G = [\langle g_i, g_j\rangle]_{ij}$ é simétrica no caso real ($\langle g_i,g_j\rangle = \langle g_j,g_i\rangle$), e hermitiana no caso complexo($\langle g_i,g_j\rangle = \overline{\langle g_j,g_i\rangle}$). É ainda definida positiva, porque para qualquer $\mathbf{v} \neq \mathbf{0}$

$$\begin{aligned} \mathbf{v}^* G \mathbf{v} &= \sum_{i=1}^n \bar{v_i}(G\mathbf{v})_i = \sum_{i=1}^n \bar{v_i} \sum_{j=1}^n G_{ij} v_j = \sum_{i=1}^n \sum_{j=1}^n \bar{v_i} \langle g_i, g_j\rangle v_j \\ &= \left\langle \sum_{i=1}^n v_i g_i, \sum_{j=1}^n g_j v_j \right\rangle = \langle g_v, g_v\rangle = ||g_v||^2 > 0, \end{aligned}$$

considerando que a função $g_v = \sum_{i=1}^n v_i g_i$ não é nula porque g_i são linearmente independentes.

Observação 16. A função $g \in S$ obtida pela solução do sistema normal é denominada projecção de f sobre S, escrevendo-se

$$g = \text{proj}_S(f) = \sum_{k=1}^n a_k g_k.$$

[4]Mostramos a equivalência. É suficiente porque $\psi = g - \varphi \in S$ logo $\langle \psi, f-g\rangle = 0$ implica

$$||f-g||^2 = ||f-\varphi||^2 - ||\psi||^2 \leq ||f-\varphi||^2, \text{ porque}$$

$$||f-\varphi||^2 = \langle (f-g)+\psi, (f-g)+\psi\rangle = ||f-g||^2 + 2\text{Re}\,\langle \psi, f-g\rangle + ||\psi||^2 = ||f-g||^2 + ||\psi||^2,$$

e é necessária porque se $\langle \phi, f-g\rangle \neq 0$ para algum $\phi \in S$, tomando $\varphi = g - \psi \in S$ com $\psi = -\left\langle \hat{\phi}, f-g\right\rangle \hat{\phi}$ (aqui $\hat{\phi} = \phi/||\phi||$), obtemos $\langle \psi, f-g\rangle = -\left\langle \left\langle \hat{\phi}, f-g\right\rangle \hat{\phi}, f-g\right\rangle = -\left|\left\langle \hat{\phi}, f-g\right\rangle\right|^2 = -||\psi||^2 < 0.$

$$||f-\varphi||^2 = ||f-g||^2 + 2\ \text{Re}\,\langle \psi, f-g\rangle + ||\psi||^2 = ||f-g||^2 - 2||\psi||^2 + ||\psi||^2$$

o que seria absurdo, pois $||f-\varphi|| < ||f-g||$ e g deveria ser mínimo.

Quando a base $\hat{g}_1, \ldots, \hat{g}_n$ é ortonormada temos $\langle \hat{g}_i, \hat{g}_j \rangle = \delta_{ij}$ e a matriz G é a identidade, pelo que obtemos directamente $a_k = \langle \hat{g}_k, f \rangle$ e assim a projecção é dada por

$$g = \sum_{k=1}^{n} \langle \hat{g}_k, f \rangle \, \hat{g}_k.$$

Observação 17. O erro da aproximação $||f - g||$, é determinado imediatamente pela norma, com a solução g obtida.

1.10.1 Ortonormalização e Separabilidade

A inversa da matriz do sistema normal permite obter uma base ortonormada. Com efeito, sendo Γ_{ij} as entradas da matriz inversa de G,

$$\left\langle \sum_{k=1}^{n} \Gamma_{ik} g_k, \sum_{j=1}^{n} \Gamma_{ij} g_j \right\rangle = \sum_{k=1}^{n} \bar{\Gamma_{ik}} \left\langle g_k, \sum_{j=1}^{n} \Gamma_{ij} g_j \right\rangle$$

$$\delta_{ij} = \sum_{k=1}^{n} \Gamma_{ik} G_{kj} = \sum_{k=1}^{n} \Gamma_{ik} \langle g_k, g_j \rangle = \left\langle \sum_{k=1}^{n} \Gamma_{ik} g_k, g_j \right\rangle =$$

Lembramos que podemos sempre construir uma base ortonormada a partir de uma qualquer base, usando o método de Gram-Schmidt.

• *Processo de ortonormalização de Gram-Schmidt*:

Dada uma base inicial $\{\varphi_1, \cdots, \varphi_n\}$, construímos uma base ortonormal $\{\hat{\psi}_1, \cdots, \hat{\psi}_n\}$:

(i) $\psi_1 = \varphi_1$; $\hat{\psi}_1 = \frac{\psi_1}{||\psi_1||}$; e depois iteramos $(k = 2, \cdots, n)$:

(ii) $\psi_k = \varphi_k - \sum_{j=1}^{k-1} \left\langle \varphi_k, \hat{\psi}_j \right\rangle \hat{\psi}_j$;$\hat{\psi}_k = \frac{\psi_k}{||\psi_k||}$;

No entanto este processo de ortogonalização[5] tem um problema de mau condicionamento, resultante de um natural cancelamento subtractivo. Quando aplicado a uma base polinomial pode ser simplificado, levando aos denominados *polinómios ortogonais*.

O espaço de Hilbert H diz-se separável se admitir uma base ortonormada numerável $\hat{\phi}_1, \ldots, \hat{\phi}_n, \cdots$. Ou seja, qualquer elemento de H pode ser escrito na forma

Nessa situação $\hat{\phi}_1, \ldots, \hat{\phi}_n, \cdots$ podemos escrever qualquer $f \in H$ através da expansão de Fourier

$$f = \sum_{k=1}^{\infty} \left\langle \hat{\phi}_k, f \right\rangle \hat{\phi}_k$$

uma generalização da expansão em série de Fourier. Este é o limite da sucessão de somas finitas

$$f_n = \sum_{k=1}^{n} \left\langle \hat{\phi}_k, f \right\rangle \hat{\phi}_k$$

[5]Note que para $m < k$, como $\hat{\psi}_j$ são ortonormadas:

$$\left\langle \psi_k, \hat{\psi}_m \right\rangle = \left\langle \varphi_k, \hat{\psi}_m \right\rangle - \sum_{j=1}^{k-1} \left\langle \varphi_k, \hat{\psi}_j \right\rangle \left\langle \hat{\psi}_j, \hat{\psi}_m \right\rangle = \left\langle \varphi_k, \hat{\psi}_m \right\rangle - \sum_{j=1}^{k-1} \left\langle \varphi_k, \hat{\psi}_j \right\rangle \delta_{jm} = \left\langle \varphi_k, \hat{\psi}_m \right\rangle - \left\langle \varphi_k, \hat{\psi}_m \right\rangle = 0.$$

que correspondem à solução do sistema normal limitando a base, já que no caso de base ortonormada $\left\langle \hat{\phi}_i, \hat{\phi}_j \right\rangle = \delta_{ij}$ e a matriz do sistema normal seria a identidade.

Teorema 9. *Num espaço de Hilbert definido pela base ortonormada* $(\hat{\phi}_n)$*, temos a desigualdade de Bessel*

$$||f||^2 \geq \sum_{k=1}^{n} \left| \left\langle \hat{\phi}_k, f \right\rangle \right|^2 = ||f_n||^2$$

verificando-se $||f - f_n||^2 = ||f||^2 - ||f_n||^2$*, o que no caso limite dá a igualdade de Parseval*

$$||f||^2 = \sum_{k=1}^{\infty} \left| \left\langle \hat{\phi}_k, f \right\rangle \right|^2 .$$

Demonstração. Basta reparar que

$$\begin{aligned} ||f_n||^2 &= \left\langle \sum_{k=1}^{n} \left\langle \hat{\phi}_k, f \right\rangle \hat{\phi}_k, \sum_{j=1}^{n} \left\langle \hat{\phi}_j, f \right\rangle \hat{\phi}_j \right\rangle = \sum_{k=1}^{n} \sum_{j=1}^{n} \overline{\left\langle \hat{\phi}_k, f \right\rangle} \left\langle \hat{\phi}_j, f \right\rangle \left\langle \hat{\phi}_k, \hat{\phi}_j \right\rangle \\ &= \sum_{k=1}^{n} \sum_{j=1}^{n} \overline{\left\langle \hat{\phi}_k, f \right\rangle} \left\langle \hat{\phi}_j, f \right\rangle \delta_{kj} = \sum_{k=1}^{n} \left| \left\langle \hat{\phi}_k, f \right\rangle \right|^2 \end{aligned}$$

e como a base é ortogonal, $\left\langle f_n, \hat{\phi}_k \right\rangle = 0$, para $k > n$, logo$\langle f_n, f - f_n \rangle = 0$ e aplica-se o teorema de Pitágoras

$$||f||^2 = ||f_n + f - f_n||^2 = ||f_n||^2 + ||f - f_n||^2,$$

implica $||f||^2 \geq ||f_n||^2$. Quando $n \to \infty$ temos $||f - f_n|| \to 0$, e de $||f||^2 = \lim ||f_n||^2 + \lim ||f - f_n||^2$ resulta a igualdade de Parseval.

□

1.10.2 Caso discreto

Distinguimos os dois casos habituais, um discreto e outro contínuo. No caso do produto interno discreto consideramos vectores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$ (ou $\mathbb{C}^m$), associados aos valores das funções em nós de colocação $x_1, \cdots, x_m$, ou seja $\mathbf{u} = (u_1, \cdots, u_m) = (u(x_1), \cdots, u(x_m))$, e definimos

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^* \mathbf{v} = \sum_{k=1}^{m} \overline{u(x_k)} v(x_k)$$

no caso em que consideramos todos os pesos unitários. Podemos no entanto definir num caso mais geral:

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{Q}} = \mathbf{u}^* \mathbf{Q} \mathbf{v} = \sum_{i,j=1}^{m} \overline{u(x_i)} Q_{ij} v(x_j)$$

onde $\mathbf{Q}$ é uma matriz definida positiva simétrica (ou hermitiana), que contém uma distribuição de pesos. Quando $\mathbf{Q}$ é a identidade obtemos o caso do produto interno habitual, num outro caso simples, é uma matriz diagonal com entradas positivas $Q_{kk} = w_k > 0$. O facto da matriz ser definida positiva implica a propriedade $\langle \mathbf{u}, \mathbf{u} \rangle = \mathbf{u}^* \mathbf{Q} \mathbf{u} > 0$ quando $\mathbf{u} \neq \mathbf{0}$, e portanto $||\mathbf{u}|| = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}$. Por outro lado ser simétrica (hermitiana) implica $\langle \mathbf{v}, \mathbf{u} \rangle^* = (\mathbf{v}^* \mathbf{Q} \mathbf{u})^* = \mathbf{u}^* \mathbf{Q}^* \mathbf{v} = \mathbf{u}^* \mathbf{Q} \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle$.

Relembramos ainda que definindo $V_{ij} = \varphi_j(x_i)$, temos matrizes $\mathbf{V}$ (de dimensão $m \times n$), podemos ver como o sistema normal discreto generaliza o sistema de Vandermonde. Com efeito, a matriz do sistema normal fica

$$\Phi_{ij} = \langle \varphi_i, \varphi_j \rangle = \sum_{k=1}^{m} \overline{\varphi_i(x_k)} \varphi_j(x_k) = \sum_{k=1}^{m} \bar{V_{ki}} V_{kj} = [\mathbf{V}^* \mathbf{V}]_{ij}$$

ou seja, $\Phi = \mathbf{V}^* \mathbf{V}$.

Escrevendo ainda $F_i = \langle \varphi_i, f \rangle = \sum_{k=1}^{m} \overline{\varphi_i(x_k)} f(x_k) = \sum_{k=1}^{m} \bar{V_{ki}} f_k = [\mathbf{V}^* \mathbf{f}]_i$ temos

$$\Phi \mathbf{a} = \mathbf{F} \Leftrightarrow \mathbf{V}^* \mathbf{V} \mathbf{a} = \mathbf{V}^* \mathbf{f}$$

e concluímos que do sistema de Vandermonde $\mathbf{V}\mathbf{a} = \mathbf{f}$ (com $\mathbf{V}$ de dimensão $m \times n$), chegamos ao sistema normal multiplicando por $\mathbf{V}^*$ em ambos os lados da equação (é aliás isso que permitirá equilibrar a equação, tornando a matriz quadrada $n \times n$, ao multiplicar pela conjugada, doutra forma teríamos mais equações do que incógnitas).

Observação 18. No caso em que $\mathbf{Q} \neq \mathbf{I}$ é semelhante

$$\Phi_{ij} = \langle \varphi_i, \varphi_j \rangle_{\mathbf{Q}} = \sum_{k,r=1}^{m} \overline{\varphi_i(x_k)} Q_{kr} \varphi_j(x_r) = \sum_{k,r=1}^{m} \bar{V_{ki}} Q_{kr} V_{rj} = [\mathbf{V}^* \mathbf{Q} \mathbf{V}]_{ij}$$

e como $F_i = \langle \varphi_i, f \rangle_{\mathbf{Q}} = \sum_{k,r=1}^{m} \overline{\varphi_i(x_k)} Q_{kr} f(x_r) = \sum_{k,r=1}^{m} \bar{V_{ki}} Q_{kr} f_r = [\mathbf{V}^* \mathbf{Q} \mathbf{f}]_i$ então o sistema normal fica

$$\mathbf{V}^* \mathbf{Q} \mathbf{V} \mathbf{a} = \mathbf{V}^* \mathbf{Q} \mathbf{f}$$

resumindo-se a multiplicar por $\mathbf{V}^* \mathbf{Q}$ (à esquerda), ambos os lados do sistema de Vandermonde.

1.10.3 Caso contínuo

No caso do produto interno $L^2(a, b)$ consideramos uma função peso w tal que $w(x) > 0$ (q.t.p.).

$$\langle u, v \rangle_w = \int_a^b w(t) \bar{u}(t) v(t) dt$$

Notamos que no caso $w \equiv 1$ isto corresponde ao produto interno $L^2(a, b)$ habitual, mas geralmente considera-se $L^2_w = \{u : u/\sqrt{w} \in L^2(a, b)\}$.
Há produtos internos conhecidos pela sua associação a polinómios ortogonais:

- (Chebyshev) produto interno no intervalo $(-1, 1)$, com $w(x) = (1 - x^2)^{-1/2}$.
- (Hermite) produto interno em $\mathbb{R}$ com $w(x) = e^{-x^2}$.
- (Laguerre) produto interno em $\mathbb{R}^+$ com $w(x) = x^\alpha e^{-x^2}$, com $\alpha > -1$.

Para evitar a resolução do sistema normal, podem procurar-se funções que sejam ortogonais, e assim reduzam o sistema a uma matriz diagonal, ou ainda à identidade sendo ortonormadas.

1.11 Polinómios ortogonais

Teorema 10. *Consideremos polinómios q_k mónicos de grau k. Dados $q_0, q_1 : \langle q_0, q_1\rangle_w = 0$, então a sucessão de polinómios ortogonais é definida pela fórmula de recorrência*

$$q_{k+1}(x) = xq_k(x) - \langle \hat{q}_k, tq_k\rangle_w \hat{q}_k(x) - \frac{||q_k||_w^2}{||q_{k-1}||_w^2} q_{k-1}(x)$$

onde $\hat{q}_k = q_k/||q_k||$ definirá a base ortonormada correspondente (notação:$(tq_k)(t) = tq_k(t)$).

Demonstração. Demonstramos por indução, sendo válido para os iniciais pois $\langle q_0, q_1\rangle_w = 0$. Admitimos assim que temos já uma base ortogonal $\{q_0, q_1, \cdots, q_n\}$ e vamos obter q_{n+1}, que sendo mónico pode escrever-se na forma

$$q_{k+1}(x) = xq_k(x) + c_k q_k(x) + \cdots + c_0 q_0(x).$$

Pretendemos encontrar os c_k que fazem $\langle q_j, q_{k+1}\rangle_w = 0$, para qualquer $j \leq k$.
(i) Para $j = k$ obtemos $\langle q_k, q_{k+1}\rangle_w = \langle q_k, tq_k\rangle_w + \langle q_k, c_k q_k\rangle_w + 0 + \ldots + 0$, pois $\langle q_k, q_j\rangle_w = 0$ para $j < k$. Para a ortogonalidade $\langle q_k, q_{k+1}\rangle_w = 0$, resulta

$$c_k \langle q_k, q_k\rangle_w = -\langle q_k, tq_k\rangle_w \Leftrightarrow c_k = \frac{-\langle q_k, tq_k\rangle_w}{||q_k||_w^2}.$$

(ii) Para $j = k-1$ obtemos $\langle q_{k-1}, q_{k+1}\rangle_w = \langle q_{k-1}, tq_k\rangle_w + 0 + \langle q_{k-1}, c_{k-1}q_{k-1}\rangle_w + 0 + \ldots + 0$, pois $\langle q_{k-1}, q_j\rangle_w = 0$ para $j \neq k-1$. Para a ortogonalidade $\langle q_{k-1}, q_{k+1}\rangle_w = 0$, resulta

$$c_{k-1} \langle q_{k-1}, q_{k-1}\rangle_w = -\langle q_{k-1}, tq_k\rangle_w \Leftrightarrow c_{k-1} = \frac{-\langle q_{k-1}, tq_k\rangle_w}{||q_{k-1}||_w^2}$$

no caso do produto interno considerado em $(a, b) \subseteq \mathbb{R}$, a expressão simplifica-se[6]: $\langle q_{k-1}, tq_k\rangle_w = \langle tq_{k-1}, q_k\rangle_w$ e como por hipótese $q_k(t) = tq_{k-1}(t) + p_{k-1}(t)$, obtemos por ortogonalidade[7] $\langle p_{k-1}, q_k\rangle_w = 0$ e por isso:

$$\langle q_k, q_k\rangle_w = \langle tq_{k-1}, q_k\rangle_w + \langle p_{k-1}, q_k\rangle_w = \langle tq_{k-1}, q_k\rangle_w ,$$

concluindo-se que $c_{k-1} = \frac{-\langle q_{k-1}, tq_k\rangle_w}{||q_{k-1}||_w^2} = \frac{-||q_k||_w^2}{||q_{k-1}||_w^2}$.
(iii) Finalmente, para $j \leq k-2$ temos $\langle q_j, q_{k+1}\rangle_w = \langle q_j, tq_k\rangle_w + 0 + \langle q_j, c_j q_j\rangle_w + 0 + \ldots + 0$, o que implica (em $\mathbb{R}$):

$$c_j ||q_j||_w^2 = -\langle q_j, tq_k\rangle_w = -\langle tq_j, q_k\rangle_w = 0$$

pois tq_j tem grau $j + 1 < k$, o que implica $c_j = 0$ para $j \leq k-2$. □

Exercício 22. Mostre que se o intervalo for $(a, b) = (-R, R)$, e $q_0(x) = 1, q_1(x) = x$, então a fórmula de recorrência simplifica-se quando o peso w é uma função par:

$$q_{k+1}(x) = xq_k(x) - \frac{||q_k||_w^2}{||q_{k-1}||_w^2} q_{k-1}(x)$$

e os polinómios q_k são pares (resp. ímpares) quando k é par (resp. ímpar).

[6] Mesmo no caso complexo:
$\langle q_{k-1}, tq_k\rangle_w = \int_a^b w(t)\overline{q_{k-1}(t)}tq_k(t)dt = \int_a^b w(t)\overline{tq_{k-1}(t)}q_k(t)dt = \langle tq_{k-1}, q_k\rangle_w$

[7] O polinómio p_{k-1} pode ser escrito na combinação da base de q_j com $j < k$, logo ortogonais a q_k. Aliás, de forma geral, q_k será ortogonal a qualquer polinómio de grau inferior a k.

Resolução: Basta reparar que (assumimos o caso real)

$$\langle \hat{q}_k, tq_k \rangle_w = \frac{1}{||q_k||_w} \int_{-R}^{R} w(t) q_k(t)^2 t \, dt = 0$$

porque $w(t)q_k(t)^2 t$ é ímpar. Assim, por indução, q_{2k+1} é ímpar pois resulta de soma de xq_{2k} que é impar (poisq_{2k}é par) com Cq_{2k-1} que é ímpar (note que $q_0 = 1$ par e $q_1(x) = x$ é ímpar. Analogamente q_{2k+2} será par.

Exemplo 5. Considerando o produto interno com $w \equiv 1$ em $[-1, 1]$, obtemos os polinómios de Legendre P_n :

$$P_{n+1}(x) = xP_n(x) - \frac{||P_n||^2}{||P_{n-1}||^2} P_{n-1}(x),$$

e como $P_0(x) = 1, P_1(x) = x$, obtemos $P_2(x) = xP_1(x) - \frac{||P_1||^2}{||P_0||^2} P_0(x) = x^2 - \frac{\int_{-1}^{1} x^2 dx}{\int_{-1}^{1} 1 dx} = x^2 - \frac{1}{3}$. De forma semelhante, obtemos os seguintes $P_3(x) = x^3 - \frac{3}{5}x$, podendo mesmo estabelecer-se uma fórmula de recorrência:

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

Exemplo 6. Considerando o produto interno com $w \equiv (1 - x^2)^{-1/2}$ em $]-1, 1[$, obtemos os polinómios de Chebyshev $\tilde{T}_n$ (na forma mónica):

$$\tilde{T}_0(x) = 1, \tilde{T}_1(x) = x, \tilde{T}_2(x) = x^2 - \frac{1}{2}, \tilde{T}_3(x) = x^3 - \frac{3}{4}, ...$$

sendo mais conhecidos na sua forma não mónica

$$T_n(x) = \cos(n \arccos(x)),$$

com a relação de recorrência $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$.

Teorema 11. *As raízes dos polinómios ortogonais estão no intervalo* (a, b).

Demonstração. O polinómio mónico q_n pode ser escrito em termos das suas raízes $z_1, \ldots, z_n$

$$q_n(x) = (x - z_1) \cdots (x - z_n) = \underbrace{(x - z_1) \cdots (x - z_m)}_{p_m(x)} p_{n-m}(x),$$

em que designamos $z_1, \ldots, z_m \in [a, b]$ as m raízes no intervalo. Pretendemos mostrar que $m = n$, e notamos que p_{n-m} não tendo raízes em $[a, b]$, não muda de sinal nesse intervalo. Portanto,

$$\langle q_n, p_m \rangle_w = \int_a^b w(x) q_n(x) p_m(x) dx = \int_a^b w(x) p_m(x)^2 p_{n-m}(x) dx,$$

e como $w > 0$, $p_m^2 \geq 0$, o sinal da função integranda é o de p_{n-m} e o integral não é nulo. Porém, como q_n é ortogonal a todos os polinómios de grau inferior a n, o produto interno só não é nulo, $\langle q_n, p_m \rangle_w \neq 0$, quando $n = m$.

Finalmente mostramos que as raízes são simples. Se, por exemplo, z_1 fosse raiz dupla, com $z_1 = z_2$, então definindo $p_{n-2}(x) = (x - z_3) \cdots (x - z_n)$ implicaria que $\langle q_n, p_{n-2} \rangle_w = 0$ (pois p_{n-2} tem grau inferior), o que contradiria:

$$\langle q_n, p_{n-2} \rangle_w = \int_a^b w(x) q_n(x) p_{n-2}(x) dx = \int_a^b w(x)(x - z_1)^2 (x - z_3)^2 \cdots (x - z_n)^2 dx > 0,$$

porque tratar-se-ia do integral de uma função não negativa e não nula. □

1.11.1 Fórmulas de Integração de Gauss

A partir dos n valores das raízes dos polinómios ortogonais, que servem como nós, é possível encontrar fórmulas de quadratura de grau $2n-1$.

Teorema 12. *Seja P_n um polinómio ortogonal de grau n, com raízes $z_1, \cdots, z_n$. Para a integração*

$$I_w(f) = \int_a^b w(x)f(x)dx$$

a fórmula de quadratura de Gauss (em que os nós são as raizes de P_n)

$$Q_w(f) = \alpha_1 f(z_1) + \cdots + \alpha_n f(z_n)$$

tem grau $2n-1$, e os pesos são obtidos por $\alpha_k = I_w(L_k)$, em que L_k são os polinómios base de Lagrange (os nós de interpolação são os z_k).

Demonstração. Sendo p_{2n-1} um polinómio de grau $\leq 2n-1$, queremos ver que $I_w(p_{2n-1}) = Q_w(p_{2n-1})$.
Para esse efeito, consideramos a divisão de polinómios

$$p_{2n-1}(x) = q_{n-1}(x)P_n(x) + r_{n-1}(x)$$

em que q_{n-1} é o quociente (grau $n-1$) e r_{n-1} tem grau $\leq n-1$. Assim, como $P_n(z_k) = 0$,

$$Q_w(p_{2n-1}) = \sum \alpha_k \left(q_{n-1}(z_k)P_n(z_k) + r_{n-1}(z_k)\right) = \sum \alpha_k r_{n-1}(z_k) = Q_w(r_{n-1}).$$

Ora, r_{n-1} (com grau $\leq n-1$) pode ser escrito pela fórmula de Lagrange

$$r_{n-1}(x) = r_{n-1}(z_1)L_1(x) + ...r_{n-1}(z_n)L_n(x)$$

e temos $I_w(r_{n-1}) = r_{n-1}(z_1)I_w(L_1) + ...r_{n-1}(z_n)I_w(L_n) = Q_w(r_{n-1})$. Assim, $Q_w(p_{2n-1}) = Q_w(r_{n-1}) = I_w(r_{n-1})$, falta apenas ver que $I_w(r_{n-1}) = I_w(p_{2n-1})$:

$$I_w(p_{2n-1}) = I_w(q_{n-1}P_n) + I_w(r_{n-1}) = I_w(r_{n-1}),$$

porque $I_w(q_{n-1}P_n) = \langle q_{n-1}, P_n \rangle_w = 0$, pois P_n é ortogonal aos polinómios de grau $n-1$. Provámos assim que tem pelo menos grau $2n-1$, e não tem grau $2n$ porque $Q_w(P_n^2) = 0 \neq ||P_n||_w^2 = I_w(P_n^2)$. □

Exercício 23. Calcule uma fórmula de Gauss $Q_c(f)$ para a aproximação de

$$I_c(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}dx$$

que seja exacta para polinómios de grau $2n-1$.

Resolução: Basta reparar que $w(x) = (1-x^2)^{-1/2}$ é o peso para os polinómios de Chebyshev. Assim sendo $z_1, ..., z_n$ os zeros de T_n (que são os nós de Chebyshev) obtemos

$$Q_c(f) = \alpha_1 f(z_1) + \cdots + \alpha_n f(z_n)$$

em que os pesos α_k podem ser obtidos pelo método dos coeficientes indeterminados, resolvendo $Q_c(t^k) = I_c(t^k)$, para $k = 1, ..., n$, ou através de

$$\alpha_k = I_c(L_k) = \int_{-1}^{1} \frac{L_k(x)}{\sqrt{1-x^2}}dx = \langle 1, L_k \rangle_w .$$

Salientamos ainda que se $p_m(x) = a_0T_0(x) + \cdots + a_mT_m(x)$, temos $I_c(p_m) = 2\pi a_0$ porque

$$I_c(p_n) = \langle 1, p_n \rangle_w = a_0 \langle 1, T_0 \rangle_w + ... + a_n \langle 1, T_n \rangle_w = a_0(2\pi) + 0 + ... + 0,$$

já que $T_0 \equiv 1$, e todos os outros T_k lhe são ortogonais.

1.12 Outras bases ortogonais

Para além dos polinómios ortogonais, é bem conhecida a base de funções na série de Fourier, que é ortogonal em $[-\pi, \pi]$ (com $w = 1$),

$$S_n = \langle 1, \sin(t), \cos(t), ..., \sin(nt), \cos(nt) \rangle$$

verificando-se $\langle 1, 1 \rangle = 2\pi$, $\langle \sin(kt), \sin(kt) \rangle = \langle \cos(kt), \cos(kt) \rangle = \pi$, o que permite obter facilmente a expansão de Fourier

$$f(t) \approx f_n(t) = \frac{1}{2\pi} \langle 1, f \rangle + \frac{1}{\pi} \sum_{k=1}^{n} \langle \sin(kt), f \rangle \sin(kt) + \langle \cos(kt), f \rangle \cos(kt)$$

ou ainda na forma complexa

$$f(t) \approx f_n(t) = \frac{1}{2\pi} \sum_{k=0}^{n} \langle \exp(ikt), f \rangle \exp(ikt)$$

que converge em $L^2(-\pi, \pi)$.

1.13 Aproximação em Espaços de Banach

No caso de espaços métricos, podemos não ter associado um produto interno, como é o caso do espaço da funções contínuas $C[a, b]$ com a norma do máximo:

$$||f||_\infty = \max_{x \in [a,b]} |f(x)|$$

neste caso, dado um subespaço $S = \langle \varphi_1, ..., \varphi_n \rangle$, e uma função $f \notin S$, o objectivo será encontrar a função $g \in S$ que minimiza a distância

$$dist(f, g) = ||f - g||_\infty$$

mas não havendo produto interno, não temos nenhum sistema normal associado, que permita uma solução rapidamente.

1.13.1 Melhor aproximação uniforme (mini-max)

O problema de minimização da distância ao subespaço, é muitas vezes designado Mini-Max, pois trata-se de encontrar $g \in S$:

$$||f - g||_\infty = \min_{\varphi \in S} ||f - \varphi||_\infty = \min_{\varphi \in S} \max_{x \in [a,b]} |f(x) - \varphi(x)|,$$

e a função g é denominada a *melhor aproximação uniforme* de f em S.

Ao contrário do que acontece no caso da norma associada a um produto interno, neste caso não há garantia de unicidade da melhor aproximação uniforme, pelo menos no caso geral. Iremos ver que há unicidade quando consideramos $S = \mathbb{P}_n$.

Exemplo 7. (Contra-exemplo de unicidade no caso geral). Pretendemos aproximar $f(x) = x$ em $S = \langle 1, t^2 \rangle$, no intervalo $[-1, 1]$. Ou seja, queremos encontrar $g(x) = a + bx^2$ que minimize a distância

$$||f - g||_\infty = \max_{x\in[-1,1]} |(a + bx^2) - x|.$$

Começamos por notar que nos extremos do intervalo ($x = \pm 1$), obtemos

$$|f(-1) - g(-1)| = |a + b + 1|, \quad |f(1) - g(1)| = |a + b - 1|,$$

o máximo destes dois valores é $|a+b|+1$ (porque se $a+b \geq 0$ é $a+b+1$, e se $a+b \leq 0$ é $-(a+b-1)$), e assim a distância mínima possivel será 1, quando $a = -b$. A função $g \equiv 0$ permite imediatamente obter uma solução minimizante pois $||f - 0||_\infty = \max_{[-1,1]} |x| = 1$. No entanto, podemos encontrar mais soluções, para isso vamos ver se há máximo local no interior do intervalo para $f - g$ com $g(x) = b(x^2 - 1)$. Como

$$0 = (f - g)'(z) = 1 - 2bz \Leftrightarrow z = \tfrac{1}{2b} \quad (b\neq 0),$$

a função é monótona se $z \notin (-1, 1)$, ou seja sempre que $|b| \leq \frac{1}{2}$. Assim, os máximos estão nos extremos já analisados, e todas as funções $g(x) = b(x^2 - 1), g \in S$, são minimizantes desde que $|b| \leq \frac{1}{2}$.

Esta situação é bastante diferente da que ocorreria num espaço de Hilbert, pois pelos mínimos quadrados, o sistema normal teria segundo membro nulo, pois$\langle 1, t\rangle = 0, \langle t^2, t\rangle = 0$, consequentemente a única solução minimizante seria $g \equiv 0$.

Observação 19. A melhor aproximação uniforme pode não ter solução única, como já vimos no contra-exemplo. De um modo geral, dado um subespaço de funções $S = \langle g_1, \cdots, g_n\rangle$, a unicidade é garantida se e só se for verificada a independência linear para qualquer conjunto de n nós distintos $X = \{x_1, ..., x_n\} \in [a, b]$, é a denominada *Condição de Haar*, ou seja a matriz de Vandermonde generalizada é sempre invertivel:

$$\det \begin{bmatrix} g_1(x_1) & \cdots & g_n(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \cdots & g_n(x_n) \end{bmatrix} \neq 0, \tag{1.13.1}$$

para quaisquer nós distintos $\{x_1, ..., x_n\} \in [a, b]$. Notamos que isso não era válido no contraexemplo com $S = \langle 1, t^2\rangle$, pois $\det \begin{bmatrix} 1 & x_1^2 \\ 1 & x_2^2 \end{bmatrix} = x_2^2 - x_1^2 = 0$ pode ser obtido com $x_1 = -x_2 \neq x_2$ sempre que considerarmos um intervalo da forma $[-a, a]$.

No caso de considerarmos como subespaço os polinómios completos (sem ausência de nenhum grau intermédio), ou seja $\mathbb{P}_n$ polinómios de grau menor ou igual a n, definidos pela base canónica $\mathbb{P}_n = \langle 1, t, ..., t^n\rangle$, a matriz de Vandermonde é sempre invertível, para quaisquer nós distintos, verificando-se a unicidade. Iremos agora ver como se pode caracterizar a melhor aproximação uniforme.

Teorema 13. *(de La Vallée-Poussin). Seja $f \in C[a, b]$, e $p_n \in \mathbb{P}_n$ um polinómio que para $n+2$ pontos: $a \leq x_0 < x_1 < ... < x_{n+1} \leq b$ verifique*

$$f(x_k) - p_n(x_k) = (-1)^k e_k$$

sempre com $e_k > 0$ (ou alternativamente sempre $e_k < 0$), então

$$\min_k |e_k| \leq \min_{q\in\mathbb{P}_n} ||f - q||_\infty$$

Demonstração. Suponhamos por absurdo que para $q_n \in \mathbb{P}_n$ tinhamos $||f - q_n||_\infty < \min_k |e_k|$, então considerando

$$(q_n - p_n)(x_k) = (f - p_n)(x_k) - (f - q_n)(x_k) = (-1)^k e_k - (f - q_n)(x_k)$$

e como $|(f - q_n)(x_k)| < |e_k|$ o sinal de $(q_n - p_n)(x_k)$ é o mesmo de $(-1)^k e_k$, por isso a oscilação entre os dois polinómios $q_n - p_n$ implica $n+1$ zeros, cada um em $[x_k, x_{k+1}]$ (com $k = 0, ..., n+1$) e sendo de grau $\leq n$, tem que ser $q_n - p_n \equiv 0$. Isto é absurdo porque

$$|f(x_k) - p_n(x_k)| = |e_k| > ||f - q_n||_\infty = ||f - p_n||_\infty.$$

□

Exemplo 8. Por exemplo, considerando a aproximação de $f(x) = x^{1/2}$, por $p_1(x) = x + 1/5$ no intervalo $[0, 1]$, ao escolher 3 pontos (o polinómio é de grau 1), $x_0 = 0, x_1 = 1/4, x_2 = 1$ vemos que

$$(f - p_1)(0) = -\frac{1}{5} = e_0, (f - p_1)(\frac{1}{4}) = \frac{1}{20} = -e_1, (f - p_1)(1) = -\frac{1}{5} = e_2,$$

e concluímos que $\frac{1}{20} = \min\{|e_0|, |e_1|, |e_2|\} \leq \min_{q \in \mathbb{P}_1} ||f - q||_\infty$. O teorema seguinte mostra que, como $|e_k|$ são diferentes, isto não garante que seja a melhor aproximação (ver também exercício seguinte).

Teorema 14. *(da equioscilação, de Chebyshev). Um polinómio p_n de grau $\leq n$ é a melhor aproximação uniforme de $f \in C[a, b]$ no espaço $\mathbb{P}_n$ se e só se existem (pelo menos) $n+2$ pontos $x_0, ..., x_{n+1} \in [a, b]$, distintos tais que verificam a equioscilação:*

$$(f - p_n)(x_k) = \pm(-1)^k ||f - p_n||_\infty \tag{1.13.2}$$

Essa melhor aproximação é única nos polinómios $\mathbb{P}_n$.

Demonstração. ($\Leftarrow$) Verificando p_n a propriedade de oscilação com constante $|e_k| = ||f - p_n||_\infty$, pelo teorema de La Vallée-Poussin temos

$$||f - p_n||_\infty \leq \min_{q \in \mathbb{P}_n} ||f - q||_\infty = d$$

e portanto $p_n \in \mathbb{P}_n$ é um polinómio de mínimo. ($\Rightarrow$) Para a condição necessária, ver e.g. [Atkinson].
(Unicidade) Supondo que p_n e q_n são melhores aproximações também seria a média pois:

$$||f - \tfrac{1}{2}(p_n + q_n)||_\infty \leq ||\tfrac{1}{2}(f - p_n)||_\infty + ||\tfrac{1}{2}(f - q_n)||_\infty \leq \tfrac{1}{2}(d + d) = d$$

e pela condição necessária haveria $z_0, ..., z_{n+1} \in [a, b]$, onde muda de sinal

$$\pm(-1)^k d = f(z_k) - \tfrac{1}{2}(p_n + q_n)(z_k) = \tfrac{1}{2}(f - p_n)(z_k) + \tfrac{1}{2}(f - q_n)(z_k)$$

e como $\frac{1}{2}|(f - q_n)(z_k)| \leq \frac{1}{2}d$, a oscilação dá-se nesses pontos z_k também para $f - p_n$ (ou analogamente para $f - q_n$). Tal como no teorema de La Vallée-Poussin, obtemos

$$(q_n - p_n)(z_k) = (f - p_n)(z_k) - (f - q_n)(z_k) = \pm(-1)^k d - (f - q_n)(z_k)$$

e daqui, há $n + 1$ zeros, e $p_n = q_n$. □

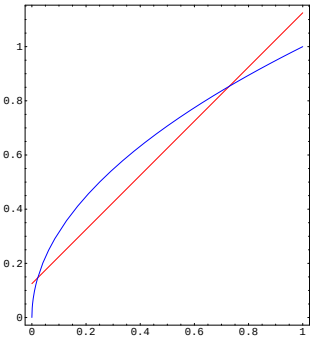

Figura 1.13.1: Melhor aproximação uniforme de $\sqrt{x}$ pela recta $x + \frac{1}{8}$, em [0,1].

Observação 20. Daqui concluímos que é sempre possível encontrar uma melhor aproximação em $\mathbb{P}_n$, pelo que aumentando o grau do polinómio essa aproximação será melhor ou igual, e pelo Teorema de Weierstrass a aproximação polinomial converge, diminuindo a distância até zero.

Exercício 24. Mostre que a melhor aproximação uniforme de $f \in C[a,b]$ por uma constante é dada por $c = \frac{f(x_m)+f(x_M)}{2}$ em que x_m é o ponto de mínimo e x_M o ponto de máximo em $[a,b]$.

Resolução: Como $r(x) = f(x) - c$ verifica $r(x_m) = f(x_m) - c = \frac{f(x_m)-f(x_M)}{2} = -d$, e por outro lado $r(x_M) = f(x_M) - c = \frac{f(x_M)-f(x_m)}{2} = d$, resta ver que $d = ||f - c||_\infty$ para concluir que é a melhor aproximação pelo teorema de Chebyshev. Isso é imediato pois

$$f(x_m) \leq f(x) \leq f(x_M) \Rightarrow -d = f(x_m) - c \leq f(x) - c \leq f(x_M) - c = d \Rightarrow |f(x) - c| \leq d.$$

Também obteríamos isso imediatamente pelo algoritmo de Remes, quando $X^{(0)} = \{x_m, x_M\}$, pois

$$\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} f(x_m) \\ f(x_M) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f(x_m) \\ f(x_M) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} f(x_m) + f(x_M) \\ f(x_m) - f(x_M) \end{bmatrix}.$$

Exercício 25. Determine a melhor aproximação $\mathbb{P}_1$ da função $f(x) = \sqrt{x}$ no intervalo $[0,1]$.

Resolução: Pelo critério de equioscilação, procuramos que $f(x_k) - p_1(x_k) = (-1)^k d$, sendo incógnitas d, x_0, x_1, x_2. Neste caso, para $p_1(x) = a + bx$, vemos que

$$f'(t) - p_1'(t) = \tfrac{1}{2}t^{-1/2} - b = 0 \Leftrightarrow t = \frac{1}{4b^2}$$

havendo apenas um ponto crítico, usamos as extremidades como pontos adicionais, obtendo o sistema não linear

$$\begin{cases} f(0) - p_1(0) = d \\ f(t) - p_1(t) = -d \\ f(1) - p_1(1) = d \end{cases} \begin{cases} -a = d \\ \frac{1}{2b} - a - \frac{1}{4b} = -d \\ 1 - a - b = d \end{cases} \begin{cases} a = -d \\ \frac{1}{4b} = -2d \Leftrightarrow d = -\frac{1}{8} \\ b = 1 \end{cases}$$

ou seja, $p_1(x) = \frac{1}{8} + x$, com $||f - p_1||_\infty = -d = \frac{1}{8}$.

Nem sempre é possível obter tão directamente a melhor aproximação, que envolverá cálculos não lineares, pelo que noutros casos podemos usar o algoritmo de Remes.

Algoritmo de Remes

(i) Definimos um conjunto $X^{(0)} = \{x_0, \cdots, x_{n+1}\}$ pré-definindo os pontos[8].

(ii) Dado o conjunto $X^{(m)}$, resolvemos o sistema linear

[8]Uma escolha inicial possivel são os nós de Chebyshev.

$a_0 + a_1 x_k + ... + a_n x_k^n + (-1)^k d = f(x_k)$

$$\begin{bmatrix} 1 & x_0 & \cdots & x_0^n & (-1)^0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n+1} & \cdots & x_{n+1}^n & (-1)^{n+1} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ a_n \\ d \end{bmatrix} = \begin{bmatrix} f(x_0) \\ \vdots \\ \vdots \\ f(x_{n+1}) \end{bmatrix}$$

(iii) Definimos $p_n(x) = a_0 + a_1 x + ... + a_n x^n$, e procuramos $X^{(m+1)} = \{x_0^{'}, \cdots, x_{n+1}^{'}\}$ que maximizem $|f(x) - p_n(x)|$, o que pode ser feito resolvendo $f'(x) - p_n'(x) = 0$ (quando $f \in C^1[a,b]$), adicionando os extremos a, b se necessário.

(iv) Se $X^{(m+1)} = X^{(m)}$ o critério de equioscilação é verificado, senão regressamos a (ii) com o novo conjunto de pontos $X^{(1)}$.

Exercício 26. Aplique o algoritmo de Remes para determinar a recta que é a melhor aproximação uniforme de $f(x) = x^{3/2}$ em $[0,1]$.

Resolução: Iniciamos com $X^{(0)} = \{0, \frac{1}{2}, 1\}$, e resolvemos o sistema linear

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & \frac{1}{2} & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ (\frac{1}{2})^{3/2} \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} a_0 \\ a_1 \\ d \end{bmatrix} = \frac{1}{8} \begin{bmatrix} \sqrt{2}-2 \\ 8 \\ 2-\sqrt{2} \end{bmatrix}$$

obtemos $p_1(x) = \frac{1}{8}(\sqrt{2}-2) + x$, e a função $r(x) = f(x) - p_1(x)$ verifica $r'(x) = \frac{3}{2}x^{1/2} - 1 = 0$ quando $x = \frac{4}{9}$. Este ponto, juntamente com os extremos, definem $X^{(1)} = \{0, \frac{4}{9}, 1\}$ e temos

$$\begin{aligned} (f-p_1)(X^{(1)}) &= \left(0-\tfrac{1}{8}(\sqrt{2}-2),\ \tfrac{8}{27}-\tfrac{1}{8}(\sqrt{2}-2)-\tfrac{4}{9},\ 1-\tfrac{1}{8}(\sqrt{2}-2)-1\right) \\ &= (-0.0732..,\ 0.0749..,-0.0733..) = (-e_0, e_1, -e_2), \end{aligned}$$

(pelo teorema de La Vallée-Poussin sabemos que a distância será maior que$|e_0| = |e_2| = \frac{1}{8}(\sqrt{2}-2) < |e_1|$). Como a condição de equioscilação não é verificada, efectuamos novo passo:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & \frac{4}{9} & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{8}{27} \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} a_0 \\ a_1 \\ d \end{bmatrix} = \frac{1}{27} \begin{bmatrix} -2 \\ 27 \\ 2 \end{bmatrix}$$

obtendo agora $p_1(x) = x - \frac{2}{27}$, como $r'(x) = \frac{3}{2}x^{1/2} - 1$, mantém-se o ponto crítico interno $x = \frac{4}{9}$. Ou seja $X^{(2)} = X^{(1)}$, e o algoritmo pára. De facto $p_1(x) = x - \frac{2}{27}$ é a melhor aproximação uniforme em $\mathbb{P}_1$, pois ao resolver o sistema linear já verificámos a condição de equioscilação de Chebyshev, e a distância obtida

$$||f - p_1||_\infty = d = \frac{2}{27} \approx 0.074...$$

Nestes casos mais simples, o método de Remes pode dar a solução em poucas iteradas. Notamos que pode ser computacionalmente delicado definir os pontos de máximo para $|f - p_n|$, quando n é grande.

Observação 21. No caso de procurar a melhor aproximação uniforme linear, é útil usar a expressão pré-determinada da inversa:

$$\begin{bmatrix} 1 & x_0 & 1 \\ 1 & x_1 & -1 \\ 1 & x_2 & 1 \end{bmatrix}^{-1} = \frac{1}{2(x_2 - x_0)} \begin{bmatrix} x_1 + x_2 & x_2 - x_0 & -(x_0 + x_1) \\ -2 & 0 & 2 \\ x_2 - x_1 & x_0 - x_2 & x_1 - x_0 \end{bmatrix}$$

neste último exercício como só mudámos o $x_1$usámos a expressão

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & x_1 & -1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} x_1 + 1 & 1 & -x_1 \\ -2 & 0 & 2 \\ 1 - x_1 & -1 & x_1 \end{bmatrix}.$$

Para encontrar a melhor aproximação uniforme em $\mathbb{P}_2$ já seria necessário considerar uma matriz 4×4.

1.13.2 Nós de Chebyshev

Na fórmula de erro de interpolação polinomial (de Lagrange):

$$E(x) = (f - p_n)(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W_{n+1}(x)$$

aparece o termo $W_{n+1}(x) = (x-x_0)\cdots(x-x_n)$ em que a escolha dos nós desempenha um papel importante, já que não depende de f. Ou seja, pela simples escolha dos nós de interpolação $\{x_0, \cdots, x_n\}$, podemos minimizar o valor $||W_{n+1}||_\infty$.

Como W_{n+1} é um polinómio mónico, que escrevemos $W_{n+1}(x) = x^{n+1} - q_n(x)$, a questão pode ser vista no contexto da melhor aproximação polinomial, para minimizar a distância entre $f(x) = x^{n+1}$ e os polinómios $\mathbb{P}_n$. A solução deste problema de minimização está relacionada com os *polinómios de Chebyshev*, definidos no intervalo $[-1, 1]$ por

$$T_n(x) = \cos(n \arccos(x)).$$

Vimos que os polinómios de Chebyshev verificam $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. É claro que T_n é um polinómio de grau n em que o coeficiente de grau n resulta das sucessivas multiplicações por 2, ficando 2^{n-1}. O polinómio fica mónico se dividirmos por esse termo, ou seja

$$\tilde{T}_{n+1}(x) = \frac{1}{2^n} T_{n+1}(x)$$

é o polinómio mónico associado a T_{n+1}.

As n raízes dos polinómios de Chebyshev T_n (ou $\tilde{T}_n$) também são fáceis de obter:

$$t_k = \cos(\frac{2k+1}{2n}\pi), \text{ para } k = 0, ..., n-1, \tag{1.13.3}$$

pois $T_n(t_k) = \cos(n \arccos(t_k)) = \cos(n\frac{2k+1}{2n}\pi) = \sin(k\pi) = 0$.

Estes valores t_k são os denominados nós de Chebyshev e são a escolha optimal para minimizar o erro de interpolação polinomial no intervalo $[-1, 1]$. Isso é consequência da melhor aproximação de x^{n+1} pelo Teorema de Chebyshev.

Proposição 10. *Mostre que $||W_n||_\infty$ é minimizado quando se escolhem t_k (os nós de Chebyshev de T_n).*

Demonstração. (Exercício): Já vimos que este problema está relacionado com minimizar a distância de x^n a $\mathbb{P}_{n-1}$, usando a factorização de um polinómio mónico nas suas raízes $x_0, ..., x_{n-1}$:

$$W_n(x) = (x - x_0)\cdots(x - x_n) = x^n - q_w(x),$$

queremos demonstrar que a distância é mínima com $\tilde{T}_n(x) = (x-t_0)\cdots(x-t_{n-1}) = x^n - q_T(x)$.

Como T_n é definido por um co-seno, atinge os mínimos/máximos (que são -1,+1), alternadamente nos pontos

$$t'_k = \cos(k\frac{\pi}{n}), \text{ para } k = 0, ..., n,$$

porque $T_n(t'_k) = \cos(n \arccos(t'_k)) = \cos(k\pi) = (-1)^k$, ou analogamente $\tilde{T}_n(t'_k) = \frac{(-1)^k}{2^{n-1}}$. Como $||T_n||_\infty = 1$ (é definido por um coseno), ou ainda $||\tilde{T}_n||_\infty = \frac{1}{2^{n-1}}$, obtemos

$$T_n(t'_k) = (-1)^k ||T_n||_\infty \Leftrightarrow \tilde{T}_n(t'_k) = (-1)^k ||\tilde{T}_n||_\infty, \text{ para } k = 0, ..., n,$$

o resto $\tilde{T}_n = x^n - q_T$ verifica a propriedade de equioscilação em $n+1$ pontos, e pelo Teorema de Chebyshev, que minimizámos a distância a x^n, para qualquer $q_w \in \mathbb{P}_{n-1}$:

$$||(x-t_0)\cdots(x-t_{n-1})||_\infty = ||\tilde{T}_n||_\infty = ||x^n - q_T||_\infty \leq ||x^n - q_w||_\infty = ||W_n||_\infty.$$

□

É consequência deste resultado que a escolha dos nós de Chebyshev permite minimizar o erro de interpolação, no termo w_{n+1} que não depende da função. Assim, quando se escolhem nós de Chebyshev no intervalo $[-1,1]$ temos:

$$||f - p_n||_\infty \leq \frac{||f^{(n+1)}||_\infty}{(n+1)!}||\tilde{T}_{n+1}|| = \frac{||f^{(n+1)}||_\infty}{2^n(n+1)!}. \tag{1.13.4}$$

No caso geral podemos sempre levar a interpolação num outro intervalo $[a,b]$ para $[-1,1]$ considerando a transformação linear:

$$x(t) = a + \frac{b-a}{2}(t+1),$$

exportando os nós de Chebyshev para o intervalo $[a,b]$, considerando $x_k = x(t_k)$. Isto é equivalente a fazer interpolação para a função $\tilde{f}(t) = f(x(t))$, obtido aí o polinómio interpolador $\tilde{p}_n$ em $[-1,1]$ podemos regressar a $[a,b]$ fazendo $p_n(x) = \tilde{p}_n(t(x))$, usando a inversa $t(x) = 2\frac{x-a}{b-a} - 1$,

$$f(x_k) = f(x(t_k)) = \tilde{f}(t_k) = \tilde{p}_n(t_k) = \tilde{p}_n(t(x_k)) = p_n(x_k).$$

Observação 22. Dada uma função $f \in C[a,b]$ os nós de Chebyshev não garantem a melhor aproximação, essa seria obtida pela melhor aproximação uniforme. A minimização do resto W_n garante apenas que é a melhor escolha quando se incluem todas as funções analíticas (ou C^∞, por ser válida a fórmula do erro com quaisquer derivadas. A convergência será muito rápida quando as derivadas são limitadas (por exemplo, senos ou cosenos), sendo consequência directa da expressão (1.13.4) a ordem $O(\frac{1}{2^n n!})$.

No caso em que f é função analítica em $[-1,1]$ com uma expansão em série de potências $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(y)}{n!}(x-y)^n$ válida para $x, y \in [-1,1]$, isto implica em particular que $a_n = \frac{|f^{(n)}(y)|}{n!}$ converge para zero (considerando $|x-y| = 1$), e portanto $||f - p_n||_\infty = o(2^{-n})$. Neste caso a função será ainda muito regular e a convergência rápida. A convergência da interpolação com nós de Chebyshev pode ser ainda demonstrada quando $f \in C^2[-1,1]$, mas não para qualquer $f \in C[-1,1]$.

Esta propriedade optimal de aproximação com os nós de Chebyshev leva à sua escolha na iniciação do algoritmo de Remes.

1.13.3 Convergência da interpolação polinomial

Recordamos o teorema de Weierstrass, que garante a convergência da aproximação polinomial num intervalo compacto.

Teorema 15. *(Weierstrass). Seja $f \in C[a,b]$, dado $\varepsilon > 0$ existe $p_n \in \mathbb{P}_n$: $||f - p_n||_\infty < \varepsilon$.*

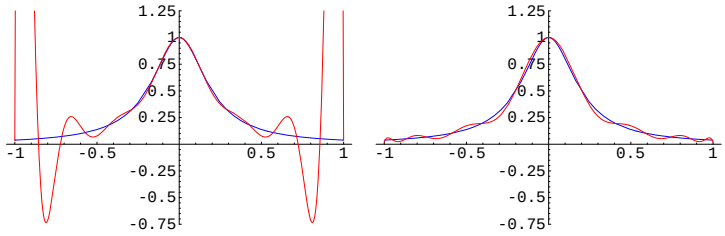

Figura 1.13.2: Exemplo de Runge: aproximação com nós igualmente espaçados (à esquerda), e com nós de Chebyshev (à direita).

Demonstração. A demonstração habitual pode ser encontrada em vários livros de Análise Real. Aqui apenas notamos que é possível uma demonstração construtiva usando em $[0,1]$ os polinómios de Bernstein

$$\beta_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k} \text{ e a aproximação } p_n(x) = \sum_{k=0}^{n} f(\tfrac{k}{n})\beta_{n,k}(x),$$

mostrando-se que a convergência será $O(\frac{1}{\sqrt{n}})$. □

A partir dos teoremas de Weierstrass e Chebyshev mostramos que é possível encontrar nós que permitam a convergência da interpolação.

Teorema 16. *(Marcinkiewicz). Dada uma função $f \in C[a,b]$, existe uma sucessão de nós de interpolação $\mathbf{x}^{(n)} = \{x_0^{(n)}, ..., x_n^{(n)}\}$, associados a um polinómio interpolador $p_n \in \mathbb{P}_n$ que convergirá uniformemente para f (quando $n \to \infty$).*

Demonstração. Pelo Teorema de Weierstrass, dado $\varepsilon > 0$, consideramos $q_n \in \mathbb{P}_n$ tal que $||f - q_n||_\infty < \varepsilon$. Esta não é necessariamente a melhor aproximação de f, por isso pelo Teorema de Chebyshev se consideramos a melhor aproximação $p_n \in \mathbb{P}_n$ verifica-se para $n+2$ pontos

$$f(z_k) - p_n(z_k) = (-1)^k ||f - p_n||, \text{ com } k = 0, ..., n+1.$$

Esta alternância de sinal mostra que $f - p_n$ tem $n+1$ zeros, $x_k \in [z_k, z_{k+1}]$ (para $k = 0, ..., n$), o que significa que $f(x_k) = p_n(x_k)$, e é estes pontos são os nós de interpolação. □

Observação 23. Em sinal oposto, é possível mostrar que dada uma sucessão de nós de interpolação existe uma função $f \in C[a,b]$ para a qual o polinómio interpolador não converge uniformemente. Um exemplo conhecido é o *exemplo de Runge.* Ao considerar nós igualmente espaçados no intervalo $[-1,1]$, os sucessivos polinómios interpoladores para a função

$$f(x) = \frac{1}{1+(5x)^2}$$

apresentam oscilações em que a amplitude se acentua com o aumento do grau do polinómio, não havendo convergência. Se considerarmos os nós de Chebyshev essas oscilações têm amplitude pequena e ajustam-se à curva (ver figura seguinte).

Capítulo 2

Determinação de vectores e valores próprios

2.1 Introdução

Seja E um espaço vectorial. Dizemos que $\lambda \in \mathbb{C}$ é um *valor próprio* de uma *aplicação linear* $A : E \to E$ se:

$$\exists v \in E, v \neq 0 : Av = \lambda v,$$

e a $v \in E$ chamamos *vector próprio* de A associado a λ.

Um mesmo valor próprio λ pode ter associados varios vectores próprios, que geram um subespaço vectorial, designado *subespaço próprio* S_λ associado a λ. Para qualquer $u \in S_\lambda$ é óbvio que $Au = \lambda u$.

Podemos considerar sempre uma base ortonormada em S_λ. Ao longo de cada elemento da base u a aplicação A fica invariante e comporta-se como uma aplicação linear a uma dimensão (i.e: como uma "recta" de inclinação λ). Quando um dos valores próprios é $\lambda = 0$, o subespaço próprio associado é o próprio núcleo (*kernel*) da aplicação A.

No caso geral, $S_\lambda = Ker(A - \lambda I)$.

Lembramos que se dois valores próprios λ, μ são distintos, então os vectores próprios associados a λ são independentes dos que estão associados a μ. Basta reparar que se $0 \neq v \in S_\lambda \cap S_\mu$, então $\lambda v = Av = \mu v \Rightarrow (\lambda - \mu)v = 0 \Rightarrow \lambda = \mu$.

Apenas nos interessa considerar o caso em que o espaço vectorial E tem dimensão finita N, que podemos identificar a um certo $\mathbb{R}^N$. No caso de operadores em dimensão infinita, o processo habitual é aproximar o operador linear por uma matriz (operador linear de dimensão finita) e aí determinar os valores próprios. Ou seja, 'formalmente' consideramos $A_n \to A$, e ao determinar $\lambda_n : A_n v_n = \lambda_n v_n$, obtemos uma sucessão tal que $\lambda_n \to \lambda$. *Note-se que isto é apenas possível quando o problema é regular e está demonstrada a dependência contínua.*

Começamos por rever algumas propriedades algébricas dos valores próprios em dimensão finita.

Como $S_\lambda = Ker(A - \lambda I) \neq \{0\}$, λ é valor próprio de A se e só se

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

o que define uma equação polinomial. Encontrando as raízes desta equação podemos obter a decomposição

$$p_A(\lambda) = (\lambda - \lambda_1)...(\lambda - \lambda_N)$$

em que $\lambda_1, ..., \lambda_N$ são os valores próprios de A. Podemos ter raízes múltiplas nessa equação e, nesse caso, dizemos que λ é um valor próprio com *multiplicidade algébrica* p se λ for uma raiz com multiplicidade p. Distinguimos multiplicidade algébrica de *multiplicidade geométrica*, que determina a dimensão do subespaço próprio S_λ. A multiplicidade geométrica nem sempre coincide com algébrica, para ilustrar esse facto, podemos dar como exemplo a matriz

$$\begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix}$$

onde $\lambda = 1$ é um valor próprio de multiplicidade algébrica 2, raiz da equação $(\lambda - 1)^2 = 0$, mas que tem apenas multiplicidade geométrica 1, no caso de $\varepsilon \neq 0$, porque tem apenas um vector próprio independente, $v = (1, 0)$, e que no caso $\varepsilon = 0$ tem multiplicidade geométrica 2.

Sabemos que a multiplicidade geométrica é sempre menor ou igual que a algébrica. No entanto, enquanto que a soma das multiplicidades algébricas é sempre igual à dimensão da matriz N, a soma das multiplicidades geométricas pode variar muito com pequenas variações das entradas da matriz... basta ver o exemplo anterior!

• Uma propriedade importante do polinómio característico é o Teorema de Hamilton-Cayley, que afirma

$$p_A(A) = 0,$$

ou seja, a potência A^N pode ser obtida pela combinação linear das potências de grau inferior $I, A, A^2, ..., A^{N-1}$.

• Outras propriedades importantes são aquelas que relacionam valores próprios de diferentes matrizes.

Proposição 11. *Se duas matrizes A, B são* semelhantes*, ou seja, se existe uma matriz P invertível*

$$B = P^{-1}AP$$

(P é a matriz mudança de base), então os polinómios característicos são iguais. Portanto, os valores próprios coincidem com a sua multiplicidade, e temos:
v é vector próprio (associado a um valor próprio λ) de B se e só se *Pv for vector próprio (associado ao mesmo valor próprio λ) de A.*

Demonstração. Basta reparar que

$$p_B(\lambda) = \det(\lambda I - B) = \det(\lambda P^{-1}P - P^{-1}AP) = \det(P^{-1}(\lambda I - A)P) = \det(\lambda I - A) = p_A(\lambda),$$

porque $\det(P^{-1}) = 1/\det(P)$. A segunda afirmação resulta de

$$Bv = P^{-1}APv = P^{-1}(\lambda Pv) = \lambda v. \square$$

□

Observação 24. (decomposição - formas de Schur e Jordan). Podemos mesmo obter uma decomposição em que os valores próprios são os elementos da diagonal de uma matriz. A decomposição na *forma normal de Schur* diz-nos que existe uma matriz unitária U tal que:

$$T = U^* A U$$

é uma matriz triangular superior, e portanto $p_A(\lambda) = (\lambda - t_{11})...(\lambda - t_{NN})$. No caso de A se tratar de uma matriz hermitiana, podemos obter T diagonal, ou seja

$$U^* A U = diag(\lambda_1, ..., \lambda_N)$$

em que vectores próprios associados a $\lambda_1, ..., \lambda_N$ formam uma base ortonormada do espaço.

No caso mais geral, apenas podemos obter a decomposição na forma canónica de Jordan:

$$P^{-1}AP = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{n_2}(\lambda_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{n_r}(\lambda_r) \end{bmatrix} ; \; J_{n_i}(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & 0 \dots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & \lambda_i \end{bmatrix}_{n_i \times n_i}$$

em que n_i corresponde à multiplicidade algébrica do valor próprio λ_i. O caso que nos interessará especialmente é aquele em a matriz A é diagonalizável, ou seja em que os blocos J_{n_i} têm apenas um elemento.

Observação 25. (matrizes Hermitianas). No caso em que A é uma matriz hermitiana, ou seja $A = A^*$, os valores próprios são reais e a forma normal de Schur assegura que existe uma matriz unitária U tal que é possível a *decomposição espectral*:

$$Ax = UDU^*x = \lambda_1(u_1 \cdot x)\, u_1 + ... + \lambda_N(u_N \cdot x)\, u_N$$

em que $u_1, ..., u_N$ são vectores próprios (ortonormais entre si) associados aos valores próprios $\lambda_1, ..., \lambda_N$. A matriz U é uma matriz de mudança de base formada por esses vectores próprios, enquanto que a matriz D é a matriz diagonal com os respectivos valores próprios. Trata-se de um caso em que a matriz é diagonalizável.

Não é difícil verificar que neste caso

$$A^k x = \lambda_1^k(u_1 \cdot x)\, u_1 + ... + \lambda_N^k(u_N \cdot x)\, u_N$$

Ora, isto permite definir, a partir da expansão em série de Taylor, funções analíticas (inteiras) em que a variável é uma matriz!

Assim, se $f(x) = \alpha_0 + \alpha_1 x + ... + \alpha_n x^n + ...$ obtemos

$$f(A)x = (\alpha_0 I + \alpha_1 A + ... + \alpha_n A^n + ...)x = f(\lambda_1)(u_1 \cdot x)\, u_1 + ... + f(\lambda_N)(u_N \cdot x)\, u_N$$

o que permite, por exemplo, definir a exponencial de uma matriz, a partir dos seus valores próprios:

$$e^A x = e^{\lambda_1}(u_1 \cdot x)\, u_1 + ... + e^{\lambda_N}(u_N \cdot x)\, u_N$$

Esta representação tem especial importância ao resolvermos um sistema de equações diferenciais $u'(t) = Au(t)$, pois nesse caso $u(t) = e^{At}u(0)$.

Reparando que os vectores próprios definem operadores de projecção $P_i x = (u_i \cdot x)\, u_i$ (assumimos $||u_i|| = 1$), podemos também escrever A na forma:

$$Ax = (\lambda_1 P_1 + ... + \lambda_N P_N)x,$$

e daqui $f(A)x = (f(\lambda_1)P_1 + ... + f(\lambda_N)P_N)x$. ((Como informação, observamos que esta representação pode ser generalizada a casos particulares de operadores lineares em espaços de dimensão infinita (operadores auto-adjuntos compactos...))

Observação 26. (valores singulares). Até aqui apenas falámos de valores próprios, noção que se aplica a uma matriz quadrada. Podemos introduzir uma noção adaptada a matrizes não quadradas. Sendo $B \in \mathbb{C}^M \times \mathbb{C}^N$ uma matriz.$M \times N$ com valores complexos (N pode ser igual a M) iremos considerar a matriz quadrada obtida pelo produto da matriz adjunta $B^* = \bar{B}^\top$ por ela própria. Dessa forma obtemos uma matriz quadrada $A = B^*B$ de dimensão $N \times N$, que será hermitiana e semi-definida positiva. Os valores próprios de A serão positivos ou nulos e é através desses valores que definimos valores singulares:

• Dizemos que $\mu \geq 0$ é *valor singular* de B se μ^2 for valor próprio de $A = B^*B$.

(i) Note-se que os núcleos de A e B coincidem, ie. $Ker(A) = Ker(B)$, porque se $Bx = 0$ é óbvio que $Ax = B^*Bx = 0$; e reciprocamente, se $Ax = 0$ temos

$$x^*Ax = x^*B^*Bx = (Bx)^*Bx = ||Bx||_2^2 = 0,$$

o que implica $Bx = 0$.□

Repare-se que isto significa que o número de valores singulares positivos (contando com a multiplicidade geométrica) será igual à característica da matriz B, e como é claro, os restantes valores singulares serão nulos.

(ii) A norma euclidiana de uma matriz não quadrada é dada por $||B||_2 = \sqrt{\rho(B^*B)}$, e assim concluímos que a norma euclidiana de uma matriz será igual ao maior valor singular.

(iii) A principal propriedade, é a decomposição em valores singulares (o análogo da decomposição espectral), que garante a existência de N vectores ortonormais $u_1, u_2, ...u_N \in \mathbb{C}^N$, e de M vectores ortonormais $v_1, ..., v_M \in \mathbb{C}^M$:

$$Bx = \mu_1(u_1 \cdot x)v_1 + ... + \mu_r(u_r \cdot x)v_r,$$

em que r é a característica da matriz B (os μ_i restantes seriam nulos). Note-se que $Bu_k = \mu_k v_k$, e que $B^*v_k = \mu_k u_k$. Quando um sistema da forma $Bx = y$ tem solução (ou soluções), então

$$x = \frac{1}{\mu_1}(y \cdot v_1)u_1 + ... + \frac{1}{\mu_r}(y \cdot v_r)u_r.$$

• Os valores singulares têm especial interesse na aproximação de sistemas mal condicionados. Por exemplo, estão relacionados com o problema de aproximação de dados pelo método dos mínimos quadrados, que iremos abordar no último capítulo.(Para maior detalhe, consultar p.ex.[9].)

2.1.1 Valores próprios e o polinómio característico

Já vimos que sendo A uma matriz $N \times N$, encontrar os valores próprios de A é encontrar as raizes $\lambda \in \mathbf{C}$:

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

em que $p_A(\lambda)$ é o polinómio característico de grau N, e isto corresponde a resolver uma equação polinomial.

• Encarando o determinante como forma multilinear, temos

$$p_A(\lambda) = \det(\lambda I - A) = \det(\lambda e^{(1)} - a^{(1)}, ..., \lambda e^{(N)} - a^{(N)})$$

em que $a^{(k)}$ são as linhas da matriz A e $e^{(k)}$ as linhas da matriz identidade (i.e: o vector k da base canónica).

Ora, se desenvolvermos $\det(\lambda e^{(1)} - a^{(1)}, ..., \lambda e^{(N)} - a^{(N)})$, obtemos

$$p_A(\lambda) = \lambda^N \det(e^{(1)}, ..., e^{(N)}) + ... + (-1)^N \det(a^{(1)}, ..., a^{(N)})$$

e reparamos que no termo constante $\det(a^{(1)}, ..., a^{(N)}) = \det(A)$. Por outro lado, como

$$p_A(\lambda) = (-1)^N \lambda_1 ... \lambda_N + ... - (\lambda_1 + ... + \lambda_N)\lambda^{N-1} + \lambda^N,$$

isto implica imediatamente que os termos constantes têm que ser iguais, ou seja $\det(A) = \lambda_1 ... \lambda_N$. Da mesma forma, podemos obter

$$\lambda_1 + ... + \lambda_N = \det(a^{(1)}, e^{(2)}, ..., e^{(N)}) + ... + \det(e^{(1)}, ..., e^{(N-1)}, a^{(N)}) = a_{11} + ... + a_{NN} = \mathrm{tr}(A),$$

ou seja, a soma dos valores próprios é igual a $\mathrm{tr}(A)$, o traço da matriz A.

Portanto, podemos concluir as relações

$$\begin{aligned} \mathrm{tr}(A) &= \lambda_1 + ... + \lambda_N, \\ \det(A) &= \lambda_1 ... \lambda_N. \end{aligned}$$

• Escrevendo o polinómio característico na forma

$$p_A(\lambda) = \alpha_1 + \alpha_2 \lambda + ... + \alpha_N \lambda^{N-1} + \lambda^N,$$

acabamos de concluir que os valores dos coeficientes α_k são obtidos pelo cálculo de determinantes, o que significa que há uma dependência contínua dos α_k face aos valores dos elementos da matriz. Sendo os valores próprios as raízes $\lambda_1, ..., \lambda_N$, do polinómio característico, resta saber se há uma dependência contínua das raízes face à variação dos coeficientes, para concluir que os valores próprios dependem de forma contínua das entradas da matriz. De facto isso verifica-se[1]:

Lema 2.1. *Seja $p(x) = \alpha_1 + \alpha_2 x + ... + \alpha_N x^{N-1} + x^N$, com $\alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{C}^N$, e seja $z = (z_1, ..., z_N) \in \mathbb{C}^N$ um vector que tem as N raízes de p. Consideremos agora uma perturbação do polinómio, $\tilde{p}$, com coeficientes em $\tilde{\alpha}$ e raízes em $\tilde{z}$ (ordenadas convenientemente, de forma a que $\tilde{z}_k$ seja a componente mais próxima de z_k). Então,*

$$\tilde{\alpha} \to \alpha \Rightarrow \tilde{z} \to z,$$

ou seja, há uma dependência contínua das raízes face aos coeficientes.

[1]Uma demonstração alternativa, usando o teorema de Rouché pode ser vista em [11].

Demonstração. Usando o teorema fundamental da álgebra,

$$p(x) = (x - z_1)...(x - z_N) = \alpha_1 + \alpha_2 x + ... + \alpha_N x^{N-1} + x^N$$

e a igualdade entre os polinómios permite escrever um sistema com N equações relacionando cada α_k como função contínua dos valores $z_1, ..., z_N$. Por exemplo, $\alpha_1 = (-1)^N z_1...z_N$, ou $\alpha_N = -(z_1 + ... + z_N)$. Ou seja, com notação vectorial, podemos escrever

$$\alpha = \mathcal{P}(z),$$

em que $\mathcal{P}$ é uma função vectorial de $\mathbb{C}^N$ em $\mathbb{C}^N$,

$$\mathcal{P}(z_1, ..., z_N) = ((-1)^N z_1...z_N, \cdots\cdots, -(z_1 + ... + z_N)).$$

Esta função é claramente contínua e a menos de permutação na lista $(z_1, ..., z_N)$ também é injectiva. Aliás, considerando a relação de equivalência $w \triangleq z$ quando $\sigma(z) = w$, onde σ é uma permutação dos coeficientes de z, podemos definir a bijectividade de

$$\begin{array}{rccc} \mathcal{P}: & \mathbb{C}^N/\triangleq & \longrightarrow & \mathbb{C}^N \\ & \dot{z} & \longmapsto & \alpha \end{array}$$

pelo teorema fundamental da álgebra. Como se trata de uma aplicação contínua e bijectiva a sua inversa é também contínua (i.e. trata-se de um homeomorfismo). Concluímos assim que com uma ordenação apropriada das raízes é possível obter o resultado. □

Corolário 2.2. *Os valores próprios são funções contínuas dos elementos da matriz.*

2.2 Teorema de Gerschgorin

Podemos começar por retirar alguma informação acerca da localização dos valores próprios usando o teorema do Ponto Fixo. Com efeito, reparamos que se $\lambda \neq 0$, podemos escrever

$$Av = \lambda v \Leftrightarrow v = \frac{A}{\lambda} v$$

e se $||\frac{A}{\lambda}|| < 1$, temos uma contracção, logo a única solução será $v = 0$. Assim, para termos soluções não nulas, e consequentemente valores próprios, é necessário que $\lambda \leq ||A||$ (o caso $\lambda = 0$ é trivial). Isto reflecte a propriedade que já tinhamos visto acerca do raio espectral

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

No caso de A ser uma matriz hermitiana é também possí vel obter uma minoração de forma simples,

$$x^*.Ax = x^*U^*DUx \leq x^*U^*(\lambda_{\max} I)Ux = \lambda_{\max} x^*.x = \lambda_{\max}||x||_2^2$$

o que significa que

$$\rho(A) \geq \max_{||u||_2=1} ||u^*Au||_2$$

Exercício 27. (Quociente de Rayleigh). Mostre que se A for hermitiana então o maior valor próprio verifica

$$\lambda_{\max} = \max_{x \neq 0} \frac{x^* A x}{x^* x}.$$

No entanto, estes resultados podem ser melhorados. O próximo teorema permite obter informações *a priori*, mais concretas, acerca da localização dos valores próprios, através dos elementos da matriz.

Teorema 17. *(Gerschgorin).*

a) *Um valor próprio λ de uma matriz A verifica uma das seguintes desigualdades:*

$$|a_{kk} - \lambda| \leq \sum_{j=1, j\neq i}^{N} |a_{kj}| = r_k, \ (k = 1, .., N)$$

o que significa que os valores próprios pertencem a bolas fechadas com centro na diagonal e raio r_k, ou seja, $\lambda \in \bigcup_{k=1}^{N} \bar{B}(a_{kk}, r_k)$.

b) *Para além disso, se a reunião de m bolas forma uma componente conexa, haverá exactamente m valores próprios nessa componente (consideramos $m \geq 1$).*

c) *O mesmo argumento é válido se considerarmos linhas ao invés de colunas!*

Demonstração. **a)** Um vector próprio v associado ao valor próprio λ verifica

$$[Av]_i = \sum_{j=1}^{N} a_{ij} v_j = \lambda v_i \Leftrightarrow \sum_{j=1, j\neq i}^{N} a_{ij} v_j + a_{ii} v_i = \lambda v_i$$

e daqui obtemos

$$\sum_{j=1, j\neq i}^{N} a_{ij} v_j = (\lambda - a_{ii}) v_i$$

e portanto

$$|\lambda - a_{ii}| \, |v_i| \leq \sum_{j=1, j\neq i}^{N} |a_{ij}| \, |v_j|.$$

Considerando agora o índice k para o qual $|v_k| = \max_{i=1,\ldots,N} |v_i| = ||v||_\infty$ obtemos

$$|\lambda - a_{kk}| \, ||v||_\infty \leq \sum_{j=1, j\neq i}^{N} |a_{kj}| \, |v_j| \leq \sum_{j=1, j\neq i}^{N} |a_{kj}| \, ||v||_\infty$$

e assim, dividindo por $||v||_\infty \neq 0$ (porque é um valor próprio), obtemos o resultado.

b) Para mostrar a segunda parte, usamos argumentos analíticos. Consideramos um "*segmento formado por matrizes*"

$$A_t = D + t(A - D), \ (t \in [0, 1]),$$

que começa na matriz $D = diag(a_{11}, ..., a_{NN})$, quando $t = 0$, e termina em A, quando $t = 1$.

Essas matrizes A_t têm valores próprios associados $\Lambda_1(t), ..., \Lambda_N(t)$ que vão definir linhas contínuas (caminhos) no plano complexo. As matrizes A_t têm a mesma diagonal que a matriz A, e as outras entradas estão multiplicadas por t. Temos $\Lambda_i(0) = a_{kk}$, e também $\Lambda_i(1) = \lambda_i$.

Pelo que vimos em a), concluímos que os valores próprios $\lambda_i(t)$ pertencem à reunião das bolas, $\bigcup_k \bar{B}(a_{kk}, t\, r_k) \subseteq \bigcup_k \bar{B}(a_{kk},\, r_k)$.

Pelo corolário do lema anterior as funções $\Lambda_k : [0,1] \to \mathbf{C}$ são contínuas, consequentemente transformam conexos em conexos, logo $\Lambda_k([0,1])$ é conexo, e por outro lado sabemos que tem que pertencer a $\bigcup_k \bar{B}(a_{kk},\, r_k)$. Isto implica que tem que pertencer a uma componente conexa dessa reunião. Assim, $\Lambda_k(1) = \lambda_k$ pertencem exactamente à componente conexa que contém $\Lambda_k(0) = a_{kk}$ e o resultado está provado.

c) Basta reparar que os valores próprios de A^T coincidem com os de A, porque $\det(A - \lambda I) = \det((A - \lambda I)^T) = \det(A^T - \lambda I)$. □

Observação 27. Quando a matriz é real, o polinómio característico tem coeficientes reais. Assim, nesse caso, os valores próprios complexos aparecem como pares conjugados. Logo, caso se conclua que há apenas um valor próprio numa componente conexa, ele terá que obrigatoriamente ser real. Isso acontece frequentemente quando a componente conexa é uma bola disjunta das restantes.

Exemplo 9. Consideremos um primeiro caso em que a matriz é

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

Na primeira figura são representadas as três bolas $\bar{B}(-4,2), \bar{B}(1,1), \bar{B}(3,4)$, bem como a localização dos valores próprios, no plano complexo. Estas três bolas são as que se obtêm fazendo uma análise por linhas com o Teorema de Gerschgorin. Como a primeira bola é disjunta das outras duas conclui-se que há um valor próprio em $\bar{B}(-4,2)$ (que é obrigatoriamente real, pela observação anterior) e dois valores próprios na reunião das outras duas, $\bar{B}(1,1) \cup \bar{B}(3,4) = \bar{B}(3,4)$, o que é confirmado na figura. Repare-se que na bola $\bar{B}(1,1)$ não há nenhum valor próprio.
Na segunda figura é feita uma análise por colunas. Nesse caso, obtemos $\bar{B}(-4,0), \bar{B}(1,5), \bar{B}(3,2)$. No entanto, reparamos que $\bar{B}(4,0)$ não é mais que o ponto $z = -4$, que é obviamente um valor próprio, pois basta considerar $v = (1,0,0)$, para termos $Av = -4v$. Isso acontece sempre nos casos em a matriz tem uma coluna (ou linha) em que o elemento não nulo está na diagonal. Assim, a análise por colunas permite concluir que há um valor próprio $\lambda_1 = -4$ e que os dois restantes, λ_2, λ_3 estão em $B(1,5)$. Juntando esta informação com a obtida por linhas, podemos concluir-se que $\lambda_1 = -4$, e que $\lambda_2, \lambda_3 \in \bar{B}(3,4)$, o que é confirmado nas figuras. Note-se que, neste caso, uma análise através da regra de Laplace permitiria resultados imediatos.

Exemplo 10. Consideremos agora um outro exemplo em que a matriz é

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

Neste caso, fazendo uma análise por linhas, obtemos as bolas $B_1 = \bar{B}(-4, 2), B_2 = \bar{B}(-4, 2\sqrt{2}), B_3 = \bar{B}(2 + 4i, 2), B_4 = \bar{B}(4, 4)$. A reunião tem duas componentes conexas, uma formada por $B_1 \cup B_2 = B_2$, que é disjunta da outra formada por $B_3 \cup B_4$. Pelo teorema concluímos que há dois valores próprios em cada uma destas componentes. Façamos agora uma análise por colunas. Nesse caso temos $B'_1 = \bar{B}(-4, 2\sqrt{2}), B'_2 = \bar{B}(-4, 4), B'_3 = \bar{B}(2+4i, 2), B'_4 = \bar{B}(4, -2)$. Aqui há três componentes conexas, uma formada pela reunião $B'_1 \cup B'_2 = B'_2$ e duas outras formadas pelas bolas B'_3 e B'_4. Podemos assim concluir que há um valor próprio $\lambda_3 \in B'_3$, outro $\lambda_4 \in B'_4$ e que os outros dois $\lambda_1, \lambda_2 \in \bar{B}(-4, 4) = B'_1 \cup B'_2$. Intersectando esta informação com a informação obtida por linhas, podemos mesmo concluir que $\lambda_1, \lambda_2 \in \bar{B}(-4, 2\sqrt{2})$.

Como curiosidade, apresentamos o gráfico da evolução dos valores próprios da matriz D para a matriz A, através do segmento de matrizes formadas por $A_t = D + t(A - D)$, que foi utilizado na demonstração do teorema. Consideremos a matriz

$$A = \begin{bmatrix} -1 & 2 & -2 \\ -4 & 3 & -2 \\ 5 & 5 & 7 + \alpha i \end{bmatrix}$$

Podemos ver, no primeiro gráfico, para $\alpha = 1$, a evolução da posição dos valores próprios desde a diagonal D, que tem valores próprios $\Lambda_1(0) = -1, \Lambda_2(0) = 3, \Lambda_3(0) = 7 + i$, a que correspondem os pontos $x_1 = (-1, 0), x_2 = (3, 0), x_3 = (7, 1)$, até aos valores próprios da matriz final A. Repare-se na evolução do valor próprio Λ_1, que começando no centro x_1 acaba por sair da bola $\bar{B}(-1, 4)$. Isto retrata bem que pode haver bolas (definidas pela análise de linhas ou colunas) onde não há valores próprios, apenas podemos garantir a existência de valores próprios na componente conexa. No segundo gráfico mostramos a mesma evolução, mas para $\alpha = 0$. Repare-se que para um certo t os valores das trajectórias de Λ_1 e Λ_2 coincidem, o que corresponde a um valor próprio com multiplicidade 2. A partir desse t as trajectórias tomam sempre valores complexos conjugados, como é característico das matrizes reais; note-se que é indiferente dizer que a trajectória de Λ_1 é continuada para o valor próprio com parte imaginária positiva ou negativa, o que interessa é que a trajectória é contínua.

Exemplo 11. Terminamos com um exemplo em que os valores próprios estão na fronteira das bolas,

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

Quer seja feita uma análise por linhas ou colunas o resultado é o mesmo. Se $\alpha = 2$ (primeira figura) conclui-se que deverá haver dois valores próprios na bola $\bar{B}(-2, 1)$ e dois valores próprios em $\bar{B}(1, 1) \cup \bar{B}(2, 1)$. Neste caso, como se pode observar, os valores próprios estão exactamente sobre a fronteira das componentes conexas, o que ilustra que a estimativa não poderia dar raios inferiores. Se $\alpha = 1$ (segunda figura) temos uma situação em que há uma translação da situação anterior para uma situação de contacto num único ponto. A reunião das bolas fechadas tem apenas uma componente conexa, que é ela própria. No entanto, nesta situação, em que a reunião das bolas abertas tem duas componentes conexas (o que significa, na prática, que há

apenas intersecção num ponto), seguindo a demonstração do teorema, podemos concluir que a trajectória de um valor próprio $\Lambda_1(t)$ pertenceria sempre à bola $\bar{B}(-2,t)$, para $t<1$, e portanto (sendo contínua) no instante $t=1$ apenas poderia estar na fronteira. É o que se passa neste caso. Da mesma forma, a conclusão do teorema mantém-se válida num caso geral, quando há intersecção num único ponto. Basta ver o que se passa nesse ponto, analisando se se trata ou não de um valor próprio (e qual a sua multiplicidade) e concluir para as componentes.

2.3 Método das Potências

Estando interessados em encontrar os valores próprios de uma matriz podemos pensar imediatamente num processo – encontrar as raízes do polinómio característico. Para esse efeito podemos usar qualquer método que vimos, como sejam os métodos de Bernoulli, de Newton, da Secante, ou de Steffensen. O primeiro apenas nos dá a maior raiz real, o segundo necessita do cálculo da derivada de um determinante, e os outros dois necessitam do cálculo em cada iterada de um determinante, o que é bastante moroso, para além de serem precisas boas aproximações iniciais...

Poderíamos ainda simplificar o processo determinando exactamente o polinómio através de interpolação polinomial usando apenas o cálculo de $N+1$ determinantes, o que reduziria o número de cálculos... mas mesmo este processo pode ser demasiado moroso.

Vamos começar por ver um processo extremamente simples, o *método das potências*, de von Mises (1929), que no entanto funciona apenas em circunstâncias particulares! É o método mais simples e pode ser encarado como um método de ponto fixo, em que se procura um vector próprio u de norma 1 (associado a um valor próprio $\lambda \neq 0$) no conjunto $S=\{x\in\mathbb{R}^N : ||x||=1\}$.

Escrevendo

$$Au=\lambda u \Leftrightarrow u=\frac{Au}{\lambda},$$

e reparando que $||Au||=||\lambda u||=|\lambda|$, obtemos

$$u=\frac{|\lambda|}{\lambda}\frac{Au}{||Au||}.$$

O método iterativo poderia ficar

$$u^{(n+1)}=\frac{|\lambda|}{\lambda}\frac{Au^{(n)}}{||Au^{(n)}||},$$

mas isso implicava um conhecimento a priori do argumento $\theta_\lambda \in [0,2\pi[$ do valor próprio, caso λ fosse um número complexo, pois $\frac{|\lambda|}{\lambda}=e^{-\theta_\lambda i}$.

No entanto, no caso de se tratar de um valor próprio real $\frac{|\lambda|}{\lambda}=\pm 1$, e a situação é mais fácil de resolver... sob certas condições. Devemos começar por reparar que, havendo sempre mais que um valor próprio, a convergência de uma tal sucessão não estaria *a priori* bem determinada. É preciso impor restrições.

• Admitiremos assim que a matriz é diagonalizável e que um dos valores próprios é dominante, ou seja,

$$|\lambda_1|>\max_{i=2,\ldots,N}|\lambda_i|,$$

e também que esse valor próprio dominante, λ_1, é real.

Observação 28. A condição de ser diagonalizável pode ser verificada imediatamente se a matriz for hermitiana. Nesse caso, a matriz tem valores próprios reais e basta mostrar que há um valor próprio maior que os restantes. Isso pode ser provado pelo teorema de Gerschgorin, mas, na prática, esta situação é quase sempre verificada, a menos que haja valores próprios com multiplicidade superior a 1, ou em que há um outro valor próprio dominante simétrico.

• Verificadas estas condições, podemos estabelecer o *método das potências*[2]**:**

$$\begin{cases} u^{(0)} : ||u^{(0)}|| = 1, \\ u^{(n+1)} = \sigma_n \frac{Au^{(n)}}{||Au^{(n)}||}, \end{cases}$$

em que $\sigma_n = \pm 1$ é o sinal da componente com maior módulo do vector $Au^{(n)}$. Todas as iteradas são vectores unitários para a norma considerada.

Habitualmente, considera-se a normalização usando a norma do máximo (também é frequente usar a norma euclidiana), que será a adoptada no que se segue.

A iterada inicial $u^{(0)}$ é normalmente um valor aleatório, de forma a que na base ortonormada $v_1, ..., v_N$ formada pelos vectores próprios tenha componente não nula relativamente ao vector próprio associado ao valor próprio dominante. Ou seja, exigimos que

$$u^{(0)} = \alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N$$

com $\alpha \neq 0$.

Proposição 12. *Seja A uma matriz diagonalizável (em particular, hermitiana) com um valor próprio dominante λ_1 real: $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_N|$. Se a iterada inicial $u^{(0)}$ tiver a componente $\alpha \neq 0$, o método das potências converge para v_1, e uma aproximação para o valor próprio dominante é*

$$\lambda^{(n)} = \frac{[Au^{(n)}]_i}{u_i^{(n)}}$$

[2]Uma outra possibilidade é considerar simplesmente

$$\begin{cases} x^{(0)} \in \mathbf{R}^d, \\ x^{(n+1)} = Ax^{(n)}, \end{cases}$$

e apenas normalizar no final dividindo por $||Ax^{(n)}||$ já que a divisão sucessiva por $||Ax^{(n)}||$ tem apenas como objectivo evitar a divergência, mantendo sempre o vector com norma 1.

É aliás fácil verificar que se μ_i são escalares,

$$\mu_n A(...(\mu_1 A(\mu_0 x^{(0)})...) = \mu_n ... \mu_0 A^n x^{(0)}$$

e portanto a normalização no final leva ao mesmo resultado. Como podemos ver, se $u^{(n)} = \frac{x^{(n)}}{||x^{(n)}||}$,

$$\frac{Au^{(n)}}{||Au^{(n)}||} = \frac{||x^{(n)}||}{||Ax^{(n)}||} A(\frac{x^{(n)}}{||x^{(n)}||}) = \frac{Ax^{(n)}}{||Ax^{(n)}||}.$$

No entanto, se não for efectuada a normalização, as iteradas podem tomar valores que crescem muito rapidamente e surgem problemas de cálculo e precisão.

para qualquer índice i (desde que $u_i^{(n)} \neq 0$*), sendo normalmente escolhido o índice com componente igual a 1. Estabelecemos, também, a estimativa de erro*[3]

$$||v_1 - u^{(n)}||_\infty \leq C \left|\frac{\lambda_2}{\lambda_1}\right|^n,$$

(onde a constante $C > 0$ *depende directamente de* $\frac{1}{|\alpha|}$*) e também*

$$|\lambda_1 - \lambda^{(n)}| \leq C' \left|\frac{\lambda_2}{\lambda_1}\right|^n.$$

Demonstração. [4] Seja $v_1, ..., v_N$ a base de valores próprios unitária (i.e. $||v_k||_\infty = 1$), associada aos valores próprios $\lambda_1, ..., \lambda_N$.

Como supomos que $u^{(0)} = \alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N$, com $\alpha \neq 0$, vamos considerar o subconjunto fechado de $\mathbb{R}^N$

$$S_\alpha = \{\alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N : \alpha_2, ..., \alpha_N \in \mathbb{R}\},$$

que é um subespaço afim. Para qualquer $x \in S_\alpha$, podemos considerar a decomposição $x = \alpha v_1 + \tilde{x}$.

Como podemos escrever $x = x'_1 v_1 + x'_2 v_2 + ... + x'_N v_N$, vamos considerar a norma em $\mathbb{R}^N$

$$||x||_1^* = |x'_1| + ... + |x'_N|,$$

que é equivalente a qualquer outra norma de $\mathbb{R}^N$. Por exemplo, temos $c_1||.||_\infty \leq ||.||_1^* \leq c_2||.||_\infty$, com $c_1 = 1, c_2 = N||P||_\infty$, em que P é a matriz de mudança para a base canónica, $Pv_k = e_k$, porque

$$\begin{aligned} ||x||_\infty &= ||x'_1 v_1 + ...x'_N v_N||_\infty \leq |x'_1| + ... + |x'_N| = ||x||_1^* \\ ||x||_1^* &= |x'_1| + ... + |x'_N| \leq N \max_{k=1,...,N} |x'_k| = N||Px||_\infty \leq N||P||_\infty \, ||x||_\infty, \end{aligned}$$

já que $x'_1 e_1 + ... + x'_N e_N = P(x'_1 v_1 + ... + x'_N v_N)$, e portanto $x'_k = [Px]_k$.

Note-se ainda que $||v_k||_1^* = 1$.

• Define-se a aplicação $T : S_\alpha \to S_\alpha$

$$Tx = \frac{Ax}{\lambda_1}$$

(note-se que $\lambda_1 \neq 0$, senão todos os valores próprios seriam nulos).

Note-se que $T(S_\alpha) = S_\alpha$, pois quando $x \in S_\alpha$ temos $x = \alpha v_1 + x'_2 v_2 + ... + x'_N v_N = \alpha v_1 + \tilde{x}$, e é fácil ver que

$$Tx = \frac{A}{\lambda_1}(\alpha v_1 + \tilde{x}) = \frac{1}{\lambda_1}(\alpha A v_1 + A\tilde{x}) = \alpha \frac{\lambda_1}{\lambda_1} v_1 + \frac{1}{\lambda_1} A\tilde{x} \in S_\alpha.$$

[3] Esta estimativa de erro é válida em qualquer norma, já que como todas as normas são equivalentes,

$$||v_1 - u^{(n)}|| \leq c_2 ||v_1 - u^{(n)}||_\infty,$$

e trata-se apenas de ajustar a constante.

[4] A demonstração clássica é eventualmente bastante mais simples (e.g. [1]), mas fornece menos informação.

porque $A\tilde{x} = A(x'_2 v_2 + ... + x'_N v_N) = x'_2\lambda_2 v_2 + ... + x'_N \lambda_N v_N$.

• Vejamos agora que se trata de uma contracção:

$$\begin{aligned} ||Tx - Ty||_1^* &= ||\tfrac{1}{\lambda_1} A\tilde{x} - \tfrac{1}{\lambda_1} A\tilde{y}||_\infty = \tfrac{1}{|\lambda_1|} ||(x'_2 - y'_2)\lambda_2 v_2 + ... + (x'_N - y'_N)\lambda_N v_N||_1^* = \\ &= \left|\tfrac{\lambda_2}{\lambda_1}\right| ||(x'_2 - y'_2)v_2 + (x'_3 - y'_3)(\tfrac{\lambda_3}{\lambda_2})v_3 ... + (x'_N - y'_N)(\tfrac{\lambda_N}{\lambda_2})v_N||_1^* = \\ &= \left|\tfrac{\lambda_2}{\lambda_1}\right| \left(|x'_2 - y'_2| + |x'_3 - y'_3|\,|\tfrac{\lambda_3}{\lambda_2}| + ... + |x'_N - y'_N|\,|\tfrac{\lambda_N}{\lambda_2}| \right) \tfrac{1}{N} \leq \left|\tfrac{\lambda_2}{\lambda_1}\right| \, ||x - y||_1^*. \end{aligned}$$

Portanto, pelo Teorema do Ponto Fixo existe um único $z \in S_\alpha : Tz = z$, esse valor é $z = \alpha v_1$ e temos

$$||z - T^n u^{(0)}||_1^* \leq \left|\frac{\lambda_2}{\lambda_1}\right|^n \, ||z - u^{(0)}||_1^*.$$

Como a norma $||.||_1^*$ é equivalente a $||.||_\infty$, temos

$$||z - T^n u^{(0)}||_\infty \leq \frac{c_2}{c_1} \left|\frac{\lambda_2}{\lambda_1}\right|^n \, ||z - u^{(0)}||_\infty,$$

em que $c_2 \geq c_1 > 0$ são as constantes que determinam a equivalência entre as normas.

Por outro lado, temos $||\frac{\mathbf{a}}{||\mathbf{a}||} - \frac{\mathbf{b}}{||\mathbf{b}||}|| \leq 2||\frac{\mathbf{a}}{||\mathbf{a}||} - \beta\mathbf{b}||, \forall \beta \geq 0$ (exercício[5]). Aplicando esta desigualdade com $\mathbf{a} = z, \mathbf{b} = T^n u^{(0)}, \beta = \frac{1}{||z||_\infty} = \frac{1}{|\alpha|}$,obtemos

$$||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty}||_\infty \leq 2||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||z||_\infty}||_\infty,$$

logo,

$$||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty}||_\infty \leq \frac{2}{|\alpha|}\frac{c_2}{c_1} \left|\frac{\lambda_2}{\lambda_1}\right|^n \, ||z - u^{(0)}||_\infty.$$

Basta agora reparar que $T^n u^{(0)} = \frac{A^n u^{(0)}}{\lambda_1^n}$, e portanto

$$\frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty} = \frac{A^n u^{(0)}}{\lambda_1^n}||\frac{\lambda_1^n}{A^n u^{(0)}}||_\infty = \left(\frac{|\lambda_1|}{\lambda_1}\right)^n \frac{A^n u^{(0)}}{||A^n u^{(0)}||_\infty} = (\pm 1)^n \sigma_1 ... \sigma_n \frac{A u^{(n)}}{||A u^{(n)}||_\infty} = \pm u^{(n+1)},$$

e por outro lado $\frac{z}{||z||_\infty} = \frac{\alpha}{|\alpha|} v_1 = \pm v_1$ (os sinais coincidem devido à construção), pelo que

$$||v_1 - u^{(n+1)}||_\infty \leq 2\frac{c_2}{c_1} \left|\frac{\lambda_2}{\lambda_1}\right|^n \, ||v_1 - \frac{u^{(0)}}{|\alpha|}||_\infty.$$

---

[5]Quando $||a|| = 1$, temos a desigualdade

$$||a - \frac{b}{||b||}|| \leq 2||a - b||,$$

porque

$$||a - \frac{b}{||b||}|| \leq ||a - b|| + ||b - \frac{b}{||b||}||$$

e como

$$||b - \frac{b}{||b||}|| = \frac{1}{||b||}||b\,(||b|| - 1)\,|| = |\,||b|| - 1|\,,$$

quando $||a|| = 1$ temos

$$||a - \frac{b}{||b||}|| \leq ||a - b|| + |\,||b|| - ||a||\,| \leq 2||a - b||.$$

Basta agora considerar $a = \frac{\mathbf{a}}{||\mathbf{a}||}$ e $b = \beta\mathbf{b}$.

Como $||v_1 - \frac{u^{(0)}}{|\alpha|}||_\infty = ||\frac{\alpha_2}{|\alpha|}v_2 + ... + \frac{\alpha_N}{|\alpha|}v_N||_\infty \leq \frac{1}{|\alpha|}||u^{(0)}||_1^*$, e como $||u^{(0)}||_1^* \leq c_2||u^{(0)}||_\infty = c_2$, obtemos

$$||v_1 - u^{(n+1)}||_\infty \leq \frac{K}{|\alpha|}\left|\frac{\lambda_2}{\lambda_1}\right|^n$$

em que a constante $K = \frac{2c_2^2}{c_1}$ só poderia ser calculada se fossem conhecidos os valores e vectores próprios. Repare-se que se $\alpha \to 0$ o majorante tende para infinito[6], o que coloca em evidência que a componente segundo o vector próprio dominante não pode ser nula.

Finalmente, como $\lambda^{(n)} = [Au^{(n)}]_j$, em que j é o índice correspondente a $u_j^{(n)} = 1$, e temos $\lambda_1 = [Av_1]_j$, obtemos

$$|\lambda_1 - \lambda^{(n)}| = |[Av_1]_j - [Au^{(n)}]_j| \leq ||Av_1 - Au^{(n)}||_\infty \leq ||A||_\infty||v_1 - u^{(n)}||_\infty.$$

□

Observação 29. No caso de ser utilizada a norma $||.||_2$ para a normalização, a estimativa obtida pode ser explícita no caso de matrizes simétricas, já que a matriz mudança de base P será unitária.

Observação 30. A restrição à escolha do vector inicial $u^{(0)}$ não é significativa. Dificilmente acontece a situação improvável de ser exactamente um vector cuja componente segundo v_1 fosse nula. Se usarmos números com todos os decimais (p. ex. gerados aleatoriamente), é praticamente uma situação inexistente, já que os próprios erros de arredondamento fazem aparecer uma pequena componente.

No entanto, em casos mais simples, quando usamos números inteiros, podemos cair facilmente em situações em que isso acontece. Apresentamos um exemplo simples. Consideremos a matriz

$$M = \begin{bmatrix} 5 & -7 & 7 \\ 6 & -9 & 8 \\ 6 & -7 & 6 \end{bmatrix}$$

Se começarmos com $u^{(0)} = (1, 1, 0)$, obtemos $u^{(n)} \to v = (0.5, 1, 0.5)$. No entanto, o limite obtido não é o vector próprio associado ao valor próprio dominante! A matriz tem valores próprios $\lambda_1 = 5, \lambda_2 = -2, \lambda_3 = -1$, e é fácil ver que $Mv = -2v$, portanto v é o vector próprio associado ao valor próprio λ_2 e não a λ_1. A razão é simples, como os vectores próprios são $v_1 = (1, 1, 1), v_2 = (0.5, 1, 0.5), v_3 = (0, 1, 1)$, temos $u^{(0)} = 2v_2 - v_3$, ou seja, a componente segundo v_1 é $\alpha = 0$. Uma simples perturbação, usando $u^{(0)} = (1, 1, \varepsilon)$ com $\varepsilon \neq 0$, mesmo muito pequeno, já será suficiente para que o método convirja para v_1.

Exemplo 12. Consideremos a matriz:

$$A = \begin{bmatrix} -2 & 1 & 0 & 2 \\ -1 & 10 & -1 & -1 \\ -1 & 1 & -2 & 1 \\ 1 & 0 & -1 & 2 \end{bmatrix}$$

[6]Isto não significa que não é possível majorar $||v_1 - u^{(n+1)}||_\infty$, pois é óbvio que será sempre menor que 2. Apenas significa que não há convergência para zero.

Figura 2.3.1:

Pelo Teorema de Gerschgorin, aplicado a linhas, concluímos que os valores próprios devem estar na reunião das bolas

$$B_1 = \bar{B}(-2,3), B_2 = \bar{B}(10,3), B_3 = \bar{B}(-2,3), B_4 = B(2,2).$$

Imediatamente vemos que irá haver duas componentes conexas, uma que será B_2, onde haverá apenas um valor próprio real, e a outra componente que será a reunião das três restantes bolas, o que se resume a $B_1 \cup B_4$, onde haverá três valores próprios (ver a primeira figura). Feita uma análise por colunas, obtemos as bolas

$$B_1' = \bar{B}(-2,3), B_2' = \bar{B}(10,2), B_3' = \bar{B}(-2,2), B_4' = B(2,4),$$

e concluímos que há apenas um valor próprio real em B_2' e três valores próprios em $B_1' \cup B_4'$. Intersectando a informação, concluímos que há um valor próprio real $\lambda_1 \in [8,12]$ e três valores próprios $\lambda_2, \lambda_3, \lambda_4 \in \bar{B}(-2,3) \cup \bar{B}(2,2)$, onde este último domínio é obtido pela intersecção de $B_1 \cup B_4$ com $B_1' \cup B_4'$.
Podemos concluir que o valor próprio real λ_1 é um valor próprio dominante, porque $|\lambda_1| \geq 8$ e $|\lambda_2|, |\lambda_3|, |\lambda_4| \leq 5$.
Admitindo que a matriz é diagonalizável, estamos nas condições de convergência do método das potências. Partimos do vector inicial $u^{(0)} = (0,1,0,0)$, escolha que foi orientada pelo vector próprio dominante associado à matriz diagonal. Obtemos

$$u^{(1)} = \sigma_0 \frac{Au^{(0)}}{||Au^{(0)}||_\infty} = + \frac{(1,10,1,0)}{||(1,10,1,0)||_\infty} = (\frac{1}{10}, 1, \frac{1}{10}, 0)$$

e sucessivamente $u^{(2)} = (0.0816, 1\ , 0.0714, 0)$, $u^{(3)} = (0.0846, 1\ , 0.077, 0.0009)$. Em $u^{(3)}$ todos os dígitos apresentados já são correctos. Calculando $Au^{(3)} = (0.832, 9.835, 0.758, 0.0082)$, obtemos a aproximação $\lambda^{(3)} = 9.835$, que é dada pelo valor da segunda componente (pois é nessa que $u^{(3)}$ tem valor unitário... poderíamos obter outras aproximações dividindo a componente $Au_k^{(3)}$ por $u_k^{(3)}$, mas esta é a mais simples de obter). O valor exacto do valor próprio dominante é $\lambda_1 = 9.83703$. Como $\lambda_2 = 2.34741$, temos uma convergência linear com o factor $|\frac{\lambda_2}{\lambda_1}| = 0.238$, o que nos permitiria escrever

$$||v_1 - u^{(n)}||_\infty \leq 0.238^n C.$$

Como avaliar a constante C?
– Conhecendo o valor exacto do vector próprio, podemos avaliar os erros, e admitindo que

$$||v_1 - u^{(n)}||_\infty = |\frac{\lambda_2}{\lambda_1}|^n C_n,$$

colocamos num gráfico os valores de C_n obtidos (figura à esquerda), e podemos concluir que os valores de C_n são inferiores a 0.1, aproximando-se de 0.03. ((Na figura da direita colocamos em evidência a dependência da constante C do valor α, componente segundo o vector próprio dominante. É considerada a matriz M usada na observação anterior e o vector inicial $u^{(0)} = (1, 1, \frac{1}{453})$. Com este vector inicial temos $\alpha = \frac{1}{453} \neq 0$, e o método converge para o vector próprio dominante. Como podemos ver no gráfico os valores de C_n tendem para um valor próximo de

450, e para outros α o valor da constante seria próximo de $\frac{1}{\alpha}$. Como previsto na estimativa de erro, a constante depende directamente de $\frac{1}{\alpha}$, e isso é aqui verificado))

– Como a priori não conhecemos os valores exactos, podemos no entanto obter informações avaliando o comportamento de $||u^{(n+1)} - u^{(n)}||_\infty$ para n suficientemente grande. Notamos que

$$||u^{(n+1)} - u^{(n)}||_\infty \leq ||u^{(n+1)} - v_1||_\infty + ||v_1 - u^{(n)}||_\infty,$$

e então

$$||u^{(n+1)} - u^{(n)}||_\infty \leq C_{n+1}|\frac{\lambda_2}{\lambda_1}|^{n+1} + C_n|\frac{\lambda_2}{\lambda_1}|^n.$$

Admitindo que $C_n \leq C$, então

$$||u^{(n+1)} - u^{(n)}||_\infty \leq 2C|\frac{\lambda_2}{\lambda_1}|^n.$$

Assim, é usual considerar a razão

$$K_n = \frac{||u^{(n+1)} - u^{(n)}||_\infty}{||u^{(n)} - u^{(n-1)}||_\infty} \approx \frac{2C|\frac{\lambda_2}{\lambda_1}|^n}{2C|\frac{\lambda_2}{\lambda_1}|^{n-1}} = |\frac{\lambda_2}{\lambda_1}|,$$

o que permite não apenas ter informação acerca da rapidez de convergência, mas também avaliar $|\lambda_2|$, já que $|\lambda_2| \sim K_n|\lambda_1|$. Com efeito, partindo dos valores $u^{(1)}, u^{(2)}, u^{(3)}$ calculados, poderíamos obter

$$K_2 = \frac{||u^{(3)} - u^{(2)}||_\infty}{||u^{(2)} - u^{(1)}||_\infty} = \frac{0.007328}{0.02857} = 0.2565,$$

o que não difere muito do valor 0.2386. Como tinhamos obtido $\lambda^{(3)} = 9.835$, retiramos $|\lambda_2| \sim 0.2386 \times 9.835 = 2.3466$, o que é uma aproximação muito razoável, já que $\lambda_2 = 2.34741$.

2.4 Método das iterações inversas

Este método é semelhante ao método das potências, mas baseia-se num conhecimento prévio da localização dos valores próprios. Continuamos a assumir uma diagonalização com os valores próprios que consideraremos reais (normalmente trabalharemos com matrizes hermiteanas). O método das potências apenas permitia aproximar o valor próprio dominante. Aqui consideramos qualquer um, mas precisamos de um conhecimento *a priori* sobre ele, que pode advir do Teorema de Gerschgorin ou de uma aproximação pelo método das potências.

Assim, é suposto termos λ como aproximação do valor próprio λ_m que pretendemos calcular. Logo,

$$Av = \lambda_m v \Leftrightarrow (A - \lambda I)v = (\lambda_m - \lambda)v \Leftrightarrow \frac{v}{\lambda_m - \lambda} = (A - \lambda I)^{-1}v,$$

e portanto se v é valor próprio de A, também é de $(A - \lambda I)^{-1}$. No entanto, os valores próprios são diferentes, λ_m é valor próprio de A e $\mu_m = \frac{1}{\lambda_m - \lambda}$ é valor próprio de $(A - \lambda I)^{-1}$ para o mesmo vector próprio!

A partir de uma iterada inicial $x^{(0)}$ (... com componente não nula no vector próprio), obtemos o *método das iterações inversas* (ou método de deflação de Wielandt)

$$x^{(n+1)} = \sigma_n \frac{(A - \lambda I)^{-1}x^{(n)}}{||(A - \lambda I)^{-1}x^{(n)}||_\infty}$$

em que σ_n é o sinal da componente de maior módulo de $(A-\lambda I)^{-1}x^{(n)}$. Reparamos, mais uma vez, tratar-se uma iteração do ponto fixo, pois como vimos,

$$Av = \lambda_m v \Leftrightarrow \frac{v}{\lambda_m - \lambda} = (A-\lambda I)^{-1}v,$$

e daqui obtemos $||(A-\lambda I)^{-1}v|| = \frac{||v||}{|\lambda_m - \lambda|}$. Assim:

$$\frac{v}{||v||} = \frac{\lambda_m - \lambda}{|\lambda_m - \lambda|}\frac{(A-\lambda I)^{-1}v}{||(A-\lambda I)^{-1}v||}$$

e mais uma vez substituimos $\frac{\lambda_m - \lambda}{|\lambda_m - \lambda|}$ pelo sinal da componente que determina o módulo (que designamos por σ).

A maneira para calcular de calcular as sucessivas iteradas baseia-se numa única factorização

$$A - \lambda I = LU,$$

seguida de sucessivas resoluções de sistemas (para cada n) :

$$LUw = x^{(n)} \Leftrightarrow \left\{ \begin{array}{c} Ly = x^{(n)} \\ Uw = y \end{array} \right.$$

o valor w é $(A-\lambda I)^{-1}x^{(n)}$. Assim, obtemos $x^{(n+1)} = \sigma_n w/||w||_\infty$.

Reparamos que o método das iterações inversas dá-nos uma aproximação do vector próprio v, para calcularmos uma aproximação do valor próprio devemos fazer:

$$\lambda_m \sim \frac{[Ax^{(n)}]_i}{[x^{(n)}]_i}$$

De forma semelhante ao que conseguimos no método das potências podemos mostrar a convergência deste método desde que

$$L = \frac{|\lambda_m - \lambda|}{\min_{i\neq m}|\lambda_i - \lambda|} < 1.$$

Este resultado pode ser facilmente verificado se repararmos que isto corresponde a considerar

$$\max_{i\neq m}\frac{1}{|\lambda_i - \lambda|} < \frac{1}{|\lambda_m - \lambda|}$$

o que significa que $\frac{1}{|\lambda_m - \lambda|}$ é valor próprio dominante de $(A-\lambda I)^{-1}$. Depois basta aplicar o resultado obtido para o método das potências.

Exemplo 13. Consideremos a matriz

$$A = \left[\begin{array}{cccc} -15 & 0 & 1 & 1 \\ 2 & 10 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array} \right]$$

Pretendemos aproximar o valor próprio que se encontra no intervalo $[8, 12]$ e escolhemos $\lambda = 9$. Ficamos com a factorização

$$A - \lambda I = \begin{bmatrix} -24 & 0 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 1 & 1 & -8 & 1 \\ 1 & 1 & 1 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/12 & 1 & 0 & 0 \\ -1/24 & 1 & 1 & 0 \\ -1/24 & 1 & \frac{-23}{193} & 1 \end{bmatrix} \begin{bmatrix} -24 & 0 & 1 & 1 \\ 0 & 1 & 1/12 & 1/12 \\ 0 & 0 & -\frac{193}{24} & \frac{23}{24} \\ 0 & 0 & 0 & -\frac{1530}{193} \end{bmatrix} = LU.$$

que iremos utilizar para calcular $(A - \lambda I)^{-1}$.

Escolhendo como iterada inicial $u^{(0)} = (0, 1, 0, 0)$. A escolha deve-se às mesmas razões que as justificadas no exemplo anterior, atendendo a que agora procuramos o valor próprio que está na bola com centro em 10.

Resolvendo $LUw = u^{(0)}$, obtemos $(0.0117, 0.9765, 0.1412, 0.1412)$, portanto

$$\begin{aligned} u^{(1)} &= \sigma_0 \frac{(A - \lambda I)^{-1} u^{(0)}}{||(A - \lambda I)^{-1} u^{(0)}||_\infty} = \frac{(0.0117, 0.9765, 0.1412, 0.1412)}{||(0.0117, 0.9765, 0.1412, 0.1412)||_\infty} \\ &= (0.0120482, 1., 0.144578, 0.144578) \end{aligned}$$

neste caso $Au^{(1)} = (0.1084, 10.024, 1.301, 1.301)$ e portanto $\lambda^{(1)} = 10.024...$
Continuando com as iterações, $u^{(2)} = (0.00975434, 1., 0.123194, 0.123194)$, portanto $\lambda^{(2)} = 10.0195...$, $\lambda^{(3)} = 10.0202$ e aproximaríamos rapidamente o valor correcto $\lambda = 10.0201$.

2.5 Métodos de Factorização

Um dos métodos mais utilizados para a determinação de valores próprios é o método QR, de Francis, que foi precedido por um outro método semelhante, devido a Rutishauser – o método LR, apresentado no final dos anos 50. A ideia principal destes métodos consiste em efectuar uma factorização da matriz num produto de matrizes mais simples, trocar a ordem do produto e obter uma nova matriz a que será aplicado o mesmo esquema!

Estes métodos baseiam-se na semelhança entre matrizes, pois escrevendo

$$B = P^{-1} A P,$$

a matriz A tem os mesmos valores próprios que B. Portanto, a ideia consiste em efectuar a iteração

$$A_{n+1} = P_n^{-1} A_n P_n,$$

começando com $A_0 = A$, se no limite tivermos uma matriz cujo cálculo dos valores próprios é simples (por exemplo, uma matriz triangular) então o problema fica simplificado, ou resolvido.

Alternativamente, estes métodos podem ser encarados como resultantes de uma factorização das matrizes. Assim, se for possível efectuar uma factorização do tipo $A_n = X_n Y_n$, em que X_n é invertível, bastará considerar $A_{n+1} = Y_n X_n$ para termos

$$A_{n+1} = X_n^{-1} A_n X_n,$$

porque $Y_n = X_n^{-1} A_n$.

2.5.1 Método LR

No caso do método LR, de Rutishauser, efectuamos uma factorização $A = LU$ que por tradição é designada LR (left-right ao invés de lower-upper). Assim, começando com $A_0 = A$, e tendo obtido

$$A_n = L_n U_n$$

definimos a nova iterada como sendo

$$A_{n+1} = U_n L_n,$$

o que também significa que consideramos a iteração $A_{n+1} = L_n^{-1} A_n L_n$. Reparamos que a matriz A_{n+1} é semelhante a A_n e por isso os valores próprios são os mesmos, subsequentemente os mesmos que os de $A_0 = A$.

Se o método convergir, é suposto que a sucessão de matrizes A_n tenda para uma matriz triangular superior, cuja diagonal irá conter os valores próprios. No entanto, não é fácil obter condições de convergência para este método, podendo ser bastante instável. Sabe-se (cf. [10]) que se A for simétrica e definida positiva há convergência.

2.5.2 Método QR

• O método QR, de Francis, é baseado numa factorização menos conhecida

$$A = QR$$

em que Q é uma matriz unitária (ou seja, $QQ^* = Q^*Q = I$) e R uma matriz triangular superior.

Proposição 13. *A factorização $A = QR$ é única, a menos de produto por uma matriz diagonal, cujas entradas têm módulo 1.*

Demonstração. Supondo que $A = Q_1R_1 = Q_2R_2$, então $R_1R_2^{-1} = Q_1^*Q_2$, o que significa que a matriz triangular superior $R_1R_2^{-1}$ seria uma matriz ortogonal (porque $Q_1^*Q_2$ é). No entanto, as únicas matrizes nestas condições são matrizes diagonais, logo $R_1R_2^{-1} = D$, ou seja $R_1 = DR_2$ e $Q_1^*Q_2 = D$, ou seja $Q_2 = Q_1D$. Verifica-se que essa diagonal verifica $DD^* = Q_1^*Q_2Q_2^*Q_1 = I$, ou seja $|d_{ii}| = 1$. □

• Construção da factorização QR através de matrizes de Householder.

Uma matriz de Householder é uma matriz do tipo

$$H = I - 2vv^*$$

em que $v : ||v||_2 = 1$, ou seja $v^*v = 1$ (note-se que v^*v é uma matriz 1×1, identificada com um número, mas vv^* já é uma matriz $N \times N$).

As matrizes de Householder são unitárias porque $HH^* = H^*H = (I - 2vv^*)(I - 2vv^*) = I - 4vv^* + 4vv^*vv^* = I$.

Podemos considerar vectores $v^{(k)} = (0, .., 0, v_k, ..., v_N)$, que irão definir matrizes de Householder H_k. É possível efectuar a decomposição QR :

$$R = H_{N-1}...H_1A,$$
$$Q^* = H_{N-1}...H_1,$$

já que é fácil verificar que $QR = H_1^*...H_{N-1}^* H_{N-1}...H_1 A = A$, faltando apenas ver que $H_{N-1}...H_1 A$ é triangular superior calculando $v^{(k)}$ (cf.[1]).

- O método QR consiste em começar com $A_0 = A$, e tendo factorizado

$$A_n = Q_n R_n,$$

definir uma nova iterada

$$A_{n+1} = R_n Q_n,$$

ou seja, $A_{n+1} = Q_n^* A_n Q_n$ que é uma matriz semelhante a A_n.

Teorema 18. *(Francis) Se a matriz A for invertível e os seus valores próprios tiverem módulos diferentes, $|\lambda_1| > ... > |\lambda_N| > 0$, a matriz é diagonalizável, ou seja, $A = P^{-1}DP$. Se P admitir uma factorização $P = LU$, a sucessão de matrizes (A_n) converge para uma matriz triangular superior cujos valores da diagonal serão os valores próprios de A. Os vectores próprios associados encontram-se na matriz unitária Q.*

Demonstração. Ver, por exemplo, [3]. ■ □

No caso mais geral, pode convergir para uma matriz quase triangular (por blocos), cujos valores próprios são razovelmente fáceis de calcular. O método tem ainda normalmente a particularidade de apresentar os valores próprios ordenados, estando na primeira linha o maior e na última o mais pequeno. A rapidez de convergência para zero dos elementos não diagonais depende da relação $|\frac{\lambda_k}{\lambda_{k+1}}|$, o que pode constituir um obstáculo à rapidez do método, quando alguns valores próprios têm módulos semelhantes. Por isso é usada uma técnica de aceleração de convergência que veremos mais à frente.

Exemplo 14. Consideramos a factorização QR da matriz A,

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 3 & 2 \\ 2 & -2 & -4 \end{bmatrix} = \overbrace{\begin{bmatrix} \sqrt{2/3} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix}}^{Q} \overbrace{\begin{bmatrix} 2\sqrt{6} & \sqrt{3/2} & 0 \\ 0 & 5/\sqrt{2} & 3\sqrt{2} \\ 0 & 0 & \sqrt{3} \end{bmatrix}}^{R},$$

não especificando os cálculos inerentes... A partir deste ponto calculamos $A_1 = RQ$, e obtemos uma nova matriz cuja diagonal é $\{4.5, -0.5, -1\}$, estes são os primeiros valores que aproximam os valores próprios de A. Voltamos a efectuar a decomposição $A_1 = Q_1 R_1...$ que por razões óbvias não será aqui colocada. Calculando $A_2 = R_1 Q_1$, obtemos na diagonal os valores $\{4.52, -3.52, 2\}$. Procedendo de forma semelhante nas iteradas seguintes, obtemos ao fim de 7 iterações, na diagonal de A_7, os valores $\{5.07..., -3.69..., 1.62...\}$, que não estão muito longe dos valores próprios correctos $\{5, -3.64..., 1.64...\}$. A matriz A_7 já é próxima de uma matriz triangular superior,

$$A_7 = \begin{bmatrix} 5.07467 & -1.89079 & 0.853412 \\ 0.343404 & -3.69541 & -3.5415 \\ 0.001455 & -0.0368519 & 1.62073 \end{bmatrix}.$$

O valor absoluto do maior elemento da subdiagonal determina, normalmente, um majorante do erro da aproximação.

Observação 31. (método de Jacobi). Outra possibilidade de obter a factorização QR é usar matrizes de rotações no plano ao invés de matrizes de Householder, ideia que também é usada no método de Jacobi. O método de Jacobi é válido para matrizes reais simétricas e baseia-se na utilização de matrizes de rotação

$$U = \begin{bmatrix} 1 & 0 & & \cdots & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & & \vdots \\ & \sin(\theta) & \cos(\theta) & & \\ \vdots & & & 1 & 0 \\ & & & \ddots & \\ 0 & \cdots & & 0 & 1 \end{bmatrix},$$

mas não falaremos dele aqui (ver, por exemplo, [9]).

Ainda uma outra possibilidade para a efectuar a factorização QR é considerar o processo de ortonormalização de Gram-Schmidt, que no entanto é instável numericamente, devido ao cancelamento subtractivo.

2.5.3 Método QR com deslocamento

Os métodos de factorização são computacionalmente dispendiosos em termos de tempo (aproximadamente $\frac{2}{3}N^3$ operações por iteração) e como já referimos a sua convergência pode ser lenta. Uma possibilidade para acelerar a convergência destes métodos é utilizar uma técnica de deslocamento (ou *shift*), reparando que considerando

$$\tilde{A} = A - \alpha I$$

se uma matriz B for semelhante a $\tilde{A}$ então $\tilde{B} = B + \alpha I$ será semelhante a A, porque

$$B = P^{-1}\tilde{A}P = P^{-1}(A - \alpha I)P = P^{-1}AP - \alpha I.$$

Assim, para o método QR, podemos estruturar os passos usando um deslocamento α_n diferente, em cada passo, de forma a que efectuamos primeiro a decomposição QR da matriz $A_n - \alpha_n I$ e depois trocamos a ordem somando $\alpha_n I$. Ou seja,

$$\begin{aligned} A_n - \alpha_n I &= Q_n R_n, \\ A_{n+1} &= R_n Q_n + \alpha_n I, \end{aligned}$$

ficando com

$$A_{n+1} = Q_n^*(A_n - \alpha_n I)Q_n + \alpha_n I = Q_n^* A_n Q_n,$$

e desta forma, A_{n+1} continua a ser uma matriz semelhante a A_n. A escolha do deslocamento α_n é discutida em [13] e uma das possibilidades é considerar α_n como sendo o elemento de menor módulo da diagonal (normalmente o último).

Observação 32. Apesar de ser o método mais utilizado para o cálculo de valores próprios, o método QR com *shift* tem resistido à demonstração da sua convergência no caso mais geral (cf.[3]).

Observação 33. O *Mathematica* tem implementadas as rotinas Eigenvalues e Eigenvectors, que permitem o cálculo de valores e vectores próprios de matrizes, usando um método QR com *shift*. A factorização QR pode ser obtida usando a rotina QRDecomposition.(o resultado é uma lista com a transposta da matriz Q e com a matriz R).

2.6 Condicionamento do cálculo de valores próprios

Apresentamos agora um resultado relativo ao condicionamento do cálculo de valores próprios.

Teorema 19. *(Bauer-Fike). Seja A uma matriz hermitiana. No caso de $\tilde{A}$ ser uma aproximação (hermitiana) de A, temos o resultado*

$$\forall j \, \exists i : |\lambda_i - \tilde{\lambda}_j| \leq ||A - \tilde{A}||_2 \tag{2.6.1}$$

em que λ_i são os valores próprios de A e $\tilde{\lambda}_j$ os de $\tilde{A}$.

No caso mais geral, em que há a matriz tem forma canónica de Jordan diagonal, $A = P^{-1}DP$ (com $D = diag(\lambda_1, ..., \lambda_N)$), temos

$$\forall j \, \exists i : |\lambda_i - \tilde{\lambda}_j| \leq cond_\infty(P)||A - \tilde{A}||_\infty. \tag{2.6.2}$$

(o que também é válido para algumas outras normas, como $||.||_1, ||.||_2$).

Demonstração. i) Começamos por ver que o resultado sai facilmente para a norma $||.||_\infty$ (ou mesmo para $||.||_1$).

Seja $B = P(A - \tilde{A})P^{-1}$, temos $B = D - C$ em que $C = P\tilde{A}P^{-1}$ tem os valores próprios de $\tilde{A}$. Pelo teorema de Gerschgorin, aplicado a $C = D - B$, sabemos que dado um valor próprio $\tilde{\lambda}_j$ de C existe uma linha i :

$$|\lambda_i - b_{ii} - \tilde{\lambda}_j| \leq \sum_{k \neq i} |b_{ik}|,$$

e portanto

$$|\lambda_i - \tilde{\lambda}_j| \leq \sum_k |b_{ik}| \leq ||B||_\infty \leq ||P||_\infty ||A - \tilde{A}||_\infty ||P^{-1}||_\infty.$$

ii) Para mostrar que é válido para a norma $||.||_2$, vemos que

$$\min_{i=1,...,N} |\lambda_i - \tilde{\lambda}| \leq \text{cond}_2(P)||A - \tilde{A}||_2,$$

para qualquer valor próprio $\tilde{\lambda}$, e a partir daqui podemos aplicar de novo o teorema de Gerschgorin para concluir o teorema.

Suponhamos que $\tilde{\lambda} \neq \lambda_i$ para qualquer i (senão seria trivial, pois o mí nimo seria zero) e seja $\tilde{v}$ um vector próprio de $\tilde{A}$.

Como $\tilde{A}\,\tilde{v} = \tilde{\lambda}\tilde{v}$,

$$(\tilde{\lambda}I - A)\tilde{v} = (\tilde{A} - A)\tilde{v} \tag{2.6.3}$$

e substituindo A, temos $(\tilde{\lambda}I - A)\tilde{v} = (\tilde{\lambda}I - P^{-1}DP)\tilde{v} = P^{-1}(\tilde{\lambda}I - D)P\tilde{v}$ o que implica, por (2.6.3), que

$$(\tilde{\lambda}I - D)P\tilde{v} = P(\tilde{A} - A)P^{-1}P\tilde{v}.$$

Como $\tilde{\lambda} \neq \lambda_i$, a matriz diagonal $\tilde{\lambda}I - D$ tem inversa, e obtemos

$$P\tilde{v} = (\tilde{\lambda}I - D)^{-1}P(\tilde{A} - A)P^{-1}P\tilde{v}.$$

Notando que $||(\tilde{\lambda}I - D)^{-1}||_2 = \rho((\tilde{\lambda}I - D)^{-1}) = \frac{1}{\min|\tilde{\lambda}-\lambda_i|}$ (o que também é válido para outras normas ditas 'monótonas'), temos

$$||P\tilde{v}||_2 \leq \frac{1}{\min|\tilde{\lambda} - \lambda_i|} ||P(\tilde{A} - A)P^{-1}||_2 \, ||P\tilde{v}||_2$$

o que origina

$$\min_{i=1,\ldots,N} |\lambda_i - \tilde{\lambda}| \leq ||P||_2||P^{-1}||_2||A - \tilde{A}||_2.$$

No caso de matrizes hermitianas, basta referir que pela decomposição na forma normal de Schur podemos encontrar matrizes P unitárias tal que $A = P^*DP$, pelo que $||P||_2 = ||P^*||_2 = 1$. □

Observação 34. A propriedade que provámos traduz também o bom condicionamento no cálculo de valores próprios para as matrizes hermitianas. Para outro tipo de matrizes, o cálculo dos valores próprios poderá ser um problema mal condicionado, dependendo do número de condição da matriz P.

Como a estimativa do número de condição de P não é normalmente possível (se P fosse conhecido também seriam os seus valores próprios), apenas temos a informação da possibilidade de ocorrerem problemas de condicionamento no cálculo dos valores próprios.

2.7 Cálculo de raízes polinomiais

Terminamos este capí tulo referindo que um excelente processo de obter resultados acerca das raí zes de polinómios é a utilização da noção de matriz companheira de um polinómio.

Definição 5. Dizemos que $\mathcal{C}$ é a matriz companheira do polinómio $p(x) = a_0 + a_1x + ... + a_{N-1}x^{N-1} + x^N$, se

$$\mathcal{C} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{N-2} & -a_{N-1} \end{bmatrix}$$

notando que o polinómio característico de $\mathcal{C}$ é exactamente p.

Esta noção pode ser aplicada para a localização das raízes de polinómios através do teorema de Gerschgorin (ver exercício 2, no final do capítulo) ou mesmo para aproximá-las usando um qualquer método de valores próprios, já que identificar os valores próprios de C é equivalente a determinar as raízes de p. Deste facto retiramos que a *determinação de valores próprios é um problema teoricamente equivalente à resolução de equações algébricas.*

Exemplo 15. Tomemos como exemplo o método das potências aplicado a $\mathcal{C}$. Executar a iteração

$$x^{(n+1)} = \mathcal{C}\, x^{(n)}$$

é equivalente a considerar

$$\begin{cases} x_i^{(n+1)} = x_{i+1}^{(n)} & \text{se } i = 1, ..., N-1, \\ x_N^{(n+1)} = -a_0 x_1^{(n)} - ... - a_{N-1} x_N^{(n)} & \text{caso } i = N. \end{cases}$$

Reparamos assim que $x_1^{(n)} = x_2^{(n-1)} = ... = x_N^{(n-N+1)}$, $x_2^{(n)} = ... = x_N^{(n-N+2)}$, etc... de um modo geral, $x_i^{(n)} = x_N^{(n-N+i)}$, o que corresponde a substituir valores na iterada n por valores em iteradas anteriores.

Ora, designando $y_k = x_N^{(k-N+1)}$, obtemos $x_i^{(n)} = y_{n+i-1}$, pelo que o sistema anterior reduz-se à equação às diferenças

$$y_{n+N} = -a_0 y_n - ... - a_{N-1} y_{n+N-1}.$$

A mesma equação às diferenças que encontrámos no método de Bernoulli.

Para concluirmos que o *método de Bernoulli aparece como um caso particular do método das potências*, reparamos que no caso do método das potências consideramos como aproximação do valor próprio dominante[7]:

$$\lambda^{(n)} = \frac{[\mathcal{C}\, x^{(n)}]_1}{x_1^{(n)}} = \frac{x_1^{(n+1)}}{x_1^{(n)}} = \frac{y_{n+1}}{y_n},$$

ou seja, a mesma aproximação que consideramos no método de Bernoulli para a raiz dominante!

Outros métodos para valores próprios levam a outras aproximações, não havendo necessariamente um método especí fico para polinómios que lhes corresponda, como neste caso aconteceu com o método de Bernoulli.

Observação 35. Como curiosidade reparamos que a matriz inversa da matriz companheira é

$$\mathcal{C}^{-1} = \begin{bmatrix} -\frac{a_1}{a_0} & -\frac{a_1}{a_0} & \cdots & -\frac{a_{N-1}}{a_0} & -\frac{1}{a_0} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

que tem associada como polinómio característico $q(y) = \frac{1}{a_0} + \frac{a_{N-1}}{a_0} y + ... + \frac{a_1}{a_0} y^{N-1} + y^N$ cujas raízes são as inversas de $p(x)$, como vimos num exercício do Capítulo 2 (basta tomar $y = 1/x$). Isto é perfeitamente natural, já que é claro que os valores próprios da matriz inversa são os inversos da original.

[7] Ver também a nota de rodapé anterior, considerando o método das potências sem a normalização sucessiva! Consideramos aqui a primeira componente, mas para qualquer componente j obteríamos

$$\lambda^{(n)} = \frac{x_j^{(n+1)}}{x_j^{(n)}} = \frac{y_{n+j}}{y_{n+j-1}}$$

o que corresponde ao mesmo resultado.

2.8 Exercícios

1. (Método de Krylov) Considere o seguinte método, baseado na aplicação do teorema de Hamilton-Cayley, para encontrar o polinómio característico de uma matriz A de dimensão N :

• Calcular A^k, para $k = 2, ..., N$

• Determinar os coeficientes α_i tais que $\alpha_0 I + \alpha_1 A + ... + \alpha_{N-1} A^{N-1} + A^N = 0$.

a) Indique uma estimativa do número de operações $(*, /)$ necessárias a esse cálculo.

b) Use este método para determinar a equação característica de uma matriz 2×2.

c) Ao invés de calcular A^k, considere um vector inicial $x^{(0)}$ e defina $x^{(k)} = Ax^{(k-1)}$. Apresente um processo equivalente para determinar o polinómio característico. Comente quanto ao número de operações e quanto à solubilidade do sistema.

2. Considere a matriz companheira do polinómio com coeficientes reais $p(x) = a_0 + a_1 x + ... + a_{n-1} x^{n-1} + x^n$.

a) Mostre que se $|a_{n-1}| > 1 + M$, com $M = \max\{|a_0|, |a_1| + 1, ..., |a_{n-2}| + 1\}$ então existe uma e uma só raiz real dominante em $[-a_{n-1} - 1, -a_{n-1} + 1]$, e que as restantes se encontram na bola $\{|z| \leq M\}$.

b) Considere $p(x) = 2 - 6x^2 + 4x^3 - 16x^4 + 2x^5$.

Localize as raí zes dominante num intervalo de comprimento 2 e as restantes numa bola de raio 1.

Determine aproximadamente a raiz dominante usando duas iterações do método das potências.

3. Seja A uma matriz real $N \times N$, que verifica:

$$|a_{ii} - a_{jj}| > r_i + r_j, \forall i, j = 1, ..., N \ (i \neq j)$$

em que

$$r_k = -|a_{kk}| + \sum_{j=1}^{N} |a_{kj}|$$

Mostre que os valores próprios da matriz são reais.

4. Considere uma matriz $A \in \mathbb{C}^N \times \mathbb{C}^N$ e várias sucessões $\mu^{(k)} \in l^1$. Supondo que

$$|a_{ii}| > ||\mu^{(i)}||_1 \ \ \forall i = 1, \ldots, N$$

$$|a_{ij}| \leq |\mu_j^{(i)}| \ \ \forall i, j = 1, \ldots, N, \ \ (i \neq j)$$

a) Mostre que a matriz A é invertível.

b) Mostre que se A for hermitiana e tiver a diagonal positiva, então é definida positiva e o raio espectral verifica

$$\rho(A) \leq \max_{i=1,...,N} \left(|a_{ii}| + ||\mu^{(i)}||_1 \right).$$

c) Mostre que é possível resolver o sistema $Ax = b$, para qualquer $b \in \mathbb{R}^N$, usando o método de Jacobi, e que se verifica:

$$||x - x^{(n)}||_\infty \leq \frac{L^n}{1 - L} \frac{||b||_\infty}{K}$$

considerando $x^{(0)} = 0$, com

$$L = 1 - \min_{i=1,\ldots,n} \left(\frac{|\mu_i^{(i)}|}{||\mu^{(i)}||_1}\right), \quad \text{e com } K = \min_{i=1,\ldots,n} ||\mu^{(i)}||_1.$$

5. Considere a matriz

$$A = \begin{bmatrix} 6 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & -1 & -1 \end{bmatrix}.$$

a) Aplicando o T. de Gerschgorin determine um domínio em $\mathbb{C}$ onde se encontram os valores próprios de A.

b) Conclua que existe um valor próprio dominante para A, e determine uma aproximação utilizando o método das potências.

c) Diga qual o raio espectral da matriz $A/10$? O que pode concluir acerca da convergência do seguinte método:

6. Considere a matriz

$$\begin{bmatrix} -1+i & 1 & 1 \\ 1 & -1-i & 1 \\ 1 & 0 & 3+4i \end{bmatrix}$$

a) Indique um domí nio do plano complexo onde se situam os valores próprios.

b) Determine um majorante para o módulo do determinante da matriz.

c) Entre que valores se pode situar o raio espectral da matriz? A matriz é invertível?

7. Considere a matriz

$$\begin{bmatrix} 8 & 1 & -1 \\ 1 & -3 & 1 \\ 0 & 1/2 & 1 \end{bmatrix}$$

a) Justifique que todos os valores próprios da matriz são reais, e indique intervalos que os contenham.

b) Verifique que a matriz possui um valor próprio dominante e aproxime-o considerando três iteradas do método das potências, usando como vector inicial $v^{(0)} = (1, 0, 0)$.

8. Considere a matriz

$$A = \begin{bmatrix} 10 & 3 - 2\cos(b) & \cos(b) \\ 1 & 25 & 5\sin(a) \\ 1 & 5\sin(a) + \sin(b) & 50 \end{bmatrix}$$

a) Localize os valores próprios de A usando o teorema de Gerschgorin.

b) Indique os valores de b para os quais podemos obter uma decomposição $A = LL^T$, em que L é uma matriz triangular inferior real.

c) Para que valores de $h \in \mathbb{R}^3$ é possível utilizar o método de Jacobi para resolver um sistema $Ax = h$? Indique uma estimativa de erro para $||e^{(n)}||_\infty$ em função de $||h||_\infty$, sabendo que $x^{(0)} = 0$.

9. Considere uma matriz $A \in \mathbb{C}^N \times \mathbb{C}^N$ e várias sucessões $\mu^{(k)} \in l^1$. Supondo que

$$|a_{ii}| > ||\mu^{(i)}||_1 \ \ \forall i = 1, \ldots, N$$

$$|a_{ij}| \leq |\mu_j^{(i)}| \ \ \forall i, j = 1, \ldots, N, \ \ (i \neq j)$$

a) Mostre que a matriz A é invertível.

b) Mostre que se A for hermitiana e tiver a diagonal positiva, então é definida positiva e o raio espectral verifica

$$\rho(A) \leq \max_{i=1,\ldots,N} \left(|a_{ii}| + ||\mu^{(i)}||_1 \right).$$

c) Mostre que é possível resolver o sistema $Ax = b$, para qualquer $b \in \mathbb{R}^N$, usando o método de Jacobi, e que se verifica:

$$||x - x^{(n)}||_\infty \leq \frac{L^n}{1-L} \frac{||b||_\infty}{K}$$

considerando $x^{(0)} = 0$, com

$$L = 1 - \min_{i=1,\ldots,n} \left(\frac{|\mu_i^{(i)}|}{||\mu^{(i)}||_1} \right), \quad \text{e com } K = \min_{i=1,\ldots,n} ||\mu^{(i)}||_1.$$

10. Suponha que obteve

$$A = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 0 & -1 \end{bmatrix}.$$

a) Calcule a primeira iteração pelo método QR.

b) Compare as aproximações dos valores próprios com os valores próprios correctos.

11. Considere

$$A = \begin{bmatrix} a\cos(\theta) & -a\sin(\theta) \\ a\sin(\theta) & a\cos(\theta) \end{bmatrix}.$$

Qual a factorização QR de A? Como se processa a iteração do método QR neste caso? Calcule os valores próprios de A e justifique.

12. Mostre que se os elementos de uma matriz forem números racionais então os valores próprios dessa matriz não podem ser números transcendentes.

Capítulo 3

Resolução de equações diferenciais ordinárias

3.1 Problema de Cauchy unidimensional

3.1.1 Problema de Cauchy e formulação integral

O exemplo mais simples de equação diferencial é escrito na forma

$$y'(t) = f(t),$$

cuja solução quando $y(t_0) = y_0$, é dada pela simples integração de f :

$$y(t) = y_0 + \int_{t_0}^{t} f(s)ds.$$

Os métodos para resolver equações diferenciais mais gerais podem usar ideias de integração numérica, e podem também ser usados para integração numérica. Consideramos mais geralmente o Problema de Cauchy

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0 \end{cases} \tag{3.1.1}$$

em que $f(t, y)$ depende de duas variáveis (a primeira t é o tempo, a segunda y refere as ocorrências de $y(t)$ no segundo membro. Associado a este problema está a sua forma integral equivalente[1]

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))ds, \tag{3.1.2}$$

onde a solução y é colocada como uma equação de ponto fixo, $y = Ay$, onde A é o operador integral definido no segundo membro. Este formato equivalente, quando f é Lipschitziana, permite concluir sobre a existência e unicidade de solução numa vizinhança do instante inicial t_0, por aplicação do Teorema do Ponto Fixo (de Banach):

[1]Para verificar a equivalência basta derivar a forma integral, e aplicar a fórmula de Barrow na equação diferencial, atendendo à condição inicial.

Teorema 20. *(Picard-Lindelöf). Consideremos $V_\varepsilon = B(t_0, \varepsilon_1) \times B(y_0, \varepsilon_2)$, uma vizinhança centrada nos dados iniciais (t_0, y_0), tal que f é contínua (na primeira variável) e Lipschitziana (na segunda variável) em $\bar{V}_\varepsilon$, ou seja, existe $L > 0$:*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|, \quad \forall (t, y_1), (t, y_2) \in \bar{V}_\varepsilon.$$

Então existe um $\tau = \min\{\varepsilon_1, \varepsilon_2/||f||_\infty\}$, que define o intervalo $\bar{B}(t_0, \tau)$ onde a solução $y \in C^1(\bar{B}(t_0, \tau))$ do problema de Cauchy existe e é única. O método do ponto fixo definido pela sucessão de funções (x_n) com $x_0(t) = y_0$,

$$x_{n+1}(t) = y_0 + \int_{t_0}^{t} f(s, x_n(s))ds,$$

converge uniformemente para y no intervalo $\bar{B}(t_0, \tau)$.

A iteração do ponto fixo, é neste contexto designada por *iteração de Picard*, e apesar de definir um método geral para a resolução de equações, é computacionalmente menos eficaz do que outros métodos que iremos ver.

A condição de f ser Lipschitziana na $2^{\underline{a}}$ variável pode ser verificada simplesmente exigindo limitação na derivada da $2^{\underline{a}}$ componente, ou seja, quando $f \in C^1$:

$$|\frac{\partial f}{\partial y}(t, y)| \leq L, \quad \forall (t, y) \in \bar{V}_\varepsilon,$$

Notamos que para garantir existência podemos apenas exigir que f seja contínua em $\bar{V}_\varepsilon$ (Teorema de Peano), mas isso não garante a unicidade. Ilustramos a não unicidade com um contraexemplo.

Exemplo 16. (contra-exemplo de unicidade para f apenas contínua)

Consideremos $t_0 = 1$, $y_0 = 0$, $f(t, y) = 2\sqrt{y}$, temos

$$\begin{cases} y'(t) = 2\sqrt{y(t)} \\ y(1) = 0 \end{cases}$$

e verificamos que $y(t) = (t-p)^2$ é solução da equação diferencial, pois $y'(t) = 2(t-p) = 2\sqrt{y(t)}$, para $t \geq p$, qualquer que seja o $p \geq 1$. Só $p = 1$ verifica $y(1) = 0$, mas notamos que podemos também definir

$$\begin{cases} y_p(t) = (t-p)^2, & \text{se } t \geq p \\ y_p(1) = 0, & \text{se } t < p \end{cases}$$

todas estas funções $y_p \in C^1(\mathbb{R})$ verificam o problema de Cauchy, para $p \geq 1$, e não há assim unicidade de solução. Isto não contradiz o Teorema de Picard-Lindelöf, porque

$$\frac{\partial f}{\partial y}(t, y) = \frac{\partial}{\partial y}(2\sqrt{y}) = \frac{1}{\sqrt{y}} \underset{(y \to 0)}{\longrightarrow} \infty,$$

e f não pode ser Lipschitziana perto de zero. Este caso é ainda ilustrativo, porque para $y_0 \neq 0$, este problema não ocorreria, e mudando essa condição inicial f já seria Lipschitziana, garantindo unicidade. A existência é em qualquer caso garantida pela continuidade de f, devido ao Teorema de Peano.

A restrição a um intervalo pequeno $[t_0 - \tau, t_0 + \tau]$ não é uma limitação teórica, porque podemos sempre definir novo problema de Cauchy avançando (ou retrocedendo) no tempo, considerando um novo $\tilde{t_0} = t_0 + \tau$. Sucessivamente podemos avançar no tempo até que as condições deixem de poder ser verificadas. Ou seja, o Teorema de Picard-Lindelöf garante existência e unicidade local, mas a sua aplicação sucessiva permite uma continuação até que ocorra uma descontinuidade de y.

Exemplo 17. (limitação da continuação da solução)

Consideramos a equação $y'(t) = y(t)^2$, onde a expressão $f(t, y) = y^2$ não faz antever nenhum problema de extensão da solução. Porém, notamos que quando $y(0) = y_0$, a solução única é dada por

$$y(t) = \frac{y_0}{1 - y_0 t}$$

e assim quando $t = \frac{1}{y_0}$ haverá uma singularidade, e um limite a partir do qual a solução não poderá ser estendida. Isto é ilustrativo da limitação da continuação da solução quando f não é limitada.

Por exemplo, para $y_0 = 1$, temos $y(t) = \frac{1}{1-t}$, e quando $t = 1 - \alpha$, temos $y(1 - \alpha) = \frac{1}{\alpha}$ que tende para infinito quando α é pequeno. Aplicando aí o Teorema de Picard-Lindelöf podemos ver que o novo intervalo obtido não permite passar o limite $t = 1$. Com efeito, $||f||_\infty = (\frac{1}{\alpha} + \varepsilon_2)^2$ e assim $\tau = \min\{\varepsilon_1, \frac{\varepsilon_2}{(1/\alpha+\varepsilon_2)^2}\}$, e sendo $0 < \alpha < \varepsilon_1$ obtemos $\tau = \frac{\varepsilon_2}{(1/\alpha+\varepsilon_2)^2} = \alpha\frac{\alpha\varepsilon_2}{(1+\alpha\varepsilon_2)^2} < \alpha$. Ou seja, ainda que a função f seja limitada e Lipschitziana em todos os intervalos fechados, o teorema devolve um intervalo de dimensão inferior, que evita chegar à singularidade.

Faz-se ainda notar que $f(t, y) = y$ leva à solução exponencial, por isso, $f(t, y) > y$ vão levar a crescimentos superiores ao da exponencial. E da mesma forma, se $f(t, y) > y^2$ então $y(t) > \frac{y_0}{1-y_0 t}$. Por exemplo, com isto podemos antecipar que se $y_0 = 1$, a solução de $y'(t) = e^{y(t)}$ irá "explodir" nalgum $t \leq 1$, o que ocorre de facto em $t = e^{-1}$, porque a solução é $y(t) = -\log(e^{-1} - t)$.

3.1.2 Casos particulares

Há várias situações em que é possível encontrar soluções explícitas das equações diferenciais. Quando f depende apenas de t, já vimos que se reduz a um problema de primitivação. De forma semelhante, quando f depende apenas da segunda variável (diz-se que a equação é autónoma), também podemos obter uma solução explícita por primitivação.

Proposição 14. *Se f não depende de t, e G é primitiva de $1/f(x)$, com inversa bem definida, a solução do problema de Cauchy é*

$$y(t) = G^{-1}(t - t_0 + G(y_0))$$

Demonstração. Com efeito, da equação diferencial $y' = f(y)$, obtemos por divisão

$$\frac{y'(t)}{f(y(t))} = 1 \implies \int_{t_0}^{t} \frac{y'(s)}{f(y(s))} ds = \int_{t_0}^{t} 1 ds = t - t_0$$

e agora notamos que se $G(x) = \int \frac{1}{f(x)} dx$ então

$$\frac{d}{dt} G(y(t)) = y'(t) G'(y(t)) = \frac{y'(t)}{f(y(t))}.$$

Consequentemente,

$$t - t_0 = \int_{t_0}^{t} \frac{y'(s)}{f(y(s))} ds = \int_{t_0}^{t} \frac{d}{dt} G(y(s)) ds = G(y(t)) - G(y_0),$$

e caso seja possível a inversão de G, a solução pode ser explícita por

$$y(t) = G^{-1}(t - t_0 + G(y_0)).$$

□

Exemplo 18. Consideramos $y'(t) = y(t)^p$, e como $1/f(x) = 1/x^p$ tem primitiva

$$G(x) = \int x^{-p} = \frac{x^{1-p}}{1-p}$$

com inversa $G^{-1}(z) = (z - pz)^{1/(1-p)}$, concluímos que

$$y(t) = ((t - t_0)(1-p) + y_0^{1-p})^{1/(1-p)}$$

expressão que já tinhamos encontrado quando $p = 2$. Note-se que se $y(0) = 1$, temos a singularidade em $t = \frac{1}{p-1}$, quando $p > 1$ (e não ocorre singularidade se $p \leq 1$).

Exemplo 19. Quando $y'(t) = ay(t) + b$, temos $f(t, y) = ay + b$, e obtemos $G(x) = \int \frac{1}{ax+b} dx = \frac{1}{a} \log(ax + b)$, resultando em $G^{-1}(z) = \frac{e^{az} - b}{a}$, e assim

$$y(t) = G^{-1}(t - t_0 + G(y_0)) = \tfrac{1}{a}(e^{a(t-t_0) + aG(y_0) - b}) = \frac{1}{a}(e^{a(t-t_0)}(ay_0 + b) - b).$$

Num caso linear mais geral,

$$y'(t) = a(t)y(t) + b(t)$$

temos $f(t, y) = a(t)y + b(t)$, e a função f depende de ambas as variáveis, pelo que o processo anterior de primitivação não pode ser aplicado, mas usando $A(t) = \int_{t_0}^{t} a(s) ds$, é possível encontrar:

$$y(t) = \left(y_0 + \int_{t_0}^{t} b(s) e^{-A(s)} ds \right) e^{A(t)},$$

que é uma forma explícita dependente apenas da primitivação das funções.

3.2 Sistemas e Equações de Ordem Superior

3.2.1 Sistemas de EDO's

O formato do problema de Cauchy pode ser adaptado para o caso vectorial, em que a função y depende apenas de t, mas tem como imagem um vector

$$\mathbf{y}(t) = (y_1(t), \ldots, y_N(t)),$$

atendendo a que $\mathbf{y}'(t) = (y_1'(t), \ldots, y_N'(t))$, e o sistema de equações diferenciais passa a ser definido também por uma função $\mathbf{f}$ com segunda componente vectorial. O problema de Cauchy fica assim

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases}$$

em que $\mathbf{y}_0$ é também um vector.

O Teorema de Picard-Lindelof pode ser aplicado de forma semelhante, exigindo que $\mathbf{f}$ seja Lipschitziana em $\mathbf{y}$, ou ainda que as derivadas a partir da segunda componente sejam limitadas. Eventuais dificuldades de adaptação ao caso escalar resultam da impossibilidade de divisão vectorial...

Um exemplo importante é aquele em que

$$\begin{cases} \mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t), \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases}$$

onde $\mathbf{A}$ é uma matriz $N \times N$. No caso escalar a solução seria $y(t) = y(t_0)e^{At}$, e neste caso também podemos escrever

$$\mathbf{y}(t) = e^{\mathbf{A}t}\mathbf{y}(t_0),$$

se dermos significado à noção de exponencial de uma matriz. Isso pode ser feito pela série de Taylor, mas é mais adequado, e eficaz, fazê-lo através dos valores e vectores próprios, pela decomposição espectral.

Admitindo que $\mathbf{A}$ é diagonalizável, consideramos $\mathbf{y}(t)$ escrita na base dos seus vectores próprios, $\mathbf{v}_1, \cdots, \mathbf{v}_N$, ou seja

$$\mathbf{y}(t) = y_{v_1}(t)\mathbf{v}_1 + \cdots + y_{v_N}(t)\mathbf{v}_N,$$

pelo que

$$\begin{aligned} \mathbf{y}'(t) &= y_{v_1}'(t)\mathbf{v}_1 + \cdots + y_{v_N}'(t)\mathbf{v}_N \\ \mathbf{A}\mathbf{y}(t) &= y_{v_1}(t)\lambda_1\mathbf{v}_1 + \cdots + y_{v_N}(t)\lambda_N\mathbf{v}_N \end{aligned}$$

implica, componente a componente $y_{v_k}'(t) = \lambda_k y_{v_k}(t)$, ou seja,

$$y_{v_k}(t) = y_{v_k}(t_0)e^{\lambda_k(t-t_0)}$$

obtendo-se explicitamente

$$\mathbf{y}(t) = y_{v_1}(t_0)e^{\lambda_1(t-t_0)}\mathbf{v}_1 + \cdots + y_{v_N}(t_0)e^{\lambda_N(t-t_0)}\mathbf{v}_N.$$

Observação 36. Sendo diagonalizável podemos escrever $\mathbf{A} = \mathbf{P^{-1}DP}$, em que $\mathbf{P}$ é a matriz mudança da base canónica para a base dos vectores próprios, e $\mathbf{D}$ é a matriz diagonal com os valores próprios, as operações polinomiais são transferidas para operações na diagonal de $\mathbf{D}$. Com efeito $\mathbf{A}^m = (\mathbf{P^{-1}D}^m\mathbf{P})$, pois por indução, é verdade para $m = 0, 1$, e sendo válido para m é válido para $m + 1$:

$$\mathbf{A}^{m+1} = \mathbf{A}^m\mathbf{A} = (\mathbf{P^{-1}D}^m\mathbf{P})(\mathbf{P^{-1}DP}) = (\mathbf{P^{-1}D}^m\mathbf{DP}) = (\mathbf{P^{-1}D}^{m+1}\mathbf{P}).$$

De forma semelhante, sendo $\mathbf{A}_1 = \mathbf{P^{-1}D_1P}$, $\mathbf{A}_2 = \mathbf{P^{-1}D_2P}$, verifica-se uma linearidade

$$\alpha\mathbf{A}_1 + \beta\mathbf{A}_2 = \alpha\mathbf{P^{-1}D_1P} + \beta\mathbf{P^{-1}D_2P} = \mathbf{P^{-1}}(\alpha\mathbf{D_1} + \beta\mathbf{D_2})\mathbf{P}.$$

Concluímos assim que para qualquer polinómio q

$$q(\mathbf{A}) = \mathbf{P}^{-1}q(\mathbf{D})\mathbf{P},$$

e de um modo geral para qualquer função f escrita na forma de série de potências, temos

$$f(\mathbf{A}) = \mathbf{P}^{-1}f(\mathbf{D})\mathbf{P},$$

com $f(\lambda_k)$ nos elementos da diagonal. Em particular, para $exp(\mathbf{A})$ a diagonal terá e^{λ_k}, conforme obtido antes.

3.2.2 Equações de Ordem Superior

As equações de ordem superior podem ser reduzidas a sistemas de equações através de introdução de novas variáveis.

Ou seja, vamos considerar que $y_0 = y$, que $y_1 = y'$, e sucessivamente $y_m = y^{(m)}$.

$$\begin{cases} y_0(t) = y(t) \\ \vdots \\ y_N(t) = y^{(N)}(t) \end{cases} \implies \begin{cases} y_0'(t) = y_1(t) = y'(t) \\ \vdots \\ y_{N-1}'(t) = y_N(t) = y^{(N)}(t) \end{cases}$$

Assim, quando escrevemos uma equação diferencial de ordem N genérica

$$y^{(N)}(t) = g(t, y(t), \cdots, y^{(N-1)}(t))$$

pela transformação de funções, fica equivalente a

$$y_{N-1}'(t) = g(t, y_0(t), \cdots, y_{N-1}(t))$$

e esta é a última equação de um sistema de N equações diferenciais

$$\begin{cases} y_0' = y_1 \\ \vdots \\ y_{N-2}' = y_{N-1} \\ y_{N-1}' = g(t, y_0, \cdots, y_{N-1}) \end{cases} \Leftrightarrow \mathbf{y}' = \left(\begin{bmatrix} y_0 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix} \right)' = \begin{bmatrix} y_1 \\ \vdots \\ y_{N-1} \\ g(t, y_0, \cdots, y_{N-1}) \end{bmatrix} = \mathbf{f}(t, \mathbf{y})$$

que está reduzido à forma vectorial anterior. Tudo o que vimos antes é assim aplicável, fazendo notar que a condição inicial é

$$\mathbf{y}(t_0) = (y(t_0), \cdots, y^{(N-1)}(t_0)),$$

ou seja, o problema de Cauchy é definido com os valores da função e de todas as derivadas até à ordem $N-1$, no ponto t_0.

Exemplo 20. (Equações com coeficientes constantes)

Um importante caso particular é

$$y^{(N)}(t) = a_0 y(t) + \cdots + a_{N-1} y^{(N-1)}(t)$$

o que corresponde a $g(t, \mathbf{y}) = a_0 y_0 + \cdots + a_{N-1} y_{N-1}$, porque podemos escrever na forma matricial

$$\mathbf{y}' = \left(\begin{bmatrix} y_0 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix} \right)' = \begin{bmatrix} y_1 \\ \vdots \\ y_{N-1} \\ a_0 y_0 + \cdots + a_{N-1} y_{N-1} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ a_0 & \cdots & \cdots & \cdots & a_{N-1} \end{bmatrix} \begin{bmatrix} y_0 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix} = \mathbf{A}\mathbf{y}$$

em que $\mathbf{A}$ é a matriz companheira do polinómio característico

$$p(\lambda) = \lambda^N - a_{N-1}\lambda^{N-1} - \cdots - a_0,$$

e as suas raízes λ_k são os valores próprios de $\mathbf{A}$ levando às soluções (quando é diagonalizável)

$$\mathbf{y}(t) = y_{v_1}(t_0) e^{\lambda_1 (t-t_0)} \mathbf{v}_1 + \cdots + y_{v_N}(t_0) e^{\lambda_N (t-t_0)} \mathbf{v}_N,$$

o que leva à solução na primeira componente $y_0(t) = y(t)$, e podemos deixar os coeficientes indeterminados

$$y(t) = C_1 e^{\lambda_1 t} + \cdots + C_N e^{\lambda_N t}$$

podendo ser deduzidos C_k por um sistema linear com as condições iniciais em t_0. Note-se que estes coeficientes podem ser complexos, quando as raízes são complexas.

Se houver raízes múltiplas, a matriz companheira não é diagonalizável, as repetições de λ_k levariam à mesma função $e^{\lambda_k t}$, e é necessário considerar funções independentes $t^p e^{\lambda_k t}$ com $p = 0, \ldots, m_k - 1$, onde m_k é a multiplicidade de λ_k.

Por exemplo, no caso $y'''(t) = 5y''(t) - 8y'(t) + 4y(t)$, o polinómio característico associado é

$$p(\lambda) = \lambda^3 - 5\lambda^2 + 8\lambda - 4$$

cujas raízes são $\lambda_1 = \lambda_2 = 2, \lambda_3 = 1$. A repetição de $\lambda = 2$, leva a considerar e^{2t}, mas também te^{2t},

$$y(t) = C_1 e^{2t} + C_2 t e^{2t} + C_3 e^t,$$

e se impusermos $y(0) = 0, y'(0) = 0, y''(0) = -1$, os coeficientes resultam do sistema

$$\begin{cases} C_1 + C_3 = 0 \\ 2C_1 + C_2 + C_3 = 0 \\ 4C_1 + 4C_2 + C_3 = -1 \end{cases} \quad \begin{cases} C_1 = 1 \\ C_2 = -1 \\ C_3 = -1 \end{cases}$$

obtendo-se a solução $y(t) = e^{2t} - te^{2t} - e^t$.

3.3 Métodos de Taylor e Runge-Kutta

Existem vários processos para obter métodos numéricos para a resolução de equações diferenciais ordinárias. Alguns vão usar a expansão de Taylor, que consideraremos aqui, outros vão usar a forma integral, outros podem ainda usar as fórmulas de diferenciação numérica. Começamos por ver Método de Euler, que é o mais simples, e pode ser deduzido pelos três processos mencionados.

De um modo geral, é comum a todos estes métodos considerar uma sucessão de tempos, com um espaçamento $h > 0$, partindo do tempo inicial t_0. Ou seja, definimos

$$t_k = t_0 + kh,$$

e vamos associar y_k ao valor obtido para $y(t_k)$. Procurando atingir um instante $T = t_n$, ao fim de n passos, consideramos

$$h = \frac{t_n - t_0}{n}.$$

O valor y_0 é exacto, pois assumimos que a condição inicial é exacta, mas logo y_1 depende do método e não coincide exactamente com $y(t_1)$. Os métodos unipasso usam o valor obtido para y_k para definir y_{k+1}, o que significa que como y_k não será igual a $y(t_k)$, esse erro acumula no cálculo de y_{k+1}. Mesmo para os métodos multipasso, que veremos depois, os valores de y_{k+1} assentam nos valores anteriores que trazem um erro acumulado, podendo até ficar instáveis.

Definição 6. Definimos erro local de discretização $e_{k+1} = y(t_{k+1}) - y_{k+1}$ em que y_{k+1} é calculado assumindo que $y(t_k) = y_k$. Este valor é diferente do erro global definido por $E_{k+1} = y(t_{k+1}) - y_{k+1}$, quando o y_{k+1} é baseado na iteração anterior y_k, que não é exacta, acumulando o erro local em cada passo.

Observação 37. No decurso do texto consideramos derivadas de y ou f ficando subentendido que se exige a regularidade necessária para que essas derivadas estejam definidas, evitando a redundância de estar sempre a exigir a regularidade implicitamente necessária.

3.3.1 Método de Euler

Vejamos 3 maneiras diferentes de deduzir o Método de Euler:

1º) Considerarmos a expansão de Taylor em torno de t_k :

$$y(t_{k+1}) = y(t_k + h) = y(t_k) + hy'(t_k) + \underbrace{\frac{1}{2}h^2 y''(\xi_k)}_{erro\,local O(h^2)},$$

Assumindo $y(t_k) = y_k$ temos $y'(t_k) = f(t_k, y(t_k)) = f(t_k, y_k)$, e desprezando o termo $e_k = \frac{1}{2}h^2 y''(\xi_k)$ que será o erro local, o Método de Euler resume-se à iteração:

$$\begin{cases} y_0 \\ y_{k+1} = y_k + hf(t_k, y_k) \end{cases} \tag{3.3.1}$$

2º) Consideramos pelas diferenças progressivas

$$y'(t_k) \approx \frac{y(t_{k+1}) - y(t_k)}{h},$$

substituindo $y(t_k) = y_k$ e assim $y'(t_k) = f(t_k, y_k)$, o valor y_{k+1} resulta de

$$f(t_k, y_k) = \frac{y_{k+1} - y_k}{h},$$

3º) Consideramos a formulação integral

$$y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} f(s, y(s))ds,$$

aplicando ao integral a regra do rectângulo $\int_a^b f(s)ds \approx (b-a)f(a)$, obtemos

$$y_{k+1} - y_k \approx (t_{k+1} - t_k)f(t_k, y(t_k)) = hf(t_k, y_k).$$

Exemplo 21. Aplicação do método de Euler a $y'(t) = \alpha y(t)$, com $y(0) = y_0$, onde conhecemos a solução $y(t) = y_0 e^{\alpha t}$.

Como $f(t, y) = \alpha y$, a iteração resume-se a $y_{k+1} = y_k + h\alpha y_k = (1 + \alpha h)y_k$.

Concluímos que $y_n = (1 + \alpha h)^n y_0$, e notando que $h = \frac{1}{n}t_n$,

$$y_n = (1 + \frac{\alpha t_n}{n})^n y_0 \approx e^{\alpha t_n} y_0,$$

lembrando que $e^x \approx (1 + \frac{x}{n})^n$ para n grande. Aliás, esta conhecida aproximação da exponencial é assim definida pela convergência do método de Euler.

De forma semelhante, o Método de Euler pode ser aplicado a sistemas, em cada componente y_j de $\mathbf{y}$ temos

$$y_j(t_{k+1}) = y_j(t_k) + hy_j'(t_k) + \underbrace{\frac{1}{2}h^2 y_j''(\xi_{j,k})}_{erro\,local O(h^2)},$$

e como $y_j'(t_k) = f_j(t_k, \mathbf{y}(t_k))$, obtemos vectorialmente o *Método de Euler para sistemas*:

$$\mathbf{y}_{k+1} = \mathbf{y}_{k+1} + h\mathbf{f}(t, \mathbf{y}_k) \tag{3.3.2}$$

Exemplo 22. Aplicação do método de Euler ao sistema

$$\begin{cases} y_0'(t) = y_1(t) \\ y_1'(t) = -y_0(t) \end{cases}$$

com $y_0(0) = 0$, $y_1(0) = 1$, ou o que é equivalente a

$$y''(t) = -y(t),$$

com $y(0) = 0$, $y'(0) = 1$. Temos

$$\mathbf{y}' = \left(\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \right)' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \mathbf{A}\mathbf{y} = \mathbf{f}(t, \mathbf{y})$$

o polinómio característico é $\lambda^2 + 1 = 0$, e a solução geral será $y(t) = C_1 e^{it} + C_2 e^{-it}$. Com as condições iniciais dadas obtemos $C_1 = -C_2 = \frac{1}{2i}$, e assim $y(t) = \sin(t) = y_0(t)$, sendo $y_1(t) = y'(t) = \cos(t)$.

A aplicação do Método de Euler resume-se à iteração

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{A}\mathbf{y}_k = (\mathbf{I} + h\mathbf{A})\mathbf{y}_k$$

e portanto

$$\mathbf{y}_n = (\mathbf{I} + h\mathbf{A})^n \mathbf{y}_0 = \begin{bmatrix} 1 & h \\ -h & 1 \end{bmatrix}^n \mathbf{y}_0 = \begin{bmatrix} 1 & \frac{t_n}{n} \\ -\frac{t_n}{n} & 1 \end{bmatrix}^n \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Este método dá um vector com uma aproximação do seno e coseno de t_n, mas é tão ineficaz quanto a fórmula obtida para a exponencial. De qualquer forma podemos escrever

$$\begin{bmatrix} \sin(x) \\ \cos(x) \end{bmatrix} = \lim_{n\to\infty} \begin{bmatrix} 1 & \frac{x}{n} \\ -\frac{x}{n} & 1 \end{bmatrix}^n \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Exercício 28. Considere o problema $y_0' = y_1, y_1' = y_2, y_2' = y_0$, com $(y_0, y_1, y_2)(0) = (1, 0, 0)$. Obtenha

$$y_0(t) = \frac{1}{3}(e^t + 2e^{-t/2}\cos(\frac{\sqrt{3}}{2}t)),$$

que verifica $y''' = y$, e tal como no exemplo anterior, apresente as expressões limite resultantes da aplicação do método de Euler vectorial.

3.3.2 Métodos de Taylor

Para obter métodos de ordem superior podemos considerar uma expansão de Taylor superior.

Por exemplo, considerando

$$y(t_{k+1}) = y(t_k) + y'(t_k)h + \frac{1}{2}y''(t_k)h^2 + \underbrace{\frac{1}{6}y'''(\xi_k)h^3}_{erro\,local O(h^3)}$$

e usando a expressão para $f \in C^1$,

$$y''(t) = \frac{d}{dt}f(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + f(t, y(t))\frac{\partial f}{\partial y}(t, y(t))$$

obtemos o Método de Taylor de segunda ordem:

$$y_{k+1} = y_k + hf_k + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_k, y_k) + f_k\frac{\partial f}{\partial y}(t_k, y_k)\right) \tag{3.3.3}$$

onde abreviámos usando a notação

$$f_k = f(t_k, y_k).$$

A expansão de Taylor pode ser usada para métodos de ordem superior, mas exige um cálculo explícito de derivadas parciais de f, o que nem sempre é possível ou conveniente realizar.

Surgiu por isso a ideia de substituir esse cálculo por expressões aproximadas sem comprometer que o erro local $e_k = \frac{1}{6}y'''(\xi_k)h^3$, se mantivesse na ordem $O(h^3)$.

Tratam-se dos *Métodos de Runge-Kutta*.

3.3.3 Métodos de Runge-Kutta (ordem 2)

Recorremos de novo à expansão de Taylor, mas agora de $f \in C^2$, em duas variáveis:

$$f(t_k+\varepsilon_1, y_k+\varepsilon_2) = f(t_k,y_k) + \varepsilon_1 \frac{\partial f}{\partial t}(t_k,y_k) + \varepsilon_2 \frac{\partial f}{\partial y}(t_k,y_k) + \varepsilon_1\varepsilon_2 \frac{\partial^2 f}{\partial t \partial y}(\xi_k^t, \xi_k^y), \tag{3.3.4}$$

e reparamos que escolhendo $\varepsilon_1 = \frac{h}{2}$, $\varepsilon_2 = \frac{h}{2} f_k$, obtemos

$$f(t_k+\frac{h}{2}, y_k+\frac{h}{2}f_k) = f_k + \frac{h}{2}\frac{\partial f}{\partial t}(t_k,y_k) + \frac{h}{2} f_k \frac{\partial f}{\partial y}(t_k,y_k) + \frac{h^2}{4} f_k \frac{\partial^2 f}{\partial t\partial y}(\xi_k^t,\xi_k^y),$$

e multiplicando por h reparamos que podemos substituir o segundo membro na expressão do Método Taylor de segunda ordem

$$y_{k+1}^T = y_k + hf_k + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_k,y_k) + f_k\frac{\partial f}{\partial y}(t_k,y_k)\right) = y_k + hf(t_k+\frac{h}{2}, y_k+\frac{h}{2}f_k) + \underbrace{\tfrac{h^3}{4} f_k \tfrac{\partial^2 f}{\partial t\partial y}(\xi_k^t,\xi_k^y)}_{erro\, O(h^3)}.$$

Aparece um termo adicional em $O(h^3)$, mas que não afecta a ordem de aproximação, pois o y_{k+1}^T do método de Taylor já tinha essa mesma ordem de erro de discretização.

Assim, no Método de Runge-Kutta do Ponto-Médio (de segunda ordem), consideramos

$$y_{k+1} = y_k + h\, f(t_k+\frac{h}{2}, y_k+\frac{h}{2}f_k) \tag{3.3.5}$$

e notamos que o erro de discretização resulta de

$$y_{k+1}^{RK} = y_k^T - \underbrace{\frac{h^3}{4} f_k \frac{\partial^2 f}{\partial t\partial y}(\xi_k^t,\xi_k^y)}_{erro\, O(h^3)} = y(t_{k+1}) - \underbrace{\frac{1}{6}y'''(\xi_k)h^3 - \frac{h^3}{4} f_k \frac{\partial^2 f}{\partial t\partial y}(\xi_k^t,\xi_k^y)}_{erro\, O(h^3)}$$

ou seja, o erro local de discretização para este método é

$$e_{k+1}^{RK} = y(t_{k+1}) - y_{k+1}^{RK} = h^3\left(\frac{y'''(\xi_k)}{6} + \frac{f_k}{4}\frac{\partial^2 f}{\partial t\partial y}(\xi_k^t,\xi_k^y)\right) = O(h^3),$$

admitindo que $f \in C^2(V_\epsilon)$.

Este não é o único Método de Runge-Kutta de segunda ordem que se pode obter.

Com efeito, de modo geral queremos que

$$\begin{aligned} \alpha f_k + \beta f(t_k+\varepsilon_1, y_k+\varepsilon_2) &= (\alpha+\beta)f_k + \beta\varepsilon_1\frac{\partial f}{\partial t}(t_k,y_k) + \beta\varepsilon_2\frac{\partial f}{\partial y}(t_k,y_k) \\ &= hf_k + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_k,y_k) + f_k\frac{\partial f}{\partial y}(t_k,y_k)\right), \end{aligned}$$

o que leva a um sistema subdeterminado, com 3 equações e 4 incógnitas, que deixamos em função de $\theta \neq 0$:

$$\begin{cases} \alpha+\beta = h \\ \beta\varepsilon_1 = \frac{1}{2}h^2 \\ \beta\varepsilon_2 = \frac{1}{2}h^2 f_k \end{cases} \qquad \begin{cases} \alpha = h - \frac{h}{2\theta}, \quad \beta = \frac{h}{2\theta}, \\ \varepsilon_1 = \theta h \\ \varepsilon_2 = \theta h f_k \end{cases}$$

Portanto, podemos definir para cada $\theta \neq 0$, uma *Regra de Runge-Kutta de ordem 2*:

$$y_{k+1} = y_k + (1 - \frac{1}{2\theta})hf_k + \frac{h}{2\theta} f\left(t_k + \theta h, y_k + \theta h f_k\right) \tag{3.3.6}$$

– A regra RK do ponto-médio, que já vimos, é obtida escolhendo $\theta = \frac{1}{2}$.
– A regra RK de Heun (ou Euler modificado) é obtida com $\theta = 1$, ficando

$$y_{k+1} = y_k + \frac{1}{2}hf_k + \frac{1}{2}h\, f\left(t_k + h, y_k + hf_k\right) \tag{3.3.7}$$

– Uma outra possibilidade menos usada é $\theta = \frac{2}{3}$.

Observação 38. Estes métodos podem ainda ser obtidos através da formulação integral, entendendo que em

$$y_{k+1} - y_k = \int_{t_k}^{t_{k+1}} f(s, y(s))ds$$

podemos considerar a regra de quadratura

$$\int_{t_k}^{t_{k+1}} g(s)ds \approx w_1 g(t_k) + w_2 g(t_k + \theta h),$$

que, para $w_1 = (1 - \frac{1}{2\theta})h$, $w_2 = \frac{h}{2\theta}$, tem pelo menos grau 1. Ou seja, é válida a aproximação

$$y_{k+1} - y_k = \int_{t_k}^{t_{k+1}} f(s, y(s))ds \approx w_1 f(t_k, y_k) + w_2 f(t_k + \theta h, y(t_k + \theta h))$$

a dedução fica completa considerando a aproximação de $y(t_k + \theta h)$ pelo método de Euler, pois $y(t_k + \theta h) \approx y_k + \theta h f_k$.

No caso em que $\theta = \frac{1}{2}$ estamos a usar a regra de integração do ponto-médio (o que justifica o nome), e quando $\theta = 1$ estamos a usar a regra dos trapézios.

3.3.4 Métodos de Runge-Kutta (ordem 4)

Os métodos de Runge-Kutta de ordem superior recorrem à mesma ideia, substituir as derivadas no Método de Taylor por avaliações em pontos adequados. Assim, a expressão do Método de Taylor de ordem 3 poderá ser substituída por uma avaliação em 3 pontos adequados, e da mesma forma os Métodos de Runge-Kutta de ordem 4 vão aproximar o Método de Taylor de Ordem 4, usando 4 pontos adequados, sem afectar a ordem do erro de discretização, que será $O(h^5)$.

Evitando a extensa dedução, colocamos aqui o algoritmo para o *Método de Runge-Kutta de ordem 4* mais utilizado

$$\begin{aligned} y_{k+1} &= y_k + \frac{h}{6}(F_1 + 2F_2 + 2F_3 + F_4) \\ &\quad \begin{cases} F_1 = f_k, & F_2 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}F_1) \\ F_3 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}F_2), & F_4 = hf(t_k + h, y_k + hF_3) \end{cases} \end{aligned}$$

Observação 39. Este é o método mais usado, porque permite ordem 4 com quatro avaliações de f em quatro passos. Para ordem 5 Butcher mostrou que não seria possível fazê-lo com cinco avaliações em cinco passos, pelo que se perderia eficiência face ao de ordem 4.

Como podemos observar pela estrutura deste algoritmo, ou de outros métodos de Runge-Kutta, é perfeitamente imediata a sua aplicação a sistemas de equações, passando a sua forma à notação vectorial.

Observação 40. (Tabelas de Butcher) Para efeitos de "economia de espaço" é habitual escrever abreviadamente os Métodos de Runge-Kutta sob a forma de tabelas (chamadas de Butcher).

Se o método se dividir em m passos, ficando na forma

$$y_{k+1} = y_k + h(\beta_1 F_1 + \ldots + \beta_m F_m),$$

com

$$F_m = f(t_k + \tau_m h,\, y_k + \alpha_{m1} h F_1 + \cdots + \alpha_{m,m-1} h F_{m-1})$$

a tabela de Butcher será

τ_1	0	$\cdots$	0	0
τ_2	$\alpha_{2,1}$	$\ddots$	$\ddots$	0
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
τ_m	$\alpha_{m,1}$	$\cdots$	$\alpha_{m,1}$	0
	β_1	$\cdots$	β_{m-1}	β_m

.

Terá zeros na parte triangular superior no caso dos métodos explícitos (também há métodos RK implícitos).

Para o Método RK-4 a tabela de Butcher é

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

,

para um RK-3 é

0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0
1	-1	2	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

, e para RK-2 é genericamente

0	0	0
θ	θ	0
	$1-\frac{1}{2\theta}$	$\frac{1}{2\theta}$

.

3.3.5 Espaçamento adaptativo

Do ponto de vista computacional, é importante poder variar o espaçamento h de forma a controlar o erro.

Querendo manter um erro local aproximado a um $\varepsilon \approx 0$, uma técnica é avaliar o erro com h usando um espaçamento $h/2$. Comparamos valor y_{k+1} obtido com h, com um valor melhor $y_{k+2\frac{1}{2}}$que resulta de aplicar duas iterações com $h/2$. O erro local será aproximadamente $\tilde{e}_k = y_{k+2\frac{1}{2}} - y_{k+1}$ e verificamos se $|\tilde{e}_k| \approx \varepsilon$. Podemos mudar o espaçamento em função dessa diferença, automaticamente, considerando um novo $\tilde{h} = \min\{h\frac{\varepsilon}{|\tilde{e}_k|}, h_{\max}\}$ onde $h_{\max}$ aparece apenas para que o espaçamento não seja demasiado grande.

Outra possibilidade consiste em usar um método de ordem superior para avaliar y_{k+1}, o que é normalmente melhor que duas iterações do mesmo método. De resto, o procedimento é semelhante. Um método especialmente eficaz é o de Runge-Kutta-Fehlberg, que combina o método de ordem 4 com uma previsão de ordem 5, conseguindo usar os mesmos pontos para avaliação da função f.

3.4 Ordem de consistência e convergência

Consideramos um *método unipasso* dado pela expressão genérica

$$y_{k+1} = y_k + h\Phi(t_k, y_k), \tag{3.4.1}$$

onde Φ depende do método, e assumiremos Φ ser Lipschitziana na $2^{\underline{a}}$ variável.

Exemplo 23. No Método de Euler temos simplesmente $\Phi = f$, e Φ é Lipschitz quando f for.

No Método de Runge-Kutta do ponto-médio temos $\Phi(t, y) = f(t+\frac{h}{2}, y+\frac{h}{2}f(t, y))$, e também podemos ver que Φ é Lipschitz quando f for:

$$\begin{aligned} |\Phi(t,y) - \Phi(t,x)| &= |f(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)) - f(t+\frac{h}{2}, x+\frac{h}{2}f(t,x))| \\ &\leq L_f|y+\frac{h}{2}f(t,y) - (x+\frac{h}{2}f(t,x))| \\ &\leq L_f|y-x| + \frac{h}{2}L_f|f(t,y) - f(t,x)| \leq (L_f + \frac{h}{2}L_f^2)|y-x| \end{aligned}$$

em que L_f é a constante de Lispchitz de f, tendo-se para este Φ a constante $L_\Phi = L_f + \frac{h}{2}L_f^2$.

De forma idêntica, o Método de Heun pode escrever-se com $\Phi(t, y) = \frac{1}{2}f(t, y) + \frac{1}{2}f(t+h, y+hf(t, y))$, e igualmente se verifica que $L_\Phi = L_f + \frac{h}{2}L_f^2$, sendo Φ Lipschitz quando f for.

Definição 7. Dizemos que o método unipasso é *consistente de ordem* r se

$$\frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k)) = O(h^r). \tag{3.4.2}$$

Diz-se consistente desde que a diferença seja um $o(1)$, em particular, basta qualquer $r > 0$.

Esta definição equivale a dizer que o método tem consistência de ordem p se o erro de discretização local for $O(h^{r+1})$, porque se assumirmos que $y_k = y(t_k)$ então $y_{k+1} = y(t_k) + h\Phi(t_k, y(t_k))$

$$e_{k+1} = y(t_{k+1}) - y_{k+1} = y(t_{k+1}) - (y(t_k) + h\Phi(t_k, y(t_k))) = O(h^{r+1}),$$

e a divisão por h dá exactamente a definição de consistência de ordem r.

Observação 41. Notamos assim que os Métodos de Taylor ou Runge-Kutta com erro local em $O(h^{r+1})$ têm ordem de consistência r. O método de Euler tem assim ordem de consistência 1, e a expansão de Taylor de ordem r dará um método de ordem de consistência r, que não será alterada pela aproximação de Runge-Kutta, pois o erro local mantém-se em $O(h^{r+1})$.

Definição 8. Dizemos que o método unipasso tem *ordem de convergência* r se

$$E_n = y(t_n) - y_n = O(h^r), \tag{3.4.3}$$

em que t_n é um instante fixo e $h = \frac{1}{n}(t_n - t_0)$. Será *convergente* desde que $E_n = o(1)$.

Teorema 21. *Um método unipasso em que Φ é Lipschitziana (na $2^{\underline{a}}$ variável), com ordem de consistência r tem ordem de convergência r.*

Demonstração. Distinguimos o valor $y_{k+1} = y_k + h\Phi(t_k, y_k)$ que assume um valor anterior y_k inexacto, de $Y_{k+1} = y(t_k) + h\Phi(t_k, y(t_k))$, obtido assumindo que o valor anterior seria exacto, ou seja, $y_k = y(t_k)$.

Assim, separamos o erro global em duas partes, pela desigualdade triangular

$$|E_{k+1}| = |y(t_{k+1}) - y_{k+1}| \leq |y(t_{k+1}) - Y_{k+1}| + |Y_{k+1} - y_{k+1}| \tag{3.4.4}$$

Pela definição de ordem de consistência, o erro local será $O(h^{r+1})$, ou seja

$$|e_{k+1}| = |y(t_{k+1}) - Y_{k+1}| \leq Ch^{r+1},$$

por outro lado, vemos que

$$\begin{aligned} |Y_{k+1} - y_{k+1}| &= |y(t_k) + h\Phi(t_k, y(t_k)) - (y_k + h\Phi(t_k, y_k))| \\ &\leq |y(t_k) - y_k| + h\,|\Phi(t_k, y(t_k)) - \Phi(t_k, y_k)| \\ &\leq |E_k| + hL_\Phi\,|y(t_k) - y_k| = (1 + hL_\Phi)|E_k| \end{aligned}$$

onde usámos a hipótese de Φ ser Lipschitziana na $2^{\underline{a}}$ variável. Substituindo em (3.4.4), obtemos

$$|E_{k+1}| \leq Ch^{r+1} + (1 + hL_\Phi)|E_k|,$$

ou seja, temos uma relação recursiva entre $|E_{k+1}|$ e $|E_k|$. Ora, não é difícil mostrar que se

$$c_{k+1} \leq A + Bc_k$$

então $c_n \leq A(1 + B + \ldots + B^{n-1}) + B^n c_0$. Com efeito, é válido para $n = 1$, e por indução, sendo válido para $n - 1$, obtemos

$$c_n \leq A + Bc_{n-1} = A + B(A(1 + B + \ldots + B^{n-2}) + B^{n-1}c_0) = A(1 + B + \ldots + B^{n-1}) + B^n c_0.$$

Podemos ainda simplicar a expressão escrevendo

$$c_n \leq A\frac{B^n - 1}{B - 1} + B^n c_0$$

Aplicando ao nosso caso, com $c_k = |E_k|$, $A = Ch^{r+1}$, $B = 1 + hL_\Phi$, tem-se

$$|E_n| \leq Ch^{r+1}\frac{B^n - 1}{B - 1} + B^n|E_0|.$$

Não havendo erro inicial, $E_0 = 0$, temos também $B - 1 = hL_\Phi$, e podemos ainda usar $1 + x \leq e^x$ para simplificar, pois $B = 1 + hL_\Phi \leq e^{hL_\Phi}$, e assim $B^n \leq e^{nhL_\Phi} = e^{(t_n - t_0)L_\Phi}$. Ou seja,

$$|E_n| \leq Ch^{r+1}\frac{e^{(t_n - t_0)L_\Phi} - 1}{hL_\Phi} = Ch^r\frac{e^{(t_n - t_0)L_\Phi} - 1}{L_\Phi} = Kh^r,$$

onde K é uma constante que não depende de h, concluindo-se a ordem de convergência ser r. □

Observação 42. No caso do Método de Euler, como temos $|e_k| = \frac{1}{2}h^2 y''(\xi_k) \leq \frac{1}{2}||y''||_\infty h^2$, temos o valor $C = \frac{1}{2}||y''||_\infty$, e como $L_\Phi = L_f = ||\frac{\partial f}{\partial y}||_\infty$, a majoração do erro global é

$$|E_n| \leq \frac{1}{2}||y''||_\infty\frac{e^{(t_n - t_0)L_f} - 1}{L_f}h.$$

Exemplo 24. Em função de α, estudar a ordem de convergência do método $y_{k+1} = y_k + hf(t_k + \alpha, y_k + \alpha f_k)$.

Como se trata de um método unipasso, pelo teorema, basta analisar a ordem de consistência. Notamos que quando $\alpha = 0$ é o método de Euler, e quando $\alpha = h/2$, trata-se do método RK do ponto médio.

Aqui $\Phi(t, y) = f(t+\alpha, y+\alpha f(t, y))$, que será Lipschitziana com $L_\Phi = L_f + \alpha L_f^2$ (ver exemplo de RK-ponto-médio).

Analisando

$$\frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k)) \quad = \quad y'(t_k) + \frac{h}{2}y''(t_k) + \frac{h^2}{6}y'''(\xi_k) - f(t_k + \alpha, y_k + \alpha f(t_k, y_k))$$

lembramos que

$$y'(t_k) + \frac{h}{2}y''(t_k) = f(t_k, y(t_k)) + \frac{h}{2}\frac{\partial f}{\partial t}(t_k, y(t_k)) + \frac{h}{2}f(t_k, y(t_k))\frac{\partial f}{\partial y}(t_k, y(t_k))$$

e que

$$f(t_k + \alpha, y_k + \alpha f(t_k, y_k)) = f(t_k, y(t_k)) + \alpha\frac{\partial f}{\partial t}(t_k, y(t_k)) + \alpha f(t_k, y(t_k))\frac{\partial f}{\partial y}(t_k, y(t_k)) + O(\alpha^2),$$

portanto

$$y'(t_k) + \tfrac{h}{2}y''(t_k) - f\big(t_k+\alpha,\, y_k+\alpha f(t_k,y_k)\big) = (\alpha - \tfrac{h}{2})\Big(\tfrac{\partial f}{\partial t}\big(t_k,y(t_k)\big) + f\big(t_k,y(t_k)\big)\tfrac{\partial f}{\partial y}\big(t_k,y(t_k)\big)\Big) + O(\alpha^2).$$

Assim,

$$\tfrac{y(t_{k+1})-y(t_k)}{h} - \Phi(t_k,y(t_k)) = (\alpha - \tfrac{h}{2})\Big(\tfrac{\partial f}{\partial t}\big(t_k,y(t_k)\big) + f\big(t_k,y(t_k)\big)\tfrac{\partial f}{\partial y}\big(t_k,y(t_k)\big)\Big) + O(\alpha^2) + \tfrac{h^2}{6}y'''(\xi_k),$$

e esta expressão será um $O(h^2)$ se $\alpha = \frac{h}{2}$, e aí terá ordem de consistência e convergência 2 (caso do RK-ponto-médio). Caso contrário, se $\alpha = O(h^p)$, com $p > 0$, terá ordem de convergência 1, se $p \geq 1$, ou ordem de convergência $p < 1$. Por exemplo nos casos $\alpha = 0$ (Euler), ou $\alpha = h$, terá convergência de ordem 1.

Observação 43. Conforme já referimos a propósito dos Métodos de Runge-Kutta, para além da ordem de convergência, há outros aspectos importantes na eficácia dos métodos numéricos, sendo um deles o número de operações envolvido em cada iteração. A função f pode ser complicada, não ser dada explicitamente, e a sua pode ser avaliação morosa. Assim é conveniente não envolver derivadas de f nem implicar muitas avaliações de f. Basta ver que duas iterações do Método de Euler levariam ao método $y_{k+1} = y_k + \frac{h}{2}f_k + \frac{h}{2}f(t_k, y_k + \frac{h}{2}f_k)$ que aparentava ser 2 vezes mais rápido do que o Método de Euler, mas mantinha-se de ordem 1 com duas avaliações, ao contrário do Método de Heun que é muito semelhante, mas com as mesmas duas avaliações permite ordem 2.

3.5 Métodos implícitos

Os métodos que vimos até aqui são considerados explícitos, já que o valor y_{k+1} é obtido explicitamente na expressão $y_{k+1} = y_k + h\Phi(t_k, y_k)$. Há ainda uma outra classe de métodos, designados implícitos, em que o valor y_{k+1} se encontra definido na forma de uma equação em que y_{k+1} é uma incógnita.

Definição 9. Os métodos unipasso implícitos são da forma

$$y_{k+1} = y_k + h\Phi(t_k, y_k, y_{k+1}) \tag{3.5.1}$$

em que Φ é ainda Lipschitziana (na $2^{\underline{a}}$ e $3^{\underline{a}}$ componente).

Observação 44. A definição de consistência para os métodos implícitos é análoga à dos métodos explícitos, acrescentando a componente de Φ em função de y_{k+1}. Uma simples adaptação permite ainda obter o Teorema que garante a mesma ordem de consistência e convergência.

Exemplo 25. O exemplo mais simples é o Método de Euler Implícito

$$y_{k+1} = y_k + hf(t_{k+1}, y_{k+1}) \tag{3.5.2}$$

onde $\Phi(t, y, x) = f(t+h, x)$. Este método resulta de considerar uma aproximação por diferenças regressivas (em vez de progressivas, como acontecia no Método de Euler explícito), ou ainda pela expansão de Taylor em t_{k+1} (ao invés de t_k), ou ainda na forma integral usando a regra do rectângulo no ponto b (ao invés de a), ou seja $\int_a^b f(s)ds \approx (b-a)f(b)$. Notamos que mudamos a aproximação, mas o erro local continua a ter a mesma ordem, e portanto o Método de Euler Implícito tem ordem 1, de consistência e convergência.

Outro método, já com ordem de convergência 2, é o Método dos Trapézios Implícito

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})) \tag{3.5.3}$$

em que $\Phi(t, y, x) = \frac{1}{2}(f(t, y)+f(t+h, x))$. Este método resulta de aplicar a Regra dos Trapézios à forma integral

$$y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))ds = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})) + O(h^3),$$

pois o erro local $O(h^3)$ permite uma ordem de consistência 2.

Observação 45. A aplicação do Método de Euler Implícito a $y'(t) = \alpha y(t)$, resulta em

$$y_{k+1} = y_k + h\alpha y_{k+1}$$

de onde obtemos $(1 - \alpha h)y_{k+1} = y_k$, e portanto recursivamente (escolhendo $h \neq \alpha^{-1}$)

$$y_n = \frac{1}{(1 - \alpha h)^n} y_0.$$

O valor dado pelo Método de Euler implícito dá um resultado substancialmente diferente do obtido pelo Método de Euler explícito, onde vimos $y_n = (1 + \alpha h)^n y_0$.

Para $\alpha < 0$ e $y_0 = 1$, a solução exacta $y(t) = e^{\alpha t}$ deveria tender para zero, quando t fosse suficientemente grande.

Porém, suponhamos que o valor de $-\alpha > 0$ é alto, de tal forma que o passo $h > 0$ não é suficientemente pequeno, ou seja

$$h > \frac{2}{-\alpha} \Leftrightarrow \alpha h < -2.$$

Pelo método explícito, como $1+\alpha h < -1$ teremos $y_n = (1+h\alpha)^n$ oscilando, tendendo para infinito, sem se aproximar de zero.

Ao contrário, usando o método implícito, como $\alpha < 0$, temos $1 - h\alpha > 1$, e obtemos $y_n = \frac{1}{(1-h\alpha)^n} \to 0$, estando de acordo com a solução.

Ou seja, se a solução fosse $y(t) = e^{-400t}$, então um passo pequeno como $h = 0.005$ levaria à aproximação $(-1)^n$, muito afastada da solução $e^{-400t} \approx 0$, (com $t > 1$), se o método fosse explícito, mas permitiria uma aproximação $\frac{1}{(-3)^n} \approx 0$, se o método fosse implícito.

3.5.1 Noção de A - estabilidade

A observação anterior faz notar que, quando $|\alpha|$ é grande, poderá ser necessário um h demasiado pequeno para obter uma aproximação satisfatória, quando se usam métodos explícitos. Não está em causa a convergência do método, que foi provada. Mas essa convergência assume que o h pode ser tão pequeno quanto necessário, e na prática isso pode implicar um número exagerado de iterações. Quando queremos usar um h não exageradamente pequeno pode ser conveniente usar métodos implícitos, quando a equação diferencial é considerada "rígida". De um modo geral as equações dizem-se rígidas (stiff), se tiverem um valor da constante de Lispchitz $L_f = ||\frac{\partial f}{\partial y}||_\infty$ elevado, o que acontece para $y(t) = e^{\alpha t}$ quando $L_f = |\alpha|$ é grande.

Para avaliar a robustez dos métodos numéricos a este tipo de problemas, introduziu-se a noção de A-estabilidade.

Definição 10. A região de A-estabilidade (estabilidade absoluta) de um método$\mathcal{M}$, é definida por

$$\mathcal{A}_{\mathcal{M}} = \{\alpha \in \mathbb{C} : \quad \sup_n(y_n^\alpha) < \infty\},$$

onde (y_n^α) é a sucessão definida pelo método $\mathcal{M}$, quando aplicada a $y'(t) = \alpha y(t)$, com $h = 1$, $y_0 = 1$.

Habitualmente o método é considerado A-estável se a região de estabilidade inclui $\mathbb{C}_- = \{\alpha : Re(\alpha) \leq 0\}$. Uma característica dos métodos implícitos é a de terem uma região de A-estabilidade ilimitada, enquanto os métodos explícitos têm essa região confinada a um pequeno domínio limitado, não podendo ser A-estáveis (teorema de Dahlquist). Vejamos alguns exemplos.

Exemplo 26. Considerando a expressão do Método de Euler explícito $y_n = (1+\alpha h)^n y_0$, vemos que quando $h = 1$, $y_0 = 1$, a sucessão fica limitada apenas se $|1 + \alpha| \leq 1$, logo

$$\mathcal{A}_{Euler-Explicito} = \{\alpha \in \mathbb{C} : |1 + \alpha| \leq 1\} = \bar{B}(-1, 1),$$

ou seja a sua região de A-estabilidade é a bola fechada $\bar{B}(-1, 1)$.

No caso do Método de Euler Implícito, como $y_n = (1 - \alpha h)^{-n} y_0$, obtemos

$$\mathcal{A}_{Euler-Implicito} = \{\alpha \in \mathbb{C} : |1 - \alpha| \geq 1\} = \mathbb{C}\backslash B(1, 1),$$

ou seja é a região ilimitada, complementar à bola $B(1, 1)$.

Vejamos agora o caso do Método de Runge-Kutta do Ponto-Médio, de ordem 2. Quando $f(t, y) = \alpha y$, temos:

$$y_{k+1} = y_k + hf\left(t_k + \frac{h}{2}, y_k + \frac{h}{2}f_k\right) = y_k + h\alpha\left(y_k + \frac{h}{2}(\alpha y_k)\right) = (1 + h\alpha + \frac{h^2\alpha^2}{2})y_k,$$

e com $h = 1, y_0 = 1$, obtemos recursivamente $y_n = (1 + \alpha + \frac{\alpha^2}{2})^n$, e assim

$$\mathcal{A}_{RK2-pt.medio} = \{\alpha \in \mathbb{C} : |1 + \alpha + \alpha^2/2| \leq 1\},$$

que se trata de uma região limitada, contida em $[-2, 0] \times [-2, 2]$.

Finalmente, pelo Método dos Trapézios implícito obtemos $y_{k+1} = y_k + \frac{h}{2}(\alpha y_k + \alpha y_{k+1})$, de onde resulta

$$y_{k+1} = \frac{1 + \frac{\alpha h}{2}}{1 - \frac{\alpha h}{2}} y_k \implies \mathcal{A}_{Trap} = \{\alpha \in \mathbb{C} : |1 + \alpha/2| \leq |1 - \alpha/2|\} = \{\alpha : Re(\alpha) \leq 0\} = \mathbb{C}_-.$$

3.5.2 Implementação de um Método Implícito

Nos simples exemplos que vimos até aqui foi possível obter uma expressão de y_{k+1}, mas isso não é possível na maioria das vezes. Por exemplo, no Método de Euler Implícito bastará que a equação $x = y_k + hf(t_{k+1}, x)$ não tenha solução algébrica para x, o que ocorre para a generalidade de funções f.

Assim, de forma geral, consideramos a função

$$g(x) = y_k + h\Phi(t_k, y_k, x) \quad \implies y_{k+1} = g(y_{k+1}),$$

e assim y_{k+1} será o ponto fixo, $x = g(x)$ mais próximo de y_k.

A equação $x = g(x)$ não tendo solução algébrica, deve ser resolvida por métodos numéricos, por exemplo, o Método do Ponto Fixo ou o Método de Newton.

Aplicando o Método do Ponto Fixo, definimos uma sucessão (x_m) com

$$x_0 = y_k, \quad x_{m+1} = g(x_m).$$

A condição de convergência local da sucessão do ponto fixo é definida pela contractividade, ou seja $|g'(x)| \leq K < 1$ para $x \approx y_k$.

Exemplo 27. No caso do Método de Euler Implícito com implementação da iteração do Ponto Fixo, obtemos

$$x_0 = y_k; \quad x_{m+1} = g(x_m) = y_k + hf(t_{k+1}, x_m)$$

e a condição de contractividade local será $h < \frac{1}{L_f}$, porque: $|g'(x)| = h|\frac{\partial f}{\partial y}(t + h, x)| \leq hL_f = K < 1$.

No caso do Método dos Trapézios Implícito, a implementação da iteração do Ponto Fixo, leva a

$$x_0 = y_k; \quad x_{m+1} = g(x_m) = y_k + \frac{h}{2}(f_k + f(t_k + h, x_m))$$

e a condição de contractividade local será $h < \frac{2}{L_f}$, porque: $|g'(x)| = \frac{h}{2}|\frac{\partial f}{\partial y}(t + h, x)| \leq \frac{h}{2}L_f = K < 1$.

Se aplicarmos o Método de Newton aos métodos implícitos, a iteração ficaria $x_0 = y_k$,

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

Cada iteração do Método de Newton exigiria um cálculo de f, mas também da derivada, não sendo muito utilizado.

Observação 46. Uma questão que se coloca nos métodos implícitos é o número de operações considerado, já que cada cálculo de x_m exigirá uma nova avaliação de f. Por isso não se devem calcular mais do que algumas iterações, devendo x_2 ou x_3 já ser suficientes para o efeito. Para evitar estes cálculos adicionais, procura-se usar uma melhor aproximação inicial x_0, numa técnica que se chama *preditor-corrector*.

3.5.3 Métodos Preditor-Corrector

Para melhorar a eficácia dos métodos implícitos, procura-se inicializar a iteração do ponto fixo com um valor x_0 que resulta da aplicação de um método explícito. Temos assim um par preditor-corrector, preditor será o método explícito que inicializa a iteração, e corrector será o método implícito usado.

Sendo $y_{k+1} = y_k + h\Phi_C(t_k, y_k, y_{k+1})$ o método implícito (corrector), e $y_{k+1} = y_k + h\Phi_P(t_k, y_k)$ o método explícito (preditor), consideramos a iteração do ponto fixo:

$$\begin{cases} x_0 = y_k + h\Phi_P(t_k, y_k) \\ x_{m+1} = y_k + h\Phi_C(t_k, y_k, x_m) \end{cases}$$

Exemplo 28. (Método de Heun surgindo do par Euler-Trapézios como preditor-corrector).

Considerando o Método dos Trapézios Implícito inicializado por Euler explícito:

$$\begin{cases} x_0 = y_k + hf(t_k, y_k) \\ x_{m+1} = y_k + \frac{h}{2}(f_k + f(t_{k+1}, x_m)\,) \end{cases}$$

Por substituição imediata obtemos

$$x_1 = y_k + \frac{h}{2}(f_k + f(t_{k+1}, y_k + hf_k)\,)$$

e assim se considerarmos $y_{k+1} = x_1$ obtemos a expressão do Método de Heun logo na primeira iterada. Podemos prosseguir com a iteração do ponto fixo, por motivo da A-estabilidade, mas sob a perspectiva da ordem de convergência, este método não passará de ordem 2.

3.6 Métodos Multipasso

A ideia dos métodos multipasso consiste em usar não apenas o valor y_k, mas também os anteriores $y_{k-1}, y_{k-2}, \ldots$ para formar o novo valor y_{k+1}. Têm a grande vantagem de usando os valores já calculados de f, nas iterações anteriores, apenas precisarem duma nova avaliação de f em cada iteração.

Definição 11. Os métodos multipasso (lineares) explícitos são da forma

$$y_{k+1} = \sum_{m=0}^{p-1} \alpha_{-m} y_{k-m} + h\Phi(t_k, y_k, \cdots, y_{k-p+1}) \tag{3.6.1}$$

em que p é o passo (quando $p = 1$, $\alpha_0 = 1$, temos os anteriores métodos unipasso). Φ considera-se Lipschitziana (a partir da 2ª componente):

$$|\Phi(t, y_0, \cdots, y_{-p+1}) - \Phi(t, x_0, \cdots, x_{-p+1})| \leq L_\Phi \sum_{m=0}^{p-1} |y_{-m} - x_{-m}|.$$

Notamos desde já que a expressão dos métodos multipasso ao implicar o conhecimento dos valores anteriores exige uma inicialização apropriada.

Num método de passo p é exigido o conhecimento inicial de $y_0, y_1, \ldots, y_{p-1}$. Essa inicialização deve ser efectuada por um outro método com a mesma ordem de consistência. Por exemplo, podemos inicializar com um método de Runge-Kutta.

3.6.1 Métodos de Adams-Bashforth

Os primeiros métodos multipasso introduzidos (no final do Séc. XIX) foram os Métodos de Adams, que na sua forma explícita são designados como Adams-Bashforth. Podemos deduzi-los através da forma integral

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))ds,$$

aplicando uma regra de integração que usa nós fora do intervalo de integração:

$$I(g) = \int_{t_k}^{t_{k+1}} g(s)ds \approx Q(g) = w_0 g(t_k) + \ldots + w_{-p+1} g(t_{k-p+1}).$$

Procuramos os pesos $w_0, \ldots, w_{-p+1}$ de forma a que a regra tenha pelo menos grau $p-1$. Isso pode ser obtido pelo método dos coeficientes indeterminados, ou ainda lembrando a interpolação de Lagrange, por

$$w_m = \int_{t_k}^{t_{k+1}} L_m(s)ds$$

em que L_m são os polinómios base de Lagrange. Basta lembrar que podemos escrever qualquer polinómio q de grau $p-1$ através do interpolador $q(s) = \sum_{m=0}^{p-1} q(t_{k-m}) L_m(s)$, e assim

$$I(q) = \int_{t_k}^{t_{k+1}} q(s)ds = \int_{t_k}^{t_{k+1}} \sum_{m=0}^{p-1} q(t_{k-m}) L_m(s)ds = \sum_{m=0}^{p-1} q(t_{k-m}) \underbrace{\int_{t_k}^{t_{k+1}} L_m(s)ds}_{w_m} = Q(q).$$

Usando a nova regra de quadratura, temos

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(s, y(s))ds \approx y(t_k) + w_0 f(t_k, y(t_k)) + \ldots + w_{-p+1} f(t_{k-p+1}, y(t_{k-p+1}))$$

de onde surgem os Métodos de Adams-Bashforth de passo p :

$$y_{k+1} = y_k + w_0 f_k + \ldots + w_{-p+1} f_{k-p+1} \tag{3.6.2}$$

em que os valores w_m são os obtidos na regra de integração.

Exemplo 29. Método de Adams-Bashforth de passo 2.

Apenas temos que determinar os valores de w_0 e w_{-1} de forma a que $I(1) = Q(1)$, $I(s) = Q(s)$. Isso pode ser obtido pelo sistema do método dos coeficientes indeterminados, ou ainda com

$$w_0 = \int_{t_k}^{t_{k+1}} L_0(s)ds = \int_{t_k}^{t_{k+1}} \frac{s - t_{k-1}}{t_k - t_{k-1}} ds = \left[\frac{(s - t_{k-1})^2}{2h} \right]_{t_k}^{t_{k+1}} = \frac{4h^2}{2h} - \frac{h^2}{2h} = \frac{3}{2}h.$$

$$w_{-1} = \int_{t_k}^{t_{k+1}} L_1(s)ds = \int_{t_k}^{t_{k+1}} \frac{s-t_k}{t_{k-1}-t_k}ds = \left[\frac{(s-t_k)^2}{-2h}\right]_{t_k}^{t_{k+1}} = -\frac{h}{2}.$$

Conclui-se assim que o Método de Adams-Bashforth de passo 2 é dado por

$$y_{k+1} = y_k + \tfrac{3h}{2}f_k - \tfrac{h}{2}f_{k-1}, \tag{3.6.3}$$

e como referimos, o método deverá ser inicializado para obter y_1, por exemplo pelo Método de Heun, ou outro de ordem 2, já que como iremos ver este método de passo duplo tem ordem 2.

3.6.2 Métodos de Adams-Moulton

Os métodos de Adams-Moulton são multipasso implícitos, considerando também t_{k+1} como nó de integração. Ou seja, passamos a ter mais um ponto de integração

$$Q(g) = w_1 g(t_{k+1}) + w_0 g(t_k) + \ldots + w_{-p+1} g(t_{k-p+1}).$$

Por isso a forma geral dos Métodos de Adams-Moulton de passo p passa a ser

$$y_{k+1} = y_k + w_1 f_{k+1} + w_0 f_k + \ldots + w_{-p+1} f_{k-p+1} \tag{3.6.4}$$

em que os valores w_m são os obtidos na nova regra de integração, de forma similar ao caso anterior.

Exemplo 30. Método de Adams-Moulton de passo 2.

Temos que determinar agora os valores de w_1, w_0 e w_{-1} de forma a que $I(1) = Q(1)$, $I(s) = Q(s)$, $I(s^2) = Q(s^2)$. Notamos que estamos a exigir uma regra de grau superior, porque dispomos de mais um nó de integração. Por consequência este método terá ordem superior ao de Adams-Bashforth de passo 2. Para obter os pesos, usamos ainda os polinómios de Lagrange:

$$w_0 = \int_{t_k}^{t_{k+1}} \frac{s-t_k}{t_{k+1}-t_k}\frac{s-t_{k-1}}{t_{k+1}-t_{k-1}}ds = \frac{5}{12}h,\ w_1 = \int_{t_k}^{t_{k+1}} \frac{s-t_{k+1}}{t_k-t_{k+1}}\frac{s-t_{k-1}}{t_k-t_{k-1}}ds = \frac{2}{3}h,$$

$$w_{-1} = \int_{t_k}^{t_{k+1}} \frac{s-t_{k+1}}{t_{k-1}-t_{k+1}}\frac{s-t_k}{t_{k-1}-t_k}ds = -\frac{1}{12}h,$$

Conclui-se assim que o Método de Adams-Moulton de passo 2 é dado por

$$y_{k+1} = y_k + \tfrac{5h}{12}f_{k+1} + \tfrac{2h}{3}f_k - \tfrac{h}{12}f_{k-1}, \tag{3.6.5}$$

e também deverá ser inicializado para obter y_1, mas neste caso por um método de ordem 3, já que como iremos ver este método de passo duplo implícito tem ordem 3.

3.6.3 Consistência dos métodos multipasso

A consistência dos métodos multipasso é definida de forma semelhante à dos unipasso, e incluímos já o caso implícito.

Definição 12. Dizemos que um método multipasso tem consistência de ordem r se verificar

$$\frac{1}{h}\left(y(t_{k+1}) - \sum_{m=0}^{p-1} \alpha_{-m} y(t_{k-m})\right) - \Phi(t_k, y(t_{k+1}), y(t_k), \ldots, y(t_{k-p+1})) = O(h^r),$$

dizendo-se consistente desde que a diferença seja $o(1)$, o que se verifica para qualquer $r > 0$.

Proposição 15. *Os métodos de Adams-Bashforth de passo p têm ordem de consistência p, e os de Adams-Moulton têm consistência $p+1$.*

Demonstração. Nos métodos de Adams-Bashforth $\alpha_0 = 1$, e $\alpha_{-m} = 0$ para os restantes.

Ora, como

$$\begin{aligned} h\Phi(t_k, y(t_k), \ldots, y(t_{k-p+1})) &= w_0 f(t_k, y(t_k)) + \ldots + w_{-p+1} f(t_{k-p+1}, y(t_{k-p+1})) \\ &= Q(f(\cdot, y(\cdot))) = Q(y'), \end{aligned}$$

e ainda $y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} y'(s)ds = I(y')$, basta mostrar que

$$y(t_{k+1}) - y(t_k) - h\Phi(t_k, y(t_k), \ldots, y(t_{k-p+1})) = I(y') - Q(y') = O(h^{p+1}),$$

o que resulta da regra de integração ter esse erro. O erro de integração vem da integração do erro de interpolação, ou seja

$$\begin{aligned} I(y') - Q(y') &= \int_{t_k}^{t_{k+1}} y'[t_k, \ldots, t_{k-p+1}, s] \prod_{m=0}^{p-1} (s - t_{k-m}) ds \\ &= y'[t_k, \ldots, t_{k-p+1}, \xi] \underbrace{\prod_{m=0}^{p-1} (\xi - t_{k-m})}_{O(h^p)} \underbrace{\int_{t_k}^{t_{k+1}} 1 ds}_{h} = O(h^{p+1}), \end{aligned}$$

notando que $\xi \in (t_k, t_{k+1})$, logo $\prod_{m=0}^{p-1} |\xi - t_{k-m}| \leq \prod_{m=0}^{p-1} |t_{k+1} - t_{k-m}| = \prod_{m=0}^{p-1} (m+1)h = p! h^p$.

A demonstração para os métodos de Adams-Moulton é semelhante, notando que havendo um nó extra, teremos

$$\begin{aligned} I(y') - Q(y') &= \int_{t_k}^{t_{k+1}} y'[t_k, \ldots, t_{k-p+1}, s] \prod_{m=-1}^{p-1} (s - t_{k-m}) ds \\ &= y'[t_k, \ldots, t_{k-p+1}, \xi] \underbrace{\prod_{m=-1}^{p-1} (\xi - t_{k-m})}_{O(h^{p+1})} \underbrace{\int_{t_k}^{t_{k+1}} 1 ds}_{h} = O(h^{p+2}). \end{aligned}$$

□

Poderíamos ser levados a concluir imediatamente a convergência de ordem p, mas no caso dos métodos multipasso, há que ter em atenção uma outra estabilidade, também chamada *zero-estabilidade*. A noção de convergência nos métodos multipasso é exactamente a mesma, dizemos que há convergência de ordem r se $E_n = y(t_n) - y_n = O(h^r)$.

Exercício 29. Considere um método multipasso da forma geral

$$y_{k+1} = \sum_{m=0}^{p-1} \alpha_{-m} y_{k-m} + h \sum_{m=-1}^{p-1} \beta_{-m} f_{k-m}.$$

Usando a expansão de Taylor, mostre que o método tem consistência de ordem r se verificar o *critério de coeficientes*:

$$\sum_{m=-1}^{p-1} \alpha_{-m} = 0 \quad \wedge \quad \sum_{m=-1}^{p-1} \alpha_{-m} m^s = s \sum_{m=-1}^{p-1} \beta_{-m} m^{s-1} \text{ (para } s = 1, \ldots, r),$$

com $\alpha_1 = -1$. Verifique este critério nos métodos Adams-Bashforth e Adams-Moulton de passo 2.

Resolução: Como $\alpha_1 = -1$, podemos escrever a igualdade

$$-\sum_{m=-1}^{p-1} \alpha_{-m} y_{k-m} = h \sum_{m=-1}^{p-1} \beta_{-m} f_{k-m}$$

e portanto para o método ser de ordem r devemos ter

$$-\frac{1}{h} \sum_{m=-1}^{p-1} \alpha_{-m} y(t_{k-m}) - \sum_{m=-1}^{p-1} \beta_{-m} y'(t_{k-m}) = O(h^r).$$

Consideramos agora o desenvolvimento de Taylor de $y(t_{k-m})$ e de $y'(t_{k-m})$:

$$y(t_{k-m}) = \sum_{s=0}^{r} \frac{(-mh)^s}{s!} y^{(s)}(t_k) + O(h^{r+1})$$

$$y'(t_{k-m}) = \sum_{s=1}^{r} \frac{(-mh)^{s-1}}{(s-1)!} y^{(s)}(t_k) + O(h^r)$$

Substituindo, obtemos

$$-\tfrac{1}{h} \sum_{m=-1}^{p-1} \alpha_{-m} \left(\sum_{s=0}^{r} \tfrac{(-mh)^s}{s!} y^{(s)}(t_k) + O(h^{r+1}) \right) - \sum_{m=-1}^{p-1} \beta_{-m} \left(\sum_{s=1}^{r} \tfrac{(-mh)^{s-1}}{(s-1)!} y^{(s)}(t_k) + O(h^r) \right) =$$

$$= -\frac{1}{h} y(t_k) \sum_{m=-1}^{p-1} \alpha_{-m} - \sum_{s=1}^{r} y^{(s)}(t_k) \frac{(-h)^{s-1}}{s!} \left(-\sum_{m=-1}^{p-1} \alpha_{-m} m^s + \sum_{m=-1}^{p-1} \beta_{-m} s m^{s-1} \right) + O(h^r)$$

será um $O(h^r)$ desde que satisfaça o critério:

$$\sum_{m=-1}^{p-1} \alpha_{-m} = 0 \quad \wedge \quad \sum_{m=-1}^{p-1} \alpha_{-m} m^s = s \sum_{m=-1}^{p-1} \beta_{-m} m^{s-1} \text{ (para } s = 1, \ldots, r).$$

Verificar para Adams-Bashforth de passo 2. Temos $\alpha_1 + \alpha_0 + \alpha_{-1} = -1 + 1 + 0 = 0$. Para $s = 1$, temos $-\alpha_1 + 0\alpha_0 + \alpha_{-1} = 1$, é igual a $\beta_{-1} + \beta_0 + \beta_1 = 0 + \frac{3}{2} - \frac{1}{2} = 1$. Para $s = 2$, temos $\alpha_1 + 0\alpha_0 + \alpha_{-1} = -1$, igual a $2(-\beta_{-1} + 0\beta_0 + \beta_1) = 2(0 + 0 - \frac{1}{2}) = -1$. Mas para $s = 3$, já temos $-\alpha_1 + 0\alpha_0 + \alpha_{-1} = 1$, diferente de $3(\beta_{-1} + 0\beta_0 + \beta_1) = 3(0 + 0 - \frac{1}{2}) = -\frac{3}{2}$. Conclui-se a ordem 2.

Verificar para Adams-Moulton de passo 2. Temos $\alpha_1 + \alpha_0 + \alpha_{-1} = -1 + 1 + 0 = 0$. Para $s = 1$, $-\alpha_1 + 0\alpha_0 + \alpha_{-1} = 1$, é igual a $\beta_{-1} + \beta_0 + \beta_1 = \frac{5}{12} + \frac{8}{12} - \frac{1}{12} = 1$. Para $s = 2$, temos $\alpha_1 + 0\alpha_0 + \alpha_{-1} = -1$, igual a $2(-\beta_{-1} + 0\beta_0 + \beta_1) = 2(-\frac{5}{12} + 0\frac{8}{12} - \frac{1}{12}) = -1$. Para $s = 3$, $-\alpha_1 + 0\alpha_0 + \alpha_{-1} = 1$, é igual a $3(\beta_{-1} + 0\beta_0 + \beta_1) = 3(\frac{5}{12} + 0\frac{8}{12} - \frac{1}{12}) = 1$. Para $s = 4$ temos $\alpha_1 + 0\alpha_0 + \alpha_{-1} = -1$, diferente de $4(-\beta_{-1} + 0\beta_0 + \beta_1) = 4(-\frac{5}{12} + 0\frac{8}{12} - \frac{1}{12}) = -2$. Conclui-se a ordem 3.

Observação 47. O critério de consistência com $s = 1$, do exercício anterior, ou seja, $\sum_{m=-1}^{p-1} \alpha_{-m} = 0 \quad \wedge \quad \sum_{m=-1}^{p-1} m\alpha_{-m} = \sum_{m=-1}^{p-1} \beta_{-m}$, dá-nos não apenas a condição suficiente para ser de ordem 1, mas mesmo a condição necessária para ser consistente. Porque a expansão mínima

$$y(t_{k-m}) = y(t_k) - mhy'(t_k) + o(h), \quad y'(t_{k-m}) = y'(t_k) + o(1),$$

leva a

$$-\frac{1}{h}\sum_{m=-1}^{p-1}\alpha_{-m}(y(t_k)-mhy'(t_k)+o(h))-\sum_{m=-1}^{p-1}\beta_{-m}(y'(t_k)+o(1))=$$

$$=-\frac{1}{h}y(t_k)\sum_{m=-1}^{p-1}\alpha_{-m}+y'(t_k)\left(\sum_{m=-1}^{p-1}\alpha_{-m}m-\sum_{m=-1}^{p-1}\beta_{-m}\right)+o(1)$$

que é $o(1)$ se e só se a condição se verificar.

3.6.4 Estabilidade e Convergência dos Métodos Multipasso

Para avaliar a estabilidade dos métodos multipasso, devemos considerar a estabilidade da equação às diferenças associada:

$$y_{n+1}=\sum_{m=0}^{p-1}\alpha_{-m}y_{n-m} \tag{3.6.6}$$

que reflecte como o método multipasso se comportaria ainda que $\Phi=0$. O que se pretende estudar é a propagação de pequenas perturbações na solução. No caso dos métodos unipasso teríamos apenas $y_{n+1}=y_n$, e portanto se $y_0\approx 0$ então $y_n=y_0\approx 0$. No entanto, no caso geral isso não passa assim, podemos ter os dados iniciais próximos de zero e a solução da equação às diferenças crescer para infinito, ou seja, não haverá estabilidade.

Para esse efeito começamos por notar que $y_n=r^n$ é solução da equação às diferenças se

$$r^{n+1}=\sum_{m=0}^{p-1}\alpha_{-m}r^{n-m}$$

ou seja, dividindo por $r^{n-p+1}\neq 0$,

$$r^p=\alpha_0 r^{p-1}+\ldots+\alpha_{-p+1} \tag{3.6.7}$$

e portanto temos uma equação característica, polinomial, que determina as soluções. Haverá p soluções, que designaremos $r_1,\cdots,r_p$. Cada $y_n=r_j^n$ é assim uma solução da equação às diferenças, e uma combinação linear será ainda solução

$$y_n=c_1r_1^n+\cdots+c_pr_p^n$$

e podemos determinar os valores $c_1,\ldots,c_p$ com base nos valores iniciais $y_0,\ldots,y_{p-1}$, através de um sistema linear que é uma matriz de Vandermonde, invertível, desde que $r_1,\cdots,r_p$ sejam distintos.

Obtemos assim, a solução geral no caso em que as raízes r_j são distintas. No caso em que não são distintas é preciso completar com outras soluções.

Vemos que sempre que algum $|r_j|>1$ a sucessão $|y_n|\to\infty$, desde que o coeficiente c_j não seja nulo. Ou seja, basta uma pequena perturbação dos valores iniciais para que a solução não seja limitada. Portanto neste caso há instabilidade.

Quando r_j é raiz de multiplicidade m é preciso fazer uma ligeira alteração, devemos considerar a combinação com as soluções particulares $\tilde{y}_n=n^kr_j^n$ com $k=0,...,m-1$. Isto só tem efeito no caso em que $|r_j|=1$ e é raiz múltipla, porque $|\tilde{y}_n|=n^k|r_j|^n\to\infty$ se e só se $|r_j|\geq 1$. Resumimos estes resultados na seguinte proposição.

Proposição 16. *A equação às diferenças (3.6.6) é estável se e só se as raízes da equação característica (3.6.7) verificarem*

$$|r_j| \leq 1 \text{ e, sendo múltipla } |r_j| < 1.$$

A equação às diferenças (3.6.6) acaba por determinar a estabilidade do método, porque vemos que é independente de Φ e por consequência de f, depende apenas dos factores recursivos que não estão multiplicados por h. Assim a sua influência não desaparece quando $h \to 0$, e contribuem decisivamente para a instabilidade do método.

Definição 13. Um método multipasso é considerado estável se a equação às diferenças associada for estável. Por isso só será estável se estiverem definidas as condições da proposição anterior (as raízes simples da equação característica devem ter módulo não superior a 1, e as múltiplas devem ter módulo inferior a 1).

Por consequência, os métodos unipasso, em que a equação característica era $r = 1$, raiz simples, eram sempre estáveis.

Proposição 17. *Os métodos de Adams são estáveis.*

Demonstração. É uma imediata consequência de a esses métodos estar associada a equação às diferenças $y_{n+1} = y_n$ idêntica à dos métodos unipasso. □

Exemplo 31. Consideramos o método de passo duplo

$$y_{k+1} = 4y_k - 3y_{k-1} - 2hf_k,$$

e verificamos que é pelo menos consistente de ordem 1, pois

$$\frac{y(t_{k+1}) - 4y(t_k) + 3y(t_{k-1})}{h} + 2f(t_k, y(t_k)) = O(h),$$

porque $y(t_{k+1}) = y(t_k) + hf(t_k, y(t_k)) + O(h^2)$, $3y(t_{k-1}) = 3y(t_k) - 3hf(t_k, y(t_k)) + O(h^2)$. Este método pode ser visto como resultado de aplicar a fórmula de diferenças progressivas de ordem 2 em t_{k-1} (como a fórmula não é aplicada em t_k perde uma ordem na aproximação).

No entanto, a equação às diferenças associada é $y_{k+1} = 4y_k - 3y_{k-1}$, que tem $r^2 = 4r - 3$ como equação característica, com raízes $r_1 = 1, r_2 = 3$. Vemos que $|r_2| > 1$, e portanto o método não será estável, apesar de ser consistente. Isso não permite que o método seja convergente.

Aplicando este método à equação $y' = -y$, cuja solução é $y(t) = y_0 e^{-t}$, obtemos

$$y_{k+1} = 4y_k - 3y_{k-1} + 2hy_k = (4 + 2h)y_k - 3y_{k-1}$$

cuja solução geral é

$$y_k = c_1 r_1^k + c_2 r_2^k$$

com $r_1 = 2-h-\sqrt{1-4h+h^2} = 1 + O(h)$, $r_2 = 2-h+\sqrt{(2-h)^2-3} = 3 + O(h)$, quando $h \to 0$.

No entanto, vemos que $y_n = c_1(1 + O(h))^n + c_2(3 + O(h))^n \to \infty$, excepto se distorcessemos os valores iniciais com $r_1 y_0 = y_1$ de forma a que $c_2 = 0$. Portanto, neste caso $h \to 0$ não assegura a convergência do método.

Teorema 22. *Os métodos multipasso são convergentes se e só se forem estáveis e consistentes. Se o método multipasso for estável e consistente de ordem r (e também inicializado com essa ordem), então será convergente de ordem r.*

Demonstração. A demonstração de que sendo estável e consistente de ordem r será convergente de ordem r, segue uma linha semelhante à que fizémos para o caso unipasso. A demonstração de que se trata de condição necessária e suficiente usa outro tipo de argumentos. Para mais detalhes ver p.ex. Numerical Analysis (1998, Springer) de R. Kress. □

Corolário 3.1. *Os métodos de Adams-Bashforth de passo p são convergentes de ordem p, e os de Adams-Moulton de passo p têm ordem de convergência $p+1$.*

Demonstração. Consequência imediata do teorema anterior, dos métodos serem estáveis, e da sua ordem de consistência. □

Observação 48. (Métodos Multipasso como Preditores-Correctores). Um processo habitual na utilização dos Métodos de Adams-Moulton é considerá-los como correctores, usando os Métodos de Adams-Bashforth como preditores.

Exercício 30. Mostre que o método Leapfrog ("salto-ao-eixo")

$$y_{k+1} = y_{k-1} + 2hf_k \tag{3.6.8}$$

tem ordem de convergência 2.

Resolução: Este método resulta da aproximação por diferenças centradas. É estável porque a equação às diferenças associada $y_{k+1} = y_{k-1}$ também é ($r^2 = 1$ tem $r_1 = -1, r_2 = 1$, raízes distintas). É também fácil ver que tem ordem de consistência 2.

$$\frac{y(t_{k+1})-y(t_{k-1})}{h} - 2f(t_k, y(t_k)) = \frac{2y'(t_k)+O(h^3)}{h} - 2f(t_k, y(t_k)) = O(h^2).$$

Como é estável e tem ordem de consistência 2, terá ordem de convergência 2.

Observação 49. (Métodos BDF). Uma outra maneira de deduzir métodos multipasso consiste em usar fórmulas de diferenciação numérica. Já vimos o caso do Leapfrog, mas também vimos no Exemplo 31 que isso corre o risco de levar a métodos instáveis.

Uma classe de métodos que usam diferenciação numérica são os Métodos BDF (*backwards differentiation formula*), que são implícitos, usando fórmulas de diferenciação regressiva. O primeiro exemplo será o Método de Euler Implícito, que tem ordem 1 e já estudámos. Vejamos um outro exemplo, de ordem 2, considerando a fórmula de diferenciação regressiva (exercício no capítulo sobre diferenciação):

$$y'(t) = \frac{3y(t) - 4y(t-h) + y(t-2h)}{2h} + O(h^2)$$

aplicada ao ponto t_{k+1}, retiramos $f_{k+1} = \frac{3y_{k+1}-4y_k+y_{k-1}}{2h}$, e obtemos o *Método BDF de ordem 2*

$$y_{k+1} = \frac{4}{3}y_k - \frac{1}{3}y_{k-1} + \frac{2h}{3}f_{k+1} \tag{3.6.9}$$

que é implícito, estável (as raízes da equação característica associada são $r_1 = \frac{1}{3}, r_2 = 1$), e tem consistência de ordem 2 (exercício).

Os restantes Métodos BDF, de ordem superior, até ordem 6, podem ser obtidos de maneira semelhante. A partir de ordem 6 falham na condição de estabilidade, não sendo convergentes.

Estes métodos têm sido utilizados pela sua eficácia em termos da A-estabilidade, nos problemas rígidos. Neste caso vemos que a região de A-estabilidade será dada por

$$\mathcal{A} = \{\alpha \in \mathbb{C} : |2 \pm \sqrt{1+2\alpha}| \leq |2\alpha - 3|\} \supset \mathbb{C}\backslash(0,4) \times (-3,3)_i$$

porque usando $f_k = \alpha y_k$ obtemos $y_{k+1} = \frac{4}{3}y_k - \frac{1}{3}y_{k-1} + \frac{2h}{3}\alpha y_{k+1}$, e daqui com $h = 1$, temos a equação característica

$$(1 - \tfrac{2}{3}\alpha)r^2 = \tfrac{4}{3}r - \tfrac{1}{3} \implies r = \frac{-2 \pm \sqrt{1+2\alpha}}{2\alpha - 3},$$

e as raízes devem ter módulo não superior a 1. Comparando com o Método dos Trapézios Implícito este método tem maior região de A-estabilidade.

3.7 Problemas de Fronteira

Ao contrário dos problemas de valor inicial, definidos pelo Problema de Cauchy, os problemas de fronteira podem ter condições em ambas as extremidades de um intervalo.

O caso mais habitual é o dos problemas de fronteira de segunda ordem, genericamente definidos por:

$$\begin{cases} y''(t) = g(t, y(t), y'(t)), \\ y(a) = y_a\,, \quad y(b) = y_b. \end{cases} \tag{3.7.1}$$

Notamos que não podemos tratar este problema da mesma forma que os Problemas de Cauchy, porque sendo um problema de 2ª ordem, temos informação sobre $y(a)$, mas falta-nos informação sobre $y'(a)$, e analogamente para o extremo b. Uma parte das informações está na extremidade a, e outra na extremidade b. Consideramos aqui o caso em que são impostas condições sobre a função nas extremidades (chamado *problema de Dirichlet)*, mas também se poderiam considerar condições sobre as derivadas $y'(a)$ e $y'(b)$ (chamado *problema de Neumann),* ou ainda mistas. Os processos podem ser generalizados sem dificuldade. Também iremos focar neste caso de ordem 2, onde são impostas duas condições nas extremidades, mas os problemas podem ser de ordem m, colocando m condições no total, repartidas em ambas as extremidades (se todas as condições ficassem definidas numa extremidade, então tratava-se de um problema de Cauchy).

3.7.1 Método do Tiro

O Método do Tiro consiste em reduzir o problema de fronteira a uma equação em que intervêm problemas de Cauchy. Podemos assim usar os métodos vistos anteriormente para resolver os problemas de Cauchy intervenientes. Vejamos a sua aplicação ao problema de fronteira definido em (3.7.1).

A ideia consiste em definir uma função

$$\phi(x) = y_x(b),$$

em que y_x é a solução do problema de Cauchy, em que x é tal que $y'_x(a) = x$. Ou seja, temos os problemas de Cauchy auxiliares:

$$\begin{cases} y''_x(t) = g(t, y_x(t), y'_x(t)), \\ y_x(a) = y_a\,, \quad y'_x(a) = x. \end{cases}$$

Consoante o valor de x, teremos uma solução diferente y_x, que respeita a equação diferencial e a condição sobre a, faltando verificar a condição sobre b. O nosso objectivo é variar x de forma a que verifique também a condição sobre b. Ou seja, pretendemos que $y_x(b) = y_b$, o que corresponde a resolver a equação:

$$\phi(x) = y_b.$$

Trata-se de uma equação não linear, onde podem ser aplicados métodos numéricos, por exemplo, os métodos da bissecção, secante ou Newton.

Aplicando o método da secante, consideramos dois valores iniciais, x_0 e x_1, e iteramos

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

notando que em cada nova iteração devemos calcular $\phi(x_n)$, o que corresponde a resolver um problema de Cauchy.

Exemplo 32. Consideramos um problema de 2ª ordem simples

$$\begin{cases} y''(t) = \beta^2 y(t), \\ y(a) = y_a\,, \quad y(b) = y_b. \end{cases}$$

Vamos considerar o problema de Cauchy auxiliar

$$\begin{cases} y_x''(t) = \beta^2 y_x(t), \\ y_x(a) = y_a\,, \quad y_x'(a) = x. \end{cases}$$

onde sabemos, neste caso que a solução é da forma

$$y_x(t) = C_1(x)e^{-\beta t} + C_2(x)e^{\beta t}.$$

Podemos determinar $C_1(x)$ e $C_2(x)$ através das condições iniciais:

$$y_a = C_1(x)e^{-\beta a} + C_2(x)e^{\beta a}, \quad x = -\beta C_1(x)e^{-\beta a} + \beta C_2(x)e^{\beta a},$$

obtendo a expressão geral

$$y_x(t) = y_a \cosh(\beta(t-a)) + \frac{x}{\beta}\sinh(\beta(t-a)),$$

e assim $\phi(x) = y_x(b) = y_a \cosh(\beta(b-a)) + \frac{x}{\beta}\sinh(\beta(b-a))$, pelo que basta resolver $\phi(x) = y_b$. Neste caso é imediato:

$$x = \beta(y_b - y_a \cosh(\beta(b-a)))/\sinh(\beta(b-a)).$$

Conclui-se que a solução é

$$y(t) = y_a \cosh(\beta(t-a)) + (y_b - y_a \cosh(\beta(b-a)))\frac{\sinh(\beta(t-a))}{\sinh(\beta(b-a))}.$$

3.7.2 Método das Diferenças Finitas

O método das diferenças finitas considera uma divisão do intervalo $[a, b]$ em pontos igualmente espaçados

$$t_k = a + kh, \text{ com } h = \frac{b-a}{n},$$

e em cada um destes pontos podemos aplicar uma aproximação da derivada e 2ª derivada por diferenciação numérica.

Por exemplo, usando a aproximação de 2ª ordem por diferenças centradas

$$\begin{aligned} y'(t_k) &= \frac{y(t_{k+1}) - y(t_{k-1})}{2h} + O(h^2) \\ y''(t_k) &= \frac{y(t_{k+1}) - 2y(t_k) + y(t_{k-1})}{h^2} + O(h^2) \end{aligned}$$

podemos substituir em cada ponto $y''(t_k) = g(t_k, y(t_k), y'(t_k))$, com $y_k \approx y(t_k)$ obtemos o sistema

$$\begin{aligned} \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} &= g\left(t_k, y_k, \tfrac{y(t_{k+1}) - y(t_{k-1})}{2h}\right), \text{ para } k = 1, \ldots, n-1 \\ &\quad \text{com } y_0 = y(t_0) = y_a,\ y_n = y(t_n) = y_b. \end{aligned}$$

Este sistema é geralmente não linear e pode ser resolvido, por exemplo, pelo método de Newton (ou Broyden).

Quando a função g é linear (na 2ª e 3ª componente), o sistema é linear e pode ser resolvido de forma algébrica.

Exemplo 33. (Caso linear)

Consideramos o problema de 2ª ordem

$$\begin{cases} y''(t) = c(t) + d(t)y(t), \\ y(a) = y_a\,, \quad y(b) = y_b. \end{cases}$$

em que $g(t, y, z) = c(t) + d(t)y$, é linear em y, e aplicando a aproximação de 2ª ordem já vista, obtemos

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = c_k + d_k y_k \qquad {\scriptstyle (k=1,\ldots,n-1)}$$

em que abreviamos $c_k = c(t_k)$, $d_k = d(t_k)$, e temos $y_0 = y_a$, $y_n = y_b$.

Podemos escrever estas equações na forma $y_{k-1} - (2 + h^2 d_k)y_k + y_{k+1} = h^2 c_k$ obtendo o sistema linear:

$$\begin{bmatrix} -(2+h^2d_1) & 1 & 0 & \cdots & 0 \\ 1 & -(2+h^2d_2) & 1 & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -(2+h^2d_{n-1}) \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ y_{n-1} \end{bmatrix} = \begin{bmatrix} h^2c_1 - y_a \\ h^2c_2 \\ \vdots \\ h^2c_{n-2} \\ h^2c_{n-1} - y_b \end{bmatrix}$$

Notamos que a matriz é tridiagonal (e é invertível se $d_k > 0$, pois tem a diagonal dominante), o que permite uma resolução facilitada do sistema linear.

ANEXO

Nome do Método	**Expressão**	**Ordem**
Euler	$y_{k+1} = y_k + hf_k$	1
Euler Implícito	$y_{k+1} = y_k + hf_{k+1}$	1
RK2-Ponto-Médio	$y_{k+1} = y_k + h\, f(t_k + \frac{h}{2}, y_k + \frac{h}{2} f_k)$	2
RK2-Heun	$y_{k+1} = y_k + \frac{h}{2}(\, f_k + f(t_k + h, y_k + hf_k)\,)$	2
Trapézios (implícito)	$y_{k+1} = y_k + \frac{h}{2}(\, f_k + f_{k+1}\,)$	2
Leapfrog (explícito)	$y_{k+1} = y_{k-1} + 2hf_k$	2
BDF-2 (implícito)	$y_{k+1} = \frac{4}{3}y_k - \frac{1}{3}y_{k-1} + \frac{2h}{3}f_{k+1}$	2
Adams-Bashforth (passo 2)	$y_{k+1} = y_k + \frac{3h}{2}f_k - \frac{h}{2}f_{k-1}$	2
Adams-Moulton (passo 2)	$y_{k+1} = y_k + \frac{5h}{12}f_{k+1} + \frac{2h}{3}f_k - \frac{h}{12}f_{k-1}$	3
Adams-Bashforth (passo 3)	$y_{k+1} = y_k + \frac{23h}{12}f_k - \frac{4h}{3}f_{k-1} + \frac{5h}{12}f_{k-2}$	3
RK-3	$\begin{cases} y_{k+1} = y_k + \frac{h}{6}(\, F_1 + 4F_2 + F_3)\,) \\ F_1 = f_k;\ F_2 = hf(t_k + \frac{h}{2}, y_k + \frac{h}{2}F_1) \\ F_3 = f(t_k + h, y_k - hF_1 + 2hF_2) \end{cases}$	3
Adams-Moulton (passo 3)	$y_{k+1} = y_k + \frac{3h}{8}f_{k+1} + \frac{19h}{24}f_k - \frac{5h}{24}f_{k-1} + \frac{h}{24}f_{k-2}$	4
Adams-Bashforth (passo 4)	$y_{k+1} = y_k + \frac{55h}{24}f_k - \frac{59h}{24}f_{k-1} + \frac{37h}{24}f_{k-2} - \frac{3h}{8}f_{k-3}$	4
RK-4	$\begin{cases} y_{k+1} = y_k + \frac{h}{6}(\, F_1 + 2F_2 + 2F_3 + F_4)\,) \\ F_1 = f_k;\ F_2 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}F_1) \\ F_3 = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}F_2) \\ F_4 = f(t_k + h, y_k + hF_3) \end{cases}$	4
etc...		

Bibliografia

[1] Atkinson K. E.; An Introduction to Numerical Analysis. *Wiley & Sons*, New York, 1989.

[2] Burden R., Faires J.; Numerical Analysis, 5th. ed. *PWS Publishers*, Boston, 1993.

[3] Ciarlet P. G.; Introduction à l'analyse numérique matricielle et à l'optimisation. *Masson*, Paris, 1988.

[4] Conte, S. D., de Boor C.; Elementary numerical analysis. *McGraw-Hill*, Singapore, 1981.

[5] Démidovitch B., Maron I.; Eléments de calcul numérique. *MIR*, Moscovo, 1973.

[6] Golub G. H., Van Loan C.F.; Matrix Computations. *John Hopkins*, Baltimore, 1985.

[7] Henrici P.; Elements of numerical analysis. *Wiley & Sons*, New York, 1964.

[8] Isaacson E., Keller H. B.; Analysis of numerical methods. *Wiley & Sons*, New York, 1966.

[9] Kress R.; Numerical Analysis. *Springer-Verlag*, New York, 1998.

[10] Nougier J. P.; Méthodes de calcul numérique, 2nd. ed. *Masson*, Paris, 1985.

[11] Ortega J. M.; Numerical Analysis, a second course. *SIAM*, Philadelphia, 1990.

[12] Scheid F.; Análise numérica. *McGraw-Hill*, Lisboa, 1991.

[13] Stoer J., Bulirsch R.; Introduction to Numerical Analysis, 2nd ed. *Springer Texts in Appl. Math.*, 1993.

[14] Zeidler E.; Nonlinear Functional Analysis and its Applications. *Springer-Verlag*, New York, 1993.