Análise Comparativa de Métodos Numéricos para Equações Algébrico-Diferenciais

Neyva Maria Lopes Romeiro

Orientador: Prof. Dr. Antonio Castelo Filho

DEDALUS - Acervo - ICMC

30300042854

Dissertação apresentada ao Instituto de Ciências Matemáticas de São Carlos, da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Ciências - Área: *"Ciências de Computação e Matemática Computacional"*.

São Carlos - Novembro de 1994

240607

Class. -	T
Cutt. -	R763a
Tombo	1544

A meu Pai.

"Que os nossos esforços desafiem as nossas impossibilidades".
(Charles Chaplin)

Agradecimentos

Ao Professor, Antonio Castelo Filho, pela orientação, incentivo e apoio indispensáveis para a realização deste trabalho.

À minha família, que mesmo estando longe, sempre deu o apoio necessário neste período de estudos.

Aos professores do Instituto de Ciências Matemáticas de São Carlos, da Universidade de São Paulo, que muito me ajudaram.

Às secretárias da pós graduação, Beth e Laura, pela simpatia, atenção e eficiência com que sempre me atenderam.

Aos meus amigos que direta e indiretamente contribuíram com seu apoio durante estes anos em São Carlos, em especial: Augusto, Sadao, Rô, Claudinha, Tchelo, Marquinhos, Ana, Di, Vanderlei, Vera, Magda, Lú, Rosana, Sandra, Neuza, Ricardo.

À CAPES pelo apoio financeiro fornecido durante todo o desenvolvimento deste trabalho.

Análise Comparativa de Métodos Numéricos de Equações Algébrico-Diferenciais

Neyva Maria Lopes Romeiro

Orientador: Prof. Dr. Antonio Castelo Filho

Resumo

Este trabalho aborda, apenas, sistemas de Equações Algébrico-Diferenciais (EAD) de índice zero ou um.

Estudaremos as propriedades de ordem e convergência do método diferenças para trás (BDF) e principalmente do método de Runge-Kutta Implícito (RKI).

Também, descreveremos os algoritmos provenientes destes métodos, quando aplicados em sistemas de EAD de índice zero ou um.

Por último, faremos uma comparação numérica entre os método BDF e método de Runge-Kutta Implícito usando dois códigos representativos, sendo eles DASSL e RADAU5, respectivamente.

Comparative Analysis of Numerical Methods in Differential- Algebraic Equations

Neyva Maria Lopes Romeiro

Adviser: Prof. Dr. Antonio Castelo Filho

Abstract

This work is concerned with the numerical solution of Differential-Algebraic Equations (DAE) of index zero and one. Among the numerical methods for solving DAE's we give special attention to Backward Differentiation Formulas (BDF) and Implicit Runge-Kutta (IRK) methods. A detailed study of order of convergence for these methods is presented. A description of the algorithms employed for solving DAE's of index zero and one is also considered.

Finally we compare the performance of the two methods by using two known codes: DASSL and RADAU5. Numerical results are presented.

Conteúdo

1	Introdução	01
2	Teoria das EAD	03
2.1	Introdução	03
2.2	Solubilidade e Índice	04
2.3	EAD Lineares com Coeficientes Constantes	05
2.4	EAD Lineares com Coeficientes Variando com o Tempo	07
2.4.1	Solubilidade e Índice	07
2.4.2	Formas Estruturais	08
2.5	EAD não Lineares	10
2.5.1	Solubilidade e Índice	10
2.5.2	Formas Estruturais	11
3	Método BDF	13
3.1	Introdução	13
3.2	Métodos de Passos Múltiplos	13
3.3	Convergência do Método BDF	15
3.3.1	Sistemas Lineares com Coeficientes Constantes	15
3.3.2	Sistemas Semi Explícitos	15

3.3.3	Sistemas Completamente Implícitos	16
4	Métodos de Runge-Kutta Implícito	18
4.1	Introdução	18
4.2	Sistemas Lineares com Coeficientes Constantes	20
4.3	Sistemas não Lineares	27
5	Algoritmos	56
5.1	Introdução	56
5.2	Algoritmos e Estratégias no DASSL	57
5.2.1	Fórmula Básica	57
5.2.2	Solução do Sistema não Linear	60
5.3	Algoritmos e Estratégias no RADAU5	61
5.3.1	Solução do Sistema não Linear	61
5.3.2	Valor Inicial para a Iteração de Newton	65
6	Resultados Numéricos	67
6.1	Introdução	67
6.2	Equação do Pêndulo	68
6.3	Equação de Van der Pol	70
6.4	Equação da Esfera	72
6.5	Equação Teste	75
7	Conclusão	77
	Bibliografia	79

Capítulo 1

Introdução

Chamaremos de Equações Algébrico-Diferenciais (EAD) a um sistema que envolve equações algébricas e/ou diferenciais. Este tipo de equação surge naturalmente em muitas aplicações, tais como:

- Equações do Movimento Descrito por Pêndulos;
- Simulação de Trabalhos Elétricos;
- Engenharia de Controle;
- Dinâmica de Fluidos;
- Sistemas Mecânicos Restritos e outros.

As EAD apresentam dificuldades numéricas e analíticas quando não ocorrem por meio de uma Equação Diferencial Ordinária (EDO) ou Equações Puramente Algébricas.

No Capítulo 2 apresentaremos uma teoria para compreendermos algumas das diferenças existentes entre EAD e EDO.

Neste capítulo, também, definiremos o conceito de índice das EAD. Geralmente as EAD de índices elevados tornam o problema, numericamente, mais difícil. Gostaríamos de ressaltar que este trabalho aborda apenas sistemas de índice zero ou um.

Ainda, neste capítulo, apresentamos três classes de EAD:

- EAD lineares com coeficientes constantes;

- EAD lineares com coeficientes variando com o tempo;
- EAD não lineares.

Temos ainda outras definições e resultados importantes, neste capítulo, que sem estes não poderíamos concluir nosso trabalho.

No Capítulo 3 estudaremos uma das classes dos métodos de passos múltiplos, conhecido como diferenciação para trás (*backward differentiation formulas*) ou método BDF. Estes métodos são, amplamente, usados para a integração de equações diferenciais stiff.

Este trabalho faz uma abordagem rápida sobre o método BDF, maiores detalhes podem ser encontrados em [PIRES - 94]. Nossa ênfase maior, neste trabalho, é sobre os métodos de Runge-Kutta Implícitos.

No Capítulo 4 estudaremos os métodos de Runge-Kutta Implícito (RKI). Recentemente tem sido desenvolvido muitos estudos sobre este método devido as seguintes razões: a combinação de ordem elevada com boa estabilidade; permite uma seleção simples do passo dado; é um método de início próprio; e outras.

Na seção 4.3 apresentaremos alguns lemas que nos permitirá mostrar o principal resultado deste capítulo: o teorema que fornece condições sobre os coeficientes do método de RKI para a convergência deste quando aplicado a um sistema de EAD completamente implícito de índice um.

No Capítulo 5 analisaremos o desenvolvimento de algoritmos para resolver sistemas de EAD usando os métodos BDF e RKI e, em particular, os algoritmos utilizados em dois códigos, já consagrados, o DASSL e o RADAU5.

A implementação do método BDF surgiu, com grande reconhecimento, da importância da EAD no meio científico e em aplicações na engenharia. O DASSL é um código que utiliza o método BDF. Este código é indicado para resolver sistemas completamente implícito de EAD de índice zero ou um. Podemos, também, aplicar este código em EAD semi explícita, como veremos no Capítulo 6.

O código RADAU5 utiliza o método de RKI em sistemas semi explícitos. Este código, também, obtém resultados satisfatórios para sistemas de EAD semi explícito de índice maior ou igual a dois.

No Capítulo 6 apresentaremos resultados numéricos de alguns exemplos de sistemas de EAD de índice zero ou um. Os resultados destes exemplos são obtidos usando os códigos DASSL e RADAU5.

No Capítulo 7 faremos uma conclusão do trabalho de um modo geral, assim como as comparações dos resultados obtidos no Capítulo 6, avaliando a estabilidade dos métodos BDF e Runge-Kutta Implícito.

Capítulo 2

Teoria das EAD

2.1 Introdução

Neste capítulo desenvolveremos uma teoria necessária para compreendermos algumas das diferenças entre equações algébrico-diferenciais e equações diferenciais ordinárias. Também veremos algumas classes de EAD incluindo o caso de EAD lineares com coeficientes constantes, EAD lineares com coeficientes variando com o tempo e EAD não lineares, onde veremos as definições de solubilidade e índice em cada um dos casos. Esta teoria encontra-se em [BRENAN - 89].

As demonstrações de alguns dos teoremas, citados neste capítulo, podem ser encontradas em [PIRES - 94].

Inicialmente, apresentaremos algumas definições necessárias para a compreensão de resultados posteriores.

Definição 2.1.1 Um Homeomorfismo entre X e Y é uma aplicação $\varphi: X \rightarrow Y$ contínua, inversível com inversa contínua. Se φ e φ^{-1} são de classe C^k , dizemos que φ é um Homeomorfismo de classe C^k entre X e Y . Um Homeomorfismo de classe C^k , com $k \geq 1$ é dito Difeomorfismo de classe C^k .

Definição 2.1.2 $M \subset R^k$ é uma Variedade de classe C^r de dimensão n se para todo $x \in M$, existem vizinhanças abertas $U \subset R^k$ de x , $V \subset R^n$ e um Homeomorfismo de classe C^r , $\varphi: V \rightarrow U \cap M$.

O exemplo da Figura 2.1.1 é de uma variedade de dimensão dois, ou também uma superfície.

Figura 2.1.1 - Variedade de dimensão dois

2.2 Solubilidade e Índice

A função $y(t)$ é uma solução da EAD

$$F(t, y, y') = 0 \quad (2.2.1)$$

em um intervalo I se y é continuamente diferenciável em I e satisfaz (2.2.1) para todo $t \in I$.

Definição 2.2.1 Seja I um intervalo aberto de $\mathbb{R}$, Ω um conjunto conexo aberto de $\mathbb{R}^{2m+1}$ e F uma função diferenciável de Ω em $\mathbb{R}^m$. Então a EAD (2.2.1) é solúvel em I sobre Ω se existe uma família r -dimensional de soluções $\phi(t, c)$ definidas em um conjunto conexo aberto $I \times \tilde{\Omega}$, $\tilde{\Omega} \subset \mathbb{R}^r$ tal que

- i) $\phi(t, c)$ está definida em I para cada $c \in \tilde{\Omega}$;
- ii) $(t, \phi(t, c), \phi'(t, c)) \in \Omega$ para $(t, c) \in I \times \tilde{\Omega}$;
- iii) Se $\psi(t)$ é qualquer outra solução e $(t, \psi(t, c), \psi'(t, c)) \in \Omega$, então $\psi(t) = \phi(t, c)$ para algum $c \in \tilde{\Omega}$;
- iv) o gráfico de ϕ como uma função de (t, c) é uma variedade $(r+1)$ -dimensional.

Definição 2.2.2 Denotamos o índice de uma EAD por v , onde v é o número mínimo de vezes que a equação (2.2.1), ou parte dela, deve ser diferenciada em relação a t para determinar y' como uma função contínua de y e t .

Exemplo 2.2.1 Seja a EAD

$$\begin{aligned}x' &= f(t, x, y) \\ 0 &= g(t, x, y).\end{aligned}$$

Diferenciando a segunda equação obtemos

$$g_x(t, x, y)x' + g_y(t, x, y)y' + g_t(t, x, y) = 0.$$

Se g_y é não singular temos

$$\begin{aligned}x' &= f(t, x, y) \\ y' &= -g_y^{-1}(t, x, y)(g_x(t, x, y)f(t, x, y) + g_t(t, x, y))\end{aligned}$$

Portanto, a EAD acima tem índice um.

2.3 EAD Lineares com Coeficientes Constantes

Consideraremos a EAD linear com coeficiente constante

$$Ax' + Bx = f, \tag{2.3.1}$$

onde A e B são matrizes $m \times m$.

Definição 2.3.1 Se λ é um parâmetro complexo, então $(\lambda A + B)$ é a matriz pencil.

Definição 2.3.2 Se $\det(\lambda A + B)$ não é identicamente nulo como uma função de λ , então a matriz pencil (ou pencil) é regular.

Quando o pencil é regular, dizemos que a EAD também é regular.

Definição 2.3.3 Uma matriz tem índice de nilpotência k se $N^k = 0$ e $N^{k-1} \neq 0$.

Definição 2.3.4 O pencil $(\lambda A + B)$ é um pencil nilpotente quando $\det(\lambda A + B)$ é constante.

Teorema 2.3.1 *Suponha que $(\lambda A + B)$ é um pencil regular. Então existem matrizes não singulares P e Q tais que*

$$PAQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad PBQ = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}, \quad (2.3.2)$$

onde N é uma matriz nilpotente com índice de nilpotência k e I a matriz identidade. No caso especial em que A é não singular, $PAQ = I$ e $PBQ = C$. Se $\det(\lambda A + B)$ é identicamente constante, então (2.3.2) reduz a $PAQ = N$ e $PBQ = I$.

Considerando que as mudanças de coordenadas P e Q proporcionadas por este teorema sejam aplicadas à EAD (2.3.1), isto é, considerando $x = Qy$, então (2.3.1) torna-se $AQy' + BQy = f$. Multiplicando a esquerda por P , temos

$$PAQy' + PBQy = Pf. \quad (2.3.3)$$

Fazendo $Pf = [f_1, f_2]^t$ e usando o teorema 2.3.1, obtemos

$$\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} y_1' \\ y_2' \end{bmatrix} + \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

ou

$$y_1' + Cy_1 = f_1 \quad (2.3.4a)$$

$$Ny_2' + y_2 = f_2, \quad (2.3.4b)$$

onde (2.3.4a) é uma EDO cuja solução existe para qualquer valor inicial de y_1 e qualquer função f_1 contínua e, (2.3.4b) tem apenas uma solução, pois

$$Ny_2' + y_2 = (ND + I)y_2 = f_2,$$

assim $y_2 = (ND + I)^{-1} f_2$, onde $D = \frac{d}{dt}$.

Mas

$$(ND + I)^{-1} = (I - (-ND))^{-1} = \sum_{i=0}^{\infty} (-ND)^i = \sum_{i=0}^{k-1} (-ND)^i = \sum_{i=0}^{k-1} (-1)^i N^i D^i,$$

onde k é o índice de nilpotência de N e, assim $(ND)^i = 0$ para $i \geq k$.

Desta forma

$$y_2 = (ND + 1)^{-1} f_2 = \sum_{i=0}^{k-1} (-1)^i N^i f_2^i \quad (2.3.5)$$

onde $f_2^i = \frac{d^i f_2}{dt^i}$.

Teorema 2.3.2 *A EAD linear com coeficiente constante (2.3.1) é solúvel se, e somente se $(\lambda A + B)$ é um pencil regular.*

Teorema 2.3.3 *Suponha $Ax' + Bx = f$ sendo uma EAD linear com coeficiente constante solúvel de índice $v \geq 1$. Então, cada iteração da resolução para restrição explícita e diferenciação destas restrições reduz o índice por 1, tal que, depois de v diferenciações a EAD será reduzida em uma EDO.*

2.4 EAD Lineares com Coeficientes Variando com o Tempo

As EAD lineares com coeficientes variando com o tempo têm a forma

$$A(t)x'(t) + B(t)x(t) = f(t) \quad (2.4.1)$$

e está definida no intervalo I .

2.4.1 Solubilidade e Índice

Definição 2.4.1 *O sistema (2.4.1) com A e B matrizes $m \times m$ é solúvel em I se, para toda f m -vezes diferenciável, existe pelo menos uma solução para (2.4.1) continuamente diferenciável. Além disso, as soluções são definidas em I e determinadas para algum valor em $t \in I$.*

Definição 2.4.2 *Suponha que (2.4.1) é solúvel em I , o índice de (2.4.1) é o número mínimo de diferenciações necessárias para determinar x' em função de x e t .*

Definição 2.4.3 *Se (2.4.1) é uma EAD regular, então o índice local em t , denotado por $v_1(t)$, é o índice do pencil $(\lambda A(t) + B(t))$.*

2.4.2 Formas Estruturais

Considerando a mudança de coordenada em (2.4.1) dada por $x = Q(t)y$ e também multiplicando por $P(t)$, obtemos

$$PAQy' + (PAQ' + PBQ)y = Pf, \quad (2.4.2)$$

onde P e Q são não singulares no intervalo de interesse e pelo menos tão suave quanto os coeficientes da EAD.

Teorema 2.4.1 *Suponha que (2.4.1) é uma EAD solúvel em I . Então as seguintes afirmações são equivalentes*

- i) (2.4.1) é uma EDO implícita;
- ii) $\nu = 0$;
- iii) $\nu_l \equiv 0$;
- iv) $\nu_l(t_0) = 0$ para algum $t_0 \in I$,

onde ν é o índice e ν_l é o índice local de (2.4.1).

Teorema 2.4.2 *Suponha que (2.4.1) é solúvel, então $\nu = 1$ se, e somente se $\nu_l \equiv 1$.*

Estes resultados mostram que ν e ν_l são os mesmos para sistemas solúveis de índice zero ou um. Isto não ocorre em sistemas de índice maior ou igual a dois [CAMPBELL - 82].

Teorema 2.4.3 *Suponha que (2.4.1) é solúvel em I . As seguintes afirmações são preservadas quando transformamos (2.4.1) num sistema analiticamente equivalente a (2.4.2)*

- i) $\nu_l \equiv 0$;
- ii) $\nu_l \equiv 1$;
- iii) $\nu_l(t) \geq 2$ para todo $t \in I$.

Se $\nu_l(t) \geq 2$ em I , então é possível encontrar um sistema analiticamente equivalente tal que $\nu_l = 2$ em um subconjunto aberto denso em I [GEAR - 83]. Tais resultados mostram que para muitos propósitos podemos considerar três classes de EAD solúveis, sendo elas

- i) EDO implícitas;
- ii) índice um;
- iii) índice dois.

Definição 2.4.4 *O sistema (2.4.1) é semi explícito se estiver na forma*

$$\begin{aligned} x_1' + B_{11}(t)x_1 + B_{12}(t)x_2 &= f_1 \\ B_{21}(t)x_1 + B_{22}(t)x_2 &= f_2. \end{aligned} \quad (2.4.3)$$

A EAD semi explícita (2.4.3) tem índice um se, e somente se B_{22} é não singular para todo t .

Definição 2.4.5 A EAD (2.4.1) está na forma Hessenberg de dimensão r se pode ser escrita como

$$\begin{bmatrix} I & 0 & & & 0 \\ 0 & I & & & \\ & & I & & \\ & & & I & \\ 0 & & & & 0 \end{bmatrix} \begin{bmatrix} x_1' \\ \vdots \\ x_r' \end{bmatrix} + \begin{bmatrix} B_{11} & * & * & B_{1,r-1} & B_{1r} \\ B_{21} & * & * & B_{2,r-2} & 0 \\ 0 & * & * & * & \\ & & * & * & \\ 0 & & 0 & B_{r,r-1} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_r \end{bmatrix},$$

onde x_i são vetores, B_{ij} são matrizes e o produto $B_{r,r-1}B_{r-1,r-2}\dots B_{1r}$ é não singular.

As formas Hessenberg de dimensão 2 ou 3 são as mais comuns.

Definição 2.4.6 Dizemos que o sistema (2.4.1) está na forma canônica padrão se

$$\begin{bmatrix} I & 0 \\ 0 & N(t) \end{bmatrix} x' + \begin{bmatrix} C(t) & 0 \\ 0 & I \end{bmatrix} x = f(t), \quad (2.4.4)$$

onde N é estritamente triangular inferior (ou superior).

A definição acima não exige que a matriz N tenha posto ou índice constante.

Definição 2.4.7 Uma propriedade de uma matriz é uma propriedade estrutural se todo elemento não nulo da matriz pode ser trocado por variáveis independentes e a propriedade ainda vale em um conjunto aberto de valores destas variáveis.

Teorema 2.4.4 Suponha que (2.4.1) é solúvel em I , então (2.4.1) é analiticamente equivalente a

$$\begin{bmatrix} I & G \\ 0 & N \end{bmatrix} z' + \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} z = \begin{bmatrix} g \\ h \end{bmatrix}, \quad (2.4.5)$$

onde $Nz_2' + z_2 = h$ tem solução única para cada função h . Além disso, existe uma família contável de intervalos abertos disjuntos I_i , tal que $\bigcup I_i$ é denso em I . O sistema $Nz_2' + z_2 = h$ é analiticamente equivalente a algum sistema na forma canônica padrão $Mw' + w = f$, onde M é estruturalmente nilpotente.

Definição 2.4.8 A EAD é uma cadeia triangular de r -ésima ordem se tem a forma

$$\begin{bmatrix} A_{11} & 0 & . & . & 0 \\ A_{21} & A_{22} & . & . & . \\ A_{31} & A_{32} & A_{33} & . & . \\ * & * & * & * & 0 \\ A_{r1} & * & * & * & A_{rr} \end{bmatrix} \begin{bmatrix} x'_1 \\ \vdots \\ x'_r \end{bmatrix} + \begin{bmatrix} B_{11} & 0 & . & . & 0 \\ B_{21} & B_{22} & . & . & . \\ B_{31} & * & * & . & . \\ * & * & * & * & 0 \\ B_{r1} & * & * & * & B_{rr} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_r \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_r \end{bmatrix}$$

e cada subsistema $A_{ii}x'_i + B_{ii}x_i = f_i - \sum_{j=1}^{i-1} (A_{ij}x'_j + B_{ij}x_j)$ é de índice zero, ou índice um, ou tem a forma Hessenberg ou está na forma canônica padrão.

Qualquer sistema que é analiticamente equivalente a uma cadeia triangular é solúvel.

2.5 EAD não Lineares

Muitas das EAD importantes são não lineares. A teoria usada para EAD linear com coeficiente variando com o tempo é de grande importância para desenvolvermos uma teoria análoga em sistemas não lineares. A não linearidade terá apenas resultados locais. Também será mais difícil estabelecermos condições necessárias e suficientes, usualmente, teremos que nos satisfazer apenas com condições suficientes.

2.5.1 Solubilidade e Índice

Consideremos a EAD não linear (2.2.1) e o seguinte sistema de equações

$$\begin{aligned} F(t, y, y') &= 0 \\ \frac{d}{dt} F(t, y, y') &= 0 \\ &\vdots \\ \frac{d^{j-1}}{dt^{j-1}} F(t, y, y') &= 0 \end{aligned} \tag{2.5.1}$$

que, ainda pode ser escrita como

$$\begin{aligned} F_{[0]}(t, y, y') &= 0 \\ F_{[1]}(t, y, y', y'') &= 0 \\ &\vdots \\ F_{[j-1]}(t, y, y', \dots, y^{(j-1)}, y^{(j)}) &= 0 \end{aligned} \tag{2.5.2}$$

onde, por exemplo, $F_{[1]} = F_{y'}y'' + F_{y'}y' + F_t$. Podemos escrever (2.5.2), também, na forma

$$F_j(t, y, y_j) = 0, \quad (2.5.3)$$

onde $y_j = [y', \dots, y^{(j)}]'$.

Definição 2.5.1 O índice v de (2.2.1) é o menor número v tal que F_{v+1} determina apenas a variável y' como uma função contínua de y e t .

Definição 2.5.2 O índice local v_l de (2.2.1) para $(\hat{t}, \hat{y}, \hat{y}')$ é o índice do pencil $\lambda F_{y'}(\hat{t}, \hat{y}, \hat{y}') + F_y(\hat{t}, \hat{y}, \hat{y}')$.

Proposição 2.5.1 Suponha que $F_{y'}$ tem posto constante. Então $v = 1$ se, e somente se $v_l \equiv 1$.

Finalmente, temos que a EAD semi explícita

$$\begin{aligned} x_1' &= F_1(x_1, x_2, t) \\ 0 &= F_2(x_1, x_2, t) \end{aligned}$$

tem índice um se, e somente se $\frac{\partial F_2}{\partial x_2}$ é não singular.

2.5.2 Formas Estruturais

Definição 2.5.3 A EAD (2.5.1) possui a forma Hessenberg de dimensão r se pode ser escrita como

$$\begin{aligned} x_1' &= F_1(x_1, x_2, \dots, x_r, t) \\ x_2' &= F_2(x_1, x_2, \dots, x_{r-1}, t) \\ &\vdots \\ x_i' &= F_i(x_{i-1}, x_i, \dots, x_{r-1}, t), \quad 3 \leq i \leq r-1 \\ &\vdots \\ 0 &= F_r(x_{r-1}, t) \end{aligned}$$

e $(\partial F_r / \partial x_{r-1})(\partial F_{r-1} / \partial x_{r-2}) \dots (\partial F_2 / \partial x_1)(\partial F_1 / \partial x_r)$ é não singular.

Assumindo que os F_i são suficientemente diferenciáveis, a forma Hessenberg de dimensão r tem índice e índice local r .

Definição 2.5.4 A EAD (2.2.1) está na forma canônica padrão de dimensão r se pode ser escrita como

$$\begin{aligned}x_1' &= F_1(x_1, t) \\ N(x_1, x_2, t)x_2' &= x_2 + h(x_1, t),\end{aligned}$$

onde N é uma matriz bloco $r \times r$ estritamente triangular inferior.

Definição 2.5.5 A EAD (2.2.1) é uma cadeia triangular inferior de ordem r se tem a forma

$$\begin{aligned}F_1(x_1', x_1, t) &= 0 \\ F_2(x_1', x_2', x_1, x_2, t) &= 0 \\ &\vdots \\ F_r(x_1', \dots, x_r', x_1, \dots, x_r, t) &= 0\end{aligned}$$

e o sistema definido pelas funções F_i é uma EDO implícita, ou tem índice um, ou está na forma Hessenberg, ou na forma canônica padrão em relação a variável x_i .

Uma cadeia triangular é sempre solúvel.

Capítulo 3

Método BDF

3.1 Introdução

Neste capítulo estudaremos as propriedades de convergência e ordem dos métodos de passos múltiplos lineares, em particular, o método BDF aplicado em sistemas de Equações Algébrico-Diferenciais na forma

$$F(t, y, y') = 0, \quad (3.1.1)$$

onde F e y tem a mesma dimensão. Assumiremos que F é continuamente diferenciável.

Definição 3.1.1 $f(x, h)$ é $O(h^p)$ se existe $k > 0$ tal que $\|f(x, h)\| \leq kh^p$ para h suficientemente pequeno.

Veremos, também, que a solução do método BDF de k passos é $O(h^k)$ para sistemas de índice um.

3.2 Métodos de Passos Múltiplos

Aqui, daremos as linhas gerais do trabalho de Griepentrog e März [GRIEPENTROG - 86] sobre as aplicações de métodos de passos múltiplos lineares e métodos one-leg para sistemas de EAD de índice um.

Para a EDO não linear

$$y' = g(t, y) \quad (3.2.1)$$

o método de passo múltiplo linear tem a forma

$$\sum_{j=0}^k a_j y_{n-j} - h \sum_{j=0}^k b_j g(t_{n-j}, y_{n-j}) = 0. \quad (3.2.2)$$

O método one-leg para (3.2.1) envolve apenas uma avaliação de g e tem a forma

$$\frac{1}{h} \sum_{j=0}^k a_j y_{n-j} - g\left(\sum_{j=0}^k b_j t_{n-j}, \sum_{j=0}^k b_j y_{n-j}\right) = 0. \quad (3.2.3)$$

O método BDF pode ser considerado como sendo um método de passo múltiplo linear ou um método one-leg, basta tomar nas equações (3.2.2) e (3.2.3), $b_0 = 1$ e $b_j = 0$, $j = 1, 2, \dots, k$.

A primeira técnica geral para soluções numéricas de EAD, proposta por Gear [GEAR - 71], foi utilizando o método BDF. Este método foi utilizado, inicialmente, em sistemas de equações diferenciais acoplados a equações algébricas

$$\begin{aligned} y' &= f(t, x, y) \\ 0 &= g(t, x, y), \end{aligned} \quad (3.2.4)$$

onde y e g têm a mesma dimensão. O método BDF trata as variáveis algébricas e diferenciais do problema (3.2.4) de maneira análoga. Assim este método pode ser estendido para a generalização de qualquer sistema completamente implícito (3.1.1)

O método BDF mais simples é o método de Euler implícito, que consiste em substituir a derivada em (3.1.1) por uma diferença para trás

$$F\left(t_n, y_n, \frac{y_n - y_{n-1}}{h}\right) = 0, \quad (3.2.5)$$

onde $h = t_i - t_{i-1}$, $i = 1, 2, \dots, n$. O sistema de equações não lineares resultante para y_n em cada passo é usualmente resolvido pelo método de Newton.

O método BDF de k passos consiste em substituir y' pela derivada do polinômio que interpola a solução sobre os pontos $t_n, t_{n-1}, \dots, t_{n-k}$ avaliados em t_n . Assim,

$$F\left(t_n, y_n, \frac{\rho y_n}{h}\right) = 0, \quad (3.2.6)$$

onde $\rho y_n = \sum_{i=0}^k \alpha_i y_{n-i}$ e α_i , $i = 0, 1, \dots, k$, são os coeficientes do método BDF.

Da teoria de equações diferenciais ordinárias temos que o método BDF de k passos é estável para $k \leq 6$ e instável para $k \geq 7$, [LAMBERT - 73].

3.3 Convergência do Método BDF

Aqui, estudaremos os resultados sobre a convergência do método BDF aplicado em sistemas de EAD de índice um. O nosso interesse, em particular nesta seção, é sobre o resultado de convergência do método BDF aplicado ao sistema completamente implícito.

3.3.1 Sistemas Lineares com Coeficientes Constantes

Para o resultado de convergência do método BDF aplicado ao sistema de EAD linear com coeficiente constante, temos o seguinte teorema, cuja demonstração encontra-se em [BRENAN - 89], assim como as demais demonstrações deste capítulo.

Teorema 3.3.1 *O método BDF de k passos, com $k < 7$ e h fixo, aplicado ao sistema de EAD linear com coeficiente constante e índice v é convergente de ordem $O(h^k)$ depois de $(v - 1)k + 1$ passos.*

3.3.2 Sistemas Semi Explícitos

Sistemas semi explícitos de índice um são subconjuntos importantes do sistema de índice um, que surge frequentemente em problemas de Ciências e Engenharia. Estes sistemas são escritos na forma

$$y' = f(t, x, y) \quad (3.3.1a)$$

$$0 = g(t, x, y), \quad (3.3.1b)$$

onde

$$g_y(t, x, y) \quad (3.3.2)$$

é inversível na vizinhança da solução de (3.3.1).

Definição 3.3.1 *Um método de passo múltiplo linear aplicado a (3.3.1) é dado por*

$$\sum_{j=0}^k a_j y_{n-j} = h \sum_{j=0}^k b_j f(x_{n-j}, y_{n-j}, t_{n-j}) \quad (3.3.3a)$$

$$0 = g(x_n, y_n, t_n). \quad (3.3.3b)$$

Os métodos de passos múltiplos aplicados às EAD semi explícitas de índice um são estáveis e convergente para EAD de mesma ordem, assim como no caso de EDO. Este resultado segue do teorema da função implícita. Especificamente, g pode ser resolvida para y_n em termos de x_n , isto é $y_n = G(t_n, x_n)$, então, y_n pode ser inserida na fórmula de

passos múltiplos (3.3.3a) para produzir a mesma equação diferença das obtidas quando aplicamos o método linear de passos múltiplos na EDO

$$y' = f(x, G(x, t), t). \quad (3.3.4)$$

Em princípio, a EAD semi explícita pode ser resolvida por qualquer método de passo múltiplo linear que seja apropriado para a EDO (3.3.4).

3.3.3 Sistemas Completamente Implícitos

Sistemas completamente implícitos de índice um, têm a forma

$$\begin{aligned} F(t, y, y') &= 0 \\ y(t_0) &= y_0. \end{aligned} \quad (3.3.5)$$

Assumiremos que F é uma função suficientemente suave e que, existe uma solução suave $y(t)$ satisfazendo os valores iniciais dados.

Definição 3.3.2 A EAD não linear (3.3.5) é de índice uniforme um se o índice do sistema de coeficientes constantes $Aw'(t) + Bw(t) = g(t)$, onde $A = F_y(\hat{t}, \hat{y}, \hat{y}')$ e $B = F_{y'}(\hat{t}, \hat{y}, \hat{y}')$, é um para todo $(\hat{t}, \hat{y}, \hat{y}')$ na vizinhança do gráfico da solução e, além disso, se

- i) as derivadas parciais de A com respeito a t, y, y' existem e são limitadas na vizinhança da solução;
- ii) o posto de A é constante na vizinhança da solução.

Para iniciarmos a análise de convergência do método BDF, sem perda de generalidade, é suficiente assumirmos que F têm a forma

$$F(t, y, y') = \begin{bmatrix} f(t, y, y') \\ g(t, y) \end{bmatrix}. \quad (3.3.6)$$

Para compreendermos o motivo disto ser verdade, suponha que o posto de F_y é $r < m$ onde m é a dimensão de y . Então existe uma submatriz não singular $r \times r$ de F_y . Suponha que as equações e as variáveis tenham sido reordenadas de modo que f são as primeiras das r equações em F e g são as últimas das $m-r$ equações em F , com

$$\frac{\partial f}{\partial y'} = \begin{bmatrix} \frac{\partial f}{\partial y'_1} & \frac{\partial f}{\partial y'_2} \end{bmatrix},$$

onde $y = [y'_1, y'_2]^T$ e $\partial f / \partial y'_1$ não singular.

Então, pelo teorema da função implícita, f pode ser resolvida para y_1' em termos de y_1, y_2, y_2' e, a expressão para y_1' , substituída nas últimas das $m-r$ equações, para obtermos uma relação implícita entre y_1 e y_2 . A principal observação é que y_2 não precisa envolver esta relação ou, então seríamos capazes de resolver F para a componente adicional de y' , que contraria a suposição do posto de F_y . Assim F pode ser escrita na forma (3.3.6).

A reordenação das equações e variáveis implicadas por (3.3.6) pode não ser suficiente sobre todo intervalo de integração. Entretanto, se F é de índice uniforme um, podemos cobrir o intervalo original com um conjunto finito de subintervalos onde F tem esta forma na vizinhança da solução. Através destes subintervalos, a solução do método BDF não se altera ao reescrevermos F na forma (3.3.6). Assim, se o método BDF converge para (3.3.6) em um subintervalo, então teremos a convergência para (3.3.5) neste subintervalo.

Lema 3.3.1 *Seja y um vetor m -dimensional, então se F é uniforme de índice um na forma (3.3.6), sempre existirá matrizes $P(t, y(t), y'(t))$ e $Q(t, y(t), y'(t))$ tais que*

$$PAQ = \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix}, \quad PBQ = \begin{bmatrix} C(t, y(t), y'(t)) & 0 \\ 0 & I_{m_2} \end{bmatrix} \quad (3.3.7)$$

onde $A(t, y(t), y'(t)) = F_y(t, y(t), y'(t))$, $B(t, y(t), y'(t)) = F_y(t, y(t), y'(t))$, $m = m_1 + m_2$ e satisfazem

- i) $Q(t, y(t), y'(t))$ e $Q^{-1}(t, y(t), y'(t))$ existem e são limitadas para todo $(t, y(t), y'(t))$ resolvendo (3.3.5),
- ii) $Q^{-1}(t_1, y(t_1), y'(t_1)) Q(t_2, y(t_2), y'(t_2)) = I_m + O(t_2 - t_1)$,
- iii) $C(t_1, y(t_1), y'(t_1)) = C(t_2, y(t_2), y'(t_2)) + O(t_2 - t_1)$.

Teorema 3.3.2 *Considere (3.3.5) sendo uma EAD uniforme de índice um no intervalo $I = [t_0, t_0 + T]$. Então a solução numérica de (3.3.5) pelo método BDF de k passos, com $k < 7$ e h fixo, converge com $O(h^k)$ de exatidão se todos valores iniciais são $O(h^k)$ e se a iteração de Newton em cada passo é resolvida com $O(h^{k+1})$ de exatidão.*

A demonstração deste teorema encontra-se com mais detalhes em [PIRES - 94].

Capítulo 4

Método Runge-Kutta Implícito

4.1 Introdução

Estudaremos aqui, as propriedades de ordem, estabilidade e convergência do método de RKI aplicado a sistemas de EAD na forma

$$F(t, y, y') = 0, \quad (4.1.1)$$

onde os valores iniciais $y(t_0)$ e $y'(t_0)$ são dados.

Muitas vezes, estes métodos não alcançam a mesma ordem de exatidão em sistemas algébricos-diferenciais como o fazem em sistemas puramente diferenciais.

Muitos resultados deste capítulo, foram obtidos em [PETZOLD - 86], [HAIRER - 89] e [HAIRER - 91].

O método RKI para solução da EDO

$$y' = f(t, y) \quad (4.1.2)$$

é dado por

$$\begin{aligned} Y_i &= f\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j\right), \quad i = 1, 2, \dots, M \\ y_n &= y_{n-1} + h \sum_{i=1}^M b_i Y_i \end{aligned} \quad (4.1.3)$$

onde $h = t_n - t_{n-1}$.

Considerando este método aplicado ao sistema de EAD (4.1.1), temos

$$\begin{aligned} F\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j, Y_i\right) &= 0, \quad i = 1, 2, \dots, M \\ y_n &= y_{n-1} + h \sum_{i=1}^M b_i Y_i. \end{aligned} \quad (4.1.4)$$

O método de RKI, também, pode ser denotado pelo diagrama de Butcher

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1M} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_M & a_{M1} & a_{M2} & \cdots & a_{MM} \\ \hline & b_1 & b_2 & \cdots & b_M \end{array} = \begin{array}{c|c} c & \mathcal{A} \\ \hline & b' \end{array} \quad (4.1.5)$$

onde $\mathcal{A} = (a_{ij})$, $b = [b_1, b_2, \dots, b_M]^t$ e $c = [c_1, c_2, \dots, c_M]^t$.

Consideraremos apenas métodos cuja matriz Runge-Kutta $\mathcal{A} = (a_{ij})$, em (4.1.5), seja não singular.

Nos preocuparemos aqui, apenas com sistemas de EAD de índice um. Uma EAD linear

$$A(t)y'(t) + B(t)y(t) = g(t), \quad (4.1.6)$$

tem índice um se existem matrizes, dependendo do tempo, não singulares P e Q tais que

$$P(t)A(t)Q(t) = \begin{bmatrix} I_{m1} & 0 \\ 0 & 0 \end{bmatrix}, \quad P(t)B(t)Q(t) = \begin{bmatrix} C(t) & 0 \\ 0 & I_{m2} \end{bmatrix} \quad (4.1.7)$$

Esta transformação separa o sistema em partes "diferencial" e "algébrica". No caso de sistemas não lineares, associamos as matrizes A e B com F_y e F_y , respectivamente.

4.2 Sistemas Lineares com Coeficientes Constantes

Derivaremos condições necessárias e suficientes para assegurar que o erro de truncamento local do método de RKI alcance a ordem quando aplicado a um sistema linear com coeficiente constante.

Considere o sistema de EAD (4.1.1) e o método de RKI de M-estágios (4.1.4) aplicado a este sistema onde a matriz $A = (a_{ij})$ é não singular. Sendo y_n determinado por (4.1.4), y_{n-1} e t_{n-1} conhecidos, o método de RKI pode ser escrito na forma

$$y_n = y_{n-1} + h\psi(t_{n-1}, y_{n-1}, h). \quad (4.2.1)$$

Partindo deste procedimento necessitamos de algumas definições.

Definição 4.2.1 O erro local d_n do método (4.1.4) é dado por

$$y(t_n) = y(t_{n-1}) + h\psi(t_{n-1}, y(t_{n-1}), h) - d_n. \quad (4.2.2)$$

Definição 4.2.2 O método de RKI (4.1.4) é estritamente estável para a EAD (4.1.1) se a diferença entre o passo Runge-Kutta perturbado

$$F\left(t_{n-1} + c_i h, z_{n-1} + h \sum_{j=1}^M a_{ij} Z_j' + \delta_n^{(i)}, Z_i'\right) = 0, \quad i = 1, 2, \dots, M \quad (4.2.3)$$

$$z_n = z_{n-1} + h \sum_{i=1}^M b_i Z_i' + \delta_n^{(M+1)}$$

sendo $z_0 = y_0 + \delta_0^{(M+1)}$, $\|\delta_n^{(i)}\| = \Delta$, $i = 1, 2, \dots, (M+1)$, e o passo Runge-Kutta (4.1.4) satisfazem $\|z_n - y_n\| \leq K_0 \Delta$, onde $0 < h \leq h_0$ e K_0, h_0 são constantes que dependem apenas do método e da EAD.

Consideraremos a EAD linear com coeficiente constante

$$Ay' + By = g(t). \quad (4.2.4)$$

Assumindo que o índice é um, então existem matrizes não singulares P e Q tais que

$$PAQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad PBQ = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}. \quad (4.2.5)$$

Aplicando o método de RKI em (4.2.4), obtemos

$$AY'_i + B \left(y_{n-1} + h \sum_{j=1}^M a_{ij} Y'_j \right) = g(t_{n-1} + c_i h), \quad i = 1, 2, \dots, M \quad (4.2.6)$$

$$y_n = y_{n-1} + h \sum_{i=1}^M b_i Y'_i.$$

Considerando $\tilde{y}_n = Q^{-1} y_n$, $\tilde{Y}'_i = Q^{-1} Y'_i$, $[\tilde{g}_1, \tilde{g}_2]' = \tilde{g}(t) = Pg(t)$ e pré-multiplicando (4.2.6) por P , obtemos

$$PAQ\tilde{Y}'_i + PBQ \left(\tilde{y}_{n-1} + h \sum_{j=1}^M a_{ij} \tilde{Y}'_j \right) = \tilde{g}(t_{n-1} + c_i h), \quad (4.2.7)$$

$$\tilde{y}_n = \tilde{y}_{n-1} + h \sum_{i=1}^M b_i \tilde{Y}'_i.$$

Desta forma as partes diferencial e algébrica do sistema são completamente desacopladas. Assim, é suficiente estudarmos separadamente estas partes para obtermos um entendimento geral das EAD lineares com coeficientes constantes.

Desde que tenhamos assumido que o índice é um, o sistema (4.2.4) divide-se em um sistema de equações diferenciais e em um sistema de equações algébricas, onde o sistema de equações algébricas é

$$\tilde{y}_2 = \tilde{g}_2(t), \quad (4.2.8)$$

que por simplicidade denotaremos $\tilde{y}_2 = y$ e $\tilde{g}_2 = g$.

Aplicando o método de RKI, definido em (4.1.4), na equação (4.2.8), temos

$$y_{n-1} + h \sum_{j=1}^M a_{ij} Y'_j = g(t_{n-1} + c_i h), \quad i = 1, 2, \dots, M \quad (4.2.9a)$$

$$y_n = y_{n-1} + h \sum_{i=1}^M b_i Y'_i. \quad (4.2.9b)$$

Devido a $\mathcal{A} = (a_{ij})$ ser não singular, determinamos $Y' = [Y'_1, Y'_2, \dots, Y'_M]'$, unicamente, por (4.2.9a).

Para encontrar o erro local consideraremos

$$y_{n-1} = g(t_{n-1}), \quad (4.2.10)$$

assim, substituindo (4.2.10) em (4.2.9a), obtemos

$$g(t_{n-1}) + h \sum_{j=1}^M a_{ij} Y'_j = g(t_{n-1} + c_i h)$$

ou

$$\sum_{j=1}^M a_{ij} Y'_j = \frac{1}{h} (g(t_{n-1} + c_i h) - g(t_{n-1})),$$

que fornece o sistema de equações

$$\begin{aligned} a_{11}Y'_1 + a_{12}Y'_2 + \dots + a_{1M}Y'_M &= (g(t_{n-1} + c_1 h) - g(t_{n-1}))/h \\ a_{21}Y'_1 + a_{22}Y'_2 + \dots + a_{2M}Y'_M &= (g(t_{n-1} + c_2 h) - g(t_{n-1}))/h \\ &\vdots \\ a_{M1}Y'_1 + a_{M2}Y'_2 + \dots + a_{MM}Y'_M &= (g(t_{n-1} + c_M h) - g(t_{n-1}))/h \end{aligned}$$

cujas forma matricial é

$$\begin{bmatrix} a_{11} & \dots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{M1} & \dots & a_{MM} \end{bmatrix} \begin{bmatrix} Y'_1 \\ \vdots \\ Y'_M \end{bmatrix} = \begin{bmatrix} (g(t_{n-1} + c_1 h) - g(t_{n-1}))/h \\ \vdots \\ (g(t_{n-1} + c_M h) - g(t_{n-1}))/h \end{bmatrix}.$$

Sendo $\mathcal{A} = (a_{ij})$ não singular e considerando $\mathcal{A}^{-1} = (\tilde{a}_{ij})$ temos

$$\begin{bmatrix} Y'_1 \\ \vdots \\ Y'_M \end{bmatrix} = \begin{bmatrix} \tilde{a}_{11} & \dots & \tilde{a}_{1M} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{M1} & \dots & \tilde{a}_{MM} \end{bmatrix} \begin{bmatrix} (g(t_{n-1} + c_1 h) - g(t_{n-1}))/h \\ \vdots \\ (g(t_{n-1} + c_M h) - g(t_{n-1}))/h \end{bmatrix},$$

que pode ser escrito na forma

$$Y' = \mathcal{A}^{-1}G, \quad (4.2.11a)$$

onde

$$G = \begin{bmatrix} (g(t_{n-1} + c_1 h) - g(t_{n-1}))/h \\ \vdots \\ (g(t_{n-1} + c_M h) - g(t_{n-1}))/h \end{bmatrix}. \quad (4.2.11b)$$

Agora, substituindo (4.2.10) na equação (4.2.9b), obtemos

$$y_n = g(t_{n-1}) + h \sum_{i=1}^M b_i Y'_i = g(t_{n-1}) + h(b_1 Y'_1 + b_2 Y'_2 + \dots + b_M Y'_M)$$

$$= g(t_{n-1}) + h[b_1, \dots, b_M] \begin{bmatrix} Y_1' \\ \vdots \\ Y_M' \end{bmatrix} = g(t_{n-1}) + hb^t \mathcal{A}^{-1} G. \quad (4.2.12)$$

Assim, o erro local é dado por

$$d_n = g(t_{n-1}) + hb^t \mathcal{A}^{-1} G - g(t_n). \quad (4.2.13)$$

Expandindo os termos de (4.2.13) em série de Taylor em torno de t_{n-1} temos

$$d_n = g(t_{n-1}) + hb^t \mathcal{A}^{-1} \begin{bmatrix} \left[\left(g(t_{n-1}) + g'(t_{n-1})c_1 h + g''(t_{n-1}) \frac{(c_1 h)^2}{2} + \dots \right) - g(t_{n-1}) \right] / h \\ \vdots \\ \left[\left(g(t_{n-1}) + g'(t_{n-1})c_M h + g''(t_{n-1}) \frac{(c_M h)^2}{2} + \dots \right) - g(t_{n-1}) \right] / h \end{bmatrix} \\ - \left(g(t_{n-1}) + g'(t_{n-1})(t_n - t_{n-1}) + g''(t_{n-1}) \frac{(t_n - t_{n-1})^2}{2} + \dots \right)$$

sendo $h = t_n - t_{n-1}$ obtemos

$$d_n = - \left(hg' + \frac{h^2}{2} g'' + \frac{h^3}{6} g''' + \dots \right) + b^t \mathcal{A}^{-1} \begin{bmatrix} c_1 hg' + \frac{(c_1 h)^2}{2} g'' + \frac{(c_1 h)^3}{6} g''' + \dots \\ \vdots \\ c_M hg' + \frac{(c_M h)^2}{2} g'' + \frac{(c_M h)^3}{6} g''' + \dots \end{bmatrix} \\ = - \sum_{r=1}^k \frac{h^r}{r!} g^{(r)}(t_{n-1}) + b^t \mathcal{A}^{-1} \sum_{r=1}^k \frac{c^j h^r}{r!} g^{(r)}(t_{n-1}) + R_d$$

onde $R_d = - \sum_{r=k+1}^{\infty} \frac{h^r}{r!} g^{(r)}(t_n) (1 - b^t \mathcal{A}^{-1} c^j)$ e $c^j = [c_1^j, c_2^j, \dots, c_M^j]^t$, $j = 1, 2, \dots, k$, assim

$$d_n = - \sum_{r=1}^k \frac{h^r}{r!} g^{(r)}(t_{n-1}) (1 - b^t \mathcal{A}^{-1} c^j) + R_d$$

supondo $b^t \mathcal{A}^{-1} c^j = 1$, $j = 1, 2, \dots, k$ (faremos um comentário mais geral desta condição na seção 4.3), temos $d_n = R_d$ e

$$R_d = - \sum_{r=k+1}^{\infty} \frac{h^r}{r!} g^{(r)}(t_n) (1 - b^t \mathcal{A}^{-1} c^j) = \frac{h^{k+1}}{(k+1)!} g^{(k+1)}(\xi) (1 - b^t \mathcal{A}^{-1} c^{k+1}).$$

Considerando que $\|g^{(k+1)}\| \leq M_{k+1}$, podemos majorar R_d da seguinte forma

$$\|R_d\| \leq \frac{h^{k+1}}{(k+1)!} M_{k+1} (1 - |b' \mathcal{A}^{-1} c^{k+1}|). \quad (4.2.14)$$

Portanto, $d_n = O(h^{k+1})$ se $b' \mathcal{A}^{-1} c^j = 1$, $j = 1, 2, \dots, k$.

Definição 4.2.3 A ordem algébrica do método de RKI (4.2.11) é igual a k_a se $d_n = O(h^{k_a+1})$, para toda equação (4.2.8) com $g(t)$ suficientemente suave.

Teorema 4.2.1 A ordem algébrica do método de RKI (4.1.4) é k_a se, e somente se os coeficientes do método satisfazem

$$b' \mathcal{A}^{-1} c^j = 1, \quad j = 1, 2, \dots, k_a. \quad (4.2.15)$$

A seguir, examinaremos a propagação dos erros para o método de RKI aplicado em sistemas lineares com coeficientes constantes [PETZOLD - 86]. Resolvendo (4.2.10) para o método de RKI com perturbação (4.2.3) temos

$$\begin{aligned} z_{n-1} + h \sum_{j=1}^M a_{ij} Z'_j + \delta_n^{(i)} &= g(t_{n-1} + c_i h), \quad i = 1, 2, \dots, M \\ z_n &= z_{n-1} + h \sum_{i=1}^M b_i Z'_i + \delta_n^{(M+1)}. \end{aligned} \quad (4.2.16)$$

Subtraindo (4.2.16) de (4.2.9), considerando $e_n = y_n - z_n$ e $E'_i = Y'_i - Z'_i$, obtemos

$$e_{n-1} + h \sum_{j=1}^M a_{ij} E'_j - \delta_n^{(i)} = 0, \quad i = 1, 2, \dots, M \quad (4.2.17a)$$

$$e_n = e_{n-1} + h \sum_{i=1}^M b_i E'_i - \delta_n^{(M+1)}. \quad (4.2.17b)$$

Resolvendo (4.2.17a) para E'_i temos

$$\begin{aligned} h[a_{11}E'_1 + a_{12}E'_2 + \dots + a_{1M}E'_M] &= \delta_n^{(1)} + e_{n-1} \\ &\vdots \\ h[a_{M1}E'_1 + a_{M2}E'_2 + \dots + a_{MM}E'_M] &= \delta_n^{(M)} + e_{n-1}. \end{aligned}$$

Matricialmente temos

$$h \begin{bmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MM} \end{bmatrix} \begin{bmatrix} E'_1 \\ \vdots \\ E'_M \end{bmatrix} = \begin{bmatrix} \delta_n^{(1)} \\ \vdots \\ \delta_n^{(M)} \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} e_{n-1},$$

ainda, podendo ser escrito como

$$E'_i = \frac{1}{h} \mathcal{A}^{-1} (\delta_n - \epsilon_M e_{n-1}), \quad (4.2.18)$$

onde $E'_i = [E'_1, E'_2, \dots, E'_M]^t$, $\delta_n = [\delta_n^{(1)}, \delta_n^{(2)}, \dots, \delta_n^{(M)}]^t$ e $\epsilon_M = [1, 1, \dots, 1]^t$.

Substituindo em (4.2.17b) temos

$$e_n = (1 - b^t \mathcal{A}^{-1} \epsilon_M) e_{n-1} + b^t \mathcal{A}^{-1} \delta_n - \delta_n^{(M+1)} \quad (4.2.19)$$

Assim, mostramos o seguinte resultado:

Teorema 4.2.2 *O método de RKI (4.1.4) é estritamente estável para EAD lineares com coeficientes constantes de índice um se, e somente se os coeficientes do método satisfazem*

$$|1 - b^t \mathcal{A}^{-1} \epsilon_M| < 1. \quad (4.2.20)$$

Definindo a constante de estabilidade r por $r = 1 - b^t \mathcal{A}^{-1} \epsilon_M$, dizemos que o método de RKI satisfaz a condição estrita de estabilidade se $|r| < 1$ e é satisfeita a condição de estabilidade se $|r| \leq 1$.

A estimativa para o erro local torna-se simples, para sistemas de índice um, agora que conhecemos o erro de truncamento local e a propriedade de estabilidade do método. Considerando que δ_n e $\delta_n^{(M+1)}$ representam, em cada estágio, o erro de truncamento local, excluindo o erro de arredondamento e, de acordo com a suposição do erro local (4.2.10), segue por (4.2.20) que

$$e_n = (b^t \mathcal{A}^{-1} \delta_n - \delta_n^{(M+1)}). \quad (4.2.21)$$

Como o erro local d_n foi denotado como sendo $O(h^{k_a+1})$ e de acordo com a suposição de que $e_n = d_n$ temos

$$b^t \mathcal{A}^{-1} \delta_n - \delta_n^{(M+1)} = O(h^{k_a+1}).$$

Em seguida, reescrevemos a expressão para o erro global dado na equação (4.2.19), isto é,

$$e_n = re_{n-1} + (b' \mathcal{A}^{-1} \delta_n - \delta_n^{(M+1)}).$$

Assim,

$$\begin{aligned} e_{n-1} &= re_{n-2} + (b' \mathcal{A}^{-1} \delta_{n-1} - \delta_{n-1}^{(M+1)}) \\ e_{n-2} &= re_{n-3} + (b' \mathcal{A}^{-1} \delta_{n-2} - \delta_{n-2}^{(M+1)}) \\ &\vdots \\ e_1 &= re_0 + (b' \mathcal{A}^{-1} \delta_1 - \delta_1^{(M+1)}) \end{aligned}$$

Portanto, e_n torna-se

$$e_n = r^n e_0 + \sum_{i=0}^{n-1} r^i (b' \mathcal{A}^{-1} \delta_{n-i} - \delta_{n-i}^{(M+1)}). \quad (4.2.22)$$

Observe que, o erro global será igual ao erro local se $r = 0$ em (4.2.22). Para $|r| < 1$, segue de (4.2.22) que

$$\|e_n\| \leq |r|^n \|e_0\| + \left| \frac{1-r^n}{1-r} \right| M h^{k_a+1},$$

onde M é uma constante positiva. Visto que r independe de h e $\lim_{n \rightarrow \infty} \frac{1-r^n}{1-r} = \frac{1}{1-r}$ temos, neste caso, que a ordem do erro global é a mesma do erro local na parte "algébrica do sistema". A partir disto temos os seguintes resultados.

Definição 4.2.4 A ordem do método de RKI (4.1.4) aplicado a sistemas de EAD com coeficientes constantes, é igual a k_c se o método converge com erro global $O(h^{k_c})$ para todo sistema linear com coeficientes constantes (4.2.4) de índice um e com $g(t)$ suficientemente suave.

Teorema 4.2.3 Suponha que o método de RKI (4.1.4) satisfaça a condição de estabilidade (4.2.20) e a matriz $\mathcal{A}$ dos coeficientes do método é não singular. Então a ordem do erro global do método de RKI, aplicado a sistemas de EAD lineares com coeficientes constantes, é k_c , dado por

$$k_c = \min(k_a + 1, k_d), \quad (4.2.23)$$

onde k_d é a ordem do método para sistemas puramente diferenciais.

4.3 Sistemas não Lineares

Nesta seção, consideraremos as propriedades de ordem, estabilidade e convergência dos métodos de RKI, aplicados a sistemas de EAD completamente implícitos de índice um, na forma $F(t, y, y') = 0$, onde os valores iniciais, consistentes, para $y(t_0)$ e $y'(t_0)$ são dados.

Os métodos de RKI, em geral, não obtém a mesma ordem de exatidão para sistemas completamente implícitos de índice um como ocorre em sistemas lineares. Esta perda adicional de exatidão deve-se ao fato da mistura que pode ocorrer entre a solução da parte diferencial e da parte algébrica do sistema.

Definição 4.3.1 O i -ésimo erro de truncamento local $\delta_i^{(n)}$, em t_n , do método de RKI de M -estágios é dado por

$$\delta_i^{(n)} = y(t_{n-1}) + h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) - y(t_{n-1} + c_i h), \quad i = 1, 2, \dots, M \quad (4.3.1a)$$

$$\delta_{M+1}^{(n)} = y(t_{n-1}) + h \sum_{i=1}^M b_i y'(t_{n-1} + c_i h) - y(t_n). \quad (4.3.1b)$$

Definição 4.3.2 A condição de ordem k_I , do método de RKI de M -estágios (4.1.4), é dada por

$$k_I = \min(k_1, k_2, \dots, k_M, k_{M+1}), \quad (4.3.2)$$

com $\delta_i^{(n)} = O(h^{k_i+1})$, $i = 1, 2, \dots, (M+1)$.

Considere $c_i = \sum_{j=1}^M a_{ij}$ e as condições algébricas $C(q)$, $B(q)$, $D(q)$ e $A_1(q)$

$$C(q): \sum_{j=1}^M a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \quad i = 1, 2, \dots, M, \quad k = 1, 2, \dots, q$$

$$B(q): \sum_{i=1}^M b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, 2, \dots, q \quad (4.3.3)$$

$$D(q): \sum_{i=1}^M b_i c_i^{k-1} a_{ij} = \frac{b_j}{k} (1 - c_j^k), \quad j = 1, 2, \dots, M, \quad k = 1, 2, \dots, q$$

$$A_1(q): b^T A^{-1} c^k = 1, \quad k = 1, 2, \dots, q.$$

O método de RKI tem condição de ordem k_j se $C(k_j)$ e $B(k_j)$ forem válidos. Além disso, se $C(q)$ e $B(q+1)$ forem válidos, a ordem k_d da EDO é pelo menos $q+1$. $A_1(q)$ corresponde a condição de ordem para sistemas lineares com coeficientes constantes de índice um (ver seção 4.2). Isto pode ser visto, facilmente pela expansão de (4.3.1) em série de Taylor em torno de t_{n-1} (veja [FRANK - 85a] e [FRANK - 85b]).

Por meio das equações algébricas $C(q)$, $B(q)$ e $D(q)$ podemos obter várias classes de métodos. Uma classe destes métodos é o RADAU IIA, que será discutido no Capítulo 5.

Aqui, mostraremos alguns resultados de importância fundamental para a demonstração do principal teorema deste capítulo, fornecendo condições sobre os coeficientes do método de RKI para convergência quando aplicado num sistema de EAD de índice um.

Definição 4.3.3 *Sejam as matrizes $A_{m \times n}$ e $B_{r \times s}$. O produto tensorial de Kronecker é a matriz de dimensão $m.r \times n.s$ definida por*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Lema 4.3.1 *Sejam as matrizes A_M inversível e I a matriz identidade de dimensão n . Então $(A \otimes I)^{-1} = A^{-1} \otimes I$.*

Demonstração: Pela definição de produto tensorial, temos

$$A \otimes I = \begin{bmatrix} a_{11}I & \cdots & a_{1M}I \\ \vdots & \ddots & \vdots \\ a_{M1}I & \cdots & a_{MM}I \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{11} \end{pmatrix} & \cdots & \begin{pmatrix} a_{1M} & & \\ & \ddots & \\ & & a_{1M} \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} a_{M1} & & \\ & \ddots & \\ & & a_{M1} \end{pmatrix} & \cdots & \begin{pmatrix} a_{MM} & & \\ & \ddots & \\ & & a_{MM} \end{pmatrix} \end{bmatrix}$$

Agora, considerando P como sendo a matriz de permutação

$$P = [e_1, e_{n+1}, e_{2n+1}, \dots, e_{(M-1)n+1}, e_2, e_{n+2}, e_{2n+2}, \dots, e_{(M-1)n+2}, \dots, e_n, e_{2n}, e_{3n}, \dots, e_{Mn}],$$

onde e_i são os vetores da base canônica de R^{Mn} , então

$$P(\mathcal{A} \otimes I)P = \begin{bmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{11} \end{pmatrix} & \cdots & \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} & \cdots & \begin{pmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MM} \end{pmatrix} \end{bmatrix} = \begin{bmatrix} \mathcal{A} & & \\ & \ddots & \\ & & \mathcal{A} \end{bmatrix}.$$

Mas

$$P(\mathcal{A} \otimes I)^{-1}P = [P(\mathcal{A} \otimes I)P]^{-1} = \begin{bmatrix} \mathcal{A}^{-1} & & \\ & \ddots & \\ & & \mathcal{A}^{-1} \end{bmatrix} = P(\mathcal{A}^{-1} \otimes I)P.$$

Portanto, $(\mathcal{A} \otimes I)^{-1} = (\mathcal{A}^{-1} \otimes I)$ ■

Lema 4.3.2 *Seja*

$$\bar{T}_n = \begin{bmatrix} \hat{T}_1 + O(h) & h^2 T_2 \\ h^2 T_3 & h(\hat{T}_4 + O(h)) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & hI \end{bmatrix} \begin{bmatrix} \hat{T}_1 + O(h) & h^2 T_2 \\ hT_3 & \hat{T}_4 + O(h) \end{bmatrix},$$

onde $\mathcal{A}$ é inversível, $\hat{T}_1 = I + h\mathcal{A} \otimes C$, $\hat{T}_4 = \mathcal{A} \otimes I$, T_2 e T_3 são matrizes cujos elementos tem $O(1)$. Então

$$\bar{T}_n^{-1} = \begin{bmatrix} \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix},$$

para h suficientemente pequeno.

Demonstração: Considerando $X = O(h) = hE$, onde E é limitada, então $(\hat{T}_1 + X)^{-1} = (I + h\mathcal{A} \otimes C + hE)^{-1} = (I + h(\mathcal{A} \otimes C + E))^{-1} = \sum_{i=0}^{\infty} (-1)^i [h(\mathcal{A} \otimes C + E)]^i$ para $h < \|\mathcal{A} \otimes C + E\|$. Desta forma $(\hat{T}_1 + X)$ é inversível, mais ainda, para $h < \|\mathcal{A} \otimes C\|$ temos que $\hat{T}_1$ é inversível. Agora, $\hat{T}_4$ é inversível pelo fato da matriz $\mathcal{A}$ ser inversível e pelo Lema 4.3.1.

Sejam X e Y de $O(h)$, logo

$$\bar{T}_n = \begin{bmatrix} I & 0 \\ 0 & hI \end{bmatrix} \begin{bmatrix} \hat{T}_1 + X & h^2 T_2 \\ hT_3 & \hat{T}_4 + Y \end{bmatrix}.$$

Assim,

$$\bar{T}_n^{-1} = \begin{bmatrix} \hat{T}_1 + X & h^2 T_2 \\ h T_3 & \hat{T}_4 + Y \end{bmatrix}^{-1} \begin{bmatrix} I & 0 \\ 0 & Ih^{-1} \end{bmatrix}.$$

Sabemos que $(\hat{T}_1 + X)$ é inversível, logo podemos escrever

$$\begin{aligned} (\hat{T}_1 + X)^{-1} &= (\hat{T}_1 \hat{T}_1^{-1} \hat{T}_1 + \hat{T}_1 \hat{T}_1^{-1} X)^{-1} = [\hat{T}_1 (I + \hat{T}_1^{-1} X)]^{-1} = (I + \hat{T}_1^{-1} X)^{-1} \hat{T}_1^{-1} \\ &= (I - (-\hat{T}_1^{-1} X))^{-1} \hat{T}_1^{-1} = \left[\sum_{i=0}^{\infty} (-\hat{T}_1^{-1} X)^i \right] \hat{T}_1^{-1} \\ &= (I - \hat{T}_1^{-1} X + (\hat{T}_1^{-1} X)^2 - (\hat{T}_1^{-1} X)^3 + \dots) \hat{T}_1^{-1} \\ &= \hat{T}_1^{-1} - \hat{T}_1^{-1} X \hat{T}_1^{-1} + (\hat{T}_1^{-1} X)^2 \hat{T}_1^{-1} - (\hat{T}_1^{-1} X)^3 \hat{T}_1^{-1} + \dots = \hat{T}_1^{-1} + A, \end{aligned}$$

onde A é $O(h)$. De maneira análoga temos $(\hat{T}_4 + Y)^{-1} = \hat{T}_4^{-1} + D$, onde D é $O(h)$.

Mostraremos que $\bar{T}_n$ é inversível e que

$$\bar{T}_n^{-1} = \begin{bmatrix} \hat{T}_1^{-1} + A & B \\ C & \hat{T}_4^{-1} + D \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Ih^{-1} \end{bmatrix} = \begin{bmatrix} \hat{T}_1^{-1} + A & h^{-1} B \\ C & \frac{\hat{T}_4^{-1}}{h} + \frac{D}{h} \end{bmatrix}. \quad (4.3.4)$$

De fato,

$$\begin{bmatrix} \hat{T}_1 + X & h^2 T_2 \\ h^2 T_3 & h(\hat{T}_4 + Y) \end{bmatrix} \begin{bmatrix} \hat{T}_1^{-1} + A & h^{-1} B \\ C & \frac{\hat{T}_4^{-1}}{h} + \frac{D}{h} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Assim,

$$\begin{aligned} &\begin{bmatrix} (\hat{T}_1 + X)(\hat{T}_1^{-1} + A) + h^2 T_2 C & (\hat{T}_1 + X)h^{-1} B + h T_2 \hat{T}_4^{-1} + h T_2 D \\ h^2 T_3(\hat{T}_1^{-1} + A) + h(\hat{T}_4 + Y)C & h T_3 B + (\hat{T}_4 + Y)(\hat{T}_4^{-1} + D) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I + \hat{T}_1 A + X \hat{T}_1^{-1} + X A + h^2 T_2 C & h^{-1} \hat{T}_1 B + h^{-1} X B + h T_2 \hat{T}_4^{-1} + h^2 T_2 D \\ h^2 T_3 \hat{T}_1^{-1} + h^2 T_3 A + h \hat{T}_4 C + h Y C & h T_3 B + I + \hat{T}_4 D + Y \hat{T}_4^{-1} + Y D \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned}$$

Portanto, obtemos as seguintes equações

$$I + \hat{T}_1 A + X \hat{T}_1^{-1} + X A + h^2 T_2 C = I \quad (4.3.5a)$$

$$h^{-1} \hat{T}_1 B + h^{-1} X B + h T_2 \hat{T}_4^{-1} + h T_2 D = 0 \quad (4.3.5b)$$

$$h^2 T_3 \hat{T}_1^{-1} + h^2 T_3 A + h \hat{T}_4 C + h Y C = 0 \quad (4.3.5c)$$

$$h T_3 B + I + \hat{T}_4 D + Y \hat{T}_4^{-1} + Y D = I \quad (4.3.5d)$$

De (4.3.5a) temos

$$\begin{aligned} I + \hat{T}_1 A + X \hat{T}_1^{-1} + X A + h^2 T_2 C &= I \\ (\hat{T}_1 + X) A &= -(X \hat{T}_1^{-1} + h^2 T_2 C) \\ A &= -(\hat{T}_1 + X)^{-1} (X \hat{T}_1^{-1} + h^2 T_2 C) \end{aligned} \quad (4.3.6)$$

Substituindo (4.3.6) em (4.3.5c) obtemos

$$\begin{aligned} h^2 T_3 \hat{T}_1^{-1} - h^2 T_3 (\hat{T}_1 + X)^{-1} (X \hat{T}_1^{-1} + h^2 T_2 C) + h \hat{T}_4 C + h Y C &= 0 \\ h^2 T_3 \hat{T}_1^{-1} - h^2 T_3 (\hat{T}_1 + X)^{-1} X \hat{T}_1^{-1} - h^4 T_3 (\hat{T}_1 + X)^{-1} T_2 C + h \hat{T}_4 C + h Y C &= 0 \\ (h \hat{T}_4 + h Y - h^4 T_3 (\hat{T}_1 + X)^{-1} T_2) C &= h^2 T_3 (\hat{T}_1 + X)^{-1} X \hat{T}_1^{-1} - h^2 T_3 (\hat{T}_1 + X)^{-1} X \hat{T}_1^{-1} \\ (\hat{T}_4 + Y - h^3 T_3 (\hat{T}_1 + X)^{-1} T_2) C &= h T_3 (\hat{T}_1 + X)^{-1} X \hat{T}_1^{-1} - h T_3 (\hat{T}_1 + X)^{-1} X \hat{T}_1^{-1} \end{aligned}$$

e observando as ordens, de cada lado da igualdade, temos $C = O(h)$. Voltando em (4.3.6), temos $A = O(h)$.

De (4.3.5b) temos

$$\begin{aligned} h^{-1} \hat{T}_1 B + h^{-1} X B + h T_2 \hat{T}_4^{-1} + h T_2 D &= 0 \\ (\hat{T}_1 + X) B &= -h^2 T_2 (\hat{T}_4^{-1} + D) \\ B &= -(\hat{T}_1 + X)^{-1} h^2 T_2 (\hat{T}_4^{-1} + D) \end{aligned}$$

$$= -(\hat{T}_1 + X)^{-1} h^2 T_2 \hat{T}_4^{-1} - (\hat{T}_1 + X)^{-1} h^2 T_2 D \quad (4.3.7)$$

Substituindo (4.3.7) em (4.3.5d) obtemos

$$-h^3 T_3 (\hat{T}_1 + X)^{-1} T_2 (\hat{T}_4^{-1} + D) + \hat{T}_4 D + Y \hat{T}_4^{-1} + Y D = 0$$

$$-h^3 T_3 (\hat{T}_1 + X)^{-1} T_2 \hat{T}_4^{-1} - h^3 T_3 (\hat{T}_1 + X)^{-1} T_2 D + \hat{T}_4 D + Y \hat{T}_4^{-1} + Y D = 0$$

$$\left(\hat{T}_4 + Y - h^3 T_3 (\hat{T}_1 + X)^{-1} T_2 \right) D = h^3 T_3 (\hat{T}_1 + X)^{-1} T_2 \hat{T}_4^{-1} - Y \hat{T}_4^{-1}$$

e observando as ordens, de cada igualdade, temos $D = O(h)$. Voltando em (4.3.7), temos $B = O(h^2)$.

Portanto, de (4.3.4), temos

$$\bar{T}_n^{-1} = \begin{bmatrix} \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \blacksquare$$

Lema 4.3.3 Sejam $r = 1 - b' \mathcal{A}^{-1} \epsilon_M$, $\mathcal{A}^{-1} = (\tilde{a}_{ij})$ $i, j = 1, \dots, M$, $b' = [b_1, \dots, b_M]$ e $\epsilon_M = [1, \dots, 1]^t$. Então $(b' \otimes I_{m_2})(\mathcal{A} \otimes I_{m_2})^{-1}(\epsilon_M \otimes I_{m_2}) = (1 - r)I_{m_2}$.

Demonstração:

$$\begin{aligned} (b' \otimes I_{m_2})(\mathcal{A} \otimes I_{m_2})^{-1}(\epsilon_M \otimes I_{m_2}) &= (b' \otimes I_{m_2})(\mathcal{A}^{-1} \otimes I_{m_2})(\epsilon_M \otimes I_{m_2}) \\ &= [b_1 I_{m_2}, \dots, b_M I_{m_2}] \begin{bmatrix} \tilde{a}_{11} I_{m_2} & \cdots & \tilde{a}_{1M} I_{m_2} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{M1} I_{m_2} & \cdots & \tilde{a}_{MM} I_{m_2} \end{bmatrix} \begin{bmatrix} 1 I_{m_2} \\ \vdots \\ 1 I_{m_2} \end{bmatrix} \\ &= [b_1 I_{m_2}, \dots, b_M I_{m_2}] \begin{bmatrix} (\tilde{a}_{11} + \cdots + \tilde{a}_{1M}) I_{m_2} \\ \vdots \\ (\tilde{a}_{M1} + \cdots + \tilde{a}_{MM}) I_{m_2} \end{bmatrix} \\ &= [b_1 (\tilde{a}_{11} + \cdots + \tilde{a}_{1M}) I_{m_2} + \cdots + b_M (\tilde{a}_{M1} + \cdots + \tilde{a}_{MM}) I_{m_2}] \\ &= [b_1 (\tilde{a}_{11} + \cdots + \tilde{a}_{1M}) + \cdots + b_M (\tilde{a}_{M1} + \cdots + \tilde{a}_{MM})] I_{m_2} \end{aligned}$$

$$\begin{aligned}
&= [b_1, \dots, b_M] \begin{bmatrix} (\tilde{a}_{11} + \dots + \tilde{a}_{1M}) I_{m_2} \\ \vdots \\ (\tilde{a}_{M1} + \dots + \tilde{a}_{MM}) I_{m_2} \end{bmatrix} \\
&= [b_1, \dots, b_M] \begin{bmatrix} \tilde{a}_{11} & \dots & \tilde{a}_{1M} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{M1} & \dots & \tilde{a}_{MM} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} I_{m_2} \\
&= b' \mathcal{A}^{-1} \in_M I_{m_2} = [1 - (1 - b' \mathcal{A}^{-1} \in_M)] I_{m_2} = (1 - r) I_{m_2} \blacksquare
\end{aligned}$$

Lema 4.3.4 Considerando $b_i' = b' \otimes I_{m_i}$, $Z_i = \in_M \otimes I_{m_i}$, $i = 1, 2$, $\hat{T}_1 = I + h\mathcal{A} \otimes C$, $\hat{T}_4 = \mathcal{A} \otimes I$ e $\hat{S}_1 = I_{m_1} \otimes C$, então

$$S_n = \begin{bmatrix} I_{m_1} & 0 \\ 0 & I_{m_2} \end{bmatrix} - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 Z_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} Z_2 + O(1) \end{bmatrix}$$

pode ser escrito na forma $S_n = K + O(h)$, onde $K = \begin{bmatrix} I_{m_1} & 0 \\ 0 & rI_{m_2} \end{bmatrix}$.

Demonstração:

$$\begin{aligned}
S_n &= \begin{bmatrix} I_{m_1} & 0 \\ 0 & I_{m_2} \end{bmatrix} - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 Z_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} Z_2 + O(1) \end{bmatrix} \\
&= \begin{bmatrix} I_{m_1} & 0 \\ 0 & I_{m_2} \end{bmatrix} - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 Z_1 + O(h) & O(h) \\ O(1) & \frac{(1-r)}{h} I_{m_2} + O(1) \end{bmatrix}
\end{aligned}$$

pois, vimos no Lema 4.3.3 que $b_2' \hat{T}_4^{-1} Z_2 = (1-r) I_{m_2}$. Sendo $b_1' \hat{T}_1^{-1} \hat{S}_1 Z_1 = O(1)$ temos

$$\begin{aligned}
S_n &= \begin{bmatrix} I_{m_1} & 0 \\ 0 & I_{m_2} \end{bmatrix} - h \begin{bmatrix} O(1) + O(h) & O(h) \\ O(1) & \frac{(1-r)}{h} I_{m_2} + O(1) \end{bmatrix} \\
&= \begin{bmatrix} I_{m_1} + O(h) & O(h^2) \\ O(h) & I_{m_2} - (1-r) I_{m_2} + O(h) \end{bmatrix} \\
&= \begin{bmatrix} I_{m_1} & 0 \\ 0 & rI_{m_2} \end{bmatrix} + \begin{bmatrix} O(h) & O(h^2) \\ O(h) & O(h) \end{bmatrix} = K + O(h) \blacksquare
\end{aligned}$$

Lema 4.3.5 *Considere*

$$K = \begin{bmatrix} I_{m1} & 0 \\ 0 & rI_{m2} \end{bmatrix}, \quad -1 \leq r < 1.$$

Então

$$(K + O(h))^n = \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix},$$

para $n \leq 1/h$.

Demonstração (Por indução): Considerando

$$K = \begin{bmatrix} I_{m1} & 0 \\ 0 & rI_{m2} \end{bmatrix} \quad \text{e} \quad O(h) = \begin{bmatrix} O(h) & O(h) \\ O(h) & O(h) \end{bmatrix},$$

então,

$$K + O(h) = \begin{bmatrix} I_{m1} & 0 \\ 0 & rI_{m2} \end{bmatrix} + \begin{bmatrix} O(h) & O(h) \\ O(h) & O(h) \end{bmatrix} = \begin{bmatrix} I + O(h) & O(h) \\ O(h) & rI + O(h) \end{bmatrix}$$

e portanto, válido para $n = 1$.

Agora, suponhamos que

$$(K + O(h))^{n-1} = \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^{n-1}I + O(h) \end{bmatrix}$$

é verdadeiro.

Mostraremos que é válido

$$(K + O(h))^n = \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix},$$

isto é

$$\begin{aligned} (K + O(h))^n &= (K + O(h))^{n-1} (K + O(h)) \\ &= \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^{n-1}I + O(h) \end{bmatrix} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & rI + O(h) \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
&= \begin{bmatrix} I + O(h) & O(h) + r O(h) \\ O(h) + r^{n-1} O(h) & r^{n-1} r I + r^{n-1} r O(h) + r O(h) + O(h^2) \end{bmatrix} \\
&= \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix} \blacksquare
\end{aligned}$$

Teorema 4.3.1 Suponha que $F(y, y', t) = 0$ tenha índice uniforme um, que o método de RKI satisfaça a condição de estabilidade $|r| \leq 1$, que os erros nas condições iniciais sejam $O(h^G)$ e que os erros na iteração de Newton sejam $O(h^{G+\delta})$ onde $\delta = 1$ se $|r| = 1$, caso contrário, $\delta = 0$ e $G > 1$. Então o erro global satisfaz $\|e_n\| = O(h^G)$, onde

$$G = \begin{cases} q, & \text{se } C(q) \text{ e } B(q) \text{ são válidos} \\ q+1, & \text{se } C(q), B(q+1) \text{ são válidos e } -1 \leq r < 1 \\ q+1, & \text{se } C(q), B(q+1), A_1(q+1) \text{ são válidos e } r = 1 \end{cases}$$

Demonstração: Considere o método de RKI .

$$\begin{aligned}
F\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y'_j, Y'_i\right) &= 0, \quad i = 1, 2, \dots, M, \\
y_n &= y_{n-1} + h \sum_{i=1}^M b_i Y'_i,
\end{aligned} \tag{4.3.8}$$

cuja solução numérica satisfaz

$$F\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y'_j, Y'_i\right) = \bar{\eta}^{(i)}, \quad 1 \leq i \leq M \tag{4.3.9a}$$

$$y_n = y_{n-1} + h \sum_{i=1}^M b_i Y'_i \tag{4.3.9b}$$

onde $\bar{\eta}^{(i)}$, $i = 1, 2, \dots, M$, representam os erros na iteração de Newton.

A solução verdadeira satisfaz

$$F\left(t_{n-1} + c_i h, y(t_{n-1}) + h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) - \delta_i, y'(t_{n-1} + c_i h)\right) = 0 \tag{4.3.10a}$$

$$y(t_n) = y(t_{n-1}) + h \sum_{i=1}^M b_i y'(t_{n-1} + c_i h) - \delta_{M+1}, \tag{4.3.10b}$$

onde δ_i , $i = 1, 2, \dots, M+1$ são os erros de truncamento local em t_n .

Subtraindo a equação (4.3.9a) de (4.3.10a) obtemos

$$F\left(t_{n-1} + c_i h, y(t_{n-1}) + h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) - \delta_i, y'(t_{n-1} + c_i h)\right) - F\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y'_j, Y'_i\right) = \overline{\eta}^{(i)}$$

e usando o desenvolvimento de Taylor em torno de $(t_{n-1} + c_i h, y(t_{n-1} + c_i h), y'(t_{n-1} + c_i h))$ temos

$$\begin{aligned} & F_y \left(y(t_{n-1} + c_i h) - y(t_{n-1}) - h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) + \delta_i \right) + \\ & + \frac{1}{2} \left(y(t_{n-1} + c_i h) - y(t_{n-1}) - h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) + \delta_i \right)^2 F_{yy} \left(y(t_{n-1} + c_i h) - y(t_{n-1}) - h \sum_{j=1}^M a_{ij} y'(t_{n-1} + c_j h) + \delta_i \right) \\ & + \dots - F_y \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y'_j \right) - F_{y'} (y'(t_{n-1} + c_i h) - Y'_i) \\ & - \frac{1}{2} \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y'_j \right)^2 F_{yy} \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y'_j \right) \\ & - \frac{1}{2} \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y'_j \right)^2 F_{yy'} (y'(t_{n-1} + c_i h) - Y'_i) \\ & - (y'(t_{n-1} + c_i h) - Y'_i)^2 F_{y'y'} (y'(t_{n-1} + c_i h) - Y'_i) - \dots + \overline{\eta}^{(i)}. \end{aligned} \quad (4.3.11)$$

$$\text{Logo, } F_y \left(y_{n-1} - y(t_{n-1}) + h \sum_{j=1}^M a_{ij} (Y'_j - y'(t_{n-1} + c_j h)) + \delta_i \right) + F_{y'} (Y'_i - y'(t_{n-1} + c_i h)) = \eta_i.$$

Subtraindo, também, (4.3.9b) de (4.3.10b) temos

$$y_n - y(t_n) = y_{n-1} - y(t_{n-1}) + h \sum_{i=1}^M b_i (Y'_i - y'(t_{n-1} + c_i h)) + \delta_{M+1}.$$

Considerando $E'_i = Y'_i - y'(t_{n-1} + c_i h)$, $E_i = Y_i - y(t_{n-1} + c_i h)$ e $e_n = y_n - y(t_n)$, observando que η_i é a soma dos resíduos na iteração de Newton e os termos de mais alta ordem em e_{n-1} e E'_i , obtemos o sistema de equações

$$A_i E'_i + B_i \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E'_j + \delta_i \right) = \eta_i, \quad i = 1, 2, \dots, M \quad (4.3.12a)$$

$$e_n = e_{n-1} + h \sum_{i=1}^M b_i E'_i + \delta_{M+1}, \quad (4.3.12b)$$

onde $A_i = F_{y'}$ e $B_i = F_y$ avaliados no ponto $(t_{n-1} + c_i h, y(t_{n-1} + c_i h), y'(t_{n-1} + c_i h))$.

Podemos escrever (4.3.12a) na forma

$$\begin{cases} A_1 E'_1 + B_1 \{e_{n-1} + h[a_{11} E'_1 + a_{12} E'_2 + \dots + a_{1M} E'_M] + \delta_1\} = \eta_1 \\ A_2 E'_2 + B_2 \{e_{n-1} + h[a_{21} E'_1 + a_{22} E'_2 + \dots + a_{2M} E'_M] + \delta_2\} = \eta_2 \\ \vdots \\ A_M E'_M + B_M \{e_{n-1} + h[a_{M1} E'_1 + a_{M2} E'_2 + \dots + a_{MM} E'_M] + \delta_M\} = \eta_M \end{cases}$$

ou

$$\begin{cases} (A_1 + h a_{11} B_1) E'_1 + a_{12} h B_1 E'_2 + \dots + a_{1M} h B_1 E'_M = -B_1 (e_{n-1} + \delta_1) + \eta_1 \\ a_{21} h B_2 E'_1 + (A_2 + a_{22} h B_2) E'_2 + \dots + a_{2M} h B_2 E'_M = -B_2 (e_{n-1} + \delta_2) + \eta_2 \\ \vdots \\ a_{M1} h B_M E'_1 + a_{M2} h B_M E'_2 + \dots + (A_M + a_{MM} h B_M) E'_M = -B_M (e_{n-1} + \delta_M) + \eta_M \end{cases}$$

ou ainda, na forma matricial

$$\begin{bmatrix} A_1 + a_{11} h B_1 & a_{12} h B_1 & \dots & a_{1M} h B_1 \\ a_{21} h B_2 & A_2 + a_{22} h B_2 & \dots & a_{2M} h B_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} h B_M & a_{M2} h B_M & \dots & A_M + a_{MM} h B_M \end{bmatrix} \begin{bmatrix} E'_1 \\ E'_2 \\ \vdots \\ E'_M \end{bmatrix} = - \begin{bmatrix} B_1 (e_{n-1} + \delta_1) \\ B_2 (e_{n-1} + \delta_2) \\ \vdots \\ B_M (e_{n-1} + \delta_M) \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_M \end{bmatrix} \quad (4.3.13)$$

Considere $A = F_{y'}$ e $B = F_y$ sendo avaliados em $(t_n, y(t_n), y'(t_n))$, P e Q sendo as matrizes que transformam A e B na forma canônica de Kronecker. Sejam também P_i e Q_i as matrizes que transformam A_i e B_i na forma canônica de Kronecker, isto é,

$$PAQ = \begin{bmatrix} I_{m1} & 0 \\ 0 & 0 \end{bmatrix}; \quad PBQ = \begin{bmatrix} C & 0 \\ 0 & I_{m2} \end{bmatrix}; \quad P_i A_i Q_i = \begin{bmatrix} I_{m1} & 0 \\ 0 & 0 \end{bmatrix}; \quad P_i B_i Q_i = \begin{bmatrix} C_i & 0 \\ 0 & I_{m2} \end{bmatrix},$$

como no Lema 3.3.1.

Fazendo $\tilde{e}_{n-1} = Q^{-1} e_{n-1}$, $\tilde{E}_i = Q^{-1} E'_i$, $\tilde{\delta}_i = Q^{-1} \delta_i$ e $\tilde{\eta}_i = P_i \eta_i$, então (4.3.13) torna-se

$$\begin{bmatrix} A_1 + a_{11} h B_1 & a_{12} h B_1 & \dots & a_{1M} h B_1 \\ a_{21} h B_2 & A_2 + a_{22} h B_2 & \dots & a_{2M} h B_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} h B_M & a_{M2} h B_M & \dots & A_M + a_{MM} h B_M \end{bmatrix} \begin{bmatrix} Q \tilde{E}_1 \\ Q \tilde{E}_2 \\ \vdots \\ Q \tilde{E}_M \end{bmatrix} = - \begin{bmatrix} B_1 (Q \tilde{e}_{n-1} + Q \tilde{\delta}_1) \\ B_2 (Q \tilde{e}_{n-1} + Q \tilde{\delta}_2) \\ \vdots \\ B_M (Q \tilde{e}_{n-1} + Q \tilde{\delta}_M) \end{bmatrix} + \begin{bmatrix} P_1^{-1} \tilde{\eta}_1 \\ P_2^{-1} \tilde{\eta}_2 \\ \vdots \\ P_M^{-1} \tilde{\eta}_M \end{bmatrix}$$

Multiplicando à esquerda pela matriz diagonal $\text{Diag}(P_1, \dots, P_M)$ obtemos

$$\begin{bmatrix} P_1(A_1 + a_{11}hB_1)Q & a_{12}hP_1B_1Q & \cdots & a_{1M}hP_1B_1Q \\ a_{21}hP_2B_2Q & P_2(A_2 + a_{22}hB_2)Q & \cdots & a_{2M}hP_2B_2Q \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}hP_MB_MQ & a_{M2}hP_MB_MQ & \cdots & P_M(A_M + a_{MM}hB_M)Q \end{bmatrix} \begin{bmatrix} \tilde{E}'_1 \\ \tilde{E}'_2 \\ \vdots \\ \tilde{E}'_M \end{bmatrix} = - \begin{bmatrix} P_1B_1Q(\tilde{e}_{n-1} + \tilde{\delta}_1) \\ P_2B_2Q(\tilde{e}_{n-1} + \tilde{\delta}_2) \\ \vdots \\ P_MB_MQ(\tilde{e}_{n-1} + \tilde{\delta}_M) \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \\ \vdots \\ \tilde{\eta}_M \end{bmatrix}$$

que, ainda, pode ser escrita na forma

$$\begin{bmatrix} P_1A_1Q_1(Q_1^{-1}Q) + a_{11}hP_1B_1Q_1(Q_1^{-1}Q) & \cdots & a_{1M}hP_1B_1Q_1(Q_1^{-1}Q) \\ a_{21}hP_2B_2Q_2(Q_2^{-1}Q) & \cdots & a_{2M}hP_2B_2Q_2(Q_2^{-1}Q) \\ \vdots & \ddots & \vdots \\ a_{M1}hP_MB_MQ_M(Q_M^{-1}Q) & \cdots & P_MA_MQ_M(Q_M^{-1}Q) + a_{MM}hP_MB_MQ_M(Q_M^{-1}Q) \end{bmatrix} \begin{bmatrix} \tilde{E}'_1 \\ \tilde{E}'_2 \\ \vdots \\ \tilde{E}'_M \end{bmatrix} = - \begin{bmatrix} P_1B_1Q_1(Q_1^{-1}Q)(\tilde{e}_{n-1} + \tilde{\delta}_1) \\ P_2B_2Q_2(Q_2^{-1}Q)(\tilde{e}_{n-1} + \tilde{\delta}_2) \\ \vdots \\ P_MB_MQ_M(Q_M^{-1}Q)(\tilde{e}_{n-1} + \tilde{\delta}_M) \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \\ \vdots \\ \tilde{\eta}_M \end{bmatrix},$$

ou

$$\begin{bmatrix} X_1 + a_{11}hW_1 & a_{12}hW_1 & \cdots & a_{1M}hW_1 \\ a_{21}hW_2 & X_2 + a_{22}hW_2 & \cdots & a_{2M}hW_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}hW_M & a_{M2}hW_M & \cdots & X_M + a_{MM}hW_M \end{bmatrix} \begin{bmatrix} \tilde{E}'_1 \\ \tilde{E}'_2 \\ \vdots \\ \tilde{E}'_M \end{bmatrix} = - \begin{bmatrix} W_1(\tilde{e}_{n-1} + \tilde{\delta}_1) \\ W_2(\tilde{e}_{n-1} + \tilde{\delta}_2) \\ \vdots \\ W_M(\tilde{e}_{n-1} + \tilde{\delta}_M) \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \\ \vdots \\ \tilde{\eta}_M \end{bmatrix}, \quad (4.3.14)$$

sendo $W_i = P_iB_iQ_i(Q_i^{-1}Q)$ e $X_i = P_iA_iQ_i(Q_i^{-1}Q)$.

Considerando a suavidade de Q_i^{-1} , Q , C_i e C pelo Lema 3.3.1 temos

$$Q_i^{-1}Q = I + O(t_{n-1} - t_n - hc_i) = I + O((1 - c_i)h) = I + O(h),$$

$$C_i = C + O(h).$$

Agora, considerando

$$Q_i^{-1}Q = \begin{bmatrix} I_{m1} + U_i & Z_i \\ R_i & I_{m2} + V_i \end{bmatrix},$$

onde U_i , Z_i , R_i e V_i são matrizes cujos elementos são $O(h)$ obtemos

$$W_i = \begin{bmatrix} C_i & 0 \\ 0 & I_{m_2} \end{bmatrix} (I + O(h)) = \begin{bmatrix} C + D_i & 0 \\ 0 & I_{m_2} \end{bmatrix} \begin{bmatrix} I_{m_1} + U_i & Z_i \\ R_i & I_{m_2} + V_i \end{bmatrix}$$

$$= \begin{bmatrix} (C + D_i)(I_{m_1} + U_i) & (C + D_i)Z_i \\ R_i & I_{m_2} + V_i \end{bmatrix};$$

$$X_i = \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix} (I + O(h)) = \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{m_1} + U_i & Z_i \\ R_i & I_{m_2} + V_i \end{bmatrix} = \begin{bmatrix} I_{m_1} + U_i & Z_i \\ 0 & 0 \end{bmatrix}.$$

Considere $\tilde{E}_i' = [\tilde{E}_i^{(1)}, \tilde{E}_i^{(2)}]^t$, sendo que $\tilde{E}_i^{(1)}$ tem dimensão m_1 e $\tilde{E}_i^{(2)}$ tem dimensão m_2 . Fazendo $\tilde{\delta}_i = [\tilde{\delta}_i^{(1)}, \tilde{\delta}_i^{(2)}]^t$, $\tilde{\eta}_i = [\tilde{\eta}_i^{(1)}, \tilde{\eta}_i^{(2)}]^t$, $i = 1, \dots, M$ e $\tilde{e}_{n-1} = [\tilde{e}_{n-1}^{(1)}, \tilde{e}_{n-1}^{(2)}]^t$, temos que (4.3.14) torna-se

$$\begin{bmatrix} \begin{bmatrix} I_{m_1} + U_1 & Z_1 \\ 0 & 0 \end{bmatrix} + a_1 h \begin{bmatrix} (C + D_1)(I_{m_1} + U_1) & (C + D_1)Z_1 \\ R_1 & I_{m_2} + V_1 \end{bmatrix} & \dots & a_M h \begin{bmatrix} (C + D_1)(I_{m_1} + U_1) & (C + D_1)Z_1 \\ R_1 & I_{m_2} + V_1 \end{bmatrix} \\ \vdots & \ddots & \vdots \\ a_M h \begin{bmatrix} (C + D_M)(I_{m_1} + U_M) & (C + D_M)Z_M \\ R_M & I_{m_2} + V_M \end{bmatrix} & \dots & \begin{bmatrix} I_{m_1} + U_M & Z_M \\ 0 & 0 \end{bmatrix} + a_{MM} h \begin{bmatrix} (C + D_M)(I_{m_1} + U_M) & (C + D_M)Z_M \\ R_M & I_{m_2} + V_M \end{bmatrix} \end{bmatrix}^x$$

$$\begin{bmatrix} \tilde{E}_1^{(1)} \\ \tilde{E}_1^{(2)} \\ \vdots \\ \tilde{E}_M^{(1)} \\ \tilde{E}_M^{(2)} \end{bmatrix} = - \begin{bmatrix} \begin{bmatrix} (C + D_1)(I_{m_1} + U_1) & (C + D_1)Z_1 \\ R_1 & I_{m_2} + V_1 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}_1^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}_1^{(2)} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} (C + D_M)(I_{m_1} + U_M) & (C + D_M)Z_M \\ R_M & I_{m_2} + V_M \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}_M^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}_M^{(2)} \end{bmatrix} \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_1^{(1)} \\ \tilde{\eta}_1^{(2)} \\ \vdots \\ \tilde{\eta}_M^{(1)} \\ \tilde{\eta}_M^{(2)} \end{bmatrix}.$$

Obtendo, assim o seguinte sistema

$$\begin{aligned} & [I_{m_1} + U_1 + a_{11}h(C + D_1)(I_{m_1} + U_1)]\tilde{E}_1^{(1)} + (Z_1 + a_{11}h(C + D_1)Z_1)\tilde{E}_1^{(2)} \\ & + \dots + a_{1M}h(C + D_1)(I_{m_1} + U_1)\tilde{E}_M^{(1)} + a_{1M}h(C + D_1)Z_1\tilde{E}_M^{(2)} = \\ & = -(C + D_1)(I_{m_1} + U_1)(\tilde{e}_{n-1}^{(1)} + \tilde{\delta}_1^{(1)}) - (C + D_1)Z_1(\tilde{e}_{n-1}^{(2)} + \tilde{\delta}_1^{(2)}) + \tilde{\eta}_1^{(1)} \end{aligned}$$

$$a_{11}hR_1\tilde{E}_1^{(1)} + a_{11}h(I_{m2} + V_1)\tilde{E}_1^{(2)} + \dots + a_{1M}hR_1\tilde{E}_M^{(1)} + a_{1M}h(I_{m2} + V_1)\tilde{E}_M^{(2)} =$$

$$= -R_1(\tilde{e}_{n-1}^{(1)} + \tilde{\delta}_1^{(1)}) - (I_{m2} + V_1)(\tilde{e}_{n-1}^{(2)} + \tilde{\delta}_1^{(2)}) + \tilde{\eta}_1^{(2)}$$

⋮

$$a_{M1}h(C + D_M)(I_{m1} + U_M)\tilde{E}_1^{(1)} + a_{M1}h(C + D_M)Z_M\tilde{E}_1^{(2)}$$

$$+ \dots + [I_{m1} + U_M + a_{MM}h(C + D_M)(I_{m1} + U_M)]\tilde{E}_M^{(1)} + (Z_M + a_{MM}h(C + D_M)Z_M)\tilde{E}_M^{(2)} =$$

$$= -(C + D_M)(I_{m1} + U_M)(\tilde{e}_{n-1}^{(1)} + \tilde{\delta}_M^{(1)}) - (C + D_M)Z_M(\tilde{e}_{n-1}^{(2)} + \tilde{\delta}_M^{(2)}) + \tilde{\eta}_M^{(1)}$$

$$a_{M1}hR_M\tilde{E}_1^{(1)} + a_{M1}h(I_{m2} + V_M)\tilde{E}_1^{(2)} + \dots + a_{MM}hR_M\tilde{E}_M^{(1)} + a_{MM}h(I_{m2} + V_M)\tilde{E}_M^{(2)} =$$

$$= -R_M(\tilde{e}_{n-1}^{(1)} + \tilde{\delta}_M^{(1)}) - (I_{m2} + V_M)(\tilde{e}_{n-1}^{(2)} + \tilde{\delta}_M^{(2)}) + \tilde{\eta}_M^{(2)}$$

que pode ser escrito na forma matricial, após algumas trocas de linhas,

$$\begin{bmatrix} \alpha_1 + a_{11}h(C + D_1)\alpha_1 & \dots & a_{1M}h(C + D_1)\alpha_1 & Z_1 + a_{11}h(C + D_1)Z_1 & \dots & a_{1M}h(C + D_1)Z_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{M1}h(C + D_M)\alpha_M & \dots & \alpha_M + a_{MM}h(C + D_M)\alpha_M & a_{M1}h(C + D_M)Z_M & \dots & Z_M + a_{MM}h(C + D_M)Z_M \\ \hline a_{11}hR_1 & \dots & a_{1M}hR_1 & a_{11}h\beta_1 & \dots & a_{1M}h\beta_M \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{M1}hR_M & \dots & a_{MM}hR_M & a_{M1}h\beta_M & \dots & a_{MM}h\beta_M \end{bmatrix} \mathbf{x}$$

$$\begin{bmatrix} \tilde{E}_1^{(1)} \\ \vdots \\ \tilde{E}_M^{(1)} \\ \tilde{E}_1^{(2)} \\ \vdots \\ \tilde{E}_M^{(2)} \end{bmatrix} = - \begin{bmatrix} (C + D_1)\alpha_1 & & & (C + D_1)Z_1 & & \\ & \ddots & & & \ddots & \\ & & (C + D_M)\alpha_M & & & (C + D_M)Z_M \\ \hline R_1 & & & \beta_1 & & \\ & \ddots & & & \ddots & \\ & & R_M & & & \beta_M \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}_1^{(1)} \\ \vdots \\ \tilde{e}_{n-1}^{(1)} + \tilde{\delta}_M^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}_1^{(2)} \\ \vdots \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}_M^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\eta}_1^{(1)} \\ \vdots \\ \tilde{\eta}_M^{(1)} \\ \tilde{\eta}_1^{(2)} \\ \vdots \\ \tilde{\eta}_M^{(2)} \end{bmatrix}$$

(4.3.15)

onde $\alpha_i = I_{m1} + U_i$ e $\beta_i = I_{m2} + V_i$.

Analisando (4.3.15), observamos que tanto a matriz do lado esquerdo quanto a do lado direito são divididas em quatro blocos, assim,

- do primeiro bloco da matriz do lado esquerdo temos

$$\begin{aligned}
 & \begin{bmatrix} \alpha_1 + a_{11}h(C + D_1)\alpha_1 & \cdots & a_{1M}h(C + D_1)\alpha_1 \\ \vdots & \ddots & \vdots \\ a_{M1}h(C + D_M)\alpha_M & \cdots & \alpha_M + a_{MM}h(C + D_M)\alpha_M \end{bmatrix} = \\
 & = \begin{bmatrix} I_{m1} + U_1 + a_{11}h(C + D_1)(I_{m1} + U_1) & \cdots & a_{1M}h(C + D_1)(I_{m1} + U_1) \\ \vdots & \ddots & \vdots \\ a_{M1}h(C + D_M)(I_{m1} + U_M) & \cdots & I_{m1} + U_M + a_{MM}h(C + D_M)(I_{m1} + U_M) \end{bmatrix} \\
 & = \begin{bmatrix} I_{m1} & & \\ & \ddots & \\ & & I_{m1} \end{bmatrix} + h \begin{bmatrix} a_{11}C & \cdots & a_{1M}C \\ \vdots & \ddots & \vdots \\ a_{M1}C & \cdots & a_{MM}C \end{bmatrix} \\
 & + \begin{bmatrix} U_1 + a_{11}h(CU_1 + D_1 + D_1C_1) & \cdots & a_{1M}h(CU_1 + D_1 + D_1C_1) \\ \vdots & \ddots & \vdots \\ a_{M1}h(CU_M + D_M + D_MC_M) & \cdots & U_M + a_{MM}h(CU_M + D_M + D_MC_M) \end{bmatrix} \\
 & = I + hA \otimes C + O(h) = \hat{T}_1 + O(h);
 \end{aligned}$$

- do segundo bloco

$$\begin{aligned}
 & \begin{bmatrix} Z_1 + a_{11}h(C + D_1)Z_1 & \cdots & a_{1M}h(C + D_1)Z_1 \\ \vdots & \ddots & \vdots \\ a_{M1}h(C + D_M)Z_M & \cdots & Z_M + a_{MM}h(C + D_M)Z_M \end{bmatrix} = \\
 & = \begin{bmatrix} Z_1 + a_{11}hCZ_1 + a_{11}hD_1Z_1 & \cdots & a_{1M}hCZ_1 + a_{1M}hD_1Z_1 \\ \vdots & \ddots & \vdots \\ a_{M1}hCZ_M + a_{M1}hD_MZ_M & \cdots & Z_M + a_{MM}hCZ_M + a_{MM}hD_MZ_M \end{bmatrix} \\
 & = h \begin{bmatrix} a_{11}CZ_1 & \cdots & a_{1M}CZ_1 \\ \vdots & \ddots & \vdots \\ a_{M1}CZ_M & \cdots & a_{MM}CZ_M \end{bmatrix} + \begin{bmatrix} Z_1 + a_{11}hD_1Z_1 & \cdots & a_{1M}hD_1Z_1 \\ \vdots & \ddots & \vdots \\ a_{M1}hD_MZ_M & \cdots & Z_M + a_{MM}hD_MZ_M \end{bmatrix} \\
 & = h^2O(1) + O(h) = h^2T_2;
 \end{aligned}$$

- do terceiro bloco

$$\begin{bmatrix} a_{11}hR_1 & \cdots & a_{1M}hR_1 \\ \vdots & \ddots & \vdots \\ a_{M1}hR_M & \cdots & a_{MM}hR_M \end{bmatrix} = h \begin{bmatrix} a_{11}R_1 & \cdots & a_{1M}R_1 \\ \vdots & \ddots & \vdots \\ a_{M1}R_M & \cdots & a_{MM}R_M \end{bmatrix} = h^2O(1) = h^2T_3;$$

- do quarto bloco

$$\begin{aligned} \begin{bmatrix} a_{11}h\beta_1 & \cdots & a_{1M}h\beta_1 \\ \vdots & \ddots & \vdots \\ a_{M1}h\beta_M & \cdots & a_{MM}h\beta_M \end{bmatrix} &= \begin{bmatrix} a_{11}h(I_{m_2} + V_1) & \cdots & a_{1M}h(I_{m_2} + V_1) \\ \vdots & \ddots & \vdots \\ a_{M1}h(I_{m_2} + V_M) & \cdots & a_{MM}h(I_{m_2} + V_M) \end{bmatrix} \\ &= h \left\{ \begin{bmatrix} a_{11}I_{m_2} & \cdots & a_{1M}I_{m_2} \\ \vdots & \ddots & \vdots \\ a_{M1}I_{m_2} & \cdots & a_{MM}I_{m_2} \end{bmatrix} + \begin{bmatrix} a_{11}V_1 & \cdots & a_{1M}V_1 \\ \vdots & \ddots & \vdots \\ a_{M1}V_M & \cdots & a_{MM}V_M \end{bmatrix} \right\} = h(A \otimes I + O(h)) \\ &= h\hat{T}_4 + O(h^2); \end{aligned}$$

- do primeiro bloco da matriz do lado direito, temos

$$\begin{aligned} \begin{bmatrix} (C + D_1)\alpha_1 & & \\ & \ddots & \\ & & (C + D_M)\alpha_M \end{bmatrix} &= \begin{bmatrix} (C + D_1)(I_{m_1} + U_1) & & \\ & \ddots & \\ & & (C + D_M)(I_{m_1} + U_M) \end{bmatrix} \\ &= \begin{bmatrix} 1C & & \\ & \ddots & \\ & & 1C \end{bmatrix} + \begin{bmatrix} CU_1 + D_1 + D_1U_1 & & \\ & \ddots & \\ & & CU_M + D_M + D_MU_M \end{bmatrix} \\ &= I \otimes C + O(h) = \hat{S}_1 + O(h); \end{aligned}$$

- do segundo e do terceiro bloco

$$\begin{bmatrix} (C + D_1)Z_1 & & \\ & \ddots & \\ & & (C + D_M)Z_M \end{bmatrix} = hS_2 \quad \text{e} \quad \begin{bmatrix} R_1 & & \\ & \ddots & \\ & & R_M \end{bmatrix} = hS_3;$$

- finalmente, para concluirmos a observação, temos do quarto bloco

$$\begin{bmatrix} I_{m_2} + V_1 & & \\ & \ddots & \\ & & I_{m_2} + V_M \end{bmatrix} = \begin{bmatrix} I_{m_2} & & \\ & \ddots & \\ & & I_{m_2} \end{bmatrix} + \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_M \end{bmatrix} = I + O(h) = S_4.$$

Portanto, a equação (4.3.15) torna-se

$$\begin{bmatrix} T_1 & h^2T_2 \\ h^2T_3 & hT_4 \end{bmatrix} \begin{bmatrix} \tilde{E}^{(1)} \\ \tilde{E}^{(2)} \end{bmatrix} = - \begin{bmatrix} S_1 & hS_2 \\ hS_3 & S_4 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} \quad (4.3.16)$$

onde

$$\begin{aligned}\tilde{E}^{(i)} &= [\tilde{E}_1^{(i)}, \tilde{E}_2^{(i)}, \dots, \tilde{E}_M^{(i)}]^t; \\ \tilde{\delta}^{(i)} &= [\tilde{\delta}_1^{(i)}, \tilde{\delta}_2^{(i)}, \dots, \tilde{\delta}_M^{(i)}]^t; \\ \tilde{\eta}^{(i)} &= [\tilde{\eta}_1^{(i)}, \tilde{\eta}_2^{(i)}, \dots, \tilde{\eta}_M^{(i)}]^t; \\ \tilde{e}_{n-1}^{(i)} &= [\tilde{e}_{n-1}^{(i)}, \tilde{e}_{n-1}^{(i)}, \dots, \tilde{e}_{n-1}^{(i)}]^t,\end{aligned}$$

para $i = 1, 2$ e

$$\begin{aligned}T_1 &= \hat{T}_1 + O(h); \\ T_4 &= \hat{T}_4 + O(h); \\ S_1 &= \hat{S}_1 + O(h); \\ S_4 &= I + O(h),\end{aligned}$$

sendo $\hat{T}_1 = I + hA \otimes C$, $\hat{T}_4 = A \otimes I$ e $\hat{S}_1 = I \otimes C$.

Considere

$$\bar{T}_n = \begin{bmatrix} T_1 & h^2 T_2 \\ h^2 T_3 & h T_4 \end{bmatrix},$$

temos, pelo Lema 4.3.2, que a inversa de $\bar{T}_n$ é

$$\bar{T}_n^{-1} = \begin{bmatrix} \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix}. \quad (4.3.17)$$

Usando (4.3.17) para resolver (4.3.16), em termos de $\tilde{E}^{(1)}$ e $\tilde{E}^{(2)}$ temos

$$\begin{aligned}\begin{bmatrix} \tilde{E}^{(1)} \\ \tilde{E}^{(2)} \end{bmatrix} &= - \begin{bmatrix} T_1 & h^2 T_2 \\ h^2 T_3 & h T_4 \end{bmatrix}^{-1} \left\{ \begin{bmatrix} S_1 & h S_2 \\ h S_3 & S_4 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} - \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} \right\} \\ &= - \begin{bmatrix} \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} S_1 & h S_2 \\ h S_3 & S_4 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} + \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} \\ &= - \begin{bmatrix} \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} + \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix}. \quad (4.3.18)\end{aligned}$$

Multiplicando (4.3.12b) por $Q^{-1} = Q_n^{-1}$ temos

$$Q_n^{-1}e_n = \tilde{e}_{n-1} + h \sum_{i=1}^M b_i \tilde{E}_i' + \tilde{\delta}_{M+1} \quad (4.3.19)$$

que pode ser escrito

$$\begin{aligned} Q_n^{-1}e_n &= \tilde{e}_{n-1} + hb_1 \tilde{E}_1' + hb_2 \tilde{E}_2' + \dots + hb_M \tilde{E}_M' + \tilde{\delta}_{M+1} \\ &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} + hb_1 \begin{bmatrix} \tilde{E}_1'^{(1)} \\ \tilde{E}_1'^{(2)} \end{bmatrix} + hb_2 \begin{bmatrix} \tilde{E}_2'^{(1)} \\ \tilde{E}_2'^{(2)} \end{bmatrix} + \dots + hb_M \begin{bmatrix} \tilde{E}_M'^{(1)} \\ \tilde{E}_M'^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} + h \begin{bmatrix} b_1 \tilde{E}_1'^{(1)} + b_2 \tilde{E}_2'^{(1)} + \dots + b_M \tilde{E}_M'^{(1)} \\ b_1 \tilde{E}_1'^{(2)} + b_2 \tilde{E}_2'^{(2)} + \dots + b_M \tilde{E}_M'^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} + h \begin{bmatrix} b' \otimes I_{m_1} & 0 \\ 0 & b' \otimes I_{m_2} \end{bmatrix} \begin{bmatrix} \tilde{E}'^{(1)} \\ \tilde{E}'^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} + h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \begin{bmatrix} \tilde{E}'^{(1)} \\ \tilde{E}'^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix}, \end{aligned} \quad (4.3.20)$$

onde $b' = [b_1, b_2, \dots, b_M]$, $b_1' = b' \otimes I_{m_1}$ e $b_2' = b' \otimes I_{m_2}$.

Inserindo (4.3.18) em (4.3.20) temos

$$\begin{aligned} Q_n^{-1}e_n &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} - h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \left\{ \begin{bmatrix} \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} - \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} \right\} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} - h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \begin{bmatrix} \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} + \tilde{\delta}^{(1)} \\ \tilde{e}_{n-1}^{(2)} + \tilde{\delta}^{(2)} \end{bmatrix} + h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \end{aligned}$$

como $\tilde{e}_{n-1}^{(1)}$ tem dimensão m_1 e $\tilde{e}_{n-1}^{(2)}$ tem dimensão m_2 temos

$$\begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} I_{m_1} \\ \vdots \\ I_{m_1} \end{pmatrix} \tilde{e}_{n-1}^{(1)} \\ \begin{pmatrix} I_{m_2} \\ \vdots \\ I_{m_2} \end{pmatrix} \tilde{e}_{n-1}^{(2)} \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 1I_{m_1} \\ \vdots \\ 1I_{m_1} \end{pmatrix} \tilde{e}_{n-1}^{(1)} \\ \begin{pmatrix} 1I_{m_2} \\ \vdots \\ 1I_{m_2} \end{pmatrix} \tilde{e}_{n-1}^{(2)} \end{bmatrix} = \begin{bmatrix} (\epsilon_M \otimes I_{m_1}) \tilde{e}_{n-1}^{(1)} \\ (\epsilon_M \otimes I_{m_2}) \tilde{e}_{n-1}^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} \epsilon_M \otimes I_{m1} & 0 \\ 0 & \epsilon_M \otimes I_{m2} \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix}$$

onde

$$\epsilon_M = [1, 1, \dots, 1]^t, \quad Z_1 = \epsilon_M \otimes I_{m1} \quad \text{e} \quad Z_2 = \epsilon_M \otimes I_{m2}.$$

Assim,

$$\begin{aligned} Q_n^{-1} e_n &= \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} - h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \begin{bmatrix} \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} \\ &\quad - h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \begin{bmatrix} \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} \tilde{\delta}^{(1)} \\ \tilde{\delta}^{(2)} \end{bmatrix} + h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix} \\ &= \left\{ I - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \right\} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} \tilde{\delta}^{(1)} \\ \tilde{\delta}^{(2)} \end{bmatrix} \\ &\quad + h \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix} \bar{T}_n^{-1} \begin{bmatrix} \tilde{\eta}^{(1)} \\ \tilde{\eta}^{(2)} \end{bmatrix} + \begin{bmatrix} \tilde{\delta}_{M+1}^{(1)} \\ \tilde{\delta}_{M+1}^{(2)} \end{bmatrix}. \end{aligned}$$

Logo

$$Q_n^{-1} e_n = S_n Q_n^{-1} e_{n-1} - h U_n \tilde{\delta}^{(n)} + \tilde{\delta}_{M+1}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)}, \quad (4.3.21)$$

onde

$$S_n = \left\{ I - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \right\};$$

$$U_n = \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 + O(h) & O(h) \\ O(1) & b_2' \frac{\hat{T}_4^{-1}}{h} + O(1) \end{bmatrix};$$

$$\mathcal{K} = \begin{bmatrix} b_1' & 0 \\ 0 & b_2' \end{bmatrix};$$

$$\tilde{\delta}^{(n)} = [\tilde{\delta}^{(1)}, \tilde{\delta}^{(2)}]^t; \quad \tilde{\eta}^{(n)} = [\tilde{\eta}^{(1)}, \tilde{\eta}^{(2)}]^t \quad \text{e} \quad \tilde{\delta}_{M+1}^{(n)} = [\tilde{\delta}_{M+1}^{(1)}, \tilde{\delta}_{M+1}^{(2)}]^t.$$

Pelo Lema 4.3.4, temos $S_n = K + O(h)$, onde $K = \begin{bmatrix} I_{m1} & 0 \\ 0 & rI_{m2} \end{bmatrix}$.

Resolvendo (4.3.21), em função de e_n temos

$$\begin{aligned} e_n &= Q_n S_n Q_n^{-1} e_{n-1} - h Q_n U_n \tilde{\delta}^{(n)} + \tilde{\delta}_{M+1}^{(n)} + h Q_n \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \\ &= Q_n S_n Q_n^{-1} e_{n-1} + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right). \end{aligned}$$

Assim,

$$\begin{aligned} e_{n-1} &= Q_{n-1} S_{n-1} Q_{n-1}^{-1} e_{n-2} + Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) \\ e_{n-2} &= Q_{n-2} S_{n-2} Q_{n-2}^{-1} e_{n-3} + Q_{n-2} \left(\tilde{\delta}_{M+1}^{(n-2)} - h U_{n-2} \tilde{\delta}^{(n-2)} + h \mathcal{K} \bar{T}_{n-2}^{-1} \tilde{\eta}^{(n-2)} \right) \\ &\vdots \\ e_1 &= Q_1 S_1 Q_1^{-1} e_0 + Q_1 \left(\tilde{\delta}_{M+1}^{(1)} - h U_1 \tilde{\delta}^{(1)} + h \mathcal{K} \bar{T}_1^{-1} \tilde{\eta}^{(1)} \right) \end{aligned}$$

e, portanto, e_n torna-se

$$\begin{aligned} e_n &= Q_n S_n Q_n^{-1} \left[Q_{n-1} S_{n-1} Q_{n-1}^{-1} e_{n-2} + Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) \right] \\ &\quad + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right) \\ &= (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) e_{n-2} + (Q_n S_n Q_n^{-1}) Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) \\ &\quad + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right) \\ &= (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) \left[Q_{n-2} S_{n-2} Q_{n-2}^{-1} e_{n-3} + Q_{n-2} \left(\tilde{\delta}_{M+1}^{(n-2)} - h U_{n-2} \tilde{\delta}^{(n-2)} + h \mathcal{K} \bar{T}_{n-2}^{-1} \tilde{\eta}^{(n-2)} \right) \right] \\ &\quad + (Q_n S_n Q_n^{-1}) Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right) \\ &= (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) (Q_{n-2} S_{n-2} Q_{n-2}^{-1}) e_{n-3} \\ &\quad + (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) Q_{n-2} \left(\tilde{\delta}_{M+1}^{(n-2)} - h U_{n-2} \tilde{\delta}^{(n-2)} + h \mathcal{K} \bar{T}_{n-2}^{-1} \tilde{\eta}^{(n-2)} \right) \\ &\quad + (Q_n S_n Q_n^{-1}) Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right) \\ &= \dots = (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) \dots (Q_1 S_1 Q_1^{-1}) e_0 + Q_n \left(\tilde{\delta}_{M+1}^{(n)} - h U_n \tilde{\delta}^{(n)} + h \mathcal{K} \bar{T}_n^{-1} \tilde{\eta}^{(n)} \right) \\ &\quad + (Q_n S_n Q_n^{-1}) Q_{n-1} \left(\tilde{\delta}_{M+1}^{(n-1)} - h U_{n-1} \tilde{\delta}^{(n-1)} + h \mathcal{K} \bar{T}_{n-1}^{-1} \tilde{\eta}^{(n-1)} \right) + \dots \\ &\quad + (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) \dots (Q_2 S_2 Q_2^{-1}) Q_1 \left(\tilde{\delta}_{M+1}^{(1)} - h U_1 \tilde{\delta}^{(1)} + h \mathcal{K} \bar{T}_1^{-1} \tilde{\eta}^{(1)} \right) \end{aligned}$$

$$= \prod_{j=0}^{n-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1} e_0 + \sum_{i=0}^{n-1} \left[\left(\prod_{j=0}^{i-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1} \right) Q_{n-i} \left(\tilde{\delta}_{M+1}^{(n-i)} - h U_{n-i} \tilde{\delta}^{(n-i)} + h \mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)} \right) \right], \quad (4.3.22)$$

onde $\prod_{j=0}^{i-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1}$ é definido como sendo a identidade para $i = 0$. Agora,

$$\begin{aligned} \prod_{j=0}^{n-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1} &= (Q_n S_n Q_n^{-1}) (Q_{n-1} S_{n-1} Q_{n-1}^{-1}) \dots (Q_1 S_1 Q_1^{-1}) \\ &= Q_n (S_n Q_n^{-1} Q_{n-1}) (S_{n-1} Q_{n-1}^{-1} Q_{n-2}) \dots (S_2 Q_2^{-1} Q_1) S_1 Q_1^{-1} \\ &= Q_n (K + O(h)) (I + O(h)) (K + O(h)) (I + O(h)) \dots (K + O(h)) Q_1^{-1} \end{aligned}$$

visto que,

$$S_i = (K + O(h)), \quad Q_i^{-1} Q_{i-1} = (I + O(h))$$

pelo Lema 3.3.1 e observando que

$$(K + O(h))(I + O(h)) = K + KO(h) + O(h) + O(h^2) = K + O(h)$$

temos

$$\prod_{j=0}^{n-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1} = Q_n (K + O(h))^n Q_1^{-1}.$$

Analogamente, temos

$$\prod_{j=0}^{i-1} Q_{n-j} S_{n-j} Q_{n-j}^{-1} = Q_n (K + O(h))^i Q_{n-i+1}^{-1}.$$

Portanto, podemos reescrever (4.3.22)

$$e_n = Q_n (K + O(h))^n Q_1^{-1} e_0 + \sum_{i=0}^{n-1} Q_n (K + O(h))^i Q_{n-i+1}^{-1} Q_{n-i} \left(\tilde{\delta}_{M+1}^{(n-i)} - h U_{n-i} \tilde{\delta}^{(n-i)} + h \mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)} \right)$$

multiplicando à esquerda por Q_n^{-1} temos

$$\begin{aligned} \tilde{e}_n &= Q_n^{-1} Q_n (K + O(h))^n Q_1^{-1} e_0 + \sum_{i=0}^{n-1} Q_n^{-1} Q_n (K + O(h))^i Q_{n-i+1}^{-1} Q_{n-i} \left(\tilde{\delta}_{M+1}^{(n-i)} - h U_{n-i} \tilde{\delta}^{(n-i)} + h \mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)} \right) \\ &= (K + O(h))^n (I + O(h)) \tilde{e}_0 + \sum_{i=0}^{n-1} (K + O(h))^i (I + O(h)) \left(\tilde{\delta}_{M+1}^{(n-i)} - h U_{n-i} \tilde{\delta}^{(n-i)} + h \mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)} \right) \end{aligned}$$

Considerando

$$\hat{U}_n = \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 & 0 \\ 0 & b_2' \frac{\hat{T}_4^{-1}}{h} \end{bmatrix},$$

temos

$$\begin{aligned} \tilde{e}_n &= (K + O(h))^n (I + O(h)) \tilde{e}_0 + \sum_{i=0}^{n-1} (K + O(h))^i (I + O(h)) (\tilde{\delta}_{M+1}^{(n-i)} - h \hat{U}_{n-i} \tilde{\delta}^{(n-i)}) \\ &+ \sum_{i=0}^{n-1} (K + O(h))^i (I + O(h)) (h \mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)}) + \sum_{i=0}^{n-1} (K + O(h))^i (I + O(h)) h (\hat{U}_{n-i} - U_{n-i}) \tilde{\delta}^{(n-i)}. \end{aligned} \quad (4.3.23)$$

Devido a $Q(t, y(t), y'(t))$ ser uma função contínua em t , para sistemas de índice uniforme um e assumindo que a solução $y(t)$ da EAD é suficientemente suave, de maneira que $\tilde{\delta}^{(n-i)}$ e $\tilde{\delta}_{M+1}^{(n-i)}$ são contínuas, então $\tilde{\delta}_{M+1}^{(n-i)} - h \hat{U}_{n-i} \tilde{\delta}^{(n-i)}$ também é uma função contínua em t . Da mesma forma, devemos assumir que $\mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)}$ é uma função contínua em t , que é verdade sob moderadas suposições da suavidade de y , cujos erros de Newton são desconsiderados. Finalmente, pelas definições 4.3.1 e 4.3.2, temos que $\|\tilde{\delta}^{(n-i)}\| = O(h^{k_I+1})$.

Usando o fato de que $\delta_{M+1}^{(n-i)} = O(h^{k_d+1})$ e $b_1' \mathcal{A}^{-1} \delta^{(n-i)} - \delta_{M+1}^{(n-i)} = O(h^{k_a+1})$ (ver seção 4.2) temos

$$\begin{aligned} \tilde{\delta}_{M+1}^{(n-i)} - h \hat{U}_{n-i} \tilde{\delta}^{(n-i)} &= \begin{bmatrix} \tilde{\delta}_{M+1}^{(n-i)} \\ \tilde{\delta}_{M+1}^{(n-i)} \end{bmatrix} - h \begin{bmatrix} b_1' \hat{T}_1^{-1} \hat{S}_1 & 0 \\ 0 & \frac{b_2' \hat{T}_4^{-1}}{h} \end{bmatrix} \begin{bmatrix} \tilde{\delta}^{(n-i)} \\ \tilde{\delta}^{(n-i)} \end{bmatrix} = \begin{bmatrix} \tilde{\delta}_{M+1}^{(n-i)} - h b_1' \hat{T}_1^{-1} \hat{S}_1 \tilde{\delta}^{(n-i)} \\ \tilde{\delta}_{M+1}^{(n-i)} - b_2' \hat{T}_4^{-1} \tilde{\delta}^{(n-i)} \end{bmatrix} \\ &= \begin{bmatrix} O(h^{k_d+1}) - h O(1) O(h^{k_I+1}) \\ O(h^{k_a+1}) \end{bmatrix} = \begin{bmatrix} O(h^{\min(k_d+1, k_I+2)}) \\ O(h^{k_a+1}) \end{bmatrix}, \end{aligned}$$

onde $b_2' \hat{T}_4^{-1} = b_1' \mathcal{A}^{-1} \otimes I$.

Pelo Lema 4.3.5 temos, para $r \neq 1$,

$$(K + O(h))^i = \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix}.$$

Considerando $\bar{\eta}_1$ e $\bar{\eta}_2$ como sendo os erros de Newton e supondo que os termos não lineares de mais alta ordem satisfazem $\|\tilde{\eta}^{(i)}\| \leq \epsilon_i$ (aqui, não estamos incluindo os erros de

Newton que são $O(h^G)$, se $|r| < 1$ e $O(h^{G+1})$, se $|r| = 1$ e os erros nas condições iniciais satisfazem $\|\tilde{e}_0^{(i)}\| = O(\zeta_i)$, $i = 1, 2$, então, de (4.3.23), temos

$$\begin{aligned}
 \begin{bmatrix} \tilde{e}_n^{(1)} \\ \tilde{e}_n^{(2)} \end{bmatrix} &= \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix} \begin{bmatrix} O(\zeta_1) \\ O(\zeta_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h^{\min(k_d+1, k_I+2)}) \\ O(h^{k_a+1}) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} h \begin{bmatrix} b_1' \hat{T}_1^{-1} + O(h) & O(h) \\ O(h) & b_2' \hat{T}_4^{-1} h^{-1} + O(1) \end{bmatrix} \begin{bmatrix} O(\epsilon_1) + O(\bar{\eta}_1) \\ O(\epsilon_2) + O(\bar{\eta}_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} h \begin{bmatrix} O(h) & O(h) \\ O(1) & O(1) \end{bmatrix} \begin{bmatrix} O(h^{k_I+1}) \\ O(h^{k_I+1}) \end{bmatrix} \\
 &= \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix} \begin{bmatrix} O(\zeta_1) \\ O(\zeta_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h^{\min(k_d+1, k_I+2)}) \\ O(h^{k_a+1}) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h) & O(h^2) \\ O(h^2) & O(1) \end{bmatrix} \begin{bmatrix} O(\epsilon_1) + O(\bar{\eta}_1) \\ O(\epsilon_2) + O(\bar{\eta}_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h^{k_I+3}) \\ O(h^{k_I+2}) \end{bmatrix} \\
 &= \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^n I + O(h) \end{bmatrix} \begin{bmatrix} O(\zeta_1) \\ O(\zeta_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h^{\min(k_d+1, k_I+2)}) \\ O(h^{k_a+1}) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h \epsilon_1) + O(h \bar{\eta}_1) + O(h^2 \epsilon_2) + O(h^2 \bar{\eta}_2) \\ O(h^2 \epsilon_1) + O(h^2 \bar{\eta}_1) + O(\epsilon_2) + O(\bar{\eta}_2) \end{bmatrix} \\
 &+ \sum_{i=0}^{n-1} \begin{bmatrix} I + O(h) & O(h) \\ O(h) & r^i I + O(h) \end{bmatrix} \begin{bmatrix} O(h^{k_I+3}) \\ O(h^{k_I+2}) \end{bmatrix}
 \end{aligned} \tag{4.3.24}$$

Assim, separando em partes algébricas e diferenciais, temos os seguintes casos a analisar

i) $-1 < r < 1$:

$$\begin{aligned} \tilde{e}_n^{(1)} = & O(\zeta_1) + O(h\zeta_2) + O(h^{k_d}) + O(h^{k_I+1}) + O(h^{k_a+1}) + O(\epsilon_1) + O(\epsilon_2) \\ & + O(\bar{\eta}_h) + O(\bar{\eta}_2); \end{aligned} \quad (4.3.25a)$$

$$\begin{aligned} \tilde{e}_n^{(2)} = & O(h\zeta_1) + O(h\zeta_2) + O(h^{k_d+1}) + O(h^{k_I+2}) + O(h^{k_a+1}) + O(h\epsilon_1) + O(\epsilon_2) \\ & + O(h\bar{\eta}_h) + O(\bar{\eta}_2). \end{aligned} \quad (4.3.25b)$$

Consequentemente,

$$e_n = O(\zeta) + O(\bar{\eta}) + O(\epsilon) + O(h^G) \quad (4.3.26)$$

onde

$$G = \min(k_d, k_I + 1).$$

ii) $r = -1$:

$$\begin{aligned} \tilde{e}_n^{(1)} = & O(\zeta_1) + O(h\zeta_2) + O(h^{k_d}) + O(h^{k_I+1}) + O(h^{k_a+1}) + O(\epsilon_1) + O(\epsilon_2) \\ & + O(\bar{\eta}_h) + O(\bar{\eta}_2); \end{aligned}$$

$$\begin{aligned} \tilde{e}_n^{(2)} = & O(h\zeta_1) + O(\zeta_2) + O(h^{k_d+1}) + O(h^{k_I+2}) + O(h^{k_a+1}) + O(h\epsilon_1) + O(\epsilon_2) \\ & + O(h\bar{\eta}_h) + O(\bar{\eta}_2). \end{aligned}$$

Consequentemente,

$$e_n = O(\zeta) + O(\bar{\eta}) + O(\epsilon) + O(h^G) \quad (4.3.27)$$

onde

$$G = \min(k_d, k_I + 1).$$

O melhor resultado que podemos esperar para $\tilde{e}_n^{(2)}$, neste caso, é devido ao cancelamento da parte algébrica na soma de (4.3.23), que aparece devido a alteração de sinal no bloco direito de K , quando $r = -1$, junto com a suposição de suavidade. Isto pode ser visto através do agrupamento dos termos da soma de (4.3.23) e o limitante do resultado da soma.

iii) $r = 1$:

Para $i \leq n$ temos $(K + O(h))^i = O(1)$.

Assim, separando (4.3.23) em partes algébrica e diferencial, obtemos

$$\tilde{e}_n^{(1)} = O(\zeta_1) + O(\zeta_2) + O(h^{k_d}) + O(h^{k_I+1}) + O(h^{k_a}) + O(\epsilon_1) + O(h \epsilon_2) \\ + O(\bar{\eta}_h) + O(h\bar{\eta}_h);$$

$$\tilde{e}_n^{(2)} = O(\zeta_1) + O(\zeta_2) + O(h^{k_d}) + O(h^{k_I+1}) + O(h^{k_a}) + O(h \epsilon_1) + O(\epsilon_2) \\ + O(h\bar{\eta}_h) + O\left(\frac{\bar{\eta}_h}{h}\right).$$

Consequentemente,

$$e_n = O(\zeta) + O\left(\frac{\bar{\eta}}{h}\right) + O(\epsilon) + O(h^G) \quad (4.3.28)$$

onde

$$G = \begin{cases} \min(k_d, k_I), & \text{se } k_a \geq k_I \\ \min(k_d, k_I + 1), & \text{se } k_a \geq k_I + 1. \end{cases}$$

O melhor resultado que podemos esperar, neste caso, dos termos $O(\epsilon_1)$ e $O(\epsilon_2)$ em $\tilde{e}_n^{(2)}$ aparece devido a $b^t \mathcal{A}^{-1} \epsilon_M = 1 - r = 0$. Assim, a suavidade de $\tilde{\eta}$ (excluindo o erro de Newton) implica que o termo $O(h^{-1} \tilde{\eta}^{(2)})$ em $\mathcal{K} \bar{T}_{n-i}^{-1} \tilde{\eta}^{(n-i)}$ desaparece.

Assim, para problemas lineares ($\epsilon = 0$), obtemos o resultado do teorema observando que

$$\begin{aligned} C(q), B(q) &\Rightarrow k_I = q \\ C(q), B(q+1) &\Rightarrow k_a \geq q+1. \end{aligned}$$

Para sistemas não lineares, podemos assumir, sem perda de generalidade, que F tem a forma

$$F(t, y, y') = \begin{bmatrix} f(t, y, y') \\ g(t, y) \end{bmatrix}.$$

Então, a coordenada $\eta_i^{(1)}$ consiste dos resíduos da iteração de Newton e termos de mais alta ordem da forma

$$\left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)' f_{yy} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right),$$

$$\left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)' f_{yy'} E_i', \quad E_i' f_{y'y'} E_i'.$$

Isto pode se visto por (4.3.11). Observando que para sistemas não lineares os termos de mais alta ordem satisfazem $\|\tilde{\eta}^{(i)}\| \leq \epsilon_i$, assim

$$\begin{aligned} \eta_i^{(1)} &= \frac{1}{2} \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y_j' \right)' f_{yy'} (Y_i' - y'(t_{n-1} + c_i h)) \\ &+ \frac{1}{2} \left(y_{n-1} - y(t_{n-1}) + h \sum_{j=1}^M a_{ij} (Y_j' - y'(t_{n-1} + c_j h)) + \delta_i \right)' f_{yy} \left(y_{n-1} - y(t_{n-1}) + h \sum_{j=1}^M a_{ij} (Y_j' - y'(t_{n-1} + c_j h)) + \delta_i \right) \\ &+ \frac{1}{2} (y'(t_{n-1} + c_i h) - Y_i')' f_{y'y'} (Y_i' - y'(t_{n-1} + c_i h)) + \bar{\eta}^{(1)} \\ &= \frac{1}{2} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)' f_{yy} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right) \\ &- \frac{1}{2} \left(y(t_{n-1} + c_i h) - y_{n-1} - h \sum_{j=1}^M a_{ij} Y_j' \right)' f_{yy'} E_i' - \frac{1}{2} E_j' f_{y'y'} E_j' + \bar{\eta}^{(1)}. \end{aligned} \quad (4.3.29)$$

Podemos escrever

$$\begin{aligned} -y(t_{n-1} + c_i h) + y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j' &= \\ &= -y(t_{n-1}) - c_i h y'(t_{n-1}) - \dots + y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j' \\ &= y_{n-1} - y(t_{n-1}) - c_i h y'(t_{n-1}) - \dots + h \sum_{j=1}^M a_{ij} Y_j' \\ &= e_{n-1} - c_i h (y'(t_{n-1} + c_i h) + \dots) - \dots + h \sum_{j=1}^M a_{ij} Y_j' \end{aligned}$$

mas, lembrando que $c_i = \sum_{j=1}^M a_{ij}$ (pela definição de condições algébricas), temos

$$\begin{aligned} & -y(t_{n-1} + c_i h) + y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j' = \\ & = e_{n-1} + \sum_{j=1}^M a_{ij} (Y_j' - y'(t_{n-1} + c_i h) - \dots) - \dots \\ & = e_{n-1} + \sum_{j=1}^M a_{ij} (Y_j' - y'(t_{n-1} + c_i h) + \delta_j) + \delta_i \\ & = e_{n-1} + \sum_{j=1}^M a_{ij} (Y_j' - y'(t_{n-1} + c_i h)) + \underline{\delta_i} \end{aligned}$$

onde $\underline{\delta_i} = \sum_{j=1}^M a_{ij} \delta_j + \delta_i$. Assim (4.3.29) torna-se

$$\begin{aligned} \eta_i^{(1)} &= \frac{1}{2} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)^t f_{yy} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right) \\ & - \frac{1}{2} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \underline{\delta_i} \right)^t f_{yy'} E_i' - \frac{1}{2} E_i'^t F_{y'y} E_i' + \bar{\eta}^{(1)}. \end{aligned} \quad (4.3.30)$$

Usando o mesmo procedimento, temos que $\eta_i^{(2)}$ consiste dos resíduos da iteração de Newton e termos de mais alta ordem da forma

$$\left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)^t g_{yy} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right).$$

Logo,

$$\eta_i^{(2)} = \frac{1}{2} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right)^t g_{yy} \left(e_{n-1} + h \sum_{j=1}^M a_{ij} E_j' + \delta_i \right) + \bar{\eta}^{(2)}. \quad (4.3.31)$$

Como P_n tem a forma dada no Lema 3.3.1 e $\tilde{\eta}_n = P_n \eta_n$, segue que, $\tilde{\eta}_n^{(1)}$ e $\tilde{\eta}_n^{(2)}$ são compostos de termos da mesma forma que $\eta_n^{(1)}$ e $\eta_n^{(2)}$, respectivamente.

De (4.3.26), (4.3.27) e (4.3.28) temos que

$$e_n = O(h^G) + O(\epsilon_1) + O(\epsilon_2), \quad (4.3.32)$$

e segue de (4.3.18) que

$$E_i' = O(h^{G-1}) + O(\epsilon_1) + O\left(\frac{\epsilon_2}{h}\right). \quad (4.3.33)$$

Considerando

$$f_{yy} = \begin{bmatrix} f_{yy}^1 & f_{yy}^2 \\ g_{yy}^1 & g_{yy}^2 \end{bmatrix}, \quad f_{yy'} = \begin{bmatrix} f_{yy'}^1 & f_{yy'}^2 \\ 0 & 0 \end{bmatrix}, \quad f_{y'y'} = \begin{bmatrix} f_{y'y'}^1 & f_{y'y'}^2 \\ 0 & 0 \end{bmatrix}$$

e substituindo em (4.3.32), temos

$$\begin{aligned} \eta_i^{(1)} &= \frac{1}{2} \left[\left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right)^t \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right)^t \right] x \\ &\quad \left[\begin{array}{l} f_{yy}^1 \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right) + f_{yy}^2 \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right) \\ g_{yy}^1 \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right) + g_{yy}^2 \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right) \end{array} \right] - \\ &\quad - \frac{1}{2} \left[\left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right)^t \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right)^t \right] \begin{bmatrix} f_{yy'}^1 E_i^{(1)} + f_{yy'}^2 E_i^{(2)} \\ 0 \end{bmatrix} \\ &\quad - \frac{1}{2} \begin{bmatrix} E_i^{(1)t} & E_i^{(2)t} \end{bmatrix} \begin{bmatrix} f_{yy'}^1 E_i^{(1)} + f_{yy'}^2 E_i^{(2)} \\ 0 \end{bmatrix} \\ &= \frac{1}{2} \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right)^t f_{yy}^1 \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right) \\ &\quad + \frac{1}{2} \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right)^t f_{yy}^2 \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right) \\ &\quad + \frac{1}{2} \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right)^t g_{yy}^1 \left(e_{n-1}^{(1)} + h \sum_{j=1}^M a_{ij} E_j^{(1)} + \delta_i^{(1)} \right) \\ &\quad + \frac{1}{2} \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right)^t g_{yy}^2 \left(e_{n-1}^{(2)} + h \sum_{j=1}^M a_{ij} E_j^{(2)} + \delta_i^{(2)} \right) \end{aligned}$$

$$-\frac{1}{2}\left(e_{n-1}^{(1)} + h\sum_{j=1}^M a_{ij}E_j^{(1)} + \delta_i^{(1)}\right)^t f_{yy}^1 E_i^{(1)} + \frac{1}{2}\left(e_{n-1}^{(1)} + h\sum_{j=1}^M a_{ij}E_j^{(1)} + \delta_i^{(1)}\right)^t f_{yy}^2 E_i^{(2)}$$

Logo, por (4.3.2) e (4.3.3), temos

$$\|\tilde{\eta}_i^{(1)}\| \leq C_1 \left\| e_{n-1} + h\sum_{j=1}^M a_{ij}E_j' + \delta_i \right\|^2 + C_2 \|E_i'\|^2 \leq K_1 \left(h^{G-1} + \epsilon_1 + \frac{\epsilon_2}{h} \right)^2.$$

Analogamente, temos

$$\|\tilde{\eta}_i^{(2)}\| \leq K_2 (h^G + \epsilon_1 + \epsilon_2)^2$$

Agora, determinamos ϵ_1 e ϵ_2 como solução de

$$\begin{aligned} \epsilon_1 &= K_3 \left(h^{2(G-1)} + \epsilon_1^2 + \frac{\epsilon_2^2}{h^2} + \frac{\epsilon_1 \epsilon_2}{h} + h^{G-1} \epsilon_1 + h^{G-2} \epsilon_2 \right) \\ \epsilon_2 &= K_4 (h^{2G} + \epsilon_1^2 + \epsilon_2^2 + \epsilon_1 \epsilon_2 + h^G \epsilon_1 + h^G \epsilon_2), \end{aligned} \quad (4.3.34)$$

resolvendo (4.3.34) por uma iteração funcional $\hat{\epsilon} = G(\hat{\epsilon})$ com valor inicial $\hat{\epsilon}^{(0)} = O(h^{2(G-1)})$. Para $G \geq 3$, podemos usar o teorema da contração para concluirmos que $\hat{\epsilon} = O(h^{2(G-1)})$ e a demonstração está concluída. Para $G = 2$, não podemos aplicar o teorema diretamente, pois $\left\| \frac{\partial G}{\partial \hat{\epsilon}} \right\| = O(1)$. Mas se fizermos $\hat{\epsilon}_1 = \epsilon_1$ e $\hat{\epsilon}_2 = \frac{\epsilon_2}{\sqrt{h}}$, podemos aplicar a estratégia acima. Finalmente o resultado segue de (4.3.32) ■

Capítulo 5

Algoritmos

5.1 Introdução

Neste capítulo analisaremos o desenvolvimento de algoritmos para resolver sistemas de EAD, usando os métodos BDF e RKI e, em particular, os algoritmos utilizados em dois códigos já consagrados, o DASSL e o RADAU5.

O código DASSL, de Petzold [PETZOLD - 82], utiliza o método BDF. DASSL é indicado, de um modo geral, para resolver sistemas completamente implícito de EAD de índice zero ou um. Em geral, este código pode falhar em sistemas de índice elevado.

O código RADAU5, de Hairer e Wainner [HAIRER - 91], utiliza o método RADAU IIA de 3-estágios.

O método RADAU IIA é uma das várias classes existentes do método de RKI, obtido a partir das condições algébricas $B(q)$, $C(q)$ e $D(q)$ (ver equações (4.3.3)), onde estas condições satisfazem

$$B(2M-1), C(M), D(M-1), c_M = 1 \text{ e } b_i = a_{Mi}, \quad i = 1, \dots, M.$$

O código RADAU5 pode ser aplicado para resolver sistemas semi explícitos de índices zero, um, dois ou três. Mas, como já comentamos, neste trabalho abordaremos, apenas, sistemas de índice zero ou um.

Na seção 5.2 desenvolveremos o algoritmo básico usado no DASSL, incluindo as fórmulas do predictor, corretor e a solução de sistemas não lineares usando o método de Newton modificado.

Na seção 5.3 nos restringiremos a comentar sobre a implementação do método de RKI. Descreveremos como é calculado o sistema não linear, usando o método de Newton modificado e como escolher o valor inicial para a iteração de Newton.

5.2 Algoritmos e Estratégias no DASSL

DASSL é um código para resolver sistemas de equações algébrico-diferenciais, de índice zero ou um, na forma

$$\begin{aligned} F(t, y, y') &= 0 \\ y(t_0) &= y_0 \\ y'(t_0) &= y'_0 \end{aligned} \tag{5.2.1}$$

onde F , y e y' são vetores n -dimensionais.

O DASSL aproxima a derivada usando fórmulas de diferenças para trás (BDF) de k -ésima ordem, $k = 1, \dots, 5$ e, em todo passo, é escolhido a ordem k e o tamanho do passo h_{n+1} baseado no procedimento da solução.

5.2.1 Fórmula Básica

Aqui, descreveremos a fórmula básica usada no DASSL. Para encontrar uma aproximação para a solução em t_{n+1} , primeiramente, supomos um valor inicial para a solução através da avaliação do polinômio preditor e a sua derivada em t_{n+1} . O polinômio preditor w_{n+1}^P é um polinômio que interpola y_{n-i} , isto é,

$$w_{n+1}^P(t_{n-i}) = y_{n-i}, \quad i = 0, 1, 2, \dots, k. \tag{5.2.2}$$

Os valores previstos para y e y' são obtidos pela avaliação $w_{n+1}^P(t_{n+1})$ e $w_{n+1}'^P(t_{n+1})$. Assim,

$$\begin{aligned} y_{n+1}^{(0)} &= w_{n+1}^P(t_{n+1}) \\ y_{n+1}'^{(0)} &= w_{n+1}'^P(t_{n+1}). \end{aligned} \tag{5.2.3}$$

A aproximação y_{n+1} , para a solução em t_{n+1} , é a solução da fórmula do corretor e é, finalmente, reconhecida pelo DASSL. A fórmula usada é de coeficiente principal fixo de k -ésima ordem do método BDF.

A solução para a fórmula do corretor é o vetor y_{n+1} dado pelo polinômio corretor que interpola o polinômio predictor em k pontos igualmente espaçados $t_{n+1} - ih$, $i = 1, \dots, k$ e satisfaz (5.2.1), isto é,

$$\begin{aligned} w_{n+1}^C(t_{n+1}) &= y_{n+1} \\ w_{n+1}^C(t_{n+1} - ih_{n+1}) &= w_{n+1}^P(t_{n+1} - ih_{n+1}), \quad 1 \leq i \leq k \\ F(t_{n+1}, w_{n+1}^C(t_{n+1}), w_{n+1}'(t_{n+1})) &= 0. \end{aligned} \quad (5.2.4)$$

Na próxima subseção descreveremos como a solução y_{n+1} , definida implicitamente pelas condições (5.2.4), é determinada através da solução do sistema de equações não lineares, usando o método de Newton.

Os valores do predictor $y_{n+1}^{(0)}$, $y_{n+1}'^{(0)}$ e do corretor y_{n+1} em t_{n+1} são definidos em termos dos polinômios que interpola a solução em poucos passos. Segundo as idéias de Krogh [KROGH - 74] e Shampine e Gordon [SHAMPINE - 75], estes polinômios são representados no DASSL em termos de diferenças divididas modificada em y . Mais precisamente, as quantidades descritas passo a passo são dadas por

$$\begin{aligned} \psi_i(n+1) &= h_{n+1} + h_n + \dots + h_{n+2-i} = t_{n+1} - t_{n+1-i}, \quad i \geq 1 \\ \alpha_i(n+1) &= h_{n+1} / \psi_i(n+1), \quad i \geq 1 \\ \beta_1(n+1) &= 1 \\ \beta_i(n+1) &= \frac{\psi_1(n+1) \psi_2(n+1) \dots \psi_{i-1}(n+1)}{\psi_1(n) \psi_2(n) \dots \psi_{i-1}(n)}, \quad i > 1 \\ \phi_1(n) &= y_n \\ \phi_i(n) &= \psi_1(n) \psi_2(n) \dots \psi_{i-1}(n) [y_n, y_{n-1}, \dots, y_{n-i+1}], \quad i > 1 \\ \phi_i^*(n) &= \beta_i(n+1) \phi_i(n), \quad i \geq 1 \\ \rho_1(n+1) &= 1 \\ \sigma_i(n+1) &= \frac{h_{n+1}^i (i-1)!}{\psi_1(n+1) \psi_2(n+1) \dots \psi_i(n+1)}, \quad i > 1 \\ \gamma_1(n+1) &= 0 \\ \gamma_i(n+1) &= \gamma_{i-1}(n+1) + \alpha_{i-1}(n+1) / h_{n+1}, \quad i > 1 \end{aligned} \quad (5.2.5)$$

$$\alpha_s = -\sum_{j=1}^k \frac{1}{j}$$

$$\alpha^0(n+1) = -\sum_{j=1}^k \alpha_j(n+1).$$

As diferenças divididas são definidas pelas relações de recorrências

$$[y_n] = y_n$$

$$[y_n, y_{n-1}, \dots, y_{n-k}] = \frac{[y_n, y_{n-1}, \dots, y_{n-k+1}] - [y_{n-1}, y_{n-2}, \dots, y_{n-k}]}{t_n - t_{n-k}}.$$

O polinômio do preditor é dado em termos de diferenças divididas por

$$\begin{aligned} w_{n+1}^P(t) &= y_n + (t - t_n)[y_n, y_{n-1}] + (t - t_n)(t - t_{n-1})[y_n, y_{n-1}, y_{n-2}] + \dots \\ &\quad + (t - t_n)(t - t_{n-1}) \cdots (t - t_{n-k+1})[y_n, y_{n-1}, \dots, y_{n-k}]. \end{aligned} \quad (5.2.6)$$

Avaliando w_{n+1}^P em t_{n+1} e reescrevendo (5.2.6) em termos da notação de (5.2.5) temos

$$\begin{aligned} w_{n+1}^P(t_{n+1}) &= y_n + (t_{n+1} - t_n)[y_n, y_{n-1}] + (t_{n+1} - t_n)(t_{n+1} - t_{n-1})[y_n, y_{n-1}, y_{n-2}] + \dots \\ &\quad + (t_{n+1} - t_n)(t_{n+1} - t_{n-1}) \cdots (t_{n+1} - t_{n-k+1})[y_n, y_{n-1}, \dots, y_{n-k}] \\ &= y_n + \psi_1(n+1)[y_n, y_{n-1}] + \psi_1(n+1)\psi_2(n+1)[y_n, y_{n-1}, y_{n-2}] + \dots \\ &\quad + \psi_1(n+1)\psi_2(n+1) \cdots \psi_k(n+1)[y_n, y_{n-1}, \dots, y_{n-k}] \\ &= y_n + \frac{\psi_1(n+1)}{\psi_1(n)} \psi_1(n)[y_n, y_{n-1}] + \dots \\ &\quad + \frac{\psi_1(n+1) \cdots \psi_k(n+1)}{\psi_1(n) \cdots \psi_k(n)} \psi_1(n) \cdots \psi_k(n)[y_n, y_{n-1}, \dots, y_{n-k}] \\ &= \beta_1(n+1)\phi_1(n) + \dots + \beta_{k+1}(n+1)\phi_{k+1}(n) \\ &= \phi_1^*(n) + \phi_2^*(n) + \dots + \phi_k^*(n) = \sum_{i=1}^{k+1} \phi_i^*(n). \end{aligned}$$

Logo, a fórmula do preditor é $y_{n+1}^{(0)} = \sum_{i=1}^{k+1} \phi_i^*(n)$.

Agora, diferenciando (5.2.6) e avaliando em t_{n+1} , encontramos, após algumas manipulações,

$$y'_{n+1}^{(0)} = \sum_{i=1}^{k+1} \gamma_i(n+1) \phi_i^*(n).$$

Para encontrarmos a fórmula do corretor temos, das equações (5.2.2), (5.2.3) e (5.2.4), que

$$w_{n+1}^C(t) - w_{n+1}^P(t) = b(t)(y_{n+1} - y_{n+1}^{(0)}) \quad (5.2.7)$$

onde

$$\begin{aligned} b(t_{n+1} - ih_{n+1}) &= 0 \quad i = 1, 2, \dots, k \\ b(t_{n+1}) &= 1. \end{aligned}$$

Diferenciando a equação (5.2.7) e avaliando em t_{n+1} obtemos a equação do corretor para a implementação do coeficiente principal fixo,

$$\alpha_s(y_{n+1} - y_{n+1}^{(0)}) + h_{n+1}(y'_{n+1} - y'_{n+1}^{(0)}) = 0, \quad (5.2.8)$$

onde α_s está definido em (5.2.5).

Resolvendo (5.2.8) para y'_{n+1} e substituindo em (5.2.2) temos

$$F\left(t_{n+1}, y_{n+1}, y_{n+1}^{(0)} - \frac{\alpha_s}{h_{n+1}}(y_{n+1} - y_{n+1}^{(0)})\right) = 0. \quad (5.2.9)$$

5.2.2 Solução do Sistema não Linear

A equação do corretor (5.2.9) deve ser resolvida para y_{n+1} em cada passo. Aqui, descreveremos como é feita no DASSL.

Para simplificar a notação, reescreveremos a equação do corretor (5.2.9) como sendo

$$F(t, y, \alpha y + \beta) = 0, \quad (5.2.10)$$

com $\alpha = \frac{\alpha_s}{h_{n+1}}$ e $\beta = y_{n+1}^{(0)} - \alpha y_{n+1}^{(0)}$, sendo todas as variáveis avaliadas em t_{n+1} .

O corretor é resolvido usando o método de Newton modificado. Isto nos dá

$$y^{(m+1)} = y^{(m)} - G^{-1}F(t, y^{(m)}, \alpha y^{(m)} + \beta), \quad (5.2.11)$$

onde $y^{(0)}$ é dado por (5.2.6) e G é a matriz

$$G = \alpha F_y(t, y^{(0)}, \alpha y^{(0)} + \beta) + F_y(t, y^{(0)}, \alpha y^{(0)} + \beta).$$

Uma decisão, muito importante para garantirmos a eficiência de um código como DASSL, é quando terminamos a iteração do corretor (5.2.11).

Em DASSL o critério de parada é dado por

$$\|y^* - y^{(m+1)}\| \leq \frac{\rho}{1-\rho} \|y^{(m+1)} - y^{(m)}\|,$$

onde y^* é a solução da equação do corretor e ρ é uma estimativa do raio de convergência, sendo que esta estimativa é dada por

$$\rho = \left(\frac{\|y^{(m+1)} - y^{(m)}\|}{\|y^{(1)} - y^{(0)}\|} \right)^{1/m}.$$

5.3 Algoritmos e Estratégias no RADAU5

O método de RKI de M-estágio (4.1.4) envolve sistemas não lineares $F(t, y, y') = 0$ de dimensão $m \times M$, onde m é a dimensão de y . Uma solução para este sistema é a implementação do método de RKI, que mostraremos na subseção 5.3.1, inspirada pela técnica usada por Hairer e Wainner em seu código RADAU5.

O código RADAU5 é indicado para resolver sistemas semi explícitos de equações algébrico-diferenciais cuja forma é

$$By' = f(x, y), \quad y(x_0) = y_0,$$

onde B é uma matriz constante quadrada.

5.3.1 Solução do Sistema não Linear

Aplicando o método de RKI na EAD (4.1.1) obtemos

$$F\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j, Y_i\right) = 0 \quad i = 1, \dots, M. \quad (5.3.1)$$

Podemos, ainda, escrever este sistema na forma

$$\begin{aligned} 0 &= F\left(t_{n-1} + c_1 h, y_{n-1} + h \sum_{j=1}^M a_{1j} Y_j, Y_1\right) \\ &\vdots \\ 0 &= F\left(t_{n-1} + c_M h, y_{n-1} + h \sum_{j=1}^M a_{Mj} Y_j, Y_M\right). \end{aligned}$$

O método de Newton, aplicado a este sistema, é dado por

$$Z^{(k+1)} = Z^{(k)} - DG^{-1}(Z^{(k)})G(Z^{(k)}) \quad (5.3.2a)$$

sendo

$$Z^{(k)} = (Y_1^{(k)}, Y_2^{(k)}, \dots, Y_M^{(k)})^t;$$

$$G(Z^{(k)}) = \left[F\left(t_{n-1} + c_1 h, y_{n-1} + h \sum_{j=1}^M a_{1j} Y_j^{(k)}, Y_1^{(k)}\right), \dots, F\left(t_{n-1} + c_M h, y_{n-1} + h \sum_{j=1}^M a_{Mj} Y_j^{(k)}, Y_M^{(k)}\right) \right]^t;$$

$$DG(Z^{(k)}) = \begin{bmatrix} F_{y'}^1 + h a_{11} F_y^1 & \dots & h a_{1M} F_y^1 \\ \vdots & \ddots & \vdots \\ h a_{M1} F_y^M & \dots & F_{y'}^M + h a_{MM} F_y^M \end{bmatrix},$$

onde

$$F_y^i = F_y\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j^{(k)}, Y_i^{(k)}\right)$$

e

$$F_{y'}^i = F_y\left(t_{n-1} + c_i h, y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j^{(k)}, Y_i^{(k)}\right),$$

para $i = 1, \dots, M$.

Ao calcularmos a matriz $DG(Z^{(k)})$, em cada iteração de Newton, obtemos um custo muito alto, para evitarmos isto, usamos o método de Newton modificado. Este método faz com que a matriz $DG(Z^{(k)})$ seja a mesma em todas as iterações, isto é $F_y^* = f_y(t_{n-1}, y_{n-1})$ e $F_{y'}^* = B$. Desta forma $DG(Z^{(k)})$ torna-se

$$DG(Z^{(k)}) = \begin{bmatrix} F_y^* + ha_{11}F_y^* & \cdots & ha_{1M}F_y^* \\ \vdots & \ddots & \vdots \\ ha_{M1}F_y^* & \cdots & F_y^M + ha_{MM}F_y^* \end{bmatrix} = I \otimes F_y^* + hA \otimes F_y^*$$

Logo, para resolvermos (5.3.1), usamos a iteração de Newton modificado, isto é

$$\begin{aligned} (I \otimes F_y^* + hA \otimes F_y^*)S^{(k)} &= -G(Z^{(k)}) \\ Z^{(k+1)} &= Z^{(k)} + S^{(k)}. \end{aligned} \quad (5.3.2b)$$

Temos que a matriz $(I \otimes F_y^* + hA \otimes F_y^*)$ é inversível. De fato, pelo Lema 3.3.1, temos

$$PF_y^*Q = \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix} \text{ e } PF_y^*Q = \begin{bmatrix} C & 0 \\ 0 & I_{m_2} \end{bmatrix}.$$

Assim, multiplicando $(I \otimes F_y^* + hA \otimes F_y^*)$ à esquerda pela matriz $I_m \otimes P$ e à direita pela matriz $I_m \otimes Q$ obtemos

$$\begin{aligned} (I_m \otimes P)(I \otimes F_y^* + hA \otimes F_y^*)(I_m \otimes Q) &= \\ &= \begin{bmatrix} P & & \\ & \ddots & \\ & & P \end{bmatrix} \left\{ \begin{bmatrix} F_y^* & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & F_y^* \end{bmatrix} + \begin{bmatrix} ha_{11}F_y^* & \cdots & ha_{1M}F_y^* \\ \vdots & \ddots & \vdots \\ ha_{M1}F_y^* & \cdots & ha_{MM}F_y^* \end{bmatrix} \right\} \begin{bmatrix} Q & & \\ & \ddots & \\ & & Q \end{bmatrix} \\ &= \begin{bmatrix} PF_y^*Q & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & PF_y^*Q \end{bmatrix} + h \begin{bmatrix} a_{11}PF_y^*Q & \cdots & a_{1M}PF_y^*Q \\ \vdots & \ddots & \vdots \\ a_{M1}PF_y^*Q & \cdots & a_{MM}PF_y^*Q \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} + h \begin{bmatrix} a_{11} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} & \cdots & a_{1M} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \\ & \ddots & \\ a_{M1} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} & \cdots & a_{MM} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \end{bmatrix} \end{aligned}$$

Considere R sendo uma matriz de permutação, onde

$$R = [e_1, \dots, e_{m_1}, e_{m_1+1}, \dots, e_{m_1+m_2}, \dots, e_{(M-1)m_1+1}, \dots, e_{(M-1)m_1+m_2}, e_{m_1+1}, \dots, e_m, e_{m+m_1+1}, \dots, e_{2m}, \dots, e_{(M-1)m+m_1+1}, \dots, e_{Mm}]$$

e $m = m_1 + m_2$.

Logo

$$\begin{aligned}
 & R(I_m \otimes P)(I \otimes F_y^* + h\mathcal{A} \otimes F_y^*)(I_m \otimes Q)R = \\
 & = \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} + h \begin{bmatrix} \begin{bmatrix} a_{11}C & \cdots & a_{1M}C \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MM} \end{bmatrix} & \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} & \begin{bmatrix} a_{11}C & \cdots & a_{1M}C \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MM} \end{bmatrix} \end{bmatrix} \\
 & = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + h \begin{bmatrix} \mathcal{A} \otimes C & 0 \\ 0 & \mathcal{A} \otimes I \end{bmatrix} = \begin{bmatrix} I + h\mathcal{A} \otimes C & 0 \\ 0 & h\mathcal{A} \otimes I \end{bmatrix}
 \end{aligned}$$

Portanto,

$$R(I_m \otimes P)(I \otimes F_y^* + h\mathcal{A} \otimes F_y^*)(I_m \otimes Q)R = \begin{bmatrix} I + h\mathcal{A} \otimes C & 0 \\ 0 & h\mathcal{A} \otimes I \end{bmatrix}$$

implicando em

$$(I \otimes F_y^* + h\mathcal{A} \otimes F_y^*)^{-1} = R(I_m \otimes Q^{-1}) \begin{bmatrix} I + h\mathcal{A} \otimes C & 0 \\ 0 & h\mathcal{A} \otimes I \end{bmatrix}^{-1} (I_m \otimes P^{-1})R$$

Concluindo, assim, que $(I \otimes F_y^* + h\mathcal{A} \otimes F_y^*)$ é inversível para $h < \frac{1}{\|\mathcal{A} \otimes C\|}$.

Pelo fato da convergência ser linear, isto é, $\|S^{(k+1)}\| \leq \rho \|S^{(k)}\|$, $\rho < 1$, o critério de parada usado em RADAU5 é dado por

$$\|Z^{(k+1)} - Z^*\| \leq \frac{\rho}{1-\rho} \|Z^{(k+1)} - y^{(k)}\|, \quad (5.3.3)$$

onde ρ é uma estimativa do raio de convergência

$$\rho_k = \frac{\|Z^{(k+1)} - Z^{(k)}\|}{\|Z^{(k)} - Z^{(k-1)}\|}, \quad k \geq 1, \quad (5.3.4)$$

O erro iteração não deve ser maior que a discretização do erro local. Sendo, usualmente, considerado como sendo próximo de uma certa tolerância (Tol). Por esta razão, a iteração para quando

$$\eta_k \|S^{(k)}\| \leq k.Tol, \quad (5.3.5)$$

com $\eta_k = \frac{\rho_k}{1-\rho_k}$ e aceitando $Z^{(k+1)}$ como sendo uma aproximação de Z^* . Esta estratégia só pode ser aplicada após, no mínimo, duas iterações.

5.3.2 Valor Inicial para a Iteração de Newton

Seja $Y_i = y_{n-1} + h \sum_{j=1}^M a_{ij} Y_j$, $i = 1, \dots, M$ e considere $z_i = Y_i - y_{n-1}$, onde $z_i = O(h)$, pois

$$z_i = h \sum_{j=1}^M a_{ij} Y_j, \quad i = 1, \dots, M. \quad (5.3.6)$$

Logo, uma escolha natural e simples do valor inicial para a iteração de Newton (5.3.2) pode ser

$$z_i^{(0)} = 0, \quad i = 1, \dots, M. \quad (5.3.7)$$

Porém, em geral, escolhas melhores são possíveis. Se o método de RKI satisfaz a condição $C(q)$ (Capítulo 4), para algum $q < M$, então

$$z_i = y(t_{n-1} + c_i h) - y_{n-1} + O(h^{q+1}). \quad (5.3.8)$$

Suponha, agora que $c_i \neq 0$, $i = 1, 2, 3$, pois estamos usando o método RADAU IIA de 3-estágios, e considere o polinômio interpolador de grau 3 definido por

$$\begin{aligned} Q(0) &= 0 \\ Q(c_i) &= z_i^{\text{old}}, \quad i = 1, 2, 3, \end{aligned}$$

onde z_i^{old} são determinados no passo anterior.

Desde que o erro na interpolação é $O(h^{q+1})$, obtemos junto com (5.3.8) $y(t_{n-1} + c_i h) - y_{n-1} - Q(t) = O(h^{q+1})$.

Usando valores de $Q(t)$, além do intervalo $[0,1]$ (ver Figura 5.3.1), obtemos

$$z_i^{(0)} = Q(1 + wc_i) + y_n - y_{n-1}, \quad i = 1, 2, 3 \quad \text{e} \quad w = \frac{h_{new}}{h_{old}}. \quad (5.3.9)$$

Experimentos numéricos com o método RADAU IIA de 3-estágios mostram que (5.3.9) produz uma convergência mais segura que (5.3.7).

Figura 5.3.1 - Polinômio $Q(t)$

Capítulo 6

Resultados Numéricos

6.1 Introdução

Neste capítulo apresentaremos alguns exemplos de sistemas de EAD de índice zero ou um, usaremos os códigos DASSL e RADAU5 para obtermos a solução de cada exemplo. Os exemplos que apresentaremos aqui, são

- Equação do Pêndulo;
- Equação de Van der Pol;
- Equação da Esfera;
- Equação Teste.

Estes exemplos foram escolhidos devido as características de suas equações.

A equação do pêndulo é uma EAD não linear semi explícita de índice três mas, aqui, usaremos a equação após duas diferenciações reduzindo, assim, o índice em um, como veremos na seção 6.2.

A equação de Van der Pol é uma EAD não linear semi explícita de índice zero.

E a equação da Esfera é uma EAD não linear implícita de índice zero.

Quanto a equação Teste, que é uma EAD linear de índice zero, foi escolhida de tal forma que possamos analisar a estabilidade dos métodos.

Relembrando que o código RADAU5 é indicado para resolver, apenas, sistemas na forma semi explícita devemos, então, reescrever a equação da esfera nesta forma. Quanto ao código de DASSL, não temos este problema e portanto, apresentaremos os resultados da equação da esfera usando a forma semi explícita e implícita, o que será discutido na seção 6.4.

6.2 Equação do Pêndulo

Seja a EAD semi explícita não linear

$$p' = u \quad (6.2.1a)$$

$$q' = v \quad (6.2.1b)$$

$$mu' = -p\lambda \quad (6.2.1c)$$

$$mv' = -q\lambda - g \quad (6.2.1d)$$

$$0 = p^2 + q^2 - l^2. \quad (6.2.1e)$$

Nesta formulação o sistema (6.2.1) tem índice três. Diferenciando (6.2.1e) temos

$$0 = pu + qv. \quad (6.2.2)$$

Assim, as equações (6.2.1a)-(6.2.1d) e (6.2.2) formam a EAD não linear semi explícita de índice dois. Diferenciado mais uma vez e usando (6.2.1e) obtemos

$$0 = m(u^2 + v^2) - gq - l^2\lambda, \quad (6.2.3)$$

que corresponde ao sistema de índice um, isto é, (6.2.1a)-(6.2.1d) e (6.2.3) formam a EAD não linear semi explícita de índice um.

Como já comentamos, daremos os resultados numéricos apenas da formulação do equação da pêndulo de índice um. Para nossa análise usaremos o intervalo $[0,10]$, consideraremos $l = 1$, $g = 1$, $m = 1$ e valores iniciais consistentes

$$p(0) = 1, \quad q(0) = 0, \quad u(0) = 0, \quad v(0) = 0 \quad \text{e} \quad \lambda(0) = 0.$$

A solução obtida pelo DASSL com uma tolerância para o erro local $ATOL = RTOL = 1.0D - 5$ encontra-se na Tabela 6.2.1,

x	p	q
0.00	0.1000000000E+01	0.0000000000E+00
1.00	0.8795015006E+00	-0.4757984996E+00
2.00	-0.2041872563E+00	-0.9789636645E+00
3.00	-0.9688347838E+00	-0.2477426641E+00
4.00	-0.9990188721E+00	-0.4276027400E-01
5.00	-0.6851283926E+00	-0.7283317595E+00
6.00	0.5645477174E+00	-0.8255042201E+00
7.00	0.9961859922E+00	-0.8689733379E-01
8.00	0.9853129756E+00	-0.1698778608E+00
9.00	0.3682141699E+00	-0.9296837137E+00
10.00	-0.8115772951E+00	-0.5844822734E+00

Tabela 6.2.1: Solução Equação do Pêndulo usando o DASSL

cujas estatísticas do resultado do método são:

- 2337 avaliações da função;
- 172 avaliações da jacobiana;
- 1692 passos computados, sendo que 21 foram rejeitados.

A solução obtida pelo RADAU5 com uma tolerância para o erro local $ATOL = RTOL = 1.0D - 5$ encontra-se na Tabela 6.2.2,

x	p	q
0.00	0.1000000000E+01	0.0000000000E+00
1.00	0.8795328332E+00	-0.4758128248E+00
2.00	-0.2042128917E+00	-0.9789203122E+00
3.00	-0.9688684308E+00	-0.2476518781E+00
4.00	-0.9991221412E+00	-0.4267367969E-01
5.00	-0.6852563897E+00	-0.7283108499E+00
6.00	0.5646441095E+00	-0.8253279488E+00
7.00	0.9962728515E+00	-0.8655024711E-01
8.00	0.9855612854E+00	-0.1695384548E+00
9.00	0.3685838004E+00	-0.9295984457E+00
10.00	-0.8117276987E+00	-0.5839949224E+00

Tabela 6.2.2: Solução Equação do Pêndulo usando o RADAU5

cujas estatísticas do resultado do método são:

- 538 avaliações da função;
- 53 avaliações da jacobiana;
- 62 passos computados, sendo que 57 foram aceitos e 2 rejeitados;
- 62 decomposições LU foram calculadas, ou seja, quantidade de vezes que o sistema linear é resolvido;
- 160 substituições para frente e para trás de ambos sistemas dados por L e U.

6.3 Equação de Van der Pol

Seja a EAD não linear semi explícita de índice zero

$$\begin{aligned} y' &= z \\ \mathcal{E}z' &= (1 - y^2)z - y. \end{aligned} \quad (6.3.1)$$

Para a nossa análise usaremos o intervalo $[0,2]$ e consideraremos valores iniciais consistentes

$$y(0) = 2, \quad z(0) = -0.66,$$

com $\mathcal{E} = 1.0D - 6$.

A solução obtida pelo DASSL com $ATOL = RTOL = 1.0D - 5$ encontra-se na Tabela 6.3.1,

x	y	z
0.00	0.2000000000E+01	-0.6600000000E+00
0.10	0.1931343453E+01	-0.7074288674E+00
0.20	0.1858188832E+01	-0.7575585837E+00
0.30	0.1779380243E+01	-0.8214305129E+00
0.40	0.9069492018E+00	-0.9069492018E+00
0.50	0.1596748936E+01	-0.1030407605E+01
0.60	0.1484551264E+01	-0.1233128055E+01
0.70	0.1342853719E+01	-0.1671758691E+01
0.80	0.1083748147E+01	-0.6207686279E+01
0.90	-0.1936393879E+01	0.7042404907E+00
1.00	-0.1863586036E+01	0.7535888277E+00
1.10	-0.1785232238E+01	0.8162717651E+00
1.20	-0.1699644783E+01	0.8998517064E+00
1.30	-0.1604063507E+01	0.1019729996E+01
1.40	-0.1493272169E+01	0.1214168340E+01
1.50	-0.1354585174E+01	0.1622434687E+01
1.60	-0.1120353612E+01	0.4389615940E+01
1.70	0.1941421458E+01	-0.7010980566E+00
1.80	0.1868953098E+01	-0.7496773787E+00
1.90	0.1791021802E+01	-0.8112373648E+00
2.00	0.1706022520E+01	-0.8929650991E+00

Tabela 6.3.1: Equação de Van der Pol usando o DASSL

cujas estatísticas do resultado do método são:

14365 avaliações da função;
 1233 avaliações da jacobiana;
 8385 passos computados, sendo que 273 foram rejeitados.

A solução obtida pelo RADAU5 com $ATOL = RTOL = 1.0D - 5$ encontra-se na Tabela 6.3.2,

x	y	z
0.00	0.2000000000E+01	-0.6600000000E+00
0.10	0.1931361529E+01	-0.7074151877E+00
0.20	0.1858204676E+01	-0.7575443999E+00
0.30	0.1779394566E+01	-0.8214201331E+00
0.40	0.1693209270E+01	-0.9069428212E+00
0.50	0.1596768703E+01	-0.1030401438E+01
0.60	0.1484575388E+01	-0.1233048474E+01
0.70	0.1342892320E+01	-0.1671615654E+01
0.80	0.1083921569E+01	-0.6195343946E+01
0.90	-0.1936443329E+01	0.7042084938E+00
1.00	-0.1863646716E+01	0.7535455796E+00
1.10	-0.1785295380E+01	0.8162134722E+00
1.20	-0.1699713795E+01	0.8997617000E+00
1.30	-0.1604147430E+01	0.1019596637E+01
1.40	-0.1493387337E+01	0.1213935300E+01
1.50	-0.1354755772E+01	0.1621820474E+01
1.60	-0.1120811876E+01	0.4373845558E+01
1.70	0.1941502817E+01	-0.7010468156E+00
1.80	0.1869058815E+01	-0.7496098398E+00
1.90	0.1791154047E+01	-0.8111194791E+00
2.00	0.1706167729E+01	-0.8928088236E+00

Tabela 6.3.2: Equação de Van der Pol usando o RADAU5

cuja estatísticas do resultado do método são:

- 3473 avaliações da função;
- 294 avaliações da jacobiana;
- 476 passos computados, sendo que 463 foram aceitos e 12 rejeitados;
- 379 decomposições LU foram calculadas, ou seja, quantidade de vezes que o sistema linear é resolvido;
- 1003 substituições para frente e para trás de ambos sistemas dados por L e U.

6.4 Equação da Esfera

Seja a EAD não linear implícita

$$x^2 + y^2 + y'^2 = 1. \quad (6.4.1)$$

Como RADAU5 resolve apenas sistemas na forma semi explícita, reescreveremos (6.4.1) como sendo

$$y' = \pm \sqrt{1 - x^2 - y^2}, \quad (6.4.2)$$

que é a equação da esfera na forma semi explícita.

Para a nossa análise usaremos o intervalo $[0,1]$ e consideraremos valor inicial consistente $y(0) = 0$. Veremos que a equação da esfera tem singularidade quando $x = \pm 0.7975005D+00$, também observaremos que o código DASSL passa por esta singularidade o mesmo não ocorre com o código RADAU5.

A solução obtida pelo DASSL, usando a parte positiva da equação (6.4.2), isto é, $y' = \sqrt{1 - x^2 - y^2}$, com $ATOL = RTOL = 1.0D - 6$ encontra-se na Tabela 6.4.1,

x	y
0.00	0.0000000000E+00
0.05	0.4997917108E-01
0.10	0.9983347389E-01
0.15	0.1494383789E+00
0.20	0.1986698677E+00
0.25	0.2474050826E+00
0.30	0.2955219490E+00
0.35	0.3429002283E+00
0.40	0.3894216083E+00
0.45	0.4349694214E+00
0.50	0.4794299565E+00
0.55	0.5226920156E+00
0.60	0.5646476071E+00
0.65	0.6051919185E+00
0.70	0.6442235985E+00
0.75	0.6816448558E+00
0.80	0.7173620816E+00
0.85	0.7512861753E+00
0.90	0.7833323466E+00
0.95	0.8134206801E+00
1.00	0.8414759471E+00

Tabela 6.4.1: Equação da Esfera usando o DASSL

cujas estatísticas do resultado do método são:

1327 avaliações da função;
677 avaliações da jacobiana;
1011 passos computados, sendo que 10 foram rejeitados.

A solução obtida pelo DASSL, usando equação implícita (6.4.1), com $ATOL = RTOL = 1.0D - 6$ encontra-se na Tabela 6.4.2,

x	y
0.00	0.0000000000E+00
0.05	0.4997917108E-01
0.10	0.9983347389E-01
0.15	0.1494383780E+00
0.20	0.1986698593E+00
0.25	0.2474050324E+00
0.30	0.2955218577E+00
0.35	0.3429000580E+00
0.40	0.3894212026E+00
0.45	0.4349689708E+00
0.50	0.4794295539E+00
0.55	0.5226917431E+00
0.60	0.5646474364E+00
0.65	0.6051917037E+00
0.70	0.6442231276E+00
0.75	0.6816439762E+00
0.80	0.7173610736E+00
0.85	0.7512851310E+00
0.90	0.7833314321E+00
0.95	0.8134198527E+00
1.00	0.8414749838E+00

Tabela 6.4.2: Equação da Esfera usando o DASSL

cujas estatísticas do resultado do método são:

1185 avaliações da função;
677 avaliações da jacobiana;
1009 passos computados, e nenhum passo rejeitado.

A solução obtida pelo RADAU5, usando a parte positiva da equação (6.4.2), com $ATOL = RTOL = 1.0D - 6$ encontra-se na Tabela 6.4.3,

x	y
0.00	0.0000000000E+00
0.05	0.4995822516E-01
0.10	0.9966642867E-01
0.15	0.1488724154E+00
0.20	0.1973223552E+00
0.25	0.2447573847E+00
0.30	0.2909126948E+00
0.35	0.3355140951E+00
0.40	0.3782748086E+00
0.45	0.4188897108E+00
0.50	0.4570282005E+00
0.55	0.4923192885E+00
0.60	0.5243338768E+00
0.65	0.5525404551E+00
0.70	0.5762208042E+00
0.75	0.5942204117E+00

Tabela 6.4.3: Equação da Esfera usando o RADAU5

cujas estatísticas do resultado do método são:

- 474 avaliações da função;
- 42 avaliações da jacobiana;
- 114 passos computados, sendo que 65 foram aceitos e 48 rejeitados;
- 87 decomposições LU foram calculadas,, ou seja, quantidade de vezes que o sistema linear é resolvido;
- 136 substituições para frente e para trás de ambos sistemas dados por L e U.

Como comentamos no início desta seção, o código RADAU5 para quando $x = 0.7975005D+00$ e $h = 0.3455723D-15$.

6.5 Equação Teste

Seja a EAD linear

$$y' = \lambda y \quad (6.5.1)$$

Para a nossa análise usaremos o intervalo $[-10,10]$ com $\lambda = 50$ e consideraremos valor inicial consistente $y(0) = 1$.

Neste exemplo, verificaremos a estabilidade dos métodos.

A solução obtida pelo DASSL com $ATOL = RTOL = 1.0D - 6$ encontra-se na Tabela 6.5.1,

x	y	Erro
-10.00	-0.1860210341E-16	0.2610976758E+200
-9.00	-0.2215223637E-16	0.5997005308E+179
-8.00	-0.2570236933E-16	0.1342041424E+158
-7.00	-0.1037954355E-14	0.1045314372E+138
-6.00	-0.2365326141E-14	0.4594471877E+116
-5.00	-0.3692697927E-14	0.1383452519E+95
-4.00	-0.1319363747E-13	0.9533687826E+73
-3.00	-0.2620051925E-12	0.3651591470E+53
-2.00	-0.4420550209E-11	0.1188295646E+33
-1.00	-0.2861105987E-09	0.1483399234E+13
0.00	1.0000000000E+01	0.0000000000E+00
1.00	0.5185555442E+22	0.1639270486E-03
2.00	0.2689010223E+44	0.3322328557E-03
3.00	0.1394407196E+66	0.5005458370E-03
4.00	0.7230807133E+87	0.6688877475E-03
5.00	0.3749591459+109	0.8372833836E-03
6.00	0.1944379863+131	0.1005684491E-02
7.00	0.1008273322+153	0.1174109406E-02
8.00	0.5228479980+174	0.1342589501E-02
9.00	0.2711269035+196	0.1511078328E-02
10.00	0.1405949675+218	0.1679588226E-02

Tabela 6.5.1: Equação Teste usando o DASSL

cujas estatísticas do resultado do método são:

58279 avaliações da função;
 879 avaliações da jacobiana;
 55741 passos computados, sendo que 140 foram rejeitados;

A solução obtida pelo RADAU5 com $ATOL = RTOL = 1.0D - 6$ encontra-se na Tabela 6.5.2,

x	y	Erro
-10.00	0.4747760941E-14	0.6663920309E+203
- 9.00	0.7362201193E-13	0.1993079114E+183
- 8.00	-0.4903333626E-13	0.2560260791E+161
- 7.00	-0.1449712942E-12	0.1459992693E+140
- 6.00	0.4055127546E-14	0.7876786782E+116
- 5.00	0.4308464538E-11	0.1614146685E+98
- 4.00	0.2045322112E-11	0.1477944393E+76
- 3.00	-0.9244712585E-11	0.1288444450E+55
- 2.00	-0.4479671822E-11	0.1204188262E+33
- 1.00	0.3813428523E-10	0.1977150395E+12
0.00	0.1000000000E+01	0.0000000000E+00
1.00	0.5184706104E+22	0.1109998195E-06
2.00	0.2688117384E+44	0.9001393152E-07
3.00	0.1393710033E+66	0.3246253219E-06
4.00	0.7225975277E+87	0.2088606472E-06
5.00	0.3746456379E+109	0.4709285563E-06
6.00	0.1942427112E+131	0.3691675860E-06
7.00	0.1007091395E+153	0.5046256164E-06
8.00	0.5221472676E+174	0.5720042493E-06
9.00	0.2707179942E+196	0.6150909877E-06
10.00	0.1403593230E+218	0.7212839936E-06

Tabela 6.5.2: Equação Teste usando o RADAU5

cuja estatísticas do resultado do método são:

- 22042 avaliações da função;
- 2 avaliações da jacobiana;
- 5038 passos computados, sendo que 5038 foram aceitos e nenhum rejeitado;
- 69 decomposições LU foram calculadas,, ou seja, quantidade de vezes que o sistema linear é resolvido;
- 5668 substituições para frente e para trás de ambos sistemas dados por L e U.

Capítulo 7

Conclusão

Neste trabalho estudamos EAD, apenas, de índices zero ou um. O motivo disto é que queríamos fazer comparações com os resultados dos métodos BDF e RKI usando os códigos DASSL e RADAU5, respectivamente.

Os métodos de passos múltiplos, em particular o método BDF, aplicados em EAD semi explícita de índice um são estáveis e convergentes para EAD com a mesma ordem de convergência de uma EDO.

Muitas vezes os métodos de Runge-Kutta Implícito não alcançam a mesma ordem de exatidão em EAD como o fazem em EDO.

Vimos que os códigos DASSL e RADAU5 podem ser aplicados para resolver sistemas de EAD de índices zero ou um. Ambos os códigos resolvem os sistemas não lineares usando o método de Newton com algumas modificações.

Ao fazermos a análise dos resultados dos exemplos apresentados no Capítulo 6 observamos que:

i) O código DASSL resolve sistemas de EAD tanto na forma semi explícita quanto na forma implícita, o mesmo não ocorre com o RADAU5, que resolve apenas sistemas de EAD na forma semi explícita;

ii) O DASSL apresenta melhores resultados quando usamos a EAD na forma implícita, como pode ser visto na equação da esfera. Também, pode ser observado que o custo na forma implícita é menor;

iii) No caso da equação da esfera, o código RADAU5 é muito mais estável perto da singularidade, enquanto que o DASSL passa por esta singularidade.

Ao rodarmos os exemplos numéricos vimos que o código RADAU5 exige muito mais informações, sobre a EAD, que o DASSL, tornando assim o código mais estável, em compensação não é necessária tantas informações para utilizar o código DASSL, facilitando o trabalho do usuário.

Os códigos em discussão podem ser calculados usando o seguinte processo:

- o próprio código calcula as derivadas parciais;
- o usuário fornece as derivadas parciais,

em ambos os casos foi observado que não há diferenças significativas do resultado.

Para os exemplos apresentados no Capítulo 6 podemos observar que o custo do método BDF em relação ao método de Runge-Kutta é mais alto.

Bibliografia

[BRENAN - 89]: BRENNAN, K. E., CAMPBELL, S. L. & PETZOLD, L. R., *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North-Holland (1989).

[CAMPBELL - 82]: CAMPBELL, S. L., *Singular Systems of Differential-Equations II*, Pitman (1982).

[FRANK - 85a]: FRANK, R., SCHNEID, J. & UEBERHUBER, C.W., *Stability properties of implicit Runge-Kutta methods*, SIAM J. Numer. Anal., 22 (1985), 497 - 514.

[FRANK - 85b]: FRANK, R., SCHNEID, J. & UEBERHUBER, C.W., *Order results for implicit Runge-Kutta methods applied to stiff systems*, SIAM J. Numer. Anal., 22 (1985), 515 - 534.

[GEAR - 71]: GEAR, C. W., *The simultaneous numerical solution of differential-algebraic equations*, IEEE Trans. Circuit Theory, CT - 18 (1971), 89 - 95.

[GEAR - 83]: GEAR, C. W. & PETZOLD, L. R., *Differential / algebraic systems and matrix pencils* in Matrix Pencils, edited by B. Kagstrom and A. Ruhe, Lecture Notes in Mathematics 973, Springer-Verlag (1983), 75 - 89.

[GRIEPIENTROG - 86]: GRIEPIENTROG, E. & MÄRZ, R., *Differential / Algebraic Equations and their Numerical Treatment*, Teubner-texte Zur Mathematik, Band 88, 1986.

- [HAIRER - 89]: HAIRER, E., LUBICH, C. & ROCHE, M., *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Springer-Verlag (1989).
- [HAIRER - 91]: HAIRER, E. & WANNER, G., *Solving Ordinary Differential Equations II - Stiff and Differential-Algebraic Problems*, Springer-Verlag (1991).
- [KROGH - 74]: KROGH, F.T., *Changing step in the integration of differential equations using modified divided differences*, Proc. Conf. Num. Solution of ODEs, Lecture Notes in Mathematics No. 362, Springer-Verlag, New York, (1974).
- [LAMBERT - 73]: LAMBERT, J. D., *Computational Methods in Ordinary Differential Equations*, Wiley (1973).
- [PETZOLD - 82]: PETZOLD, L., *Differential / algebraic equations are not ODEs*, SIAM J. Sci. Stat. Comput., 3 (1982), 367 - 384.
- [PETZOLD - 86]: PETZOLD, L., *Order results for implicit Runge-Kutta methods applied to differential / algebraic systems*, SIAM J. Numer. Anal., 23 (1986), 836 - 852.
- [PIRES - 94]: PIRES, A.A., *Métodos Numéricos para a Solução de Equações Algébrico-Diferenciais*, Instituto de Ciências Matemáticas de São Carlos, da Universidade de São Paulo, São Carlos, (1994).
- [SHAMPINE - 75]: L.F. SHAMPINE, L.F. & GORDON, M.K., *Computer Solution of Ordinary Differential Equations*, W.H. Freeman and Co., (1975).