# Numerical Solution of Linear Ordinary Differential Equations and Differential-Algebraic Equations by Spectral Methods

## Thesis


**DEPARTMENT OF MATHEMATICS AND COMPUTING**

**OPEN UNIVERSITY**

# Numerical Solution of Linear Ordinary Differential Equations and Differential-Algebraic Equations by Spectral Methods

**by**

**Masoud Saravi**

**This thesis is submitted for the degree of PhD in Numerical Analysis**

**September 2007**

In the name of God

# Numerical Solution of linear Ordinary Differential Equations and Differential-Algebraic Equations by Spectral Methods

by

Masoud Saravi

This thesis is submitted in fulfilment of the requirements of the Degree of Doctor of Philosophy of the Open University in Numerical Analysis

September 2007

Msaravi2001@yahoo.com

# Acknowledgements

**I would like to express my thanks to my external supervisor Prof. E. Babolian, and internal supervisors Dr R. England and Dr M. Bromilow for their constructive criticism and encouragement throughout this endeavour, particularly Dr R. England for providing access to most of the books I needed during the research for this thesis. My thanks also go to Dr U. Grimm for his continuing assistance, and Miss L. J. Smith for providing most of the letters which I requested. Let me express my appreciation to Mr M. Rastegari for his expert guidance of my computer programming, and all those people who assisted or encouraged me, in particular, my wife and my children Sina and Sayeh who cooperated with me during these years.**

**M. Saravi**

**Sept 2007**

# Table of Contents

# Abstract

This thesis involves the implementation of spectral methods, for numerical solution of linear Ordinary Differential Equations (ODEs) and linear Differential-Algebraic Equations (DAEs).

First we consider ODEs with some ordinary problems, and then, focus on those problems in which the solution function or some coefficient functions have singularities. Then, by expressing weak and strong aspects of spectral methods to solve these kinds of problems, a modified pseudo-spectral method which is more efficient than other spectral methods is suggested and tested on some examples.

We extend the pseudo-spectral method to solve a system of linear ODEs and linear DAEs and compare this method with other methods such as Backward Difference Formulae (BDF), and implicit Runge-Kutta (RK) methods using some numerical examples. Furthermore, by using appropriate choice of Gauss-Chebyshev-Radau points, we will show that this method can be used to solve a linear DAE whenever some of coefficient functions have singularities by providing some examples. We also used some problems that have already been considered by some authors by finite difference methods, and compare their results with ours.

Finally, we present a short survey of properties and numerical methods for solving DAE problems and then we extend the pseudo-spectral method to solve DAE problems with variable coefficient functions. Our numerical experience shows that spectral and pseudo-spectral methods and their modified versions are very promising for linear ODE and linear DAE problems with solution or coefficient functions having singularities.

In section 3.2, a modified method for solving an ODE is introduced which is new work. Furthermore, an extension of this method for solving a DAE or system of ODEs which has been explained in section 4.6 of chapter four is also a new idea and has not been done by anyone previously.

In all chapters, wherever we talk about ODE or DAE we mean linear.

# Chapter 1

# Spectral Approximation

## 1.1 Introduction

As we know, most differential equations concerning physical phenomenon can not be solved in terms of known functions, and, even when they can, sometimes their closed form solution is so complicated that using it to obtain an image or to examine the structure of the system is impossible. Consequently, is hardly surprising that polynomial approximation often has an important role when one wants to approximate a given function $u(x)$. One of the fundamental theorems related to this is the *Weierstrass approximation theorem*, [1] which states that:

If the function $u$ on $[a,b]$ is continuous, then, for any $\varepsilon > 0$, there exists a polynomial $p_n(x)$ of degree $n$, such that on this interval for sufficiently large $n$,

$$|p_n(x) - u(x)| < \varepsilon \quad \text{for all } x \in [a,b].$$

In this chapter we shall consider from a general point of view, the problem of approximating a function in terms of an orthogonal system of polynomials which guarantees *spectral accuracy*.

Spectral accuracy happens when the $n^{th}$ coefficient of the expansion decays faster than any inverse power of $n$. Spectral accuracy is attainable for the Fourier series expansion of periodic functions $\in C^{\infty}$ (if $f \in C^{m}[a,b]$ for $m \geq 2$ then we call $f$ a smooth function, and by $f \in C^{\infty}[a,b]$ we mean $f$ is infinitely smooth). The property of spectral accuracy is also attainable for smooth but non-periodic functions provided that the expansion functions are chosen properly.

It is shown that, the eigenfunctions of a singular Sturm-Liouville operator allow spectral accuracy in the expansion of any smooth function.

The expansion in terms of an orthogonal system introduces a linear transformation between the approximated function $u$ and the finite sequence of its expansion coefficients $\{\hat{u}_n\}$.This is usually called the *finite transform* of $u$ between physical space and transform space. If the orthogonal system is complete in a suitable Hilbert space, this transform can be inverted. Hence, functions can be described both through their values in physical space and through their coefficients in transform space. The expansion coefficients depend on all values of $u$ in physical space; hence, they can rarely be computed exactly. A finite number of approximate expansion coefficients can be easily computed using the values of $u$ at a finite number of selected points, usually the nodes of high precision *quadrature formulae*. This procedure defines a *discrete transform* between the set of values of $u$ at the quadrature points and the set of approximations, or discrete coefficients. With a proper choice of the quadrature formulae, the finite series defined by the discrete transform is actually the interpolant of $u$ at the quadrature nodes. If the properties of accuracy (in particular the spectral accuracy) are retained in replacing the finite transform with the discrete transform, then the interpolant series can be used instead of the truncated series in approximating functions. For some of the most common orthogonal systems (Fourier and Chebyshev polynomials) the discrete transform can be computed in a "fast" way, i.e., with an operation count $(5/2)N\log_2 N$, where $N$ is the number of polynomials, rather than with $2N^2$ operations required by a matrix-vector multiplication.

*Fast discrete transforms* for other orthogonal systems have been suggested (Orszag(1986))[12], but their utility in practical computations is, at present, unproven.

In this chapter we shall describe in detail those orthogonal systems which guarantee spectral accuracy.

## 1.2 Orthogonal Systems of Polynomials

Orthogonal polynomials play the most important role in spectral methods, so it is useful to understand some general properties of them. Given an interval $(a,b)$ and a weight function w(x)

which is nonnegative on $(a,b)$ and w$\in L^1\ (a,b)$, we define the weighted *Sobolev space* $L^2_w(a,b)$ by

$$L^2_w(a,b)=\left\{f:\int_a^b|f(x)|^2w(x)dx<\infty\right\}. \tag{1.1}$$

It is obvious that $(.,.)_w$ defined by

$$(f,g)_w = \int_a^b f(x)g(x)w(x)dx \tag{1.2}$$

is an inner product on $L^2_w(a,b)$. Hence, $\| f \|_{L^2_w} = (f,f)^{1/2}_w$ . Two functions $f$ and $g$ are said to be orthogonal in $L^2_w(a,b)$ if $(f,g)_w$=0.

A sequence of polynomials $\{p_n\}_{n=0}^{\infty}$ with $p_n$ of degree $n$ is said to be orthogonal in $L^2_w(a,b)$ if $(p_i,p_j)_w$=0, when $i\neq j$.

Let $\{p_n\}_{n=o}^{\infty}$ be a sequence of polynomials, mutually orthogonal over the interval $(a,b)$ with respect to a weight function $w$, that is;

$$\int_a^b p_n(x)p_m(x)w(x)dx=0,\ m\neq n. \tag{1.3}$$

If $m=n$, then we have $\|p_n\|=\left[\int_a^b p_n{}^2(x)w(x)dx\right]^{1/2}$, which is called the *norm* of the orthogonal sequence of polynomials $\{p_n\}_{n=0}^{\infty}$.

The infinite sequence $\{p_n(x)\}_0^{\infty}$ is called complete in the underlying space $S$ if each function in $S$ has a unique expansion of the form $\sum_0^{\infty} a_i p_i(x)$ with $a_i\in\Re$ or C, $i=0,1,2,....$

The classical Weierstrass theorem implies that such a sequence of polynomials is complete in the space $L^2_w(a,b)$.

Since $\{p_n\}_{n=o}^{\infty}$ is complete, it follows that for any u $\in L^2_w(a,b)$ , we can write

$$u=\sum_{n=0}^{\infty}\hat{u}_n p_n, \tag{1.4}$$

where the expansion coefficients $\hat{u}_n$ are defined as

$$\hat{u}_n=\frac{1}{\|p_n\|_w^2}\int_a^b u(x)p_n(x)w(x)dx\,. \tag{1.5}$$

For an integer $N>0$, the truncated series of $u$ of order $N$ is the polynomial

$$p_N u=\sum_{n=0}^{N}\hat{u}_n p_n. \tag{1.6}$$

Due to (1.3), $p_N u$ is the orthogonal projection of $u$ upon $\overline{\mathrm{P}}_N$, the space of all polynomials of degree $\leq N$, in the inner product (1.3), i.e., $(p_N u, v)_w=(u,v)$, for all $v\in\overline{\mathrm{P}}_N$.

The completeness of sequence of polynomials $\{p_n\}_{n=0}^{\infty}$ means that for all

$u\in L_w^2(a,b)$, $\|u-p_N u\|_w\to 0$ as $N\to\infty$.

The zeros of the orthogonal polynomials play an important role in the implementation of spectral methods. The main result concerning the zeros of orthogonal polynomials is that;

"the zeros of $p_n$ separate the zeros of $p_{n+1}$, and that the polynomial $p_n$ has $n$ distinct zeros on $(a,b)$". It is also well known that these polynomials satisfy a recurrence relation of the form

$$xp_n=\alpha_n p_{n+1}+\beta_n p_n+\gamma_n p_{n-1}\quad, \tag{1.7}$$

where $\alpha_n>0$, $\beta_n,\gamma_n$ are constants. [2].

## 1.3 Sturm-Liouville Problems

The importance of Sturm-Liouville problems for the spectral methods lies in the fact that in the spectral approximation the solution of a differential problem is usually regarded as a finite expansion of eigenfunctions of a suitable Sturm-Liouville problem. We recall that a Sturm-Liouville problem is an eigenvalue problem of the form

$$
\begin{aligned}
&(pu')' + (q + \lambda w)u = 0 \\
&a_1 u(a) + a_2 u'(a) = 0 \\
&b_1 u(b) + b_2 u'(b) = 0
\end{aligned} \tag{1.8}
$$

The parameter $\lambda$ is independent of $x$ while $p, q$ and $w$ are real-valued functions of $x$. To ensure the existence of solutions we let $q$ and $w$ be continuous and $p$ be continuously differentiable, strictly positive in *(a,b)* and continuous at $x = a, b$; and $q$ is continuous, non-negative and bounded in *(a,b)*; the weight function *w* is continuous, non-negative and integrable over *(a,b)*.

The values of $\lambda$ for which the Sturm-Liouville problem has a non-trivial solution are called the *eigenvalues*, and the corresponding solutions are called *eigenfunctions*. For example, the functions $\sin(x/2), \sin(3x/2), \ldots$ are *eigenfunctions* of the Sturm-Liouville problem

$$
\begin{aligned}
&u'' + \lambda u = 0, x \in [0, \pi] \\
&u(0) = 0 \\
&u'(\pi) = 0,
\end{aligned}
$$

corresponding to the eigenvalues ¼, 9/4, ….

The Sturm- Liouville problems of interest in spectral methods are such that the expansion of infinitely smooth functions in terms of their eigenfunctions guarantees spectral accuracy.

A smooth function can be approximated by cosine series on *(a,b)* with spectral accuracy if and only if all its odd derivatives vanish at boundary. This is due to the fact that the coefficient *p(x)* in the operator does not vanish at the boundary in this case, i.e., Sturm-Liouville problem is *regular* (The Sturm-Liouville problem is called *regular* in the interval $[a,b]$ if the functions $p$ and /or $w$ are positive in $[a,b]$). If $p$ vanishes at the boundary the problem is called *singular*. The spectral accuracy is ensured if the problem is *singular*. A mathematical proof of these facts is given in Sec. 9.2[5]. Expansions based on eighenfunctions of a Sturm-Liouville problem that is singular at $x=a$ do not normally exhibit the Gibbs phenomenon at $x=a$. In applications, we encounter piecewise smooth functions frequently. In this case, the approximation is not uniform. An overshoot and

undershoot always appears across discontinuities. Such a phenomenon is called the Gibbs phenomenon.

The important conclusion is that eigenfunction expansions based on Sturm-Liouville problems that are singular at $x=a$ and at $x=b$ converge at a rate governed by the smoothness of the function being expanded not by any special boundary conditions satisfied by the function [4].

Among the singular Sturm-Liouville problems, particular importance rests with those problems whose eigenfunctions are algebraic polynomials because of the efficiency with which they can be evaluated and differentiated numerically. It is also proven in Sec. 9.2 [5] that the Jacobi polynomials are precisely the only polynomials arising as eigenfunctions of a singular Sturm-Liouville problem.

Let $\{\phi_n(x)\}_{n=1}^{\infty}$ be an orthogonal sequence of complete square-integrable functions with a positive weight function $w$ on $[a,b]$, then the solution of a Sturm-Liouville problem (1.8) can be expanded in a uniformly convergent series of the form

$$u=\sum_{n=1}^{\infty}\hat{u}_n\phi_n(x)\,, \tag{1.9}$$

where $\hat{u}_n$, the coefficients of expansion, are given by

$$\hat{u}_n=\left(\int_a^b w(x)\phi_n^2(x)dx\right)^{-1}\int_a^b w(x)u(x)\phi_n(x)dx\,. \tag{1.10}$$

A nice discussion about eigenfunction expansions and proof of a more general theorem can be found in E. C. Titchmarsh [3].

Later, in relation to Sturm-Liouville problems, we will discuss orthogonal polynomials such as Legendre, Chebyshev and Jacobi polynomials. They have special importance because, for sufficiently smooth functions, spectral accuracy is guaranteed. In other words, the $n^{th}$ coefficient of the expansion decays faster than any finite inverse power of $n$ [4]. This property does not hold, in general, for all Sturm-Liouville problems. In the next section we discuss the close relationship

between orthogonal polynomials and Gauss-type quadrature rules because of their spectral accuracy in comparison to other numerical integration rules.

## 1.4 Gauss-type Quadrature rules

For solutions of differential equations by spectral methods, it is necessary to evaluate integrals numerically. We use integration formulae on $[a,b]$ of the type

$$\int_a^b f(x)w(x)dx \cong \sum_{j=0}^{N} w_j f(x_j)\,. \tag{1.11}$$

Let $x_0, x_1, ..., x_N$, be $N$+1 distinct points in [$a$, $b$]. We may choose coefficients(weights) $\{w_j\}_{j=0}^{N}$ such that (1.11) is exact for polynomials of degree $\leq N$. More precisely, we set

$$w_j = \int_a^b l_j(x)w(x)dx \tag{1.12}$$

with $l_j(x) = \prod_{i\neq j}\left(\frac{x-x_i}{x_j-x_i}\right)$, being the Lagrange polynomial associated with nodes $\{x_j\}_{j=0}^{N}$. Since, $l_j(x_j)=1$, and $l_j(x_i)=0,\ i\neq j$, then $f(x)=\sum_{j=1}^{N} f(x_j)l_j(x)$, if $f(x)$ is a polynomial of degree $\leq N$, and in general

$$\int_a^b f(x)\,w(x)\,dx \cong \sum_{j=1}^{N} f(x_j)\int_a^b l_j(x)\,w(x)\,dx = \sum_{j=1}^{N} w_j f(x_j)\,.$$

Hence, we have

$$\int_a^b p(x)w(x)dx = \sum_{j=0}^{N} p(x_j)w_j,$$

where $p \in \overline{P}_N$, where $\overline{P}_N$ is the space of all polynomials of degree $\leq N$. However, if we are free to choose nodes $x_0, x_1, x_2, ..., x_N$, we can expect the quadrature formulae of the above form to be exact for polynomials of degree $\leq 2N+1$.

We assume that $\{p_n\}_{n=0}^{\infty}$ is a sequence of orthogonal polynomials with respect to a weight function on $(a,b)$.

***Theorem 1.*** *(Gauss quadrature)* :

(i) Let $x_0, x_1, ..., x_N$ be the zeros of $p_{N+1}$ and define

$w_j (j = 0,1,...,N)$ by (1.12). Then $w_j > 0$ for $j = 0,1,...,N$ and

$$\int_a^b p(x)w(x)dx = \sum_{j=1}^{N} p(x_j)w_j, \text{ for all } \quad p \in \overline{\mathrm{P}}_{2N+1}. \tag{1.13}$$

(ii) It is not possible to find $x_j, w_j, j = 0,...,N$ such that (1.13) holds for all polynomials $p \in \overline{\mathrm{P}}_{2N+2}$. However, it is difficult to enforce any boundary condition since end points $a$ and $b$ are not among the Gauss nodes. Therefore, we need generalized Gauss quadratures with some enforced boundary conditions. To enforce the boundary condition at one end point, we should use the Gauss-Radau quadrature. Other generalized Gauss quadrature derivative boundary conditions can also be constructed similarly, (see[6], [2]).

Suppose we would like to include the left end point $a$ in the quadrature. We choose $\alpha_N = -\frac{p_{N+1}(a)}{p_N(a)}$, and set

$$q(x) = p_{N+1}(x) + \alpha_N p_N(x). \tag{1.14}$$

Hence $q(a) = 0$ and we can write

$$q(x) = (x-a)q_N(x).$$

It is obvious that $q_N \in \overline{\mathrm{P}}_N$, and for any $r_{N-1} \in \overline{\mathrm{P}}_{N-1}$ we have

$$\int_a^b q_N(x)r_{N-1}(x)w(x)(x-a)dx =$$

$$\int_a^b (p_{N+1}(x) + \alpha_N p_N(x))r_{N-1}(x)w(x)dx = 0. \tag{1.15}$$

Hence $\{q_N\}$ is a sequence of orthogonal polynomials with respect to the weight function $(x-a)w(x)$ .

***Theorem 2.*** *(Gauss-Radau quadrature)* : Let $x_0, x_1, ..., x_N$ be the zeros of $(x-a)q_N$ and $w_j (j = 0,1,..., N)$ defined by (1.12).Then $w_j > 0$ for $j = 0,1,..., N$ and

$$\int_a^b p(x)w(x)dx = \sum_{j=0}^{N} p(x_j)w(x_j), \text{ for all } p \in \overline{P}_{2N} \ . \tag{1.16}$$

A second Gauss-Radau quadrature can be constructed if we want to include the end point $b$ instead of the left end point $a$.We now consider the Gauss-Lobatto quadrature whose nodes include the two end points. We choose $\alpha_N$ and $\beta_N$ such that

$$p_{N+1}(x) + \alpha_N p_N(x) + \beta_N p_{N-1}(x) = 0, \text{ for } x = a, b,$$

and set

$$s_{N-1}(x) = \frac{p_{N+1}(x) + \alpha_N p_N(x) + \beta_N p_{N-1}(x)}{(x-a)(b-x)}.$$

Hence, $s_{N-1} \in \overline{P}_{N-1}$ and for any $r_{N-2} \in \overline{P}_{N-2}$ , we have

$$\int_a^b s_{N-1}(x) r_{N-1}(x) w(x)(x-a)(b-x)dx =$$

$$\int_a^b (p_{N+1}(x) + \alpha_N p_N(x) + \beta_N p_{N-1}(x)) r_{N-2}(x) w(x)dx = 0 \tag{1.17}$$

Hence, $\{s_N\}$ is a sequence of orthogonal polynomials with respect to weight function $(x-a)(b-x)\, w(x)$ .

***Theorem 3.*** *(Gauss-Lobatto quadrature)* : Let $\{x_j\}_{j=0}^{N}$ be the $N+1$ zeros of

$(x-a)(b-x)\, s_{N-1}(x)$ and $w_j (j = 0,1,..., N)$ defined by (1.12). Then $w_j > 0$ for $j = 0,1,..., N$, and

$$\int_a^b p(x)w(x)dx = \sum_{j=0}^{N} p(x_j)w_j \text{ ,for all } p \in \overline{P}_{2N-1} \ . \tag{1.18}$$

The nodes of Gauss-type formulae play an important role in collocation approximations. They are precisely the collocation points at which the differential equations are required to be satisfied identically (see [5]). We assume here that a weight function $w$ is given, together with the corresponding sequence of orthogonal polynomials $p_n$. For a given $N \geq 0$, we denote by $x_0, x_1, ..., x_N$ the nodes of the $N+1$-point integration formulae of Gauss, Gauss-Radau or Gauss-Lobatto type, and by $w_0, w_1, ..., w_N$ the corresponding weights.

In a collocation method the fundamental representation of a smooth function $u$ on $(a,b)$ is in terms of its values at the discrete Gauss-type points. Derivatives of the function are approximated by analytic derivatives of the interpolating polynomial. The interpolating polynomial is denoted by $I_N u$. It is an element of $\overline{\mathrm{P}}_N$ and satisfies

$$I_N u(x_j) = u(x_j), \; j = 0,1,...,N. \tag{1.19}$$

Let $w(x) > 0$ be a weight function, and $\{x_j, w_j\}_{j=0}^{N}$ be the set of quadrature points (e.g. Gauss, Gauss-Radau or Gauss- Lobatto points) and associated weights. For $u$ and $v \in \overline{\mathrm{P}}_N$ continuous on $[a,b]$, we define

$$\langle u,v \rangle_N = \sum_{j=0}^{N} u(x_j) v(x_j) w_j \quad . \tag{1.20}$$

Then $\langle .,. \rangle_N$ is a discrete inner product in $\overline{\mathrm{P}}_N$ and $\|.\|$ defined by $\|u\|_N = \langle u,u \rangle_N^{1/2}$, is a norm in $\overline{\mathrm{P}}_N$.

In particular, Gauss, Gauss-Radau and Gauss-Lobatto quadrature formulae imply that

$$\langle u,v \rangle_N = (u,v)_w, \text{ for } u,v \in \overline{\mathrm{P}}_{2N+\delta} \tag{1.21}$$

where $\delta = 1,0,-1$ respectively for Gauss, Gauss-Radau and Gauss-Lobatto quadrature.

Let $u$ be a continuous function on $[a,b]$. The interpolation polynomial associated with $\{x_j\}_{j=0}^{N}$, is defined as a polynomial of degree at most $N$ to satisfy (1.19). Hence, we may write

$$I_N u(x) = \sum_{i=0}^{N} \tilde{u}_i p_i(x). \tag{1.22}$$

Obviously, we have

$$u(x_j) = I_N u(x_j) = \sum_{i=0}^{N} \tilde{u}_i p_i(x_j). \tag{1.23}$$

Thus, $\{\tilde{u}_k\}$ are called the discrete coefficients of $u$ determined by

$$\tilde{u}_k = \frac{1}{\gamma_k} \sum_{j=0}^{N} u(x_j) p_k(x_j) w_j \quad , \tag{1.24}$$

where

$$\gamma_k = \sum_{j=0}^{N} p_k^2(x_j) w_j . \tag{1.25}$$

Equations (1.23) and (1.24) enable one to transform freely between physical space $\{u(x_j)\}$ and transform space $\{\tilde{u}_k\}$. We shall call such a transformation a "*discrete polynomial transform*" associated with the weight function $w$ and the nodes $x_0, x_1, ..., x_N$. For any continuous $v$, (1.23) gives

$$(I_N u, v)_N = (u, v)_N \ . \tag{1.26}$$

This shows that the interpolant $I_N u$ is the projection of $u$ upon $\overline{\mathrm{P}}_N$ with respect to the discrete inner product (1.20).

The discrete polynomial coefficients $\tilde{u}_k$ can be expressed in terms of the continuous coefficients $\hat{u}_k$ as follows:

$$\tilde{u}_k = \hat{u}_k + \frac{1}{\gamma_k} \sum_{j>N} (p_j, p_k)_N \hat{u}_j . \tag{1.27}$$

This formulae is an easy consequence of (1.24). Equivalently, one can write

$$I_N u = p_N u + R_N u, \tag{1.28}$$

where

$$R_N u = \sum_{k=0}^{N} \left( \frac{1}{\gamma_k} \sum_{j>N} (p_j, p_k)_w \hat{u} \right) p_k \tag{1.29}$$

can be considered the *aliasing error* due to interpolation. The aliasing error is orthogonal to the truncation error, $u - p_N u$, so that

$$\|u - I_N u\|_w^2 = \|u - p_N u\|_w^2 + \|R_N u\|_w^2 \ . \tag{1.30}$$

In general, $(p_j, p_k) \neq 0$ for all $j > N$. Thus the $k^{th}$ mode of the algebraic interpolant of $u$ depends on the $k^{th}$ mode of $u$ and all the modes whose wave number is larger than $N$. The aliasing error has a simple expression for the Chebyshev interpolation points.

In the rest of this chapter we shall restrict our attention to some special class of orthogonal polynomials.

## 1. Jacobi polynomials

Jacobi polynomials are the most general case of classical orthogonal polynomials, which are denoted by $J_n^{\alpha,\beta}(x)$ and, generated from (1.8) with

$p(x) = (1-x)^{\alpha+1}(1+x)^{\beta+1}, q(x) = 0,\ w(x) = (1-x)^{\alpha}(1+x)^{\beta}$ for $\alpha, \beta > -1, (a,b) = (-1,1)$, they are

$$J_n^{\alpha,\beta}(x) = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n+\beta}{n-k} (x-1)^k (x+1)^{n-k} \ , \tag{1.31}$$

normalized by $J_n^{\alpha,b}(1) = \Gamma(n+\alpha+1)\Gamma(\alpha+1)/n!$, where $\Gamma(x)$ is the Gamma function. In fact, we shall mainly be concerned with two special cases of Jacobi polynomials, namely Legendre polynomials which correspond to $\alpha = \beta = 0$ and Chebyshev polynomials which correspond to $\alpha = \beta = -1/2$. Any generic treatments of Jacobi polynomials apply in particular to both Legendre and Chebyshev polynomials. A property of fundamental importance is the following:

***Theorem 4.*** The Jacobi polynomials satisfy the following singular Sturm-Liouville problem:

$$(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d}{dx}\left\{(1-x)^{\alpha+1}(1+x)^{\beta+1}\frac{d}{dx}J_n^{\alpha,\beta}(x)\right\}+n(n+1+\alpha+\beta)J_n^{\alpha,\beta}(x)=0.$$

An immediate consequence of the above relation is that there exists $\lambda$ such that

$$-\frac{d}{dx}\left\{(1-x)^{\alpha+1}(1+x)^{\beta+1}\frac{d}{dx}J_n^{\alpha,\beta}(x)\right\}=\lambda J_n^{\alpha,\beta}(x)w(x).$$

We can simply show that $\lambda=n(n+1+\alpha+\beta)$.

The orthogonality condition of Jacobi polynomials gives

$$\int_{-1}^{1}J_m^{\alpha,\beta}(x)J_n^{\alpha,\beta}(x)(1-x)^{\alpha}(1+x)^{\beta}\,dx=0\ \text{, for } m\neq n\,. \tag{1.32}$$

One derives immediately from theorem 4 and (1.32) the following result:

$$\int_{-1}^{1}(1-x)^{\alpha+1}(1+x)^{\beta+1}\frac{dJ_m^{\alpha,\beta}}{dx}\frac{dJ_n^{\alpha,\beta}}{dx}dx=0\ \text{, for } m\neq n\,. \tag{1.33}$$

The above relation indicates that $\frac{d}{dx}J_n^{\alpha,\beta}$ forms a sequence of orthogonal polynomials with weight function $w(x)=(1-x)^{\alpha+1}(1+x)^{\beta+1}$. Hence, by the uniqueness, we find that $\frac{d}{dx}J_n^{\alpha,\beta}$ is proportional to $J_{n-1}^{\alpha+1,\beta+1}$.

***Theorem 5.*** *(Rodrigues' formulae)*

$$(1-x)^{\alpha}(1+x)^{\beta}J_n^{\alpha,\beta}(x)=\frac{(-1)^n d^n}{2^n n!dx^n}\left[(1-x)^{n+\alpha}(1+x)^{n+\beta}\right]. \tag{1.34}$$

When $\alpha=\beta>-1$, the corresponding Jacobi polynomials are called Gegenbauer polynomials or ultraspherical polynomials. In this case, one derives from Rodrigues' formulae that $J_n^{\alpha,\alpha}$ is an odd function for $n$ odd and an even function for $n$ even.

## 2. Legendre polynomials

The Legendre polynomials, denoted by $L_k(x)$, are the eigenfunctions of the singular Sturm-Liouville problem with $p(x)=1-x^2, q(x)=0$, $w(x)=1$ and $(a,b)=(-1,1)$ and the boundary conditions $u(\pm 1)$ be finite. Since $p(x)/w(x)=1-x^2$ and $p'(x)/w(x)=-2x$ are both finite for $|x|\le 1$, it follows that the Legendre series expansion of infinitely differentiable functions converges faster than algebraically. When a discontinuous function is expanded in Legendre series, the rate of convergence is no longer faster than algebraic. In the neighbourhood of a discontinuity, a Gibbs phenomenon occurs whose local structure is the same as that for Fourier series.

The three-term recurrence relation for Legendre polynomials reads

$$(n+1)L_{n+1}(x)=(2n+1)xL_n(x)-nL_{n-1}(x), n\ge 1. \tag{1.35}$$

with $L_0(x)=1, L_1(x)=x$.

We present here a collection of essential formulae for Legendre polynomials. For proofs the reader may refer to Szego (1939)[12].

As we said the Legendre polynomials are Jacobi polynomials with $\alpha=\beta=0$. Hence, they are the eigenfunctions of the singular Sturm-Liouville problem

$$\left((1-x^2)L_n'(x)\right)'+n(n+1)L_n(x)=0, x\in(-1,1). \tag{1.36}$$

$L_k(x)$ is even if $k$ is even and odd if $k$ is odd. If $L_k(x)$ is normalized then for any $k$

$$L_k(x)=\frac{1}{2^k}\sum_{j=0}^{[k/2]}(-1)^j\binom{k}{j}\binom{2k-2j}{k}x^{k-2j}. \tag{1.37}$$

An important property of Legendre polynomials is the following

$$\int_{-1}^{x} L_n(t)dt = \frac{1}{2n+1}(L_{n+1}(x) - L_{n-1}(x)), n \geq 1 \ . \tag{1.38}$$

We use the above recursive relation for computing derivatives of the Legendre polynomials:

$$L_n(x) = \frac{1}{2n+1}(L'_{n+1}(x) - L'_{n-1}(x)), \ n \geq 1 \ . \tag{1.39}$$

We can derive from the above formulae that

$$L_n'(x) = \sum_{\substack{k \\ n+k \text{ odd}}}^{n-1} (2k+1)L_k(x) \ , \tag{1.40}$$

$$L_n''(x) = \sum_{\substack{k=0 \\ n+k \text{ even}}}^{n-2} (k+\tfrac{1}{2})(n(n+1) - k(k+1))L_k(x) \ . \tag{1.41}$$

We also derive relevant properties

$$|L_k(x)| \leq 1, |L'_k(x)| \leq \frac{k(k+1)}{2} \ , \tag{1.42}$$

$$L_k(\pm 1) = (\pm 1)^k \ , \ L'_k(\pm 1) = \tfrac{1}{2}(\pm 1)^{k-1} k(k+1) \ , \tag{1.43}$$

$$L''_k(\pm 1) = (\pm 1)^k (k-1)k(k+1)(k+2)/8 \ , \tag{1.44}$$

$$\int_{-1}^{1} L_k^2(x)dx = \frac{2}{2k+1} \ . \tag{1.45}$$

The expansion of any $u \in L^2(-1,1)$ in terms of $L_k$'s is

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k L_k(x) \ , \ \hat{u}_k = (k+\tfrac{1}{2}) \int_{-1}^{1} u(x) L_k(x)dx \ . \tag{1.46}$$

We consider now discrete Legendre series. Since explicit formulae for the quadrature nodes are not known, such points have to be computed numerically as zeros of appropriate polynomials. For

the Legendre series, the Gauss quadrature points and weights can be derived from theorems 1,3 (see also[2],[7]).

(i) For the Gauss-Legendre quadrature: $\{x_j\}_{j=0}^{N}$ are the zeros of $L_{N+1}(x)$, and

$$w_j = \frac{2}{(1-x_j^2)(L'_{N+1}(x_j))^2} \quad . \tag{1.47}$$

(ii) For the Legendre-Gauss-Radau quadrature: $\{x_j\}_{j=0}^{N}$ are zeros of

$L_N(x) + L_{N+1}(x)$, and

$$w_0 = \frac{2}{(N+1)^2} \quad , \; w_j = \frac{1}{(N+1)^2}\frac{1-x_j}{\left(L_N(x_j)\right)^2} \, , 1 \le j \le N \quad . \tag{1.48}$$

(iii) For the Legendre-Gauss-Lobatto quadrature: $\{x_j\}_{j=0}^{N}$ are zeros of

$$(1-x^2)L'_N(x), \text{ and } w_j = \frac{2}{N(N+1)(L_N(x_j))^2} \quad , \quad 0 \le j \le N \quad . \tag{1.49}$$

The normalization factors $\gamma_k$ introduced in (1.25) are given by

$\gamma_k = \left(k+\frac{1}{2}\right)^{-1}$, for $k < N$. And $\gamma_N$ is given by

$(N+\frac{1}{2})^{-1}$, for Gauss and Gauss-Radau formulae, and (150)

$\frac{2}{N}$ , for Gauss-Lobatto formulae .

In implementation of the spectral method, one often needs to evaluate derivatives or form derivative matrices. The derivatives can be evaluated either in the frequency space or in the physical space.

There are obvious difficulties if $u$ has discontinuities. In this case, the approximation is not uniform.

a) Differentiation in frequency space. Given $u = \sum_{k=0}^{N} \hat{u}_k L_k \in \overline{\mathrm{P}}_N$, we can write

$$u' = \sum_{k=1}^{N} \hat{u}_k L_k' = \sum_{k=0}^{N} \hat{u}_k^{(1)} L_k \quad \text{with } \hat{u}_N^{(1)} = 0.$$

Then from (1.39) we find

$$u' = \sum_{k=0}^{N} \hat{u}_k^{(1)} L_k = \hat{u}_0^{(1)} + \sum_{k=1}^{N-1} \hat{u}_k^{(1)} \frac{1}{2k+1} (L_{k+1}' - L_{k-1}')$$

$$= \frac{\hat{u}_{N-1}^{(1)}}{2N-1} L_N' + \sum_{k=1}^{N-2} \left\{ \frac{\hat{u}_{k-1}^{(1)}}{2k-1} - \frac{\hat{u}_{k+1}^{(1)}}{2k+3} \right\} L_k' \ .$$

Comparing the coefficients of $L_k'$ , we find that the coefficients $\{\hat{u}_k^{(1)}\}$ of $u'$ are determined by the recursive relation:

$$\hat{u}_N^{(1)} = 0, \hat{u}_{N-1}^{(1)} = (2N-1)\hat{u}_N, \quad \hat{u}_{k-1}^{(1)} = \left( \hat{u}_k + \frac{\hat{u}_{k+1}^{(1)}}{2k+3} \right) (2k-1), k = N-1, N-2, ..., 1 \ . \tag{1.51}$$

Higher derivatives can be obtained by repeatedly applying the above formulae.

b) Derivative matrices in physical space. Given $u \in \overline{P_N}$ and its values at a set of collocation points $\{x_j\}_{j=0}^N$. Let $\{l_j(x)\}_{j=0}^N$ be the sequence of Lagrange polynomials relative to $\{x_j\}_{j=0}^N$.Then,

$$u^{(m)}(x) = \sum_{j=0}^{N} u^{(m-1)}(x_j) l_j'(x), \ m \geq 1 \ . \tag{1.52}$$

Setting $d_{kj} = l_j'(x_k)$, and $D = (d_{kj})_{k,j=0,1,...,N}$, and

$$\bar{u}^{(m)} = \left( u^{(m)}(x_0), u^{(m)}(x_1), ..., u^{(m)}(x_N) \right)^T, m = 0,1,2,...$$

we can obtain from (1.52) :

$$u^{(m)}(x_k) = \sum_{j=0}^{N} d_{kj} u^{(m-1)}(x_j) \text{ or } \bar{u}^{(m)} = D\bar{u}^{(m-1)} \ , \tag{1.53}$$

which implies that $\bar{u}^{(m)} = D^m \bar{u}^{(0)}$.

Hence, the derivatives of $u$ in the physical space are totally determined by the matrix $D$.The above discussion is valid for any set of collocation points. In the implementation of spectral methods, one is interested mostly in using the Gauss-Lobatto points or Gauss-Legendre points. Below we provide an explicit expression for the derivative matrix in the case when $\{x_j\}_{j=0}^N$ are Legendre-Gauss-Lobatto points.

Let $\{x_j\}_{j=0}^N$ be the zeros of $(1-x^2)L_N'(x)$. Then,

$$l_j(x) = -\frac{1}{N(N+1)L_N(x_j)} \frac{(1-x^2)L_N'(x)}{x-x_j},\ j = 0,1,\ldots,N, \tag{1.54}$$

and

$$d_{kj} = \frac{L_N(x_k)}{L_N(x_j)(x-x_j)},\ k \neq j \neq 0,\ N, \tag{1.55}$$

$$d_{kj} = N(N+1)/4,\ for\ k = j = 0\ and\ d_{kj} = -N(N+1)/4,\ for\ k = j = N,$$
$$and\ 0,\ for\ 1 \leq k = j \leq N-1.$$

***Remark* 1:** The derivative matrix is a full matrix and so $O(N^2)$ flops are needed to compute $\{u'(x_j)\}_{j=0}^N$ from $\{u(x_j)\}_{j=0}^N$ by using the derivative matrix.

Since $u^{(N+1)}(x) = 0$ for any $u \in \overline{P}_N$, we have $D^{N+1}\bar{u}^{(0)} = 0$ for any $\bar{u}^{(0)} \in \Re^{N+1}$. Hence, the only eigenvalue of $D$ is zero which has a multiplicity of order $N+1$.

## 3. Chebyshev polynomials

The Chebyshev polynomials of the first kind, denoted by $\{T_k(x)\}_{k=0}^\infty$ are the eigenfunctions of the singular Sturm-Liouville problem

$$\sqrt{1-x^2}\left(\sqrt{1-x^2}T_k'(x)\right)' + k^2 T_k(x) = 0, x \in (-1,1), \tag{1.56}$$

which is (1.8) with $p(x)=(1-x^2)^{1/2}, q(x)=0,\ w(x)=(1-x^2)^{-1/2}$, and the boundary conditions that $T_k(\pm 1)$ be finite.

Three-term recurrence for the Chebyshev polynomials is:

$$\begin{aligned} &T_0(x)=1, T_1(x)=x, \\ &T_{n+1}(x)=2xT_n(x)-T_{n-1}(x), n\geq 1. \end{aligned} \tag{1.57}$$

The Chebyshev polynomials are the Jacobi polynomials with $\alpha=\beta=-1/2$, and satisfy the orthogonality relation

$$\int_{-1}^{1} T_k(x)T_j(x)/\sqrt{1-x^2}\,dx=\frac{c_k\pi}{2}\delta_{kj}\ , \tag{1.58}$$

where $c_0=2$ and $c_k=1$, for $k\geq 1$.

The Chebyshev polynomials can be expanded in power series as

$$T_k(x)=\frac{k}{2}\sum_{l=0}^{\lfloor k/2\rfloor}(-1)^k\frac{(k-l-1)!}{l!(k-2l)!}(2x)^{k-2l}; \tag{1.59}$$

where, $\lfloor k/2\rfloor$ denotes the integral part of $k/2$.

From the fact that $\cos(n\cos^{-1}x)$ is a polynomial of degree $n$ and the trigonometric relation

$$\cos(n+1)\theta+\cos(n-1)\theta=2\cos\theta\cos n\theta, \tag{1.60}$$

we find that $\cos(n\cos^{-1}x)$ satisfies also the three-term recurrence relation (1.57). Hence,

$$T_n(x)=\cos(n\cos^{-1}x),\ n=0,1,... \tag{1.61}$$

This explicit representation allows us to derive easily many useful properties of the Chebyshev polynomials. In fact, letting $\theta=\cos^{-1}x$, it follows from (1.61)

$$2T_n(x) = \frac{1}{n+1}T'_{n+1}(x) - \frac{1}{n-1}T'_{n-1}(x),\ n \geq 2\ . \tag{1.62}$$

It can be easily shown by using (1.61) that

$$|T_n(x)| \leq 1, |T'_n(x)| \leq n^2\ ,$$

$$2T_m(x)T_n(x) = T_{m+n}(x) + T_{m-n}(x)\ ,\ m \geq n \geq 0\ . \tag{1.63}$$

One can also derive from (1.58) that,

$$T'_n(\pm 1) = (\pm 1)^{n-1} n^2, \tag{1.64a}$$

$$T''_n(\pm 1) = \tfrac{1}{3}(\pm 1)^n n^2 (n^2 - 1)\ . \tag{1.64b}$$

Moreover, we can derive from (1.62) that,

,

$$T'_n(x) = 2n \sum_{\substack{k=0 \\ n+k\ \text{odd}}}^{n-1} \frac{1}{c_k} T_k(x) \tag{1.65a}$$

$$T''_n(x) = \sum_{\substack{k=0 \\ n+k\ \text{even}}}^{n-2} \frac{1}{c_k} n(n^2 - k^2) T_k(x)\ . \tag{1.65b}$$

The Chebyshev expansion of a function $u \in L^2_w(-1,1)$ is

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k T_k(x)\ ;\ \hat{u}_k = \frac{2}{\pi c_k} \int_{-1}^{1} u(x) T_k(x) w(x)\, dx\ . \tag{1.66}$$

If we define the even periodic function $\tilde{u}$ by $\tilde{u}(\theta) = u(\cos\theta)$, then $\tilde{u}(\theta) = \sum_{k=0}^{\infty} \tilde{u}_k \cos k\theta$. Hence, the Chebyshev series for $u$ corresponds to a cosine series for $\tilde{u}(\theta)$ .

It is easy to verify that if $u(x)$ is infinitely differentiable on [–1, 1], then $\tilde{u}(\theta)$ is infinitely differentiable and periodic with all derivatives on $[0, 2\pi]$.

For the Chebyshev series, one can determine from (1.13), (1.16) and (1.18) the quadrature points and weights (see also[8]).

(i) For Chebyshev-Gauss quadrature:

$$x_j = \cos\frac{(2j+1)\pi}{2N+2}, w_j = \frac{\pi}{N+1}, 0 \le j \le N \ . \tag{1.67}$$

(ii) For Chebyshev-Gauss-Radau quadrature:

$$x_j = \cos\frac{2\pi j}{2N+1}, w_j = \frac{\pi}{2N+1}, j = 0, and\ \frac{2\pi}{2N+2}, 1 \le j \le N. \tag{1.68}$$

(iii) For Chebyshev-Gauss-Lobatto quadrature:

$$x_j = \cos\frac{\pi j}{N}, w_j = \frac{\pi}{\widetilde{c}_j N}, 0 \le j \le N, \tag{1.69}$$

where $\widetilde{c}_0 = \widetilde{c}_N = 2$ and $\widetilde{c}_j = 1$ for $j = 1,2,...,N-1$.

Note that, for simplicity of the notation, these points are arranged in descending order, namely, $x_N < x_{N-1} < ... < x_1 < x_0$.

As in the Legendre case, we present here some detailed results on the interpolation operator $I_N$ based on the Chebyshev-Gauss-Lobatto points $\{x_j\}_{j=0}^N$. For any function $u$ which is continuous on [-1, 1], we have

$$u(x_j) = I_N u(x_j) = \sum_{k=0}^{N} \widetilde{u}_k T_k(x_j). \tag{1.70a}$$

In this case, (1.24) reads:

$$\widetilde{u}_k = \frac{2}{\widetilde{c}_k N} \sum_{j=0}^{N} \frac{1}{\widetilde{c}_j} u(x_j) \cos\frac{kj\pi}{N}. \tag{1.70b}$$

The most important practical feature of Chebyshev series is that the discrete Chebyshev transforms (1.70a) and (1.70b) can be performed in $O(N \log_2 N)$ operations.

a) Differentiation in frequency space

Given $u=\sum_{k=0}^{N}\widetilde{u}_k T_k \in \overline{\mathrm{P}}_N$, we derive from (1.62) that

$$u'=\sum_{k=1}^{N}\widetilde{u}_k T_k'=\sum_{k=0}^{N}\widetilde{u}_k^{(1)}T_k=\widetilde{u}_0^{(1)}+u_{(1)}^{(1)}(x)+\sum_{k=2}^{N-1}\widetilde{u}_k^{(1)}\left(\frac{T_{k+1}'}{2(k+1)}-\frac{T_{k-1}'}{2(k-1)}\right)$$

$$=\frac{\widetilde{u}_{N-1}^{(1)}}{2N}T_N'+\sum_{k=1}^{N-2}\frac{1}{2k}\left(c_{k-1}\widetilde{u}_{k-1}^{(1)}-\widetilde{u}_{k+1}^{(1)}\right)T_k', \tag{1.71}$$

where $c_0=2$ and $c_k=1$ for $k\geq 1$. Comparing the coefficients of $T_k'$ we find that the Chebyshev coefficients $\{\widetilde{u}_k^{(1)}\}$ of $u'$ are determined by the recursive relation:

$$\begin{aligned}&\widetilde{u}_N^{(1)}=0,\widetilde{u}_{N-1}^{(1)}=2N\widetilde{u}_N\ ,\\&\widetilde{u}_{k-1}^{(1)}=(2k\widetilde{u}_k+\widetilde{u}_{k+1}^{(1)})/c_{k-1}\ ,k=N-1,N-2,...,1\ .\end{aligned} \tag{1.72}$$

Higher derivatives can be obtained by repeatedly applying the above formulae.

## b) Derivative matrices in physical space

To compute the derivative matrix in physical space, we can use the same notations as in the Legendre case except that now we choose $\{x_j=\cos\frac{j\pi}{N}\}$ to be the Chebyshev-Gauss-Lobatto points.

The Lagrange polynomials associated to the Chebyshev-Gauss-Lobatto points are

$$l_j(x)=\frac{(-1)^j(x^2-1)T_N'(x)}{\widetilde{c}_j N^2(x-x_j)}\ ,\ 0\leq j\leq N\ . \tag{1.73}$$

The derivative matrix $(d_{kj}=l_j'(x_k))$ is given by

$$\begin{aligned}&d_{kj}=\frac{\widetilde{c}_k(-1)^{k+j}}{\widetilde{c}_j(x_k-x_j)}\ ,j\neq k\\&d_{kk}=\frac{x_k}{2(1-x_k^2)}\ ,k=1,2,...,N\\&d_{00}=-d_{NN}=(2N^2+1)/6\ ,\end{aligned} \tag{1.74}$$

where $\tilde{c}_k = 1$ for $1 \le k \le N-1$ and $\tilde{c}_0 = \tilde{c}_N = 2$ .

***Remark 2*:** Remark 1 applies also to the Chebyshev case. However, in the Chebyshev case, a more efficient alternative for computing derivatives is to proceed in the frequency space as described earlier in this section:

(i) Compute the discrete Chebyshev coefficients $\{\tilde{u}_k\}$ of $u$ from $u(x_j) = \sum_{k=0}^{N} \tilde{u}_k T_k(x_j)$;

(ii) Compute the discrete Chebyshev coefficients $\{\tilde{u}_k^{(1)}\}$ of $u'$ using (1.72);

(iii) Compute $u'(x_j)$ from $u'(x_j) = \sum_{k=0}^{N} \tilde{u}_k^{(1)} T_k(x_j)$.

The cost of this approach is only $O(N \log N)$.

The most important feature of Chebyshev series is that their convergence properties are not affected by the values of $u(x)$ or its derivatives at the boundaries $x = \pm 1$ but only by the smoothness of $u(x)$ and its derivatives throughout $-1 \le x \le 1$. While Chebyshev expansions do not exhibit the Gibbs phenomenon at the boundaries $x = \pm 1$, they do exhibit the phenomenon at any interior discontinuity of $u(x)$.

In the end of this section we mention although spectral accuracy can be achieved for expansion in Jacobi polynomials(see Sec 9.6.1[5]), they have seen comparatively little use, aside, of course, from the special cases of Chebyshev $(\alpha = \beta = -1/2)$ and Legendre $(\alpha = \beta = 0)$ polynomials.

We also mention the Legendre series expansion of infinitely differentiable functions converges faster than algebraically. But, when a discontinuous function is expanded in Legendre series, the rate of convergence is no longer faster than algebraic.

## 1.5 Review of Fourier Transforms

A very large class of important computational problems falls under the general rubric of "*Fourier transform methods*" or "*spectral methods*".

For some of these problems, the Fourier transform is simply an efficient computational tool for accomplishing certain common manipulation of data. It is a very good example of a spectral method.

In this section we give a review of the Fourier transform and study the Fourier expansion for $2\pi$ -periodic functions.

Suppose $u$ is a $2\pi$ -periodic function. Let us expand $u$ as

$$u(x) \sim \sum_{k=-\infty}^{\infty} \hat{u}_k e^{ikx} .$$

By taking the inner product, defined by

$$(u,v) = \tfrac{1}{2\pi} \int_0^{2\pi} u(x)\bar{v}(x)dx , \tag{1.75}$$

we find that

$$\hat{u}_k = \tfrac{1}{2\pi} \int_0^{2\pi} u(x)e^{-ikx}dx . \tag{1.76}$$

The $\hat{u}_k$ are called the *Fourier coefficients,* or *Fourier transform* of $u$ at wavenumber $k$.

The integrals in (1.76) exist if $u$ is Riemann integrable which is ensured, for instance, if $u$ is bounded and piecewise continuous in $(0,2\pi)$. More generally, the Fourier coefficients are defined for any function which is integrable in the sense of Lebesgue (see Appendix A).

It is possible as well to introduce a *Fourier cosine transform* and a *Fourier sine transform* of $u$ respectively, through the formulae

$$a_k = \tfrac{1}{2\pi} \int_0^{2\pi} u(x)\cos kx dx \ , \ k = 0, \pm 1, \pm 2, ... \tag{1.77}$$

and

$$b_k = \tfrac{1}{2\pi} \int_0^{2\pi} u(x)\sin kx dx \ , \ k = \pm 1, \pm 2, ... \tag{1.78}$$

The three Fourier transforms of $u$ are related by the formulae $\hat{u}_k = a_{|k|} - ib_{|k|}$ for $k = 0, \pm 1, \pm 2, ....$

Note that the Chebyshev series is a Fourier cosine series with a change of variable. If $u$ is smooth, then its Fourier coefficients decay very fast. That is, the error between $u$ and its $N$-th order truncated Fourier series decay faster than algebraically in $1/N$, when $u$ is infinitely smooth and periodic with all its derivatives. Indeed, by taking integration by part $n$ times from (1.76), we will get

$$\hat{u}_k = \frac{1}{(-ik)^n}\frac{1}{2\pi}\int_0^{2\pi} u^{(n)}(x)e^{-ikx}dx\,. \tag{1.79}$$

Thus, if $u \in C^n$, then $\hat{u}_k = O\left(|k|^{-n}\right)$.

When $u$ is not so smooth, say in $L^1$, we still have $\hat{u} \to 0$ , as $|k| \to \infty$.

This is a consequence of the *Riemann-Lebesgue lemma* which states:

If $f$ is in $L^1(a,b)$, then

$\hat{f}_A := \int_a^b f(x)\sin(Ax)dx \to 0$ *as* $A \to \infty$ .

Let us denote the partial sum of the Fourier expansion by $u_N$:

$$u_N(x) = \sum_{k=-N}^{k=N} \hat{u}_k e^{ikx}\ . \tag{1.80}$$

We recall the following results about the convergence of Fourier series.

(i) If $u$ is continuous, periodic, and of bounded variation on$[0,2\pi]$, then $u_N$ is uniformly convergent to $u$, i.e.,

$$\max_{x\in[0,2\pi]}|u(x)-u_N(x)| \to 0, \qquad \text{as } N \to \infty.$$

(ii) If $u$ is of bounded variation on$[0,2\pi]$, then $u_N$ converges pointwise to $(u(x^+)+u(x^-))/2$ for any $x \in [0,2\pi]$.

(iii) If $u$ is continuous and periodic, its Fourier series does not necessarily converge at every

point $x \in [0, 2\pi]$.

A full characterization of the functions for which the Fourier series is everywhere pointwise convergent is not known. However, a full characterization is available within the framework on Lebesgue integration for convergence in the mean.

The series $Su$ is said to be convergent in the mean (or $L^2$ – convergent) to $u$ if

$\int_0^{2\pi} |u(x) - P_N(x)|^2 dx \to 0$, as $N \to \infty$. Clearly, the convergence in the mean can be defined for square integrable functions.

If $u(x)$ is smooth and periodic, its Fourier series does not exhibit the Gibbs phenomenon. The Fourier series of such a $u(x)$ converges rapidly and uniformly.

***Theorem* 6.** If $u$ is a $2\pi$ -periodic function and $u \in C^\infty$, then for any $n > 0$ there exists a constant $C_n$ such that

$$|u_N(x) - u(x)| \le C_n N^{-n} . \tag{1.81}$$

The constant $C_n$ is in general not big, as compared with the term $N^{-n}$. Hence, the approximation (1.80) is highly efficient for smooth functions. As we mentioned before, the accuracy property (1.81) is called spectral accuracy.

The Fourier transform maps a $2\pi$ -periodic function $u$ into its Fourier coefficients $\{\hat{u}_k\}_{k=-\infty}^{\infty}$. We may view the Fourier transform as a map from $L^2$ space into $l^2$ space. The function spaces $L^2$ and $l^2$ are defined below;

$$L^2 := \left\{ u : u \text{ is } 2\pi - \text{periodic and } \int_0^{2\pi} |u(x)|^2 dx < \infty \right\}, \tag{1.82}$$

with inner product given by (1.75) .The space $l^2$ is defined as

$$l^2 := \left\{ (a_k)_{k=-\infty}^{\infty} : \sum_k |a_k|^2 < \infty \right\}, \tag{1.83}$$

with inner product $(a,b)=\sum_k a_k\bar{b}_k$ .

It is easy to check that $e^{ikx}$ are orthogonal in $L^2$. From this, we have

$0\le(u-u_N,u-u_N)=\|u\|^2-\sum_{|k|<N}|\hat{u}_k|^2$, for any N.

Or, equivalently

$$\sum_{|k<N|}|\hat{u}_k|^2\le\|u\|^2 . \tag{1.84}$$

This is called the Bessel inequality. It says that the Fourier transform maps

continuously from $L^2$ to $l^2$ .

***Theorem 7.*** *(Isometry property)* The Fourier transform is an isometry from $L^2$ to $l^2:(u,v)=\sum_k \hat{u}_k\overline{\hat{v}_k}$ .

The isometry property says that: the Fourier transformation preserves the inner product.When $u=v$ in the above isometry property, we obtain the following *Parseval identity.*

For $u\in L^2$, we have , $\|u\|^2=\sum_k|\hat{u}_k|^2$ .

***Theorem 8.*** If $u\in L^2$, then $u_N\to u$ in $L^2$ as $N\to\infty$.

***Theorem 9.*** If $u$ is a function of bounded variation, then

$$u_N(x)=\sum_{k=-N}^{N}\tilde{u}_k e^{ikx}\to\frac{1}{2}(u(x^+)+u(x^-)).$$

In many practical applications, numerical methods based upon the Fourier transform can not be implemented in precisely the way suggested by the standard treatment of Fourier transform. Some of the difficulties are: the Fourier coefficients of an arbitrary function are not known in closed form and must therefore be approximated in some way; and there needs to be an efficient way to

recover in physical space the information that is calculated in transform space. The key to overcoming these difficulties is the use of the discrete Fourier transform.

Given a $2\pi$ -periodic function $u$. Let us sample $u$ by $u_j = u(x_j)$, where $x_j = {}^{2\pi j}/_{N}$. Define the discrete Fourier transform for the sampled data by

$$\tilde{u}_k = {}^{1}/_{N} \sum_{j=0}^{N-1} u_j e^{-ikx_j} \ . \tag{1.85}$$

This is exactly the trapezoidal approximation for numerical integration of the Fourier coefficients ${}^{1}/_{2\pi} \int_0^{2\pi} u(x) e^{-ikx} dx$. When $u \in C^{\infty}$, according to the Euler-MacLaurin summation formulae for periodic functions,

$$\left| {}^{1}/_{2\pi} \int_0^{2\pi} u(x) e^{-ikx} dx - {}^{1}/_{N} \sum_{j=0}^{N-1} u_j e^{-ikx_j} \right| = O(N^{-n}), \tag{1.86}$$

for any $n$.Thus, the discrete Fourier coefficients can approximate Fourier coefficients, with spectral accuracy, provided the underlying function belongs to $C^{\infty}$. From $\tilde{u}_k$, we define

$$I_N u(x) = \sum_{k=-\frac{N}{2}}^{k=\frac{N}{2}-1} \tilde{u}_k e^{ikx} \ . \tag{1.87}$$

We claim that $I_N u(x_j) = u(x_j)$. In other words, $I_N u$ is a trigonometric interpolant of $u$ at $\{x_j\}_{j=0}^{N-1}$.

To see this, we substitute the formulae for $\tilde{u}_k$ into the formulae for $I_N u(x)$:

$$\begin{aligned} I_N u(x) &= \sum_{k=-\frac{N}{2}}^{k=\frac{N}{2}-1} {}^{1}/_{N} \sum_{k=0}^{N-1} u_j e^{ik(x-x_j)} \\ &= {}^{1}/_{N} \sum_{j=0}^{N-1} D_N(x-x_j) u_j \ , \end{aligned} \tag{1.88}$$

where

$$D_N(x) = \sum_{k=-\frac{N}{2}}^{k=\frac{N}{2}-1} e^{ikx} = e^{-ix/2} \frac{\sin(Nx/2)}{\sin(x/2)}.$$

We find that

$$D_N(x_j) = \begin{cases} 1, & j=0 \\ 0, & j\neq 0 \end{cases}.$$

Hence, $I_N u(x_j) = u_j$.

Let $S_N$ be the space of the trigonometric polynomials of degree $\leq N/2$:

$$S_N = \text{span}\left\{E_k(x) = e^{ikx} : -N/2 \leq k < N/2\right\}.$$

In this space, the inner product defined by (1.75) is equivalent to the discrete inner product,

$$(u,v) = 1/N \sum_{j=0}^{N-1} u_j \bar{v}_j .$$

It is easy to check that $\{E_k(x)\}_{-N/2\leq k<N/2}$ are orthogonal in both inner products. Hence, these two inner products are identical for any $u, v \in S_N$. Again, from orthogonality of $\{E_k(x)\}$, we have the isometry $(u,v)_N = \sum_{-N/2\leq k<N/2} u_k \bar{v}_k$, and the Parseval identity : $1/N \sum_{j=0}^{N} |u_j|^2 = \sum_{-N/2\leq k<N/2} |\tilde{u}_k|^2$.

Given a $2\pi$-periodic function $u$, the mapping

$$P_N u(x) = \sum_{-N/2\leq k<N/2} \hat{u}_k e^{ikx}$$

is an orthogonal projection from $L^2(-\pi,\pi)$ to $S_N$. On the other hand, the interpolation operator, $I_N u$:

$$I_N u(x) = \sum_{-N/2\leq k<N/2} \tilde{u}_k e^{ikx}$$

is a projection onto $S_N$, and is characterized by $I_N u(x_j) = u(x_j)$, $j = 0,1,\ldots, N-1$.

The difference between $P_N$ and $I_N$ is called "*aliasing error*". It can be characterized as follows.

First,

$$\tilde{u} = \frac{1}{N}\sum_{j=1}^{N-1} u(x_j)e^{-ikx_j} = \frac{1}{N}\sum_{j=1}^{N-1}\sum_{l=-\infty}^{\infty}\hat{u}_l e^{i(l-k)x_j} = \sum_{m=-\infty}^{\infty}\hat{u}_{k+mN}$$

$$= \hat{u}_k + \sum_{\substack{-\infty \\ m\neq 0}}^{\infty}\hat{u}_{k+mN}\,.$$

From the orthogonality of $E_k$ in $L^2$, we see that

$$R_N u = I_N u - P_N u = \sum_{-N/2\le k<N/2}\left(\sum_{-\infty,m\neq 0}^{\infty}\hat{u}_{k+mN}\right)E_k\,.$$

Since $P_N$ is an orthogonal projection we have

$$\|u - I_N u\|^2 = \|u - P_N u\|^2 + \|R_N u\|^2\,.$$

It is not difficult to find the approximation error for $P_N$. Indeed, Let $H^s$ denote the *Sobolov space* of order $s$: $H^s = \{u : u \text{ is } 2\pi - periodic, u, ..., u^{(s)} \in L^2\}$,

with the norm $\|u\|_{H^s}^2 = \sum_{m=0}^{s}\|u^{(m)}\|^2$. From the Parseval identity, this norm is

equivalent to $\sum_k \left(1+|k|^2\right)^s |\hat{u}_k|^2$.

We have the following approximation theorem:

***Theorem* 10.** If $u \in H^s$, then $\|u - P_N u\| \le CN^{-s}\|u^{(s)}\|$.

For the interpolation operator, we have a similar result. In other words, the aliasing error has the same spectral error as that of the truncated Fourier polynomial for smooth functions.

***Theorem* 11.** If $u \in H^s, s \ge 1$, then $\|u - I_N u\| \le CN^{-s}\|u^{(s)}\|$.

This theorem was proved by Kreiss and Oliger (1979)[9].

How much computation is involved in computing the discrete Fourier transform of $N$ points? The discrete Fourier transform can, in fact, be computed in $O(N\log N)$ operations with an algorithm called the: "*Fast FourierTransform*" , or *FFT* . The *FFT* is a particular way of factoring and rearranging the terms in sums of the discrete Fourier transform brought to the attention of the scientific community by Cooley and Tukey[10]. Its importance lies in the drastic reduction in the number of numerical operations required . For $N$ time values (measurements) a direct calculation of a discrete Fourier transform would mean about $N^2$ multiplications, but the technique of Cooley and Tukey reduces the operation counts from $O(N^2)$ to $O(N\log N)$.

For example, if $N=1024(=2^{10})$, the *FFT* achieves a computational reduction by a factor of over 200 .This is why the *FFT* is called fast. Variations of *FFT* are the trigonometric representations which create Fast Cosine and Sine Transforms.

When all $u_j \in R$, then $\tilde{u}_k = \tilde{u}_{-k} = \tilde{u}_{N-k}$, for $k = 1, N/2$. Let

$$M = \begin{cases} N/2 \text{ , for even } N \\ (N+1)/2 \text{ , for odd } N \end{cases}$$

$\tilde{u}_k = c_{2k-1} - ic_{2k}, k = 1,...,M-1,$ and $c_0 = \tilde{u}_0$ and $c_{N-1} = u_{N/2}$ . Then

$$u_j = \tilde{u}_0 + (-1)^j u_{N/2} + \sum\left(\tilde{u}_k e^{ikx_j} + \overline{\tilde{u}_k} e^{-ikx_j}\right)$$

$$= c_0 + (-1)^j c_{N-1} + 2\sum_{k=1}^{M} c_{2k-1}\cos(kx_j) + c_{2k}\sin(kx_j)$$

and

$$c_0 = {}^1\!/_N \sum_{j=0}^{N-1} u_j ,$$

$$c_{2k-1} = {}^1\!/_N \sum_{j=0}^{N-1} u_j \cos(kx_j), k = 1,..., N/2 - 1 ,$$

$$c_{2k} = \frac{1}{N}\sum_{j=0}^{N-1} u_j \sin(kx_j), k = 1,...,N/2 ,$$

$$c_{N-1} = \frac{1}{N}\sum_{j=0}^{N-1} (-1)^j u_j .$$

When $u_j$ is an even sequence, i.e. $u_{N-j} = u_j , j = 1,...,N/2$, then for $k = 0,...,N/2-1$, we will have the Fourier Cosine Transform:

$$\tilde{u}_k = \frac{1}{N}\sum_{j=-N/2}^{N/2-1} u_j e^{-ikx_j} = \frac{1}{N}\left[u_0 + (-1)^k u_{N/2} + \sum_{j=1}^{N/2-1} 2u_j \cos(kx_j)\right], \quad (1.89)$$

with its inverse transform,

$$u_j = \sum_{j=-N/2}^{N/2-1} \tilde{u}_k e^{ikx_j} = u_0 + (-1)^j \tilde{u}_{N/2} + \sum_{k=1}^{N/2-1} 2u_k \cos(kx_j) .$$

But when $u_j$ is an odd sequence, i.e. $u_{N-j} = -u_j$ , $j = 0,...,N/2$, then for $k = 1,...,N/2-1$,

$$\tilde{u}_k = \frac{1}{N}\sum_{j=-N/2}^{N/2-1} u_j e^{-ikx_j} = \frac{1}{N}\sum_{j=1}^{N/2-1} 2u_j \sin(kx_j), \quad (1.90)$$

with its inverse transform,

$$u_j = \sum_{k=-N/2}^{N/2-1} \tilde{u}_k e^{ikx_j} = \sum_{k=1}^{N/2-1} 2u_k \sin(kx_j), \text{ for } j = 1,...,N/2.$$

The most efficient way to evaluate nonlinear and general non-constant terms in spectral approximations is to apply transform methods .The key idea is to apply FFT and other transforms to transform efficiently between spectral representations of a function $f(x)$ and physical-space representations of $f(x)$. With Chebyshev polynomial expansions, FFT permits the evaluation of arbitrary nonlinear and non-constant coefficients terms in order $N \log N$ arithmetic operations.

In the end of this chapter we emphasize that on a periodic interval, the sines and cosines of a Fourier series (which are the natural basis functions for all periodic problems) can be used. For non-periodic problems on a finite interval, which can always be rescaled and translated to

$x \in [-1,1]$, Chebyshev or Legendre polynomials are optional. Apparently for $x$ away from the boundaries $x = \pm 1$, the Legendre expansion has somewhat smaller errors, while near $x = \pm 1$ the Chebyshev expansion has smaller errors.

# Chapter 2

# An introduction to Spectral methods

## 2.1 Introduction

Spectral methods arise from the fundamental problem of approximation of a function by interpolation on an interval, and are very much successful for the numerical solution of ordinary or partial differential equations [13]. Since the time of Fourier (1882), spectral representations in the analytic study of differential equations have been used and their applications for numerical solution of ordinary differential equations refer, at least, to the time of Lanczos [14].

Spectral methods have become increasingly popular, especially, since the development of Fast transform methods, with applications in problems where high accuracy is desired. A survey of some applications is given in [4].

Spectral methods may be viewed as an extreme development of the class of discretization schemes for differential equations known generally as the *method of weighted residuals* (MWR) (Finlayson and Scriven (1966)) [18]. The key elements of the MWR are the trial functions (also called expansion approximating functions) which are used as basis functions for a truncated series expansion of the solution, and the test functions (also known as weight functions) which are used to ensure that the differential equation is satisfied as closely as possible by the truncated series expansion. The choice of such functions distinguishes between the three most commonly used spectral schemes, namely, Galerkin, collocation(also called pseudo-spectral) and Tau version. The Tau approach is a modification of Galerkin method that is applicable to problems with non-periodic boundary conditions. In broad terms, Galerkin and Tau methods are implemented in terms of the expansion coefficients, where as collocation methods are implemented in terms of physical space values of the unknown function.

The basis of spectral methods to solve differential equations is to expand the solution function as a finite series of very smooth basis functions, as follows

$$y_N(x) = \sum_{n=0}^{N} a_n \phi_n(x) \ , \tag{2.1}$$

in which, as pointed out in chapter 1, one of our choice of $\phi_n$, is the eigenfunctions of a singular Sturm-Liouville problem. If the solution is infinitely smooth, the convergence of spectral method is more rapid than any finite power of *1/N*. That is the produced error of approximation (2.1), when $N \to \infty$, approaches zero with exponential rate [13]. This phenomenon is usually referred to as "spectral accuracy", [4]. The accuracy of derivatives obtained by direct, term by term differentiation of a such truncated expansion naturally deteriorates [13]. Although there will be problem but for high order derivatives truncation and round off errors may deteriorate, but for low order derivatives and sufficiently high-order truncations this deterioration is negligible. So, if the solution function and coefficient functions of the differential equation are analytic on$[a,b]$, spectral methods will be very efficient and suitable. We call function $y$ is analytic on $[a,b]$if is infinitely differentiable and with all its derivatives on this interval are bounded variation.

## 2.2 Spectral methods

In this section, we are briefly going to introduce spectral methods. For this reason, first we consider the following differential equation:

$$Ly = \sum_{0}^{M} f_{M-i}(x) D^i y = f(x),\ x \in [-1,1], \tag{2.2}$$

$$\mathrm{B}y = C \ , \tag{2.3}$$

where $L = \sum_{0}^{M} f_{M-i}(x) D^i$, and $f_i$ , $i = 0,1,...,M, f$ , are known real functions of $x, D^i$ denotes $i^{th}$ order of differentiation with respect to $x$,B is a linear functional of rank $M$ and $C \in \Re^M$.

Here (2.3) can be initial, boundary or mixed conditions. The basis of spectral methods to solve this class of equations is to expand the solution function, $y$, in (2.2) and (2.3) as a finite series of very smooth basis functions, as given below

$$y_N(x) = \sum_{n=0}^{N} a_n T_n(x) \ , \tag{2.4}$$

where, $\{T_n(x)\}_0^N$ is sequence of Chebyshev polynomials of the first kind, defined in (1.61). By replacing $y_N$ in (2.2), we define the residual term by $r_N(x)$ as follows

$$r_N(x) = Ly_N - f \ . \tag{2.5}$$

In spectral methods, the main target is to minimize $r_N(x)$, throughout the domain as much as possible with regard to (2.3), and in the sense of pointwise convergence. Implementation of these methods leads to a system of linear equations with $N+1$ equations and $N+1$ unknowns $a_0 \ , a_1 \ , \ldots , a_N$.

In the rest of this section, we discuss three spectral methods, namely, Tau, Galerkin and collocation (also known as pseudo- spectral) methods, and use them for numerical solution of second order linear differential equations. It is to be noted that this discussion can be extended to the general form (2.2), (2.3) .

### *(i) Tau method*

The Tau method was invented by Lanczos in 1938[52]. The expansion functions $\phi_n \ (n = 1,2,3,\ldots)$ are assumed to be elements of a complete set of orthonormal functions. The approximate solution is assumed to be expanded in terms of those functions as $u_N = \sum_{n=1}^{N+m} a_n \phi_n$, where $m$ is the number of independent boundary constraints $Bu_N = 0$ that must be applied. Here we are going to use a Tau method developed by Clenshaw[53] for the solution of linear ODE in terms of a Chebyshev series expansion.

Consider the following differential equation:

$$P(x)y''+Q(x)y'+R(x)y=S(x)\,,x\in(-1,1)\,,\quad y(-1)=\alpha\,,\ y(1)=\beta\,. \tag{2.6}$$

First, for an arbitrary natural number $N$, we suppose that the approximate solution of equations (2.6) is given by (2.4). Our target is to find $\underline{a}=(a_0,a_1,...,a_N)^t$. For this reason, put

$$\begin{aligned} P(x)&\cong\sum_{i=0}^{N}\xi_i T_i(x),\\ Q(x)&\cong\sum_{i=0}^{N}\gamma_i T_i(x),\\ R(x)&\cong\sum_{i=0}^{N}\lambda_i T_i(x). \end{aligned} \tag{2.7}$$

Using (1.66), we can find coefficients $\xi_i$, $\gamma_i$ and $\lambda_i$ as follows:

$$\begin{aligned} \xi_i&=\frac{2}{\pi c_i}\int_{-1}^{1}\frac{P(x)T_i(x)}{\sqrt{1-x^2}}dx\\ \gamma_i&=\frac{2}{\pi c_i}\int_{-1}^{1}\frac{Q(x)T_i(x)}{\sqrt{1-x^2}}dx\\ \lambda_i&=\frac{2}{\pi c_i}\int_{-1}^{1}\frac{R(x)T_i(x)}{\sqrt{1-x^2}}dx, \end{aligned} \tag{2.8}$$

where, $c_0=2$ and $c_i=1$, for $i\geq 1$.

To compute the right-hand side of (2.8) it is sufficient to use an appropriate numerical integration method. Here, we use $(N+1)$- point Gauss-Chebyshev-Lobatto quadrature (1.69) with weights

$$\begin{aligned} w_k&=\frac{\pi}{N}\ ,\quad 1\leq k\leq N-1\,,\\ &=\frac{\pi}{2N}\,,\quad k=0\,,k=N\,, \end{aligned}$$

and nodes $x_k=\cos\frac{\pi k}{N}$, $k=0,1,...,N$. That is, we put, [15]:

$$\xi_i\cong\frac{\pi}{N}\sum_{k=0}^{N''}P(\cos(\frac{k\pi}{N}))T_i(\cos(\frac{k\pi}{N}))\,,$$

and using $T_i(x) = \cos(i\cos^{-1}x)$, we get

$$\xi_i \cong \frac{\pi}{N}\sum_{k=0}^{N''} P(\cos(\frac{k\pi}{N}))\cos(\frac{\pi ik}{N}),$$

where, notation $\sum''$ means first and last terms become half .Therefore, we will have :

$$\xi_i \cong \frac{\pi}{N}\sum_{k=0}^{N''} P(\cos(\frac{k\pi}{N}))\cos(\frac{\pi ik}{N}),$$

$$\gamma_i \cong \frac{\pi}{N}\sum_{k=0}^{N''} Q(\cos(\frac{k\pi}{N}))\cos(\frac{\pi ik}{N}), \tag{2.9}$$

$$\lambda_i \cong \frac{\pi}{N}\sum_{k=0}^{N''} R(\cos(\frac{k\pi}{N}))\cos(\frac{\pi ik}{N}).$$

Now, substituting (2.4) and (2.9) in equations (2.6), and using the fact that

$$y'(x) \cong \sum_{m=0}^{N} a_m^{(1)} T_m(x),\ a_m^{(1)} = \frac{2}{c_m}\sum_{\substack{p=m+1 \\ m+p=\text{odd}}}^{N} p a_p,\ m = 0,1,\ldots,N-1,\ a_N{}^{(1)} = 0,$$

$$y''(x) \approx \sum a_m^{(2)} T_m(x),\ a_m^{(2)} = \frac{1}{c_m}\sum_{\substack{p=m+2 \\ m+p=\text{even}}}^{N} p(p^2-m^2)a_p,\ m = 0,1,\ldots,N-2,\ a_{N-1}^{(2)} = a_N^{(2)} = 0,$$

in this manner, we get

$$\sum_{i=0}^{N}\sum_{m=0}^{N} \xi_i a_m^{(2)} T_i(x)T_m(x) + \sum_{i=0}^{N}\sum_{m=0}^{N} \gamma_i a_m^{(1)} T_i(x)T_m(x) + \sum_{i=0}^{N}\sum_{m=0}^{N} \lambda_i a_m T_i(x)T_m(x) = S(x), \tag{2.10}$$

$$\sum_{i=0}^{N} a_i T_i(-1) = \alpha,$$
$$\sum_{i=0}^{N} a_i T_i(1) = \beta. \tag{2.11}$$

Now, we multiply both sides of (2.10) by $\frac{2}{\pi c_j}\frac{T_j(x)}{\sqrt{1-x^2}}$, and integrate from -1 to 1, to obtain

$$\frac{2}{\pi c_j}\sum_{i=0}^{N}\sum_{m=0}^{N}[\xi_i a_m^{(2)}+\gamma_i a_m^{(1)}+\lambda_i a_m]\int_{-1}^{1}\frac{T_i(x)T_m(x)T_j(x)}{\sqrt{1-x^2}}dx$$

$$=\frac{2}{\pi c_j}\int_{-1}^{1}\frac{S(x)T_j(x)}{\sqrt{1-x^2}}dx,\ j=0,1,...,N-2, \tag{2.12}$$

where,

$$\int_{-1}^{1}\frac{T_i(x)T_m(x)T_j(x)}{\sqrt{1-x^2}}dx=\begin{cases}\pi & ,\ i=m=j=0\ ,\\ \frac{\pi}{2}\delta_{i,m} & ,\ i+m>0\ ,j=0\ ,\\ \frac{\pi}{4}(\delta_{j,i+m}+\delta_{j,|i-m|}) & ,\ j>0\ ,\end{cases} \tag{2.13}$$

with, $\delta_{i,j}=1$,when $i=j$, and zero when $i\neq j$ [16] .

We can also compute the integrals in the right-hand side of (2.12) by the method of numerical integration using $N+1$ -point Gauss-Chebyshev-Lobatto quadrature (1.69). Therefore, substituing (2.13) in (2.12) and using the fact that $T_i(\pm1)=(\pm1)^i$ , equations (2.12) and (2.11) make a system of $N+1$ equations for $N+1$ unknowns $a_0,a_1,...,a_N$ , and we can obtain $(a_0,a_1,...,a_N)^t$ from this system. It should be noted that the implementation of the Tau method for the numerical solution of a system of two differential equations, is similar to the method explained for differential equation (2.6), in such a way that first for an arbitrary natural number $N$ we put

$$y_1(x)\cong\sum_{i=0}^{N}a_iT_i(x),$$

$$y_2(x)\cong\sum_{i=0}^{N}a_{i+N+1}T_i(x).$$

Then expand the coefficient functions in terms of Chebyshev polynomials; with these relations in the given problem. Then multiply both sides of the resulting equations by

$$\frac{2}{\pi c_j}\int_{-1}^{1}\frac{T_j(x)}{\sqrt{1-x^2}}dx,\ j=0,1,...,N-1,$$

to have $2N$ -linear equations. Adding boundary conditions, then solve this system to have $2N+2$ coefficients $a_0, a_1, ..., a_{2N+1}$.

In chapter four we will refer to the numerical solution of a system of linear differential equations, when we apply the Tau method to the numerical solution of DAEs.

### *(ii) Pseudo-spectral method*

For implementation of the pseudo-spectral method for numerical solution of ODE, we use a matrix method which is simpler than Tau method.

We consider again the equation (2.6), and suppose that the approximate solution of this equation is given by (2.4), where $\underline{a} = (a_0, a_1, ..., a_N)^t \in \Re^{N+1}$ are expansion coefficients and $\{T_n(x)\}_0^N$ is the sequence of Chebyshev polynomials of the first kind.

Now if we put

$$y_N(x) = \sum_{k=0}^{N} a_k T_k(x), \tag{2.14}$$

then corresponding to functions $y_N, y'_N$ and $y''_N$, we can define matrices $A^{(0)}, A^{(1)}$ and $A^{(2)}$ as follows [17]:

$$y_N \equiv A^{(0)}, \quad A^{(0)} = I_{(N+1)\times(N+1)} \tag{2.15}$$

$$y'_N \equiv A^{(1)}, \quad (A)_{ij}^{(1)} = \begin{cases} (\frac{1}{c_i}) \times 2i, \text{for } j > i, i+j = odd, \\ 0, \text{otherwise}. \end{cases} \tag{2.16}$$

$$A^{(1)} = \begin{pmatrix} 0 & 1 & 0 & 3 & 0 & \dots \\ & 0 & 4 & 0 & 8 & \dots \\ & & 0 & 6 & 0 & \dots \\ & & & 0 & 8 & \dots \\ & & & & \cdot & \\ & & & & \cdot & \\ & & & & \cdot & \end{pmatrix}$$

$$y_N'' \equiv A^{(2)}\ ,\ \ (A^{(2)})_{ij} = \begin{cases} (\frac{1}{c_i}) \times (j-i)\, j\, (j+i)\ , for\ \ j > i\ , i+j = even\ , \\ 0\ , otherwise\ . \end{cases} \tag{2.17}$$

$$A^{(2)} = \begin{pmatrix} 0 & 0 & 4 & 0 & 32 & \dots \\ & 0 & 0 & 24 & 0 & \dots \\ & & 0 & 0 & 48 & \dots \\ & & & 0 & 0 & \dots \\ & & & & \cdot & \\ & & & & \cdot & \\ & & & & \cdot & \end{pmatrix},$$

where, $N \geq j, i \geq 0$, and $\quad c_i = \begin{cases} 2\ ,\ i = 0\ , \\ 1\ ,\ i > 0\ . \end{cases}$

Now using differential equation (2.6), we define matrix $AA(x)$ as follows;

$$AA(x) = P(x)A^{(2)} + Q(x)A^{(1)} + R(x)A^{(0)}\ , \tag{2.18}$$

that is,

$$AA(x) = \begin{pmatrix} R(x) & Q(x) & 4P(x) & 3Q(x) & 32P(x) & \dots \\ & R(x) & 4Q(x) & 24P(x) & 8Q(x) & \dots \\ & & R(x) & 6Q(x) & 48P(x) & \dots \\ & & & R(x) & 8Q(x) & \dots \\ & & & & R(x) & \dots \\ & & & & \cdot & \\ & & & & \cdot & \\ & & & & \cdot & \end{pmatrix}_{(N+1)\times(N+1)}$$

Hence, differential equation (2.6) converts to

$$\sum_{i=0}^{N} a_i \varphi_i(x) \cong S(x) \ , \tag{2.19}$$

in which,

$$\varphi_i(x) = \sum_{k=0}^{i} (AA)_{ki} T_k(x) \ , \tag{2.20}$$

that is,

$$\varphi_0(x) = R(x)T_0(x),$$
$$\varphi_1(x) = Q(x)T_0(x) + R(x)T_1(x),$$
$$\varphi_2(x) = 4P(x)T_0(x) + 4Q(x)T_1(x) + R(x)T_2(x),$$
$$\vdots$$

It must be noted that, if $A^{(k+2)}; k \geq 1$ is the corresponding matrix of $(k+2)th$ order differentiation

of $y_N(x)$, it follows that [17]:

$$A_{ij}^{(k+2)} = A_{ij}^{(k)} \frac{(j-i-k)(j+i+k)(j+i-k)}{4k(k+1)} \ , 0 \leq i, j \leq N.$$

Now, if we impose boundary condition of (2.6) on $y_N(x)$ we will have:

$$y_N(-1) = \alpha \Rightarrow \sum_{k=0}^{N} a_k T_k(-1) = \sum_{k=0}^{N} a_k (-1)^k = \alpha \ ,$$
$$y_N(1) = \beta \Rightarrow \sum_{k=0}^{N} a_k T_k(1) = \sum_{k=0}^{N} a_k = \beta \ .$$

So

$$\begin{bmatrix} 1 & -1 & 1 & \dots & (-1)^N \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{2.21}$$

Relation (2.21) forms a system with two equations and $N+1$ unknowns. To construct the remaining $N-1$ equations we substitute points $x_j = \cos\left(\pi j/N\right), j = 1,2,...,N-1,$ in (2.19) and put

$$\sum_{i=0}^{N} a_i \varphi_i(x_j) = S(x_j)\ ,\ j = 1,2,...,N-1, \tag{2.22}$$

to obtain $N-1$ equations.

### *(iii) Galerkin method*

This method is similar to the Tau method, where $N-1$ basis functions $\phi_2, \phi_3, ..., \phi_N$ are obtained through Chebyshev polynomials $T_2, ..., T_N$, in order to satisfy both of boundary conditions of (2.6). Then we multiply both sides of (2.19) by

$$\int_{-1}^{1} \frac{T_j(x)}{\sqrt{1-x^2}} dx, \quad j = 2,3,...,N,$$

to obtain $N-1$ equations.

In the next chapter we are will consider some ordinary differential equations with the Tau method (as representative of Tau and Galerkin methods) and the pseudo-spectral method and discuss the results.

# Chapter 3

# Numerical solutions of linear Ordinary Differential Equations

## 3.1 Introduction

Consider the following linear differential equation:

$$Ly = \sum_{i=0}^{M} f_{M-i}(x) D^i y = f(x) \ , \ x \in [-1,1] \ , \tag{3.1}$$

$$\mathrm{B}y = C \ , \tag{3.2}$$

where $f_i$ , $i = 0,1,...,M$ , $f$ , are known real functions of $x$, $D^i$ denotes $i^{th}$ order of differentiation with respect to $x$, B is a linear functional of rank $M$ and $C \in \Re^M$ .

We know that, if the function $y$ belongs to $C^{\infty}[a,b]$, then with use of eigenfunctions of a singular Sturm-Liouville problem we can approximate it in a form of finite series of eigenfunctions such that, the produced error of approximation, when $N$ tends to infinity, approaches zero with exponential rate [13]. As we know, this phenomenon is usually referred to as "spectral accuracy", [4]. As we mentioned in chapter two, the accuracy of derivatives obtained by direct, term by term differentiation of such truncated expansion naturally deteriorates [13], but for low order derivatives and sufficiently high-order truncations this deterioration is negligible. So, if the solution function and coefficient functions of the differential equation are analytic on $[a,b]$, spectral methods will be very efficient and suitable. We mention once more that the function $y$ is analytic on $[a,b]$, if belongs to $C^{\infty}[a,b]$ and with all its derivatives are of bounded variations on this interval.

We may have different cases: One in which all coefficient functions and the solution are not analytic and another in which at least one of the coefficient functions or the solution function is not analytic, all together being three different cases.

In this chapter we devote our attentions more on those problems that have non-analytical solution on $[a,b]$, or they have non-analytical coefficient functions and in the end we suggest a modified spectral method [20, 21].

## 3.2 Numerical solutions by spectral methods

We start this section with an example that has analytic coefficient functions and analytic solution function.

*Problem 3.1:* Let us consider

$$y''(x)+xy'(x)+y=x\cos(x)\,,\ x\in[-1,1],$$
$$y(-1)=\sin(-1)\,,\ y(1)=\sin(1)\,,$$

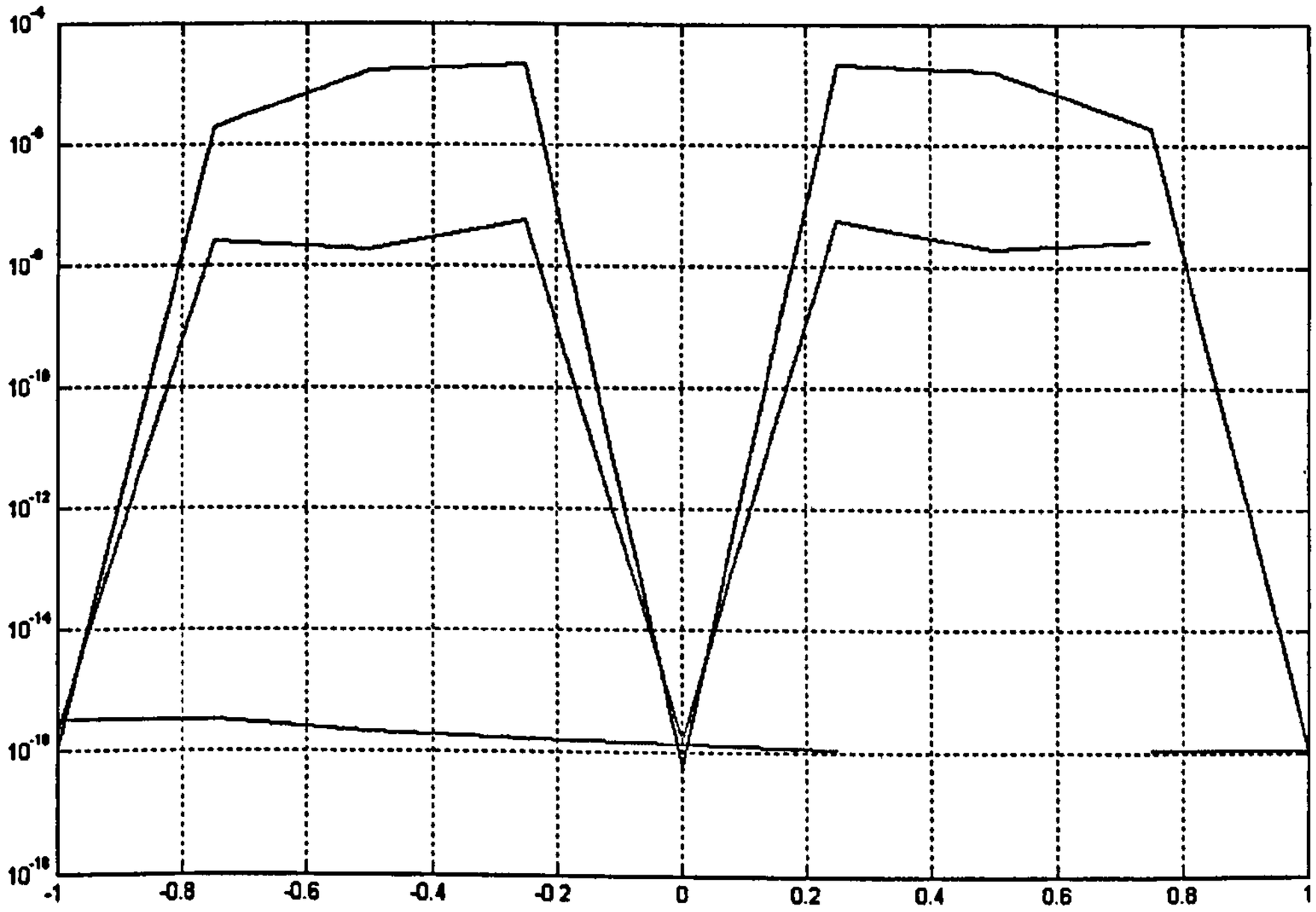
with the exact solution $y(x)=\sin(x)$. We solved it by Runge-Kutta with orders two and four and also Adams method. The maximum errors are $2.5\times10^{-4}, 2.4\times10^{-7}, 1.1\times10^{-5}$ , respectively. That is, these methods give good results for such problems. For these methods we used the same step size and step number. We also solved it by the Tau method with $N=5,8,16$, the maximum errors produced from this method are given in Table 1, where $y_\tau(x)$ means the Tau method.

Table 1

| $N$ | $\|y(x)-y_\tau(x)\|_\infty$ |
|---|---|
| 5 | $1.6\times10^{-5}$ |
| 8 | $1.6\times10^{-7}$ |
| 16 | $3.3\times10^{-16}$ |

This table shows the power of spectral methods. We also plot these results on a graph, with $N$ against the *log* of the errors in this interval. Results were shown in Fig 1.

Fig 1




Let's consider another problem.

*Problem 3.2:* Consider

$$y'' + xy' - y = f(x),\ x \in (-1,1),$$
$$y(\pm 1) = e^{\pm 5} + \sin(1),$$

where, $f(x) = (24+5x)e^{5x} + (2+2x^2)\cos(x^2) - (4x^2+1)\sin(x^2)$,

so that the exact solution is $y(x) = e^{5x} + \sin(x^2)$.

For comparison, we solved this problem by finite difference method, using the central differences for the derivatives. The mesh points are given by $x_i = -1 + ih$, $h = \frac{2}{N}$. The maximum errors given by this method are

$3.100, 7.898\times10^{-1}, 1.984\times10^{-1}, 4.968\times10^{-2}, 1.242\times10^{-2}, 3.106\times10^{-3},$

for $N = 16, 32, 64, 128, 256, 512$ , respectively.

Here the solution function and all coefficient functions are analytic on [−1,1], hence if we solve it by spectral methods, we obtain a spectral accuracy.

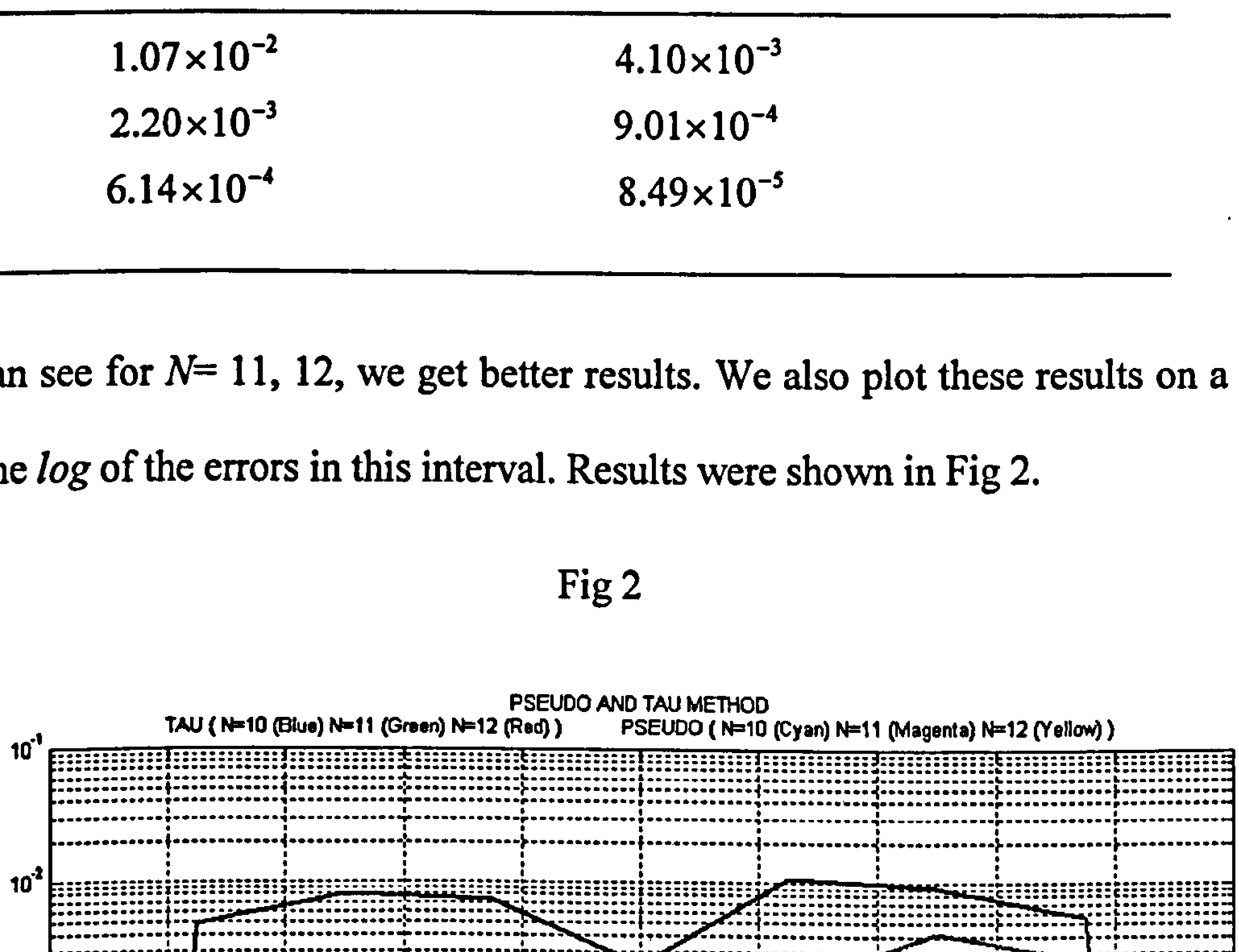
We solved it by Tau, $\tilde{y}_\tau(x)$ , (as representative of Tau and Galerkin methods) and pseudo-spectral, $\tilde{y}_{ps}(x)$, methods. Table 2 shows the results of solving this problem by these methods.

Table 2

| $N$ | $\|y(x)-\tilde{y}_\tau(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ps}(x)\|_\infty$ |
|---|---|---|
| 10 | $1.07\times10^{-2}$ | $4.10\times10^{-3}$ |
| 11 | $2.20\times10^{-3}$ | $9.01\times10^{-4}$ |
| 12 | $6.14\times10^{-4}$ | $8.49\times10^{-5}$ |

As we can see for $N$= 11, 12, we get better results. We also plot these results on a graph, with $N$ against the *log* of the errors in this interval. Results were shown in Fig 2.

Fig 2



Now we consider a problem which is a little bit different from *3.1* and *3.2*.

*Problem 3.3:* Consider

$$y''(x)+\frac{1}{x}y'(x)=\left(\frac{8}{8-x^2}\right)^2 \ , x\in(0,1)\ ,$$
$$y(1)=0\ , y'(0)=0\ ,$$

with the exact solution $y(x)=2\ln\left(\frac{7}{8-x^2}\right)$.

This problem was chosen from [19]. It was solved by the extrapolation method with maximum error of $10^{-8}$. Here we solved it by the Tau method for different values of $N$, and the results are given in Table 3.

Table 3

| N | Maximum error |
|---|---|
| 5 | $5\times10^{-5}$ |
| 15 | $2\times10^{-6}$ |
| 20 | $8\times10^{-7}$ |
| 30 | $5\times10^{-7}$ |
| 95 | $8\times10^{-8}$ |

As we see, when $N$ increases the rate of improvement of accuracy is very low. This is because of the lack of smoothness of the coefficient function. But, when we solved it by the pseudo-spectral method, since coefficient functions do not need expansion in the form of (2.9), the error produced from using (2.22), will be better than (2.12). In the above example by the Tau method with $N=95$, maximum error was about $8\times10^{-8}$, but by the pseudo-spectral method with $N=18$ we come to a maximum error about $4\times10^{-19}$. Therefore, in this case, this method is more successful than the Tau method. In this problem we do not know whether the solution function is analytic or not, although non-analyticity of solution function is recognized from the low rate of decrease in the coefficients of expansion of $y$, but we did not try to multiply both sides by $x$ because we want

to see results of these two methods for such problems. In general, it will be better to multiply both sides by $x$, even we do not know the solution function.

As we mentioned, we may have three different cases. The rest of this chapter devoted to consider these three different cases with some examples.

***Case 1:* ODEs with analytic coefficient functions and analytic solution function.**

We already considered two examples for such case. We finish with this case by considering one more example.

*Problem 3.4.* Consider

$$y''(x)+3xy'(x)+x^4y(x)=6x+9x^3+x^7,\ x\in[-1,1],$$
$$y(-1)=-1\ ,\ y(1)=1\ ,$$

with exact solution $y(x)=x^3$.

We solved it by Runge-Kutta with orders two and four and also Adams method. For these methods we used, again, the same step size and step number. The maximum errors were $1.5\times10^{-5}, 1.8\times10^{-7}, 1.2\times10^{-5}$ , respectively. We, also, solved it by shooting method with the same step size for steps $N$=14, 17. We had maximum errors $2.9\times10^{-5}, 1.3\times10^{-5}$, respectively. As we see the rate of improvement of accuracy is very low. But we used the Tau method with Legendre basis for $N=14$ and with Chebyshev basis for $N=17$ .The maximum errors were about $2.2\times10^{-13}$, and $3.4\times10^{-14}$, respectively, and we used the pseudo-spectral method with $N=14,17$, and maximum errors were $8.9\times10^{-16}$, and $4.4\times10^{-16}$, respectively. As we can see, spectral methods for solving such problems have high rate of convergency. Existence of $x^7$, indicates when $N$ get the value 7, the error becomes zero. If we observe above errors they are rounding errors.

If $y(x)$ be a non-analytic function on $[-1,1]$, and

$$y(x)\cong\sum_{i=0}^{N}a_iT_i(x), \tag{3.3}$$

then the order of infinite norm of the error of approximation (3.3) is given in Table 4 [17]

Table 4

| Function | Order of infinite norm of the error | |
|---|---|---|
| | Close to singular points | Far from singular points |
| *y, discontinuous* | 1 | $\frac{1}{N}$ |
| *y', discontinuous* | $\frac{1}{N}$ | $\frac{1}{N^2}$ |
| *y", discontinuous* | $\frac{1}{N^2}$ | $\frac{1}{N^3}$ |
| . | . | . |
| . | . | . |
| . | . | . |
| *y, anaylytic* | $e^{-CN}$ | $C>0$ |

In the regular case ( $p(x) \neq 0$), the decreased rate of expansion coefficients in (3.3) is similar to the order of infinite norm of error at the points which are far away from singular points. Here, the cases in which, at least, one of the coefficient functions or solution function is not analytic will be studied, and produced difficulties in solving differential equation (3.1) and (3.2) will be considered. Let us consider a problem of this kind.

*Case 2:* **Solution function is analytic but, at least, one of the coefficient functions is not analytic.**

*Problem 3.5:* Consider

$$y''(x)+|x|y'(x)-y(x)=|x|\exp(x)\ ,\ x\in[-1,1]\ ,$$
$$y(-1)=\exp(-1)\ ,y(1)=\exp(1)\ .$$

The exact solution is $y(x)=\exp(x)$.

In this case, if we solve (2.6) by the Tau method, the error produced from expansion of, at least, one of the coefficient functions in (2.9), in using system (2.12) (comparing with approximate error of solution function) the error is considerable and causes undesirable effect on the final solution. Now, if this problem is solved by the pseudo-spectral method, since the solution function, y, is analytic and coefficient functions do not need expansion in the form of (2.9), the error produced from using system (2.22), is much less than (2.12). Therefore, in this case, the pseudo-spectral method is more successful than the Tau method.

We solve this problem by shooting method and pseudo-spectral method, where in shooting method we used $N$ as number of steps.

For $N = 5,8,16$ the results are given in Table 5.

Table 5

| $N$ | $\|y(x)-y_{ps}(x)\|_\infty$ | $\|y(x)-y_{sh}(x)\|_\infty$ |
|---|---|---|
| 5 | $1.1\times10^{-4}$ | $1.4\times10^{-4}$ |
| 8 | $2.8\times10^{-8}$ | $2.0\times10^{-5}$ |
| 16 | $4.4\times10^{-16}$ | $1.3\times10^{-6}$ |

In this table, we used $y_{sh}(x)$ for results of shooting method.

Let us consider another problem of this kind.

*Problem 3.6:* Consider

$$y''+|x|y'+\sqrt[3]{(x^2-\frac{1}{4})^2}\,y=e^x(1+|x|+\sqrt[3]{(x^2-\frac{1}{4})^2}),\ x\in[-1,1],$$

$$y(-1)=e^{-1},\ y(1)=e,$$

with exact solution $y(x)=e^x$.

In this problem, the solution function is analytic on [−1,1] but $Q(x)$ and $R(x)$ are not analytic functions. Table 6 shows the results of solving this problem by the Tau, $\tilde{y}_\tau$, and pseudo-spectral, $\tilde{y}_{ps}$, methods.

Table 6

| $N$ | $\|y(x)-\tilde{y}_\tau(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ps}(x)\|_\infty$ |
|---|---|---|
| 8 | $3.13\times10^{-6}$ | $3.24\times10^{-8}$ |
| 11 | $6.40\times10^{-8}$ | $2.52\times10^{-12}$ |
| 16 | $3.92\times10^{-8}$ | $3.50\times10^{-18}$ |

As we can see again in this case, when $N$ increases the rate of improvement of accuracy by the Tau method is low. But, by the pseudo-spectral method, the error produced decreases rapidly.

Now we are going to consider third case, where the solution function and, at least, one of the coefficient functions are not analytic.

***Case 3:*** **The solution function and, at least, one of the coefficient functions are not analytic.**

Let us start with a problem that was chosen from [22].

*Problem 3.7:* Consider

$$y''(x)+|x|y'(x)+y(x)=|6x|+|x^3|+3x^3,\ x\in[-1,1],$$
$$y(-1)=y(1)=1,$$

with exact solution $y(x)=|x^3|$.

We solved it first by Runge-Kutta and Adams methods with different order with the same step size. The maximum errors were, nearly, 4.0 for both of these methods. We also used shooting method to solve it. Unfortunately, the method does not work for odd number of $n$, and for even value of $n$, although, it works but rate of improvement of accuracy is low, where $n=(b-a)/h$.

Let us consider this example, more carefully, by spectral methods.

In this case, since the order of infinite norm of the error produced from approximation of solution function and, at least, one of the coefficient functions follow from Table 4, system (2.12) or (2.22) does not produce accurate results. The above case results from the fact that the representing matrices $A^{(0)}, A^{(1)}, A^{(2)}$, and $AA$, because of existing a non-analytic solution function, accompany the error which is indispensable. However, if other errors are involved in the process of solving the problem, the error resulting from using the systems in the Tau method increases, and we will be far from the real solution. A modified spectral method is suggested in [22], in such a way that considerably decreases the error in setting up the system of equations. Before we go further, let's consider another problem to see what happens.

*Problem 3.8:* Consider

$$y''(x)+e^{\frac{1}{x}}y'(x)+y(x)=6x+x^3+3x^2e^{\frac{1}{x}}, x\in[-1,1],$$
$$y(-1)=-1,\ y(1)=1,$$

with exact solution, $y(x)=x^3$. Here, we have an essential singularity. Because of this singularity the shooting method does not work. If we choose N as an even number the Tau and pseudo-spectral methods do not work either, even with the modified method introduced in [22].

Here we are going to introduce another modified method which produces better results, and works very well even whenever the methods considered above do not work. The idea comes from this observation. When we checked the coefficient matrix of the resulting system of equations we found that difficulty arises from the middle row (N/2-row). This is because of the initial condition which produces (2.21) and elements in this row, are infinity. This happens whenever $j=\frac{N}{2}$ in (1.69). Now if we take the infinity, as a factor, out of the determinant of this matrix we get a matrix with two rows the same. Hence the determinant becomes zero. To avoid this difficulty we choose one of these elements close to 1. For example, 0.99999 and continue the process. We examined this choice and solved this problem with this idea and the results are given in Table 7.

Table 7

| $N$ | $\|y(x)-\tilde{y}(x)\|_{mps}$ |
|---|---|
| 5 | $2.2204\times10^{-15}$ |
| 8 | $1.6653\times10^{-15}$ |
| 12 | $1.3843\times10^{-15}$ |

Although the above idea does not work for problems with essential singularity, when we change the boundary condition, for example, from *y(1)=1* to $y(1)=1+\varepsilon$; but for any example without such a point it works. To see the results let's go back to problem 3.8. Here, we used the above idea for this problem and the results are given in Table 8.

Table 8

| $N$ | $\|y(x)-\tilde{y}_t(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ps}(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ms1}(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ms2}(x)\|_\infty$ |
|---|---|---|---|---|
| 8 | $8.98\times10^{-2}$ | $1.21\times10^{-1}$ | $3.82\times10^{-3}$ | $2.41\times10^{-3}$ |
| 15 | $1.54\times10^{-2}$ | $1.76\times10^{-2}$ | $5.61\times10^{-4}$ | $1.85\times10^{-4}$ |
| 20 | $1.68\times10^{-2}$ | $1.92\times10^{-2}$ | $1.81\times10^{-4}$ | $1.28\times10^{-4}$ |

Here, this method gives a little bit better results than those given in [22], which are notified by ms1. The notations ms1 and ms2 indicate the modified methods in [22] and here, respectively.

The main question is, why will we get better results? This improvement happens because the condition number has been reduced substantially. When we considered the condition number of the coefficient matrix before and after using this change, we found drastic changes in the condition number. For example, in problem *3.8* after using this method the condition number of infinity comes down to $4.6058\times10^{9}$. We also checked other problems, and in all cases the condition numbers reduced, at least, by a factor of $10^{7}$. Let's consider some other problems of this kind. It is

necessary to mention that the Runge-Kutta, Adams and shooting methods do not work for the following problems.

*Problem 3.9:* Consider

$$y''(x)-\frac{1}{x}y'(x)+\frac{1}{x}y(x)=|x|\ ,x\in[-1,1],$$
$$y(-1)=-1,\ y(1)=1.$$

The exact solution is $y(x)=x|x|$. We used again this idea, and tested this problem with different values of $N$, and the results are given in Table 9.

Table 9

| $N$ | $\|y(x)-\tilde{y}_t(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ps}(x)\|_\infty$ | $\|y(x)-\tilde{y}_{ms2}(x)\|_\infty$ |
|---|---|---|---|
| 5 | $8.31\times10^{-2}$ | $7.64\times10^{-2}$ | $1.92\times10^{-2}$ |
| 8 | $8.75\times10^{-1}$ | $8.86\times10^{-1}$ | $3.23\times10^{-2}$ |
| 9 | $1.54\times10^{-2}$ | $3.97\times10^{-2}$ | $1.00\times10^{-3}$ |
| 17 | $1.12\times10^{-2}$ | $2.05\times10^{-2}$ | $5.47\times10^{-4}$ |

As we can see the results with this modified method are better than the other two methods. In this problem again we get a lower condition number.

Introduction of this method will be ended by representing another example with non- analytical solution and non-analytical coefficient functions.

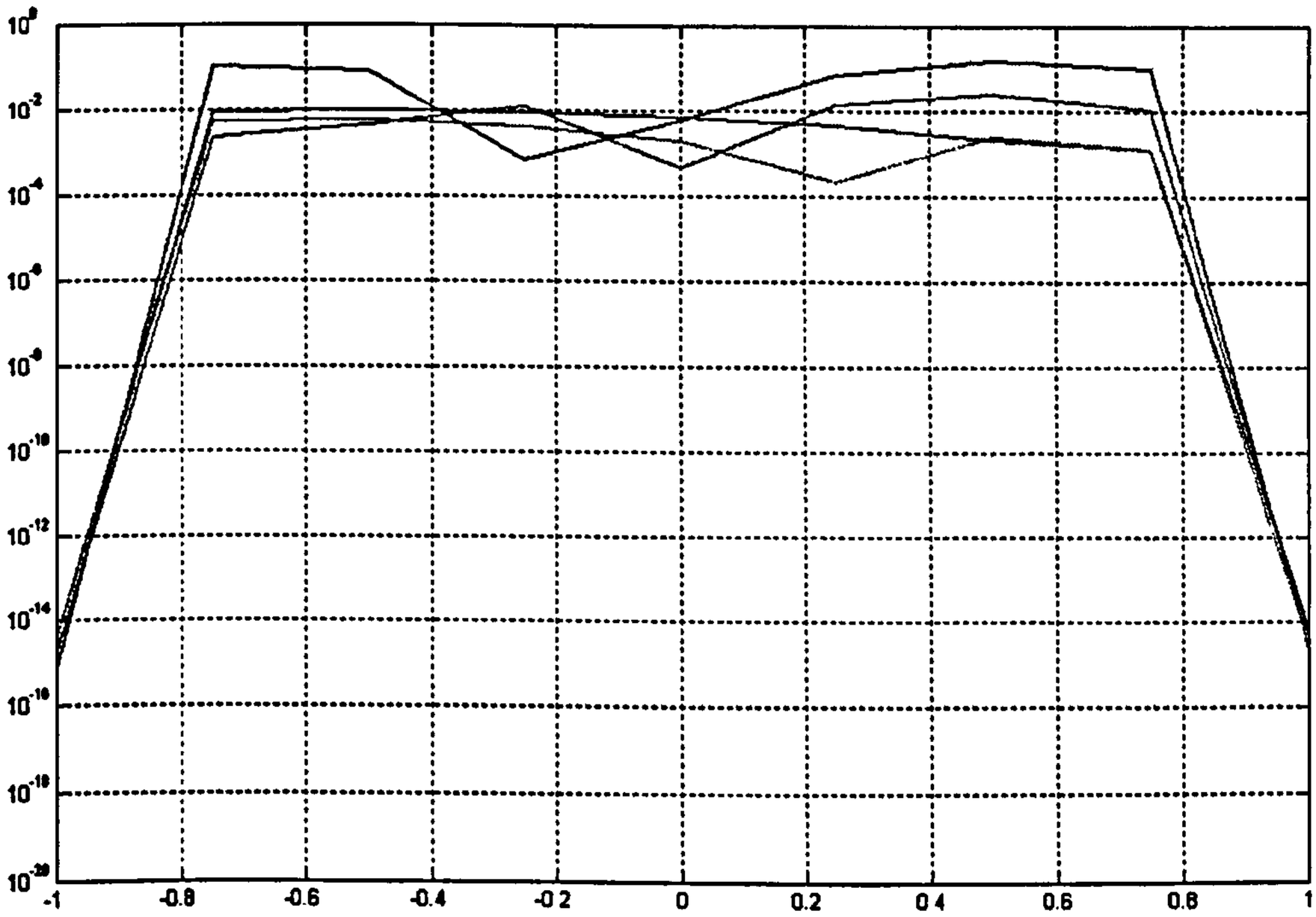
*Problem 3.10:* Consider

$$|x+\tfrac{1}{2}|y''(x)+|x-\tfrac{1}{2}|y'(x)+y(x)=|x^2-\tfrac{1}{4}|[(x^2-\tfrac{1}{4})+(6x^3-\tfrac{2}{3}x)|x-\tfrac{1}{2}|+(30x^2-\tfrac{2}{3})|x+\tfrac{1}{2}|],\ x\in[-1,1],$$
$$y(-1)=y(1)=(\tfrac{3}{4})^3,$$

with exact solution $y(x)=|(x^2-\tfrac{1}{4})^3|$.

We solved it with pseudo-spectral method and our modified method for $N$=5,8 and the maximum errors for pseudo-spectral method are $1.5\times10^{-1}$,$2.4\times10^{-2}$, respectively. But for modified method are $1.0\times10^{-3}$, $5.9\times10^{-4}$. The results were illustrated in Fig 3.

Fig 3




We observe that the results may be not very good, but still the method we described above is an improvement.

The next chapter is devoted to pseudo-spectral method to solve DAEs.

# Chapter 4

## Numerical solutions of Differential-Algebraic Equations

### 4.1 Introduction

Systems of Differential-Algebraic Equations (DAEs) are systems of differential equations (sometimes also referred to as descriptor, singular or semi-state systems), where the unknown functions satisfy both differential and additional algebraic equations. In other words, they consist of a set of differential equations with additional algebraic constraints. These systems which are given in the most general form $F(x,x',t) = 0$, arise naturally in many areas of science and engineering, such as robotics (via Lagrange's equation with interdependent coordinates), biomechanics, control theory, electrical engineering (via Kirchoff's law), and fluid dynamics (via Navier-Stokes equations for incompressible flow).

Many physical phenomenon are most naturally modelled as a combination of ordinary differential and algebraic equations. Models of chemical processes, for example, typically describe the dynamic balance of mass and energy while additional algebraic equations account for thermodynamic equilibrium relations or steady-state assumptions [23]. There has been an increased interest in several areas in exploiting the advantages of working directly with these implicit models. Multi-body systems is one area, in which methods for solving DAE are of special interest (A multi-body system is a mechanical system, that consists of one or more rigid or elastic bodies).

In the 1960's engineers working on electrical circuits or multi-body systems realized that solving a differential equation with constraints is more involved than solving one without constraints; that is, the constrained case can not in general be reduced to the unconstrained case by some standard tricks. The first paper which introduced a way to attack these problems was written by C. W. Gear in 1971[24]. There also the name "differential-algebraic equation" was introduced.

Comparing DAE with Ordinary Differential Equations (ODE), there are both numerical and analytical difficulties which do not occur with ODE. The first practical numerical methods for certain classes of DAE are the Backward Differentiation Formulae (BDF) and implicit Runge-Kutta methods [25]. But these methods can not be applied to approximate the solution to all DAE [26, 27]. Sometimes a DAE can be converted into a system of ODEs. However, numerical stability of the system is often undermined in the process so that, even if all DAE can be converted into ODE, it is usually not always desirable to do so. In the early 70s C. W. Gear started to write about using BDF-methods in connection with such problems. Now BDF-methods are widely used in computer codes that solve ODEs and DAE, such as the DASSL(Differential-Algebraic System Solver Library) which is designed for solving initial value problems of the implicit form $F(t,y,y')=0$ with index zero or one, and the LSODI(Liver more Solver for ODEs, Implicit) code which is similar to DASSL in that it is based on BDF methods [28].

A considerable amount of research has been done on numerical methods for DAE. Ascher, Petzold, Campbell and Gear have carried out extensive work on numerical solution of this class of equations. Differentiation plays an important role in both the analysis and numerical solution of DAE. The index, which will be defined shortly, is one measure of how singular a DAE system is . Increasing index implies more complex behaviour. Usually DAEs are difficult to solve if they have a high index, that is, an index greater than one. Several techniques exist for index reduction and consistent initialization of higher index DAE [29]. In this chapter, the general definition of a system of DAE is presented and then narrow our focus to those DAEs in Hessenberg form of size 2 and 3. The concept of index as characterization of DAEs is briefly mentioned.

## 4.2 Differential-algebraic equations

In this section we give a definition of DAE, and introduce the concept of index, and of DAE in Hessenberg form.

***Definition 1*:** The general or *fully implicit* DAE is a vector equation of the form

$$F(t,x(t),x'(t))=0 \tag{4.1}$$

where $t \in \mathrm{R}, x \in \mathrm{R}^n, F : \mathrm{R} \times \mathrm{R}^n \times \mathrm{R}^n \to \mathrm{R}^n, \frac{\partial F}{\partial x} = F_x$ ($F_x$ is the Jacobian of $F$ with respect to $x$, which may be singular or not) and $\frac{\partial F}{\partial x'} = F_{x'}$ is singular for all $x(t), x'(t), t$.

The other forms of DAEs may be

$$x' = f(t, x), \tag{4.2}$$

called *explicit form,* and

$$x' = f(t, x), \tag{4.3a}$$
$$0 = g(x), \tag{4.3b}$$

called *semi-explicit form* or an ODE with constraints.

The idea of using ODE methods for direct solution of DAE systems, was first stated by Gear [43]. This can be done with a simple algorithm based on Euler's method. In this method $x'(t_{n+1})$ will be approximated by a BDF of $x(t)$ and (4.1) becomes

$$F(t_{n+1}, x_{n+1}, (x_{n+1} - x_n)/(t_{n+1} - t_n)) = 0.$$

One of the aspects of ODE methods for direct solution of DAE is that these methods preserve sparsity of the resulting system. For example, a class of DAE systems which are simple to solve, are those systems which, indeed, are implicit ODE systems that have been changed. That is, if in (4.1) $\frac{\partial F}{\partial x'}$ is non singular, then the system can be written in an explicit form as (4.2). But if $\frac{\partial F}{\partial x'}$ is a sparse matrix, its inverse may not be sparse so it is preferred to solve the initial system directly. Difficulties often occur when $\frac{\partial F}{\partial x'}$ is singular. A simple class of such systems is linear DAE with constant coefficients, which has the form $Ax' + Bx = f$ .

A function $x(t)$ is a solution of (4.1) in the interval $I$, if $x$ is continuously differentiable on $I$ and satisfies (4.1) for $t \in I$. In this chapter, we are concerned with the case where the solutions exist and are uniquely defined on the interval of interest: how ever, not all initial values for $x$ admit a smooth solution. This concept of solvability is made more precise in the next definition. But first of all, we say a problem is called *solvable* if it has at least one solution. This definition seems

natural but it should be noted that the term solvability is used only for systems which have a unique solution when consistent initial conditions are provided. If the solution of initial value problem is not unique, then further conditions have to be specified to single out specific desired solutions.

***Definition2:*** Consider the general non-linear DAE given in (4.1). Let $I$ be an open subinterval of $\Re$, $\Omega$ a connected open subset of $\Re^{2m+1}$, and F a differentiable function from $\Omega$ to $\Re^m$. Then the DAE (4.1) is solvable on $I$ in $\Omega$ if there is an $r$-dimensional family of solutions $\phi(t,c)$ defined on a connected open set $I \times \tilde{\Omega}$, $\tilde{\Omega} \subset \Re^r$, such that:

i) $\varphi(t)$ is defined on all of $I$ for each $c \in \tilde{\Omega}$.

ii) $(t, \phi(t,c), \phi'(t,c)) \in \Omega$ for $(t,c) \in I \times \tilde{\Omega}$.

iii) if $\varphi(t)$ is any other solution with $(t, \varphi(t), \varphi'(t)) \in \Omega$, then $\varphi(t) = \phi(t,c)$ for some $c \in \tilde{\Omega}$.

iv) the graph of $\phi$ as a function of $(t,c)$ is an $(r+1)$-dimensional manifold.

As we know a property known as the index plays a key role in the classification of the behaviour of DAE.

***Definition 3*:** The *index* of DAE is the minimum number of times that all or part of (4.1) must be differentiated with respect to $t$ in order to determine $x'$ as a continuous function of $x$ and $t$.The class of index zero DAE is the set of all ODE. DAE with index zero or one are generally much simpler than DAE with index two or higher. A naive method for solving DAE can be constructed based on reducing the index of the problem through repeated differentiation of the constraint equations. Once an index zero DAE is obtained then the problem has been converted from a DAE to a system of ODE, and can be solved numerically with an established solver such as MATLAB's ode45. This process is called *index reduction*, and may be applied to a system for lowering the index from an initially high value down to e.g. index one. Let us illustrate it by an example. We look at a semi-explicit DAEs in the form

that is a *semi-explicit* DAE, or an ODE with *constraints*. The index is 1, if $\partial g / \partial z$ is non-singular.

In the general case, each component of $y$ may contain differential and algebraic components, which makes the numerical solution of such high-index problems harder and more risky. The semi-explicit form is decoupled in this sense. On the other hand, any DAE of the form (4.1) can be written in the semi-explicit form (4.5) but with the index increased by 1, upon defining $y' = z$, which gives

$$y' = z,$$

$$0 = F(t, y, z).$$

Needless to say, this rewriting alone does not make the problem easier to solve. The converse transformation is also possible: given a semi-explicit index-2 DAE system (4.5), let $w' = z$. It is easily shown that the system

$$x' = f(t, x, w'),$$

$$0 = g(t, x, w')$$

is an index-1 DAE and yields exactly the same solution for $x$ as (4.5).

There is a relationship between the index of semi-explicit systems and general systems that is worth stating as a 'rule of thumb' : The semi-explicit case is much like the general case of one lower index.

The general DAE system (4.1) can include problems which are not well defined in a mathematical sense, as well as problems which will result in failure for any direct discretization method. Fortunately, most of the higher-index problems encountered in practice can be expressed as a combination of more restrictive structures of ODEs coupled with constraints. In such systems the algebraic and differential variables are explicitly identified for higher-index DAEs as well, and the algebraic variables may all be eliminated (in principle) using the same number of differentiations. These are called *Hessenberg forms* of the DAE and are given below.

**Hessenberg Index-1**

$$x' = f(t,x,y), \tag{4.6a}$$

$$0 = g(t,x,y). \tag{4.6b}$$

Here the Jacobian matrix function $g_y$ is assumed to be non-singular for all $t$.This is also often referred to as a *semi-explicit index*-1 system. Semi-explicit index-1 DAE are very closely related to implicit ODE.

**Hessenberg Index-2**

$$x' = f(t,x,y), \tag{4.7a}$$

$$0 = g(t,x). \tag{4.7b}$$

Here the product of Jacobians $g_x f_y$ is assumed to be non-singular for all $t$. Note the absence of the algebraic variable $y$ from the constraints (4.7b). This is a *pure* index-2 DAE, and all algebraic variables play the role of index-2 variables.

**Hessenberg Index-3**

$$x' = f(t,x,y,z), \tag{4.8a}$$

$$y' = g(t,x,y), \tag{4.8b}$$

$$0 = h(t,y). \tag{4.8c}.$$

Here the product of three matrix functions $h_y g_x f_z$ is non-singular.

Semi-explicit index-1 systems arise in a wide variety of applications including most circuit analysis and power systems problems. Some examples of Hessenberg index-2 systems arise in the modelling of incompressible fluids, and some index-2 formulations of mechanical systems [32]. Hessenberg index-3 DAEs arise in the simulation of mechanical systems and in optimal control. For a variety of reasons, systems of index-3 and higher have proven to be very difficult to solve

numerically [29], and much recent work has focused instead on reformulating these systems as index-2 or lower.

For numerical solution of linear or non-linear DAE system with index less than or equal one, usually, numerical methods such as (4.4) can be used. But, when the index is more than one algorithms based on this method can encounter difficulties. With a little bit of connivance, we can say that the techniques based on higher order methods, such as the extrapolation method, can be used in such cases. But this can not be extended to nonlinear DAE or linear with variable coefficients. In fact, numerical methods that work for system (4.3), fail to produce a numerical solution of DAE with coefficient matrices which are time varying and systems with index higher than one [2],[33].

In this chapter we present a method that reduces a DAE with high index to a DAE with lower index. But first we study, briefly, the problems with index not more than one and linear systems with constant coefficients, with any arbitrary index.

## 4.3 Linear DAEs with constant coefficients

One class of DAE systems is linear DAE with constant coefficients which

has a form of

$$Ax' + Bx = f \tag{4.9}$$

where $A$, $B$ are $n \times n$ matrices and $f$ is sufficiently smooth.

***Definition 4*:** Assume $\lambda$ is a complex parameter and call $\lambda A + B$ a *matrix pencil.* The matrix pair $(A, B)$ is called regular (or regular pencil) of index $\lambda$ if $\det(\lambda A + B)$ is not identically zero.

***Theorem 1.*** (4.9) is solvable if $\lambda A + B$ is regular.

If $(A + \lambda B)$ is singular, then system (4.9) either has no solution or has infinitely many solutions.

Let $x = Q\,y$, and premultiply (4.9) by $P$, where $P$ and $Q$ are $n \times n$ non-singular matrices, we get

$$PAQ\,y' + PBQ\,y = Pf. \tag{4.10}$$

We note that rescaling of the equation and the unknown by non-singular matrices, does not change the solution behaviour.

Recall that a matrix $N$ is said to have *nilpotency* $k$ if $N^k = 0$ and $N^{k-1} \neq 0$. A corresponding normal form which is the so-called *Kronecker Canonical Form* (KCF), exhibits all properties of a linear DAE with constant coefficients. Converting linear systems (4.9) to a KCF, then a significant property of it will be specified [45]. In fact, the main idea of it is that, there exist non-singular matrices $P$ and $Q$ such that $(A, B)$ reduces to a KFC. The key structure theorem for (4.9) which follows from the KCF, is:

***Theorem 2.*** Let $\lambda A + B$ be regular, then there exist non singular matrices $P$ and $Q$ such that

$$PAQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, PBQ = \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \tag{4.11}$$

where $N$ is a matrix of nilpotency $k$. If $N = 0$ then define $k$=1. In the special case that $A$ is non-singular, we take $PAQ = I$, and $PBQ = C$ and define $k$=0. If $det(\lambda A + B)$ is identically constant, then (4.12) simplifies to $PAQ = N$, $PBQ = I$. Now, assume $Y = (y_1, y_2)^t$, $Pf = (f_1, f_2)^t$ and apply the coordinate changes to the DAEs (4.9) to obtain (4.10) which by (4.11) becomes

$$\begin{aligned} y_1' + Cy_1 &= f_1 \\ Ny_2' + y_2 &= f_2 \end{aligned} \tag{4.12}$$

Solution to the second equation of (4.12) is given by

$$y_2 = \sum_{i=0}^{k-1} (-1)^i N^i f_2^{(i)}.$$

If the k-step constant step-size BDF method ($k$<7) is applied to constant coefficient linear DAE system of index $m$ it converges after $(m - 1)\ k + 1$ steps with order of accuracy $O(h^k)$ [29].

Unfortunately if step size is not constant, this method breaks down in some cases [30]. Convergence and stability properties of numerical methods are dependent on the structure of the system.

## 4.4 Linear DAEs with variable coefficients

There are of the form

$$A(t)x'(t) + B(t)x(t) = f(t) \tag{4.13}$$

defined on the interval $I$. This kind of DAE which also is called a linear time-varying DAE, exhibits most of the behaviour found in the non-linear case that is not already presented in the constant coefficient case.

Using BDF method (of order one) starting at time $t_0$ with constant step size $h$, with $t_n = t_0 + nh$ gives

$$(A_n + hB_n)x_n = A_n x_{n-1} + hf_n. \tag{4.14}$$

In order for (4.14) to uniquely determine $x_n$, given $x_{n-1}$, we need $A(t_n) + hB(t_n)$ to be non singular for small $h$. Thus we need regularity of $\lambda A(t) + B(t)$ for each $t \in I$.

The system (4.13) is semi-explicit if it is in the form

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

This semi-explicit DAE has index one if and only if $B_{22}$ is non-singular for all $t$.

Many of higher index semi-explicit DAE's arising in applications have a natural structure which we call Hessenberg form [47].

***Definition* 7:** The DAE (4.13) DAE in Hessenberg form is of size r if it can be written as

$$\begin{bmatrix} I & 0 & \cdots & 0 \\ 0 & I & \cdots & \cdot \\ \cdot\cdot & & I & \cdot\cdot \\ \cdots & & & I \\ 0 & \cdots & \cdots & 0 \end{bmatrix}\begin{bmatrix} x_1' \\ x_2' \\ \cdot \\ \cdot \\ x_r' \end{bmatrix} + \begin{bmatrix} B_{11} & \cdots & B_{1,r-1} & B_{1r} \\ B_{21} & \cdots & B_{2,r-1} & 0 \\ 0 & \cdots & & \cdot \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & B_{r,r-1} & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_r \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ f_r \end{bmatrix} \tag{4.15}$$

where $B_{r,r-1}B_{r-1,r-2}\cdots B_{12}$ is non-singular.

A DAE in Hessenberg form of size $r$ is solvable and has index and local index $r$ [29].

For linear constant coefficient systems, there are several equivalent definitions of the index. These definitions are not all equivalent for linear time varying DAE. If matrix pair $(A(t), B(t))$ is non-singular, we define local index for (4.13).

***Definition* 8:** Let (4.13) be regular, then the local index at t denoted by $v_e(t)$, is the index of the pencil $\lambda A(t) + B(t)$.

We also define global index, if it exists, as follows:

with $x = Q(t)\,y$ and premultiplying (4.13) by $P(t)$, where $P(t)$, $Q(t)$ are $n\times n$

non-singular matrices, then (4.13) becomes

$$P(t)A(t)Q(t)y' + \left[P(t)B(t)Q(t) + P(t)A(t)Q'(t)\right]y = P(t)f(t). \tag{4.16}$$

Now if $P(t)$ and $Q(t)$ exist, such that

$$P(t)A(t)Q(t) = \begin{bmatrix} I & 0 \\ 0 & E \end{bmatrix}$$

$$P(t)B(t)Q(t) + P(t)A(t)Q'(t) = \begin{bmatrix} P(t) & 0 \\ 0 & I \end{bmatrix}$$

where $E$ is strictly lower (or upper) triangular, with index $m$, then we say (4.13) has global index $m$.

If the global index of (4.13) is one, then an extrapolation method for solving this system will converge [34].

Here we introduce an algorithm for reducing the index of (4.13) as follows:

(i) if A is non-singular, then stop.

(ii) premultiply by a non-singular $P(t)$ in (4.13) to obtain a maximum number of rows of zeros of $A$ and put these rows at the bottom of matrix $A$.

$$\begin{pmatrix} A_{11} \\ 0 \end{pmatrix} x' + \begin{pmatrix} B_{11} \\ B_{12} \end{pmatrix} x = P(t) f(t).$$

(iii) Differentiate the second half of the system and write

$$\begin{pmatrix} A_{11} \\ B_{12} \end{pmatrix} x' + \begin{pmatrix} B_{11} \\ B'_{12} \end{pmatrix} x = \hat{f}(t).$$

(iv) Go back to (i) and repeat.

At each iteration, the index of the system reduces by one. This will continue until the index of the system becomes zero, that is, we get a non-singular coefficient matrix. In fact, the structure of this algorithm is based on the following theorem.

***Theorem 3.*** Consider the DAE system of (4.13)

(a) if $A$ is non singular, then index of $(A, B)$ is zero.

(b) if $A$ is singular, then we choose a non singular matrix $P$, such that

$PA = \begin{bmatrix} A_1 \\ 0 \end{bmatrix}$ and $A_1$ is a full-rank matrix and $PB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$.

Now if $\begin{bmatrix} A_1 \\ B_2 \end{bmatrix}$ is non singular, then (4.13) has index one.

Proof : See [35].

Before we end this section we introduce a simple formulation for index reduction for the following problem. Consider

$$x' = Ax + By + q, \tag{4.17a}$$
$$0 = Cx + r, \tag{4.17b}$$

where, $A=(a_{ij})_{n\times n}$, $B=(b_i)_{n\times 1}$, $C=(c_i)_{1\times n}$, $q=(q_i)_{n\times 1}$, $n\geq 2$, are smooth functions of $t$ and $CB$ is non-singular for all $t\in[0,t_k]$, for some fixed $t_k$.

This problem is called the Hessenberg index-2 system.

By (4.17a), we have $y=(CB)^{-1}C[x'-Ax-q]$, and substituting into (4.17a) implies,

$$x'=Ax+B(CB)^{-1}C[x'-Ax-q]+q.$$

So problem (4.17) transforms to the system:

$$(I-B(CB)^{-1}C)[x'-Ax-q]=0 \tag{4.18a}$$

$$0=Cx+r, \tag{4.18b}$$

Here, the over determined system (4.18) will transform to a full rank DAE system with $n$ equations and $n$ unknowns which has index one.

***Theorem 4.*** The index-2 DAE system (4.18), with $n=2$, is equivalent to an index-1 DAE system given by,

$$E_0x'+E_1x=\hat{q}, \tag{4.19}$$

such that [42],

$$E_0=\begin{bmatrix}b_2 & -b_1\\ 0 & 0\end{bmatrix},\ E_1=\begin{bmatrix}b_1a_{21}-b_2a_{11} & b_1a_{22}-b_2a_{12}\\ c_1 & c_2\end{bmatrix},$$
$$\hat{q}=\begin{bmatrix}b_2q_1-b_1q_2\\ -r\end{bmatrix}.$$

This index reduction formulae can be used for a linear model problem

$$\begin{aligned}x^{(m)}&=\sum_{j=1}^{m}A_jx^{(j-1)}+By+q,\\ 0&=Cx+r.\end{aligned} \tag{4.20}$$

This DAE has index $m+1$ and will be transformed into an implicit DAE form by putting

$$y=(CB)^{-1}C[x^{(m)}-\sum_{j=1}^{m}A_jx^{(j-1)}-q] \tag{4.21}$$

and substituting it in (4.20). We obtain a DAE which has index $m$, as follows,

$$\sum_{j=0}^{m}E_jx^{(j)}=\hat{q}, \tag{4.22}$$

where $E_j(t) \in \Re^{n\times n}, j = 0,1,...,m,$ and except $E_0(t)$, other matrices are singular.

We emphasize that if (4.13) has high index (usually more than one), reducing the index causes numerical instability [46]. To remove this difficulty, regularization methods will be used to solve the reduced problem [36], [regularization seeks to convert DAE into ODE without using repeated differentiation of the constraint]. As we know that a numerical method may be applied to either the original DAE or to the enlarged system, but because of the change in the index, resulting convergence and stability properties of schemes may be quite different. In chapter three of [29], convergence, order and stability properties of linear multi-step methods applied to DAE have been studied.

Here we give a brief description about convergence which has been considered nicely in [29].

## 4.5 The conclusions about stability and convergence

When applying multi-step methods (and therefore in particular BDF methods) to semi-explicit index one DAE, they are stable and convergent to the same order of accuracy for the DAE as for the underlying ODE. These problems can therefore be solved with any linear multi-step method, which is appropriate for the underlying ODE, assuming that the constraint(s) are satisfied in each step. For the fully implicit index one system it can be shown that a constant step size BDF-method of order $k<7$ with the initial values given correct to the order of $O(h^k)$ converges with an accuracy of order $O(h^k)$ if each Newton iteration is solved to an accuracy of order $O(h^{k+1})$. If a variable step-size BDF-method is used (with step-size restriction as for the standard ODE case), then it will also be convergent for the fully implicit index one system [29].

If a semi-explicit index two system is solved with a constant step-size BDF-method of order $k \le 7$ with the initial values given correct to the order of $O(h^k)$, it converges with an accuracy of order $O(h^k)$ after $(k + 1)$ steps if each Newton iteration is solved to an accuracy of order $O(h^{k+1})$. If a variable step size BDF-method is used the same as in the index one case will happen. It will be convergent if the method is implemented in a way that it is stable for standard ODE [29].

For an index three system of Hessenberg form if a constant step-size BDF-method with $k<7$ is used with starting values at an accuracy of order $O(h^{k+1})$, and the algebraic equations are solved to an accuracy of order $O(h^{k+2})$ if $k \geq 2$ or to order $O(h^{k+3})$ if $k=1$ after $k+1$ steps it converges to an order of $O(h^k)$. If a variable step size BDF-method is used the system might fail to converge. Notice that when using variable step size there is no guarantee of convergence for DAE of index higher than two. The nature and stability of a class of nonlinear DAE systems have been considered in [37]. It was shown via an appropriate coordinate transformation that the solution of this representation is unstable about its solution manifold when the system's differential index is higher than one. We can also find papers on stabilization of DAE and invariant manifolds [38] or stabilization of constrained mechanical systems with DAE and invariant manifolds [39].

As mentioned earlier there exist codes using BDF- methods for solving DAE. Most of these codes are designed for solving index zero or index one DAE. In [29] some of these codes are described and tested. They have focused on the DASSL code. DASSL uses a BDF-method with variable step size with order up to five. It is reported that the DASSL code has successfully solved a wide variety of scientific problems, and that there are probabilities of testing out where it went wrong. They report that the finite difference calculation of the Jacobian matrix is the weakest part of the code. It also has problems with handling inconsistent initial conditions and discontinuities. A detailed description of DASSL and a little bit of the code LSODI can be found in chapter five of [29]. In the end of this paper, once again I mention several nice research works which have been carried out by: U.M. Ascher, L.R. Petzold, C. Gear, S.L. Campbell and P. Kunkel.

## 4.6 DAE with variable coefficients and the Pseudo-spectral method

According to (4.13), consider the following DAE

$$\begin{aligned} a_{11}(t)y_1'(t)+a_{12}(t)y_2'(t)+a_{13}(t)y_1+a_{14}(t)y_2(t) &= f_1(t) \\ a_{23}(t)y_1+a_{24}(t)y_2(t) &= f_2(t)' \end{aligned} \qquad (4.23a)$$

so that;

$$A(t)=\begin{pmatrix} a_{11}(t) & a_{12}(t) \\ 0 & 0 \end{pmatrix},\ B(t)=\begin{pmatrix} a_{13}(t) & a_{14}(t) \\ a_{23}(t) & a_{24}(t) \end{pmatrix},\ \text{and}\ f(t)=\begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix},$$

with initial condition,

$$y_1(-1)=\alpha, \tag{4.23b}$$

where $a_{ij}, f_1$ and $f_2$ are sufficiently smooth functions of $t$ and $\alpha$ is constant. Now, for an arbitrary natural number $N$, we suppose that the given DAE has an approximate solution

$$\begin{aligned} y_1(t) &\cong \sum_{i=0}^{N} a_i T_i(t) \\ y_2(t) &\cong \sum_{i=0}^{N} a_{N+1+i} T_i(t) \end{aligned} \tag{4.24}$$

where $\{T_i\}_{i=0}^{N}$ is the sequence of Chebyshev polynomials of the first kind and

$\underline{a}=(a_0, a_1, \ldots, a_{2N+1})^t \in \Re^{2N+2}$.

The main target is to find $\underline{a}$. As in chapter 2, put

$$y_N(x)=\sum_{k=0}^{N} a_k T_k(x),$$

then, again, corresponding to functions $y_N$ and $y_N'$, we can define $A^{(0)}$ and $A^{(1)}$ as before and let,

$$\begin{aligned} AA &= a_{11}(t)A^{(1)} + a_{13}(t)A^{(0)} \\ BB &= a_{12}(t)A^{(1)} + a_{14}(t)A^{(0)} \\ CC &= a_{23}(t)A^{(0)} \\ DD &= a_{24}(t)A^{(0)} \end{aligned} \tag{4.25}$$

$$\phi_i(t)=\begin{cases} \sum_{k=0}^{i} (AA)_{ki} T_k(t), & 0 \le i \le N, \\ \sum_{k=0}^{i} (BB)_{k(i-1-N)} T_k(t), & N+1 \le i \le 2N+1, \end{cases} \tag{4.26}$$

$$\psi_i(t)=\begin{cases} \sum_{k=0}^{i} (CC)_{ki} T_k(t), & 0 \le i \le N, \\ \sum_{k=0}^{i} (DD)_{k(i-1-N)} T_k(t), & N+1 \le i \le 2N+1, \end{cases} \tag{4.27}$$

then (4.23a) converts to

$$\sum_{i=0}^{2N+1} a_i \phi_i(t) \cong f_1(t)\,,$$
$$\sum_{i=0}^{2N+1} a_i \psi_i(t) \cong f_2(t)\,, \tag{4.28}$$

with initial conditions,

$$\sum_{i=0}^{N} a_i T_i(-1) = \sum_{i=0}^{N} a_i (-1)^i = \alpha\,, \tag{4.29}$$

Relation (4.29) forms a system with one equation and $2N+2$ unknowns. A second equation $a_{23}(-1)y_1(-1)+a_{24})-1)y_2(-1) = f_2(-1)$ is obtained from (4.23a). To construct the remaining $2N$ equations we substitute points

$$t_j = \cos(\tfrac{2\pi j}{2N-1}), j = 0,\ldots,N-1\,, \tag{4.30}$$

into (4.25) , (4.28) and put,

$$\sum_{i=0}^{2N+1} a_i \phi_i(t_j) = f_1(t_j)\,,$$
$$\sum_{i=0}^{2N+1} a_i \psi_i(t_j) = f_2(t_j)\,, \qquad j = 0,\ldots,N-1$$

to obtain $2N$ equations.

As we mentioned in chapter three, about how the Tau method can be used to solve a system of linear ODE, it should be noted that the pseudo-spectral method can also be used to solve a system of ODE such as

$$\begin{aligned} &a_{11}(t)y_1'(t)+a_{12}(t)y_2'(t)+a_{13}(t)y_1+a_{14}(t)y_2(t) = f_1(t), \\ &a_{21}(t)y_1'(t)+a_{22}(t)y_2'(t)+a_{23}(t)y_1+a_{24}(t)y_2(t) = f_2(t), \end{aligned} \qquad t \in [-1,1]$$

provided that

$$AA = a_{11}(t)A^{(1)} + a_{13}(t)A^{(0)}$$
$$BB = a_{12}(t)A^{(1)} + a_{14}(t)A^{(0)}$$
$$CC = a_{21}(t)A^{(1)} + a_{23}(t)A^{(0)}$$
$$DD = a_{22}(t)A^{(1)} + a_{24}(t)A^{(0)}$$

Even we can extend this method to systems of differential equations of any order [48]. For example;

$$\begin{aligned} &a_{11}(t)y_1''(t) + a_{12}(t)y_2''(t) + a_{13}y_1'(t) + a_{14}y_2'(t) + a_{15}y_1(t) + a_{16}y_2(t) = f_1(t) \\ &a_{21}(t)y_1''(t) + a_{22}(t)y_2''(t) + a_{23}y_1'(t) + a_{24}y_2'(t) + a_{25}y_1(t) + a_{26}y_2(t) = f_2(t) \end{aligned} \qquad t \in [-1,1],$$

which is of order two, provided that

$$AA = a_{11}(t)A^{(2)} + a_{13}(t)A^{(1)} + a_{15}(t)A^{(0)},$$
$$BB = a_{12}(t)A^{(2)} + a_{14}(t)A^{(1)} + a_{16}(t)A^{(0)},$$
$$CC = a_{21}(t)A^{(2)} + a_{23}(t)A^{(1)} + a_{25}(t)A^{(0)},$$
$$DD = a_{22}(t)A^{(2)} + a_{24}(t)A^{(1)} + a_{26}(t)A^{(0)},$$

with $y_1(a) = \alpha_1$, $y_2(a) = \beta_1, y_1(b) = \alpha_2, y_2(b) = \beta_2$ as initial conditions.

In the next section we consider some numerical examples.

## 4.7 Some numerical examples

Before consideration of the numerical solution of DAE, first we consider a numerical example for a system of differential equations and then the rest of this section is devoted to DAEs. In all examples $e_1$ and $e_2$ denote the maximum error of $y_1(t)$ and $y_2(t)$,respectively.

*Example 4.1:* Consider

$$ty_1'(t) + y_2'(t) - 2y_1(t) + e^t y_2(t) = 1 - e^{-t},$$
$$e^{-t}y_1'(t) + 2ty_2'(t) = 0,\ t \in [-1,1],$$

with initial conditions $y_1(-1) = 1, y_2(-1) = e.$

The exact solutions are $y_1(t) = t^2$, and $y_2(t) = e^{-t}$. We solved it by the pseudo-spectral method for $N = 4,7,10$, and the results are given in Table 1.

Table 1

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $3.479\times10^{-3}$ | $2.750\times10^{-3}$ |
| 7 | $4.002\times10^{-6}$ | $7.656\times10^{-7}$ |
| 10 | $8.182\times10^{-10}$ | $2.059\times10^{-10}$ |

As can be seen from the results, this method produces very good results. In this manner we can solve a system of linear ODE of order two, and extend it to any order. But in this chapter our main target is to consider examples of linear DAE. Therefore, the rest of this section is devoted to such systems.

*Example 4.2:* Consider

$$y_1'(t) - y_2'(t) - ty_1 = 0,$$
$$t(\sin t)y_1(t) + (\cos t)y_2(t) = t,\ t \in [-1,1],$$

with initial condition $y_1(-1) = \sin(-1)$.

The exact solutions are, $y_1(t) = \sin t$, and $y_2(t) = t\cos t$.

We solved it by the pseudo-spectral method and the error produced for different values of $N$ for $y_1(t)$ and $y_2(t)$ are given in Table 2.

Table 2

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $1.341\times10^{-2}$ | $6.784\times10^{-3}$ |
| 6 | $7.782\times10^{-5}$ | $4.840\times10^{-5}$ |
| 10 | $9.182\times10^{-10}$ | $4.063\times10^{-10}$ |

Let's consider another example. This example was chosen from [40].

*Example 4.3*: Consider the following problem with initial values;

$$\begin{pmatrix}1 & -t\\ 0 & 0\end{pmatrix}y'(t)+\begin{pmatrix}1 & -1-t\\ -\mu & 1+\mu t\end{pmatrix}y(t)=\begin{pmatrix}0\\ \sin t\end{pmatrix},\quad t\in[0,1],$$

with initial condition $y_1(0)=1$.

The exact solutions are, $y_1(t)=t\sin t+(1+\mu t)e^{-t}$, and $y_2(t)=\mu e^{-t}+\sin t$.

We solved it by the pseudo-spectral method. Although, this problem has index 1, Ascher showed in 1989, that for $\mu>>0$ symmetric methods of numerical solution encounter difficulty [41], and solve it with $\mu=10$. In 1994 Amodio solved it by techniques of boundary values [44]. Although he did not mention for what value of $\mu$ difficulty happens, we solved it by the pseudo-spectral method for $\mu=200$ and the results are given in Table 3. But when we increase the value of $\mu$, the error increases. In this table $e_{ps}$ and $e_A$ mean maximum error for $y_1(t)$ and $y_2(t)$ using the pseudo-spectral and Adams methods, respectively.

Table 3

| $N$ | $e_{ps}$ | $e_A$ | $h$ |
|---|---|---|---|
| 6 | $1.73\times10^{-4}$ | $1.22\times10^{-5}$ | $2\times10^{-2}$ |
| 10 | $1.29\times10^{-10}$ | $1.92\times10^{-7}$ | $5\times10^{-3}$ |
| 14 | $6.21\times10^{-14}$ | $2.41\times10^{-8}$ | $2.5\times10^{-3}$ |

As table 3 shows, results obtained by the pseudo-spectral method are much better than those for the Adams method.

*Example4.4*: Let's consider another problem with initial condition,

$$\begin{pmatrix} 0 & 0 \\ 1 & \mu t \end{pmatrix} y'(t) + \begin{pmatrix} 1 & \mu t \\ 0 & 1+\mu \end{pmatrix} y(t) = \begin{pmatrix} e^t \\ t^2 \end{pmatrix}, \quad t \in [-\tfrac{1}{2}, \tfrac{1}{2}],$$

The exact solutions are, $y_1(t) = e^t + \mu t(e^t - t^2)$, and $y_2(t) = t^2 - e^t$, with initial condition $y_1(-\frac{1}{2}) = \frac{1}{\sqrt{e}} - \frac{\mu}{2}(\frac{1}{\sqrt{e}} - \frac{1}{4})$.

This problem has global index 2 and was considered in several papers such as [40],[26],[27],[25]. Gear and Petzold in 1984 shown that [40], when $\mu$ <<-1/2, then the backward Euler method is unable to solve it numerically, and in [25], numerical methods based on finite differences encounter difficulty. In 1994 Amodio [44], solved it by techniques of boundary values, but the rate of convergence for $\mu$<-1/2 is very low. It has been shown that there is no solution of the equations defining $y_n$ using backward Euler discretization [28]. We solved it for $\mu = -1$, and examined it with different values of $N$. The results are given in Table 4

Table 4

| $N$ | $e_{ps}$ | $h$ | $e_A$ |
|---|---|---|---|
| 6 | $2.61\times10^{-6}$ | $10^{-1}$ | $7.06\times10^{-6}$ |
| 10 | $2.07\times10^{-12}$ | $1.25\times10^{-2}$ | $1.30\times10^{-7}$ |
| 14 | $9.12\times10^{-17}$ | $6.25\times10^{-3}$ | $1.66\times10^{-8}$ |

Here, again, we have better results.

*Example 4.5:* Consider

$$x_1' = \left(\alpha - \frac{1}{2-t}\right)x_1 + (2-t)\alpha z + \frac{3-t}{2-t}e^t,$$
$$x_2' = \frac{1-\alpha}{t-2}x_1 - x_2 + (\alpha-1)z + 2e^x,$$
$$0 = (t+2)x_1 + (t^2-4)x_2 - (t^2+t-2)e^t,$$

where $\alpha$ is a parameter and $t \in [0,1]$. This DAE is in pure index-2 form (4.17), and was chosen from [28]. For the initial conditions $x_1(0) = x_2(0) = 1$ we have the exact solutions

$$x_1(t) = x_2(t) = e^x,\ z(t) = -\frac{e^x}{2-t}.$$

This problem was chosen from [28]. Although two initial conditions were used, but with our method we need only one initial condition. The value of $\alpha$ has been selected as 10 and has been integrate this DAE from *t=0* to *t*=1 using the first three BDF methods. The maximum errors for different values of *h* ranging from 1/20 to 1/2560 is, nearly, $10^{-4}$ . By theorem 4, this problem can be converted to the system,

$$\begin{pmatrix} 0 & 0 \\ (\alpha-1) & \alpha(t-2) \end{pmatrix}\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} + \begin{pmatrix} t+2 & t^2-4 \\ \dfrac{\alpha-1}{2-t} & \alpha(t-2) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} =$$
$$\begin{pmatrix} t^2+t-2 \\ \dfrac{(\alpha-1)(3-t)-2\alpha(2-t)^2}{2-t} \end{pmatrix} e^t.$$

We record the results of using the pseudo-spectral method with $\alpha = 10$, and with one initial condition. The comparison between these results mentioned in Table 5, and the results in [28], show the power of the proposed method for this example.

Table 5

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 7 | $1.4867\times10^{-8}$ | $1.2458\times10^{-8}$ |
| 10 | $2.4736\times10^{-13}$ | $1.8296\times10^{-13}$ |
| 13 | $5.2301\times10^{-17}$ | $3.8627\times10^{-17}$ |

Let us consider another example. This example was chosen from [39].

*Example4.6:* Consider

$$x_1' = x_1 + (\sin\alpha t)y + (2+\frac{\sin\alpha t}{2-t})e^t,$$

$$x_2' = x_2 + (\cos\alpha t)y + (2+\frac{\cos\alpha t}{2-t})e^t, \qquad t\in[0,1],$$

$$0 = (\sin\alpha t)x_1 + (\cos\alpha t)x_2 - (\sin\alpha t+\cos\alpha t)e^t,$$

with $x_1(0) = x_2(0) = 1$, where $x_1(t) = x_2(t) = e^t$ and $y(t) = \frac{e^t}{t-2}$.

By theorem 4, we converted this problem to the system,

$$(\cos\alpha t)x_1' - (\sin\alpha t)x_2' + (\cos\alpha t)x_1 - (\sin\alpha t)x_2 = 2e^t(\cos\alpha t - \sin\alpha t),$$
$$0 = (\sin\alpha t)x_1 + (\cos\alpha t)x_2 - (\sin\alpha t+\cos\alpha t)e^t,$$

where $t\in[0,1]$. Here we choose $\alpha = 1000$, as in [39] with one initial condition, and solved it by using pseudo-spectral method with one initial condition. The results are given in Table 6.

Table 6

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 7 | $6.9462\times10^{-9}$ | $1.0646\times10^{-8}$ |
| 10 | $1.3367\times10^{-13}$ | $1.1102\times10^{-13}$ |
| 13 | $5.2301\times10^{-17}$ | $3.8627\times10^{-17}$ |

Ascher and Petzold [39], have considered this example using Baumgarte's technique with backward Euler and applying backward Euler directly to the original index-2 DAE, with $\alpha = 1000, h = 0.01$. The comparison between the general results mentioned in Table 6, and published results in [39], shows the power of pseudo-spectral method, for this example.

## 4.8 DAEs with non-analytical coefficient functions

When coefficient functions or solution functions are non-analytic, we can solve a DAE problem by use of domain decomposition [35], [37], but as chapter three, without loss of generality, fortunately, when the differential equations or the constraint have, at least, one non-analytical coefficient function, with appropriate position and choice of Gauss-Chebyshev-Radau points, we do not need to do that. In such cases we will get good results [49,51]. To observe this we continue to consider some examples which have, at least, one non-analytic coefficient.

*Example 4.7*: Consider

$$\begin{aligned} &\cos(t)y_1'(t) - \sin(t)y_2'(t) = 1 \\ &|t|y_1(t) + y_2(t) = |t|\sin(t) + \cos(t), \end{aligned} \quad ,t \in [-1,1],$$

with initial condition $y_1(-1) = \sin(-1)$.

The exact solutions are, $y_1(t) = \sin(t)$, and $y_2(t) = \cos(t)$.

Here, some coefficient functions are not analytic at zero. We used the pseudo-spectral method and the results are given in Table 7.

Table 7

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $5.426\times10^{-2}$ | $2.656\times10^{-2}$ |
| 6 | $2.488\times10^{-5}$ | $1.806\times10^{-5}$ |
| 10 | $2.077\times10^{-9}$ | $1.598\times10^{-9}$ |

Let us consider another example in which the coefficient functions are not analytic in more than one point.

*Example 4.8:* Consider

$$y_1'(t) - ty_2'(t) + y_1(t) = te^{-t},$$
$$\left|t-\tfrac{1}{2}\right| y_1(t) + \left|t+\tfrac{1}{2}\right| y_2(t) = e^{-t}\left(\left|t-\tfrac{1}{2}\right| + \left|t+\tfrac{1}{2}\right|\right) \quad t \in [-1,1],$$

with initial condition $y_1(-1) = e$.

The exact solutions are, $y_1(t) = y_2(t) = e^{-t}$.

Here, we have two points, in which the coefficient functions are not analytic. Again, we solved this example by this method and the results are given in Table 8.

Table 8

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $5.297\times10^{-3}$ | $7.165\times10^{-3}$ |
| 6 | $2.942\times10^{-5}$ | $5.249\times10^{-5}$ |
| 10 | $7.717\times10^{-9}$ | $2.261\times10^{-8}$ |

Now we consider an index-2 form (4.17) with non-analytic coefficients functions in its constraint.

*Example 4.9:* Consider

$$\tfrac{1}{2}x_1' = \sin(2t-1)x_1 + \cos(2t-1)x_2 - y + e^{(2t-1)}[\cos(2t-1) + \sin(2t-1)],$$
$$\tfrac{1}{2}x_2' = -x_1 + x_2 + e^{(2t-1)}y - e^{(2t-1)},$$
$$0 = x_1 + |t|x_2 - e^{(2t-1)}[\cos(2t-1)\sin(2t-1) + |2t-1|\cos(2t-1)], \; t \in [0,1],$$

with initial condition $x_1(0) = e^{-1}\sin(-1)$, where the exact solution is

$$x_1(t) = e^{2t-1}\sin(2t-1),\; x_2(t) = e^{(2t-1)}\cos(2t-1),\; y = e^{(2t-1)}.$$

By applying formulation of index reduction in section 4.6, and change of variable $t = \frac{(b-a)T+(b+a)}{2}$ and changing $T$ to $t$ on given DAE, we obtain,

$$\begin{aligned} \tfrac{1}{2}e^{(2t-1)}x_1' + \tfrac{1}{2}x_2' &= [e^{(2t-1)}\sin(2t-1)-1]x_1 + [e^{(2t-1)}\cos(2t-1)+1]x_2 \\ &\quad + [e^{(2t-1)}[\cos(2t-1)+\sin(2t-1)], \\ 0 &= x_1 + |2t-1|x_2 - e^{(2t-1)}[\sin(2t-1)+|2t-1|\cos(2t-1)]. \end{aligned}$$

We record results of running the pseudo-spectral method for this example, and the results are presented in Table 9.

Table 9

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $2.7764\times10^{-2}$ | $3.8124\times10^{-2}$ |
| 7 | $2.7560\times10^{-5}$ | $4.0283\times10^{-5}$ |
| 10 | $5.7678\times10^{-9}$ | $8.0090\times10^{-9}$ |

We see this method will be, also, useful for a DAE, even, when differential equations have non-analytic coefficient functions.

*Example 4.10:* Consider

$$\begin{aligned} &|t|x_1' + x_2' + x_1 = |t|\cos t, \\ &(\sin t)x_1 + (\cos t)x_2 = 1, \end{aligned}$$

where $t \in [-1,1]$, with initial condition $x_1(-1) = \sin(-1)$.

The exact solutions are $x_1(t) = \sin t$, $x_2(t) = \cos t$.

We solved it for *N=4, 7, 11,* and the results are given in Table 10.

Table 10

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $7.1356\times10^{-3}$ | $4.9252\times10^{-3}$ |
| 7 | $2.8846\times10^{-6}$ | $9.0824\times10^{-7}$ |
| 11 | $3.2327\times10^{-11}$ | $8.3095\times10^{-12}$ |

As we see this method works well be for DAEs which have non-analytic coefficient functions.

In example 4.1, we solved a system of differential equations by this method. We mention here that this method can be used for a system of differential equations, even if, there are non-analytic coefficient functions.

We end this chapter by solving a system of differential equations with non-analytic coefficient functions.

*Example 4.11:* Let us consider

$$(\cos t)x_1' - (\sin t)x_2' = 1,$$
$$|t|x_1' + x_2' + x_1 = |t|\cos t, \; t \in [-1,1],$$

with initial conditions $x_1(-1) = \sin(-1)$, $x_2(-1) = \cos(-1)$. The exact solutions are

$x_1(t) = \sin t$ , $x_2 = \cos t$ . In this system we have non-analytical coefficient functions. We used the pseudo-spectral method for this system and the results are shown in Table 11.

Table 11

| $N$ | $e_1$ | $e_2$ |
|---|---|---|
| 4 | $4.819\times10^{-3}$ | $5.701\times10^{-3}$ |
| 6 | $2.297\times10^{-5}$ | $3.827\times10^{-5}$ |
| 10 | $2.384\times10^{-10}$ | $3.926\times10^{-10}$ |

Unfortunately, for systems of differential equations of order one or more and a DAE with solution which is not analytic, the rate of convergence will be very low by this method, but it seems it may be possible to do some more work on such systems by spectral methods to have better results.

## 4.9 Some conclusions about use of Pseudo-spectral method

Numerical results of all examples in chapters three and four show the efficiency of spectral methods. In all DAE examples considered having one algebraic constraint and differentiation of order one, we obtained a spectral accuracy. Numerical results for most examples confirm good accuracy of pseudo-spectral method comparing with other methods. Rate of convergence of Adams method which is known as a good method [11], comparing with pseudo-spectral method is very low. Another advantage of this method is its flexibility compared to finite difference methods.

Unfortunately, in presence of non-analytical solution and/or coefficient functions, our modification to speed up convergence of, linear ODE problems [21] does not improve rate of convergence very much when applied to DAE problems. Research in this matter is one of our future goals.

Another research which may be possible to do is, we extend our works to unbounded intervals and Inegral-Algebraic Equations(IAEs), and more general case when we have a mixed equations of integral and differential algebraic equations.

I hope to have a suggestion to solve these kind problems in a near future.

# Appendix A

## Some Basic Mathematical Concepts

### A.1. *Hilbert, Banach and Sobolov spaces*

Let $X$ be a real vector space. An *inner product* on $X$ is a function $X \times X \to \Re$ denoted by $(u,v)$, which satisfies following properties:

*(i)* $(u,v) = (v,u)$ for all $u,v \in X$;

*(ii)* $(\alpha u + \beta v, w) = \alpha(u,w) + \beta(b,w)$ for all $\alpha, \beta \in \Re$ and all $u,v,w \in X$;

*(iii)* $(u,u) \geq 0$ for all $u \in X$;

*(iv)* $(u,u) = 0$ implies $u = 0$.

Two elements $u,v \in X$ are said to be *orthogonal* in $X$ if $(u,v) = 0$. The inner product $(u,v)$ defines a *norm* on $X$ by the relation

$$\|u\| = (u,u)^{1/2} \qquad \text{for all } u \in X.$$

The *distance* between two elements $u,v \in X$ is the positive number $\|u - v\|$.

A sequence $\{u_n\}_1^\infty$ in $X$ is called a *Cauchy sequence* if, for every positive number $\varepsilon$, there exists a positive integer $N = N(\varepsilon)$ such that $\|x_n - x_m\|_X < \varepsilon$ whenever both $m$ and $n$ exceed $N$.

A sequence in $X$ is said to *converge* to an element $u \in X$ if the distance $\|u_n - u\|$ tends to 0 as $k$ tends to $\infty$.

A normed linear space $X$ is said to be *complete* if every Cauchy sequence in $X$ converges to an element in $X$.

A *Hilbert space* is a vector space equipped with an inner product for which all the Cauchy sequence are convergent.

For example, $\Re^n$ endowed with Euclidean product

$$(u,v) = \sum_{i=1}^{n} u_i v_i$$

is a finite dimensional Hilbert space.

If $[a,b] \subset \Re$ is an interval, the space $L^2(a,b)$ is an infinite dimensional Hilbert space for the inner product

$$(u,v) = \int_a^b u(x)\,v(x)dx.$$

*Banach space*: The concept of Banach space extends that Hilbert space. Given a vector space $X$, a norm on $X$ is a function $X \to \Re$ denoted by $\|u\|$ which satisfies the following properties:

$\|u+v\| \le \|u\| + \|v\|$ for all $u, v \in X$;

$\|\lambda u\| = |\lambda| \|u\|$ for all $u \in X, \lambda \in \Re$;

$\|u\| \ge 0$ for all $u \in X$;

$\|u\| = 0$ if and only if $u$=0.

A Banach space is a linear vector space equipped with a norm for which the space is complete.

For example; $\Re^n$ endowed with the norm

$$\|u\| = \left( \sum_{i=1}^{n} |u_i|^p \right)^{1/p}$$

(with $1 \le p < \infty$) is a finite dimensional Banach space.

If $[a,b] \subset \Re$ is an interval and $1 \le p < \infty$, the space $L^p(a,b)$ is an infinite dimensional Banach space for norm

$$\|u\| = \left( \int_a^b |u(x)|^p \, dx \right)^{1/p}.$$

A *Sobolov space* of order $m$ is a space of square integrable functions that possesses $m$ derivatives that are representable as square integrable functions:

$$H^m(a,b) = \left\{ u \in L^2(a,b) \middle| \frac{\partial^k u}{\partial x^k} \in L^2(a,b), 1 \le k \le m \right\}.$$

$H^m(a,b)$ is endowed with inner product:

$$(u,v)_{H^m(a,b)} = \sum_{k=0}^{m} \int_a^b \frac{\partial^k u}{\partial x^k} \frac{\partial^k v}{\partial x^k} dx,$$

and norm: $(u,v)_{H^m(a,b)} = \sqrt{(u,u)_{H^m(a,b)}}$.

Following property can be derived:

$$H^{m+1}(a,b) \subset H^m(a,b) \subset ... \subset H^0(a,b) \equiv L^2(a,b).$$

**A.2.** ***The Lebesgue Integral and $L^p$ - spaces***

(a) *The Lebesgue (Outer) Measure*

Each set $A$ contained in $(a,b)$ can be covered by a countable union of open intervals $I_n$,i.e. $A \subset \bigcup_{n=0}^{\infty} I_n$. Taking into account this property, the *Lebesgue outer measure* $\mu(A) = \inf \sum_n |I_n|$, where $|I_n|$ denotes the length of the interval $I_n$ and the infimum is taken over all the coverings $A$ by open intervals. Note that the measure of an interval is its length.

(b) *Measurable Sets*

A set $A \subseteq (a,b)$ is said to be measurable if

$$\mu(A) + \mu(\tilde{A}) = \mu(a,b) = b - a,$$

where $\tilde{A}$ denotes the complementary set of $A$.

In Lebesgue's measure theory only measurable sets are of interest.

(c) *Simple Measurable Functions*

A function $s:(a,b)\to[0,\infty)$ is a *Simple Measurable Function* if it assumes only a finite number of values $\{s_0,...,s_n\}$ and if each set $A_i=\{x\in(a,b):s(x)=s_i\}$ is measueable.

(d) *Measurable Functions*

A positive function $u:(a,b)\to[0,\infty)$ is *Measurable* if it is the pointwise limit of simple measurable functions $s^{(n)}$ such that

$(i)\ 0\le s^{(1)}\le s^{(2)}\le ...\le u$

$(ii)\ s^{(n)}(x)\to u(x)\ \ as\ n\to\infty,\ \forall x\in(a,b).$

(e) *Lebesgue Integral*

If $s$ is a simple measurable function on $(a,b)$, we set

$$\int_a^b s\,d\mu=\sum_{i=0}^{n}s_i\mu(A_i).$$

If $u$ is a positive measurable function on $(a,b)$, we set

$$\int_a^b u\,d\mu=\sup\int_a^b s\,d\mu,$$

the supremum being taken over all the simple measurable function such that

$0\le s\le u$. The value of the right-hand side is a a non-negative number or $+\infty$. We call it the *Lebesgue Integral* of $u$ on $(a,b)$.

A positive measurable function $u$ is said to be *Lebesgue Integralable* on $(a,b)$ if

$$\int_a^b u\,d\mu<+\infty.$$

(f) *The Spaces* $L^p(a,b),\ 1\le p\le\infty$

Let *(a,b)* be a bounded interval of $\Re$ and let $1 \le p \le \infty$. We denote by $L^p(a,b)$ the space of the measurable function $u:(a,b) \to \Re$ such that

$$\int_a^b |u(x)^p| dx < +\infty.$$

Endowed with the norm

$$\|u\|_{L^p(a,b)} = \left( \int_a^b |u(x)|^p dx \right)^{1/p},$$

it is a Banach space.

For $p = \infty, L^\infty(a,b)$ is the space of the measurable function $u:(a,b) \to \Re$ such that $|u(x)|$ is bounded outside a set of measure zero.

The index $p = 2$ is of special interest, because $L^p(a,b)$ is not only a Banach space, but also Hilbert space. The inner product is

$$(u,v) = \int_a^b u(x)v(x)dx,$$

which induces the norm

$$\|u\|_{L^2(a,b)} = \left( \int_a^b |u(x)|^2 dx \right)^{1/2}.$$

(g) *The weighted Spaces* $L^p(-1,1), 1 \le p \le \infty$

Let *w(x)* be a weight function on the interval (-1,1), i.e., a continuous, strictly positive and integrable function on (-1,1). For $p < +\infty$, we denote by $L^p_w(-1,1)$ the Banach space of the measurable functions $u:(a,b) \to \Re$ such that $\int_a^b |u(x)^p| w(x) dx < +\infty$. It is endowed with the norm

$$\|u\|_{L^p_w(-1,1)} = \left( \int_a^b |u(x)|^p w(x) dx \right)^{1/p}.$$

For $p = \infty$ we set $L^{\infty}_{w}(-1,1) = L^{\infty}(-1,1)$.

The space $L^{2}_{w}(-1,1)$ is a Hilbert space for the inner product

$$(u,v)_{w} = \int_{a}^{b} u(x)\,v(x)\,w(x)dx,$$

which induces the weighted norm

$$\|u\|_{L^{2}_{w}(-1,1)} = \left( \int_{a}^{b} |u(x)|^{2}\, w(x)\, dx \right)^{1/2}.$$

# Appendix B

## Runge-Kutta Methods

### *B.1. Forward Euler*

*Runge-Kutta Methods* are in some sense a generalization of *Euler's Classic Method* for solving ODEs. Euler's method, commonly referred to by numerical analysts as *Forward Euler*, is used to solve initial-value problems (IVPs) of the form

$$x' = f(x,t), \quad x(t_0) = x_0,$$

where $x \in \Re^n, t \in [t_0, t_k] \subset \Re$.

Suppose we are interested in approximating the value $x(t_k)$. We begin by discretizing the domain of $t$ into small intervals $[t_n, t_{n+1}]$ which we refer to as *steps*. The width, $h_n = t_{n+1} - t_n$, of each step is referred to as the *(local) step-size*. We then approximate the value of $x$ at each of $t_n$ $n \geq 1$, by evaluating

$$x_{n+1} = x_n + f(x_n, t_n) h_n,$$

which can be thought of as projection along a tangent line to the (unknown) function governing $x$. We then simply repeat this process for each $x_n$ until we obtain an approximation to $x(t_k)$.

### *B.2. Runge-Kutta Methods*

#### 1. *Explicit Runge-Kutta Methods*

The general Runge-Kutta method can be written as

$$y_{n+1} = y_n + \sum_{i=1}^{v} w_i K_i, \quad n = 0(1) N-1,$$

where

$$K_i = hf(x_n + c_i h, y_n + \sum_{m=1}^{i-1} a_{im} K_m),\ c_1 = 0.$$

For $v$=1, $w$=1, we get the Euler method. This is the lowest order Runge-Kutta method.

We list a few Runge-Kutta methods with higher orders.

**(i) Second order methods**

(a) ***method*: *Improved Tangent***

$$y_{n+1} = y_n + K_2,\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}).$$

(b) ***Modified Euler method*:**

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + h, y_n + K_1).$$

**(ii) Third order methods**

(a) ***Nystrom method*:**

$$y_{n+1} = y_n + \frac{1}{8}(2K_1 + 3K_2 + 3K_3),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_1),$$
$$K_3 = hf(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_2).$$

(b) ***Heun method*:**

$$y_{n+1} = y_n + \frac{1}{4}(K_1 + 3K_3),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{1}{3}h, y_n + \frac{1}{3}K_1),$$
$$K_3 = hf(x_n + \frac{2}{3}h, y_n + \frac{2}{3}K_2).$$

(c) ***Classical method*:**

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 4K_2 + K_3),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1),$$
$$K_3 = hf(x_n + h, y_n - K_1 + 2K_2).$$

**(iii) Fourth order methods**

(a) ***Kutta method*:**

$$y_{n+1} = y_n + \frac{1}{8}(K_1 + 3K_2 + 3K_3 + K_4),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{1}{3}h, y_n + \frac{1}{3}K_1),$$
$$K_3 = hf(x_n + \frac{2}{3}h, y_n - \frac{1}{3}K_1 + K_2),$$
$$K_4 = hf(x_n + h, y_n + K_2 - K_3).$$

(b) ***Classical method*:**

$$y_{n+1} = y_n + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4),\ n = 0(1)N-1,$$
$$K_1 = hf(x_n, y_n),$$
$$K_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_1),$$
$$K_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}K_2),$$
$$K_4 = hf(x_n + h, y_n + K_3).$$

## 2. *Implicit Runge-Kutta Methods*

The general Runge-Kutta given in *Explicit form* can be modified to

$$y_{n+1} = y_n + \sum_{i=1}^{v} w_i K_i, \quad n = 0(1) N-1,$$

where

$$K_i = hf(x_n + c_i h, y_n + \sum_{m=1}^{v} a_{im} K_m).$$

With $v$ function evaluation, implicit Runge-Kutta methods of order $2v$ can be obtained. A few methods are listed.

**(i) Second order methods**

$$y_{n+1} = y_n + K_1, \quad n = 0(1) N-1,$$
$$K_1 = hf(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}).$$

**(ii) Fourth order methods**

$$y_{n+1} = y_n + \frac{1}{2}(K_1 + K_2), \quad n = 0(1) N-1,$$
$$K_1 = hf(x_n + (\frac{1}{2} - \frac{\sqrt{3}}{6})h, y_n + \frac{1}{4}K_1 + (\frac{1}{4} - \frac{\sqrt{3}}{6})K_2),$$
$$K_2 = hf(x_n + (\frac{1}{2} + \frac{\sqrt{3}}{6})h, y_n + (\frac{1}{4} + \frac{\sqrt{3}}{6})K_1 + \frac{1}{4}K_2).$$

## *B.3. Multistep Methods*

The general Multistep method can be written as

$$y_{n+1} = a_{m-1} y_m + a_{m-2} y_{n-1} + \ldots a_0 y_{n+1-m}$$
$$+ h[b_m f(x_{n+1}, y_{n+1}) + b_{m-1} f(x_n, y_n) + \ldots + b_0 f(x_{n+1-m}, y_{n+1-m}),$$

with $n = m-1(1)N-1$, where the starting values

$y_0 = \alpha, y_1 = \alpha_1, \ldots, y_{m-1} = \alpha_{m-1}$

are specified and $h=(b-a)/N$.

When $b_m = 0$, the method is called *Explicit.* But, if $b_m \neq 0$, the method is called *Implicit.*

We list here just two of multistep methods.

**(i) Fourth order Adams-Bashforth method**

$$y_0 = \alpha, y_1 = \alpha_1, y_2 = \alpha_2, y_3 = \alpha_3,$$
$$y_{n+1} = y_n + \frac{h}{24}[55f(x_n, y_n) - 59f(x_{n-1}, y_{n-1}) + 37f(x_{n-2}, y_{n-2}) - 9f(x_{n-3}, y_{n-3})],$$

for each $i$=3(1)$N$-1.

**(ii) Fourth order Adams-Moulton method**

$$y_0 = \alpha, y_1 = \alpha_1, y_2 = \alpha_2,$$
$$y_{n+1} = y_n + \frac{h}{24}[9f(x_{n+1}, y_{n+1}) + 19f(x_n, y_n) - 5f(x_{n-1}, y_{n-1}) + f(x_{n-2}, y_{n-2})],$$

for each $i$=2(1)$N$-1.

To have more information about numerical methods for numerical solution of BVPs for ODEs one can refer to [50].

### *B.4. Shooting Methods*

Consider the linear boundary-value problem,

$$y'' = p(x)y' + q(x)y + r(x),\ a \le x \le b,\ y(a) = \alpha,\ y(b) = \beta. \qquad (*)$$

To approximate the unique solution, let us first consider the initial-value problems,

$$y'' = p(x)y' + q(x)y + r(x),\ a \le x \le b,\ y(a) = \alpha,\ y'(a) = 0\ , \qquad (**)$$

$$y'' = p(x)y' + q(x)y,\ a \le x \le b,\ y(a) = 0,\ y'(a) = 1. \qquad (***)$$

If $y_1(x)$ denotes the solution to Eq.(**) and $y_2(x)$ denotes the solution to Eq.(***), it is not difficult to verify that

$$y(x) = y_1(x) + \frac{\beta - y_1(b)}{y_2(b)} + y_2(x) \qquad (****)$$

Is the unique solution to our boundary-value problem, provided, of course, that $y_2(b) \neq 0$.

The Shooting method for linear equations is based on this replacement of the boundary-value problem (**) and (***). Numerous methods are available for approximating the solutions $y_1(x)$ and $y_2(x)$, and once these approximations are available, the solution to boundary-value problem is approximated using Eq.(****).

We used the fourth-order Runge-Kutta technique to find the approximations to $y_1(x)$ and $y_2(x)$.

***B.5. Difference Methods***

There are difference equations obtained from a given differential equation. The system of equations is then solved by direct or indirect methods. Let us consider a linear second order BVP of the form

$$-y'' + f(x)y = r(x),\ x \in [a,b],$$
$$y(a) = \alpha_1, y(b) = \alpha_2.$$

We assume $f(x) \geq 0$ for $x \in [a,b]$ to ensure the existence and uniqueness of the solution. In order to compute a numerical approximation to the solution *y(x)*, we first divide the interval [*a, b*] into $N$+1 subintervals of length $h = {}^{(b-a)}/_{(N+1)}$, and at each point $x_n = a + nh,\ n = 1,2,...,N$, approximate $y''(x_n)$ by the second central difference quotient

$$y''(x_n) = {}^{1}/_{h^2}[y(x_{n+1}) - 2y(x_n) + y(x_{n-1})] + O(h^2), n = 1(1)N.$$

When this approximation is used in a given problem, we find the solution satisfies

$$\frac{1}{h^2}[y(x_{n+1})-2y(x_n)+y(x_{n-1})]+f(x_n)\,y(x_n)+O(h^2)=r(x_n)$$

at the grid points $x_1, x_2, ..., x_N$.

Dropping the error term in this equation and defining approximations $y_1, y_2, ..., y_N$ to the values of the solution at the grid points $x_j$, we get the system of $N$ equations

$$-y(x_{n-1})+2y(x_n)-y(x_{n+1})+h^2 f_n y_n = h^2 r_n.$$

The boundary conditions become $y_0 = \alpha_1$, $y_{N+1} = \alpha_2$.

If

$$\mathbf{J}=\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ . & . & . & & & \\ . & . & . & & & \\ . & . & . & & & \\ & & & -1 & 2 & -1 \\ & & & & 2 & -1 \end{bmatrix}, \mathbf{F}=\begin{bmatrix} f_1 & & & \\ & . & & \\ & & . & \\ & & & f_N \end{bmatrix}, \mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix}, \mathbf{C}=\begin{bmatrix} \alpha_1 + h^2 r_1 \\ h^2 r_2 \\ . \\ . \\ . \\ \alpha_2 + h^2 r_N \end{bmatrix},$$

then, after incorporating the boundary conditions, the system of difference equations can be written as

$$(\mathbf{J}+h^2\mathbf{F})\,\mathbf{y}=\mathbf{C}.$$

If $\det(\mathbf{J}+h^2\mathbf{F})\neq 0$, then the solution of the above system becomes

$$\mathbf{y}=(\mathbf{J}+h^2\mathbf{F})^{-1}\,\mathbf{C}.$$

# Bibliography


[1] T. J. Rivlin, *An Introduction to the Approximation of Functions*, New York,1969.

[2] P .J .Davis, and P.Rabinowitz, *Methods of Numerical Integration*, Acadamic Press, New York,1975.

[3] E. C. Titchmarsh, (1962): *Eigenfunction Expansions*, Part 1 ( Oxford, Univ. Press, London).

[4] D.Gottlieb, S.A.Orszag, *Numerical Analysis of Spectral Methods, Theory and Applications*, SIAM,Philadelphia,1982.

[5] C. Canuto, M. Y. Hussaini , A. Quarteroni , T. A. Zang , *Spectral Methods in Fluid Dynamics, Springer series in computational physics,* Springer-Verlag,NewYork,1988.

[6] C. Bernardi and Y. Maday. *Approximations Spectrales de problems aux Limites Elliptiques.* Springer-Verlag , Paris,1992.

[7] D. Funaro. *Polynomial Approximations of Diffrential Equations.* Springer-Verlag ,1992.

[8] T. J. Rivlin . *The Chebyshev polynomials* . John Wiley and Sons , 1974.

[9] Kreiss and Oliger , *"Stability of the Fourier* method", SIAM J.Numer.Anal , 16:421-33 (1979).

[10] J. W. Cooley and J. W. Tukey, *Math. computation* **19**,297-301(1965)

[11] S. A. Orszag, (1986): *Fast Eigen function Transforms, 'in Science and Computers, Advances in Mathematics* Supplementary Series, ed. By G. C. Rota (Academic Press, London, NewYork), pp.23-30.

[12] G. Szego (1939): *Orthogonal Polynomials*, Vol. 23 (AMS Coll. Publ., New York).

[13] C. Canuto, M. Hussaini, A. Quarteroni, T. Zang, *Spectral Methods in Fluid Dynamics,* Springer, Berlin, 1988.

[14] C. Lanczos, *Trigonometric interpolation of empirical and analytical functions,* J. Math. Phys. 17 (1938) 123-129.

[15] L. M. Delves and J. L. Mohamed, *Computational methods for integral equations*, Cambridge University Press, 1985.

[16] E. Babolian and L .M. Delves, *A fast Galerkin scheme for linear integro-differential equations*, IMAJ. Numer. Anal, Vol.1, pp. 193-213, 1981.

[17] B. Fornberg, *A practical guide to Pseudo-Spectral Methods*, Cambridge University press, Cambridge, 1996.

[18] A. B. Finlayson, L. E. Scriven (1966): *The method of weighted residuals-a review.* Appl. Mech. Rev. **19,** 735-748.

[19] U. M. Ascher, R. M. Mattheij and R. D. Russell, *Numerical solution of boundary value problems for ordinary differential equations,* perentice-Hall, Inc., 1988.

[20] M. Saravi, E. Babolian, R. England, M. Bromilow, *'An experimental modification of spectral methods'*, 21st Biennial Conference on Numerical Analysis.28 June-1 July, 2005,University of Dundee,UK.

[21] E. Babolian, M. Bromilow, R. England, M. Saravi, *'A modification of pseudo-spectral method for solving linear ODEs with singularity'*, AMC 188 (2007) 1260-1266.

[22] E. Babolian, M.M. Hosseini, *A modification spectral method for numerical solution of ordinary differential equations with non- analytical solution,* Appl. Math.Comput.132 (2002) 341-351.

[23] Al Rehni and F. Allgöwer (2001) *H∞ Control of Differential-Algebraic Equation System*, 6th European Control Conference, Portugal.

[24] C. W. Gear. *Simultaneous numerical solution of DAEs.* IEEE Trans. Circ. Th. 18(1):89-95, 1971.

[25] E. Hairer, C. Lubich and M.Roche, *The Numerical Solution of Differential- Algebraic Systems by Runge-Kutta methods*, Springer-Verlag, New York, 1989.

[26] C. W Gear (1971) *Numerical IVPs in ODE*, Prentice-Hall Inc., Englewood Cliffs, N.J.

[27] K.E. Brenan and B.E. Engquist (1985) *Backward Differentiation Approximations of Non-linear DAEs*, Dept. Comput. Sci. Uppsala Univ. Rpt.# 101, Uppsala, Sweden.

[28] U.M. Ascher and L.R. Petzold (1998) *Computer Methods for ODEs and DAEs*, SIAM.

[29] K. E. Brenan, S. I. Campbell and L.R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, SIAM, New York, 1989.

[30] C.W. Gear, H.H. Hsu and L. Petzold (1981) *DAEs Revisited proc. Numerical Methods for Solving Stiff IVPs*, Oberwolfach, W. Germany

[31] U.M. Ascher and P. Lin (1996) *'Sequential regularization methods for higher index DAEs with constant singularities:* The linear index-2 case', SIAM, *J. Num. Anal.*, Vol. 33.

[32] U. Ascher and L. Petzold, *Stability of computational methods for constrained dynamics systems,* SIAM J. Scient. Comput. 14 (1993), 95-120.

[33] C. W. Gear and L. R. Petzold, *ODE methods for the solution of differential- algebraic systems,* SIAM J. Numer. Anal., Vol. 21, No. 4, 1984.

[34] C. W. Gear, G.K. Gupta and B. Leimkuhler (1985) *'Automatic integration of Euler-Lagrange equations with constraints', J. Comput. Appl., Math.*, Vol. 12-12.

[35] D.C. Tarraf and H.H. Asada (2002) *On the Nature and Stability of DAEs, Preceedings of the American Control Conference Anchorage, AK*, May 8-10.

[36] U. M. Ascher and P. Lin (1991) *'Sequential regularization methods for non-linear higher index DAEs*, SIAM', *J. Sec. Comput.*, Vol. 18.

[37] U.M. Ascher (1994) *Stabilization of DAEs and Invariant Manifold*, Ascher, Hongslieng Chin. Sebastian Reich November 15.

[38] U.M. Ascher and L.R. Petzold (1994) *Stabilization of Constrained Mechanical System with DAEs and Invariant Manifolds*, Hongslieng Chin., Sebastian Reich, October 26.

[39] Peter Kunkel (2003) *Numerical Treatment of Unstructured DAEs with Arbitrary Index*. SES, Bart-Monopolt.

[40] E. Hairer and G. Wanner (1991) *Solving ODE II: Stiff and Differential- Algebraic Problems*, Springer-Verlag, Berlin.

[41] C.W. Gear and L.R. Petzold (1982) *ODE Methods for the Solution of DAEs*, Dept., Rpt. UIU CDCS-R-82-1103.

[42] C. W. Gear, *Differential-algebraic equation index transformations,* SIAM J. Sci. Stat. Comput., Vol. 9, No. 1, 1988.

[43] C. W. Gear and L.R. Petzold (1977) *ODE Methods for the Solution of DAEs*, SIAM–CBMS. Philadelphia.

[44] P. Amodio and F. Mazzia, *Boundary value methods for solution of DAEs*, Numer. Math., Vol. 66, pp.411-421, 1994.

[45] R. F. Sincovec, A. M. Erismam. E. L. Yip and M. A. Epton, *Analysis of descrip to systems using numerical algorithms*, IEEE Trans. Automat. Control, Ac-26, pp.139-147, 1981.

[46] L. Kalachev, *Boundary value problems for differential equations,* Numer. Funct.Anal. Appl., Vol 16, pp, 363-378, 1985.

[47] K. D. Clark, *A structural form for higher index semistate equations I: Theory and applications to circuit and control,* Lin. Alg. Appl., (to appear).

[48] M. Saravi, E. Babolian, R. England, M. Bromilow, M. Rastegari, '*Solution of linear ODEs system by pseudo-spectral method with coefficient singularity*', 22$^{st}$ Biennial Conference on Numerical Analysis.26-29, 2007,University of Dundee,UK.

[49] M. Saravi, E. Babolian, R. England, M. Bromilow, '*System of Linear Differential Equations and Differential-Algebraic Equations*', NUMDIFF-11, Sept 4-8, 2006, Halle, Germany.

[50] M. K. Jain, S. R. K. Iyengar, R. K. Jain, *Numerical Methods, Problems and Solutions,* 1994, New Age International (p) Ltd, Pub.

[51] M. Saravi, E. Babolian, R. England, M. Bromilow, '*System of Linear Differential Equations and Differential-Algebraic Equations*', Mathematical Methods physical Models and Simulation in Science and Technology, xxx (2007) xxx-xxx(Article in press}.

[52] C. Lanczos(1956), *Applied Analysis*, Prentice-Hall, Englewood Cliffs,N J.

[53] C. W. Clenshaw, *The numerical solution of linear DEs in Chebyshev series*, Cambridge: Philosophical Soc., 1957, 53:134-159.