**Métodos Numéricos para a
Solução de Equações
Algébrico - Diferenciais**

Augusto de Abreu Pires

orientador: Prof. Dr. Antônio Castelo Filho

Dissertação apresentada ao Instituto de Ciências Matemáticas de São Carlos, da Universidade de São Paulo, para a obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

**São Carlos
1994**

À Deus.

À minha família.

Agradecimentos

Ao professor Dr. Antônio Castelo Filho, pela orientação, apoio e atenção que sempre me dispensou durante a elaboração deste trabalho.

Ao professor Dr. José Alberto Cuminato, pela orientação durante a preparação para os exames gerais de qualificação.

Aos professores, colegas de pós - graduação e funcionários do Instituto de Ciências Matemáticas de São Carlos, da Universidade de São Paulo, que muito me ajudaram.

Aos meus colegas de sala, pela cooperação e amizade.

À Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio econômico prestado.

Finalmente, agradeço aos que de alguma forma contribuíram para a realização deste trabalho, um motivo de grande satisfação pessoal.

Métodos Numéricos para a Solução de Equações Algébrico - Diferenciais

Resumo

Este trabalho diz respeito à solução numérica de equações algébrico - diferenciais de índices zero, um e dois. Uma classe de métodos, conhecida na literatura como métodos de diferenças para trás é considerada e alguns resultados de convergência são dados. Equações algébricas - diferenciais com singularidades também são estudadas. Alguns resultados de existência e unicidade são apresentados e uma nova classe de métodos desenvolvida por [Castelo/Tavares - 94] é introduzida. Finalizaremos este trabalho com alguns exemplos numéricos que ilustram o funcionamento dos métodos considerados.

Numerical Methods for the Solution of Differential - Algebraic Equations

Augusto de Abreu Pires
Adviser: Antônio Castelo Filho

Abstract

This work is concerned with the numerical solution of differential - algebraic equations of indexes zero, one and two. A class of methods, known in the literature as backward difference methods, is considered and some convergence results are given. Differential - algebraic equations with singularities are also studied. Some existence and uniqueness results are presented and a new class of methods developed by [Castelo/Tavares - 94] is introduced. We finish this work with some numerical examples which illustrate the performance of the methods considered.

ÍNDICE

1 - INTRODUÇÃO	01
2 - TEORIA DAS EAD	02
2.1 - Introdução	02
2.2 - Solubilidade e índice	02
2.3 - EAD linear com coeficientes constantes	03
2.4 - EAD linear com coeficientes dependendo do tempo	06
2.4.1 - Solubilidade e índice	06
2.4.2 - Formas estruturais	07
2.4.3 - Teoria geral estrutural	10
2.5 - Sistemas não lineares	12
2.5.1 - Solubilidade e índice	12
2.5.2 - Formas estruturais	13
2.5.3 - Redução de índice e estabilização de restrições	14
3 - MÉTODOS BDF	17
3.1 - Introdução	17
3.2 - Convergência dos BDF	23
3.2.1 - Sistemas semi - explícitos de índice 1	23
3.2.2 - Sistemas de índice 1 implícitos	24
3.2.3 - Sistemas semi - explícitos de índice 2	39
3.3 - Algoritmo	41
4 - EAD COM SINGULARIDADES	43
4.1 - Introdução	43
4.2 - Resultados preliminares	43
4.3 - Método numérico	58
5 - RESULTADOS NUMÉRICOS	64
5.1 - Introdução	64
5.2 - Exemplos numéricos	65
6 - CONCLUSÃO	74
7 - BIBLIOGRAFIA	76

1 INTRODUÇÃO

O tema deste trabalho é Equações Algébrico - Diferenciais (EAD), as quais constituem sistemas que envolvem equações algébricas e equações diferenciais. A escolha deste assunto se deve ao fato de que a solução numérica de EAD tem sido objeto de intensas pesquisas nos últimos anos, conseguindo-se grande progresso na compreensão da estrutura matemática e na análise de métodos numéricos para sistemas de EAD.

EAD surgem frequentemente em problemas de diversas áreas de estudo, dentre as quais destacamos:

- Sistemas mecânicos restritos
- Equações do movimento descrito por pêndulos
- Problemas em robótica
- Problemas de controles ótimos com controles restritos.

Gostaríamos de ressaltar que, para este trabalho, foi escolhido como texto básico o livro *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, cujos autores são:

- K. E. Brennan - The Aerospace Corporation El Segundo
- S. L. Campbell - Department of Mathematics North Carolina State University
- L. R. Petzold - Lawrence Livermore National Laboratory.

No capítulo 2 apresentaremos um resumo da teoria das Equações Algébrico - Diferenciais, sem a qual seria impossível estudarmos métodos numéricos.

Tendo em vista que, na grande maioria dos problemas práticos, as EAD que surgem são de índice 0, 1 ou 2, ainda mais, as de índice 2 quase sempre se apresentam na forma semi - explícita (as definições de índice e forma semi - explícita serão dadas no capítulo 2), no capítulo 3, estudaremos as fórmulas de diferenças para trás (BDF - Backward Differentiation formulas) aplicados na resolução numérica de tais EAD.

No capítulo 4, estudaremos uma nova classe de métodos para solução de EAD de índice 0, 1 e algumas de índice 2, com base na teoria apresentada em [Castelo/Tavares - 94].

No capítulo 5, apresentaremos alguns resultados numéricos obtidos por estas duas classes numéricas, e no capítulo 6, faremos uma breve comparação entre estas classes.

2 TEORIA DAS EAD

2.1 Introdução

Apresentaremos, agora, algumas definições básicas necessárias para a compreensão de resultados posteriores.

Definição 2.1.1 Um Homeomorfismo entre X e Y é uma aplicação $H : X \rightarrow Y$ contínua, inversível com inversa contínua. Se H e H^{-1} são de classe C^k , dizemos que H é um Homeomorfismo de classe C^k entre X e Y . Um Homeomorfismo de classe C^k , com $k \geq 1$ é dito um Difeomorfismo de classe C^k , onde C^j é o espaço das funções j -vezes diferenciáveis com a j -ésima derivada contínua.

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 , $H : V \rightarrow U \cap M$.

Adotaremos para o restante deste trabalho, a seguinte notação, a menos que algo ao contrário seja dito:

$R^{i \times j}$: espaço das matrizes $(i \times j)$ com coeficientes reais,

R^{ij} : espaço real de dimensão ij , onde, ij denota o produto do número real i pelo número real j ($i \cdot j$).

As demonstrações dos resultados apresentados até o final deste capítulo podem ser encontradas em [Brenan/Campbell/Petzold - 89] e em [Pires - 93].

2.2 Solubilidade e índice

Uma função $y(t)$ é uma solução (clássica) da EAD não linear geral

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

onde $t \in I \subset R$, $y \in R^m$ e $F: R^{2m+1} \rightarrow R^m$, 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 R , Ω um subconjunto conexo aberto de R^{2m+1} e F uma função diferenciável de Ω em 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 $I \times \tilde{\Omega}$, $\tilde{\Omega} \subset R^r$ tal que:

- i) $\phi(t,c)$ está definida em todo 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 O número mínimo de vezes que a equação (2.2.1), ou parte dela, será diferenciada com relação a t para determinar y' como uma função contínua de y e t (ou seja, transformar a EAD em uma EDO) é o índice v da EAD (2.2.1). Este índice também é chamado de índice global.

2.3 EAD linear com coeficientes constantes

Consideremos a EAD linear com coeficientes constantes

$$Ax' + Bx = f \quad (2.3.1)$$

onde A e B são matrizes $m \times m$.

Definição 2.3.1 Seja λ um parâmetro complexo. Então $(\lambda A + B)$ é a matriz pencil.

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

Quando o pencil é um pencil regular dizemos que a EAD (2.3.1) é uma EAD regular.

Definição 2.3.3 Uma matriz N tem índice de nilpotência, ou simplesmente 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*

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

onde N é uma matriz diagonal por blocos, isto é,

$$N = \begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_k \end{bmatrix}$$

e cada N_i é da forma

$$N_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

com dimensão m_i e C é a forma canônica de Jordan da matriz $(J_1^{-1}(I - cJ_1))$ onde J_1 é o bloco de Jordan da matriz $((cA + B)^{-1} \cdot A)$ que contém todos os auto valores não nulos e $c \neq 1$ é tal que $(cA + B)$ é inversível. Além disso, considerando o sistema

$$\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} x' + \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} x = f$$

se $N=0$, então $v=1$. No caso em que A é não singular, nós temos $PAQ=I$, $PBQ=C$ e $v=0$. Se $\det(\lambda A + B)$ é identicamente constante então (2.3.2) simplifica-se para $PAQ=N$, $PBQ=I$.

Observe que a matriz N do teorema anterior tem índice de nilpotência α onde $\alpha = \max_{1 \leq j \leq k} \{\dim N_j\}$.

Observe, ainda, que o índice de nilpotência de N é exatamente o índice da EAD, como descrito anteriormente.

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

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

logo, pelo teorema 2.3.1, obtemos, com $Pf = [f_1, f_2]^t$,

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

ou seja,

$$\begin{cases} y_1' + Cy_1 = f_1 \\ Ny_2' + y_2 = f_2 \end{cases} \quad (2.3.4a)$$

$$(2.3.4b)$$

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

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

e assim

$$y_2 = (ND + I)^{-1} f_2$$

onde $D = \frac{d}{dt}$.

Como

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

onde α é o grau de nilpotência de N, temos

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

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

Do resultado acima, vemos que:

i) a solução de (2.3.1) envolve derivadas de ordem $(\alpha - 1)$ da função f, se a EAD é de índice alto ($v \geq 2$);

ii) nem todas as condições iniciais de (2.3.1) fornecem uma solução suave se $\alpha \geq 1$. Estas que conduzem a uma solução suave são chamadas condições iniciais consistentes. Disto definimos que uma condição inicial é dita ser consistente quando esta satisfaz (2.2.1) e admite uma solução suave, ou então, dizemos simplesmente que a EAD é consistente;

iii) EAD com índice alto pode ter restrições algébricas implícitas.

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

Teorema 2.3.3 *Suponha que $Ax' + Bx = f$ é uma EAD solúvel com coeficientes constantes, com índice $\nu \geq 1$. Então cada iteração do processo para explicitar restrições e então diferenciação destas restrições reduz o índice de 1. Após ν diferenciações a EAD estará reduzida a uma EDO.*

Observação: Após realizarmos o processo de diferenciação para reduzir a EAD a uma EDO, em muitos casos, nem todas soluções da EDO serão soluções da EAD.

2.4 EAD linear com coeficientes dependendo do tempo

Seja a EAD

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

onde A, B são $2m$ -vezes diferenciáveis, f é m -vezes diferenciável, $t \in I$ e I é um subintervalo de $\mathbb{R}$.

A equação (2.4.1) é a forma geral de uma EAD linear com coeficientes dependendo do tempo.

2.4.1 Solubilidade e índice

Vamos relembrar as definições de solubilidade e índice, especificamente voltadas para (2.4.1).

Definição 2.4.1 *O sistema (2.4.1) com A, B matrizes $m \times m$ é solúvel no intervalo I se para toda f m -vezes diferenciável, existe pelo menos uma solução para (2.4.1)*

continuamente diferenciável. E além disso, as soluções estão definidas em todo I e são determinadas unicamente por seus valores iniciais para qualquer $t \in I$.

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 unicamente x' como uma função de x, t .*

Definição 2.4.3 *Se (2.4.1) é uma EAD regular, então o índice local para t , denotado por $\nu_1(t)$, é o índice do pencil $(\lambda A(t) + B(t))$, que é a nilpotência da matriz N , apresentada no teorema 2.3.1.*

No caso de EAD lineares com coeficientes constantes, definimos índice como o número de iterações do processo abaixo:

- i) usando mudança de coordenadas reformulamos a EAD para termos restrições algébricas explícitas;
- ii) diferenciamos tais restrições algébricas até reduzirmos o sistema em uma EDO explícita.

Para sistemas com coeficientes dependendo do tempo este procedimento requer que a matriz $A(t)$ tenha posto constante, uma condição que, em geral, não é verdadeira.

2.4.2 Formas estruturais

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

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

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

Como P e Q são não singulares, (2.4.1) e (2.4.2) são analiticamente equivalentes.

Teorema 2.4.1 *Suponha que (2.4.1) é uma EAD solúvel no intervalo I . Então as seguintes afirmações são equivalentes:*

- i) (2.4.1) é uma EDO implícita;
- ii) $\nu = 0$;
- iii) $\nu_1 \equiv 0$;
- iv) $\nu_1(t_0) = 0$ para algum $t_0 \in I$;

onde v é o índice e v_i é o índice local de (2.4.1).

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

Os dois resultados acima mostram que $v = v_i$ para sistemas solúveis de índice zero e um. Isto não ocorre para sistemas com índices maiores ou iguais a dois.

Teorema 2.4.3 *Assuma que (2.4.1) é solúvel no intervalo I . As seguintes afirmações são preservadas quando transformamos (2.4.1) para o sistema analiticamente equivalente (2.4.2):*

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

Portanto, podemos considerar 3 classes de EAD solúveis:

EDO implícitas,
Índice 1,
Índice ≥ 2 .

Definição 2.4.4 *O sistema (2.4.1) é semi-implícito se está na forma*

$$\begin{cases} \dot{x}_1 + B_{11}(t)x_1 + B_{12}(t)x_2 = f_1 \\ B_{21}(t)x_1 + B_{22}(t)x_2 = f_2 \end{cases} \quad (2.4.3)$$

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

$$\begin{bmatrix} I & 0 & . & . & 0 \\ 0 & I & . & . & . \\ . & . & I & . & . \\ . & . & . & I & 0 \\ 0 & . & . & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ . \\ . \\ . \\ \dot{x}_r \end{bmatrix} + \begin{bmatrix} B_{11} & . & . & B_{1,r-1} & B_{1r} \\ B_{21} & . & . & B_{2,r-2} & 0 \\ 0 & . & . & . & . \\ . & . & . & . & . \\ 0 & . & 0 & B_{r,r-1} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_r \end{bmatrix} = \begin{bmatrix} f_1 \\ . \\ . \\ . \\ f_r \end{bmatrix}$$

onde x_i são vetores, B_{ij} são matrizes e o produto $B_{r,r-1}B_{r-1,r-2}\dots B_{21}B_{1r}$ é não singular.

Observação: Toda EAD na forma de Hessenberg de dimensão r é solúvel e tem índice r .

Definição 2.4.6 *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 denso aberto de valores destas variáveis.*

Ao contrário do caso linear com coeficientes constantes a estrutura do pencil $(\lambda A + B)$ não está diretamente ligada com as soluções da EAD linear com coeficientes dependendo do tempo. Na seção 2.3 vimos que se N é uma matriz constante e $N^\alpha = 0$ e $N^{\alpha-1} \neq 0$ (ou seja, N tem índice de nilpotência α), então $Nx' + x = f$ é solúvel e tem apenas uma solução que envolve $(\alpha - 1)$ derivadas de f . Agora, no caso de sistemas lineares com coeficientes dependendo do tempo podemos ter mais de uma solução devido ao fato que $(N(t))^\alpha = 0$ não necessariamente implica $\left(N \frac{d}{dt}\right)^\alpha = 0$.

Agora se $N(t)$ é estruturalmente nilpotente (no sentido da definição 2.4.6), então se $[N(t)]^\alpha = 0$ teremos $\left(N \frac{d}{dt}\right)^\alpha = 0$ e assim

$$N(t)x' + x = f,$$

que pode ser escrita como

$$(N(t)D + I)x = f$$

onde $D = \frac{d}{dt}$, tem como solução

$$x = (N(t)D + I)^{-1} f = \sum_{i=0}^{\infty} (-1)^i (N(t)D)^i f = \sum_{i=0}^{\alpha-1} (-1)^i (N(t)D)^i f.$$

Observemos que x envolve derivadas de $N(t)$ até a ordem $(\alpha - 2)$, além das derivadas de f , até a ordem $(\alpha - 1)$.

Definição 2.4.7 *Dizemos que o sistema (2.4.1) está na forma canônica padrão se está na forma*

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

onde N é estritamente triangular inferior (ou superior).

Observe que a definição anterior não exige que a matriz N tenha posto ou índice constante.

Teorema 2.4.4 *Suponha que (2.4.1) é solúvel no intervalo I , então (2.4.1) é analiticamente equivalente a*

$$\begin{cases} Z_1' + GZ_2 = g & (2.4.5a) \\ NZ_2' + Z_2 = h & (2.4.5b) \end{cases}$$

onde (2.4.5a) pode estar ausente e $NZ_2' + Z_2 = h$ tem solução única para cada $N(t)$ suficientemente suave. Além disto existe uma família contável de intervalos abertos disjuntos I_i , tal que $\cup I_i$ é denso em I e o sistema $NZ_2' + Z_2 = h$ pode ser transformado por P, Q como em (2.4.2) tal que (2.4.5b) estará na forma canônica padrão.

Definição 2.4.8 *A EAD (2.4.1) é uma cadeia triangular (inferior) de r -ésima ordem se tem a forma*

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

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

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

2.4.3 Teoria geral estrutural

Suponha que x é solução de (2.4.1), para o restante desta seção usaremos a notação $c_i(t)$ para designar $\frac{c^{(i)}(t)}{i!}$ para $i \geq 0$, onde c pode ser uma variável, uma matriz, ou uma função. Então para cada t e para cada j menor que o grau de suavidade de x , (2.4.1) torna-se

$$\sum_j A_j \delta^j \sum_j x_{j+1} \delta^j (j+1) = - \sum_j B_j \delta^j \sum_j x_j \delta^j + \sum_j f_j \delta^j$$

onde $\delta = t - t_0$. Matricialmente temos

$$\begin{bmatrix} A_0 & 0 & \cdot & 0 \\ A_1 + B_0 & 2A_0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 \\ A_{j-1} + B_{j-2} & 2A_{j-2} + B_{j-3} & \cdot & jA_0 \end{bmatrix} \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_j \end{bmatrix} = - \begin{bmatrix} B_0 \\ \cdot \\ \cdot \\ B_{j-1} \end{bmatrix} x_0 + \begin{bmatrix} f_0 \\ \cdot \\ \cdot \\ f_{j-1} \end{bmatrix}$$

ou, em notação simplificada

$$A_j x_j = -B_j x_0 + f_j. \quad (2.4.6)$$

Observemos que se A_j é singular para todo t , então (2.4.6) não determina unicamente x_j dado x_0 e t .

Definição 2.4.9 Quando A, B são $m \times m$, a matriz A_j é suavemente 1-full se existe uma matriz não singular suave $R(t)$ tal que

$$R(t)A_j(t) = \begin{bmatrix} I & 0 \\ 0 & H(t) \end{bmatrix}$$

onde a identidade I é $m \times m$.

Dado t e $x_0 = x(t)$, o sistema de equações algébricas (2.4.6) determinará unicamente $x'(t)$ quando A_j for 1-full.

Teorema 2.4.5 Suponha que A, B são 2m-vezes diferenciáveis, f é m-vezes diferenciável e (2.4.1) é solúvel em I . Então

- i) A_j tem posto constante em I para $j = m+1$;
- ii) A_j é suavemente 1-full em I para $j = m+1$;
- iii) $E(A_j) + E(B_j) = R^{mj}$ para todo $t \in I$, $1 \leq j \leq m+1$, onde $E(X)$ denota o espaço coluna da matriz X .

Observação: Supondo A, B 3m-vezes continuamente diferenciáveis, [Campbell - 87] mostra que (i)-(iii) do teorema 2.4.5 implicam que a EAD (2.4.1) é solúvel.

Proposição 2.4.1 O índice v é o menor inteiro tal que A_{v+1} é 1-full e tem posto constante.

Lema 2.4.1 *O índice v é invariante sob as transformações P e Q já conhecidas.*

2.5 Sistemas não lineares

2.5.1 Solubilidade e índice

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

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

que podemos escrever como:

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

ou ainda podemos denotar o sistema (2.5.1) por:

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

onde

$$y_j = \begin{bmatrix} y' \\ \vdots \\ y^{(j+1)} \end{bmatrix}. \quad (2.5.4)$$

O sistema não linear (2.5.3) é análogo ao sistema $A_j x_j = -B_j x_0 + f_j$ da última seção.

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

Proposição 2.5.1 Condições suficientes para (2.5.3) determinar unicamente y' como uma função contínua de y, t são que a matriz Jacobiana de F_j com respeito a y_j , denotada $\frac{\partial F_j}{\partial y_j}$, seja 1-full com posto constante e (2.5.3) seja consistente.

No caso de EAD lineares com coeficientes dependendo do tempo, esta matriz Jacobiana é precisamente a matriz A_j definida em (2.4.6).

Analogamente ao caso das EAD lineares com coeficientes dependendo do tempo pode-se mostrar que as propriedades de 1-full e posto constante da matriz Jacobiana são preservadas por mudanças de coordenadas não lineares, multiplicando-se (2.2.1) por $P(t, y)$ e fazendo $y=Q(x, t)$ com Q_x não singular.

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

Proposição 2.5.2 Suponha que F_y tem posto constante. Então $v = 1$ se e somente se $v_i = 1$.

2.5.2 Formas estruturais

Definição 2.5.3 A EAD (2.2.1) está na forma de Hessenberg de dimensão r se pode ser escrita como

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

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

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

$$\begin{cases} \dot{x}_1 = F_1(x_1, t) \\ N(x_1, x_2, t) \dot{x}_2 = x_2 + h(x_1, t) \end{cases}$$

onde N é uma matriz bloco rxr estritamente triangular inferior.

Definição 2.5.5 A EAD (2.2.1) é uma cadeia triangular inferior de ordem r se pode ser escrita como

$$\begin{cases} F_1(x'_1, x_1, t) = 0 \\ F_2(x'_1, x'_2, x_1, x_1, t) = 0 \\ \vdots \\ F_r(x'_1, \dots, x'_r, x_1, \dots, x_r, t) = 0 \end{cases}$$

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

Observação: Uma cadeia triangular é sempre solúvel.

2.5.3 Redução de índice e estabilização de restrições

Numericamente é importante que a EAD tenha índice tão pequeno quanto possível. A redução de índice pode ser feita por diferenciações das restrições, porém a solução numérica do sistema resultante nem sempre satisfaz, exatamente, as restrições originais. Então, é importante termos noção das restrições que tenham sido perdidas, por diferenciações, ou outras manipulações.

Discutiremos, agora, uma teoria devido a [Gear - 88], cuja aplicação é de grande importância.

Suponha que após uma série de diferenciações e mudanças de coordenadas aplicadas a EAD (2.2.1), obtemos a EDO

$$y' = f(t, y) \tag{2.5.5a}$$

e que havia a restrição

$$0 = g(t, y). \tag{2.5.5b}$$

A solução analítica de (2.5.5a) satisfaz a restrição (2.5.5b), ao passo que a solução numérica pode não satisfazer.

Para tentarmos corrigir este problema suponhamos que g_y tem posto linha completo, ou seja, as condições dadas por (2.5.5b) são linearmente independentes. Seja μ uma nova variável com a dimensão do posto de (g_y) e consideremos a EAD semi-explicita

$$\begin{cases} y' = f(t, y) + g_y^T \mu & (2.5.6a) \\ 0 = g(t, y) & (2.5.6b) \end{cases}$$

Teorema 2.5.1 *Assuma que g_y tem posto linha completo e que (2.5.5b) é uma caracterização da variedade de solução para (2.2.1). Mais precisamente, existe uma solução para (2.2.1) satisfazendo $y(\hat{t}) = \hat{y}$ se e somente se $g(\hat{y}, \hat{t}) = 0$ para todo $\hat{t}$ fixo. Então, as únicas soluções de (2.5.6) são, $\mu = 0$ e uma solução y de (2.5.5).*

Teorema 2.5.2 *Assuma que g_y tem posto linha completo. Então a EAD semi-explicita (2.5.6) tem índice 2.*

Um outro procedimento que às vezes, pode reduzir o índice de uma EAD é introduzir uma variável, como mostra o exemplo abaixo.

Exemplo 2.5.1 Consideremos a EAD semi-explicita com coeficientes constantes de índice 3,

$$\begin{cases} \dot{x}_2 = x_1 + f_1(t) \\ \dot{x}_3 = x_2 + f_2(t) \\ 0 = x_3 + f_3(t) \end{cases} \quad (2.5.7)$$

Introduzimos a variável $z_1' = x_1$. Assim (2.5.7) torna-se

$$\begin{cases} \dot{x}_2 - z_1' = f_1(t) \\ \dot{x}_3 = x_2 + f_2(t) \\ 0 = x_3 + f_3(t) \end{cases}$$

que é uma EAD de índice 2 em (z_1, x_2, x_3) .

Esta opção de redução é utilizável toda vez que tenhamos uma EAD com variáveis, tais como x , neste exemplo, cujas derivadas não aparecem explicitamente. Em sistemas mais complicados, isto é, com mais variáveis, a substituição pode não reduzir o índice.

Proposição 2.5.3 *Suponha que (2.2.1) é uma EAD solúvel de índice ν . Então*

$$\begin{cases} y' = z \\ 0 = F(t, y, z) \end{cases}$$

é uma EAD semi-explicita de índice $\nu + 1$.

3 MÉTODOS BDF

3.1 Introdução

Inicialmente mostraremos porque as fórmulas de diferenças para trás (métodos BDF) aparecem como a mais popular classe de métodos para a solução de EAD gerais.

A primeira técnica geral bem sucedida para a solução numérica de EAD, foi proposta por [Gear - 71]. Esta técnica, que utilizava os métodos BDF, foi inicialmente aplicada a sistemas de equações diferenciais acopladas a equações algébricas:

$$\begin{cases} x' = f(x, y, t) \\ 0 = g(x, y, t) \end{cases} \quad \begin{matrix} (3.1.1a) \\ (3.1.1b) \end{matrix}$$

onde y é um vetor da mesma dimensão que g .

Os métodos BDF tratam as variáveis algébricas e diferenciais do problema (3.1.1) de maneira análoga, logo a idéia básica desses métodos pode ser generalizada para a solução do problema completamente implícito (2.2.1).

O método BDF de primeira ordem mais simples é o método de Euler implícito, que consiste em substituir a derivada em (2.2.1) por uma diferença para trás, ou seja, (2.2.1) é transformada em:

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

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

O sistema de equações não lineares para y_n em cada passo é, então, 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 $t_n, t_{n-1}, \dots, t_{n-k}$ avaliado para t_n . Isto nos dá:

$$F\left(t_n, y_n, \frac{\rho y_n}{h}\right) = 0 \quad (3.1.2)$$

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, que podem ser encontrados em [Lambert - 73], na página 242.

Lembramos que o método BDF é uma classe especial dos métodos lineares de passo múltiplo, pois tais métodos, quando aplicados ao problema de valor inicial

$$\begin{cases} y' = f(x, y) \\ y(a) = \eta \end{cases}$$

são representados, em sua forma geral, pela expressão:

$$\sum_{j=0}^k \alpha_j y_{n-j} = h \sum_{j=0}^k \beta_j f_{n-j}, \quad \alpha_0 = 1, \quad (3.1.3)$$

e tomando $\beta_1 = \dots = \beta_k = 0$ temos a expressão

$$\sum_{j=0}^k \alpha_j y_{n-j} = h \beta_0 f_n$$

que é o método BDF geral, para o problema de valor inicial acima.

Ainda mais, quando $k=1$, temos:

$$\alpha_0 y_n + \alpha_1 y_{n-1} = h \beta_0 f_n.$$

De [Lambert - 73], temos que se $k=1$ então $\alpha_0 = -\alpha_1 = \beta_0 = 1$, assim:

$$y_n - y_{n-1} = h f_n$$

ou seja,

$$y'_n = f_n = \frac{y_n - y_{n-1}}{h}$$

que é o método de Euler implícito.

Observação: Da teoria de EDO, sabemos que um método BDF de k -passos é estável para EDO se $k \leq 6$. Assim consideraremos sempre $k \leq 6$.

Os resultados apresentados a seguir podem ser encontrados em [Brenan/Campbell/Petzold - 89].

Para ilustrar algumas das idéias usualmente empregadas em análises de convergência para sistemas de EAD mais complicados, apresentaremos o seguinte teorema.

Teorema 3.1.1 *O método BDF de k-passos ($k \leq 6$) com tamanho de passo constante h aplicado a sistemas de EAD linear com coeficientes constantes de índice v é convergente de ordem $O(h^k)$ depois de $(v-1)k+1$ passos.*

Demonstração:

Consideremos o sistema linear com coeficientes constantes (2.3.1). Discretizando (2.3.1) com o método BDF de k-passos obtemos:

$$\frac{1}{h} \sum_{j=0}^k \alpha_j A x_{n-j} + B x_n = f_n \quad (3.1.4)$$

onde $f_n = f(t_n)$.

Seja

$$\rho x_n = \sum_{j=0}^k \alpha_j x_{n-j} \quad (3.1.5)$$

então (3.1.4) pode ser escrita simplesmente como:

$$A \left(\frac{\rho x_n}{h} \right) + B x_n = f_n. \quad (3.1.6)$$

Do teorema 2.3.1 sabemos que se a matriz pencil $(\lambda A + B)$ é regular, existem matrizes não singulares P, Q tais que:

$$PAQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \quad \text{e} \quad PBQ = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}$$

onde N é uma matriz com índice de nilpotência v e I é a matriz identidade.

Seja $x_n = Q y_n$, e multiplicando o sistema discretizado (3.1.6) pela matriz não singular P , temos:

$$PAQ \left(\frac{\rho y_n}{h} \right) + PBQ y_n = P f_n \quad (3.1.7)$$

onde usamos o fato que $Q\rho = \rho Q$.

Fazendo $P f_n = g_n$, o sistema (3.1.7) pode ser escrito como dois subsistemas acoplados:

$$\begin{cases} \frac{\rho y_{1,n}}{h} + Cy_{1,n} = g_{1,n} & (3.1.8a) \\ N\left(\frac{\rho y_{2,n}}{h}\right) + y_{2,n} = g_{2,n} & (3.1.8b) \end{cases}$$

A equação (3.1.8a) é justamente a equação diferença obtida aplicando o método BDF à EDO $y_1' + Cy_1 = g_1(t)$ e sabemos que a solução numérica de uma EDO explícita obtida pelo BDF de k -passos é exata $O(h^k)$. A equação (3.1.8b) representa o método BDF aplicado ao sistema $Ny_2' + y_2 = g_2(t)$. Claramente nos falta analisar a convergência da solução numérica para o subsistema (3.1.8b). Para facilitar a notação, faremos $y_2 = y$ e assim (3.1.8b) se torna:

$$N\left(\frac{\rho y_n}{h}\right) + y_n = g_n \quad (3.1.9)$$

ou seja:

$$\left(N\frac{\rho}{h} + I\right)y_n = g_n. \quad (3.1.10)$$

No capítulo 2, derivamos a solução analítica para (3.1.10), que é:

$$y(t) = \sum_{i=0}^{v-1} (-1)^i N^i g^{(i)}(t) \quad (3.1.11)$$

onde $g^{(i)} = \frac{\partial^i g}{\partial t^i}$.

Resolvendo (3.1.10) para y_n , e notando que

$$\left(N\left(\frac{\rho}{h}\right) + I\right)^{-1} = \sum_{i=0}^{v-1} (-1)^i N^i \left(\frac{\rho}{h}\right)^i$$

obtemos:

$$y_n = \sum_{i=0}^{v-1} (-1)^i N^i \left(\frac{\rho}{h}\right)^i g_n.$$

Desde que para os coeficientes do BDF

$$\left(\frac{\rho}{h}\right)^i g_n = g_n^{(i)} + O(h^k)$$

segue que:

$$y_n = \sum_{i=0}^{v-1} (-1)^i N^i (g_n^{(i)} + O(h^k)). \quad (3.1.12)$$

Note que a solução numérica y_n não depende do valor inicial y_0 , e que de fato, depende somente dos valores de g nos últimos $[(v-1)k + 1]$ pontos. Comparando a solução numérica y_n com a solução analítica $y(t_n)$ dada em (3.1.11), vemos que o erro local e o erro global, cujas definições são como em EDO, são $O(h^k)$, concluindo, assim, a demonstração do teorema 3.1.1.

Em contraste, para métodos BDF aplicados a uma EDO explícita, o erro local é de uma ordem superior ao erro global.

Agora, nem todas EAD são resolvidas com sucesso pelos métodos BDF. Para ilustrar um problema que pode ocorrer com a aplicação dos métodos BDF a sistemas de índice alto, ou seja, maior que 1, analisaremos um exemplo simples.

Consideremos o sistema de índice 3:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_3 \\ 0 = x_1 - g(t) \end{cases}$$

A solução exata da EAD é claramente:

$$\begin{cases} x_1(t) = g(t) \\ x_2(t) = g'(t) \\ x_3(t) = g''(t) \end{cases}$$

Pelo método de Euler implícito obtemos:

$$\begin{cases} x_{1,n} = g(t_n) \\ x_{2,n} = \frac{x_{1,n} - x_{1,n-1}}{h} \\ x_{3,n} = \frac{x_{2,n} - x_{2,n-1}}{h} \end{cases}$$

Ignorando os erros de arredondamento, o valor de x_1 será determinado exatamente em todos os passos, mesmo se o valor inicial $x_{1,0}$ está errado. Agora suponha que os valores iniciais para t_0 não satisfazem o sistema original. Então os valores $x_{2,1}$ e $x_{3,1}$ obtidos

no final do primeiro passo estarão incorretos. Depois de dois passos, $x_{2,2}$ será exato $O(h)$, visto que ele é determinado pela primeira diferença dividida de $g(t)$. De qualquer maneira, $x_{3,2}$ ainda está incorreto. No terceiro passo, o valor para $x_{3,3}$ é exato $O(h)$, pois ele é obtido pela segunda diferença dividida de $g(t)$. Portanto, mesmo com valores iniciais incorretos, depois de três passos de tamanho h com o método de Euler implícito, a solução numérica é exata $O(h)$. (Se valores iniciais consistentes são dados, a solução é exata $O(h)$ no final do segundo passo).

Uma segunda dificuldade que aparece para o método BDF, diz respeito à estabilidade, quando aplicados a sistemas de EAD de índice ≥ 2 . Nós vimos que o método BDF com passo constante pode ser usado para EAD com coeficientes constantes. Agora, o que ocorre quando eles são aplicados à EAD com coeficientes variando com o tempo? De [Gear/Petzold - 83] vemos que se o índice local de nilpotência da matriz N apresentada no teorema 2.3.1, que é o índice local da EAD, é dois, teremos um problema de estabilidade dependendo dos coeficientes.

Para exemplificar o fato acima, veremos um exemplo:

Consideremos o sistema:

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix} z' + \begin{bmatrix} 1 & \eta t \\ 0 & 1 + \eta \end{bmatrix} z = \begin{bmatrix} g(t) \\ 0 \end{bmatrix} \quad (3.1.13)$$

o qual foi obtido do sistema

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} y' + y = \begin{bmatrix} g(t) \\ 0 \end{bmatrix}$$

pela mudança de variável $y = Hz$ onde $H = \begin{bmatrix} 1 & \eta t \\ 0 & 1 \end{bmatrix}$.

Facilmente vemos que o índice da EAD é 2. Se $\eta = -1$, a EAD (3.1.13) não tem uma matriz pencil regular. De qualquer modo é fácil verificar, usando as transformações

$$P = I \quad \text{e} \quad Q = \begin{bmatrix} 1 & -\frac{\eta t}{(1+\eta)} \\ 0 & \frac{1}{(1+\eta)} \end{bmatrix}$$

que seu índice local também é 2.

Se (3.1.13) é resolvida pelo método de Euler implícito, obtemos a solução numérica:

$$\begin{cases} z_{2,n} = \frac{\eta}{1+\eta} z_{2,n-1} - \frac{g_n - g_{n-1}}{h(1+\eta)} \\ z_{1,n} = g_n - \eta t_n z_{2,n} \end{cases}$$

A solução exata é

$$\begin{cases} z_2(t_n) = -g'(t_n) \\ z_1(t_n) = g(t_n) + \eta t_n g'(t_n) \end{cases}$$

e a solução numérica é instável se $\eta < -\frac{1}{2}$.

Relembramos que um método é estável se existir um $h_0 > 0$ para cada equação diferencial tal que uma mudança nos valores iniciais por uma quantidade fixa produz uma mudança limitada na solução numérica para todo $0 < h \leq h_0$.

3.2 Convergência dos métodos BDF

Nesta seção estudaremos alguns resultados sobre a convergência dos métodos BDF, quando aplicados à EAD de índice 1 e índice 2 na forma semi - explícita, uma vez que, da definição de índice, vemos que uma EAD de índice 0 é uma EDO.

3.2.1 Sistemas semi-explícitos de índice 1

Sistemas semi-explícitos de índice 1 constituem uma importante classe dentre os sistemas de índice 1 e aparecem frequentemente em problemas de engenharia. Consideremos o sistema dado por (3.1.1) onde $\left(\frac{\partial g}{\partial y}\right)^{-1}$ existe e é limitada em uma vizinhança da solução exata.

A aplicação dos métodos lineares de passo múltiplo a estes sistemas é tão simples, que descreveremos o caso geral, ao invés de nos limitarmos aos métodos BDF no resto desta seção. Analogamente a (3.1.3), a aplicação destes métodos a EAD (3.1.1) conduz ao seguinte algoritmo:

$$\begin{cases} \sum_{j=0}^k \alpha_j x_{n-j} = h \sum_{j=0}^k \beta_j f(x_{n-j}, y_{n-j}, t_{n-j}) & (3.2.1a) \\ 0 = g(x_n, y_n, t_n) & (3.2.1b) \end{cases}$$

Desta maneira, os resultados de convergência e estabilidade para EDO não stiff são ainda válidos para a EAD semi-explicita de índice 1.

3.2.2 Sistemas de índice 1 implícitos

Nesta subseção desenvolveremos o principal resultado de convergência para métodos BDF aplicados a sistemas de índice 1 implícitos.

Vamos, primeiramente, definir algumas terminologias e provar alguns resultados preliminares.

Consideremos a EAD implícita (2.2.1) sujeita a condição inicial $y(t_0) = y_0$.

Assumiremos que F é uma função suficientemente suave e que existe uma solução suave $y(t)$ satisfazendo a condição inicial dada.

Definição 3.2.1 A EAD não linear (2.2.1) é dita ser de índice 1 uniforme se o índice do sistema com coeficientes constantes

$$Aw'(t) + Bw(t) = g(t) \quad (3.2.2)$$

onde,

$$\begin{cases} A = F_y(\hat{t}, \hat{y}, \hat{y}') \\ B = F_y(\hat{t}, \hat{y}, \hat{y}') \end{cases} \quad (3.2.3)$$

é 1 para todo $(\hat{t}, \hat{y}, \hat{y}')$ em uma vizinhança do gráfico da solução, e se:

- 1- As derivadas parciais de A com respeito a t, y, y' , existem e são limitadas em uma vizinhança da solução.
- 2- O posto de A é constante em uma vizinhança da solução.

Notamos ainda, que sem perda de generalidade, F pode ser escrita como:

$$F(t, y, y') = \begin{bmatrix} F_1(t, y, y') \\ G(t, y) \end{bmatrix}. \quad (3.2.4)$$

Com efeito, suponha que $\text{posto}(F_y) = r < m$, onde m é a dimensão de y . Então, existe uma submatriz de F_y de dimensão $r \times r$ não singular. Suponha que as equações foram

reenumeradas, tal que posto $\left(\frac{\partial F_1}{\partial y'}\right) = r$ e F_1 é composta pelas r primeiras equações de F . Seja F_2 as últimas $(m-r)$ equações de F . Suponha, também, que as variáveis são reenumeradas tal que:

$$\frac{\partial F_1}{\partial y'} = \begin{bmatrix} \frac{\partial F_1}{\partial y'_1} & \frac{\partial F_1}{\partial y'_2} \end{bmatrix}$$

onde $y = [y_1^T \ y_2^T]^T$, e $\frac{\partial F_1}{\partial y'_1}$ é não singular. Então, pelo Teorema da Função Implícita, F_1 pode ser resolvida para y'_1 em termos de y_1, y_2, y'_2 e substituindo a expressão para y'_1 nas últimas $(m-r)$ equações obtemos uma relação implícita entre y_1 e y_2 . A observação chave é que y'_2 não pode ser envolvida nesta relação, ou seríamos capazes de resolver F para y' , contrariando a hipótese sobre o posto de F_y . Logo F pode ser escrita na forma (3.2.4).

As transformações efetuadas são válidas localmente. De qualquer forma, se F tem índice 1 uniforme, então, podemos cobrir o intervalo original com um conjunto finito de subintervalos, onde F tem esta forma em uma vizinhança da solução. Sobre qualquer subintervalo, a solução pelo BDF não é alterada quando reescrevemos F na forma (3.2.4).

Lema 3.2.1 Para F de índice 1 uniforme na forma (3.2.4), onde y é um vetor m dimensional, existem matrizes não singulares $P(t, y(t), y'(t))$ e $Q(t, y(t), y'(t))$ tais que:

$$PAQ = \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix} \quad (3.2.5a)$$

$$PBQ = \begin{bmatrix} C(t, y(t), y'(t)) & 0 \\ 0 & I_{m_2} \end{bmatrix} \quad (3.2.5b)$$

onde

$$\begin{cases} A(t, y(t), y'(t)) = F_y(t, y(t), y'(t)) \\ B(t, y(t), y'(t)) = F_y(t, y(t), y'(t))' \end{cases}$$

as quais satisfazem:

1- $Q(t, y(t), y'(t))$ e $Q^{-1}(t, y(t), y'(t))$ existem e são limitadas para todo $(t, y(t), y'(t))$ solução de (3.2.3),

$$2- Q^{-1}(t_1, y(t_1), y'(t_1))Q(t_2, y(t_2), y'(t_2)) = I_{m_2} + O(t_2 - t_1),$$

$$3- C(t_1, y(t_1), y'(t_1)) = C(t_2, y(t_2), y'(t_2)) + O(t_2 - t_1),$$

onde $m = m_1 + m_2$.

Demonstração:

Para F na forma (3.2.4), temos que:

$$A = \frac{\partial F}{\partial y'} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0 \end{bmatrix}$$

$$B = \frac{\partial F}{\partial y} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

Provemos que P e Q da forma

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} I_{m_1} & -(B_{11}Q_{12} + B_{12}Q_{22}) \\ 0 & I_{m_2} \end{bmatrix} \quad (3.2.6a)$$

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ B_{21} & B_{22} \end{bmatrix}^{-1} \quad (3.2.6b)$$

satisfazem as condições do lema. Começemos por notar que a matriz

$$\begin{bmatrix} A_{11} & A_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

é sempre não singular para sistemas de índice 1, fato que segue diretamente da definição de índice dada na seção 2.2.

Por outro lado, vemos que P , Q sempre transformam as matrizes A e B para a forma (3.2.5).

Finalmente as exigências de suavidade de Q , Q^{-1} e C seguem da forma de P e Q , juntamente com a suposição de suavidade de F , concluindo, assim, a demonstração do lema 3.2.1.

Lema 3.2.2 *Seja K definida por*

$$K = \begin{bmatrix} \hat{K}_{11} & \hat{K}_{12} \\ 0 & N \end{bmatrix}$$

onde $\hat{K}_{11} = K_{11} + O(h)$, $\hat{K}_{12} = O(h)$, N é nilpotente de ordem k , e

$$K_{11} = \begin{bmatrix} \gamma_1 I_{m_1} & \gamma_2 I_{m_1} & \cdots & \cdots & \gamma_k I_{m_1} \\ I_{m_1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I_{m_1} & 0 \end{bmatrix}$$

$$N = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ I_{m_2} & \ddots & & & \vdots \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I_{m_2} & 0 \end{bmatrix}$$

e γ_i são dados por $\gamma_i = \frac{-\alpha_i}{\alpha_0}$, onde α_i são os coeficientes do método BDF de k -passos, com $k < 7$.

Seja $2k \leq i \leq n$. Então

$$K^i = \begin{bmatrix} O(1) & O(h) \\ 0 & 0 \end{bmatrix}$$

onde $O(h^\delta)$, $\delta = 0, 1$ denota a matriz cujos elementos são limitados por uma constante (que independe de n) vezes h^δ , e $n \leq \frac{1}{h}$.

Demonstração:

Seja $i = qk + r$, onde q e r são inteiros e $r < k$.

Pelas hipóteses, temos que $q \geq 2$, pois se $k = 1$ então $r = 0$.

Então:

$$K^i = K^{kq} K^r.$$

Agora note que, como $N^k = 0$, então

$$K^k = \begin{bmatrix} \hat{K}_{11}^k & \hat{K}_{12}^k \\ 0 & 0 \end{bmatrix}$$

onde $\hat{K}_{12}^k = O(h)$. Temos ainda que

$$K^{kq} = K^k K^{k(q-2)} K^k.$$

Por outro lado

$$K^{k(q-2)} = \begin{bmatrix} \hat{K}_{11}^{k(q-2)} & \hat{K}_{11}^{k(q-3)} \hat{K}_{12}^k \\ 0 & 0 \end{bmatrix}.$$

Da teoria de EDO, que pode ser encontrada em [Hairer/Norsett/Wanner - 80], temos que $\hat{K}_{11}^k$ é a matriz de estabilidade dos métodos BDF aplicados à EDO e assim temos também que:

$$\hat{K}_{11}^{k(q-2)} \text{ é limitado para } k < 7 \text{ e para } n \leq \frac{1}{h}.$$

Relembrando que $\hat{K}_{12}^k = O(h)$ e multiplicando o lado direito por K^t , temos o resultado desejado, concluindo, assim, a demonstração do lema 3.2.2.

Faremos agora o principal resultado desta seção, que é o teorema de convergência para métodos BDF de k-passos ($k < 7$) quando aplicados à EAD não lineares de índice 1 implícitas.

Teorema 3.2.1 *Seja (2.2.1) uma EAD de índice 1 uniforme em um intervalo $I = [t_0, t_0 + T]$. Suponhamos que os erros nos valores iniciais são $O(h^k)$ e os erros da iteração de Newton em cada passo são $O(h^{k+1})$. Então a solução numérica de (2.2.1) por um método BDF de k-passos com tamanho h , fixo, para $k < 7$, converge com $O(h^k)$ (depois de $k+1$ passos).*

Demonstração:

Seja y_n satisfazendo

$$F\left(t_n, y_n, \frac{1}{h} \sum_{i=0}^k \alpha_i y_{n-i}\right) = 0$$

onde os α_i são os coeficientes do BDF e $t_n = nh$. Sem perda de generalidade, assumamos que F está na forma (3.2.4).

Seja $e_n = y_n - y(t_n)$ e $\rho y_n = \sum_{i=0}^k \alpha_i y_{n-i}$ para que

$$\frac{\rho y(t_n)}{h} - y'(t_n) = \frac{T_n}{h}$$

onde $T_n = -\frac{1}{(k+1)!} h^{k+1} y^{(k+1)}(t_n) + O(h^{k+2})$ é o erro de truncamento local.

Então, como $y'_n = \frac{\rho y_n}{h}$ temos que:

$$\begin{aligned} 0 &= F\left(t_n, y_n, \frac{\rho y_n}{h}\right) \\ &= F\left(t_n, y(t_n) + e_n, \frac{\rho(y(t_n) + e_n)}{h}\right) \\ &= F\left(t_n, y(t_n) + e_n, \frac{\rho y(t_n)}{h} + \frac{\rho}{h} e_n\right) \\ &= F\left(t_n, y(t_n) + e_n, y'(t_n) + \left(\frac{T_n}{h} + \frac{\rho}{h} e_n\right)\right). \end{aligned}$$

Assim, usando desenvolvimento por série de Taylor, obtemos:

$$0 = F(t_n, y(t_n), y'(t_n)) + \frac{\partial F}{\partial y} e_n + \frac{\partial F}{\partial y'} \left(\frac{T_n}{h}\right) + \frac{\partial F}{\partial y'} \left(\frac{\rho e_n}{h}\right) + \eta_n \quad (3.2.7)$$

onde $\frac{\partial F}{\partial y}$ e $\frac{\partial F}{\partial y'}$ são avaliadas para $(t_n, y(t_n), y'(t_n))$, e η_n é composto de resíduos da iteração de Newton e termos de ordem mais alta.

Sejam $A_n = \frac{\partial F}{\partial y'}$ e $B_n = \frac{\partial F}{\partial y}$. Então, podemos escrever (3.2.7) como

$$A_n \rho e_n + h B_n e_n + A_n T_n + h \eta_n = 0 \quad (3.2.8)$$

pois $F(t_n, y(t_n), y'(t_n)) = 0$.

Sejam (P_n, Q_n) as matrizes que transformam (A_n, B_n) para a forma (3.2.5).

Pela teoria apresentada na seção 2.3, vemos que:

$$(A + \lambda B)^{-1} = Q(PAQ + \lambda PBQ)^{-1}P = Q \begin{bmatrix} (I + \lambda C)^{-1} & 0 \\ 0 & (N + \lambda I)^{-1} \end{bmatrix} P = Q \begin{bmatrix} (I + \lambda C)^{-1} & 0 \\ 0 & \frac{1}{\lambda} \sum_{i=0}^{v-1} \frac{(-N)^i}{\lambda^i} \end{bmatrix} P \quad (3.2.9)$$

onde v é o índice de nilpotência de N .

Agora definimos as matrizes S_n e R_n por:

$$S_n = Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n \right)^{-1} A_n Q_n \quad \text{e} \quad R_n = Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n \right)^{-1} P_n^{-1} \quad (3.2.10)$$

Por (3.2.9) e sabendo que $v = 1$, temos:

$$S_n = \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n \right)^{-1} & 0 \\ 0 & \frac{\alpha_0}{h} I_{m_2} \end{bmatrix} P_n A_n Q_n = \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n \right)^{-1} & 0 \\ 0 & \frac{\alpha_0}{h} I_{m_2} \end{bmatrix} \begin{bmatrix} I_{m_1} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n \right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Analogamente:

$$R_n = \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n \right)^{-1} & 0 \\ 0 & \frac{\alpha_0}{h} I_{m_2} \end{bmatrix}$$

Sejam: $\tilde{e}_n = Q_n^{-1} e_n$, $\tilde{T}_n = Q_n^{-1} T_n$, $\tilde{\eta}_n = P_n \eta_n$.

Agora:

$$\rho e_n = \alpha_0 e_n + \alpha_1 e_{n-1} + \dots + \alpha_k e_{n-k}.$$

Logo:

$$\begin{aligned}\frac{\rho e_n}{\alpha_0} &= e_n + \frac{\alpha_1 e_{n-1}}{\alpha_0} + \dots + \frac{\alpha_k e_{n-k}}{\alpha_0} \\ &= e_n - \gamma_1 e_{n-1} - \dots - \gamma_k e_{n-k}, \text{ onde } \gamma_i = -\frac{\alpha_i}{\alpha_0}, \quad i = 1, \dots, k.\end{aligned}$$

De (3.2.8)

$$A_n(e_n - \gamma_1 e_{n-1} - \dots - \gamma_k e_{n-k}) + \frac{h}{\alpha_0} B_n e_n + A_n \frac{T_n}{\alpha_0} + \frac{h}{\alpha_0} \eta_n = 0$$

$$A_n e_n + \frac{h}{\alpha_0} B_n e_n = A_n(\gamma_1 e_{n-1} + \dots + \gamma_k e_{n-k}) - A_n \frac{T_n}{\alpha_0} - \frac{h}{\alpha_0} \eta_n$$

$$\left(A_n + \frac{h}{\alpha_0} B_n\right) e_n = A_n(\gamma_1 e_{n-1} + \dots + \gamma_k e_{n-k}) - A_n \frac{T_n}{\alpha_0} - \frac{h}{\alpha_0} \eta_n$$

Portanto:

$$e_n = \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n \sum_{i=1}^k \gamma_i e_{n-i} - \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n \frac{T_n}{\alpha_0} - \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} \frac{h}{\alpha_0} \eta_n.$$

Logo:

$$Q_n^{-1} e_n = Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n \sum_{i=1}^k \gamma_i e_{n-i} - Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n \frac{T_n}{\alpha_0} - Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} \frac{h}{\alpha_0} \eta_n$$

e assim:

$$\begin{aligned}\bar{e}_n &= Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n Q_n Q_n^{-1} \sum_{i=1}^k \gamma_i e_{n-i} - Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} A_n Q_n Q_n^{-1} \frac{T_n}{\alpha_0} - \\ &\quad Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} \frac{h}{\alpha_0} \eta_n\end{aligned}$$

$$= S_n Q_n^{-1} \sum_{i=1}^k \gamma_i e_{n-i} - S_n Q_n^{-1} \frac{T_n}{\alpha_0} - Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} \frac{h}{\alpha_0} \eta_n$$

$$= S_n Q_n^{-1} \sum_{i=1}^k \gamma_i e_{n-i} - S_n \frac{\tilde{T}_n}{\alpha_0} - Q_n^{-1} \left(A_n + \frac{h}{\alpha_0} B_n\right)^{-1} P_n^{-1} P_n \frac{h}{\alpha_0} \eta_n$$

$$\begin{aligned}
&= S_n Q_n^{-1} \sum_{i=1}^k \gamma_i e_{n-i} - S_n \frac{\tilde{T}_n}{\alpha_0} - R_n \frac{h}{\alpha_0} P_n \eta_n \\
&= S_n Q_n^{-1} \sum_{i=1}^k \gamma_i e_{n-i} - S_n \frac{\tilde{T}_n}{\alpha_0} - \frac{h}{\alpha_0} R_n \tilde{\eta}_n \\
&= S_n \sum_{i=1}^k Q_n^{-1} \gamma_i e_{n-i} - \left(\frac{1}{\alpha_0} \right) S_n \tilde{T}_n - \frac{h}{\alpha_0} R_n \tilde{\eta}_n
\end{aligned}$$

e assim,

$$\tilde{e}_n = S_n \sum_{i=1}^k \gamma_i (Q_n^{-1} Q_{n-i}) \tilde{e}_{n-i} - \left(\frac{1}{\alpha_0} \right) S_n \tilde{T}_n - \frac{h}{\alpha_0} R_n \tilde{\eta}_n \quad (3.2.11)$$

pois $\tilde{e}_n = Q_n^{-1} e_n$ e assim $e_n = Q_n \tilde{e}_n$, onde assumimos que $\gamma_i = -\frac{\alpha_i}{\alpha_0}$, onde α_i são os coeficientes do método BDF.

Sejam, $\tilde{e}_n = \begin{bmatrix} \tilde{e}_n^{(1)} \\ \tilde{e}_n^{(2)} \end{bmatrix}$, tal que $\tilde{e}_n^{(1)}$ tem dimensão m_1 e $\tilde{e}_n^{(2)}$ tem dimensão m_2 e

$\tilde{e}_n = [\tilde{e}_n^{(1)}, \dots, \tilde{e}_{n-k+1}^{(1)}, \tilde{e}_n^{(2)}, \dots, \tilde{e}_{n-k+1}^{(2)}]^t$, e façamos o mesmo para $\tilde{\eta}_n$ e $\tilde{T}_n$.

Denotando, $\left(I_{m_1} + \frac{h}{\alpha_0} C_n \right)^{-1} = X_n$ e $Q_n^{-1} Q_{n-i} = \begin{bmatrix} U_{n-i} & Z_{n-i} \\ W_{n-i} & V_{n-i} \end{bmatrix}$, obtemos:

$$\begin{bmatrix} \tilde{e}_n^{(1)} \\ \tilde{e}_n^{(2)} \end{bmatrix} = \gamma_1 \begin{bmatrix} X_n U_{n-1} & X_n Z_{n-1} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-1}^{(1)} \\ \tilde{e}_{n-1}^{(2)} \end{bmatrix} + \dots + \gamma_k \begin{bmatrix} X_n U_{n-k} & X_n Z_{n-k} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{e}_{n-k}^{(1)} \\ \tilde{e}_{n-k}^{(2)} \end{bmatrix} - \left(\frac{1}{\alpha_0} \right) S_n \tilde{T}_n - \left(\frac{h}{\alpha_0} \right) R_n \tilde{\eta}_n.$$

Agora, sendo $\varphi_n = -\left(\frac{1}{\alpha_0} \right) X_n \tilde{T}_n^{(1)} - \left(\frac{h}{\alpha_0} \right) X_n \tilde{\eta}_n^{(1)}$ e $\psi_n = -\left(\frac{\alpha_0}{h} \right) \tilde{\eta}_n^{(2)}$, temos:

$$\tilde{e}_n^{(1)} = \gamma_1 X_n U_{n-1} \tilde{e}_{n-1}^{(1)} + \dots + \gamma_k X_n U_{n-k} \tilde{e}_{n-k}^{(1)} + \gamma_1 X_n Z_{n-1} \tilde{e}_{n-1}^{(2)} + \dots + \gamma_k X_n Z_{n-k} \tilde{e}_{n-k}^{(2)} + \varphi_n.$$

$$\tilde{e}_n^{(2)} = 0 + \left(\frac{h}{\alpha_0} \right) \psi_n.$$

Assim podemos escrever:

$$\tilde{\mathbf{e}}_n^{(1)} = [\gamma_1 \mathbf{X}_n \mathbf{U}_{n-1} \quad \cdots \quad \gamma_k \mathbf{X}_n \mathbf{U}_{n-k}] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(1)} \end{bmatrix} + [\gamma_1 \mathbf{X}_n \mathbf{Z}_{n-1} \quad \cdots \quad \gamma_k \mathbf{X}_n \mathbf{Z}_{n-k}] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(2)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(2)} \end{bmatrix} + \varphi_n$$

$$\tilde{\mathbf{e}}_{n-1}^{(1)} = [\mathbf{I} \quad 0 \quad \cdots \quad 0] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(1)} \end{bmatrix}$$

$\vdots$
 $\vdots$

$$\tilde{\mathbf{e}}_{n-k+1}^{(1)} = [0 \quad \cdots \quad 0 \quad \mathbf{I} \quad 0] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(1)} \end{bmatrix}$$

$$\tilde{\mathbf{e}}_n^{(2)} = 0 + \left(\frac{h}{\alpha_0} \right) \psi_n$$

$$\tilde{\mathbf{e}}_{n-1}^{(2)} = [\mathbf{I} \quad 0 \quad \cdots \quad 0] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(2)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(2)} \end{bmatrix}$$

$\vdots$
 $\vdots$

$$\tilde{\mathbf{e}}_{n-k+1}^{(2)} = [0 \quad \cdots \quad 0 \quad \mathbf{I} \quad 0] \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(2)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(2)} \end{bmatrix}$$

Assim:

$$\begin{bmatrix} \tilde{\mathbf{e}}_n^{(1)} \\ \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k+1}^{(1)} \\ \tilde{\mathbf{e}}_n^{(2)} \\ \tilde{\mathbf{e}}_{n-1}^{(2)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k+1}^{(2)} \end{bmatrix} = \begin{bmatrix} \gamma_1 \mathbf{X}_n \mathbf{U}_{n-1} & \cdots & \cdots & \gamma_1 \mathbf{X}_n \mathbf{U}_{n-k} & \gamma_1 \mathbf{X}_n \mathbf{Z}_{n-1} & \cdots & \cdots & \gamma_1 \mathbf{X}_n \mathbf{Z}_{n-k} \\ \mathbf{I} & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ 0 & & \mathbf{I} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots & \mathbf{I} & 0 & \cdots & 0 \\ \vdots & & & \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & & \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \tilde{\mathbf{e}}_{n-2}^{(1)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(1)} \\ \tilde{\mathbf{e}}_{n-1}^{(2)} \\ \tilde{\mathbf{e}}_{n-2}^{(2)} \\ \vdots \\ \tilde{\mathbf{e}}_{n-k}^{(2)} \end{bmatrix} -$$

$$-\left(\frac{1}{\alpha_0}\right) \begin{bmatrix} X_n \tilde{T}_n^{(1)} \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} - \left(\frac{h}{\alpha_0}\right) \begin{bmatrix} X \tilde{\eta}_n^{(1)} \\ 0 \\ \vdots \\ 0 \\ \left(\frac{\alpha_0}{h}\right) \tilde{\eta}_n^{(2)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Pelo lema 3.2.1, temos que $Q_n^{-1}Q_{n-1} = I_m + O(h)$, donde:

$$\tilde{\epsilon}_n = K \tilde{\epsilon}_{n-1} - \left(\frac{1}{\alpha_0}\right) \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n\right)^{-1} \tilde{T}_n^{(1)} \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} - \left(\frac{h}{\alpha_0}\right) \begin{bmatrix} \left(I_{m_1} + \frac{h}{\alpha_0} C_n\right)^{-1} \tilde{\eta}_n^{(1)} \\ 0 \\ \vdots \\ 0 \\ \left(\frac{\alpha_0}{h}\right) \tilde{\eta}_n^{(2)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.2.12)$$

onde a matriz K é como no lema 3.2.2.

Resolvendo a relação de recorrência (3.2.12) obtemos:

$$\tilde{\mathbf{e}}_n = \mathbf{K}^n \tilde{\mathbf{e}}_0 - \left(\frac{1}{\alpha_0} \right) \sum_{i=0}^{n-1} \mathbf{K}^i \begin{bmatrix} \left(\mathbf{I}_{m_1} + \frac{h}{\alpha_0} \mathbf{C}_{n-i} \right)^{-1} \tilde{\mathbf{T}}_{n-i}^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \left(\frac{h}{\alpha_0} \right) \sum_{i=0}^{n-1} \mathbf{K}^i \begin{bmatrix} \left(\mathbf{I}_{m_1} + \frac{h}{\alpha_0} \mathbf{C}_{n-i} \right)^{-1} \tilde{\eta}_{n-i}^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.2.13)$$

Usando o lema 3.2.2 para limitar a grandeza dos blocos de $\mathbf{K}^i$ para $2k \leq i \leq n$ e observando a grandeza dos blocos de $\mathbf{K}^i$ para $i < 2k$, temos de (3.2.13):

$$\|\tilde{\mathbf{e}}_n^{(1)}\| \leq O(\|\tilde{\mathbf{e}}_0\|) + O(h^k) + O(\|\tilde{\eta}_n^{(1)}\|) + O(\|\tilde{\eta}_n^{(2)}\|) \quad (3.2.14a)$$

$$\|\tilde{\mathbf{e}}_n^{(2)}\| \leq O(\|\tilde{\eta}_n^{(2)}\|) \quad (3.2.14b)$$

$$\text{pois } \left(\mathbf{I}_{m_1} + \frac{h}{\alpha_0} \mathbf{C}_{n-i} \right)^{-1} = O(1).$$

Para problemas lineares $\tilde{\eta}$ é composto dos resíduos da iteração de Newton que são $O(h^{k+1})$ e a demonstração está concluída.

Para problemas não lineares, temos por (3.2.12) que:

$$\tilde{\mathbf{e}}_n^{(1)} = \begin{bmatrix} (\mathbf{K}_{11}^{(1)} + O(h)) & O(h) \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_{n-1}^{(1)} \\ \tilde{\mathbf{e}}_{n-1}^{(2)} \end{bmatrix} - \left(\frac{1}{\alpha_0} \right) \left(\mathbf{I}_{m_1} + \frac{h}{\alpha_0} \mathbf{C}_n \right)^{-1} \tilde{\mathbf{T}}_n^{(1)} - \left(\frac{h}{\alpha_0} \right) \left(\mathbf{I}_{m_1} + \frac{h}{\alpha_0} \mathbf{C}_n \right)^{-1} \tilde{\eta}_n^{(1)}.$$

Então:

$$\tilde{\mathbf{e}}_n^{(1)} = (\mathbf{K}_{11}^{(1)} + O(h)) \tilde{\mathbf{e}}_{n-1}^{(1)} + O(h) \tilde{\mathbf{e}}_{n-1}^{(2)} - \left(\frac{1}{\alpha_0} \right) O(1) \tilde{\mathbf{T}}_n^{(1)} - \left(\frac{h}{\alpha_0} \right) O(1) \tilde{\eta}_n^{(1)}$$

onde $\begin{bmatrix} (\mathbf{K}_{11}^{(1)} + O(h)) & O(h) \end{bmatrix}$ é a primeira linha da matriz $\mathbf{K}$ definida no lema 3.2.2.

Podemos escrever:

$$\tilde{\mathbf{e}}_n^{(1)} = \mathbf{K}_{11}^{(1)} \tilde{\mathbf{e}}_{n-1}^{(1)} + \mathcal{O}(h) \tilde{\mathbf{e}}_{n-1}^{(1)} + \mathcal{O}(h) \tilde{\mathbf{e}}_{n-1}^{(2)} - \left(\frac{1}{\alpha_0} \right) \mathcal{O}(1) \tilde{\mathbf{T}}_n^{(1)} - \left(\frac{h}{\alpha_0} \right) \mathcal{O}(1) \tilde{\eta}_n^{(1)}.$$

Logo:

$$\frac{\rho \tilde{\mathbf{e}}_n^{(1)}}{\alpha_0} = \sum_{i=1}^k \mathcal{O}(h) \tilde{\mathbf{e}}_{n-i}^{(1)} + \sum_{i=1}^k \mathcal{O}(h) \tilde{\mathbf{e}}_{n-i}^{(2)} - \left(\frac{1}{\alpha_0} \right) \mathcal{O}(1) \tilde{\mathbf{T}}_n^{(1)} - \left(\frac{h}{\alpha_0} \right) \mathcal{O}(1) \tilde{\eta}_n^{(1)}.$$

Assim:

$$\frac{\rho \tilde{\mathbf{e}}_n^{(1)}}{h} = \frac{\alpha_0 \sum_{i=1}^k \mathcal{O}(h) (\tilde{\mathbf{e}}_{n-i}^{(1)})}{h} + \frac{\alpha_0 \sum_{i=1}^k \mathcal{O}(h) (\tilde{\mathbf{e}}_{n-i}^{(2)})}{h} - \left(\frac{1}{h} \right) \mathcal{O}(1) \tilde{\mathbf{T}}_n^{(1)} - \left(\frac{h}{h} \right) \mathcal{O}(1) \tilde{\eta}_n^{(1)}$$

e portanto:

$$\left\| \frac{\rho \tilde{\mathbf{e}}_n^{(1)}}{h} \right\| \leq \mathcal{O}(\|\tilde{\mathbf{e}}_n^{(1)}\|) + \mathcal{O}(\|\tilde{\mathbf{e}}_n^{(2)}\|) + \mathcal{O}\left(\left\| \frac{\tilde{\mathbf{T}}_n^{(1)}}{h} \right\| \right) + \mathcal{O}(\|\tilde{\eta}_n^{(1)}\|).$$

Como $\mathcal{O}\left(\left\| \frac{\tilde{\mathbf{T}}_n^{(1)}}{h} \right\| \right) = \mathcal{O}(h^k)$, temos que:

$$\left\| \frac{\rho \tilde{\mathbf{e}}_n^{(1)}}{h} \right\| \leq \mathcal{O}(\|\tilde{\mathbf{e}}_0\|) + \mathcal{O}(h^k) + \mathcal{O}(\|\tilde{\eta}_n^{(1)}\|) + \mathcal{O}(\|\tilde{\eta}_n^{(2)}\|) \quad (3.2.15a)$$

e por (3.2.14b):

$$\left\| \frac{\rho \tilde{\mathbf{e}}_n^{(2)}}{h} \right\| \leq \mathcal{O}\left(\left\| \frac{\tilde{\eta}_n^{(2)}}{h} \right\| \right), \quad (3.2.15b)$$

que devido aos erros nos valores iniciais, estes limitantes são válidos para $n \geq k+1$.

Como $\mathbf{F}$ está na forma (3.2.4), $\eta_n^{(1)}$ consiste dos resíduos da iteração de Newton que são $\mathcal{O}(h^{k+1})$ e termos de ordem mais alta da forma:

$$\mathbf{e}_n^T \mathbf{F}_{yy} \mathbf{e}_n, \quad \left(\frac{\rho \mathbf{e}_n^T}{h} \right) \mathbf{F}_{yy} \mathbf{e}_n, \quad \left(\frac{\rho \mathbf{e}_n^T}{h} \right) \mathbf{F}_{y'y'} \left(\frac{\rho \mathbf{e}_n}{h} \right),$$

e $\eta_n^{(2)}$ consiste dos resíduos da iteração de Newton e termos da forma:

$$\mathbf{e}_n^T \mathbf{F}_{yy} \mathbf{e}_n.$$

Como P_n tem a forma dada por (3.2.6a) 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.

Suponha que os termos de alta ordem não lineares satisfazem $\|\tilde{\eta}_n^{(i)}\| \leq \varepsilon_i$, $i = 1, 2$ (isto não inclui os erros da iteração de Newton, que assumimos que são $O(h^{k+1})$), e os erros nas condições iniciais satisfazem $\|\tilde{e}_0\| = O(h^k)$. Então para concluir a prova, mostraremos que ε_i são pequenos.

Para isto, observemos que $\tilde{\eta}_n^{(1)}$ e $\tilde{\eta}_n^{(2)}$ são respectivamente das formas

$$\tilde{\eta}_n^{(1)} = \frac{1}{2} e_n^T F_{yy} e_n + \frac{1}{2} \left(\frac{\rho e_n^T}{h} \right) F_{y'y} e_n + \frac{1}{2} \left(\frac{\rho e_n^T}{h} \right) F_{y'y'} \left(\frac{\rho e_n}{h} \right) + O(h^{k+1})$$

$$\tilde{\eta}_n^{(2)} = \frac{1}{2} e_n^T F_{yy} e_n + O(h^{k+1}).$$

Consideremos

$$F_{yy} = \begin{bmatrix} F_{yy}^1 & F_{yy}^2 \\ F_{yy}^3 & F_{yy}^4 \end{bmatrix}, F_{y'y} = \begin{bmatrix} F_{y'y}^1 & F_{y'y}^2 \\ F_{y'y}^3 & F_{y'y}^4 \end{bmatrix}, F_{y'y'} = \begin{bmatrix} F_{y'y'}^1 & F_{y'y'}^2 \\ F_{y'y'}^3 & F_{y'y'}^4 \end{bmatrix},$$

e assim:

$$\begin{aligned} \tilde{\eta}_n^{(1)} = & \frac{1}{2} \begin{bmatrix} (e_n^{(1)})^t & (e_n^{(2)})^t \end{bmatrix} \begin{bmatrix} F_{yy}^1 e_n^{(1)} + F_{yy}^2 e_n^{(2)} \\ F_{yy}^3 e_n^{(1)} + F_{yy}^4 e_n^{(2)} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \rho (e_n^{(1)})^t & (e_n^{(2)})^t \end{bmatrix} \frac{1}{h} \begin{bmatrix} F_{y'y}^1 e_n^{(1)} + F_{y'y}^2 e_n^{(2)} \\ F_{y'y}^3 e_n^{(1)} + F_{y'y}^4 e_n^{(2)} \end{bmatrix} + \\ & \frac{1}{2} \begin{bmatrix} \rho (e_n^{(1)})^t & (e_n^{(2)})^t \end{bmatrix} \frac{1}{h} \begin{bmatrix} F_{y'y'}^1 \frac{\rho (e_n^{(1)})^t}{h} + F_{y'y'}^2 \frac{(e_n^{(2)})^t}{h} \\ F_{y'y'}^3 \frac{\rho (e_n^{(1)})^t}{h} + F_{y'y'}^4 \frac{(e_n^{(2)})^t}{h} \end{bmatrix} + O(h^{k+1}) \end{aligned}$$

$$\tilde{\eta}_n^{(2)} = \frac{1}{2} \begin{bmatrix} (e_n^{(1)})^t & (e_n^{(2)})^t \end{bmatrix} \begin{bmatrix} F_{yy}^1 e_n^{(1)} + F_{yy}^2 e_n^{(2)} \\ F_{yy}^3 e_n^{(1)} + F_{yy}^4 e_n^{(2)} \end{bmatrix} + O(h^{k+1}).$$

Segue que:

$$\begin{aligned}\tilde{\eta}_n^{(1)} = & \frac{1}{2} \left\{ (e_n^{(1)})^t F_{yy}^1 e_n^{(1)} + (e_n^{(1)})^t F_{yy}^2 e_n^{(2)} + (e_n^{(2)})^t F_{yy}^3 e_n^{(1)} + (e_n^{(2)})^t F_{yy}^4 e_n^{(2)} \right\} + \\ & \frac{1}{2} \left\{ \frac{\rho(e_n^{(1)})^t}{h} F_{yy}^1 e_n^{(1)} + \frac{\rho(e_n^{(1)})^t}{h} F_{yy}^2 e_n^{(2)} + \frac{\rho(e_n^{(2)})^t}{h} F_{yy}^3 e_n^{(1)} + \frac{\rho(e_n^{(2)})^t}{h} F_{yy}^4 e_n^{(2)} \right\} + \\ & \frac{1}{2} \left\{ \frac{\rho(e_n^{(1)})^t}{h} F_{yy'}^1 \frac{\rho e_n^{(1)}}{h} + \frac{\rho(e_n^{(1)})^t}{h} F_{yy'}^2 \frac{\rho e_n^{(2)}}{h} + \frac{\rho(e_n^{(2)})^t}{h} F_{yy'}^3 \frac{\rho e_n^{(1)}}{h} + \frac{\rho(e_n^{(2)})^t}{h} F_{yy'}^4 \frac{\rho e_n^{(2)}}{h} \right\} + O(h^{k+1})\end{aligned}$$

$$\tilde{\eta}_n^{(2)} = \frac{1}{2} \left\{ (e_n^{(1)})^t F_{yy}^1 e_n^{(1)} + (e_n^{(1)})^t F_{yy}^2 e_n^{(2)} + (e_n^{(2)})^t F_{yy}^3 e_n^{(1)} + (e_n^{(2)})^t F_{yy}^4 e_n^{(2)} \right\} + O(h^{k+1}).$$

Tomando normas, obtemos:

$$\begin{aligned}\|\tilde{\eta}_n^{(1)}\| \leq & C_1 \left\{ \|e_n^{(1)}\|^2 + \|e_n^{(1)}\| \|e_n^{(2)}\| + \|e_n^{(2)}\|^2 \right\} + \\ & C_2 \left\{ \left\| \frac{\rho e_n^{(1)}}{h} \right\| \|e_n^{(1)}\| + \left\| \frac{\rho e_n^{(1)}}{h} \right\| \|e_n^{(2)}\| + \left\| \frac{\rho e_n^{(2)}}{h} \right\| \|e_n^{(1)}\| + \left\| \frac{\rho e_n^{(2)}}{h} \right\| \|e_n^{(2)}\| \right\} + \\ & C_3 \left\{ \left\| \frac{\rho e_n^{(1)}}{h} \right\|^2 + \left\| \frac{\rho e_n^{(1)}}{h} \right\| \left\| \frac{\rho e_n^{(2)}}{h} \right\| + \left\| \frac{\rho e_n^{(2)}}{h} \right\|^2 \right\} + O(h^{k+1}) \\ \|\tilde{\eta}_n^{(2)}\| \leq & C_4 \left\{ \|e_n^{(1)}\|^2 + \|e_n^{(1)}\| \|e_n^{(2)}\| + \|e_n^{(2)}\|^2 \right\} + O(h^{k+1})\end{aligned}$$

onde C_1 , C_2 , C_3 e C_4 são constantes.

Temos ainda que (3.2.14a) e (3.2.14b) valem também para $\|e_n^{(1)}\|$ e $\|e_n^{(2)}\|$, respectivamente e, (3.2.15a) e (3.2.15b) valem também para $\left\| \frac{\rho e_n^{(1)}}{h} \right\|$ e $\left\| \frac{\rho e_n^{(2)}}{h} \right\|$, respectivamente, pois $\tilde{e}_n = Q_n^{-1} e_n$ onde Q_n é uma matriz não singular dada por (3.2.6b). Assim para $n \geq k+1$ temos:

$$\|\tilde{\eta}_n^{(1)}\| \leq k_1 \left(h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + \left(2 + \frac{2}{h} \right) \varepsilon_1 \varepsilon_2 + 2 \frac{\varepsilon_2^2}{h} + \frac{\varepsilon_2^2}{h^2} + 2h^k \varepsilon_1 + 2h^k \varepsilon_2 + 2h^{k-1} \varepsilon_2 \right)$$

$$\|\tilde{\eta}_n^{(2)}\| \leq k_2 (h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_1 \varepsilon_2 + 2h^k \varepsilon_1 + 2h^k \varepsilon_2)$$

onde k_1 e k_2 são constantes.

Agora determinaremos ε_1 e ε_2 como solução de

$$\varepsilon_1 = k_1 \left(h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + \left(2 + \frac{2}{h} \right) \varepsilon_1 \varepsilon_2 + 2 \frac{\varepsilon_2^2}{h} + \frac{\varepsilon_2^2}{h^2} + 2h^k \varepsilon_1 + 2h^k \varepsilon_2 + 2h^{k-1} \varepsilon_2 \right) \quad (3.2.16)$$

$$\varepsilon_2 = k_2 (h^{2k} + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_1 \varepsilon_2 + 2h^k \varepsilon_1 + 2h^k \varepsilon_2)$$

resolvendo (3.2.16) por uma iteração funcional $\hat{\varepsilon} = G(\hat{\varepsilon})$ com valor inicial $\hat{\varepsilon}^{(0)}$ satisfazendo $\hat{\varepsilon}^{(0)} = O(h^{2k})$, onde $\hat{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$.

Para $k \geq 2$, podemos usar o teorema da contração para concluir que $\hat{\varepsilon} = O(h^{2k})$, e a demonstração está concluída.

Para $k=1$, não podemos aplicar o teorema diretamente, pois $\left\| \frac{\partial G}{\partial \varepsilon} \right\| = O(1)$. Mas se fizermos

$$\begin{cases} \hat{\varepsilon}_1 = \frac{\varepsilon_1}{\sqrt{h}} \\ \hat{\varepsilon}_2 = \varepsilon_2 \end{cases}$$

então, podemos aplicar a estratégia acima. Finalmente, o resultado segue de (3.2.14), concluindo, assim, a demonstração do teorema 3.2.1.

3.2.3 Sistemas semi-explícitos de índice 2

Nesta subseção estudaremos o comportamento dos métodos BDF aplicados a sistemas semi-explícitos de índice 2. Veremos que um método BDF de k -passos ($k < 7$) é convergente com $O(h^k)$.

Consideremos o sistema não linear semi-explícito

$$\begin{cases} f(x, x', y, t) = 0 & (3.2.17a) \\ g(x, y, t) = 0 & (3.2.17b) \end{cases}$$

onde assumimos que o sistema tem índice 2, $\left(\frac{\partial f}{\partial x'} \right)^{-1}$ existe e é limitada em uma vizinhança da solução, $\left(\frac{\partial g}{\partial y} \right)$ tem posto constante, e f e g tem quantas derivadas parciais contínuas desejarmos em uma vizinhança da solução.

Teorema 3.2.2 *Suponha que o sistema semi-implícito de índice 2 (3.2.17) é resolvido numericamente por um método BDF de k -passos ($k < 7$), os erros nos valores iniciais são $O(h^k)$, e os erros na iteração de Newton satisfazem $O(h^{k+1})$. Então o método BDF de k -passos é convergente (globalmente) com $O(h^k)$, depois de $(k+1)$ passos.*

Demonstração:

Para o sistema semi-implícito de índice 2 (3.2.17) seja $w' = y$ e consideremos o sistema não linear de índice 1

$$\begin{cases} f(x, x', w', t) = 0 \\ g(x, w', t) = 0 \end{cases} \quad \begin{matrix} (3.2.18a) \\ (3.2.18b) \end{matrix}$$

É fácil ver que a solução para x , obtida resolvendo o sistema de índice 1 (3.2.18) pelo método BDF de k -passos, é a mesma encontrada quando resolvemos o sistema semi-implícito de índice 2 original (3.2.17) pelo método BDF de k -passos (fato que decorre da teoria de EDO).

Agora consideremos o erro em y . A discretização de (3.2.18) pelo método BDF de k -passos nos dá:

$$\begin{cases} f\left(x_n, \frac{\rho x_n}{h}, \frac{\rho w_n}{h}, t_n\right) = 0 \\ g\left(x_n, \frac{\rho w_n}{h}, t_n\right) = 0 \end{cases} \quad \begin{matrix} (3.2.19a) \\ (3.2.19b) \end{matrix}$$

Seja $e_n^w = w_n - w(t_n)$. Então por (3.2.15) temos que para $n \geq k+1$,

$$\left\| \frac{\rho e_n^w}{h} \right\| = O(h^k). \quad (3.2.20)$$

Seja $e_n^y = y_n - y(t_n)$. Então, temos:

$$y_n = \frac{\rho w_n}{h} = \frac{\rho w(t_n)}{h} + \frac{\rho e_n^w}{h}. \quad (3.2.21)$$

A solução analítica satisfaz:

$$y(t_n) = w'(t_n). \quad (3.2.22)$$

Subtraindo (3.2.22) de (3.2.21) e notando que $w'(t_n) - \frac{\rho w(t_n)}{h}$ é $O(h^k)$ para o método BDF de k-passos, nós obtemos:

$$e_n^y = \frac{\rho e_n^w}{h} + O(h^k) = O(h^k)$$

concluindo, assim, a demonstração do teorema 3.2.2.

3.3 Algoritmo

Faremos agora, alguns comentários sobre a implementação numérica dos métodos BDF. Esta classe numérica encontra uma aproximação para a solução clássica $y(t)$ da EAD (2.2.1).

Dentre outros modos, usando o método de Euler implícito temos que:

$$y_n' = \frac{y_n - y_{n-1}}{h}$$

onde $h = t_n - t_{n-1}$. Assim obtemos o sistema

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

o qual, é resolvido para y_n , em cada passo, por exemplo, pelo método de Newton modificado. Isto nos dá:

$$y_n^{k+1} = y_n^k - (A_n^0)^{-1} F\left(t_n^k, y_n^k, \frac{y_n^k - y_{n-1}}{h}\right)$$

onde

$$A_n^0 = F_y\left(t_n^0, y_n^0, \frac{y_n^0 - y_{n-1}}{h}\right) + \frac{1}{h} F_y\left(t_n^0, y_n^0, \frac{y_n^0 - y_{n-1}}{h}\right).$$

Como é usual, não invertemos A_n^0 , resolvemos o sistema

$$(A_n^0)S = -F\left(t_n^k, y_n^k, \frac{y_n^k - y_{n-1}}{h}\right)$$

onde $S = y_n^{k+1} - y_n^k$. Assim obtemos:

$$\begin{cases} y_n^{k+1} = S + y_n^k \\ (y_n^{k+1})' = \frac{y_n^{k+1} - y_{n-1}}{h} \end{cases}.$$

O custo de um passo desta classe numérica (método BDF) é aproximadamente $(2m^3 + r_3m^2)$, onde r_3 é o número de iterações do método de Newton modificado.

4 EAD COM SINGULARIDADES:

4.1 Introdução:

No capítulo anterior não tratamos de EAD com singularidades, isto porque os métodos BDF, quando aplicados a EAD com singularidades não apresentam um bom comportamento. Neste capítulo, então, estudaremos uma nova classe de métodos para a solução numérica de EAD, desenvolvida com base na teoria apresentada em [Castelo/Tavares - 94], que em particular, resolve EAD com singularidades satisfatoriamente.

4.2 Resultados preliminares:

Nesta subseção apresentaremos alguns resultados teóricos, necessários para o desenvolvimento desta nova classe de métodos numéricos.

Definição 4.2.1 *Dada uma EAD de índice v , sabemos que após v diferenciações esta estará reduzida a uma EDO do tipo*

$$y' = G(t, y). \quad (4.2.1)$$

Disto definimos K como o conjunto conexo, aberto, maximal onde G está definida.

Definição 4.2.2 *O índice v da EAD (2.2.1) sobre K é o número mínimo de vezes que (2.2.1), ou parte dela, será diferenciada com relação a t para determinar y' como uma função contínua de y e t , isto é, $y'(t) = G(t, y(t))$, com $G: K \rightarrow R^m$ (observando-se também a EAD).*

Vamos agora definir singularidade de uma EAD:

Definição 4.2.3 *Um ponto $P_0 = (t_0, y_0)$ é uma singularidade da EAD (2.2.1) se $P_0 \in K$, $G(t_0, y_0) = 0$ e $F(P_0, z_0) = 0$, ou P_0 está no bordo de K e $F(P_0, z_0) = 0$, para algum $z_0 \in R^m$.*

Definição 4.2.4 A imagem de uma função $c: J \rightarrow \mathbb{R}^{2m+1}$ é uma solução geral da EAD (2.2.1) se $c \in C^1(J)$ e satisfaz

$$\begin{cases} F(c(s)) = 0 \\ Y_1(s)T'(s) - Y_0'(s) = 0 \end{cases} \quad (4.2.2)$$

para todo $s \in J$ onde $c(s) = (T(s), Y_0(s), Y_1(s))$ com $T(s) \in \mathbb{R}$, $Y_0(s) \in \mathbb{R}^m$ e $Y_1(s) \in \mathbb{R}^m$. Se a função F em (2.2.1) não depende de y' , isto é, (2.2.1) é uma equação puramente algébrica então (4.2.2) se reduz à $F(c(s)) = 0$ e $c(s) = (T(s), Y_0(s))$.

Observe que se $\alpha: L \rightarrow J$ com $L \subset \mathbb{R}$ é um difeomorfismo de classe C^1 e $c: J \rightarrow \mathbb{R}^{2m+1}$ é de classe C^1 e satisfaz (4.2.2) (isto é, $c(J)$ é uma solução geral de (2.2.1)), então $c \circ \alpha: L \rightarrow \mathbb{R}^{2m+1}$ é também de classe C^1 e satisfaz (4.2.2) (isto é, $(c \circ \alpha)(L)$ é uma solução geral de (2.2.1)), mas o traço de c e $(c \circ \alpha)$ é o mesmo, isto é $c(J) = (c \circ \alpha)(L)$.

Para exemplificar as definições acima, tomemos a EAD puramente algébrica $F(t, y) = 0$ onde $F: \Omega \rightarrow \mathbb{R}$ é dada por $F(t, y) = y^2 + t^2 - 1$ e $\Omega \subset \mathbb{R}^2$.

Seu índice v é 1, pois diferenciando F com relação a t , apenas uma vez, temos:

$$2yy' + 2t = 0$$

ou seja, $y' = \frac{-2t}{2y}$. Assim a função $G(t, y)$ da definição 4.2.1 é dada por:

$$G(t, y) = \frac{-2t}{2y}.$$

Pela equação $F(t, y) = 0$ vemos que se $|t| = 1$ então $y = 0$, e assim concluímos que temos duas possibilidades para a escolha do conjunto K , da definição 4.2.1, que são:

$$K_1 = \{(t, y) \in \mathbb{R}^2 : |t| < 1, 0 < y \leq 1\}$$

e

$$K_2 = \{(t, y) \in \mathbb{R}^2 : |t| < 1, -1 \leq y < 0\}.$$

Temos que os pontos $P_1 = (0, 1)$, $P_2 = (0, -1)$, $P_3 = (1, 0)$ e $P_4 = (-1, 0)$ são pontos singulares da EAD $y^2 + t^2 - 1 = 0$, pois $G(0, 1) = G(0, -1) = 0$ e P_i , $i = 3, 4$ estão no bordo de K_1 e K_2 (e satisfazem $y^2 + t^2 - 1 = 0$).

Agora seja $c: J \rightarrow \mathbb{R}^2$ definida por:

$$c(s) = (\sin(s), \cos(s)).$$

Temos que $c \in C^1(J)$, onde $J \subset \mathbb{R}$. Agora, $F(c(s)) = 0$. Assim a imagem da função c , isto é, $c(J)$ é uma solução geral da EAD $y^2 + t^2 - 1 = 0$.

Os resultados apresentados a seguir, podem ser encontrados em [Castelo/Tavares - 94].

Teorema 4.2.1 *Considere uma solução geral de (2.2.1) dada por $c: J \rightarrow \mathbb{R}^{2m+1}$. Se $T'(s) \neq 0$ em um intervalo $I \subset J$, então $Y_0(s) = y_0(T(s))$, $Y_1(s) = \frac{d}{dT} y_0(T(s))$ em I e y_0 definida em $T(I)$ é uma solução clássica da EAD (2.2.1).*

Demonstração:

Como $T'(s) \neq 0$ em $I \subset J$, $T(I)$ é um intervalo e T é um difeomorfismo de I em $T(I)$. Assim $y_0: T(I) \rightarrow \mathbb{R}^m$ com $y_0(T(s)) = Y_0(s)$ é uma função de classe C^1 .

$$\text{Então } Y_1(s)T'(s) = Y_0'(s) = \frac{d}{dT} y_0(T(s))T'(s) \Rightarrow Y_1(s) = \frac{d}{dT} y_0(T(s)).$$

$$\text{Agora } F\left(T(s), y_0(T(s)), \frac{d}{dT} y_0(T(s))\right) = F(T(s), Y_0(s), Y_1(s)) = 0.$$

Assim, y_0 é uma solução clássica de (2.2.1) no conjunto $T(I)$, concluindo, assim, a demonstração do teorema 4.2.1.

O problema de valor inicial para a EAD (2.2.1) é definido como:

$$\begin{cases} F(t, y(t), y'(t)) = 0 \\ y(t_0) = a_0 \\ y'(t_0) = a_1 \end{cases} \quad (4.2.3)$$

onde $F(t_0, a_0, a_1) = 0$.

Seja $w: \Omega \rightarrow \mathbb{R}^{m \times (2m+1)}$, onde Ω é um subconjunto aberto conexo de $\mathbb{R}^{2m+1}$, definido por:

$$w(T, Y_0, Y_1) = \begin{pmatrix} Y_1 & -I_m & 0 \end{pmatrix}_{m \times (2m+1)} \text{ com } \begin{cases} Y_1 \in \mathbb{R}^{m \times 1} \\ I_m \in \mathbb{R}^{m \times m} \\ 0 \in \mathbb{R}^{m \times m} \end{cases}.$$

Agora podemos reescrever (4.2.2) como:

$$\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \end{cases} \quad (4.2.4)$$

$$\begin{aligned} \text{pois, } w(c(s))c'(s) &= \begin{pmatrix} Y_1(s) & -I_m & 0 \end{pmatrix} \begin{pmatrix} T'(s) & Y_0'(s) & Y_1'(s) \end{pmatrix}' = Y_1(s)T'(s) - I_m Y_0'(s) + 0 Y_1'(s) \\ &= Y_1(s)T'(s) - Y_0'(s). \end{aligned}$$

Se F não depende de y' , isto é, é um sistema puramente algébrico, definimos (4.2.4) como:

$$F(c(s)) = 0.$$

O sistema (4.2.4) é chamado o sistema diferencial associado à EAD (2.2.1).

O problema de valor inicial para (4.2.4) é definido como:

$$\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \\ c(0) = c_0 = (t_0, a_0, a_1) \end{cases} \quad (4.2.5)$$

com $F(c_0) = 0$.

Diferenciando a primeira equação de (4.2.5) com respeito a s e definindo:

$$A(c(s)) = \begin{pmatrix} DF(c(s)) \\ w(c(s)) \end{pmatrix}_{2m \times (2m+1)}$$

temos um novo problema de valor inicial:

$$\begin{cases} A(c(s))c'(s) = 0 \\ c(0) = c_0 = (t_0, a_0, a_1) \end{cases} \quad (4.2.6)$$

Note que se (2.2.1) é do tipo $F(t, y(t)) = 0$ então $A(c(s)) = DF(c(s))$.

Definição 4.2.5 Um ponto $c^0 = (t^0, a^0, a^1)$ é dito ser uma singularidade do sistema diferencial (4.2.4) se $F(c^0) = 0$ e $A(c^0)$ não tem posto máximo.

Vamos agora ver alguns exemplos:

Exemplo 4.2.1
$$\begin{cases} (y(t))^2 = t \\ y(t_0) = y_0 > 0, t_0 > 0, y_0^2 = t_0 \end{cases}$$

Neste exemplo $G(t,y) = \frac{1}{2y}$, que é sempre não nula e definida para qualquer (t,y) tal que $y \neq 0$.

Esta EAD tem índice 1 sobre os conjuntos

$$K_1 = \{(t,y): 0 < t < \infty, 0 < y < \infty\}$$

e

$$K_2 = \{(t,y): 0 < t < \infty, -\infty < y < 0\}.$$

O ponto $(0,0)$ é uma singularidade e a solução clássica é $y(t) = t^{\frac{1}{2}}$ para $t \in (0, \infty)$.

O sistema diferencial associado é

$$\begin{cases} Y_0^2(s) - T(s) = 0 \\ T(0) = t_0 \\ Y_0(0) = y_0 \end{cases}$$

que não tem nenhuma singularidade e a solução geral é dada por:

$$c(s) = \left(\left(s + t_0^{\frac{1}{2}} \right)^2, \left(s + t_0^{\frac{1}{2}} \right) \right)$$

para $s \in \mathbb{R}$, onde $c(0) = \left(t_0, t_0^{\frac{1}{2}} \right) = (t_0, y_0)$ é a condição inicial. Como $T' \left(-t_0^{\frac{1}{2}} \right) = 0$ e $c(0) = (t_0, y_0)$, usando o teorema 4.2.1 nós temos que $c(s) = (t, y(t))$ para $s > -t_0^{\frac{1}{2}}$ onde $T(s) = \left(s + t_0^{\frac{1}{2}} \right)^2$.

Exemplo 4.2.2
$$\begin{cases} (y'(t))^2 = y(t) \\ y(t_0) = y_0 > 0, t_0 > 0, y'(t_0) = z_0, z_0^2 = y_0 \end{cases}$$

Neste exemplo $G(t,y) = \sqrt{y}$.

Esta EAD tem índice 0 sobre

$$K = \{(t,y): -\infty < t < \infty, 0 < y < \infty\}$$

e o conjunto

$$\{(t,0): t \in \mathbb{R}\}$$

é formado por pontos singulares da EAD.

Uma das soluções clássicas é $y(t) = \left(\frac{(t-t_0)}{2} + y_0^{\frac{1}{2}} \right)^2$ para $t \in \mathbb{R}$.

O sistema diferencial associado é

$$\begin{cases} Y_1^2(s) - Y_0(s) = 0 \\ Y_1(s)T'(s) - Y_0'(s) = 0 \\ T(0) = t_0 \\ Y_0(0) = y_0 \\ Y_1(0) = z_0 \end{cases}$$

onde o conjunto de singularidades é

$$\{(T,0,0); T \in \mathbb{R}\}$$

e uma das soluções gerais é dada por:

$$c(s) = \left(s + t_0, \left(\frac{s}{2} + y_0^{\frac{1}{2}} \right)^2, \left(\frac{s}{2} + y_0^{\frac{1}{2}} \right) \right)$$

para $s \in \mathbb{R}$, onde $c(0) = \left(t_0, y_0, y_0^{\frac{1}{2}} \right) = (t_0, y_0, z_0)$ é a condição inicial. Como $T'(s) \neq 0$ nós temos que $c(s) = (t, y(t), y'(t))$ para qualquer s , onde $T(s) = s + t_0$.

Exemplo 4.2.3
$$\begin{cases} (y'(t))^2 = t \\ y(t_0) = y_0 > 0, t_0 > 0, y'(t_0) = z_0, z_0^2 = t_0 \end{cases}$$

Neste exemplo, $G(t,y) = \sqrt{t}$.

Esta EAD tem índice 0 sobre

$$K = \{(t,y): 0 < t < \infty, -\infty < y < \infty\}$$

e

$$\{(0,y): y \in \mathbb{R}\}$$

é o conjunto de singularidades.

A solução clássica é $y(t) = y_0 + \frac{2}{3} \left(t^{\frac{3}{2}} - t_0^{\frac{3}{2}} \right)$ para $t \in (0, \infty)$.

O sistema diferencial associado é

$$\begin{cases} Y_1^2(s) - T(s) = 0 \\ Y_1(s)T'(s) - Y_0'(s) = 0 \\ T(0) = t_0 \\ Y_0(0) = y_0 \\ Y_1(0) = z_0 \end{cases}$$

que não tem nenhuma singularidade e a solução geral é dada por:

$$c(s) = \left(\left(s + t_0^{\frac{1}{2}} \right)^2, y_0 + \frac{2}{3} \left(\left(s + t_0^{\frac{1}{2}} \right)^3 - t_0^{\frac{3}{2}} \right), s + t_0^{\frac{1}{2}} \right)$$

para $s \in \mathbb{R}$, onde $c(0) = \left(t_0, y_0, t_0^{\frac{1}{2}} \right) = (t_0, y_0, z_0)$ é a condição inicial. Novamente como

$T' \left(-t_0^{\frac{1}{2}} \right) = 0$ e $c(0) = (t_0, y_0, z_0)$ nós temos que $c(s) = (t, y(t), y'(t))$ para $s > -t_0^{\frac{1}{2}}$, onde

$$T(s) = \left(s + t_0^{\frac{1}{2}} \right)^2.$$

Teorema 4.2.2 Se DF é Lipschitz em uma vizinhança aberta de c_0 e $A(c_0)$ tem posto máximo, então existe $r > 0$ tal que o problema de valor inicial (4.2.5) tem solução única em $B = \{u: \|u - c_0\| \leq r\}$. A solução é única no sentido que o traço é único.

Demonstração:

Primeiramente, vamos mostrar que (4.2.6) é equivalente a um problema de valor inicial para uma equação diferencial ordinária, da forma:

$$\begin{cases} c'(s) = \xi(A(c(s))) \\ c(0) = c_0 \end{cases}$$

e posteriormente mostraremos que (4.2.6) é equivalente a (4.2.5).

Como $A(c_0)$ tem posto máximo, $\det(A(c_0)A'(c_0)) \neq 0$. Mas DF é uma função contínua em uma vizinhança aberta U de c_0 , então $\det(A(u)A'(u)) \neq 0$ é contínua também em U , assim existe uma vizinhança aberta $V \subset U$ de c_0 , tal que $\det(A(u)A'(u)) \neq 0$ ($A(u)$ tem posto máximo) para todo $u \in V$.

Seja $r > 0$, tal que $B = \{u \in \mathbb{R}^{2m+1}: \|u - c_0\| \leq r\} \subset V$.

Como $A(u)$ tem posto máximo em B e $A(u) \in \mathbb{R}^{2m \times (2m+1)}$, temos que o núcleo de $A(u)$ tem dimensão 1, assim existe apenas um vetor $\xi(A(u)) \in \mathbb{R}^{m+1}$, tal que $A(u)\xi(A(u)) = 0$, $\|\xi(A(u))\| = 1$ e $\det \begin{pmatrix} A(u) \\ \xi'(A(u)) \end{pmatrix} > 0$ ($\xi: \mathbb{R}^{2m \times (2m+1)} \rightarrow \mathbb{R}^{2m+1}$).

Agora, mostraremos que $\xi(A)$ é uma função localmente Lipschitz com respeito a A , para toda A com posto máximo.

Sejam $f(A, \xi) = \begin{pmatrix} A\xi \\ (\xi^t \xi - 1) \end{pmatrix}$ com (A, ξ) solução de $f(A, \xi) = 0$. O teorema da função implícita implica que ξ pode ser escrito como uma função de A , pois $\det(f_\xi(A, \xi)) = \det \begin{pmatrix} A \\ \xi^t \end{pmatrix} \neq 0$, ou seja $f_\xi(A, \xi)$ é não singular.

Temos também que $\begin{pmatrix} A \\ \xi^t \end{pmatrix}^{-1} = (A^+ \quad \xi)$ onde $A^+ = A'(AA')^{-1}$ é a inversa Moore-Penrose de A , pois:

$$\begin{pmatrix} A \\ \xi' \end{pmatrix} \begin{pmatrix} A^+ & \xi \end{pmatrix} = \begin{pmatrix} AA^+ & A\xi \\ \xi'A^+ & \xi'\xi \end{pmatrix} = \begin{pmatrix} AA^+(AA')^{-1} & 0 \\ (A\xi)'(AA')^{-1} & \|\xi\| \end{pmatrix} = I$$

além disso A^+ é uma função contínua de A , para A com posto máximo.

Mostraremos agora que $\xi'(A)X = A^+X\xi$, onde $X \in \mathbb{R}^{2m \times (2m+1)}$. Notemos que esta relação implica que ξ' é uma função contínua de A .

Fazendo $2m = n$, temos que $A \in \mathbb{R}^{n \times (n+1)}$ e $\xi: \mathbb{R}^{n \times (n+1)} \rightarrow \mathbb{R}^{n+1}$.

Sabemos que se $G = \begin{pmatrix} g_1 & \cdots & g_{n+1} \\ g_{n+2} & \cdots & g_{2n+2} \\ \vdots & & \vdots \\ g_{n(n+1)-n} & \cdots & g_{n(n+1)} \end{pmatrix}_{n \times (n+1)}$ e $L: \mathbb{R}^{n \times (n+1)} \rightarrow \mathbb{R}^{n(n+1)}$ definida por

$L(G) = \begin{pmatrix} G_1' \\ \vdots \\ G_n' \end{pmatrix}$, onde G_i é a i -ésima linha da matriz G , então L é um isomorfismo entre

$\mathbb{R}^{n \times (n+1)}$ e $\mathbb{R}^{n(n+1)}$, e lembramos que $\mathbb{R}^{ij}$ é o espaço das matrizes $(i \times j)$ com coeficientes reais e $\mathbb{R}^{ij}$ é o espaço real de dimensão ij .

Sejam então:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1(n+1)} \\ a_{21} & a_{22} & \cdots & a_{2(n+1)} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n(n+1)} \end{pmatrix}_{n \times (n+1)} \quad \text{e} \quad \xi = (\xi_1 \quad \xi_2 \quad \cdots \quad \xi_{n+1})'_{(n+1) \times 1}$$

$$\text{Assim } f(A, \xi) = \begin{pmatrix} a_{11}\xi_1 + a_{12}\xi_2 + \cdots + a_{1(n+1)}\xi_{n+1} \\ a_{21}\xi_1 + a_{22}\xi_2 + \cdots + a_{2(n+1)}\xi_{n+1} \\ \vdots \\ a_{n1}\xi_1 + a_{n2}\xi_2 + \cdots + a_{n(n+1)}\xi_{n+1} \\ \frac{\xi_1^2 + \xi_2^2 + \cdots + \xi_{n+1}^2}{2} - \frac{1}{2} \end{pmatrix}_{(n+1) \times 1}$$

Considerando $L(A)$ em vez de A , temos:

$$f_A(A, \xi) = \begin{pmatrix} \xi_1 & \xi_2 & \dots & \xi_{n+1} & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & \xi_1 & \xi_2 & \dots & \xi_{n+1} & 0 & \dots & \dots & 0 \\ & & \vdots & & & & & \ddots & & & & \\ & & \vdots & & & & & & \ddots & & & \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \xi_1 & \xi_2 & \dots & \xi_{n+1} \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \end{pmatrix}_{(n+1) \times n(n+1)}$$

$$\text{ou seja } f_A(A, \xi) = \begin{pmatrix} \xi^t & 0 & \dots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \xi^t \\ 0 & \dots & \dots & 0 \end{pmatrix}.$$

$$\text{Temos também que } f_\xi(A, \xi) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1(n+1)} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{n(n+1)} \\ \xi_1 & \xi_2 & \dots & \xi_{n+1} \end{pmatrix}_{(n+1) \times (n+1)} = \begin{pmatrix} A \\ \xi^t \end{pmatrix}.$$

$$\text{Agora sendo } A^+ = \begin{pmatrix} m_{11} & \dots & m_{1n} \\ m_{21} & \dots & m_{2n} \\ \vdots & \vdots & \vdots \\ m_{(n+1)1} & \dots & m_{(n+1)n} \end{pmatrix}_{(n+1) \times n}, \text{ como } \hat{f}_\xi^{-1}(A, \xi) = \begin{pmatrix} A^+ & \xi \end{pmatrix}, \text{ segue que}$$

$$\hat{f}_\xi^{-1}(A, \xi) = \begin{pmatrix} m_{11} & \dots & m_{1n} & \xi_1 \\ m_{21} & \dots & m_{2n} & \xi_2 \\ \vdots & \vdots & \vdots & \vdots \\ m_{(n+1)1} & \dots & m_{(n+1)n} & \xi_{n+1} \end{pmatrix}_{(n+1) \times (n+1)}$$

Assim, pelo teorema da função implícita:

$$\xi'(A) = -\hat{f}_\xi^{-1}(A, \xi) f_A(A, \xi) = - \begin{pmatrix} m_{11} & \dots & m_{1n} & \xi_1 \\ m_{21} & \dots & m_{2n} & \xi_2 \\ \vdots & \vdots & \vdots & \vdots \\ m_{(n+1)1} & \dots & m_{(n+1)n} & \xi_{n+1} \end{pmatrix} \begin{pmatrix} \xi^t(A) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \xi^t(A) \\ 0 & \dots & \dots & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} m_{11}\xi_1 & m_{11}\xi_2 & \cdots & m_{11}\xi_{n+1} & m_{12}\xi_1 & m_{12}\xi_2 & \cdots & m_{12}\xi_{n+1} & \cdots & m_{1n}\xi_1 & m_{1n}\xi_2 & \cdots & m_{1n}\xi_{n+1} \\ m_{21}\xi_1 & m_{21}\xi_2 & \cdots & m_{21}\xi_{n+1} & m_{22}\xi_1 & m_{22}\xi_2 & \cdots & m_{22}\xi_{n+1} & \cdots & m_{2n}\xi_1 & m_{2n}\xi_2 & \cdots & m_{2n}\xi_{n+1} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ m_{(n+1)1}\xi_1 & m_{(n+1)1}\xi_2 & \cdots & m_{(n+1)1}\xi_{n+1} & m_{(n+1)2}\xi_1 & m_{(n+1)2}\xi_2 & \cdots & m_{(n+1)2}\xi_{n+1} & \cdots & m_{(n+1)n}\xi_1 & m_{(n+1)n}\xi_2 & \cdots & m_{(n+1)n}\xi_{n+1} \end{pmatrix}$$

Seja $X = \begin{pmatrix} x_1 & \cdots & x_{n+1} \\ x_{n+2} & \cdots & x_{2n+2} \\ \vdots & & \vdots \\ x_{n(n+1)-n} & \cdots & x_{n(n+1)} \end{pmatrix}_{n \times (n+1)}$, e consideremos então, ao invés de X , o vetor

$L(X) \in \mathbb{R}^{n(n+1)}$, então:

$$\xi'(A)X = \begin{pmatrix} m_{11}x_1\xi_1 + \cdots + m_{11}x_{n+1}\xi_{n+1} + \cdots + m_{1n}x_{n(n+1)-n}\xi_1 + \cdots + m_{1n}x_{n(n+1)}\xi_{n+1} \\ m_{21}x_1\xi_1 + \cdots + m_{21}x_{n+1}\xi_{n+1} + \cdots + m_{2n}x_{n(n+1)-n}\xi_1 + \cdots + m_{2n}x_{n(n+1)}\xi_{n+1} \\ \vdots \\ m_{(n+1)1}x_1\xi_1 + \cdots + m_{(n+1)1}x_{n+1}\xi_{n+1} + \cdots + m_{(n+1)n}x_{n(n+1)-n}\xi_1 + \cdots + m_{(n+1)n}x_{n(n+1)}\xi_{n+1} \end{pmatrix}$$

$$= - \begin{pmatrix} m_{11}x_1 & \cdots & m_{11}x_{n+1} \\ m_{21}x_1 & \cdots & m_{21}x_{n+1} \\ \vdots & & \vdots \\ m_{(n+1)1}x_1 & \cdots & m_{(n+1)1}x_{n+1} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix} - \cdots - \begin{pmatrix} m_{1n}x_{n(n+1)-n} & \cdots & m_{1n}x_{n(n+1)} \\ m_{2n}x_{n(n+1)-n} & \cdots & m_{2n}x_{n(n+1)} \\ \vdots & & \vdots \\ m_{(n+1)n}x_{n(n+1)-n} & \cdots & m_{(n+1)n}x_{n(n+1)} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix}$$

$$= - \begin{pmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{(n+1)1} \end{pmatrix} (x_1 \ x_2 \ \cdots \ x_{n+1}) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix} - \cdots - \begin{pmatrix} m_{1n} \\ m_{2n} \\ \vdots \\ m_{(n+1)n} \end{pmatrix} (x_{(n+1)n-n} \ \cdots \ \cdots \ x_{n(n+1)}) \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix}$$

$$= - \left[\begin{pmatrix} m_{11} \\ m_{21} \\ \vdots \\ m_{(n+1)1} \end{pmatrix} (x_1 \ x_2 \ \cdots \ x_{n+1}) + \cdots + \begin{pmatrix} m_{1n} \\ m_{2n} \\ \vdots \\ m_{(n+1)n} \end{pmatrix} (x_{(n+1)n-n} \ \cdots \ \cdots \ x_{n(n+1)}) \right] \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix}$$

$$= - \begin{pmatrix} m_{11}x_1 + \cdots + m_{1n}x_{n(n+1)-n} & \cdots & \cdots & m_{11}x_{n+1} + \cdots + m_{1n}x_{n(n+1)} \\ m_{21}x_1 + \cdots + m_{2n}x_{n(n+1)-n} & & & m_{21}x_{n+1} + \cdots + m_{2n}x_{n(n+1)} \\ \vdots & & & \vdots \\ m_{(n+1)1}x_1 + \cdots + m_{(n+1)n}x_{n(n+1)-n} & \cdots & \cdots & m_{(n+1)1}x_{n+1} + \cdots + m_{(n+1)n}x_{n(n+1)} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n+1} \end{pmatrix}$$

$= -A^+X\xi$, onde agora estamos novamente considerando a matriz X .

Assim, concluímos que $\xi'(A)$ é uma função contínua de A .

Agora, se $c(s)$ é solução de $\begin{cases} c'(s) = \xi(A(c(s))) \\ c(0) = c_0 \end{cases}$, então $c(s)$ satisfaz (4.2.6), pois $c'(s) = \xi(A(c(s)))$, e também se $d(s)$ é outra solução de (4.2.6), podemos reparametrizar $d(s)$ por comprimento de arco e assim $\|d'(s)\| = 1$, e escolhendo $d(s)$ tal que $\det \begin{pmatrix} A \\ d' \end{pmatrix} > 0$ temos, agora, que $d(s)$ é solução de $\begin{cases} c'(s) = \xi(A(c(s))) \\ c(0) = c_0 \end{cases}$. Logo, o problema de valor inicial (4.2.6) é equivalente, no conjunto B , ao sistema:

$$\begin{cases} c'(s) = \xi(A(c(s))) \\ c(0) = c_0 \end{cases} \quad (4.2.7)$$

Temos que $A(v)$ é Lipschitz em B pois DF o é por hipótese e w é Lipschitz. O conjunto $\{A(v): v \in B\}$ é compacto e $\xi'(A)$ é contínua. Então, $\xi(A(v))$ é Lipschitz em B , como queríamos mostrar.

Agora, aplicando o teorema de Picard para (4.2.7) concluímos que este tem solução única definida em B . Assim (4.2.6) tem solução única em B , no sentido que o traço é único.

Para mostrar que (4.2.5) tem solução única em B observamos que toda solução de (4.2.6) é também solução de (4.2.5) pois $DF(c(s))c'(s)=0$ implica que $F(c(s))$ é constante e como $F(c_0) = 0$, temos que $F(c(s)) = 0$. Por outro lado toda solução de (4.2.5) é solução de (4.2.6), logo (4.2.5) e (4.2.6) são equivalentes, e assim (4.2.5) tem solução única, concluindo, assim, a demonstração do teorema 4.2.2.

Corolário 4.2.1 *Se DF é Lipschitz em uma vizinhança aberta de $c_0 = (t_0, a_0, a_1)$ e $A(t_0, a_0, a_1)$ tem posto máximo, então existe $r > 0$ tal que o problema de valor inicial (4.2.3) com $F(t_0, a_0, a_1) = 0$ tem uma única solução geral no conjunto:*

$$B = \{(T, Y_0, Y_1): \|(T, Y_0, Y_1) - (t_0, a_0, a_1)\| \leq r\}.$$

Note que a hipótese de posto máximo em $A(t_0, a_0, a_1)$ é essencial, como podemos ver no exemplo 4.2.2. Naquele exemplo

$$A(c(s)) = \begin{pmatrix} 0 & -1 & 2Y_1(s) \\ Y_1(s) & -1 & 0 \end{pmatrix}$$

não tem posto máximo quando $Y_1(s) = 0$, e se $c_0 = (t_0, 0, 0)$ então o exemplo 4.2.2 tem muitas soluções gerais, duas delas são dadas por

$$c(s) = (s + t_0, 0, 0)$$

e

$$c(s) = \left(s + t_0, \frac{s^2}{4}, \frac{s}{2} \right).$$

Se $F_{Y_1}(c_0)$ é não singular (isto é, a EAD tem índice 0) então $A(c_0)$ tem posto máximo, pois, sendo $c_0 = (t_0, a_0, a_1)$ então:

$$A(c_0) = \begin{pmatrix} F_T(c_0) & F_{Y_0}(c_0) & F_{Y_1}(c_0) \\ a_1 & -I & 0 \end{pmatrix}. \quad (4.2.8)$$

Agora se $F_{Y_1}(c_0)$ tem posto $(m-1)$ (isto é, a EAD tem índice maior que 0), $A(c_0)$ pode ainda ter posto máximo pois o vetor $\begin{pmatrix} F_T(c_0) \\ a_1 \end{pmatrix}$ pode recuperar o posto de $A(c_0)$.

No exemplo 4.2.3 temos que

$$A(c(s)) = \begin{pmatrix} -1 & 0 & 2Y_1(s) \\ Y_1(s) & -1 & 0 \end{pmatrix}$$

tem posto máximo mesmo que $F_{Y_1}(c(s)) = 2Y_1(s) = 0$.

Se $F_{Y_1}(c_0)$ tem posto menor que $(m-1)$, $A(c_0)$ não pode ter posto máximo. Usando mudanças de variáveis no domínio e na ordenação de F , podemos localmente escrever (2.2.1) na seguinte forma:

$$\begin{cases} f(t, y(t), z(t), y'(t)) = 0 \\ g(t, y(t), z(t)) = 0 \end{cases} \quad (4.2.9)$$

onde $f: \Omega \rightarrow \mathbb{R}^n$ e $g: \Omega \rightarrow \mathbb{R}^p$ são funções suficientemente diferenciáveis definidas no conjunto aberto conexo $\Omega \subset \mathbb{R}^{2n+p+1}$ e $n + p = m$.

Temos também para a EAD na forma (4.2.9), definições de solução análogas às anteriores:

Definição 4.2.6 Uma função $Y(t) = (y(t), z(t))^t$ com $y: I \rightarrow \mathbb{R}^n$ e $g: I \rightarrow \mathbb{R}^p$ é uma solução clássica da EAD (4.2.9) se $y \in C^1$, $z \in C^0$ em $I \subset \mathbb{R}$ e satisfaz (4.2.9).

Definição 4.2.7 A imagem de uma função $c: J \rightarrow R^{2n+p+1}$ é uma solução geral da EAD (4.2.9) se $c \in C^1$ em $J \subset R$ e satisfaz

$$\begin{cases} f(c(s)) = 0 \\ g(c(s)) = 0 \\ Y_1(s)T'(s) - Y_0(s) = 0 \end{cases} \quad (4.2.10)$$

para todo $s \in J$, onde $c(s) = (T(s), Y_0(s), Z_0(s), Y_1(s))$.

Sejam $w: \Omega \rightarrow R^N$ e $A: \Omega \rightarrow R^M$ onde $N = nx(2n+p+1)$ e $M = (2n+p)x(2n+p+1)$, definidas como:

$$w(T, Y_0, Z_0, Y_1) = \begin{pmatrix} Y_1 & -I_n & 0_{n \times p} & 0_n \end{pmatrix}_{n \times (2n+p+1)} \text{ com } \begin{cases} Y_1 \in R^{n \times 1} \\ I_n \in R^{n \times n} \\ 0_{n \times p} \in R^{n \times p} \\ 0_n \in R^{n \times n} \end{cases}$$

e

$$A(c(s)) = \begin{pmatrix} Df(c(s)) \\ Dg(c(s)) \\ w(c(s)) \end{pmatrix}_{(2n+p) \times (2n+p+1)}$$

Então os seguintes problemas de valor inicial

$$\begin{cases} f(c(s)) = 0 \\ g(c(s)) = 0 \\ w(c(s))c'(s) = 0 \\ c(0) = c_0 = (t_0, a_0, b_0, a_1) \end{cases} \quad (4.2.11)$$

e

$$\begin{cases} A(c(s))c'(s) = 0 \\ c(0) = c_0 = (t_0, a_0, b_0, a_1) \end{cases} \quad (4.2.12)$$

com $f(c_0) = 0$ e $g(c_0) = 0$ são equivalentes a (4.2.5) e (4.2.6), respectivamente.

Teorema 4.2.3 Considere uma solução geral de (4.2.9) dada por $c: J \rightarrow R^{2n+p+1}$. Se $T'(s) \neq 0$ em um intervalo $I \subset J$, então $Y_0(s) = y_0(T(s))$, $Y_1(s) = \frac{d}{dT} y_0(T(s))$, $Z_0(s) = z_0(T(s))$, em I , e $(y_0, z_0)'$ definida em $T(I)$ é uma solução clássica da EAD (4.2.9).

Teorema 4.2.4 Se DF e Dg são Lipschitz em uma vizinhança aberta de c_0 e $A(c_0)$ tem posto máximo, então existe $r > 0$ tal que o problema de valor inicial (4.2.11) tem solução única em $B = \{v: \|v - c_0\| \leq r\}$ (solução única no sentido de traço único).

Corolário 4.2.2 Se DF e Dg são Lipschitz em uma vizinhança aberta de $c_0 = (t_0, a_0, b_0, a_1)$ e $A(t_0, a_0, b_0, a_1)$ tem posto máximo, então existe $r > 0$ tal que o problema de valor inicial

$$\begin{cases} F(t, y(t), z(t), y'(t)) = 0 \\ g(t, y(t), z(t)) = 0 \\ y(t_0) = a_0 \\ y'(t_0) = a_1 \\ z(t_0) = b_0 \end{cases} \quad (4.2.13)$$

onde $\begin{cases} f(c_0) = 0 \\ g(c_0) = 0 \end{cases}$, tem uma única solução geral no conjunto:

$$B = \{(T, Y_0, Z_0, Y_1): \|(T, Y_0, Z_0, Y_1) - (t_0, a_0, b_0, a_1)\| \leq r\}.$$

As demonstrações dos teoremas 4.2.3 e 4.2.4 seguem as demonstrações dos teoremas 4.2.1 e 4.2.2 respectivamente, e por isto serão omitidas.

Note que se $f_{y'}(c_0)$ e $g_z(c_0)$ são não singulares, a EAD (4.2.9) tem índice 1 e $A(c_0)$ tem posto máximo, pois:

$$A(c(s)) = \begin{pmatrix} Df(c(s)) \\ Dg(c(s)) \\ w(c(s)) \end{pmatrix} = \begin{pmatrix} f_T & f_{Y_0} & f_{Z_0} & f_{Y_1} \\ g_T & g_{Y_0} & g_{Z_0} & 0_{p \times n} \\ Y_1 & -I_n & 0_{n \times p} & 0_{n \times n} \end{pmatrix}, \text{ e observe que } \begin{cases} f_{Y_0} = f_y \\ f_{Y_1} = f_{y'} \\ g_{Z_0} = g_z \end{cases}.$$

Mas, se $g_z(c_0)$ é singular a EAD (4.2.9) tem índice maior que 1 e $A(c_0)$ poderá ainda ter posto máximo, pois a primeira coluna poderá recuperar o seu posto. Para ilustrar este fato, consideremos o seguinte exemplo:

$$\begin{cases} (y'(t))^2 + (z(t))^2 = 1 \\ (y(t))^2 + t^2 = \frac{1}{4} \\ y'(0) = 1 \\ y(0) = \frac{1}{2} \\ z(0) = 0 \end{cases}$$

onde temos que $c_0 = \left(0, \frac{1}{2}, 0, 1\right)$ e $A(c_0) = \begin{pmatrix} 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{pmatrix}$ tem posto máximo, e observe que $g_z(c_0) = 0$ é singular.

4.3 Método numérico

Nesta subseção, explicaremos como resolver numericamente uma EAD, com base na teoria apresentada na seção 4.2.

Lembramos que, se desejamos resolver numericamente um sistema do tipo

$$F(x) = 0$$

em vez de usarmos o método de Newton,

$$x^{k+1} = x^k - (DF(x^k))^{-1} F(x^k)$$

podemos usar o método de Newton modificado,

$$x^{k+1} = x^k - (DF(x^0))^{-1} F(x^k)$$

sem perda de convergência.

Dada a EAD (2.2.1) os métodos numéricos usualmente utilizados (como o método BDF, por exemplo) encontram uma aproximação para a solução clássica $y(t)$, o que faremos através desta nova classe de métodos é encontrar uma aproximação para a solução geral $c(s)$ da EAD (2.2.1), utilizando um método PC (Previsor - Corretor).

Consideremos então o sistema diferencial (4.2.4) associado à EAD (2.2.1) e o PVI (4.2.5).

Pelo teorema 4.2.2 este PVI tem localmente uma única solução.

Por diferenciação obteve-se de (4.2.5) um outro PVI equivalente, dado por (4.2.6).

Na demonstração do teorema 4.2.2 mostramos que este PVI é equivalente ao PVI para uma equação diferencial ordinária da forma (4.2.7).

Resolveremos então, o PVI (4.2.7). Antes porém, necessitamos encontrar o vetor $\xi(A)$, que pode ser obtido da seguinte maneira.

Suponhamos que a decomposição QR de A^t

$$A^t = Q \begin{pmatrix} R \\ 0^t \end{pmatrix}$$

é dada onde Q é uma matriz $(2m+1) \times (2m+1)$ ortogonal, isto é $Q^t Q = I$ e R é uma matriz triangular superior $(2m) \times (2m)$, não singular, isto é $R_{ij} = 0$ para $i > j$ e $R_{ii} \neq 0$. Assim se z denota a última coluna de Q então

$$Az = \begin{pmatrix} R^t & 0 \end{pmatrix} Q^t z = \begin{pmatrix} R^t & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = 0$$

pois as colunas de Q são ortonormais. Ainda nos resta encontrar o sinal de z tal que $\det \begin{pmatrix} A \\ z^t \end{pmatrix} > 0$. Note que:

$$\begin{pmatrix} A^t & z \end{pmatrix} = Q \begin{bmatrix} R & 0 \\ 0^t & 1 \end{bmatrix}$$

implica que:

$$\det \begin{pmatrix} A \\ z^t \end{pmatrix} = \det \begin{pmatrix} A^t & z \end{pmatrix} = \det Q \det R.$$

Assim, $\xi(A) = \pm z$, concordando com o sinal de $\det Q \det R$, se é positivo ou negativo.

Observação: $\det R$ é o produto dos elementos diagonais de R , e seu sinal é facilmente determinado. O sinal de $\det Q$ também é facilmente obtido, pois sabemos que como Q é ortogonal, então $\det Q = \pm 1$. Por exemplo, se rotações de Givens são usadas para obter a decomposição QR de A^t então $\det Q = 1$; ou, se usamos reflexões de Householder, cada reflexão muda o sinal, e então o sinal do $\det Q$ é igual a $(-1)^p$ onde p é o número de

reflexões que são envolvidas na fatoração de A' pelo método de Householder. Em ambos os casos não existe qualquer custo computacional adicional.

Como vimos na seção 4.2, $A(c(s)) \in \mathbb{R}^{(2m) \times (2m+1)}$ e assim o custo do processo acima para determinar $\xi(A)$ é aproximadamente $16m^3$ para cada passo. No entanto, este cálculo pode ter o custo minimizado, pois, sendo $c(s)=(T(s), Y(s), Z(s))$ temos que o PVI (4.2.6) se torna:

$$\begin{cases} F_T T' + F_Y Y' + F_Z Z' = 0 \\ ZT' - Y' = 0 \end{cases}$$

e assim $c'(s)$ pode ser obtido resolvendo o sistema

$$(F_T + F_Y Z) T' + F_Z Z' = (F_T + F_Y Z \quad F_Z) \begin{pmatrix} T' \\ Z' \end{pmatrix} = 0$$

analogamente como fizemos acima para calcular $\xi(A)$, e fazendo $Y' = ZT'$. Assim,

$$\xi(A) = \pm \frac{c'(s)}{\|c'(s)\|}.$$

Ainda mais, o sinal de $\xi(A)$ pode ser escolhido arbitrariamente, uma vez que a integração numérica é feita em ambas as direções da condição inicial. Observe ainda que, como a matriz

$$(F_T + F_Y Z \quad F_Z)$$

a ser decomposta na forma QR (entre outras maneiras) na resolução do sistema acima tem dimensão $m \times (m+1)$ o custo de tal decomposição é aproximadamente $2m^3$, que agora é o custo aproximado do cálculo de $\xi(A)$.

Assim, o cálculo de $\xi(A)$ está facilmente resolvido e assim podemos resolver numericamente o PVI (4.2.7), por exemplo, utilizando o método de Euler explícito:

$$\frac{c_n - c_{n-1}}{h} = \xi(A(c_{n-1}))$$

ou seja,

$$c_n = c_{n-1} + h\xi(A(c_{n-1}))$$

que é o passo preditor.

O passo corretor é feito levando-se em conta o sistema diferencial (4.2.4), utilizando, por exemplo, o método de Euler implícito, isto é:

$$\begin{cases} F(c_n) = 0 \\ w(c_n)(c_n - c_{n-1}) = 0 \end{cases}$$

ou seja:

$$\begin{cases} F(T_n, Y_n, Z_n) = 0 \\ Z_n(T_n - T_{n-1}) - (Y_n - Y_{n-1}) = 0 \end{cases}$$

que pode ser resolvido, por exemplo, pelo método de Newton modificado. Isto nos dá:

$$c_n^{k+1} = c_n^k - (DB_n^0)^+ B_n^k,$$

onde $B_n^k = \begin{bmatrix} F(c_n^k) \\ w(c_n^k)(c_n^k - c_{n-1}) \end{bmatrix}$, $DB_n^0 = \begin{bmatrix} F_T^n & F_Y^n & F_Z^n \\ Z_n^0 & -I & T_n^0 - T_{n-1}^0 \end{bmatrix}$, com $F_\theta^n = F_\theta(T_n^0, Y_n^0, Z_n^0)$ para $\theta = T, Y, Z$.

Portanto, temos que calcular a inversa de Moore-Penrose de DB_n^0 , denotada por $(DB_n^0)^+$, e para isto devemos ter DB_n^0 com posto máximo, o que é garantido pelo resultado a seguir.

Temos que

$$DB_n^0 = A_n^0 - E_n^0 = \begin{bmatrix} F_T^n & F_Y^n & F_Z^n \\ Z_n^0 & -I & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & T_{n-1}^0 - T_n^0 \end{bmatrix}$$

onde $A_n^0 = A(c_n^0)$ tem posto máximo, e podemos exigir também que $\|(A_n^0)^+ \| \|E_n^0\| < 1$ para h suficientemente pequeno. Então DB_n^0 tem posto máximo e

$$\|DB_n^0\| \leq \frac{\|(A_n^0)^+\|}{1 - \|(A_n^0)^+\| \|E_n^0\|}$$

cuja demonstração pode ser encontrada em [Allgower/Georg - 90].

Assim, supondo que a decomposição QR de $(DB_n^0)^t$

$$(DB_n^0)^t = Q \begin{pmatrix} R \\ 0^t \end{pmatrix}$$

é dada, temos que $DB_n^0 = \begin{pmatrix} R^t & 0 \end{pmatrix} Q^t$, e como sabemos que

$$(DB_n^0)^+ = (DB_n^0)^t (DB_n^0 (DB_n^0)^t)^{-1}$$

segue que

$$(DB_n^0)^+ = Q \begin{pmatrix} R \\ 0^t \end{pmatrix} \left[\begin{pmatrix} R^t & 0 \end{pmatrix} Q^t Q \begin{pmatrix} R \\ 0^t \end{pmatrix} \right]^{-1} = Q \begin{pmatrix} R \\ 0^t \end{pmatrix} [R^t R]^{-1} = Q \begin{pmatrix} R \\ 0^t \end{pmatrix} R^{-1} (R^t)^{-1} = Q \begin{pmatrix} (R^t)^{-1} \\ 0 \end{pmatrix}.$$

Logicamente, nós não inverteremos R^t , calcularemos $\theta = Q(DB_n^0)^+ B_n^k$ resolvendo $R^t \theta = B_n^k$, e assim, $\theta = Q \begin{pmatrix} \theta \\ 0 \end{pmatrix}$.

Assim, temos que o custo de resolução pelo método de Newton modificado do sistema

$$\begin{cases} F(T_n, Y_n, Z_n) = 0 \\ Z_n(T_n - T_{n-1}) - (Y_n - Y_{n-1}) = 0 \end{cases} \quad (4.3.1)$$

é aproximadamente $(16m^3 + 4r_1 m^2)$, onde r_1 é o número de iterações do método de Newton modificado. No entanto, este custo também pode ser minimizado, pois da segunda equação do sistema (4.3.1) temos

$$Y_n = Y_{n-1} + Z_n(T_n - T_{n-1})$$

e assim, substituindo isto na primeira equação, obtemos o sistema equivalente,

$$H(T_n, Z_n) = F(T_n, Y_{n-1} + Z_n(T_n - T_{n-1}), Z_n) = 0$$

que resolveremos pelo método de Newton modificado.

Para este sistema a matriz a ser decomposta na forma QR é

$$DH(T_n^0, Z_n^0) = \begin{pmatrix} F_T + F_Y Z_n^0 & F_Y(T_n^0 - T_{n-1}) + F_Z \end{pmatrix} \in \mathbb{R}^{m \times (m+1)}.$$

Assim, o custo é aproximadamente $(2m^3 + r_2 m^2)$, onde r_2 é o número de iterações do método de Newton modificado para este sistema.

Gostaríamos de ressaltar que o passo corretor é essencial, uma vez que se resolvermos somente o PVI (4.2.7), ou seja, realizarmos somente o passo preditor, encontraremos uma solução geral $c(s)$, porém poderemos ter problemas de estabilidade, como mostra a ilustração abaixo, onde a EAD é dada por $(y(t))^2 + t^2 = 1$:

usando somente o passo preditor

usando o par preditor - corretor

5 RESULTADOS NUMÉRICOS

5.1 Introdução

Nesta seção, apresentaremos alguns resultados numéricos obtidos pelas duas classes de métodos numéricos estudadas nas seções anteriores. Ressaltamos que ambos os programas utilizados, respectivamente para a primeira classe numérica (métodos BDF), que foi apresentada no capítulo 3 e, para a segunda classe numérica, que foi apresentada no capítulo 4, utilizam dois métodos bastante conhecidos e simples, o método de Euler e o método de Newton modificado para resolução de sistemas.

Para as duas classes numéricas os programas utilizam as mesmas estratégias de controle de passo, tanto no método de Euler quanto no método de Newton modificado. A seguir, descreveremos estas estratégias:

- Controle de passo no método de Euler:

Seja

$$C = A_n^0 = F_y \left(t_n^0, y_n^0, \frac{y_n^0 - y_{n-1}^0}{h} \right) + \frac{1}{h} F_y \left(t_n^0, y_n^0, \frac{y_n^0 - y_{n-1}^0}{h} \right)$$

para a primeira classe numérica, ou

$$C = DH_n^0 = DH(T_n^0, Z_n^0) = \begin{pmatrix} H_T(T_n^0, Z_n^0) & H_Z(T_n^0, Z_n^0) \end{pmatrix}$$

para a segunda classe numérica, onde H foi dada na seção 4.3.

Sejam também, CD_{MAX} a condição máxima aceitável, h_{MIN} e h_{MAX} os tamanhos mínimo e máximo respectivamente, aceitáveis para o passo h e COND a condição da matriz C.

O controle é então feito na predição da seguinte maneira:

- se $COND > CD_{MAX}$ então diminua o tamanho do passo h.
- se $h < h_{MIN}$ então FIM.
- se COND é boa então aumente o tamanho do passo h (respeitando o limite $h \leq h_{MAX}$).

- Controle de passo através do método de Newton:

Seja

$$\text{DIST} = \|y_n^{k+1} - y_n^k\|$$

para a primeira classe numérica, ou

$$\text{DIST} = \|c_n^{k+1} - c_n^k\|$$

para a segunda classe numérica. Sejam também, D_{MAX} a distância máxima aceitável entre y_n^k e y_n^{k+1} para a primeira classe numérica e entre c_n^k e c_n^{k+1} para a segunda classe numérica, e TOL uma tolerância estabelecida.

O controle é então feito na correção da seguinte maneira:

- se $\text{DIST} > D_{\text{MAX}}$ então retorne ao preditor e diminua o tamanho do passo h .
- se $\text{DIST} < \text{TOL}$ então aumente o tamanho do passo h (respeitando o limite $h \leq h_{\text{MAX}}$).

5.2 Exemplos numéricos

Observação: Os valores expressos nas colunas erro(w) e erro(w') são aproximações para o erro cometido no cálculo de $w(t)$ e $w'(t)$ respectivamente, para $w = y, z$, nos pontos em que a solução clássica está bem definida. Em negrito destacamos a condição inicial. Além dos apresentados a seguir, outros exemplos que não constam neste trabalho, foram analisados.

$$\textbf{Exemplo 5.2.1} \quad \begin{cases} (y'(t))^2 = t \\ y(1) = 0.33333 \\ y'(1) = 1 \end{cases}$$

Este exemplo é uma EAD de índice 0, onde $t = 0$ é uma singularidade somente para $F(t, y, y') = 0$, como foi visto no exemplo 4.2.3.

Solução obtida pela primeira classe numérica:

t	y(t)	erro(y)	y'(t)	erro(y')
.	.	.	.	.
.	.	.	.	.
-0.0088291	-0.3297755	.	-0.1792389	.
-0.0088246	-0.3297763	.	0.0231341	.
.	.	.	.	.
.	.	.	.	.
-0.0074655	-0.3297537	.	-2.2782609	.
-0.0074652	-0.3297543	.	-4.5581599	.
-0.0074651	-0.3297549	.	-9.1171387	.
.	.	.	.	.
.	.	.	.	.
-0.0001941	-0.3297608	.	-0.0059080	.
-0.0001220	-0.3297612	.	0.0146152	.
-0.0000860	-0.3297607	.	0.0329190	.
-0.0000680	-0.3297601	.	-0.0012909	.
0.0000761	-0.3297603	10^{-3}	0.0105030	10^{-3}
0.0005927	-0.3297548	10^{-3}	0.0246592	10^{-4}
.	.	.	.	.
.	.	.	.	.
0.6050000	-0.0185236	10^{-3}	0.7778588	10^{-5}
0.6150000	-0.0107450	10^{-3}	0.7842600	10^{-5}
0.6250000	-0.0029024	10^{-3}	0.7906094	10^{-5}
0.6350000	0.0050036	10^{-4}	0.7969082	10^{-5}
0.6450000	0.0129726	10^{-4}	0.8031577	10^{-5}
.	.	.	.	.
.	.	.	.	.
0.9850000	0.3184175	10^{-5}	0.9924969	10^{-5}
0.9950000	0.3283424	10^{-6}	0.9975031	10^{-6}
1.0000000	0.3333300	0	1.0000000	0
1.0050000	0.3383425	10^{-6}	1.0025031	10^{-6}
1.0150000	0.3484174	10^{-5}	1.0074968	10^{-5}
1.0250000	0.3585419	10^{-5}	1.0124472	10^{-5}
1.0350000	0.3687156	10^{-5}	1.0173736	10^{-5}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

5.7249999	8.8057893	10^{-3}	2.3927015	10^{-6}
5.7349999	8.8297372	10^{-3}	2.3947903	10^{-6}
5.7449999	8.8537060	10^{-3}	2.3968773	10^{-6}

·
·
·

·
·
·

·
·
·

·
·
·

Solução obtida pela segunda classe numérica:

t	y(t)	erro(y)	y'(t)	erro(y')
·	·		·	
·	·		·	
·	·		·	
0.1271359	-0.3657196		-0.3565613	
0.1214881	-0.3637510		-0.3485514	
0.1159193	-0.3618550		-0.3404692	
·	·		·	
·	·		·	
·	·		·	
0.0092695	-0.3366173		-0.0962785	
0.0074735	-0.3364620		-0.0864496	
0.0058656	-0.3363388		-0.0765878	
·	·		·	
·	·		·	
·	·		·	
0.0007241	-0.3360787		-0.0269102	
0.0002865	-0.3360713		-0.0169273	
0.0000481	-0.3360966		-0.0069376	
0.0000093	-0.3360697	10^{-3}	0.0030547	10^{-14}
0.0001701	-0.3360676	10^{-3}	0.0130460	10^{-15}
·	·		·	
·	·		·	
·	·		·	
·	·		·	
0.6054257	-0.0200448	10^{-4}	0.7780910	10^{-12}
0.6124562	-0.0145428	10^{-4}	0.7825958	10^{-12}
0.6194826	-0.0090125	10^{-4}	0.7870721	10^{-12}
0.6265048	-0.0034543	10^{-4}	0.7915205	10^{-12}
0.6335226	0.0021313	10^{-4}	0.7959413	10^{-12}
0.6405358	0.0077443	10^{-4}	0.8003348	10^{-12}
0.6475443	0.0133840	10^{-4}	0.8047014	10^{-12}
·	·		·	
·	·		·	
·	·		·	

0.9833214	0.3166960	10^{-5}	0.9916256	10^{-13}
0.9899978	0.3233390	10^{-5}	0.9949863	10^{-13}
0.9966669	0.3299969	10^{-6}	0.9983320	10^{-15}
1.0000000	0.3333300	0	1.0000000	0
1.0033336	0.3366636	10^{-6}	1.0016654	10^{-15}
1.0099977	0.3433388	10^{-5}	1.0049864	10^{-13}
1.0166544	0.3500287	10^{-5}	1.0082928	10^{-13}
.	.	.	.	.
5.7210617	8.7858140	10^{-3}	2.3918741	10^{-15}
5.7249065	8.7950103	10^{-3}	2.3926776	10^{-15}
5.7287502	8.8042071	10^{-3}	2.3934800	10^{-15}
5.7325929	8.8134044	10^{-3}	2.3942833	10^{-15}
5.7364344	8.8226021	10^{-3}	2.3950854	10^{-15}
5.7402749	8.8318004	10^{-3}	2.3958870	10^{-15}
5.7441143	8.8409992	10^{-3}	2.3966882	10^{-15}
5.7479526	8.8501984	10^{-3}	2.3974882	10^{-15}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Exemplo 5.2.2
$$\begin{cases} (y'(t))^2 = y(t) \\ y(1) = 4 \\ y'(1) = 2 \end{cases}$$

Este exemplo é uma EAD de índice 0, onde $y = 0$ é uma singularidade para $F(t, y, y') = 0$ e para o sistema diferencial associado $\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \end{cases}$, como foi visto no exemplo 4.2.2.

Solução obtida pela primeira classe numérica:

t	y(t)	erro(y)	y'(t)	erro(y')
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
-7.4245777	4.9033808		-2.2143691	
-7.4145777	4.8812371		-2.2093634	
-7.4045777	4.8591434		-2.2043578	
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
-3.0212499	0.0000051		-0.0033127	
-3.0199999	0.0000010		-0.0002737	
-3.0174999	0.0000003		-0.0002737	
-3.0162499	0.0000000	10^{-5}	0.0007237	10^{-3}
-3.0149999	0.0000008	10^{-5}	0.0007237	10^{-3}
-3.0137499	0.0000017	10^{-5}	0.0011899	10^{-3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
-2.9749999	0.0003840	10^{-4}	0.0196806	10^{-3}
-2.9724999	0.0004332	10^{-4}	0.0208929	10^{-3}
-2.9699999	0.0004854	10^{-4}	0.0221074	10^{-3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0.9850000	3.9700872	10^{-5}	1.9925203	10^{-4}
0.9950000	3.9900124	10^{-6}	1.9975046	10^{-5}
1.0000000	4.0000000	0	2.0000000	0
1.0050000	4.0100125	10^{-6}	2.0025046	10^{-6}
1.0150000	4.0300877	10^{-5}	2.0075202	10^{-5}
1.0250000	4.0502129	10^{-5}	2.0125264	10^{-5}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
4.2749999	13.2423241	10^{-2}	3.6390072	10^{-3}
4.2849999	13.2787642	10^{-2}	3.6440107	10^{-3}
4.2949999	13.3152544	10^{-2}	3.6490141	10^{-3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10.9749999	48.8689900	10^{-2}	6.9906394	10^{-3}
10.9849999	48.9389946	10^{-2}	6.9956412	10^{-3}
10.9949999	49.0089528	10^{-2}	7.0006430	10^{-3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Solução obtida pela segunda classe numérica:

t	y(t)	erro(y)	y'(t)	erro(y')
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
-2.9733704	0.0000000	10 ⁻⁴	0.0002655	10 ⁻²
-2.9728985	0.0000003	10 ⁻⁴	0.0005908	10 ⁻²
-2.9721923	0.0000010	10 ⁻⁴	0.0010417	10 ⁻²
-2.9695582	0.0000089	10 ⁻⁴	0.0029954	10 ⁻²
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0.9847087	3.9694604	10 ⁻⁵	1.9923504	10 ⁻⁶
0.9890820	3.9781832	10 ⁻⁵	1.9945383	10 ⁻⁶
0.9934517	3.9869083	10 ⁻⁶	1.9967244	10 ⁻⁶
0.9978178	3.9956356	10 ⁻⁶	1.9989086	10 ⁻⁷
1.0000000	4.0000000	0	2.0000000	0
1.0021822	4.0043644	10 ⁻⁶	2.0010908	10 ⁻⁷
1.0065448	4.0130944	10 ⁻⁶	2.0032709	10 ⁻⁶
1.0109038	4.0218268	10 ⁻⁵	2.0054492	10 ⁻⁶
1.0152593	4.0305613	10 ⁻⁵	2.0076258	10 ⁻⁶
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
4.2830082	13.2567929	10 ⁻³	3.6409879	10 ⁻⁴
4.2856338	13.2663524	10 ⁻³	3.6423004	10 ⁻⁴
4.2882584	13.2759122	10 ⁻³	3.6436125	10 ⁻⁴
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Exemplo 5.2.3
$$\begin{cases} t^2 + (y'(t))^2 = 0.25 \\ (y(t))^2 + (z(t))^2 = 1 \\ y(0) = 0 \\ z(0) = 1 \\ y'(0) = 0.5 \end{cases}$$

Este exemplo é uma EAD de índice 1, onde $t = \pm 0.5$ são singularidades para $F(t,y,y') = 0$ e para o sistema diferencial associado $\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \end{cases}$

Solução obtida pela primeira classe numérica:

t	y(t)	z(t)	erro(y,z)	y'(t)	erro(y')
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
-0.5031928	-0.1962009	0.9805639		-0.0478155	
-0.5031831	-0.1962014	0.9805642		0.0461442	
-0.5031635	-0.1962005	0.9805635		-0.0467676	
-0.5031538	-0.1962010	0.9805634		0.0264499	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
-0.5001953	-0.1961966	0.9805647		-0.0067991	
-0.5001171	-0.1961971	0.9805647		0.0141303	
-0.5000781	-0.1961966	0.9805645		-0.0036703	
-0.4999218	-0.1961971	0.9805643	10 ⁻⁴	0.0094230	10 ⁻⁴
-0.4998437	-0.1961964	0.9805652	10 ⁻⁴	0.0126897	10 ⁻⁴
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
-0.0150000	-0.0074976	0.9999718	10 ⁻⁶	0.4997750	10 ⁻⁸
-0.0050000	-0.0024998	0.9999968	10 ⁻⁸	0.4999750	10 ⁻¹⁰
0.0000000	0.0000000	1.0000000	0	0.5000000	0
0.0050000	0.0024998	0.9999968	10 ⁻⁸	0.4999750	10 ⁻¹⁰
0.0150000	0.0074976	0.9999718	10 ⁻⁶	0.4997750	10 ⁻⁸
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
0.4997917	0.1961893	0.9805645	10 ⁻⁴	0.0154607	10 ⁻³
0.5000357	0.1961899	0.9805662		0.0022688	
0.5000686	0.1961894	0.9805657		-0.0140039	
0.5001345	0.1961893	0.9805663		-0.0021957	
0.5001511	0.1961898	0.9805661		0.0333085	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
0.5036211	0.1961746	0.9805687		0.0109145	
0.5036252	0.1961739	0.9805688		-0.1612201	
0.5036334	0.1961736	0.9805689		-0.0430939	
0.5036449	0.1961747	0.9805689		0.0685680	
0.5036829	0.1961740	0.9805682		-0.0200322	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Solução obtida pela segunda classe numérica:

t	y(t)	z(t)	erro(y,z)	y'(t)	erro(y')
.	.	.	.	.	.
.	.	.	.	.	.
0.4997237	-0.5902034	0.8072545		0.0166187	
0.4999560	-0.5902019	0.8072556		0.0066288	
0.4999886	-0.5902020	0.8072555		-0.0033635	
0.4998216	-0.5901998	0.8072571		-0.0133546	
.	.	.	.	.	.
.	.	.	.	.	.
-0.4998279	-0.1975389	0.9802950		-0.0131137	
-0.4999902	-0.1975384	0.9802951		-0.0031225	
-0.4999528	-0.1975382	0.9802951	10 ⁻³	0.0068698	10 ⁻¹²
-0.4997156	-0.1975342	0.9802959	10 ⁻³	0.0168595	10 ⁻¹³
-0.4992789	-0.1975224	0.9802983	10 ⁻³	0.0268424	10 ⁻¹³
.	.	.	.	.	.
.	.	.	.	.	.
-0.0223348	-0.0111725	0.9999375	10 ⁻⁶	0.4995002	10 ⁻¹⁴
-0.0134128	-0.0067062	0.9999775	10 ⁻⁷	0.4998200	10 ⁻¹⁵
-0.0044719	-0.0022359	0.9999975	10 ⁻⁸	0.4999800	10 ⁻¹⁴
0.0000000	0.0000000	1.0000000	0	0.5000000	0
0.0044719	0.0022359	0.9999975	10 ⁻⁸	0.4999800	10 ⁻¹⁴
0.0134128	0.0067062	0.9999775	10 ⁻⁷	0.4998200	10 ⁻¹⁵
0.0223487	0.0111725	0.9999375	10 ⁻⁶	0.4995002	10 ⁻¹⁴
.	.	.	.	.	.
.	.	.	.	.	.
0.4992789	0.1975224	0.9802983	10 ⁻³	0.0268424	10 ⁻¹³
0.4997156	0.1975342	0.9802959	10 ⁻³	0.0168859	10 ⁻¹³
0.4999528	0.1975382	0.9802951	10 ⁻³	0.0068698	10 ⁻¹²
0.4999902	0.1975384	0.9802951		-0.0031225	
0.4998279	0.1975389	0.9802950		-0.0131137	
0.4994661	0.1975437	0.9802940		-0.0230997	
.	.	.	.	.	.
.	.	.	.	.	.
-0.4998216	0.5901998	0.8072571		-0.0133546	
-0.4999886	0.5902020	0.8072555		-0.0033635	
-0.4999560	0.5902019	0.8072556		0.0066288	
-0.4997237	0.5902034	0.8072545		0.0166187	
.	.	.	.	.	.
.	.	.	.	.	.

-0.0137710	0.7784871	0.6276605	0.4998103
-0.0059637	0.7823893	0.6227895	0.4999644
0.0018195	0.7862807	0.6178693	0.4999966
0.0095778	0.7901598	0.6129007	0.4999082
.	.	.	.
.	.	.	.
.	.	.	.

6 CONCLUSÃO

Nesta seção concluiremos nosso trabalho, apresentando uma breve comparação entre as duas classes numéricas estudadas.

Vimos que o custo de um passo da primeira classe numérica (os métodos BDF) é aproximadamente $(2m^3 + r_1 m^2)$, onde r_1 é o número de iterações do método de Newton modificado, enquanto que o custo de um passo da segunda classe numérica é aproximadamente $(4m^3 + r_2 m^2)$, onde r_2 é o número de iterações do método de Newton modificado. Portanto o custo da segunda classe numérica é, aproximadamente, 2 vezes maior que o custo da primeira classe numérica, e lembramos que m é a dimensão de F .

De um modo geral, observando os exemplos apresentados na seção 5.2 vemos que os erros cometidos na avaliação de $y(t)$ e de $y'(t)$ principalmente, pela segunda classe numérica, são inferiores aos cometidos pela primeira classe numérica.

Destacamos, também, outro ponto de grande importância, que se refere às EAD que apresentam singularidades somente para $F(t, y, y') = 0$. Neste caso, a segunda classe numérica apresenta um melhor desempenho, pois enquanto que a primeira classe numérica nas vizinhanças de tais pontos apresenta problemas, estes mesmos pontos não constituem singularidades para o sistema diferencial associado à EAD $F(t, y, y') = 0$,

$$\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \end{cases}$$

que é resolvido pela segunda classe numérica.

Outro problema que ocorre com a primeira classe numérica, dependendo dos parâmetros de implementação, é que ela pode avaliar a solução em pontos onde esta não faz sentido (isto pode ser visto claramente no exemplo 5.2.1).

Do ponto de vista de convergência, temos que ambas as classes são convergentes, porém, espera-se que em um futuro não distante prove-se que a convergência da segunda classe numérica é melhor, quando comparada com a convergência da primeira classe, que estudamos no capítulo 3.

Concluiremos nosso trabalho lembrando que em regiões singulares para o sistema diferencial

$$\begin{cases} F(c(s)) = 0 \\ w(c(s))c'(s) = 0 \end{cases}$$

ambas as classes numéricas apresentam problemas, pois tais singularidades implicam em singularidades para a EAD $F(t,y,y') = 0$.

7 BIBLIOGRAFIA

- [Allgower/Georg - 90] - Allgower E. L. & Georg K., *Numerical Continuation Methods - An Introduction* - Springer - Verlag - 1990.
- [Brenan/Campbell/Petzold - 89] - Brenan K. E. & Campbell S. L. & Petzold L. R., *Numerical Solution of Initial - Value Problems in Differential - Algebraic Equations* - North - Holland - 1989.
- [Campbell - 87] - Campbell S. L., *A General Form for Solvable Linear Time Varying Singular Systems of Differential Equations*, SIAM J. Math. Anal., 18 (1987), 1101 - 1115.
- [Castelo/Tavares - 94] - Castelo Filho A. & Tavares G., *Differential - Algebraic Equations with Singularities* - 1994 (a publicar).
- [Gear - 71] - Gear C. W., *The Simultaneous Numerical Solution of Differential - Algebraic Equations*, IEEE Trans. Circuit Theory, CT - 18 (1971), 89 - 95.
- [Gear - 88] - Gear C. W., *Differential - Algebraic Equation Index Transformations*, SIAM J. Sci. Stat. Comp., 9 (1988), 39 - 47.
- [Gear/Petzold - 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.
- [Hairer/Norsett/Wanner - 80] - Hairer E. & Norsett S. P. & Wanner G., *Solving Ordinary Differential Equations I - Nonstiff Problems* - Springer - Verlag - 1980.
- [Lambert - 73] - Lambert J. D., *Computational Methods in Ordinary Differential Equations*, Wiley, 1973.
- [Pires - 93] - Pires A. A., *Métodos Numéricos para a Solução de Equações Algébrico - Diferenciais*, Exame Geral de Qualificação, Instituto de Ciências Matemáticas de São Carlos da Universidade de São Paulo, São Carlos, 1993.