Efficient Continuous Runge-Kutta Methods for Asymptotically Correct Defect Control

By

Hanan Drfoun

A Thesis Submitted to Saint Mary's University, Halifax, Nova Scotia in Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied Science.

Examination Committee:

Approved: Dr. Paul Muir, Senior Supervisor

Department of Mathematics and Computing Science

Approved: Dr. David Iron, External Examiner

Department of Mathematics and Statistics, Dalhousie University

Approved: Dr. Walt Finden, Supervisory Committee Member

Department of Mathematics and Computing Science

Approved: Dr. Bert Hartnell, Supervisory Committee Member

Department of Mathematics and Computing Science

Date: August 8, 2017

Table of Contents

List of Tables

List of Figures

Acknowledgements

This thesis would not have been possible without the help of Almighty Allah. I would like to thank him for all he has given us, for he alone is the giver and taker of all, from the new to the old.

I would like to acknowledge the financial support of the Ministry of Higher Education and Scientific Research of Libya. Great thankfulness goes to it for having given me the opportunity to pursue my masters degree in Canada.

I would like to express my sincere gratitude towards my supervisor Dr. Paul Muir for his invaluable guidance, advice and encouragement during the development of this thesis. I extend my thanks to the supervisory committee members, Dr. Walt Finden, Dr. Bert Hartnell and Dr. David Iron for reviewing and discussing this thesis.

I would like to thank all friends for their encouragement and good wishes.

Last but not least, I would like to express my deep thanks to my husband and my family for their unlimited support and encouragement. Their prayers and love was my source and motivation to continue and finish this research.

Abstract

Efficient Continuous Runge-Kutta Methods for Asymptotically Correct Defect Control

by

Hanan Drfoun

Mono-Implicit Runge-Kutta (MIRK) methods and continuous MIRK (CMIRK) methods, are used in the numerical solution of boundary value ordinary differential equations (ODEs). One way of assessing the quality of the numerical solution is to estimate its maximum defect, which is the amount by which the solution fails to satisfy the ODE. The standard approach is to perform two point sampling of the defect on each subinterval of a mesh which partitions the problem domain to estimate the maximum defect. However, the location of the maximum defect on each subinterval typically varies from subinterval to subinterval, and from problem to problem. Thus sampling at only two points typically leads to an underestimate of the maximum defect.

In this thesis, we will derive a new class of CMIRK interpolants for which the location of the maximum defect on each subinterval is the same over all subintervals and problems.

Date: August 8, 2017

Chapter 1

Introduction

Experimental and theoretical science has had a long history, but over the last 50-60 years computational science has also become a significant avenue of investigation. Almost every area of science, e.g., chemistry, biology, astronomy, engineering, etc., now has a computational component. Computational science is based on mathematical or computational models, which often involve a system of ordinary differential equations (ODEs). Such equations describe how a system will change with time. Since these equations are usually too complicated to be solved by hand, it is necessary to use numerical methods to solve them.

We will consider boundary value ordinary differential equations (BVODEs) which are systems of ODEs with boundary conditions imposed at two or more distinct points [2]. Some examples of difficult problems that arise in the study of real world phenomena in different areas of science that involve BVODEs are:

- Shock Wave in a one-dimensional nozzle flow, see example (1.17) in [2] (SWAVE problem)
  - ODE:

$$y''(t) = \left(\frac{\frac{1}{2} + \frac{\gamma}{2} - \epsilon A'(t)}{\epsilon A(t)} \right) y'(t) - \frac{y'(t)}{\epsilon A(t) y^2(t)}$$

$$-\frac{A'(t)}{\epsilon A^2(t)y(t)}\left(1-\frac{\gamma-1}{2}y^2(t)\right).$$

(1.1)

t: normalized downstream distance. $y(t)$: normalized velocity. $A(t)$: area of nozzle at t. ϵ: inverse of Reynolds number.

- Boundary conditions: $y(0) = 0.9129$, $y(1) = 0.375$.

- Swirling flow between two rotating coaxial disks, see example (1.20) in [2] (SWIRL-III problem)

  - ODEs:

$$\epsilon g''(t) = f'(t)g(t) - f(t)g'(t),$$

$$\epsilon f''''(t) = -f(t)f'(t) - g(t)g'(t)),$$

(1.2)

$f'(t)$, $g(t)$, $f(t)$: radial, angular, and axial velocities.

  - Boundary conditions:

$$f(0) = f(1) = f'(0) = f'(1) = 0,$$

$$g(0) = \Omega_0, \quad g(1) = \Omega_1.$$

Angular velocities, Ω_0 and Ω_1. ϵ: viscosity.

The process of obtaining a numerical solution to a BVODE involves computing an approximate solution at a set of mesh points, $\{t_i\}_{i=0}^{N}$, that partition the problem domain [44] using a numerical method such as Runge-Kutta (RK) method. This solution is called a discrete numerical solution. Each region, $[t_i,\ t_{i+1}]$, of the problem domain is called a subinterval. Continuous numerical methods [29] such as continuous RK are used to augment the discrete solution over each subinterval to yield a continuous numerical solution, over the entire problem domain. One class of RK methods that commonly used to provide a discrete solution to a system of BVODEs is called mono implicit RK (MIRK) methods. A MIRK scheme is of order p if it has error $O(h^p)$, ans it has stage order q where $q \leq p$ if its coefficients satisfy a set of q conditions called stage order conditions [28].

Although one can derive a MIRK method of any desired order, the resultant method can have at most stage order 3 [7]. This can be an issue when the method is applied to a stiff ODE because the order of the method can drop to its stage order. Thus, for example, even a 6^{th} order method can behave like a 3^{rd} order method. This is called order reduction [13]. Generalized MIRK (GMIRK) methods are extensions of MIRK schemes that allow the schemes to have a higher stage order. GMIRK schemes allow us to increase the number of coefficients associated with certain stages of the method that limit its stage order. These methods will not suffer from order reduction when applied to stiff ODEs. However, the GMIRK schemes have a greater number of implicit stages, and this increases the cost per subinterval associated with using these schemes above what is required for the use of a MIRK scheme.

For BVODEs, when the number of the implicit stages increases, the size of the non-linear system that must be solved changes. Rather than needing to solve a non-linear system of size $n(N+1)$ associated with $y_i, i = 0, ..., N$, the computation will require the solution of a non-linear system of size $n(N+1) + l \cdot n \cdot N$ where n is the number of ODEs, N is the number of subintervals, and l is the number of stages that are implicitly defined [13].

One way to assess the quality of an approximate solution is to examine the amount by which that solution fails to satisfy the BVODE; this is called the defect [15, 17, 23, 32, 34]. In a defect control framework, we need to estimate the maximum defect of the numerical solution on each subinterval. The user of the software provides a tolerance and the software adaptively chooses a sequence of meshes so that for the final accepted numerical solution, the estimated maximum defect is less than the user-provided tolerance; this is called defect control [21, 25].

The location of the maximum defect of a numerical solution on each subinterval generally varies from subinterval to subinterval, and from problem to problem [20]. It is computationally expensive to evaluate the defect at a large number of points in order to find the maximum defect for each subinterval. The hope is to sample the defect at only a small number of points on each subinterval but nonetheless obtain a good estimate of the maximum defect on each subinterval.

BVP_SOLVER_2 [25, 38] is a software package for the numerical solution of BVODEs that has an option for defect control. It uses just two point sampling to estimate the

maximum defect on each subinterval with the hope that one of them is close to the location of the true maximum defect. In [20], it was shown that the two point sampling approach is not reliable, and in that paper a new approach was considered that led to better estimation of the maximum defect. The new approach estimates the maximum defect using only one defect sample point per subinterval. This leads to what is known as Asymptotically Correct Defect Control (ACDC).

The approach considered in [20] was only for the sixth order case (i.e., for the case where the numerical solution has an error that is $O(h^6)$, where h is the subinterval size) and it employed an algorithm that relied upon constructing a new continuous approximate solution using a Hermite-Birkhoff interpolant based on the previously computed continuous numerical solution, i.e., a boot-strapping process [20]. We discuss this approach later in the thesis.

The main purpose of this thesis is to describe the development of families of Runge-Kutta methods that provide ACDC while being more efficient than those based on Hermite-Birkhoff interpolants.

This thesis is organized as follows. In Chapter 2, BVODEs are discussed. MIRK schemes and the associated Continuous Mono-Implicit Runge-Kutta (CMIRK) schemes are presented. We also present Runge-Kutta order conditions, continuous Runge-Kutta order conditions, and Runge-Kutta stage order conditions that are used to derive MIRK methods and CMIRK methods. The idea of defect control, the ACDC property, the derivation of the Hermite-Birkhoff interpolants via a boot-strapping

approach, and the idea of a validity check are also explained.

Chapter 3 first describes standard CMIRK schemes from orders 1 to 5 and the issue with these numerical methods regarding maximum defect estimation. It also provides numerical experiments and results obtained by applying standard CMIRK schemes to some test problems. In addition, this chapter discusses the boot-strapping approach for orders 4 and 5 and identifies an issue with the approach (namely, that the resultant interpolant uses more evaluations of the right hand side of the ODE than may be necessary). Finally, numerical experiments and results associated with applying the boot-strapping approach are provided.

Chapter 4 considers our proposed solution to the issue identified with the boot-strapping approach. In particular, we derive new CMIRK methods of orders 4 and 5 that directly have the ACDC property and that are more efficient than the boot-strapping approach in terms of the number of evaluations of the right hand side of the ODE. Also, we discuss numerical experiments and give results obtained by applying these new fourth and fifth orders CMIRK schemes to some test problems.

Next, in Chapter 5, a comparison between ACDC schemes that use the boot-strapping approach and the new ACDC CMIRK schemes, for orders 4 and 5, based on results obtained by applying these schemes to some test problems, is considered.

Chapter 6 first describes standard sixth order CMIRK schemes and provides numerical results obtained by applying these schemes to solve several test problems.

In addition, this chapter discusses the sixth order boot-strapping approach and provides numerical results associated with applying this approach. We then consider new sixth order ACDC CMIRK schemes and new sixth order continuous generalized ACDC MIRK (CGMIRK) schemes.

Finally, Chapter 7 gives the conclusions from this thesis and suggestions for future work.

Chapter 2

Ordinary Differential Equations, Runge-Kutta Methods and Asymptotically Correct Defect Control

2.1 Ordinary Differential Equations

An ODE is an equation that involves a function of one independent variable (e.g., time) and one or more derivatives of that function with respect to that independent variable. ODEs arise in mathematical models in many areas of science and engineering.

2.1.1 Boundary value ODES

The type of ODE that we consider in this thesis is called a BVODE. BVODEs are systems of ODEs with boundary conditions imposed on the solution at two or more distinct points [2]. A BVODE may not have a solution, or may have a finite number of solutions, or may have infinitely many solutions. Many problems, arising in a wide variety of application areas, give rise to mathematical models which involve BVODEs. These problems rarely have closed form solutions and computational methods are often used to estimate their approximate solution [1, 2, 3, 4, 37]. Many methods are available to carry out such computations in a robust, efficient, and reliable manner

[2, 6, 44].

In this thesis, we will assume non-linear two-point BVODEs written in first order system form with coupled boundary conditions, of the form,

$$y'(t) = f(t, y(t)), \quad g(y(a), y(b)) = 0, \tag{2.1}$$

where $t \in [a, b]$, $y : \mathrm{R} \to \mathrm{R}^n$, $f : \mathrm{R} \times \mathrm{R}^n \to \mathrm{R}^n$, and $g : \mathrm{R}^n \times \mathrm{R}^n \to \mathrm{R}^n$.

2.2 Runge-Kutta Methods

Runge-Kutta (RK) methods are numerical methods which are popular for solving systems of BVODEs (2.1) [11, 41, 44]. Let the problem interval $[a, b]$ be subdivided by a mesh $\{t_i\}_{i=0}^N$, with $a = t_0 < t_1 < \cdots < t_N = b$. Through use of the RK methods, we get a discrete numerical solution, $y_i \approx y(t_i), i = 0, \ldots, N$, by applying Newton's method to solve a nonlinear system of equations consisting of the boundary conditions and n more equations for each subinterval which depend on the RK scheme.

For a RK method, for the ith subinterval, $[t_{i-1}, t_i]$, where y_i is an approximation to the exact solution, $y(t)$, evaluated at the point t_i, and where $h_i = t_i - t_{i-1}$, we define an equation of the form

$$\phi_i = y_{i+1} - \left(y_i + h_i \sum_{r=1}^{s} b_r k_r \right), \tag{2.2}$$

where the stages are given by

$$k_r = f\left(t_i + c_r h_i, y_i + h_i \sum_{j=1}^{s} a_{rj} k_j\right), \quad r = 1, 2, \ldots, s. \tag{2.3}$$

The coefficients of this method are given in a Butcher tableau of the form

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \ldots & a_{1s} \\ c_2 & a_{21} & a_{22} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss} \\ \hline & b_1 & b_2 & \ldots & b_s \end{array} .$$

The above tableau is sometimes condensed to:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array},$$

where $c = (c_1, c_2, ..., c_s)^T$, $b = (b_1, b_2, ..., b_s)^T$, and A is the s by s matrix whose (i,j)th component is a_{ij}. Also, we usually require $c = Ae$, where e is the vector of l's of length s. This is equivalent to requiring $c_r = \sum_{j=1}^{s} a_{rj}$. Interpolants for Runge-Kutta methods have also been developed; see, e.g., [33].

2.3 Mono Implicit Runge-Kutta (MIRK) Methods

The MIRK methods [8, 9, 12, 22, 26, 30, 39, 43] are a subclass of the well-known implicit RK methods [2] and have application in the efficient numerical solution of systems of BVODEs [7, 10]. A significant property of this class of methods is that, for the discrete numerical solutions obtained by the use of BVODEs, the stage computations are explicit in y_i and y_{i+1}.

When the RK scheme is a MIRK scheme, the set of n equations associated with the ith subinterval has the form, [21]:

$$\phi_i = y_{i+1} - \left(y_i + h_i \sum_{r=1}^{s} b_r k_r \right), \tag{2.4}$$

where

$$k_r = f\left(t_i + c_r h_i, (1 - v_r) y_i + v_r y_{i+1} + h_i \sum_{j=1}^{r-1} x_{rj} k_j \right), \quad r = 1, 2, \ldots, s. \tag{2.5}$$

Note that each stage depends only on y_i, y_{i+1}, and previously defined stages. The method is defined by the number of stages, s, the coefficients, $\{v_r\}_{r=1}^{s}$ and $\{x_{rj}\}_{j=1,r=1}^{r-1,s}$, and the weights $\{b_r\}_{r=1}^{s}$. The abscissa, $\{c_r\}_{r=1}^{s}$, are defined by $c_r = v_r + \sum_{j=1}^{r-1} x_{rj}$.

The coefficients of a MIRK method are usually presented in a tableau of the form,

$$\begin{array}{c|c|ccccc} c_1 & v_1 & 0 & 0 & \dots & \dots & 0 \\ c_2 & v_2 & x_{21} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & & \vdots \\ c_s & v_s & x_{s1} & x_{s2} & \dots & x_{s,s-1} & 0 \\ \hline & & b_1 & b_2 & \dots & \dots & b_s \end{array},$$

which is sometimes condensed to the form

$$\begin{array}{c|c|c} c & v & X \\ \hline & & b^T \end{array},$$

where $v = (v_1, v_2, ..., v_s)^T$, $b = (b_1, b_2, ..., b_s)^T$, and X is the s by s strictly lower triangular matrix whose (i,j)th component is x_{ij}. It can be shown that the MIRK method (2.4), (2.5) is equivalent to the general RK method (2.2), (2.3), with $A = X + vb^T$ [19].

2.3.1 Order conditions for MIRK methods

A MIRK method is of order p if the numerical solution of the BVODE, obtained by solving (2.4), (2.5) together with the boundary conditions $g(y_0, y_N) = 0$ satisfies $|y(t_i) - y_i| = \mathrm{O}(h^p)$, where $y(t_i)$ is the exact solution evaluated at t_i [7]. A MIRK method of order p is derived by requiring its coefficients to satisfy a set of equations

called order conditions [7]. The order conditions for MIRK methods as presented in [28] are as follows:

MIRK methods of first order, must have

$$b^T e = 1, \tag{2.6}$$

where e is a vector of 1's of length s.

The order conditions for order 2 are

$$b^T e = 1, \quad b^T c = \frac{1}{2}. \tag{2.7}$$

The order conditions for third order are

$$b^T e = 1, \quad b^T c = \frac{1}{2}, \quad b^T c^2 = \frac{1}{3}, \quad b^T \left(Xc + \frac{v}{2} \right) = \frac{1}{6}, \tag{2.8}$$

where

$$c^l = \left[c_1^l, \quad c_2^l, \quad ..., \quad c_s^l \right]^T. \tag{2.9}$$

The order conditions for fourth order are (2.8) and

$$b^T c^3 = \frac{1}{4}, \quad b^T c \left(Xc + \frac{v}{2} \right) = \frac{1}{8}, \quad b^T \left(Xc^2 + \frac{v}{3} \right) = \frac{1}{12}, \quad b^T \left(X \left(Xc + \frac{v}{2} \right) + \frac{v}{6} \right) = \frac{1}{24}. \tag{2.10}$$

Table 2.1: Number of order conditions for MIRK methods of orders $p = 1, ..., 8$.

p	1	2	3	4	5	6	7	8
number of order conditions	1	2	4	8	17	37	85	200

It is obvious from Table 2.1 that the number of order conditions that a MIRK scheme must satisfy increases rapidly with the increasing order of the MIRK scheme.

2.4 Continuous Mono Implicit Runge-Kutta (CMIRK) Methods

After the discrete solution is obtained using a computation based on a MIRK scheme, a CMIRK scheme can be used on each subinterval to augment the discrete solution to obtain a C^1 continuous approximate solution over the whole problem domain. A CMIRK scheme applied on the subinterval $[t_i, t_{i+1}]$, is given, for $0 \leq \theta \leq 1$, by

$$u(t_i + \theta h_i) = y_i + h_i \sum_{r=1}^{s^*} b_r(\theta) k_r, \tag{2.11}$$

with the k_r's defined as in (2.5). In addition to the coefficients which define its stages, the scheme is defined by the weight polynomials, $\{b_r(\theta)\}_{r=1}^{s^*}$, which are polynomials in θ.

If the stages of the MIRK scheme can be stored and then reused by the CMIRK scheme, this makes the scheme more efficient. Thus there is an advantage to deriving CMIRK schemes with s stages identical to those of the MIRK scheme used before it. In this case the MIRK scheme is said to be "embedded" within the CMIRK scheme. In [28], optimal MIRK schemes, and optimal CMIRK schemes that have the optimal MIRK schemes embedded, are derived.

2.4.1 Continuous order conditions for CMIRK methods

A CMIRK method as defined in (2.11) is of order p if for the continuous numerical solution of the ODE, at $t = t_i + \theta h$, we have

$$\max_{0\leq\theta\leq 1} |y(t_i + \theta h) - u(t_i + \theta h)| = O(h^p), \tag{2.12}$$

where $y(t_i + \theta h)$ is the exact solution to the ODE evaluated at $t_i + \theta h_i$.

A pth order CMIRK scheme is derived by requiring its coefficients and weight polynomials to satisfy a set of continuous versions of the MIRK order conditions [11]. In addition, in order for the associated interpolant to have C^1 continuity, the weight polynomials must also satisfy certain continuity requirements [42].

For example, the continuous versions of the order conditions that are used to derive a standard fourth order CMIRK method are (compare with (2.10))

$$\begin{aligned}
&b^T(\theta)e = \theta, \quad b^T(\theta)c = \tfrac{\theta^2}{2}, \quad b^T(\theta)c^2 = \tfrac{\theta^3}{3}, \quad b^T(\theta)\left(Xc + \tfrac{v}{2}\right) = \tfrac{\theta^3}{6}, \quad b^T(\theta)c^3 = \tfrac{\theta^4}{4},\\
&b^T(\theta)c\left(Xc + \tfrac{v}{2}\right) = \tfrac{\theta^4}{8}, \quad b^T(\theta)\left(Xc^2 + \tfrac{v}{3}\right) = \tfrac{\theta^4}{12}, \quad b^T(\theta)\left(X\left(Xc + \tfrac{v}{2}\right) + \tfrac{v}{6}\right) = \tfrac{\theta^4}{24}.
\end{aligned} \tag{2.13}$$

The number of order conditions for CMIRK methods are the same as for MIRK schemes of the same order.

2.4.2 Stage order conditions for MIRK and CMIRK methods

Another set of conditions that can optionally be applied to a MIRK method or a CMIRK method are called stage order conditions. A pth order MIRK or CMIRK method is said to have stage order q ($q \leq p$) if its coefficients satisfy the stage order conditions [28]

$$Xc^{j-1} + \frac{v}{j} = \frac{c^j}{j}, \quad j = 1, \ldots, q. \tag{2.14}$$

In [7], it is proved that the maximum stage order for a pth order MIRK method is $\min\{p, 3\}$.

When a MIRK or CMIRK method has higher stage order, the number of order conditions is reduced. The number of order conditions for each order as given in Table 2.1, is made under the assumption that the stage order of the method is one; these numbers decrease rapidly with increasing stage order. For example, the number of order conditions associated with MIRK or CMIRK schemes reduces to the following number of order conditions, given in Table 2.2, when the MIRK or CMIRK schemes have stage order three.

Table 2.2: Number of order conditions for MIRK or CMIRK methods that have stage order three for orders $p = 1, ..., 8$.

p	1	2	3	4	5	6	7	8
number of order conditions	1	2	3	4	6	10	18	34

It is thus helpful to have MIRK and CMIRK methods with as high a stage order as possible, because the number of the order conditions that the method has to satisfy

is lower when it has a higher stage order. This means that the number of stages, k_r, required by the method will be lower (since the number of stages required by the method is related to the number of order conditions). See [28] for examples of MIRK methods of different orders.

2.5 Defect Control and the Maximum Defect Estimation Process

A common way to measure the quality of the continuous approximate solution of a BVODE is by computing its defect. The defect or residual, $\delta(t)$, is a continuous function over the problem interval that measures the amount by which a C^1 continuous numerical solution fails to satisfy the BVODE. The defect on the ith subinterval, $\delta_i(t)$, is computed by substituting the continuous numerical solution, $u_i(t) \equiv u(t_i + \theta h_i)$ (2.11), into the BVODE; this gives

$$\delta_i(t) = u_i'(t) - f(t, u_i(t)). \tag{2.15}$$

The strategy of a defect control solver is to adaptively choose a mesh such that, for the final accepted numerical solution, an estimate of the maximum defect over the entire problem domain is bounded by a user-provided tolerance. It is a fundamental requirement, therefore, that a defect control based solver be able to obtain an accurate and efficient estimate of the maximum defect on each subinterval. We can easily compute $\delta(t)$ at any point in the domain; however the bigger challenge is to determine, in an efficient manner, the maximum value of the defect on each subinterval.

When a standard CMIRK interpolant is employed for $u(t)$, the usual approach is to simply sample the defect at a small number of points on each subinterval with the hope that one of the points will be close enough to the location of the true maximum defect. In order for the maximum defect estimation process to be reasonably efficient, the number of points employed in estimating the defect must be kept reasonably small. While the software package BVP_SOLVER_2 employs two point sampling of the defect, there is no particular justification that either of the sampling points selected will be the equal to or even close to the location of the maximum defect. It was shown in [20] that the true maximum defect in some cases can exceed the estimated maximum defect by more than an order of magnitude. Thus, the software may accept a numerical solution for which the maximum defect is in fact substantially larger than the user-provided tolerance. As well the underestimation of the maximum defect can impact negatively on the overall performance of the computation because the mesh selection algorithm will not have access to a good profile of the defect over the subintervals of the mesh.

Recall that the continuous solution approximation on the ith subinterval, $u_i(t)$, is based on a CMIRK scheme (2.11). The continuous solution approximation on the ith subinterval, $u_i(t)$, is an approximation to the exact solution, $z_i(t)$, of the local initial value problem

$$z_i' = f(t, z_i), \qquad z_i(t_i) = y_i, \qquad t \in [t_i, t_{i+1}]. \tag{2.16}$$

For a method of order p, the continuous local error of $u_i(t)$ on the ith subinterval is [6]

$$u_i(t) - z_i(t) = O(h_i^{p+1}). \tag{2.17}$$

Similarly, the derivative of this numerical solution satisfies [6]

$$u_i'(t) - z_i'(t) = O(h_i^p), \tag{2.18}$$

since the variables t and θ are related by the equations $t = t_i + \theta h \Rightarrow \theta = \frac{1}{h}(t - t_i)$ and $\frac{d\theta}{dt} = \frac{1}{h}$. Hence the right hand side of (2.18) is reduced by a factor of h.

Recall that the defect of the numerical solution, $u_i(t)$, on the ith subinterval has the form

$$\delta_i(t) = u_i'(t) - f(t, u_i(t)). \tag{2.19}$$

Taking advantage of the fact that $z_i(t)$ is the exact solution of (2.16), (2.19) can be written as

$$\delta_i(t) = u_i'(t) - f(t, u_i(t)) + f(t, z_i(t)) - z_i'(t). \tag{2.20}$$

A slight rearranging of (2.20) gives

$$\delta_i(t) = u_i'(t) - z_i'(t) - (f(t, u_i(t)) - f(t, z_i(t))). \tag{2.21}$$

Assuming a Lipschitz condition [2] on f, the second term in (2.21) then can be seen to be of $O(h_i^{p+1})$ (see(2.17)) and the defect can thus be written as

$$\delta_i(t) = u_i'(t) - z_i'(t) + O(h_i^{p+1}). \tag{2.22}$$

The leading term in the defect (2.22) is thus $O(h_i^p)$ from (2.18). Furthermore the leading order term in the defect can be seen to be equal to the leading order term in the error for $u_i'(t)$.

When $u_i(t)$ is based on a CMIRK scheme, the leading error term is known from the theory of Runge-Kutta methods [5]. On the ith subinterval the defect can be expressed in an expansion that is related to the local error expansion of the approximation solution. It has the form,

$$\delta_i(t) = \left(\sum_{j=0}^{\rho} q_j(\theta) F_j \right) h_i^p + O(h_i^{p+1}), \tag{2.23}$$

where p is the order of the Runge-Kutta scheme, the $q_j(\theta)$'s are polynomials of degree p that are the derivatives of the unsatisfied continuous order conditions for order $p+1$. These depend on the CMIRK but are independent of the problem or h_i. The F_j's are elementary differentials [6] which depend only on the problem and $\rho+1$ is the number of elementary differentials of $(p+1)st$ order. As $h_i \to 0$, it is evident from (2.23) that the value of the defect will approach a linear combination of the $q_j(\theta)$ values, where the coefficients of this linear combination are the elementary differentials, F_j. Therefore, the location of the maximum will vary from subinterval to subinterval and from problem to problem. This means that on any given subinterval, we cannot make an *a priori* determination of the location of the maximum value of the leading term

of the defect.

For example, consider the SWAVE problem (1.1) using a standard fourth order CMIRK scheme. In order to observe how the defect behaves on each subinterval, we plot the absolute scaled defect of each subinterval mapped on to [0,1] on the same graph. We obtain the scaled defect by dividing the defect, which is computed across each subinterval at many points, by the maximum defect on that subinterval. This ensures that the maximum scaled defect, on each subinterval is 1. See Figure 2.1.

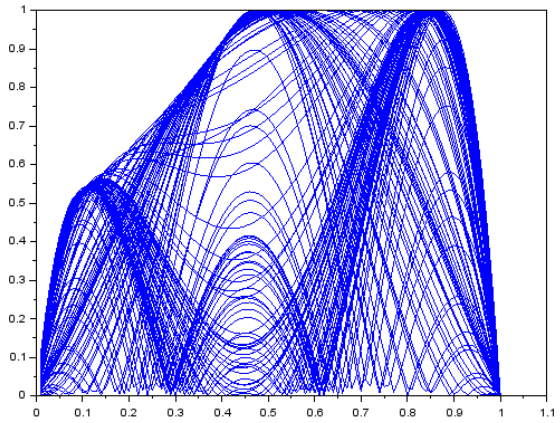

Figure 2.1: A plot of the absolute scaled defect obtained by applying a standard 4^{th} order CMIRK scheme (3.9) with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$. The location of the maximum defect varies from subinterval to subinterval.

2.6 Hermite - Birkhoff Interpolants Derived Via a Bootstrapping Process

In the previous section we saw that the standard approach for obtaining an accurate estimate of the maximum defect using a standard CMIRK scheme is either not efficient (if we sample the defect at a large number of points on each subinterval)

or (usually) not accurate (if we sample the defect at a small number of points on each subinterval). In [20], the authors describe one approach in which an interpolant with a greatly simplified expression for the leading order term in the defect is derived. Starting with a standard CMIRK scheme, they employ a boot-strapping algorithm developed in [18] to derive a special type of interpolant expressed in a Hermite-Birkhoff form [18]. This special interpolant yields a defect for which the location of the maximum defect on each subinterval can be determined (at least asymptotically) in an *a priori* manner. The maximum defect of each subinterval has the same location for different subintervals and problems; see Figure 2.2 for an example. The estimate of the maximum defect obtained in this case is said to be asymptotically correct. A numerical scheme that uses defect control based on an asymptotically correct estimate of the maximum defect is known as an Asymptotically Correct Defect Control (ACDC) scheme. For this case, the leading order term in the defect expansion is a multiple of a single polynomial in θ.

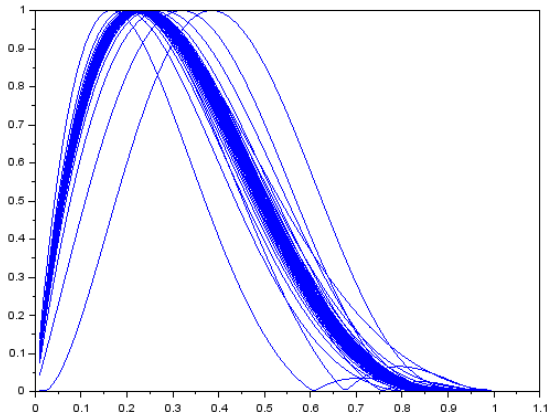

Figure 2.2: A plot of the absolute scaled defect obtained by applying a 4^{th} order Hermite-Birkhoff (2.24) scheme with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$. The absolute scaled defect has the same shape on almost every subinterval, and thus the location of the maximum defect is the same on almost every subinterval.

In [14], Ellis derived boot-strap interpolants which provide an asymptotically correct estimate of the maximum defect of the continuous numerical solution.

The *asymptotically correct* quality possessed by the Hermite-Birkhoff interpolants is a consequence of the simplification of the expression in (2.23). Based on earlier work in [16], Enright and Muir [20] derived a sixth order Hermite-Birkhoff interpolant that gives an asymptotically correct maximum defect estimate property using the bootstrapping approach. This led to an interpolant for which the highest order term in the defect expansion is a multiple of a single polynomial in θ. This implies that one point sampling for estimating the maximum defect is then possible.

The general form of a Hermite-Birkhoff scheme on the subinterval $[t_i, t_{i+1}]$, with $0 \leq \theta \leq 1$ is:

$$\tilde{u}_i(t_i + \theta h_i) \quad = \quad d_0(\theta) y_i + d_1(\theta) y_{i+1} + h_i \sum_{r=1}^{\tilde{s}^*} \tilde{b}_r(\theta) k_r, \tag{2.24}$$

where the k_r's have the same general form as (2.5), $d_0(\theta)$, $d_1(\theta)$, $\{\tilde{b}_r(\theta)\}_{r=1}^{\tilde{s}^*}$, are polynomials in θ, and $\tilde{s}^*$ is the total number of required stages. The determination of the required stages and weight polynomials as described in [20], is done by requiring the interpolant (2.24) and its derivative to satisfy certain interpolation conditions at a number of points within the i^{th} subinterval. This process is detailed in sections 3.3.1, 3.3.2 and 6.2 of this thesis, during the derivation of fourth, fifth and sixth order Hermite-Birkhoff schemes. The leading order term in the defect expansion, using a Hermite-Birkhoff scheme, is a multiple of a single polynomial in θ, namely $d_1'(\theta)$ [20].

It is a relatively straightforward process to convert the Hermite-Birkhoff form of $\tilde{u}(t)$ to its CMIRK equivalent. By substituting for $\underline{y}_{i+1}$ using the discrete formula

$$\underline{y}_{i+1} \quad = \quad \underline{y}_i + h_i \sum_{r=1}^{s} b_r \underline{k}_r, \tag{2.25}$$

in (2.24) and noting the interpolation condition $d_0(\theta) + d_1(\theta) = 1$, the CMIRK form of $\underline{\tilde{u}}(t)$ (2.24) can be obtained:

$$\underline{\tilde{u}}_i(t_i + \theta h_i) \;=\; \underline{y}_i + h_i \sum_{r=1}^{\tilde{s}^*} (b_r d_1(\theta) + \tilde{b}_r(\theta))\underline{k}_r. \tag{2.26}$$

However, it is pointed out in [20] that the lack of an explicit dependence on y_{i+1} in (2.26) means that $\underline{\tilde{u}}(t)$ may have discontinuities that are of the size of the Newton tolerance used to determine the $\{\underline{y}_i\}_{i=0}^N$ values. (The reason for this discontinuity is that (2.25) is not solved exactly, only to within the Newton tolerance; see [20] for further details.) The above substitution introduces an additional error of $O(h^{p+1})$ associated with the error for y_{i+1} from the discrete formula. On the other hand, since $\tilde{u}(t)$ in (2.24) has an explicit dependence on $k_1 = f(t_i, y_i)$ and $k_2 = f(t_{i+1}, y_{i+1})$, the interpolant and its first derivative will be continuous across each internal mesh point.

2.7 Validity Check

It is important to monitor the accuracy and robustness of the one point defect sampling process by checking the value of the defect estimate at an additional point known as a validity check sampling point. This sample point is a point where the value of the defect should be half the value of the maximum defect. Thus, the defect of the interpolant $\tilde{u}_i(t)$ is also computed at a second predetermined spot within each subinterval. The auxiliary validity check process was discussed in [20], where it is observed that the successful defect estimation rate of the sixth order Hermite-Birkhoff for the final converged mesh was around 83% for a collection of test problems. Closer examination revealed that the subintervals where the estimation failed were

relatively large and thus the associated computation wasn't within the asymptotic regime for the formula. Hence the error contribution from the higher order terms was significant enough to interfere with the dominance of the leading order term in the defect expansion. The validity check provides an additional layer of confidence for the defect sampling and control process.

Chapter 3

Implementations of Standard CMIRK Schemes For Orders 1-5 and Fourth and Fifth Order Hermite-Birkhoff Interpolants

In this chapter, we will implement some standard CMIRK schemes for orders 1-5 and apply them to several test problems. For orders 1-3, we will see that there are standard CMIRK schemes that naturally lead to ACDC schemes. For orders 4 and 5, we will see that the standard CMIRK schemes do not lead to ACDC schemes. Therefore, for orders 4 and 5 we will consider the use of Hermite-Birkhoff interpolants in order to obtain ACDC schemes.

3.1 Implementations of Standard CMIRK Schemes For Orders 1-5

Using Scilab, a powerful numerical computing environment for engineering and scientific applications [45] we implemented standard MIRK schemes, (2.4), (2.5), of orders 1 to 5 and applied them to several test problems in order to obtain a discrete numerical solution, $\{y_i\}_{i=0}^{N}$, on a mesh of points that partition the problem domain into subintervals. The discrete numerical solution is obtained by applying MIRK methods, to get a set of nonlinear equations that are solved using a Newton iteration

(using the Scilab fsolve function) with a default tolerance 10^{-10}. We then augment the discrete solution with a continuous numerical solution, $u_i(t)$, using a standard CMIRK scheme, (2.11), of the same order as the MIRK scheme, over each subinterval. In order to assess the quality of the continuous numerical solution, we then evaluate the defect, $\delta(t)$, using (2.15), and plot the absolute scaled defect for each subinterval mapped onto [0,1].

We sampled the defect at a hundred points within each subinterval of a uniform mesh of a hundred subintervals, i.e., N=100, h=0.01, where N is the number of subintervals, h is the subinterval size, and the problem interval, [a,b], is [0,1]. We also found the maximum value of the defect samples on each subinterval, and then plotted the absolute scaled defect mapped onto [0,1] for each subinterval. The scaling involved dividing each defect sample by the maximum defect, which implies that the absolute scaled defect values will be in the range [0,1]. This process requires substantial computing time and thus would not be a practical way of computing the maximum defect on each subinterval, but it allows us to see what shape the absolute scaled defect has on each subinterval, thus allowing us to determine if the scheme has the ACDC property. We can also determine the location of the maximum defect on each subinterval from this plot.

3.1.1 Test problems

We will consider numerical experiments on the following BVODEs which have been converted to systems of first order BVODE systems where necessary:

1. A linear BVODE [2]:

$$Y'(x) = AY(x) + Q(x),$$

with boundary conditions:

$$Y \begin{pmatrix} 0 \\ \\ 1 \end{pmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix},$$

where

$$A = \begin{bmatrix} 0 & \lambda \\ \\ \lambda & 0 \end{bmatrix},$$

and

$$Q(x) = \begin{bmatrix} 0 \\ \\ \lambda cos^2(\Pi x) + \frac{2}{\lambda}\Pi^2 cos(2\Pi x) \end{bmatrix}. \tag{3.1}$$

2. A simple nonlinear BVODE [40]:

$$w''(x) = \frac{3}{2}w^2(x),$$

with boundary conditions

$$w(0) = 4, w(1) = 1.$$

(3.2)

3. SWAVE problem: (1.1)

4. SWIRL-III problem: (1.2)

3.2 Numerical experiments

3.2.1 Standard first order CMIRK scheme

We consider the standard two stage, first order, stage order one, CMIRK scheme (CMIRK211), which has the tableau,

$$\begin{array}{c|c|cc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline & & b_1(\theta) & b_2(\theta) \end{array} \quad , \tag{3.3}$$

where $b_1(\theta) = -\theta(\theta^2 - \theta - 1)$ and $b_2(\theta) = \theta^2(\theta - 1)$, and which has the discrete one-stage, first order, stage order one MIRK scheme (Euler's method) (MIRK111) embeded within it; this discrete scheme has the tableau

$$\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline & & 1 \end{array} \quad . \tag{3.4}$$

We applied the above MIRK, CMIRK pair to the SWAVE problem (1.1), and obtained a plot of the absolute scaled defect for each subinterval (mapped onto [0,1]) shown in Figure 3.1.

It is clear that this CMIRK scheme has the ACDC property, since almost all the absolute scaled defect plots are the same. This happens because the leading order term in the defect expansion is a multiple of a single polynomial in θ. And this happens because any first order CMIRK scheme has only one unsatisfied order condition, namely $\left(b^T(\theta)c - \frac{\theta^2}{2}\right)$, associated with second order, appearing in the leading order term of the local error expansion. Then the leading order term in the expansion of the defect is a multiple of the derivative of this polynomial; we plot it in Figure 3.2. From Figure 3.1, we observe that the maximum defect occurs at $\theta \approx 0.5$ for each subinterval. We can explicitly verify that this is correct. The unsatisfied second order condition is $b^T(\theta)c - \frac{\theta^2}{2} = \theta^2(\theta - 1) - \frac{\theta^2}{2} = \theta^3 - \frac{3}{2}\theta^2$ for the specific CMIRK scheme we are considering. Its derivative is $3\theta^2 - 3\theta$. The maximum value of this polynomial will occur where its derivative is zero, i.e., when $6\theta - 3 = 0 \Rightarrow \theta = \frac{1}{2}$; see Figure 3.2.

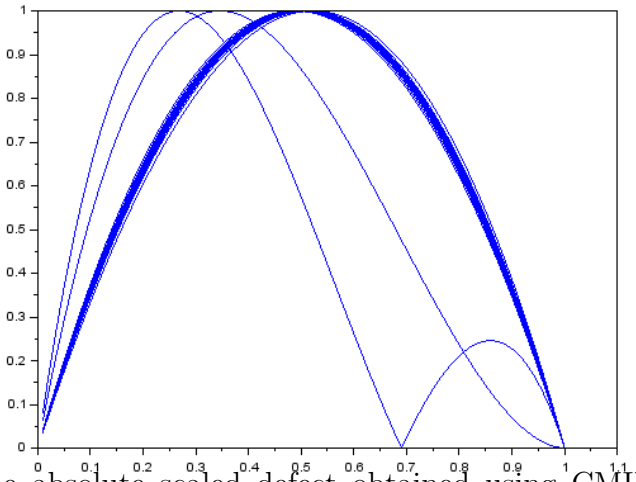

Figure 3.1: A plot of the absolute scaled defect obtained using CMIRK211 with N=100 on the test problem (1.1) with $\epsilon = 0.1$.

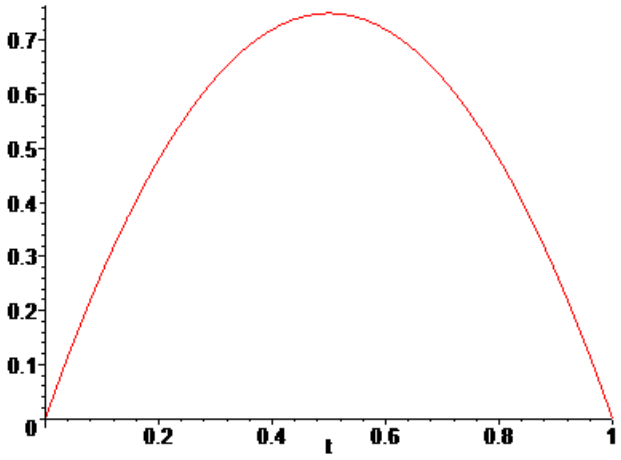

Figure 3.2: A plot of the absolute derivative of the unsatisfied second order condition, $b^T(\theta)c - \frac{\theta^2}{2}$.

We see from Figure 3.1 that the scaled defect has the same shape and in particular the same location for the maximum defect over almost all subintervals. There are a small number of subintervals where the shape of the absolute scaled defect is different. We consider this later in this chapter.

3.2.2 Standard second order CMIRK scheme

We consider the standard two stage, second order, stage order two CMIRK scheme (CMIRK222) taken from [28], (also known as the continuous trapezoidal scheme). It has the tableau,

$$\begin{array}{c|c|cc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline & & b_1(\theta) & b_2(\theta) \end{array}, \tag{3.5}$$

where $b_1(\theta) = -\frac{1}{2}\theta(\theta-2)$ and $b_2(\theta) = \frac{1}{2}\theta^2$, and it has the MIRK222 (trapezoidal) scheme embedded within it; the MIRK222 scheme has this tableau

$$\begin{array}{c|c|cc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline & & \frac{1}{2} & \frac{1}{2} \end{array} \quad . \tag{3.6}$$

We apply the MIRK222/CMIRK222 pair to the SWAVE problem (1.1), the linear problem (3.1), the SWIRL-III problem (1.2) and the simple nonlinear problem (3.2), and plot the absolute scaled defect for each subinterval (mapped onto [0,1]); see Figures 3.3, 3.5, 3.6 and 3.7. We observe that the absolute scaled defect has the same shape on almost every subinterval and for all four problems. This is because the leading order term in the defect expansion is a multiple of a single polynomial in θ. This happens because any second order CMIRK scheme that satisfies the stage order two conditions has only one unsatisfied order condition, $\left(b^T(\theta)c^2 - \frac{\theta^3}{3}\right)$ (of order 3), appearing in the leading order term of the local error expansion. Then the leading order term in the expansion of the defect will be a multiple of the derivative of this polynomial. From Figures 3.3, 3.5, 3.6 and 3.7, we see that the maximum defect using this scheme is located at $\theta \approx 0.5$ for each subinterval and all of the problems.

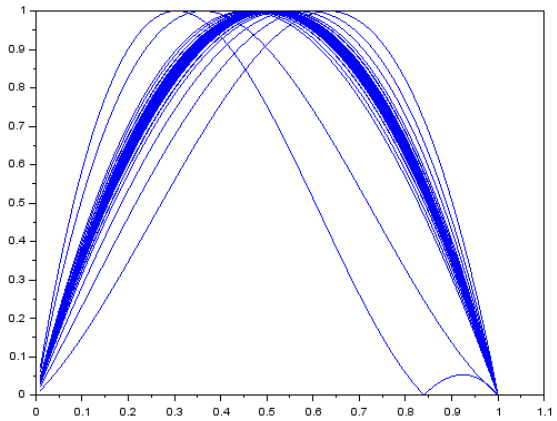

Figure 3.3: A plot of the absolute scaled defect obtained from applying CMIRK222, with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$.

We can predict this location of the maximum defect and the shape of the defect on each subinterval. As mentioned previously, the derivative of the unsatisfied order condition $\left(b^T(\theta)c^2 - \frac{\theta^3}{3}\right)$ will appear in the leading order term in the expansion of the defect. For the specific CMIRK scheme we are considering, this polynomial is $\frac{d}{d\theta}(\frac{\theta^2}{2} - \frac{\theta^3}{3}) = \theta - \theta^2$. Its maximum will occur where its derivative is zero, i.e., where $1 - 2\theta = 0 \Rightarrow \theta = \frac{1}{2}$; see Figure 3.4.

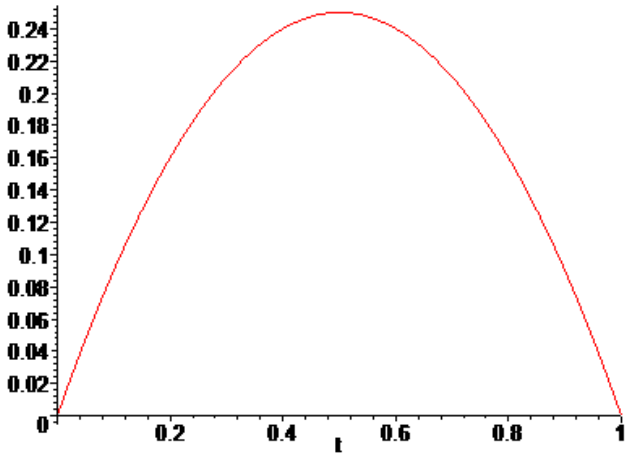

Figure 3.4: A plot of the absolute derivative of the unsatisfied order condition for order 3, $b^T(\theta)c^2 - \frac{\theta^3}{3}$.

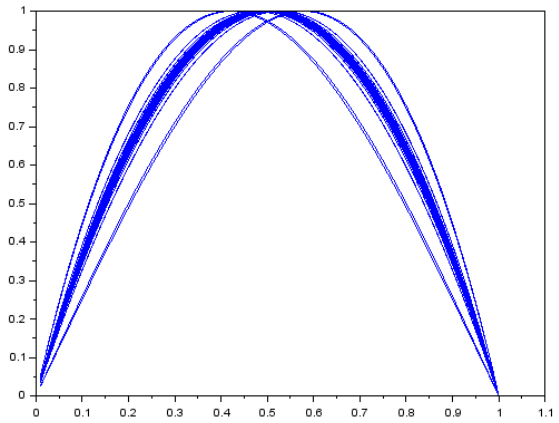

Figure 3.5: A plot of the absolute scaled defect obtained from applying CMIRK222 with N=100 to the test linear problem (3.1) with $\lambda = 1$.

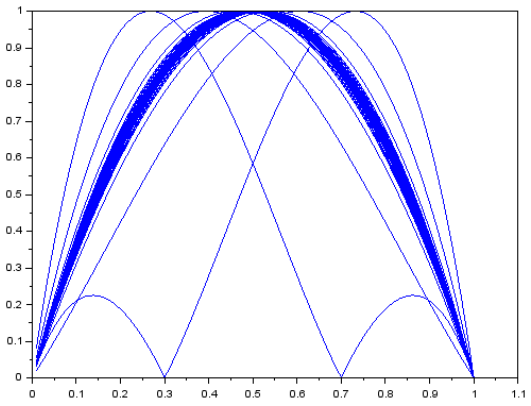

Figure 3.6: A plot of the absolute scaled defect obtained from applying CMIRK222 with N=100 to the test problem SWIRL_III (1.2) with $\epsilon = 0.01$.

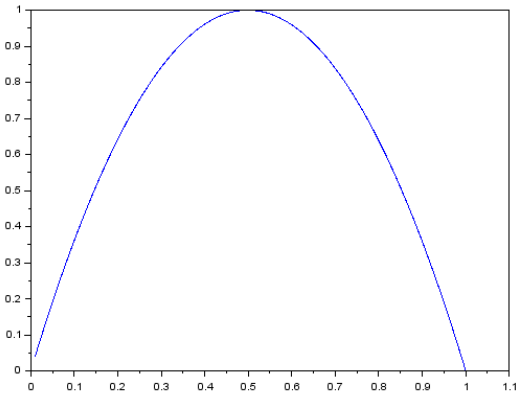

Figure 3.7: A plot of the absolute scaled defect obtained from applying CMIRK222 with N=100 to the simple nonlinear test problem (3.2).

From the above figures, we observe that the location of the maximum defect is the same over all subintervals and problems except for a few cases. By plotting the non-scaled defect over the whole problem domain, [0,1], using CMIRK222 applied to the SWIRL-III problem (1.2), (see Figure 3.8), and displaying the maximum defect of each subinterval, we investigated the subintervals that have defects with different shapes from the usual case. In particular, we found that the $19^{th}, 49^{th}, 50^{th}, 51^{th}, 52^{th}$ and 82^{th} subintervals have different shapes from others. However, we also found that for these subintervals the maximum defect is much smaller than the maximum defect of the other 94 subintervals. That is, the subintervals where the defect has a different shape are also subintervals where the defect is much smaller than the overall maximum defect, and therefore it is less important to obtain an accurate estimate of the maximum defect on these subintervals.

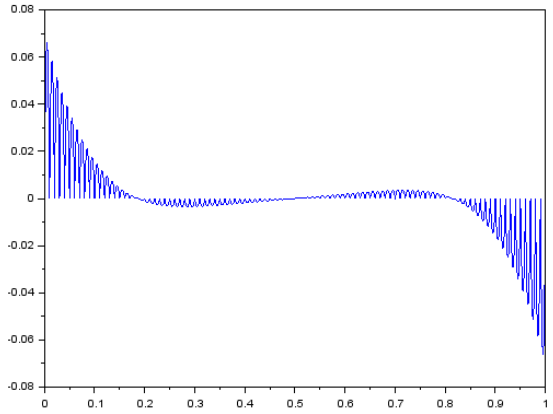

Figure 3.8: A plot of the absolute non-scaled defect obtained from applying CMIRK222 with N=100 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$. The subintervals where the defect has a different shape are also cases where the maximum defect is much smaller.

3.2.3 Standard third order CMIRK scheme

We consider a three stage, second order, stage order three CMIRK scheme (CMIRK333), which has the tableau,

$$\begin{array}{c|c|ccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{7}{27} & \frac{4}{27} & \frac{-2}{27} & 0 \\ \hline & & b_1(\theta) & b_2(\theta) & b_3(\theta) \end{array}, \tag{3.7}$$

where $b_1(\theta) = \theta(\theta-1)^2$, $b_2(\theta) = \frac{1}{4}\theta^2(-1+2\theta)$ and $b_3(\theta) = \frac{-3}{4}\theta^2(-3+2\theta)$. This scheme has the MIRK333 scheme embedded within it; the MIRK333 scheme has the tableau,

$$\begin{array}{c|c|ccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{7}{27} & \frac{4}{27} & \frac{-2}{27} & 0 \\ \hline & & 0 & \frac{1}{4} & \frac{3}{4} \end{array}. \tag{3.8}$$

We apply the MIRK333/CMIRK333 pair to the SWAVE problem (1.1). We plot the absolute scaled defect in Figure 3.10. We observe that the absolute scaled defect has the same shape on all subintervals. The maximum defect using this scheme appears to occur at $\theta \approx 0.74$ for each subinterval. This is because the leading order term in the defect expansion is a multiple of a single polynomial in θ. This happens because any 3^{rd} order CMIRK scheme that has stage order three has only one unsatisfied order condition $\left(b^T(\theta)c^3 - \frac{\theta^4}{4}\right)$ (associated with 4^{th} order) which appears in

the leading order term of the local error expansion. We can predict the shape of the absolute scaled defect on each subinterval since it will be a multiple of the derivative of $\left(b^T(\theta)c^3 - \frac{\theta^4}{4}\right)$, i.e., $\frac{d}{d\theta}\left(\frac{\theta^2}{4}(-1+2\theta) - \frac{1}{27}\left(\frac{3}{4}\theta^2(-3+2\theta)\right) - \frac{\theta^4}{4}\right) = -\frac{1}{3}\theta + \frac{4}{3}\theta^2 - \theta^3$. Its maximum will occur where its derivative, $-\frac{1}{3} + \frac{8}{3}\theta - 3\theta^2$, is zero $\Rightarrow \theta = 0.7384168123$; see Figure 3.9.

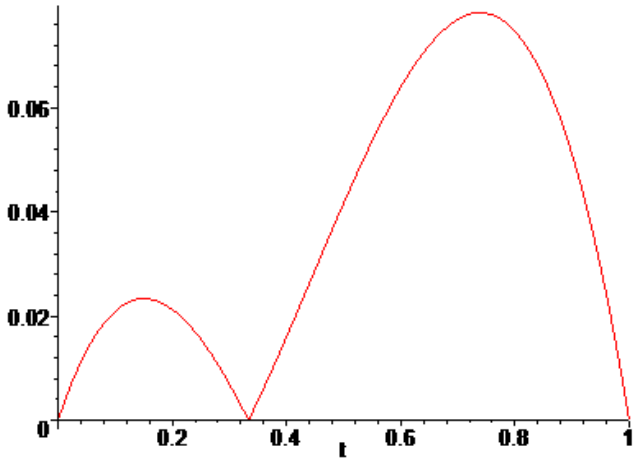

Figure 3.9: A plot of the absolute derivative of the unsatisfied order condition for order 4.

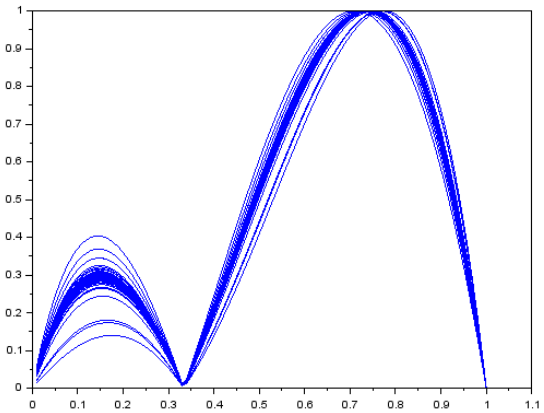

Figure 3.10: A plot of the absolute scaled defect obtained from applying CMIRK333 with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$.

3.2.4 Standard fourth order CMIRK scheme

Recall that the maximum stage order of a MIRK or CMIRK scheme is three. We consider a four stage, 4^{th} order, stage order three CMIRK scheme (CMIRK443) taken from [28], which has the tableau,

$$\begin{array}{c|c|cccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{8} & -\frac{1}{8} & 0 & 0 \\ \frac{2}{5} & \frac{2}{5} & \frac{17}{125} & -\frac{13}{125} & -\frac{4}{125} & 0 \\ \hline & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) \end{array}, \tag{3.9}$$

where

$$b_1(\theta) = -\frac{1}{12}\theta(3\theta-4)(5\theta^2-6\theta+3), \quad b_2(\theta) = \frac{1}{6}\theta^2(5\theta^2-6\theta+2),$$

$$b_3(\theta) = -\frac{2}{3}\theta^2(3\theta-2)(5\theta-6), \quad \text{and} \quad b_4(\theta) = \frac{125}{12}\theta^2(\theta-1)^2.$$

This scheme has embedded within it the discrete MIRK343 scheme [28], which has the tableau

$$\begin{array}{c|c|ccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{8} & -\frac{1}{8} & 0 \\ \hline & & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}. \tag{3.10}$$

We apply the MIRK343/CMIRK443 pair to the SWAVE problem (1.1) and the SWIRL-III problem (1.2). The plots of the absolute scaled defect are given in Figures 3.11 and 3.12. We note that the location of the maximum defect changes from subinterval to subinterval and that the scaled defect does not have the same shape on each subinterval. The location of the maximum defect varies from subinterval to subinterval and over the two problems. This happens because the leading order term in the defect expansion depends on a varying linear combination of two polynomials. This happens because any fourth order CMIRK scheme that satisfies stage order three has a leading order term in its local error expansion that depends on the two unsatisfied order conditions $\left(b^T(\theta)c^4 - \frac{\theta^5}{5}\right)$ and $\left(b^T(\theta)(Xc^3 + \frac{v}{4}) - \frac{\theta^5}{20}\right)$ associated with 5^{th} order. This results in a leading order term in the defect expansion of the form (2.23) with $\rho = 1$ that depends on a varying linear combination of two different polynomials in θ. Thus, it is impossible to determine a priori where the maximum defect will occur for each subinterval.

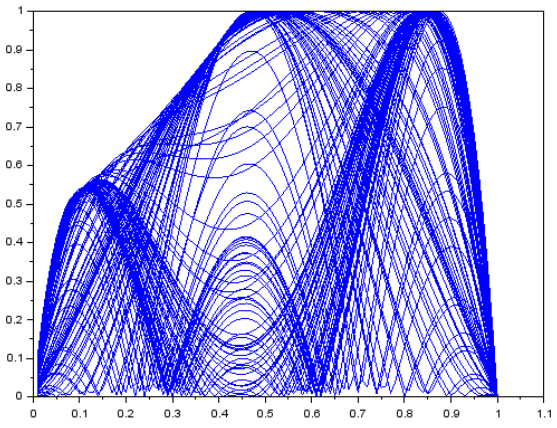

Figure 3.11: A plot of the absolute scaled defect obtained from applying CMIRK443 with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$.

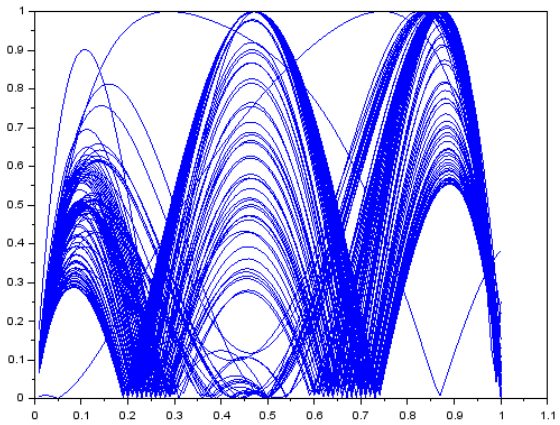

Figure 3.12: A plot of the absolute scaled defect obtained from applying CMIRK443 with N=100 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$.

However, when applying the MIRK343/CMIRK443 pair to the linear problem (3.1), or the simple nonlinear problem (3.2), we found that the absolute scaled defect plots - see Figures 3.13 and 3.14- were the same on all subintervals. This happens because the leading order term in the defect expansion is a multiple of a single polynomial in θ due to the simplicity of these two problems. However, a software package that implements defect control cannot detect if the problem is going to be sufficiently simple that the expansion of the local error will be a multiple of a single polynomial.

For each of these simple test problems, we examine the unsatisfied order conditions for 5^{th} order to see if the shape of the absolute scaled defects we observe in Figures 3.13 and 3.14 matches the shape of derivatives of either of these unsatisfied order conditions. The derivative of the first order condition $\left(b^T(\theta)c^4 - \frac{\theta^5}{5}\right)$, i.e., $\frac{d}{d\theta}\left(b^T(\theta)c^4 - \frac{\theta^5}{5}\right) = \frac{1}{5}\theta - \frac{11}{10}\theta^2 + \frac{19}{10}\theta^3 - \theta^4$. Its maximum will occur at $\frac{d}{d\theta}(\frac{1}{5}\theta - \frac{11}{10}\theta^2 +$

$\frac{19}{10}\theta^3 - \theta^4) = \frac{1}{5} - \frac{11}{5}\theta + \frac{57}{10}\theta^2 - 4\theta^3 = 0 \Rightarrow \theta = 0.8428127370$; see Figure 3.15 where we see that the plot of this polynomial has almost the same shape as the absolute scaled defect shown in Figure 3.13. The derivative of the other unsatisfied order condition $\left(b^T(\theta)(Xc^3 + \frac{v}{4}) - \frac{\theta^5}{20}\right)$, i.e., $\frac{d}{d\theta}\left(b^T(\theta)(Xc^3 + \frac{v}{4}) - \frac{\theta^5}{20}\right) = -\frac{1}{4}\theta^2 + \frac{1}{2}\theta^3 - \frac{1}{4}\theta^4$. Its maximum will occur at $\frac{d}{d\theta}(-\frac{1}{4}\theta^2 + \frac{1}{2}\theta^3 - \frac{1}{4}\theta^4) = -\frac{1}{2}\theta + \frac{3}{2}\theta^2 - \theta^3 = 0 \Rightarrow \theta = 0.5$; see Figure 3.16 where we see that the plot of this polynomial has a similar shape to the absolute scaled defect shown in Figure 3.14.

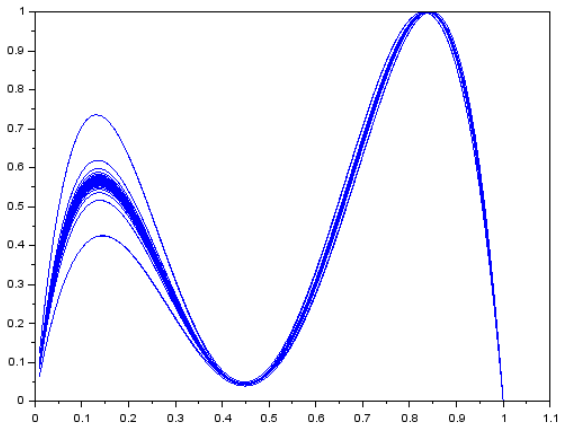

Figure 3.13: A plot of the absolute scaled defect obtained from applying CMIRK443 with N=100 to the test linear problem (3.1) with $\lambda = 1$.

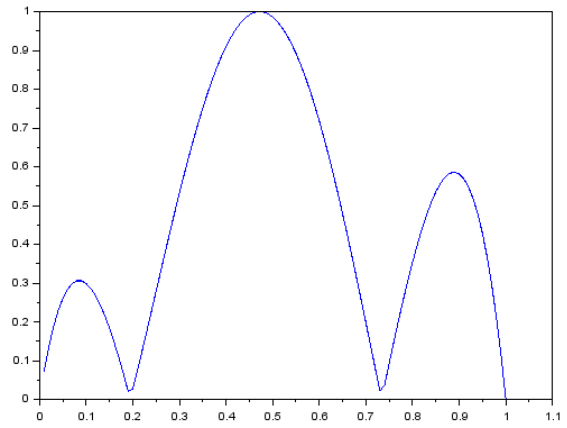

Figure 3.14: A plot of the absolute scaled defect obtained from applying CMIRK443 with N=100 to the simple nonlinear test problem (3.2).

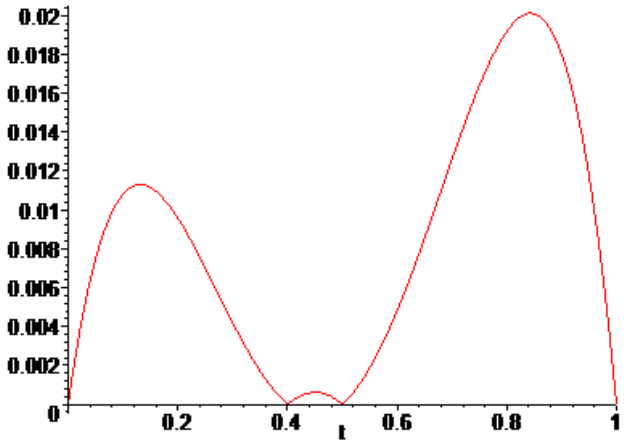

Figure 3.15: A plot of the absolute first unsatisfied order condition for order 5.

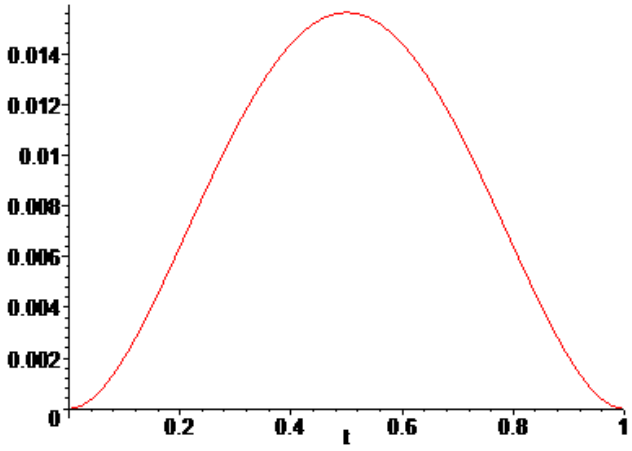

Figure 3.16: A plot of the absolute second unsatisfied order condition for order 5.

3.2.5 Standard fifth order CMIRK scheme

Recall that the maximum stage order of a fifth order CMIRK method is three. We consider a six stage, 5^{th} order, stage order three, CMIRK scheme (CMIRK653) taken from [28], which has the tableau,

$$\begin{array}{c|c|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ c_3 & v_3 & x_{31} & x_{32} & 0 & 0 & 0 & 0 \\ c_4 & v_4 & x_{41} & x_{42} & 0 & 0 & 0 & 0 \\ c_5 & v_5 & x_{51} & x_{52} & x_{53} & x_{54} & 0 & 0 \\ \frac{14}{25} & \frac{14}{25} & x_{61} & x_{62} & x_{63} & x_{64} & x_{65} & 0 \\ \hline & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) \end{array}, \tag{3.11}$$

where the coefficients for the first five rows of the tableau are given in the next tableau of the discrete five stage, 5^{th} order MIRK scheme (MIRK553) (3.12), and

$$x_{61} = -\frac{28017913\sqrt{393}}{4515625000} - \frac{493827103}{22578125000}, \quad x_{62} = \frac{37817373}{2187500000} - \frac{745481\sqrt{393}}{19687500000},$$

$$x_{63} = \frac{13007794215933}{92082812500000} + \frac{686727625023\sqrt{393}}{92082812500000},$$

$$x_{64} = -\frac{2408972902336}{9694346953125} - \frac{2652451648\sqrt{393}}{5816608171875},$$

$$x_{65} = \frac{4506347288003}{40301953125000} - \frac{10198807509\sqrt{393}}{13433984375000},$$

$$b_1(\theta) = \frac{(42919\sqrt{393} + 726581)}{279890843022336}\theta(1853738880\,\theta^4 - 3898264641\,\theta^3$$

$$-67451925\sqrt{393}\theta^3 + 1702187994\,\theta^2 + 176781154\sqrt{393}\theta^2$$

$$+1127678925\,\theta - 160912047\sqrt{393}\theta - 1037557668 + 61288332\sqrt{393}),$$

$$b_2(\theta) = -\frac{(-19675275 + 559079\sqrt{393})}{52326786664728576}\theta^2(4250789760\,\theta^3 - 5448163857\,\theta^2$$

$$+177236475\sqrt{393}\theta^2 + 1009672582\,\theta - 284790018\sqrt{393}\theta$$

$$+702696666 + 128060898\sqrt{393}),$$

$$b_3(\theta) = -\frac{(11675241621 + 610860615\sqrt{393})}{23449186109440}\theta^2(13440\,\theta^3 - 35583\,\theta^2$$

$$-75\sqrt{393}\theta^2 + 156\sqrt{393}\theta + 32044\,\theta - 10500 - 84\sqrt{393}),$$

$$b_4(\theta) = \frac{(5288000\sqrt{393} - 38154000)}{364664379807}\theta^2(13440\,\theta^3 - 35583\,\theta^2 - 75\sqrt{393}\theta^2$$

$$+156\sqrt{393}\theta + 32044\,\theta - 10500 - 84\sqrt{393}),$$

$$b_5(\theta) = -\frac{(948114929207 + 39148987645\sqrt{393})}{169641265393978560}\theta^2(2297220\,\theta^3 + 294525\sqrt{393}\theta^2$$

$$-10931079\,\theta^2 - 612612\sqrt{393}\theta + 15563192\,\theta - 7225680 + 329868\sqrt{393}),$$

$$b_6(\theta) = \frac{(1944296875 + 59765625\sqrt{393})}{119925757952}(\theta - 1)^2\theta^2(896\,\theta - 1241 + 51\sqrt{393}).$$

This scheme has embedded within it the discrete MIRK553 scheme which has the tableau

$$\begin{array}{c|c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -\frac{5}{28}+\frac{\sqrt{393}}{84} & v_3 & x_{31} & x_{32} & 0 & 0 & 0 \\ \frac{17}{20} & v_4 & x_{41} & x_{42} & 0 & 0 & 0 \\ \frac{125}{224}+\frac{\sqrt{393}}{224} & v_5 & x_{51} & x_{52} & x_{53} & x_{54} & 0 \\ \hline & & b_1 & b_2 & b_3 & b_4 & b_5 \end{array}, \tag{3.12}$$

where

$$v_3 = \frac{229}{686} - \frac{101\sqrt{393}}{6174}, \quad x_{31} = -\frac{6409}{16464} + \frac{1097\sqrt{393}}{49392},$$

$$x_{32} = -\frac{2027}{16464} + \frac{299\sqrt{393}}{49392}, \quad v_4 = \frac{3757}{4000}, \quad x_{41} = \frac{153}{8000}, \quad x_{42} = -\frac{867}{8000},$$

$$v_5 = \frac{237704435}{314703872} + \frac{1823343\sqrt{393}}{314703872}, \quad x_{51} = \frac{223029279}{10699931648} - \frac{36659445\sqrt{393}}{10699931648},$$

$$x_{52} = -\frac{1758793}{629407744} - \frac{682877\sqrt{393}}{629407744}, \quad x_{53} = \frac{164181897}{3862233088} + \frac{9504189\sqrt{393}}{3862233088},$$

$$x_{54} = -\frac{725872015625}{2815085142016} + \frac{2028935875\sqrt{393}}{2815085142016},$$

$$b_1 = -\frac{19}{272} - \frac{11\sqrt{393}}{816}, \quad b_2 = \frac{3035}{28224} + \frac{187\sqrt{393}}{84672},$$

$$b_3 = \frac{43425027}{132009920} + \frac{2257089\sqrt{393}}{132009920}, \quad b_4 = \frac{998000}{21897819} - \frac{550000\sqrt{393}}{65693457},$$

$$b_5 = \frac{5993083}{10195920} + \frac{25961\sqrt{393}}{10195920}.$$

We apply the MIRK553/CMIRK653 pair to the SWAVE problem (1.1) and the SWIRL-III problem (1.2). The plot of the absolute scaled defect is given in Figures

3.17 and 3.18. The location of the maximum defect varies from subinterval to subinterval and from problem to problem. This is because any fifth order CMIRK scheme that satisfies stage order three has four unsatisfied order conditions appearing in a varying linear combination in the leading order term of the local error expansion. These polynomials are $(b^T(\theta)c^5 - \frac{\theta^6}{6})$, $\left(b^T(\theta)(Xc^4 + \frac{v}{5}) - \frac{\theta^6}{30}\right)$, $\left(b^T(\theta)c(Xc^3 + \frac{v}{4}) - \frac{\theta^6}{24}\right)$ and $\left(b^T(\theta)(X(Xc^3 + \frac{v}{4}) + \frac{v}{20}) - \frac{\theta^6}{120}\right)$. This implies that the leading order term of the defect expansion (2.23) will be a varying linear combination of four polynomials (the derivatives of the above unsatisfied order conditions).

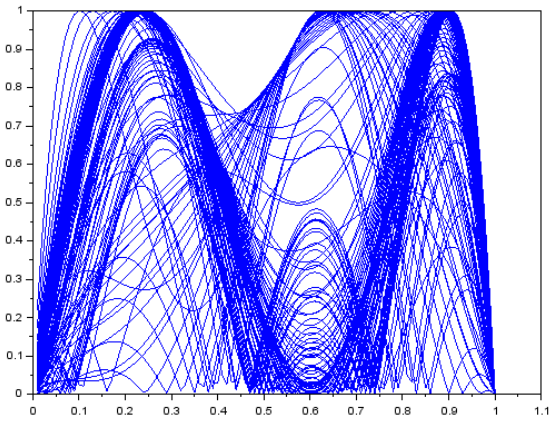

Figure 3.17: A plot of the absolute scaled defect obtained from applying CMIRK653 with N=100 to the test problem SWAVE (1.1) with $\epsilon = 0.1$.

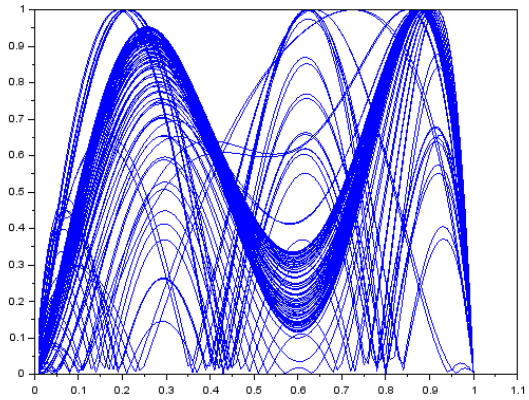

Figure 3.18: A plot of the absolute scaled defect obtained from applying CMIRK653 with N=100 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$.

3.3 Implementations of Hermite-Birkhoff interpolants

Since the standard 4^{th} and 5^{th} order CMIRK schemes do not lead to ACDC, we consider Hermite-Birkhoff interpolants (2.24) of orders four and five, and their application to the SWAVE (1.1) and SWIRL-III (1.2) problems.

3.3.1 A Fourth Order Hermite-Birkhoff Interpolant

In [14], a fourth order Hermite-Birkhoff (H-B) scheme was developed using the standard fourth order CMIRK interpolant (3.9) as a basis. The fourth order H-B scheme uses y_i, y_{i+1}, k_1, k_2, and two additional stages, k_5, k_6, constructed using the boot-strapping algorithm described in [18]. The extra stages, associated with the abscissa values $c_5 = \frac{86}{100}$ and $c_6 = \frac{93}{100}$, are based on evaluations of the standard CMIRK scheme; they have the form,

$$k_{4+j} = f(t_i + c_{4+j}h_i, u_i(t_i + c_{4+j}h_i)), \tag{3.13}$$

where $j = 1, 2$.

The Hermite-Birkhoff interpolant then has the form,

$$\begin{aligned} \tilde{u}_i(t_i + \theta h_i) &= d_0(\theta) y_i + d_1(\theta) y_{i+1} \\ &\quad + h_i \left(\tilde{b}_1(\theta) k_1 + \tilde{b}_2(\theta) k_2 + \tilde{b}_5(\theta) k_5 + \tilde{b}_6(\theta) k_6 \right), \end{aligned} \tag{3.14}$$

where $d_0(\theta)$, $d_1(\theta)$, $\tilde{b}_1(\theta)$, $\tilde{b}_2(\theta)$, $\tilde{b}_5(\theta)$, and $\tilde{b}_6(\theta)$ are weight polynomials of degree five, obtained from the interpolation conditions, $\tilde{u}_i(t_i) = y_i$, $\tilde{u}_i(t_{i+1}) = y_{i+1}$, $\tilde{u}_i'(t_i) = k_1$, $\tilde{u}_i'(t_{i+1}) = k_2$, $\tilde{u}_i'(t_{i+c_5 h_i}) = k_5$, $\tilde{u}_i'(t_{i+c_6 h_i}) = k_6$. This gives

$$d_0(\theta) = 1 - \tfrac{11997}{1024}\theta^2 + \tfrac{12949}{512}\theta^3 - \tfrac{20925}{1024}\theta^4 + \tfrac{375}{64}\theta^5,$$

$$d_1(\theta) = \tfrac{11997}{1024}\theta^2 - \tfrac{12949}{512}\theta^3 + \tfrac{20925}{1024}\theta^4 - \tfrac{375}{64}\theta^5,$$

$$\tilde{b}_1(\theta) = \theta - \tfrac{35442229}{8189952}\theta^2 + \tfrac{28704301}{4094976}\theta^3 - \tfrac{41250325}{8189952}\theta^4 + \tfrac{5375}{3968}\theta^5,$$

$$\tilde{b}_2(\theta) = -\tfrac{2291427}{100352}\theta^2 + \tfrac{3838251}{50176}\theta^3 - \tfrac{8579075}{100352}\theta^4 + \tfrac{199625}{6272}\theta^5,$$

$$\tilde{b}_5(\theta) = -\tfrac{47953125}{1078784}\theta^2 + \tfrac{74828125}{539392}\theta^3 - \tfrac{155453125}{1078784}\theta^4 + \tfrac{78125}{1568}\theta^5,$$

$$\tilde{b}_6(\theta) = \tfrac{8734375}{145824}\theta^2 - \tfrac{14359375}{72912}\theta^3 + \tfrac{31234375}{145824}\theta^4 - \tfrac{234375}{3038}\theta^5.$$

We next apply the above 4^{th} order H-B interpolant (3.14) to the SWAVE problem (1.1) and the SWIRL-III problem (1.2). (The standard 4^{th} order MIRK/CMIRK pair considered earlier are applied first; then we compute the 4^{th} order H-B interpolant as described above). The plots of the absolute scaled defects are given in Figures 3.19 and 3.20. The location of the maximum defect is the same for all subintervals and both problems, and it occurs at $\theta \approx 0.23$.

Since $u_i(t)$ is a fourth order CMIRK scheme, each evaluation of this scheme as well as the stages k_2, k_5, and k_6 (with a Lipschitz assumption on f) has an error that is $O(h_i^5)$. Therefore the error contributions of the terms $h_i k_2$, $h_i k_5$, and $h_i k_6$ are $O(h_i^6)$ while the y_i and k_1 terms are considered exact and thus contribute no data error to $\tilde{u}_i(t)$. We note also from standard interpolation theory, the interpolation error associated with $\tilde{u}_i$ is $O(h_i^6)$. Thus the term $d_1(\theta)y_{i+1}$ contributes the largest error of $O(h_i^5)$ to the new interpolant $\tilde{u}_i(t)$. The continuous local error is therefore

$$\tilde{u}_i(t) - z_i(t) = d_1(\theta) C_i {h_i}^5 + O(h_i^6), \tag{3.15}$$

where C_i is associated with the data error for y_{i+1}, and from (2.22) the defect of $\tilde{u}_i(t)$ satisfies

$$\tilde{\delta}(t) = \tilde{u}_i'(t) - z_i'(t) + O(h_i^5) = d_1'(\theta) C_i h_i^4 + O(h_i^5). \tag{3.16}$$

Therefore as h_i becomes sufficiently small, the location of the maximum defect on each subinterval for any problem will coincide with the extremum of the polynomial $d_1'(\theta)$; the local maximum of $d_1'(\theta) = \frac{11997}{512}\theta - \frac{38847}{512}\theta^2 + \frac{20925}{256}\theta^3 - \frac{1875}{64}\theta^4$. Its maximum will occur where its derivative $(\frac{11997}{512} - \frac{38847}{256}\theta + \frac{62775}{256}\theta^2 - \frac{1875}{16}\theta^3)$ is zero $\Rightarrow$ $\theta = 0.2313271928$; see Figure 3.21.

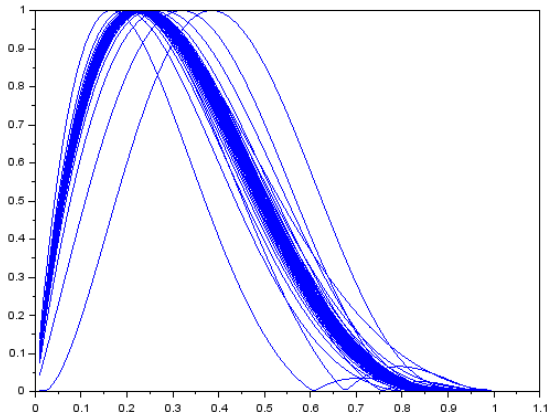

Figure 3.19: A plot of the absolute scaled defect obtained by applying the 4^{th} order Hermite-Birkhoff scheme with N=100 to the test SWAVE problem (1.1) with $\epsilon = 0.1$.

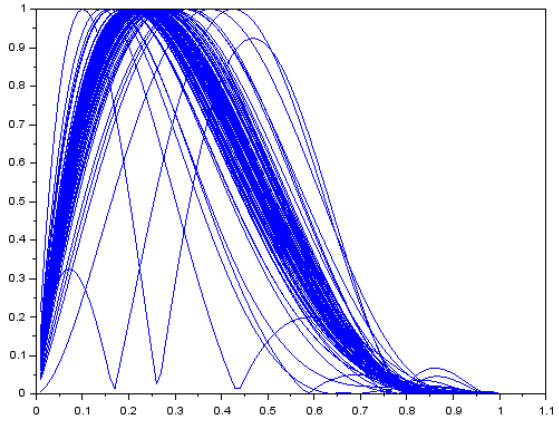

Figure 3.20: A plot of the absolute scaled defect obtained by applying the 4^{th} order Hermite-Birkhoff scheme with N=100 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$.

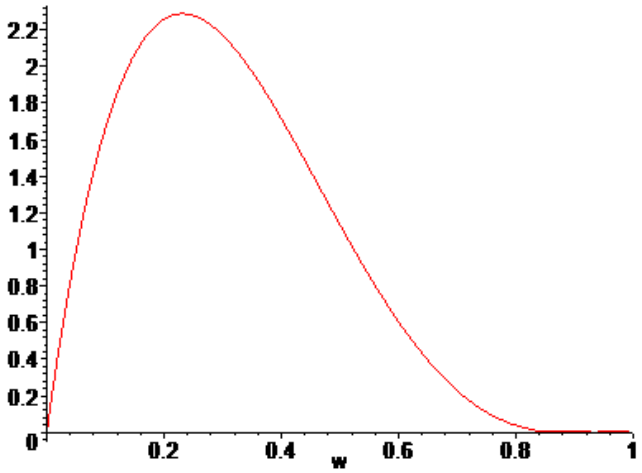

Figure 3.21: A plot of $d_1'(\theta)$ for the 4^{th} order Hermite-Birkhoff interpolant (3.15).

3.3.2 A Fifth Order Hermite-Birkhoff Interpolant

We derive a fifth order H-B interpolant by applying the same process that was used by Enright and Muir to derive the sixth order H-B interpolant in [20], and by Ellis in [14] to derive the fourth order H-B interpolant. The fifth order H-B interpolant is based on the standard fifth order CMIRK interpolant (3.11). The fifth order H-B interpolant uses y_i, y_{i+1}, k_1, k_2, and three additional stages, k_7, k_8, k_9, constructed using the boot-strapping algorithm described in [18]. The extra stages associated with the abscissa values $c_7 = \frac{7}{100}$, $c_8 = \frac{14}{100}$ and $c_9 = \frac{86}{100}$, are based on evaluations of the standard CMIRK scheme; they have the form,

$$k_{6+j} = f(t_i + c_{6+j}h_i, u_i(t_i + c_{6+j}h_i)), \tag{3.17}$$

where $j = 1, 2, 3$.

The H-B interpolant has the form,

$$\begin{aligned} \tilde{u}_i(t_i + \theta h_i) &= d_0(\theta)y_i + d_1(\theta)y_{i+1} \\ &\quad + h_i\left(\tilde{b}_1(\theta)k_1 + \tilde{b}_2(\theta)k_2 + \tilde{b}_7(\theta)k_7 + \tilde{b}_8(\theta)k_8 + \tilde{b}_9(\theta)k_9\right), \end{aligned} \tag{3.18}$$

where $d_0(\theta)$, $d_1(\theta)$, $\tilde{b}_1(\theta)$, $\tilde{b}_2(\theta)$, $\tilde{b}_7(\theta)$, $\tilde{b}_8(\theta)$, and $\tilde{b}_9(\theta)$ are weight polynomials of degree six, obtained from the interpolation conditions, $\tilde{u}_i(t_i) = y_i$, $\tilde{u}_i(t_{i+1}) = y_{i+1}$, $\tilde{u}_i'(t_i) = k_1$, $\tilde{u}_i'(t_{i+1}) = k_2$, $\tilde{u}_i'(t_{i+c_7h_i}) = k_7$, $\tilde{u}_i'(t_{i+c_8h_i}) = k_8$, and $\tilde{u}_i'(t_{i+c_9h_i}) = k_9$. This gives

$$d_0(\theta) = 1 - \tfrac{147}{199}\theta^2 + \tfrac{99414}{8557}\theta^3 - \tfrac{472650}{8557}\theta^4 + \tfrac{621000}{8557}\theta^5 - \tfrac{250000}{8557}\theta^6,$$

$$d_1(\theta) = \tfrac{147}{199}\theta^2 - \tfrac{99414}{8557}\theta^3 + \tfrac{472650}{8557}\theta^4 - \tfrac{621000}{8557}\theta^5 + \tfrac{250000}{8557}\theta^6,$$

$$\tilde{b}_1(\theta) = \theta - \tfrac{767698}{59899}\theta^2 + \tfrac{3564741697}{54088797}\theta^3 - \tfrac{2483339900}{18029599}\theta^4 + \tfrac{2236098500}{18029599}\theta^5 - \tfrac{2183875000}{54088797}\theta^6,$$

$$\tilde{b}_2(\theta) = \tfrac{13349}{795801}\theta^2 - \tfrac{29658871}{102658329}\theta^3 + \tfrac{4325550}{2575657}\theta^4 - \tfrac{852548500}{239536101}\theta^5 + \tfrac{1546375000}{718608303}\theta^6,$$

$$\tilde{b}_7(\theta) = \tfrac{1000000}{51429}\theta^2 - \tfrac{6806000000}{46440387}\theta^3 + \tfrac{57000000}{166453}\theta^4 - \tfrac{5000000000}{15480129}\theta^5 + \tfrac{5000000000}{46440387}\theta^6,$$

$$\tilde{b}_8(\theta) = -\tfrac{2546875}{359394}\theta^2 + \tfrac{42757609375}{486799173}\theta^3 - \tfrac{77402328125}{324532782}\theta^4 + \tfrac{39244375000}{162266391}\theta^5 - \tfrac{40937500000}{486799173}\theta^6,$$

$$\tilde{b}_9(\theta) = -\tfrac{1203125}{4056018}\theta^2 + \tfrac{3704984375}{784839483}\theta^3 - \tfrac{84653921875}{3662584254}\theta^4 + \tfrac{60891875000}{1831292127}\theta^5 - \tfrac{80000000000}{5493876381}\theta^6.$$

We apply the above fifth order H-B interpolant (3.19) to the SWAVE problem (1.1) and the SWIRL-III problem (1.2). The plot of the absolute scaled defects are given in Figures 3.22 and 3.23. The location of the maximum defect is the same for all subintervals and both problems, and it occurs at $\theta \approx 0.58$.

Similar to the 4^{th} order case, it can be shown that

$$\tilde{\delta}(t) = \tilde{u}_i'(t) - z_i'(t) + O(h_i^6) = d_1'(\theta) C_i h_i^5 + O(h_i^6). \tag{3.19}$$

We can therefore predict the location of the maximum defect since it will occur at the maximum of the derivative of $d_1(\theta)$, i.e., at the maximum of $d_1'(\theta) = \frac{294}{199}\theta - \frac{298242}{8557}\theta^2 + \frac{1890600}{8557}\theta^3 - \frac{3105000}{8557}\theta^4 + \frac{1500000}{8557}\theta^5$. Its maximum will occur where its derivative ($\frac{294}{199} - \frac{596484}{8557}\theta + \frac{5671800}{8557}\theta^2 - \frac{12420000}{8557}\theta^3 + \frac{7500000}{8557}\theta^4$) is zero $\Rightarrow \theta = 0.580173236$; see Figure 3.24.

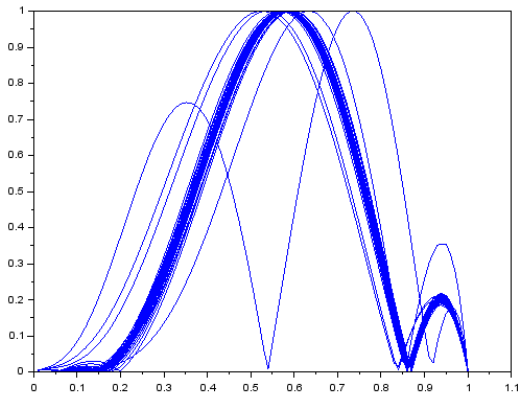

Figure 3.22: A plot of the absolute scaled defect obtained from applying the 5^{th} order Hermite-Birkhoff scheme with N=60 to the test SWAVE problem (1.1) with $\epsilon = 0.1$.

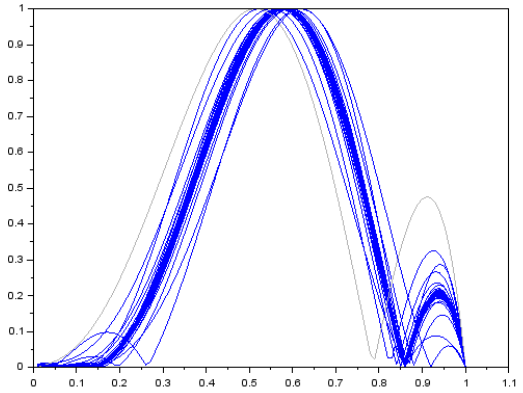

Figure 3.23: A plot of the absolute scaled defect obtained from applying the 5^{th} order Hermite-Birkhoff scheme with N=60 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$.

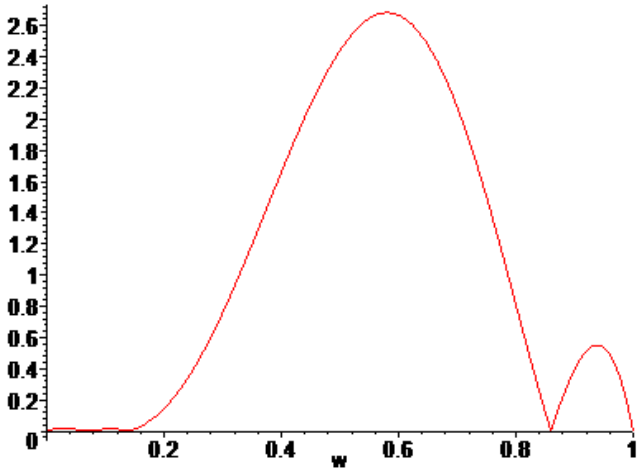

Figure 3.24: A plot of $d_1'(\theta)$ for the 5^{th} order Hermite-Birkhoff interpolant (3.19).

Chapter 4

Derivations of ACDC CMIRK Schemes for Orders 4 and 5

The use of a standard MIRK/CMIRK scheme, followed by the use of a Hermite-Birkhoff interpolant, provides a continuous numerical solution that leads to ACDC. An advantage of this approach is that it is quite general. However, a disadvantage is that the total number of stages that are used can be greater than necessary. In this chapter, we will consider an alternative approach to obtaining a continuous numerical solution that leads to ACDC. The idea is to directly obtain CMIRK schemes that lead to ACDC by deriving schemes that have only one unsatisfied order condition of the next highest order so that the leading order term in the local error expansion is a multiple of a single polynomial in θ (namely, the one unsatisfied order condition). We will refer to a CMIRK scheme that directly leads to ACDC as an ACDC CMIRK scheme.

4.1 Derivation of a Fourth Order ACDC CMIRK scheme

Some preliminary work in the derivation of a 4^{th} order ACDC CMIRK scheme was considered in [14]. In [14], Ellis derived a new fourth order CMIRK scheme through the direct approach of satisfying all but one of the continuous order conditions for the next highest order in order to simplify the expression in (2.23). In this thesis, we

will expand upon this idea to derive new CMIRK schemes that lead to ACDC.

To directly derive a fourth order CMIRK scheme that leads to an asymptotically correct defect estimate, we start by embedding the discrete MIRK343 (3.10) in a family of five stage, fourth order, stage order three CMIRK schemes with stage order vector, SOV = (4, 4, 3, 3, 4), which means that we impose stage order 4, 4, 3, 3, 4 on the first, second, third, fourth and fifth stages, respectively. The resulting Butcher tableau is

$$\begin{array}{c|c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 \\ c_4 & v_4 & x_{41} & x_{42} & x_{43} & 0 & 0 \\ c_5 & v_5 & x_{51} & x_{52} & x_{53} & x_{54} & 0 \\ \hline & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) \end{array} , \tag{4.1}$$

where x_{41}, x_{42}, and x_{43} are given in terms of c_4, v_4, due to the imposition of the stage order 3 conditions, and x_{51}, x_{52}, x_{53}, and x_{54} are given in terms of c_5, v_5, due to the imposition of the stage order 4 conditions.

We then require that the weight polynomials satisfy the standard fourth order continuous order conditions: $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$ and $b(\theta)^T c^3 = \frac{\theta^4}{4}$, and one of the two unsatisfied fifth order continuous order conditions, $b(\theta)^T c^4 = \frac{\theta^5}{5}$

and $b(\theta)^T(Xc^3 + \frac{v}{4}) = \frac{\theta^5}{20}$. (There are nine unsatisfied fifth order conditions if we consider a 4^{th} order scheme that has only stage order one. The imposition of the stage order three conditions effectively reduces the number of fifth order conditions from the nine to multiples of just two: $b(\theta)^T c^4 = \frac{1}{5}\theta^5$ and $b(\theta)^T \left(Xc^3 + \frac{v}{4}\right) = \frac{1}{20}\theta^5$.)

After solving the four fourth order conditions plus the first of the fifth order conditions, i.e., $b(\theta)^T c^4 = \frac{1}{5}\theta^5$, using the five weight polynomials $b_1(\theta)$, $b_2(\theta)$, ..., $b_5(\theta)$, we are left with the four free parameters, c_4, v_4, c_5, v_5. By (arbitrarily) choosing $c_4 = v_4 = 1/4$ and $c_5 = v_5 = 3/4$, we obtain an example of a 4^{th} order ACDC CMIRK method (CMIRK543-I) (which was derived in [14]):

$$\begin{array}{c|c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{4} & \frac{2}{16} & -\frac{1}{16} & -\frac{1}{16} & 0 & 0 \\
\frac{3}{4} & \frac{3}{4} & -\frac{1}{128} & -\frac{13}{128} & -\frac{5}{64} & \frac{3}{16} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta)
\end{array}, \tag{4.2}$$

where

$$b_1(\theta) = \frac{1}{90}\theta(90 - 375\theta + 700\theta^2 - 600\theta^3 + 192\theta^4),$$

$$b_2(\theta) = \frac{1}{90}\theta^2(-45 + 220\theta - 360\theta^2 + 192\theta^3),$$

$$b_3(\theta) = \frac{2}{15}\theta^2(-45 + 190\theta - 240\theta^2 + 96\theta^3),$$

$$b_4(\theta) = -\frac{8}{45}\theta^2(-45 + 130\theta - 135\theta^2 + 48\theta^3),$$

$$b_5(\theta) = -\frac{8}{45}\theta^2(-15 + 70\theta - 105\theta^2 + 48\theta^3).$$

An alternative approach is to start with the same CMIRK scheme (4.1), and apply the four 4^{th} order continuous order conditions and the second of the fifth order continuous order conditions $b(\theta)^T \left(Xc^3 + \frac{v}{4}\right) = \frac{1}{20}\theta^5$. This gives another example of this type of method, again with four free parameters c_4, v_4, c_5, v_5. Choosing the four free parameters $c_4 = v_4 = 1/3$ and $c_5 = v_5 = 2/3$, gives another 4^{th} order ACDC CMIRK method (CMIRK543-II):

$$
\begin{array}{c|c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{8} & -\frac{1}{8} & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{11}{81} & -\frac{7}{81} & -\frac{4}{81} & 0 & 0 \\
\frac{2}{3} & \frac{2}{3} & \frac{1}{81} & -\frac{8}{81} & -\frac{20}{81} & \frac{27}{81} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta)
\end{array}
, \tag{4.3}
$$

where

$$b_1(\theta) = -\frac{1}{120}\theta(-120 + 360\theta - 350\theta^2 + 45\theta^3 + 54\theta^4),$$

$$b_2(\theta) = -\frac{1}{120}\theta^2(-30 + 190\theta - 225\theta^2 + 54\theta^3),$$

$$b_3(\theta) = -\frac{4}{15}\theta^2(15 + 5\theta - 45\theta^2 + 27\theta^3),$$

$$b_4(\theta) = \frac{27}{40}\theta^2(10 - 10\theta - 5\theta^2 + 6\theta^3),$$

$$b_5(\theta) = \frac{27}{40}\theta^3(10 - 15\theta + 6\theta^2).$$

We discovered that when we choose values for the free coefficients of CMIRK543-I to be the same as the values for the free coefficients of CMIRK543-II we obtain the same CMIRK543 method in each case. That is, starting with the CMIRK family (4.1) with SOV=(4,4,3,3,4) and deriving a CMIRK method by imposing the four 4^{th} order continuous order conditions and either of the 5^{th} order conditions leads to the same family of ACDC CMIRK schemes. Future work will investigate why this happens.

4.2 Numerical Experiments with CMIRK543-I

When applying CMIRK543-I (4.2) to solve the SWAVE problem (1.1) and the SWIRL-III problem (1.2), we obtain continuous numerical solutions for which, for most of the subintervals, the maximum defects occur in the same location; see Figures 4.2 and 4.3. For most subintervals, the defects are multiples of the same polynomial and we can predict what the polynomial is. It will be the derivative of the lone unsatisfied 5^{th} order condition; see Figure (4.1). This happens because the leading order term in the local error expansion will be a multiple of the lone unsatisfied order condition, and the leading term in the defect is a multiple of the derivative of the leading term in the local error expansion.

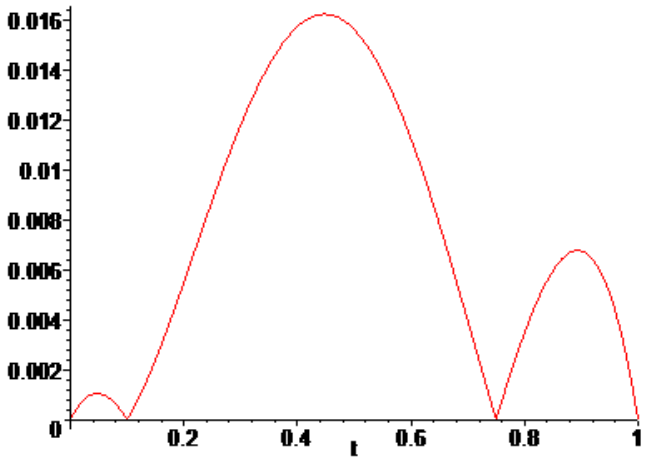

Figure 4.1: A plot of the absolute derivative of the lone unsatisfied 5^{th} order condition $b(\theta)^T(Xc^3 + \frac{v}{4}) = \frac{1}{20}\theta^5$ for the CMIRK543-I.

For most subintervals, the location of the maximum defect does not depend on the subinterval or the problem. From an inspection of Figures 4.2 and 4.3 the location of the maximum defect for CMIRK543-I is at $\theta \approx 0.44$. We can predict the location of this maximum by finding the maximum of the derivative of the unsatisfied order condition, $b(\theta)^T(Xc^3 + \frac{v}{4}) = \frac{1}{20}\theta^5$, shown in Figure 4.1. For CMIRK543-I, this polynomial is $-\frac{1}{64}\theta(\theta - 1)(4\theta - 3)(10\theta - 1)$, and its maximum is at $\theta = 0.4473760769$.

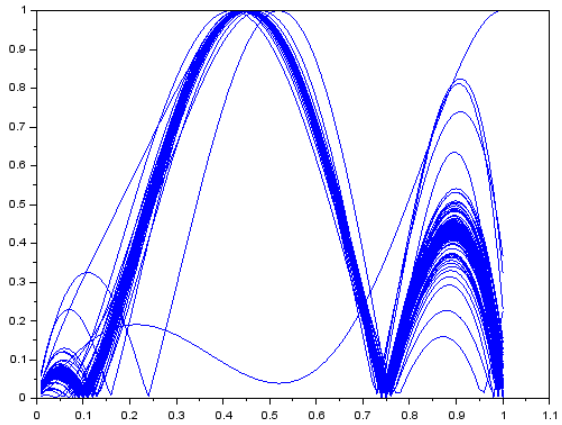

Figure 4.2: A plot of the absolute scaled defect obtained from applying the CMIRK543-I scheme with N=100 to the SWAVE problem (1.1) with $\epsilon = 0.1$.

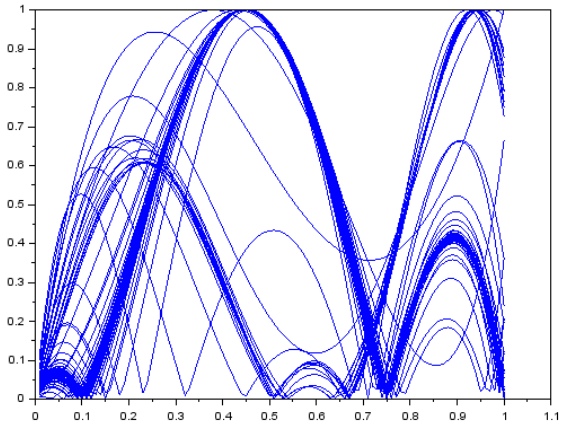

Figure 4.3: A plot of the absolute scaled defect obtained from applying the CMIRK543-I scheme with N=100 to the SWIRL-III problem (1.2) with $\epsilon = 0.01$.

4.3 Derivation of a Fifth Order ACDC CMIRK scheme

We derive a fifth order CMIRK scheme that directly gives the ACDC property, and has the discrete optimal MIRK553 (3.12) scheme embedded within it. The discrete scheme has the stage order vector SOV = (5, 5, 3, 3, 5). We will require all additional stages of the CMIRK scheme to have stage order 5. We will require the scheme to satisfy the six fifth order conditions that must be satisfied in order to obtain a 5^{th} order CMIRK scheme that has stage order 3, $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b(\theta)^T c^4 = \frac{\theta^5}{5}$, and $b(\theta)^T(Xc^3 + \frac{v}{4}) = \frac{\theta^5}{20}$. We will also require the scheme to satisfy three out of the four sixth order conditions (when a CMIRK scheme has stage order 3, there are 4 conditions of order six.)

By substituting the stage order four condition, $C4 = Xc^3 + \frac{v}{4} - \frac{c^4}{4}$, and $b(\theta)^T c^4 = \frac{\theta^5}{5}$, we can transform the fifth order condition $b(\theta)^T(Xc^3 + \frac{v}{4}) = \frac{\theta^5}{20}$ into $b^T(\theta)C4 = 0$. The four sixth order conditions are: $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$, $b^T(\theta)c(Xc^3 +$

$\frac{v}{4}) = \frac{\theta^6}{24}$, and $b^T(\theta)(X(Xc^3 + \frac{v}{4}) + \frac{v}{20}) = \frac{\theta^6}{120}$. By substituting the stage order five condition, $C5 = Xc^4 + \frac{v}{5} - \frac{c^5}{5}$, and $b^T(\theta)c^5 = \frac{\theta^6}{6}$, we can transform the sixth order condition $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$ into $b^T(\theta)C5 = 0$. By substituting the stage order four condition, $C4$, and $b^T(\theta)c^5 = \frac{\theta^6}{6}$, we can transform the sixth order condition $b^T(\theta)c(Xc^3 + \frac{v}{4}) = \frac{\theta^6}{24}$ into $b^T(\theta)cC4 = 0$. By substituting the stage order four condition, $C4$, and $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$, we can transform the sixth order condition $b^T(\theta)(X(Xc^3 + \frac{v}{4}) + \frac{v}{20}) = \frac{\theta^6}{120}$ into $b^T(\theta)XC4 = 0$.

If we were to simply impose all nine conditions (the six 5^{th} order conditions and three of the four 6^{th} order conditions) the CMIRK scheme would require 9 stages. Our goal is to derive a scheme with as few stages as possible. Here we derive a family of fifth order, stage order three CMIRK schemes with eight stages, CMIRK853, with SOV = (5, 5, 3, 3, 5, 5, 5, 5). The resulting Butcher tableau is

$$
\begin{array}{c|c|cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c_3 & v_3 & x_{31} & x_{32} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_4 & v_4 & x_{41} & x_{42} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_5 & v_5 & x_{51} & x_{52} & x_{53} & x_{54} & 0 & 0 & 0 & 0 \\
c_6 & v_6 & x_{61} & x_{62} & x_{63} & x_{64} & x_{65} & 0 & 0 & 0 \\
c_7 & v_7 & x_{71} & x_{72} & x_{73} & x_{74} & x_{75} & x_{76} & 0 & 0 \\
c_8 & v_8 & x_{81} & x_{82} & x_{83} & x_{84} & x_{85} & x_{86} & x_{87} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) & b_7(\theta) & b_8(\theta)
\end{array}
, \qquad (4.4)
$$

where the coefficients for the first five rows of the tableau are given in the tableau of the embedded discrete (MIRK553) scheme (3.12).

As mentioned earlier, the sixth, seventh and eighth stages are required to satisfy the stage order five conditions. After imposing these stage order conditions on stages six, seven and eight, we get x_{61}, x_{62}, x_{63}, x_{64} and x_{65} in terms of c_6 and v_6, and x_{72}, x_{73}, x_{74}, x_{75} and x_{76} in terms of c_7, v_7 and x_{71}, and x_{83}, x_{84}, x_{85}, x_{86} and x_{87} in terms of c_8, v_8, x_{81} and x_{82}. We then require that the weight polynomials and remaining nine free coefficients, c_6, v_6, c_7, v_7, x_{71}, c_8, v_8, x_{81} and x_{82} be chosen to satisfy the six fifth order continuous order conditions: $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$ and $b^T(\theta)C4 = 0$, and three of the four unsatisfied sixth order continuous order conditions, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)C5 = 0$, $b^T(\theta)cC4 = 0$, and

$b^T(\theta)XC4 = 0$.

The three sixth order conditions we choose to be satisfied are: $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)C5 = 0$, and $b^T(\theta)cC4 = 0$. Since the third and fourth stages have only stage order three while the other stages have stage order 5, $C4$ and $C5$ have all zero entries except for positions 3 and 4. By requiring $b_3(\theta)$ and $b_4(\theta)$ to be zero identically, it follows that the fifth order condition, $b^T(\theta)C4 = 0$, and the sixth order conditions $b^T(\theta)C5 = 0$ and $b^T(\theta)cC4 = 0$ will be automatically satisfied. This leaves us with the six conditions, $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, and $b^T(\theta)c^5 = \frac{\theta^6}{6}$, to be satisfied. Therefore, the total number of stages that we will need is eight.

We use the six weight polynomials $b_1(\theta)$, $b_2(\theta)$, $b_5(\theta)$, $b_6(\theta)$, $b_7(\theta)$, $b_8(\theta)$ to satisfy these six order conditions. This leaves us with the nine free parameters mentioned earlier. By choosing values for them, as indicated in the following tableau, we obtain an example of a 5^{th} order ACDC CMIRK method as follows:

$$
\begin{array}{c|c|cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c_3 & v_3 & x_{31} & x_{32} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_4 & v_4 & x_{41} & x_{42} & 0 & 0 & 0 & 0 & 0 & 0 \\
c_5 & v_5 & x_{51} & x_{52} & x_{53} & x_{54} & 0 & 0 & 0 & 0 \\
c_6 & v_6 & x_{61} & x_{62} & x_{63} & x_{64} & x_{65} & 0 & 0 & 0 \\
c_7 & v_7 & x_{71} & x_{72} & x_{73} & x_{74} & x_{75} & x_{76} & 0 & 0 \\
c_8 & v_8 & x_{81} & x_{82} & x_{83} & x_{84} & x_{85} & x_{86} & x_{87} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) & b_7(\theta) & b_8(\theta)
\end{array}, \tag{4.5}
$$

where the coefficients for the first five rows of the tableau are given in the tableau of the embedded discrete (MIRK553) scheme (3.12), and

$$c_6 = \frac{1}{2}, \quad v_6 = \frac{1}{2}, \quad x_{61} = -\frac{495}{73984}\sqrt{393} - \frac{1229}{73984}, \quad x_{62} = \frac{3121}{150528} - \frac{143}{451584}\sqrt{393},$$

$$x_{63} = \frac{316130859}{2112158720} + \frac{16758009}{2112158720}\sqrt{393}, \quad x_{64} = -\frac{19261875}{82725094} + \frac{295625}{744525846}\sqrt{393},$$

$$x_{65} = \frac{73074419}{924430080} - \frac{1223791}{924430080}\sqrt{393}$$

$$c_7 = \frac{1}{2} - \frac{\sqrt{7}}{14}, \quad v_7 = \frac{1}{2} - \frac{\sqrt{7}}{14}, \quad x_{71} = \frac{3}{112} + \frac{9\sqrt{7}}{1960},$$

$$x_{72} = -\frac{3555}{38416} + \frac{11}{19208}\sqrt{393} + \frac{3509}{48404160}\sqrt{393}\sqrt{7} - \frac{60211}{16134720}\sqrt{7},$$

$$x_{73} = \frac{34702587}{27491065840}\sqrt{393} + \frac{50869782}{1718191615} + \frac{2446484931}{769749843520}\sqrt{7}$$

$$+\frac{520529229}{3848749217600}\sqrt{393}\sqrt{7},$$

$$x_{74} = -\frac{295625}{1669100426}\sqrt{393} + \frac{173356875}{1669100426} + \frac{211880625}{5841851491}\sqrt{7} - \frac{3251875}{52576663419}\sqrt{393}\sqrt{7},$$

$$x_{75} = -\frac{8136237}{111022240}\sqrt{393} + \frac{112728633}{111022240} + \frac{124233703}{1165733520}\sqrt{7} - \frac{14610453}{1942889200}\sqrt{393}\sqrt{7},$$

$$x_{76} = \frac{13365}{186592}\sqrt{393} - \frac{202095}{186592} - \frac{479613}{3265360}\sqrt{7} + \frac{24079}{3265360}\sqrt{393}\sqrt{7},$$

$$c_8 = \frac{87}{100}, \quad v_8 = c_8 = \frac{87}{100}, \quad x_{81} = \frac{2707592511}{1000000000000} - \frac{1006699707\sqrt{7}}{1000000000000},$$

$$x_{82} = -\frac{51527976591}{1000000000000} - \frac{1006699707\sqrt{7}}{1000000000000},$$

$$x_{83} = \frac{3252672638312435815 23}{3172893778700000000000000} + \frac{14587453893841540569}{317289377870000000000000}\sqrt{393}$$

$$+\frac{2262384450385281729}{4532705410000000000000}\sqrt{7} + \frac{121801009211392563}{4532705410000000000000}\sqrt{393}\sqrt{7},$$

$$x_{84} = \frac{9977027915179}{919708398000000}\sqrt{7} - \frac{459372549779}{248321267460000000}\sqrt{393}\sqrt{7} - \frac{3678014263866527}{459854199000000000} +$$

$$\frac{169346904196327}{1241606337300000000}\sqrt{393},$$

$$x_{85} = -\frac{2143419845411741}{606900000000000000} - \frac{984609766854493}{86700000000000000}\sqrt{7} + \frac{101060638539809}{86700000000000000}\sqrt{393}\sqrt{7} -$$

$$\frac{788117416154783}{606900000000000000}\sqrt{393},$$

$$x_{86} = \frac{122326924577}{119000000000000}\sqrt{393} - \frac{1075079202067}{119000000000000} - \frac{284341779579}{29750000000000}\sqrt{393}\sqrt{7} +$$

$$\frac{4778781424469}{29750000000000}\sqrt{7},$$

$$x_{87} = -\frac{4822701365224561}{85509000000000000}\sqrt{7} + \frac{10441137625184357}{42754500000000000}$$

$$-\frac{4864579879858699}{2308743000000000000}\sqrt{393}\sqrt{7} + \frac{2435992347982193}{1154371500000000000}\sqrt{393},$$

and

$$b_1(\theta) = \frac{1}{1669731840}(-875 - 125\sqrt{7} + 7\sqrt{393} + \sqrt{393}\sqrt{7})\theta(90681115\theta + 326250\sqrt{7}$$

$$-18270\sqrt{393} + 2610\sqrt{393}\sqrt{7} + 1401140\theta^2\sqrt{7} - 18853800\theta^2 + 21598710\theta^3$$

$$-12900384\theta^4 + 31360000\theta^5 - 2283750 - 5415\theta\sqrt{393}\sqrt{7} - 983820\theta^3\sqrt{7} - 83720\theta^2\sqrt{393}$$

$$+60270\theta^3\sqrt{393} + 56175\theta\sqrt{393} - 969195\theta\sqrt{7} + 2688000\theta^4\sqrt{7} + 4740\theta^2\sqrt{393}\sqrt{7}$$

$$-1500\theta^3\sqrt{393}\sqrt{7} - 168000\theta^4\sqrt{393}),$$

$$b_2(\theta) = -\frac{1}{154103040}(-693 + 99\sqrt{7} - 7\sqrt{393} + \sqrt{393}\sqrt{7})(-52841600\theta + 39270\theta^2\sqrt{393}$$

$$-168000\theta^3\sqrt{393} - 163125\sqrt{7} + 101772300\theta^2 - 9137184\theta^3 + 31360000\theta^4 + 2688000\theta^3\sqrt{7}$$

$$-6478200\theta^2\sqrt{7} - 313600\theta\sqrt{393} + 27400\theta\sqrt{393}\sqrt{7} + 1141875 - 15000\theta^2\sqrt{393}\sqrt{7}$$

$$+5373800\theta\sqrt{7} + 91350\sqrt{393} - 13050\sqrt{393}\sqrt{7})\theta^2,$$

$$b_3(\theta) = 0,$$

$$b_4(\theta) = 0,$$

$$b_5(\theta) = \frac{1}{435817403280}(-14220361\sqrt{7} + 176130429\sqrt{393}\sqrt{7} - 4314649619$$

$$-142217201\sqrt{393})(-37450\theta + 3610\theta\sqrt{7} + 9135 - 1305\sqrt{7} + 62790\theta^2$$

$$-3555\theta^2\sqrt{7} - 48216\theta^3 + 1200\theta^3\sqrt{7} + 14000\theta^4)\theta^2,$$

$$b_6(\theta) = -\frac{1}{435120}\sqrt{7}(-13 + \sqrt{393})(-5509392\theta^3 + 134400\theta^3\sqrt{7} + 24885\theta^2\sqrt{393}$$

$$-25270\theta\sqrt{393} + 1870\theta\sqrt{393}\sqrt{7} + 7355985\theta^2 + 1141875 - 163125\sqrt{7} - 4522910\theta$$

$$+9135\sqrt{393} - 1305\sqrt{393}\sqrt{7} - 407910\theta^2\sqrt{7} + 428630\theta\sqrt{7} + 1568000\theta^4$$

$$-750\theta^2\sqrt{393}\sqrt{7} - 8400\theta^3\sqrt{393})\theta^2,$$

$$b_7(\theta) = -\frac{7}{591572160}(-66745 + 20351\sqrt{7} - 1939\sqrt{393} + 821\sqrt{393}\sqrt{7})(163125$$

$$+1305\sqrt{393} - 646130\theta - 3610\theta\sqrt{393} + 1050855\theta^2 + 3555\theta^2\sqrt{393} - 787056\theta^3$$

$$-1200\theta^3\sqrt{393} + 224000\theta^4)\theta^2,$$

$$b_8(\theta) = \frac{3906250}{1091762011055079}(-452473 - 6475\sqrt{393} + 87350\sqrt{7} + 1250\sqrt{393}\sqrt{7})$$

$$(-59430\theta - 350\theta\sqrt{393} + 111300\theta^2 + 420\theta^2\sqrt{393} - 96264\theta^3 - 168\theta^3\sqrt{393} + 31360\theta^4$$

$$+13125 - 1875\sqrt{7} + 105\sqrt{393} - 15\sqrt{393}\sqrt{7} + 5990\theta\sqrt{7} + 30\theta\sqrt{393}\sqrt{7} - 6915\theta^2\sqrt{7}$$

$$-15\theta^2\sqrt{393}\sqrt{7} + 2688\theta^3\sqrt{7})\theta^2.$$

(We could choose $b^T(\theta)XC4 = 0$ instead of choosing $b^T(\theta)cC4 = 0$, but that will require one more stage. We cannot choose $b^T(\theta)C5 = 0$ to be unsatisfied because

$b^T(\theta)XC4 = 0$ depends on it. If we consider the original forms of the order conditions and do not force any of the weight polynomials to be zero, then $b^T(\theta)XC4 = 0$ cannot be satisfied because $C4_3x_{53}+C4_4x_{54}$ will not equal zero, and this will lead to a CMIRK scheme with nine stages.)

4.4 Numerical Experiments with CMIRK853

The results of applying CMIRK853 (4.5) to the SWAVE problem (1.1) and the SWIRL-III problem (1.2) are shown in Figures 4.4 and 4.5.

The location of the maximum defect is the same for most of the subintervals and for both problems. It occurs at $\theta \approx 0.14$ within each subinterval.

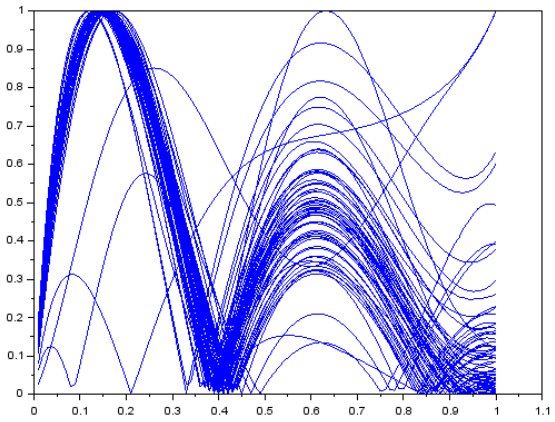

Figure 4.4: A plot of the absolute scaled defect obtained from applying the CMIRK853 scheme with N=60 to the SWAVE problem (1.1) with $\epsilon = 0.1$.

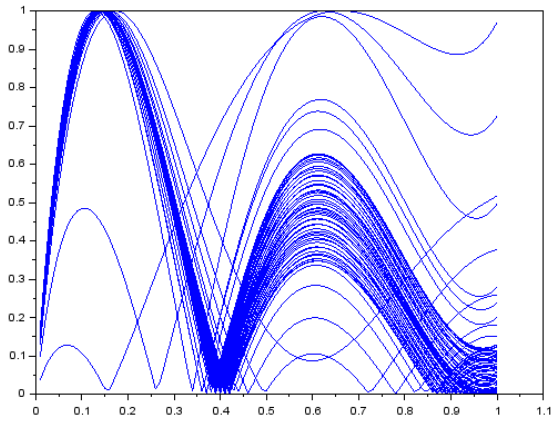

Figure 4.5: A plot of the absolute scaled defect obtained from applying the CMIRK853 scheme with N=60 to the SWIRL-III problem (1.2) with $\epsilon = 0.01$.

We can predict the shape of this absolute scaled defect and the location of the maximum. They are obtained from the derivative of the lone unsatisfied sixth order condition $b^T(\theta)XC4$ mentioned earlier. This derivative is plotted in Figure 4.6. It is $-0.6232140220\theta(0.0754657367 3\theta^3 - 0.2366997140\theta^2 + 0.2106057408\theta - 0.05113623823)(-1+\theta)$, and its maximum is at $\theta = 0.1440679896$.

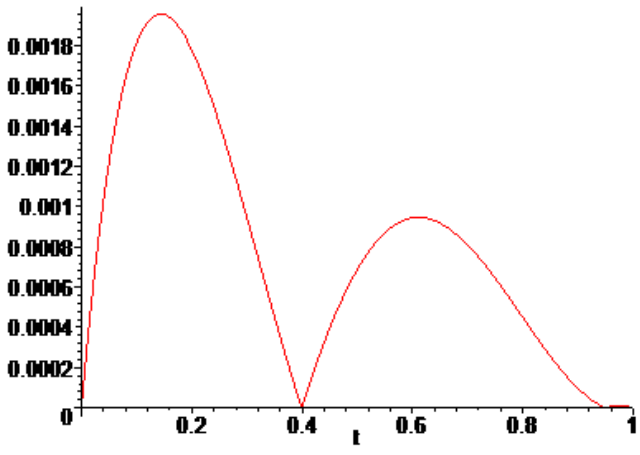

Figure 4.6: A plot of the absolute derivative of the lone unsatisfied 6^{th} order condition, $b^T(\theta)XC4$.

Chapter 5

A Comparison Between Hermite-Birkhoff Interpolants and ACDC CMIRK Schemes For Orders 4 and 5

In this chapter, we will compare ACDC CMIRK schemes with Hermite-Birkhoff interpolants for 4^{th} and 5^{th} orders applied to the SWAVE (1.1) and SWIRL-III (1.2) problems.

5.1 Comparing 4^{th} Order Schemes

We apply the 4^{th} order ACDC CMIRK543-I scheme and the 4^{th} order Hermite-Birkhoff interpolant to the SWAVE (1.1) and SWIRL-III (1.2) problems where we require the schemes to achieve approximately the same maximum defect. This is done by using a different number of subintervals, N, for each scheme. The results are presented in Tables 5.1 and 5.2. The processor is lntel(R) Core(TM) i5-3337U CPU @ 1.80GHz 1.80 GHz. The operating system is Windows 8.1. The version of Scilab is 5.5.1.

In the above tables we note that the CMIRK543-I scheme is more accurate than the 4^{th} Order H-B interpolant since it is able to solve both problems to the same

Table 5.1: A comparison between the 4^{th} order Hermite-Birkhoff interpolant and the 4^{th} order ACDC CMIRK scheme CMIRK543-I applied to the SWAVE problem. The time to solve the problem for each method is given in seconds.

Scheme	maximum defect	N	time (seconds)
4^{th} order Hermite-Birkhoff interpolant	2.6×10^{-6}	100	3.8
4^{th} Order ACDC CMIRK543-I scheme	2.6×10^{-6}	92	2.9

Table 5.2: A comparison between the 4^{th} order Hermite-Birkhoff interpolant and the 4^{th} order ACDC CMIRK scheme CMIRK543-I applied to the SWIRL-III problem. The time to solve the problem for each method is given in seconds.

Scheme	maximum defect	N	time (seconds)
4^{th} order Hermite-Birkhoff interpolant	4.8×10^{-6}	99	11.2
ACDC CMIRK543-I	4.8×10^{-6}	84	7.8

accuracy as the H-B interpolant using fewer subintervals. As well the CMIRK543-I uses fewer stages (5 vs. 6) than the H-B interpolant. These two effects lead to the CMIRK543-I scheme using less computational time than the H-B interpolant.

5.2 Comparing 5^{th} Order Schemes

We apply the 5^{th} order ACDC CMIRK853 scheme and the 5^{th} order Hermite-Birkhoff interpolant to the SWAVE (1.1) and SWIRL-III (1.2) problems where we require the schemes to achieve approximately the same maximum defect. The results are presented in Tables 5.3 and 5.4.

Table 5.3: A comparison between the 5^{th} order Hermite-Birkhoff interpolant and the 5^{th} order ACDC CMIRK scheme CMIRK853 applied to the SWAVE problem. The time to solve the problem for each method is given in seconds.

Scheme	maximum defect	N	time (seconds)
5^{th} order Hermite-Birkhoff interpolant	4×10^{-7}	57	3.878125
ACDC CMIRK853	4×10^{-7}	61	3.596875

Table 5.4: A comparison between the 5^{th} order Hermite-Birkhoff interpolant and the 5^{th} order ACDC CMIRK scheme CMIRK853 applied to the SWIRL-III problem. The time to solve the problem for each method is given in seconds.

Scheme	maximum defect	N	time (seconds)
5^{th} order Hermite-Birkhoff interpolant	4×10^{-7}	59	10.640625
ACDC CMIRK853	4×10^{-7}	66	11.134375

The ACDC CMIRK853 scheme is less accurate than the H-B interpolant and therefore requires more subintervals to obtain the same accuracy. For the SWAVE problem, the CMIRK853 scheme is still faster due to the fact that it uses fewer stages (8 vs. 9) per subinterval. However, for the SWIRL-III problem, despite using fewer stages per subinterval (8 vs. 9), the overall computational cost for the CMIRK853 scheme is larger than the cost for the H-B interpolant. Future work will optimize the CMIRK853 to be more accurate, by choosing the free coefficients to minimize the leading order term in the local error expansion for the scheme.

Chapter 6

Investigation of 6^{th} Order Standard CMIRK Schemes, Hermite-Birkhoff Interpolants, and ACDC CMIRK and CGMIRK Schemes

In this chapter, we will consider a standard sixth order CMIRK scheme, a sixth order Hermite-Birkhoff interpolant and several sixth order ACDC CMIRK schemes. We also consider several ACDC CGMIRK schemes.

6.1 A Standard Sixth Order CMIRK Scheme

We consider an eight stage, 6^{th} order, stage order three, CMIRK scheme (CMIRK863) taken from [28], which has the tableau,

$$\begin{array}{c|c|cccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2}-\frac{\sqrt{21}}{14} & \frac{1}{2}-\frac{9\sqrt{21}}{98} & x_{31} & x_{32} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2}+\frac{\sqrt{21}}{14} & \frac{1}{2}+\frac{9\sqrt{21}}{98} & x_{41} & x_{42} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & x_{51} & x_{52} & x_{53} & x_{54} & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & x_{61} & x_{62} & x_{63} & x_{64} & 0 & 0 & 0 & 0 \\
\frac{1}{2}-\frac{\sqrt{7}}{14} & \frac{1}{2}-\frac{\sqrt{7}}{14} & x_{71} & x_{72} & x_{73} & x_{74} & x_{75} & x_{76} & 0 & 0 \\
\frac{87}{100} & \frac{87}{100} & x_{81} & x_{82} & x_{83} & x_{84} & x_{85} & x_{86} & x_{87} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) & b_7(\theta) & b_8(\theta)
\end{array} , \quad (6.1)$$

where

$$x_{31} = \frac{1}{14} + \frac{\sqrt{21}}{98}, \quad x_{32} = -\frac{1}{14} + \frac{\sqrt{21}}{98}, \quad x_{41} = \frac{1}{14} - \frac{\sqrt{21}}{98}, \quad x_{42} = -\frac{1}{14} - \frac{\sqrt{21}}{98},$$

$$x_{51} = -\frac{5}{128}, \quad x_{52} = \frac{5}{128}, \quad x_{53} = \frac{7\sqrt{21}}{128}, \quad x_{54} = -\frac{7\sqrt{21}}{128},$$

$$x_{61} = \frac{1}{64}, \quad x_{62} = -\frac{1}{64}, \quad x_{63} = \frac{7\sqrt{21}}{192}, \quad x_{64} = -\frac{7\sqrt{21}}{192},$$

$$x_{71} = \frac{3}{112} + \frac{9\sqrt{7}}{1960}, \quad x_{72} = -\frac{3}{112} + \frac{9\sqrt{7}}{1960}, \quad x_{73} = \frac{3\sqrt{7}\sqrt{3}}{112} + \frac{11\sqrt{7}}{840},$$

$$x_{74} = -\frac{3\sqrt{7}\sqrt{3}}{112} + \frac{11\sqrt{7}}{840}, \quad x_{75} = \frac{88\sqrt{7}}{5145}, \quad x_{76} = -\frac{18\sqrt{7}}{343},$$

$$x_{81} = \frac{2707592511}{1000000000000} - \frac{1006699707\sqrt{7}}{1000000000000},$$

$$x_{82} = -\frac{51527976591}{1000000000000} - \frac{1006699707\sqrt{7}}{1000000000000},$$

$$x_{83} = -\frac{610366393}{75000000000} + \frac{7046897949\sqrt{7}}{1000000000000} + \frac{14508670449\sqrt{7}\sqrt{3}}{1000000000000},$$

$$x_{84} = -\frac{610366393}{75000000000} + \frac{7046897949\sqrt{7}}{1000000000000} - \frac{14508670449\sqrt{7}\sqrt{3}}{1000000000000},$$

$$x_{85} = -\frac{12456457}{1171875000} + \frac{1006699707\sqrt{7}}{109375000000},$$

$$x_{86} = \frac{3020099121\sqrt{7}}{437500000000} + \frac{47328957}{625000000}, \quad x_{87} = -\frac{7046897949\sqrt{7}}{250000000000},$$

and where

$$b_1(\theta) = -\frac{1}{2112984835740}\theta(1450\sqrt{7} + 12233)(800086000\,\theta^5 - 2936650584\,\theta^4$$

$$+63579600\sqrt{7}\theta^4 - 201404565\sqrt{7}\theta^3 + 4235152620\,\theta^3 + 232506630\sqrt{7}\theta^2$$

$$-3033109390\,\theta^2 + 1116511695\,\theta - 116253315\sqrt{7}\theta - 191568780$$

$$+22707000\sqrt{7}),$$

$$b_2(\theta) = -\frac{-10799 + 650\sqrt{7}}{29551834260}\theta^2(24962000\,\theta^4 + 473200\sqrt{7}\theta^3 - 67024328\,\theta^3$$

$$+66629600\,\theta^2 - 751855\sqrt{7}\theta^2 + 236210\sqrt{7}\theta - 29507250\,\theta + 5080365$$

$$+50895\sqrt{7}),$$

$$b_3(\theta) = b_4(\theta) = \frac{49}{64} b_5(\theta),$$

$$b_5(\theta) = \frac{4144 + 800\sqrt{7}}{2231145}\theta^2(14000\,\theta^4 - 48216\,\theta^3 + 1200\sqrt{7}\theta^3 + 62790\,\theta^2 - 3555\sqrt{7}\theta^2$$

$$+3610\sqrt{7}\theta - 37450\,\theta + 9135 - 1305\sqrt{7}),$$

$$b_6(\theta) = -\frac{-24332 + 2960\sqrt{7}}{1227278493}(\theta - 1)^2\theta^2(-1561000\,\theta^2 + 2461284\,\theta + 109520\sqrt{7}\theta$$

$$-86913\sqrt{7} - 979272),$$

$$b_7(\theta) = -\frac{49\sqrt{7}}{63747}(\theta - 1)^2\theta^2(20000\,\theta^2 - 20000\,\theta + 3393),$$

$$b_8(\theta) = -\frac{1}{889206903}(\theta - 1)^2\theta^2(35000000000\,\theta^2 - 35000000000\,\theta + 11250000000).$$

This scheme has embedded within it a discrete MIRK563 scheme [28] which has the tableau

$$\begin{array}{c|cc|ccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} - \frac{\sqrt{21}}{14} & \frac{1}{2} - \frac{9\sqrt{21}}{98} & \frac{1}{14} + \frac{\sqrt{21}}{98} & -\frac{1}{14} + \frac{\sqrt{21}}{98} & 0 & 0 & 0 \\
\frac{1}{2} + \frac{\sqrt{21}}{14} & \frac{1}{2} + \frac{9\sqrt{21}}{98} & \frac{1}{14} - \frac{\sqrt{21}}{98} & -\frac{1}{14} - \frac{\sqrt{21}}{98} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & -\frac{5}{128} & \frac{5}{128} & \frac{7\sqrt{21}}{128} & -\frac{7\sqrt{21}}{128} & 0 \\
\hline
 & & \frac{1}{20} & \frac{1}{20} & \frac{49}{180} & \frac{49}{180} & \frac{16}{45}
\end{array} . \tag{6.2}$$

We apply the MIRK563/CMIRK863 pair to the SWAVE problem (1.1) and the SWIRL-III problem (1.2). The plots of the absolute scaled defect are given in Figures 6.1 and 6.2. We note that the location of the maximum defect changes from

subinterval to subinterval and over the two problems and that the absolute scaled defect does not have the same shape on each subinterval. This happens because the leading order term in the defect expansion depends on a varying linear combination of eight polynomials. This happens because any sixth order CMIRK scheme that satisfies stage order three has a leading order term in its local error expansion that depends on the eight unsatisfied order conditions for order 7 which are:

$$b^T(\theta)\left(X\left(X\left(Xc^3+\frac{v}{4}\right)+\frac{v}{20}\right)+\frac{v}{120}\right)-\frac{\theta^7}{840}, \tag{6.3}$$

$$b^T(\theta)\left(X\left(Xc^4+\frac{v}{5}\right)+\frac{v}{30}\right)-\frac{\theta^7}{210}, \tag{6.4}$$

$$b^T(\theta)\left(X\left(c\left(Xc^3+\frac{v}{4}\right)\right)+\frac{v}{24}\right)-\frac{\theta^7}{168}, \tag{6.5}$$

$$b^T(\theta)\left(Xc^5+\frac{v}{6}\right)-\frac{\theta^7}{42}, \tag{6.6}$$

$$b^T(\theta)c\left(X\left(Xc^3+\frac{v}{4}\right)+\frac{v}{20}\right)-\frac{\theta^7}{140}, \tag{6.7}$$

$$b^T(\theta)c\left(Xc^4+\frac{v}{5}\right)-\frac{\theta^7}{35}, \tag{6.8}$$

$$b^T(\theta)c^2\left(Xc^3+\frac{v}{4}\right)-\frac{\theta^7}{28}, \tag{6.9}$$

$$b^T(\theta)c^6-\frac{\theta^7}{7}. \tag{6.10}$$

Thus, it is impossible to determine a priori where the maximum defect will occur for each subinterval and for each problem.

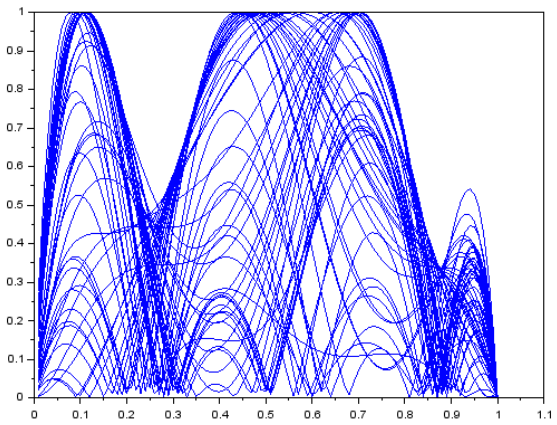

Figure 6.1: A plot of the absolute scaled defect obtained by applying CMIRK863 with N=50 to the test problem SWAVE (1.1) with $\epsilon = 0.1$.

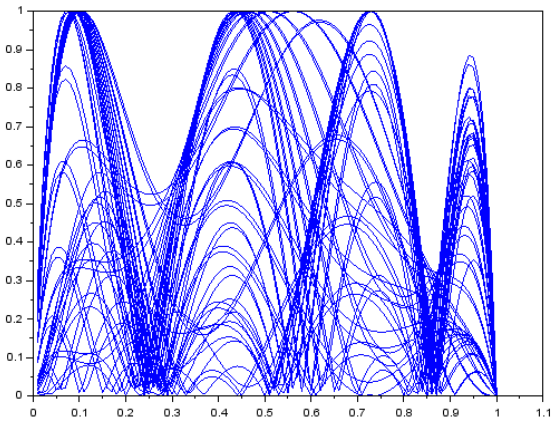

Figure 6.2: A plot of the absolute scaled defect obtained by applying CMIRK863 with N=50 to the test problem SWIRL-III (1.2) with $\epsilon = 0.01$.

However, when applying the MIRK563/CMIRK863 pair to the linear problem (3.1), or the simple nonlinear problem (3.2), we found that the scaled defect plots - see Figures 6.3 and 6.4 - were the same on all subintervals. This happens because the leading order term in the defect expansion is a multiple of a single polynomial in θ. However, we cannot determine a priori if a given problem is going to have an expansion of the local error that will be a multiple of a single polynomial.

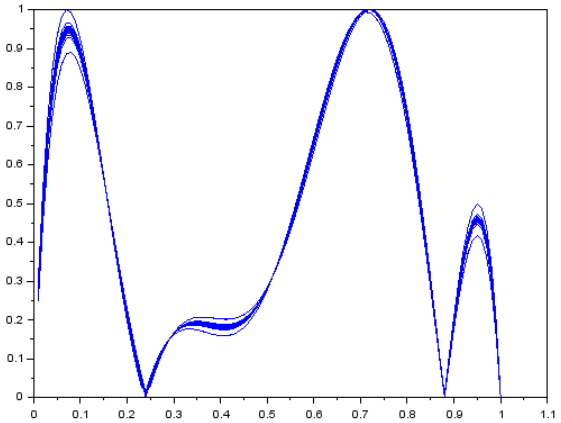

Figure 6.3: A plot of the absolute scaled defect obtained by applying CMIRK863 with N=40, to the linear test problem (3.1) with $\lambda = 1$.

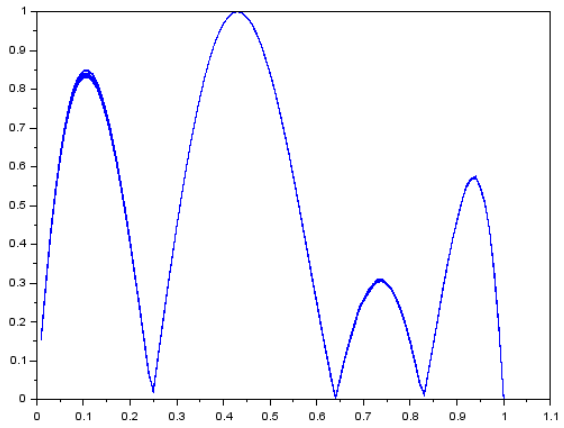

Figure 6.4: A plot of the absolute scaled defect obtained by applying CMIRK863 with N=50, to the simple nonlinear test problem (3.2).

6.2 A Sixth Order Hermite-Birkhoff Interpolant

In [20], a sixth order Hermite-Birkhoff scheme was developed using a standard sixth order CMIRK interpolant (6.1) as a basis. The sixth order Hermite-Birkhoff scheme uses y_i, y_{i+1}, k_1, k_2, and four additional stages constructed using the boot-strapping algorithm described in [18]. The extra stages, associated with the abscissa values, $c_9 = \frac{7}{100}$, $c_{10} = \frac{14}{100}$, $c_{11} = \frac{86}{100}$ and $c_{12} = \frac{93}{100}$, are based on evaluations of the basic CMIRK scheme and have the form,

$$k_{8+j} = f(t_i + c_{8+j}h_i, u_i(t_i + c_{8+j}h_i)), \tag{6.11}$$

where $j = 1, 2, 3, 4$. In (6.11), u_i is the CMIRK scheme.

The sixth order Hermite-Birkhoff interpolant has the form,

$$\begin{aligned}\tilde{u}_i(t_i+\theta h_i) &= d_0(\theta)y_i + d_1(\theta)y_{i+1} \\ &\quad + h_i\left(\tilde{b}_1(\theta)k_1 + \tilde{b}_2(\theta)k_2 + \tilde{b}_9(\theta)k_9 + \tilde{b}_{10}(\theta)k_{10} + \tilde{b}_{11}(\theta)k_{11} + \tilde{b}_{12}(\theta)k_{12}\right),\end{aligned} \tag{6.12}$$

where $d_0(\theta)$, $d_1(\theta)$, $\tilde{b}_1(\theta)$, $\tilde{b}_2(\theta)$, $\tilde{b}_9(\theta)$, $\tilde{b}_{10}(\theta)$, $\tilde{b}_{11}(\theta)$ and $\tilde{b}_{12}(\theta)$ are weight polynomials of degree seven, obtained from the interpolation conditions, $\tilde{u}_i(t_i) = y_i$, $\tilde{u}_i(t_{i+1}) = y_{i+1}$, $\tilde{u}_i'(t_i) = k_1$, $\tilde{u}_i'(t_{i+1}) = k_2$, $\tilde{u}_i'(t_{i+c_9h_i}) = k_9$, $\tilde{u}_i'(t_{i+c_{10}h_i}) = k_{10}$, $\tilde{u}_i'(t_{i+c_{11}h_i}) = k_{11}$, and $\tilde{u}_i'(t_{i+c_{12}h_i}) = k_{12}$. The application of these interpolation conditions gives

$$d_0(\theta) = \frac{1}{2379157}(1500000000\theta^5 - 2250000000\theta^4 + 689550000\theta^3 + 30225000\theta^2$$

$$+47583140\theta + 2379157)(-1+\theta)^2,$$

$$d_1(\theta) = -\frac{1}{2379157}\theta^2(-4114971 + 676683140\theta - 3598875000\theta^2 + 6689550000\theta^3$$

$$-5250000000\theta^4 + 1500000000\theta^5),$$

$$\tilde{b}_1(\theta) = \frac{1}{1398594579921}\theta(576827250000000\theta^4 - 1162635500000000\theta^3$$

$$+740998886825000\theta^2 - 160345372818750\theta + 1398594579921)(-1+\theta)^2,$$

$$\tilde{b}_2(\theta) = \frac{1}{1398594579921}\theta^2(-1+\theta)(576827250000000\theta^4 - 1144673500000000\theta^3$$

$$+714055886825000\theta^2 - 141054900831250\theta + 883120980546),$$

$$\tilde{b}_9(\theta) = -\frac{500000}{110488971813759}\theta^2(256710000000\theta^3 - 504022850000\theta^2$$

$$+298349687600\theta - 4700220651)(-1+\theta)^2,$$

$$\tilde{b}_{10}(\theta) = \frac{15625}{21384962286534}\theta^2(1456920000000\theta^3 - 2661211400000\theta^2$$

$$+1351136688800\theta - 11988758061)(-1+\theta)^2,$$

$$\tilde{b}_{11}(\theta) = \frac{15625}{21384962286534}\theta^2(1456920000000\theta^3 - 1709548600000\theta^2$$

$$+399473888800\theta - 2695770819)(-1+\theta)^2,$$

$$\tilde{b}_{12}(\theta) = -\frac{500000}{110488971813759}\theta^2(256710000000\theta^3 - 266107150000\theta^2$$

$$+60433987600\theta - 403463109)(-1+\theta)^2.$$

We apply the above sixth order Hermite-Birkhoff interpolant (6.12) to the SWAVE problem (1.1). The plot of the scaled absolute defect is given in Figure 6.5. The location of the maximum defect is the same for almost all subintervals; it occurs at $\theta \approx 0.5$.

Similar to the 4^{th} order case, it can be shown that

$$\tilde{\delta}(t) = \tilde{u}_i'(t) - z_i'(t) + O(h_i^7) = d_1'(\theta)C_i h_i^6 + O(h_i^7), \tag{6.13}$$

where $d_1'(\theta)$ is a polynomial of degree six.

Thus, we can predict the location of the maximum defect since it will occur at the maximum of $d_1'(\theta) = -\frac{2}{2379157}\theta(-4114971+676683140-3598875000^2+6689550000^3-52500000000^4+15000000000^5)-1/2379157\theta^2(67668314-7197750000+20068650000^2-21000000000^3+7500000000^4)$. The maximum of the polynomial will occur where its derivative, $\frac{265482}{76747} - \frac{406009884}{2379157}\theta + \frac{4318650000}{2379157}\theta^2 - \frac{13379100000}{2379157}\theta^3 + \frac{15750000000}{2379157}\theta^4 - \frac{6300000000}{2379157}\theta^5)$, is zero $\Rightarrow \theta = 0.5$; see Figure 6.6.

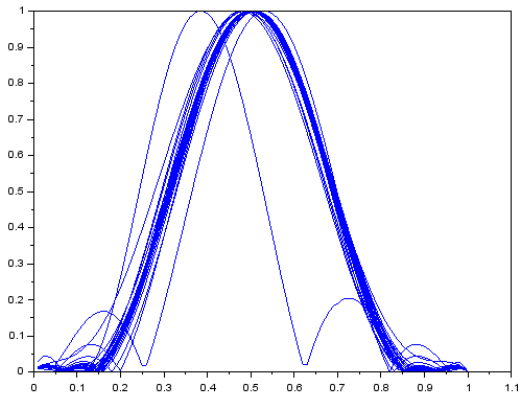

Figure 6.5: A plot of the absolute scaled defect obtained from applying the 6^{th} order Hermite-Birkhoff scheme with N=30 to the SWAVE test problem (1.1) with $\epsilon = 0.1$.

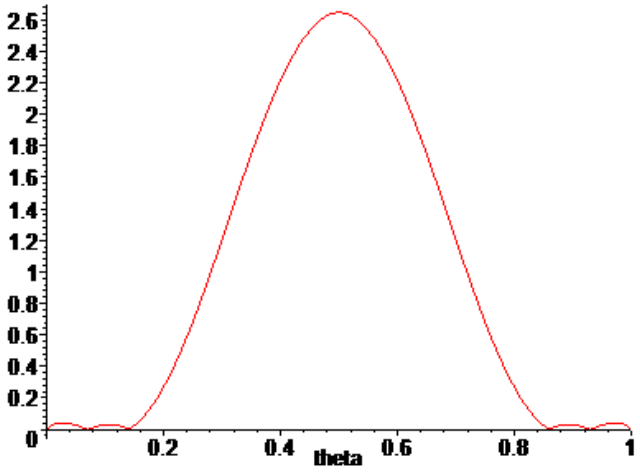

Figure 6.6: A plot of $d_1'(\theta)$ for the 6^{th} order Hermite-Birkhoff interpolant (6.13).

6.3 Derivation of a 13 Stage, Sixth Order ACDC CMIRK Scheme

We derive a sixth order ACDC CMIRK scheme that has the discrete MIRK563 (6.2) scheme embedded within it. The discrete scheme has the stage order vector SOV = (6, 6, 3, 3, 3). We will require all additional stages of the CMIRK scheme to have stage order 6, except for the sixth stage which we require to have stage order five. We also require the scheme to satisfy the 10 order conditions (associated with a sixth order

CMIRK scheme of stage order 3) as well as 7 out of the 8 order conditions associated with seventh order (that appear in the leading order term in the expansion of the local error for a sixth order scheme). The 10 sixth order conditions are ($b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)(xc^3 + \frac{v}{4}) = \frac{\theta^5}{20}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$, $b^T(\theta)c(Xc^3 + \frac{v}{4}) = \frac{\theta^6}{24}$, and $b^T(\theta)(X(Xc^3 + \frac{v}{4}) + \frac{v}{20}) = \frac{\theta^6}{120}$). The 8 seventh order conditions are:

$$b^T(\theta)\left(X\left(X\left(Xc^3 + \frac{v}{4}\right) + \frac{v}{20}\right) + \frac{v}{120}\right) - \frac{\theta^7}{840}, \tag{6.14}$$

$$b^T(\theta)\left(X\left(Xc^4 + \frac{v}{5}\right) + \frac{v}{30}\right) - \frac{\theta^7}{210}, \tag{6.15}$$

$$b^T(\theta)\left(X\left(c\left(Xc^3 + \frac{v}{4}\right)\right) + \frac{v}{24}\right) - \frac{\theta^7}{168}, \tag{6.16}$$

$$b^T(\theta)\left(Xc^5 + \frac{v}{6}\right) - \frac{\theta^7}{42}, \tag{6.17}$$

$$b^T(\theta)c\left(X\left(Xc^3 + \frac{v}{4}\right) + \frac{v}{20}\right) - \frac{\theta^7}{140}, \tag{6.18}$$

$$b^T(\theta)c\left(Xc^4 + \frac{v}{5}\right) - \frac{\theta^7}{35}, \tag{6.19}$$

$$b^T(\theta)c^2\left(Xc^3+\frac{v}{4}\right)-\frac{\theta^7}{28}, \tag{6.20}$$

$$b^T(\theta)c^6-\frac{\theta^7}{7} \quad . \tag{6.21}$$

We can rewrite several of the above order conditions as follows:

(By substitution of (6.15) and $C4$ into (6.14)), (6.14) $\Rightarrow b^T(\theta)XXC4=0$,

(By substitution of (6.17) and $C5$ into (6.15)), (6.15) $\Rightarrow b^T(\theta)XC5=0$,

(By substitution of (6.17) and $C4$ into (6.16)), (6.16) $\Rightarrow b^T(\theta)XcC4=0$,

(By substitution of $C6$ into (6.17)), (6.17) $\Rightarrow b^T(\theta)C6=0$,

(By substitution of (6.19) and $C4$ into (6.18)), (6.18) $\Rightarrow b^T(\theta)cXC4=0$,

(By substitution of (6.21) and $C5$ into (6.19)), (6.19) $\Rightarrow b^T(\theta)cC5=0$,

(By substitution of (6.21) and $C4$ into (6.20)), (6.20) $\Rightarrow b^T(\theta)c^2C4=0$,

where $C4$ is stage order four condition, $C4=Xc^3+\frac{v}{4}-\frac{c^4}{4}$, $C5$ is stage order five condition, $C5=Xc^4+\frac{v}{5}-\frac{c^5}{5}$, and $C6$ is stage order six condition, $C6=Xc^5+\frac{v}{6}-\frac{c^6}{6}$.

A less efficient CMIRK scheme for this case would have the number of stages equal to the number of order conditions that need to be satisfied. As mentioned above, in this case, it would appear that the $10+7=17$ order conditions will require the method to have 17 stages. Our goal is to derive a method with a smaller number of stages. Here we derive a family of sixth order, stage order three, ACDC CMIRK schemes with 13 stages (CMIRK1363) with SOV = (6, 6, 3, 3, 3, 5, 6, 6, 6, 6, 6, 6, 6). The Butcher tableau is

0	0	0	0	0	...	0	0	0	0	0	0	0
1	1	0	0	0	...	0	0	0	0	0	0	0
c_3	v_3	x_{31}	x_{32}	0	...	0	0	0	0	0	0	0
c_4	v_4	x_{41}	x_{42}	0	...	0	0	0	0	0	0	0
c_5	v_5	x_{51}	x_{52}	x_{53}	...	0	0	0	0	0	0	0
c_6	v_6	x_{61}	x_{62}	x_{63}	...	0	0	0	0	0	0	0
c_7	v_7	x_{71}	x_{72}	x_{73}	...	0	0	0	0	0	0	0
c_8	v_8	x_{81}	x_{82}	x_{83}	...	x_{87}	0	0	0	0	0	0
c_9	v_9	x_{91}	x_{92}	x_{93}	...	x_{97}	x_{98}	0	0	0	0	0
c_{10}	v_{10}	x_{101}	x_{102}	x_{103}	...	x_{107}	x_{108}	x_{109}	0	0	0	0
c_{11}	v_{11}	x_{111}	x_{112}	x_{113}	...	x_{117}	x_{118}	x_{119}	x_{1110}	0	0	0
c_{12}	v_{12}	x_{121}	x_{122}	x_{123}	...	x_{127}	x_{128}	x_{129}	x_{1210}	x_{1211}	0	0
c_{13}	v_{13}	x_{131}	x_{132}	x_{133}	...	x_{137}	x_{138}	x_{139}	x_{1310}	x_{1311}	x_{1312}	0
		$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	...	$b_7(\theta)$	$b_8(\theta)$	$b_9(\theta)$	$b_{10}(\theta)$	$b_{11}(\theta)$	$b_{12}(\theta)$	$b_{13}(\theta)$

, (6.22)

where the coefficients for the first five rows of the tableau are given in the tableau of the embedded discrete (MIRK563) scheme (6.2).

As mentioned earlier, we will assume the scheme has the SOV = (6, 6, 3, 3, 3, 5, 6, ... , 6). Thus, the 6^{th} stage is required to satisfy the stage order five conditions, and

the 7^{th} to 13^{th} stages are required to satisfy the stage order six conditions. Imposing the appropriate stage order conditions on stages six, seven, ..., thirteen, we get

x_{61}, x_{62}, x_{63}, x_{64} and x_{65} in terms of c_6 and v_6,

x_{71}, x_{72}, x_{73}, x_{74}, x_{75} and x_{76} in terms of c_7 and v_7,

x_{81}, x_{82}, x_{83}, x_{84}, x_{86} and x_{87} in terms of c_8, v_8 and x_{85},

x_{91}, x_{92}, x_{96}, x_{97}, x_{98} and v_9 in terms of c_9, x_{93}, x_{94} and x_{95},

x_{101}, x_{102}, x_{106}, x_{107}, x_{108}, and x_{109} in terms of c_{10}, v_{10}, x_{103}, x_{104} and x_{105},

x_{111}, x_{112}, x_{116}, x_{117}, x_{118}, and x_{119} in terms of c_{11}, v_{11}, x_{113}, x_{114}, x_{115} and x_{1110},

x_{121}, x_{122}, x_{126}, x_{127}, x_{128}, and x_{129} in terms of c_{12}, v_{12}, x_{123}, x_{124}, x_{125}, x_{1210} and x_{1211},

x_{131}, x_{132}, x_{136}, x_{137}, x_{138}, and x_{139} in terms of c_{13}, v_{13}, x_{133}, x_{134}, x_{135}, x_{1310}, x_{1311} and x_{1312}.

In addition to the ten sixth order conditions, the seven seventh order conditions which we choose to be satisfied are: (6.15) - (6.21). That is, we choose to leave (6.14) unsatisfied.

Since the third, fourth and fifth stages have only stage order three and the remaining stages have at least stage order five, the third, fourth, and fifth positions of $C4$ and $C5$ are the only ones that are non-zero. Therefore, in order to satisfy the order conditions $b^T(\theta)C4 = 0$, $b^T(\theta)cC4 = 0$, $b^T(\theta)C5 = 0$, $b^T(\theta)cC5 = 0$, and $b^T(\theta)c^2C4 = 0$, we choose the weight polynomials $b_3(\theta)$, $b_4(\theta)$, and $b_5(\theta)$ to be identically zero.

Since the third, fourth and fifth stages have only stage order three, and the sixth stage has only stage order five, the third, fourth, fifth, and sixth positions of $C6$ are

non-zero; since the rest of the stages have stage order six, the remaining positions of $C6$ are zero. Therefore, in order to satisfy the order condition $b^T(\theta)C6 = 0$, we choose the weight polynomial $b_6(\theta)$ to be identically zero.

In order for the remaining order conditions $b^T(\theta)XC4 = 0$, $b^T(\theta)XC5 = 0$, $b^T(\theta)XcC4 = 0$, $b^T(\theta)cXC4 = 0$ to be satisfied, we need to choose $b_7(\theta)$, $b_8(\theta)$, and $x_{i,3}$, $x_{i,4}$, and $x_{i,5}$ equal zero, where $i = 9, ..., 13$.

We are then left with the seven quadrature order conditions $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, and $b^T(\theta)c^6 = \frac{\theta^7}{7}$. We then require the remaining weight polynomials to satisfy them; that is, $b_1(\theta)$, $b_2(\theta)$, $b_9(\theta)$, ... ,$b_{13}(\theta)$, (seven polynomials) are used to satisfy the seven quadrature conditions.

After solving the 17 order conditions, and choosing values for the 21 free parameters as below, we obtain an example of such a method. The structure of the tableau is as in (6.22). Recall that the coefficients for the first five rows of the tableau are given in the tableau of the embedded discrete (MIRK563) scheme (6.2).The remaining coefficients are

$$c_6 = \frac{1}{4}, \quad v_6 = \frac{1}{4}, \quad x_{61} = \frac{51}{1024}, \quad x_{62} = -\frac{15}{1024}, \quad x_{63} = \frac{49}{1536} + \frac{21}{1024}\sqrt{21},$$

$$x_{64} = \frac{49}{1536} - \frac{21}{1024}\sqrt{21}, \quad x_{65} = -\frac{19}{192},$$

$$c_7 = \frac{1}{100}, \quad v_7 = \frac{1}{100}, \quad x_{71} = \frac{1102868613}{125000000000}, \quad x_{72} = -\frac{30295683}{62500000000},$$

$$x_{73} = \frac{289590147}{1250000000000}\sqrt{21} - \frac{2145735823}{1250000000000},$$

$$x_{74} = -\frac{289590147}{1250000000000}\sqrt{21} - \frac{2145735823}{1250000000000},$$

$$x_{75} = -\frac{103425487}{31250000000}, \quad x_{76} = -\frac{15580323}{9765625000},$$

$$c_8 = \frac{3}{100}, \quad v_8 = \frac{3}{100}, \quad x_{81} = \frac{9220007819}{50000000000}, \quad x_{82} = -\frac{12491798419}{14850000000000},$$

$$x_{83} = \frac{10588528959211}{290843750000000} + \frac{71579882096219}{6980250000000000}\sqrt{21},$$

$$x_{84} = \frac{10588528959211}{290843750000000} - \frac{71579882096219}{6980250000000000}\sqrt{21},$$

$$x_{85} = 0, \quad x_{86} = -\frac{9836391797}{140625000000}, \quad x_{87} = -\frac{824493889}{4422686400},$$

$$c_9 = \frac{1}{20}, \quad v_9 = -\frac{1011}{377840000}, \quad x_{91} = \frac{12766043}{755680000}, \quad x_{92} = \frac{5350837}{13062231072000},$$

$$x_{93} = 0, \quad x_{94} = 0, \quad x_{95} = 0,$$

$$x_{96} = \frac{8690353}{224436960000}, \quad x_{97} = -\frac{13420175}{1077297408}, \quad x_{98} = \frac{14683675}{322524224},$$

$$c_{10} = \frac{1}{2}, \quad v_{10} = \frac{1}{2}, \quad x_{101} = -\frac{155431}{72}, \quad x_{102} = -\frac{988673}{13136904},$$

$$x_{103} = 0, \quad x_{104} = 0, \quad x_{105} = 0,$$

$$x_{106} = -\frac{111845}{19008}, \quad x_{107} = \frac{246015625}{57024}, \quad x_{108} = -\frac{252578125}{76824}, \quad x_{109} = \frac{2076125}{1824},$$

$$c_{11} = \frac{3}{4}, \quad v_{11} = \frac{3}{4}, \quad x_{111} = -\frac{67111}{48}, \quad x_{112} = -\frac{29913}{324368},$$

$$x_{113} = 0, \quad x_{114} = 0, \quad x_{115} = 0,$$

$$x_{116} = -\frac{113959}{33792}, \quad x_{117} = \frac{94140625}{33792}, \quad x_{118} = -\frac{864453125}{409728}, \quad x_{119} = \frac{7058125}{9728},$$

$$x_{1110} = 0,$$

$$c_{12} = \frac{3}{5}, \quad v_{12} = \frac{3}{5}, \quad x_{121} = -\frac{801451}{375}, \quad x_{122} = -\frac{220567}{2534125},$$

$$x_{123} = 0, \quad x_{124} = 0, \quad x_{125} = 0,$$

$$x_{126} = -\frac{92203}{16500}, \quad x_{127} = \frac{563225}{132}, \quad x_{128} = -\frac{10386350}{3201}, \quad x_{129} = \frac{212941}{190},$$

$$x_{1210} = 0, \quad x_{1211} = 0,$$

$$c_{13} = \frac{1}{5}, \quad v_{13} = \frac{1}{5}, \quad x_{131} = -\frac{1102486}{1125}, \quad x_{132} = -\frac{6281114}{205264125},$$

$$x_{133} = 0, \quad x_{134} = 0, \quad x_{135} = 0, \quad x_{136} = -\frac{216617}{74250},$$

$$x_{137} = \frac{3494825}{1782}, \quad x_{138} = -\frac{14395100}{9603}, \quad x_{139} = \frac{148421}{285}, \quad x_{1310} = 0, \quad x_{1311} = 0,$$

$$x_{1312} = 0,$$

and where

$$b_1(\theta) = \frac{1}{756}\theta(756 - 117180\theta + 66332\theta^2 - 178269\theta^3 + 248472\theta^4 - 173600\theta^5 + 48000\theta^6),$$

$$b_2(\theta) = \frac{1}{4256}\theta^2(-126 + 2520\theta - 14693\theta^2 + 35784\theta^3 - 39200\theta^4 + 16000\theta^5),$$

$$b_3(\theta) = 0,$$

$$b_4(\theta) = 0,$$

$$b_5(\theta) = 0,$$

$$b_6(\theta) = 0,$$

$$b_7(\theta) = 0,$$

$$b_8(\theta) = 0,$$

$$b_9(\theta) = -\frac{8000}{829521}\theta^2(-1890 + 13860\theta - 40845\theta^2 + 59556\theta^3 - 42700\theta^4 + 12000\theta^5),$$

$$b_{10}(\theta) = -\frac{4}{567}\theta^2(-378 + 7308\theta - 38787\theta^2 + 80556\theta^3 - 72800\theta^4 + 24000\theta^5),$$

$$b_{11}(\theta) = -\frac{64}{14553}\theta^2(-126 + 2492\theta - 14091\theta^2 + 32508\theta^3 - 32900\theta^4 + 12000\theta^5),$$

$$b_{12}(\theta) = \frac{125}{33264}\theta^2(-630 + 12320\theta - 67515\theta^2 + 147672\theta^3 - 140000\theta^4 + 48000\theta^5),$$

$$b_{13}(\theta) = \frac{125}{66528}\theta^2(-1890 + 32760\theta - 125895\theta^2 + 209496\theta^3 - 162400\theta^4 + 48000\theta^5).$$

We can predict the shape of the defect for this scheme and the location of the maximum defect for each subinterval. The defect will be a multiple of the derivative of the one unsatisfied order condition (6.14); see Figure 6.7. The maximum defect on each subinterval will occur where the second derivative of the unsatisfied order condition is zero. The maximum is at $\theta = 0.5416525443$.

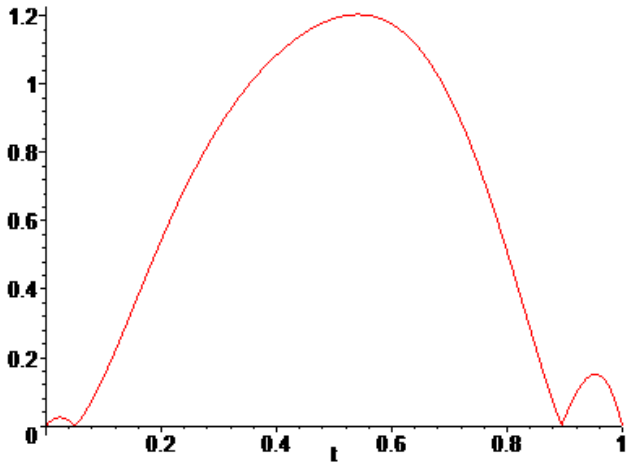

Figure 6.7: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition for CMIRK1363 (6.22).

6.4 Direct Derivation of an 11 Stage Sixth Order ACDC CMIRK Scheme

Here we derive a stage order 3, sixth order ACDC CMIRK scheme that does not have the discrete MIRK563 (6.2) scheme embedded within it. The application of the stage order three conditions gives 10 order conditions for 6^{th} order. There are also 8 order conditions associated with 7^{th} order that appear in the local error expansion for this scheme. We require the scheme to satisfy the ten sixth order conditions ($b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)(xc^3+\frac{v}{4}) = \frac{\theta^5}{20}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)(Xc^4+\frac{v}{5}) = \frac{\theta^6}{30}$, $b^T(\theta)c(Xc^3+\frac{v}{4}) = \frac{\theta^6}{24}$, and $b^T(\theta)(X(Xc^3+\frac{v}{4})+\frac{v}{20}) = \frac{\theta^6}{120}$) as well as seven out of the eight seventh order conditions (6.14) - (6.21). The seven 7^{th} order conditions we choose to satisfy are (6.15) - (6.21). That is, (6.14) will remain unsatisfied.

A less efficient CMIRK scheme for this case would have a number of stages equals to the number of order conditions, i.e., 17. Our goal is to derive a method with fewer stages. Here we derive a family of sixth order, stage order three CMIRK schemes with 11 stages, CMIRK1163, for which we impose stage order six on all but the 3^{rd}, 4^{th} 5^{th} and 6^{th} stages. The corresponding Butcher tableau is

0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0
c_3	v_3	x_{31}	x_{32}	0	0	0	0	0	0	0	0	0
c_4	v_4	x_{41}	x_{42}	x_{43}	0	0	0	0	0	0	0	0
c_5	v_5	x_{51}	x_{52}	x_{53}	x_{54}	0	0	0	0	0	0	0
c_6	v_6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	0	0	0	0	0	0
c_7	v_7	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	0	0	0	0	0
c_8	v_8	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	0	0	0	0
c_9	v_9	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	0	0	0
c_{10}	v_{10}	x_{101}	x_{102}	x_{103}	x_{104}	x_{105}	x_{106}	x_{107}	x_{108}	x_{109}	0	0
c_{11}	v_{11}	x_{111}	x_{112}	x_{113}	x_{114}	x_{115}	x_{116}	x_{117}	x_{118}	x_{119}	x_{1110}	0
		$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$	$b_5(\theta)$	$b_6(\theta)$	$b_7(\theta)$	$b_8(\theta)$	$b_9(\theta)$	$b_{10}(\theta)$	$b_{11}(\theta)$

, (6.23)

The stage order vector for this scheme will be (6,6,3,4,5,5,6,...6). After imposing the appropriate stage order conditions on the 11 stages, this gives

v_3, x_{31}, and x_{32} in terms of c_3,

v_4, x_{41}, x_{42}, and x_{43} in terms of c_4,

v_5, x_{51}, x_{52}, x_{53} and x_{54} in terms of c_5,

x_{61}, x_{62}, x_{63}, x_{64} and x_{65} in terms of c_6 and v_6,

v_7, x_{71}, x_{72}, x_{74}, x_{75} and x_{76} in terms of c_7, and x_{73},

v_8, x_{81}, x_{82}, x_{85}, x_{86} and x_{87} in terms of c_8, x_{83}, and x_{84},

x_{91}, x_{92}, x_{95}, x_{96}, x_{97}, and x_{98} in terms of c_9, v_9, x_{93}, and x_{94},

x_{102}, x_{105}, x_{106}, x_{107}, x_{108}, and x_{109} in terms of c_{10}, v_{10}, x_{101}, x_{103}, and x_{104},

x_{111}, x_{112}, x_{115}, x_{116}, x_{117}, and x_{118} in terms of c_{11}, v_{11}, x_{113}, x_{114},x_{119} and x_{1110}.

The scheme has $b_3(\theta), b_4(\theta), b_5(\theta), b_6(\theta)$ identically equal to zero. With $b_3(\theta) = b_4(\theta) = b_5(\theta) = b_6(\theta) = 0$, and applying some extra conditions on some of the free coefficients, namely, $x_{73} = x_{83} = x_{84} = x_{93} = x_{94} = x_{103} = x_{104} = x_{113} = x_{114} = 0$, and requiring c_7 to satisfy $x_{73} \times C5_3 + x_{74} \times C5_4 = 0$, we are able to satisfy the ten non-quadrature order conditions, $b^T(\theta)(xc^3 + \frac{v}{4}) = \frac{\theta^5}{20}$, $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$, $b^T(\theta)c(Xc^3 + \frac{v}{4}) = \frac{\theta^6}{24}$, $b^T(\theta)(X(Xc^3 + \frac{v}{4}) + \frac{v}{20}) = \frac{\theta^6}{120}$, $b^T(\theta)\left(X\left(Xc^4 + \frac{v}{5}\right) + \frac{v}{30}\right) - \frac{\theta^7}{210}$, $b^T(\theta)\left(X\left(c\left(Xc^3 + \frac{v}{4}\right)\right) + \frac{v}{24}\right) - \frac{\theta^7}{168}$, $b^T(\theta)\left(Xc^5 + \frac{v}{6}\right) - \frac{\theta^7}{42}$, $b^T(\theta)c\left(X\left(Xc^3 + \frac{v}{4}\right) + \frac{v}{20}\right) - \frac{\theta^7}{140}$, $b^T(\theta)c\left(Xc^4 + \frac{v}{5}\right) - \frac{\theta^7}{35}$, $b^T(\theta)c^2\left(Xc^3 + \frac{v}{4}\right) - \frac{\theta^7}{28}$. This implies that the only remaining order conditions will be the seven quadrature conditions $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, and $b^T(\theta)c^6 = \frac{\theta^7}{7}$. We then require that the associated weight polynomials and the remaining free coefficients, c_3, c_4, c_5, c_6, v_6, c_8, $c_9, v_9, c_{10}, v_{10}, x_{101}, c_{11}, v_{11}, x_{119}$, and x_{1110}, be chosen to satisfy these conditions.

After solving the 17 order conditions, and choosing values for the 15 free parameters as given below, we obtain an example of such a method.

The coefficients values are:

$$c_3 = \frac{1}{4}, \quad v_3 = \frac{5}{32}, \quad x_{31} = \frac{9}{64}, \quad x_{32} = -\frac{3}{64},$$

$$c_4 = \frac{3}{4}, \quad v_4 = \frac{81}{128}, \quad x_{41} = \frac{3}{256}, \quad x_{42} = -\frac{21}{256}, \quad x_{43} = \frac{3}{16},$$

$$c_5 = \frac{1}{2}, \quad v_5 = \frac{1}{2}, \quad x_{51} = \frac{1}{24}, \quad x_{52} = -\frac{1}{24}, \quad x_{53} = \frac{1}{6}, \quad x_{54} = -\frac{1}{6},$$

$$c_6 = \frac{1}{7}, \quad v_6 = \frac{1}{7}, \quad x_{61} = \frac{3365}{50421}, \quad x_{62} = -\frac{793}{50421}, \quad x_{63} = \frac{2752}{50421},$$

$$x_{64} = -\frac{1280}{50421}, \quad x_{65} = -\frac{1348}{16807},$$

$$c_7 = \frac{1}{2} + \frac{1}{14}\sqrt{13}, \quad v_7 = \frac{1339}{33614}\sqrt{13} + \frac{4657}{33614}, \quad x_{71} = -\frac{117}{33614}\sqrt{13} + \frac{1143}{33614},$$

$$x_{72} = -\frac{117}{33614}\sqrt{13} - \frac{801}{67228}, \quad x_{73} = 0, \quad x_{74} = 0,$$

$$x_{75} = \frac{648}{16807}\sqrt{13} + \frac{12312}{84035}, \quad x_{76} = \frac{27}{140},$$

$$c_8 = \frac{3}{7}, \quad v_8 = \frac{4743}{16807} + \frac{1440}{16807}\sqrt{13}, \quad x_{81} = \frac{1720}{50421} - \frac{328}{50421}\sqrt{13},$$

$$x_{82} = -\frac{1024}{50421} - \frac{328}{50421}\sqrt{13}, \quad x_{83} = 0, \quad x_{84} = 0,$$

$$x_{85} = \frac{1536}{84035} - \frac{3072}{218491}\sqrt{13}, \quad x_{86} = \frac{4}{35}, \quad x_{87} = -\frac{16}{273}\sqrt{13},$$

$$c_9 = \frac{2}{7}, \quad v_9 = \frac{2}{7}, \quad x_{91} = -\frac{1831}{111132} + \frac{1411}{111132}\sqrt{13}, \quad x_{92} = -\frac{23657}{889056} + \frac{1411}{889056}\sqrt{13},$$

$$x_{93} = 0, \quad x_{94} = 0, \quad x_{95} = -\frac{1016}{1715} + \frac{11288}{66885}\sqrt{13},$$

$$x_{96} = \frac{7433}{15120} - \frac{1411}{15120}\sqrt{13}, \quad x_{97} = -\frac{1411}{2268} + \frac{1411}{9828}\sqrt{13}, \quad x_{98} = \frac{4633}{6048} - \frac{1411}{6048}\sqrt{13},$$

$$c_{10} = \frac{4}{7}, \quad v_{10} = \frac{4}{7}, \quad x_{101} = 0, \quad x_{102} = -\frac{803}{15120} + \frac{37}{15120}\sqrt{13},$$

$$x_{103} = 0, \quad x_{104} = 0, \quad x_{105} = -\frac{176}{105} + \frac{592}{1365}\sqrt{13},$$

$$x_{106} = -\frac{461}{2520} + \frac{37}{360}\sqrt{13}, \quad x_{107} = \frac{185}{54} - \frac{703}{702}\sqrt{13}, \quad x_{108} = \frac{1045}{336} - \frac{37}{48}\sqrt{13},$$

$$x_{109} = -\frac{971}{210} + \frac{37}{30}\sqrt{13}, \quad c_{11} = \frac{1}{8}, \quad v_{11} = \frac{1}{8}, \quad x_{111} = \frac{948535}{42467328} + \frac{277487}{42467328}\sqrt{13},$$

$$x_{112} = -\frac{3791095}{339738624} + \frac{277487}{339738624}\sqrt{13}, \quad x_{113} = 0, \quad x_{114} = 0,$$

$$x_{115} = -\frac{455}{2048} + \frac{277487}{3194880}\sqrt{13},$$

$$x_{116} = \frac{13596863}{56623104} - \frac{13596863}{283115520}\sqrt{13}, \quad x_{117} = -\frac{13596863}{42467328} + \frac{13596863}{184025088}\sqrt{13},$$

$$x_{118} = \frac{32958527}{113246208} - \frac{13596863}{113246208}\sqrt{13}, \quad x_{119} = 0, \quad x_{1110} = 0,$$

The weight polynomials are

$$b_1(\theta) = -\frac{1}{25920}\theta(-7 + \sqrt{13})(-46830\theta - 5590900\theta^5 + 762342\theta^4 - 541275\theta^3$$

$$+214690\theta^2 + 164640\theta^6 + 5040 + 720\sqrt{13} - 5970\theta\sqrt{13} + 22710\theta^2\sqrt{13} - 43260\theta^3\sqrt{13}$$

$$+39690\theta^4\sqrt{13} - 13720\theta^5\sqrt{13}),$$

$$b_2(\theta) = \frac{1}{907200}\theta^2(7+\sqrt{13})(-5040 - 720\sqrt{13} + 59080\theta + 7480\theta\sqrt{13} - 277725\theta^2$$

$$-28455\theta^2\sqrt{13} + 643860\theta^3 + 46452\theta^3\sqrt{13} - 734020\theta^4 - 27440\theta^4\sqrt{13} + 329280\theta^5),$$

$$b_3(\theta) = 0,$$

$$b_4(\theta) = 0,$$

$$b_5(\theta) = 0,$$

$$b_6(\theta) = 0,$$

$$b_7(\theta) = -\frac{343}{754920}\theta^2(-115 + 33\sqrt{13})(-720 + 7960\theta - 34065\theta^2 + 69216\theta^3 - 66150\theta^4$$

$$+23520\theta^5),$$

$$b_8(\theta) = -343146880(-1+\sqrt{13})(-1680 - 240\sqrt{13} + 18200\theta + 2280\theta\sqrt{13} - 75495\theta^2$$

$$-7365\theta^2\sqrt{13} + 147756\theta^3 + 9324\theta^3\sqrt{13} - 136220\theta^4 - 3920\theta^4\sqrt{13} + 47040\theta^5)\theta^2,$$

$$b_9(\theta) = \frac{343}{21600}\theta^2(-3+\sqrt{13})(-1260 - 180\sqrt{13} + 12670\theta + 1570\theta\sqrt{13} - 47250\theta^2$$

$$-4395\theta^2\sqrt{13} + 84210\theta^3 + 4998\theta^3\sqrt{13} - 72030\theta^4 - 1960\theta^4\sqrt{13} + 23520\theta^5),$$

$$b_{10}(\theta) = \frac{343}{216000}\theta^2(1+\sqrt{13})(-630 - 90\sqrt{13} + 7070\theta + 890\theta\sqrt{13} - 30975\theta^2$$

$$-3090\theta^2\sqrt{13}+64890\theta^3+4326\theta^3\sqrt{13}-64190\theta^4-1960\theta 4\sqrt{13}+23520\theta^5),$$

$$b_{11}(\theta)=\frac{32768}{93578625}\theta^2(-21+4\sqrt{13})(-5040-720\sqrt{13}+35560\theta+4120\theta\sqrt{13}$$

$$-108675\theta^2-9345\theta^2\sqrt{13}+170520\theta^3+9408\theta^3\sqrt{13}-133770\theta^4-3430\theta^4\sqrt{13}+41160\theta^5).$$

We can predict the shape of the defect for this scheme and the location of the maximum defect for each subinterval. The defect will be the derivative of the unsatisfied order condition, (6.14); see Figure 6.8. The maximum defect of each subinterval will occur where the second derivative of the unsatisfied order condition is zero. It will be at $\theta = 0.8978781493$.

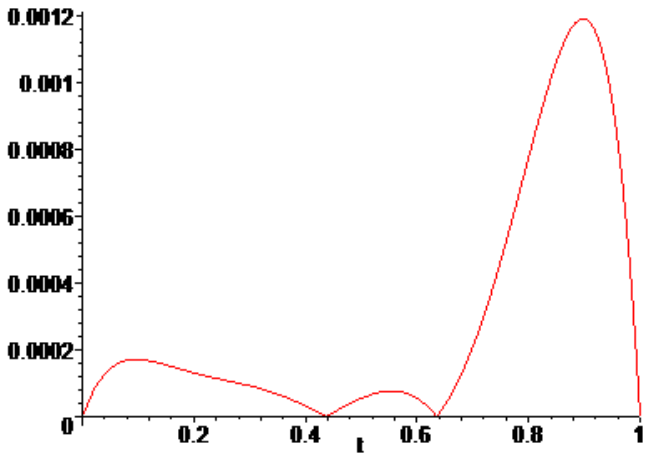

Figure 6.8: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition (6.14) for CMIRK1163 (6.23).

6.5 Derivations of Continuous Generalized ACDC MIRK (CGMIRK) Schemes

Here we derive sixth order continuous generalized ACDC MIRK (CGMIRK) schemes that have discrete GMIRK [13] schemes embedded within them. As mentioned earlier in this thesis, GMIRK schemes are extensions of MIRK schemes to allow them to have a stage order higher than three.

6.5.1 ACDC CGMIRK1064

We derive a sixth order continuous ACDC CGMIRK scheme that has the discrete GMIRK564 [13] scheme embedded within it. The discrete scheme has the stage order vector SOV = (4, 4, 4, 4, 4), and the tableau:

$$\begin{array}{c|c|ccccc|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & \\ \frac{1}{3} & -\frac{5}{27} & \frac{4}{27} & \frac{1}{27} & \frac{1}{3} & 0 & 0 & \\ \frac{2}{3} & \frac{8}{27} & \frac{2}{27} & -\frac{1}{27} & \frac{1}{3} & 0 & 0 & \\ \frac{1}{2} & -\frac{5}{8} & \frac{25}{128} & \frac{11}{128} & \frac{81}{128} & \frac{27}{128} & 0 & \\ \hline & & \frac{11}{120} & \frac{11}{120} & \frac{27}{40} & \frac{27}{40} & -\frac{8}{15} & \end{array} . \tag{6.24}$$

We will require all additional stages of the CGMIRK scheme to have stage order 6, except for the sixth stage which we will require to have stage order five. To obtain a 6^{th} order CGMIRK method with the ACDC property, we first require it to satisfy all the order conditions up to and including the sixth order. As well, we need it to

satisfy all the order conditions for 7^{th} order except one. The application of stage order four to the order conditions leaves 7 order conditions for 6^{th} order. These are $b(\theta)^T e = \theta$, $b(\theta)^T c = \frac{\theta^2}{2}$, $b(\theta)^T c^2 = \frac{\theta^3}{3}$, $b(\theta)^T c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, $b^T(\theta)(Xc^4 + \frac{v}{5}) = \frac{\theta^6}{30}$. There are 4 order conditions for order 7. These are:
$b^T(\theta)c^6 = \frac{\theta^7}{7}$, $b^T(\theta)\big(Xc^5 + \frac{v}{6}\big) = \frac{\theta^7}{42}$, $b^T(\theta)\big(X(Xc^4 + \frac{v}{5}) + \frac{v}{30}\big) = \frac{\theta^7}{210}$, $b^T(\theta)c\big(Xc^4 + \frac{v}{5}\big) = \frac{\theta^7}{35}$.

The 6^{th} stage is required to satisfy the stage order five conditions and the 7^{th} to 10^{th} stages are required to satisfy the stage order six conditions. This gives

x_{61}, x_{62}, x_{63}, x_{64} and x_{65} in terms of c_6 and v_6,

x_{71}, x_{72}, x_{73}, x_{74}, x_{75} and x_{76} in terms of c_7, and v_7,

x_{82}, x_{83}, x_{84}, x_{85}, x_{86} and x_{87} in terms of c_8, v_8, and x_{81},

x_{93}, x_{94}, x_{95}, x_{96}, x_{97}, and x_{98} in terms of c_9, v_9, x_{91}, and x_{92},

x_{104}, x_{105}, x_{106}, x_{107}, x_{108}, and x_{109} in terms of c_{10}, v_{10}, x_{101}, x_{102}, and x_{103}.

We then require the scheme to satisfy the 7 sixth order conditions plus 3 out of the 4 seventh order conditions. There are thus four ways to derive an ACDC CGMIRK1064. One way is to require the scheme to satisfy all of the seventh order conditions except $b^T(\theta)\left(X\left(Xc^4 + \frac{v}{5}\right) + \frac{v}{30}\right) = \frac{\theta^7}{210}$.

This will require the scheme to have ten stages. After solving the 10 order conditions, and choosing values for the free parameters as indicated below, we obtain an example of such a method, which has the tableau:

0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0
$\frac{1}{3}$	$-\frac{5}{27}$	$\frac{4}{27}$	$\frac{1}{27}$	$\frac{1}{3}$	0	0	0	0	0	0	0
$\frac{2}{3}$	$\frac{8}{27}$	$\frac{2}{27}$	$-\frac{1}{27}$	$\frac{1}{3}$	0	0	0	0	0	0	0
$\frac{1}{2}$	$-\frac{5}{8}$	$\frac{25}{128}$	$\frac{11}{128}$	$\frac{81}{128}$	$\frac{27}{128}$	0	0	0	0	0	0
c_6	v_6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	0	0	0	0	0
c_7	v_7	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	0	0	0	0
c_8	v_8	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	0	0	0
c_9	v_9	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	0	0
c_{10}	v_{10}	x_{101}	x_{102}	x_{103}	x_{104}	x_{105}	x_{106}	x_{107}	x_{108}	x_{109}	0
		$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$	$b_5(\theta)$	$b_6(\theta)$	$b_7(\theta)$	$b_8(\theta)$	$b_9(\theta)$	$b_{10}(\theta)$

(6.25)

The coefficients for the first five rows of the tableau are given in the tableau of the embedded discrete (GMIRK564) scheme (6.24). The remaining coefficients are

$$c_6 = \frac{7}{9}, \quad v_6 = \frac{7}{9}, \quad x_{61} = \frac{749}{26244}, \quad x_{62} = -\frac{2023}{26244}, \quad x_{63} = \frac{91}{2916},$$

$$x_{64} = -\frac{497}{2916}, \quad x_{65} = \frac{1232}{6561}$$

$$c_7 = \frac{1}{7}, \quad v_7 = \frac{1}{7}, \quad x_{71} = \frac{211927}{3294172}, \quad x_{72} = \frac{1441}{470596},$$

$$x_{73} = \frac{82053}{470596}, \quad x_{74} = \frac{170019}{470596},$$

$$x_{75} = -\frac{223984}{588245}, \quad x_{76} = -\frac{1830519}{8235430},$$

$$c_8 = \frac{1}{8}, \quad v_8 = \frac{1}{8}, \quad x_{81} = 0, \quad x_{82} = -\frac{70441}{2097152},$$

$$x_{83} = -\frac{4925907}{8388608}, \quad x_{84} = -\frac{2174067}{2883584},$$

$$x_{85} = \frac{12859}{16384}, \quad x_{86} = \frac{26499879}{83886080}, \quad x_{87} = \frac{252827701}{922746880},$$

$$c_9 = \frac{1}{99}, \quad v_9 = \frac{1}{99}, \quad x_{91} = 0, \quad x_{92} = 0,$$

$$x_{93} = \frac{36096876011}{619904625120}, \quad x_{94} = \frac{36916530077}{1108079517402}, \quad x_{95} = -\frac{136481169608}{2615222637225},$$

$$x_{96} = -\frac{59028342107}{3596977454400}, \quad x_{97} = -\frac{168197679098267}{613705578868800}, \quad x_{98} = \frac{80253382221824}{319580206268895},$$

$$c_{10} = \frac{9}{100}, \quad v_{10} = \frac{9}{100}, \quad x_{101} = 0, \quad x_{102} = 0,$$

$$x_{103} = 0, \quad x_{104} = \frac{27100294231563}{178750000000000}, \quad x_{105} = -\frac{5183586602831}{37890625000000},$$

$$x_{106} = -\frac{19265735985937623}{178600000000000000}, \quad x_{107} = \frac{22323527177885661}{50600000000000000},$$

$$x_{108} = -\frac{63914817644}{149169921875}, \quad x_{109} = \frac{885639052363337613}{11021140000000000000},$$

The weight polynomials are

$$b_1(\theta) = \frac{1}{35984685240}\theta(35984685240 - 22874262008846\theta + 3672464214886 0\theta^2$$

$$-2561371191211165\theta^3 + 7938637953902760\theta^4 - 8860475609440080\theta^5$$

$$+3236971152096000\theta^6),$$

$$b_2(\theta) = \frac{1}{26742618580860}\theta^2(-1717162713 + 1445830596850\theta - 253812691183500\theta^2$$

$$+1785795700939500\theta^3 - 506604289463340\theta^4 + 358768000008000\theta^5),$$

$$b_3(\theta) = 0,$$

$$b_4(\theta) = \frac{7371}{951975800}\theta^2(-54 + 4540\theta - 79185\theta^2 + 550236\theta^3 - 1516792\theta^4 + 950400\theta^5),$$

$$b_5(\theta) = 0,$$

$$b_6(\theta) = -\frac{59049}{6663830600}\theta^2(-54 + 4540\theta - 79185\theta^2 + 550236\theta^3 - 1516792\theta^4 + 950400\theta^5),$$

$$b_7(\theta) = \frac{16807}{14582365304400}\theta^2(-3747257262 + 3000535657700\theta - 415550646220500\theta^2$$

$$+1929992247241800\theta^3 - 2443346828907600\theta^4 + 940436529120000\theta^5),$$

$$b_8(\theta) = -\frac{16384}{318364506915}\theta^2(-183648276 + 1458328054000\theta - 193403520735000\theta^2$$

$$+8493863687160000\theta^3 - 1058431837818000\theta^4 + 40484153160000\theta^5),$$

$$b_9(\theta) = -\frac{313826716467}{154453479314789200}\theta^2(-32791878 + 6150012300\theta - 453275269500\theta^2$$

$$+14413590222000\theta^3 - 1622956164400\theta^4 + 59515728000\theta^5),$$

$$b_{10}(\theta) = \frac{12500000000}{2395947606140907}\theta^2(-1570548+1214648600\theta-139072399500\theta^2+5499719772000\theta^3$$

$$-66438903540^4 + 250948368080^5).$$

We can predict the shape of the defect for this scheme and the location of the maximum defect for each subinterval. The defect will be the derivative of the unsatisfied order condition (6.14); see Figure 6.9. The maximum defect of each subinterval will occur where the second derivative of this unsatisfied order condition is zero. It will be at $\theta = 0.6881326764$.

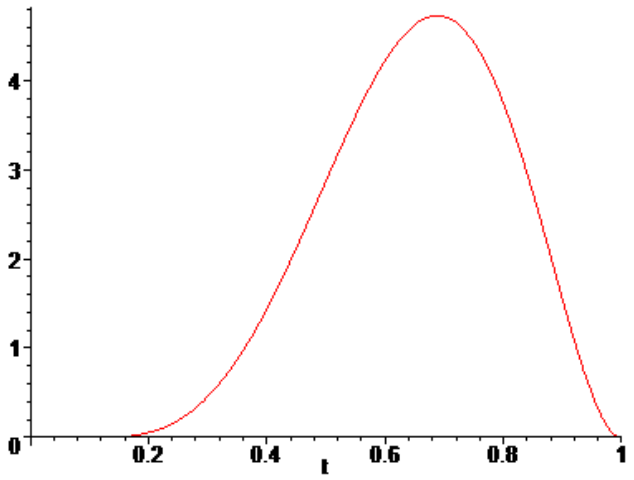

Figure 6.9: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition (6.14) for CGMIRK1064.

6.5.2 ACDC CGMIRK765

We derive a sixth order continuous generalized ACDC CGMIRK scheme that has the discrete GMIRK565 [13] scheme embedded within it. The discrete scheme has

the stage order vector SOV = (5, 5, 5, 5, 5), and the tableau:

$$\begin{array}{c|c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & -\frac{79}{625} & \frac{52}{625} & \frac{2}{625} & \frac{14}{75} & \frac{4}{75} & 0 \\
\frac{4}{5} & \frac{704}{625} & -\frac{2}{625} & -\frac{52}{625} & -\frac{4}{75} & -\frac{14}{75} & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{7}{256} & -\frac{7}{256} & \frac{125}{768} & -\frac{125}{768} & 0 \\
\hline
 & & \frac{1}{16} & \frac{1}{16} & \frac{125}{432} & \frac{125}{432} & \frac{8}{27}
\end{array} . \tag{6.26}$$

We will require all additional stages of the CGMIRK scheme to have stage order 6. To obtain a CGMIRK method with the ACDC property, we require it to satisfy the sixth order conditions plus all order conditions for order seven except one. The application of the stage order five conditions leaves 6 order conditions for 6^{th} order which are

$b^T(\theta)e = \theta, \quad b^T(\theta)c = \frac{\theta^2}{2}, \quad b^T(\theta)c^2 = \frac{\theta^3}{3}, \quad b^T(\theta)c^3 = \frac{\theta^4}{4}, \quad b^T(\theta)c^4 = \frac{\theta^5}{5}, \quad b^T(\theta)c^5 = \frac{\theta^6}{6}$. There are 2 order conditions for order 7 which are $b^T(\theta)c^6 = \frac{\theta^7}{7}, \quad b^T(\theta)\big(Xc^5 + \frac{v}{6}\big) = \frac{\theta^7}{42}$.

The 6^{th} and 7^{th} stages are required to satisfy the stage order six conditions. After imposing these conditions on stages six and seven we obtain c_6, x_{61}, x_{62}, x_{63}, x_{64} and x_{65} in terms of v_6, and x_{71}, x_{72}, x_{73}, x_{74}, x_{75} and x_{76} in terms of c_7, and v_7.

We then require the scheme to satisfy the 7 sixth order conditions as well as 1 out of the 2 seventh order conditions. There are thus two ways to derive an ACDC CGMIRK765. One way is to require the scheme to satisfy the seven quadrature

order conditions which are $b^T(\theta)e = \theta$, $b^T(\theta)c = \frac{\theta^2}{2}$, $b^T(\theta)c^2 = \frac{\theta^3}{3}$, $b^T(\theta)c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$, and $b^T(\theta)c^6 = \frac{\theta^7}{7}$, and we will call this CGMIRK765-I. The other way is to require the scheme to satisfy the six 6^{th} order quadrature order conditions given above, as well as the 7^{th} order condition $b^T(\theta)\big(Xc^5 + \frac{v}{6}\big) = \frac{\theta^7}{42}$; we will call this scheme CGMIRK765-II.

In either case, the scheme will be required to have seven stages. After solving the 7 order conditions, and choosing values for the free parameters as indicated below, we obtain an example of an CGMIRK765-I method, presented in the following tableau:

CGMIRK765-I with $v_6 = \frac{2}{5}, c_7 = \frac{3}{4}, v_7 = \frac{3}{4}$**:**

$$
\begin{array}{c|c|ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & -\frac{79}{625} & \frac{52}{625} & \frac{2}{625} & \frac{14}{75} & \frac{4}{75} & 0 & 0 & 0 \\
\frac{4}{5} & \frac{704}{625} & -\frac{2}{625} & -\frac{52}{625} & -\frac{4}{75} & -\frac{14}{75} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{7}{256} & -\frac{7}{256} & \frac{125}{768} & -\frac{125}{768} & 0 & 0 & 0 \\
\frac{2}{5} & \frac{2}{5} & \frac{183}{5000} & -\frac{117}{5000} & \frac{317}{1800} & -\frac{223}{1800} & -\frac{368}{5625} & 0 & 0 \\
\frac{3}{4} & \frac{3}{4} & \frac{333}{32768} & -\frac{843}{16384} & \frac{16375}{147456} & -\frac{44875}{294912} & \frac{559}{2304} & -\frac{2625}{16384} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) & b_7(\theta)
\end{array}
\quad , \tag{6.27}
$$

where the coefficients for the first five rows of the tableau are from the tableau of the embedded discrete (GMIRK565) scheme (6.26), and where

$$b_1(\theta) = \frac{1}{1008}\theta(1008 - 6594\theta + 22106 * t^2 - 41727\theta^3 + 44814\theta^4 - 25550\theta^5 + 6000\theta^6),$$

$$b_2(\theta) = \frac{1}{1008}\theta^2(-1008 + 8120\theta - 27069\theta^2 + 45108\theta^3 - 37100\theta^4 + 12000\theta^5),$$

$$b_3(\theta) = -\frac{125}{33264}\theta^2(-5040 + 27160\theta - 63945\theta^2 + 78036\theta^3 - 48300\theta^4 + 12000\theta^5),$$

$$b_4(\theta) = -\frac{125}{3024}\theta^2(-630 + 4970\theta - 16065\theta^2 + 25662\theta^3 - 19950\theta^4 + 6000\theta^5),$$

$$b_5(\theta) = -\frac{8}{189}\theta^2(-1008 + 7448\theta - 21987\theta^2 + 31584\theta^3 - 22050\theta^4 + 6000\theta^5),$$

$$b_6(\theta) = \frac{125}{1764}\theta^2(-630 + 4445\theta - 12390\theta^2 + 16947\theta^3 - 11375\theta^4 + 3000\theta^5),$$

$$b_7(\theta) = \frac{512}{4851}\theta^2(-336 + 2632\theta - 8421\theta^2 + 13272\theta^3 - 10150\theta^4 + 3000\theta^5).$$

We can predict the shape of the defect for this scheme (CGMIRK765-I) and the location of the maximum defect for each subinterval. The defect will be a multiple of the derivative of the unsatisfied order condition; see Figure 6.10. The maximum defect of each subinterval will occur where the second derivative of the unsatisfied order condition is zero. It will be at $\theta = 0.9191047218$.

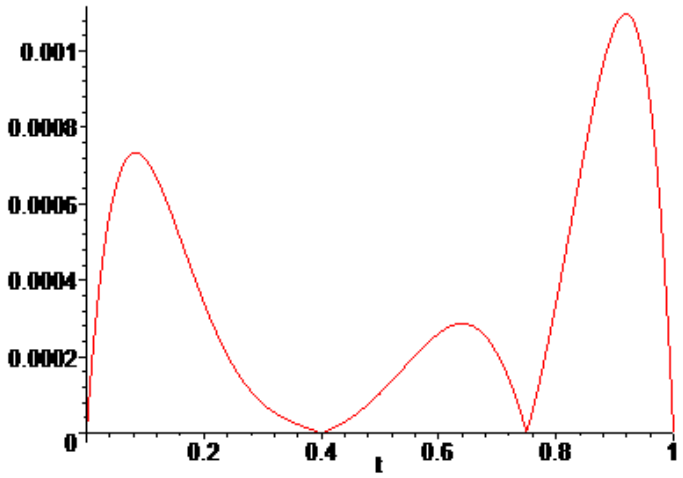

Figure 6.10: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition for CGMIRK765-I.

An example of a CGMIRK765-II method is given in the following tableau:

CGMIRK765-II with $v_6 = \frac{2}{5}, c_7 = \frac{1}{4}, v_7 = \frac{1}{4}$**:**

$$
\begin{array}{c|c|ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & -\frac{79}{625} & \frac{52}{625} & \frac{2}{625} & \frac{14}{75} & \frac{4}{75} & 0 & 0 & 0 \\
\frac{4}{5} & \frac{704}{625} & -\frac{2}{625} & -\frac{52}{625} & -\frac{4}{75} & -\frac{14}{75} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{7}{256} & -\frac{7}{256} & \frac{125}{768} & -\frac{125}{768} & 0 & 0 & 0 \\
\frac{2}{5} & \frac{2}{5} & \frac{183}{5000} & -\frac{117}{5000} & \frac{317}{1800} & -\frac{223}{1800} & -\frac{368}{5625} & 0 & 0 \\
\frac{1}{4} & \frac{1}{4} & \frac{1581}{32768} & -\frac{219}{16384} & \frac{26375}{147456} & -\frac{24875}{294912} & \frac{71}{2304} & -\frac{2625}{16384} & 0 \\
\hline
 & & b_1(\theta) & b_2(\theta) & b_3(\theta) & b_4(\theta) & b_5(\theta) & b_6(\theta) & b_7(\theta)
\end{array}
\quad . \tag{6.28}
$$

The coefficients for the first five rows of the tableau are from the tableau of the embedded discrete (GMIRK565) scheme (6.26). The weight polynomials are

$$b_1(\theta) = -\frac{1}{376}\theta(-376 + 2121\theta - 6033\theta^2 + 9112\theta^3 - 6903\theta^4 + 2050\theta^5),$$

$$b_2(\theta) = \frac{1}{20304}\theta^2(7824 - 51496\theta + 135291\theta^2 - 159684\theta^3 + 69400\theta^4)),$$

$$b_3(\theta) = \frac{125}{20304}\theta^2(1200 - 7160\theta + 16425\theta^2 - 16284\theta^3 + 5800\theta^4),$$

$$b_4(\theta) = -\frac{125}{10152}\theta^2(165 - 1055\theta + 2655\theta^2 - 2937\theta^3 + 1150\theta^4),$$

$$b_5(\theta) = \frac{128}{1269}\theta^2(174 - 1010\theta + 2223\theta^2 - 2082\theta^3 + 700\theta^4),$$

$$b_6(\theta) = -\frac{125}{10152}\theta^2(2220 - 11930\theta + 23865\theta^2 - 20622\theta^3 + 6500\theta^4),$$

$$b_7(\theta) = -\frac{256}{1269}\theta^2(-48 + 136\theta - 93\theta^2 - 48\theta^3 + 50\theta^4).$$

We can predict the shape of the defect for this scheme (CGMIRK765-II) and the location of the maximum defect for each subinterval. The defect will be a multiple of the derivative of the unsatisfied order condition; see Figure 6.11. The maximum defect of each subinterval will occur where the second derivative of the unsatisfied order condition is zero. It will be at $\theta = 0.9257407006$.

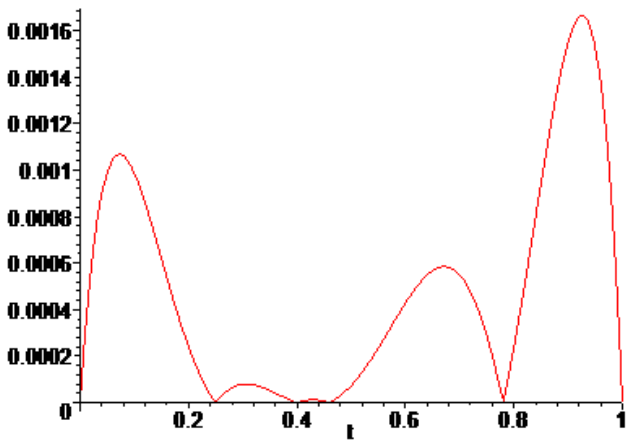

Figure 6.11: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition for CGMIRK765-II.

We also considered the case where we chose the free coefficients for CGMIRK765-I and CGMIRK765-II to be the same. In this case, the weight polynomials of the two schemes are different; however, they lead to the same shape for the derivative of the lone unsatisfied order condition in each case, and thus the same location for the maximum defect.

6.5.3 ACDC CGMIRK666

We derive a sixth order continuous ACDC CGMIRK scheme that has the discrete GMIRK666 [13] scheme embedded within it. The discrete scheme has the stage order

vector SOV = (6, 6, 6, 6, 6, 6), and it has the tableau:

$$\begin{array}{c|c|cccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & -\frac{23}{81} & \frac{23}{243} & \frac{20}{729} & -\frac{2}{9} & \frac{7}{45} & \frac{2048}{3645} & 0 \\
\frac{2}{3} & -\frac{56}{81} & \frac{32}{243} & \frac{47}{729} & \frac{1}{9} & \frac{22}{45} & \frac{2048}{3645} & 0 \\
\frac{1}{4} & -\frac{299}{1024} & \frac{783}{8192} & \frac{231}{8192} & -\frac{2187}{8192} & \frac{6561}{40960} & \frac{21}{40} & 0 \\
\frac{3}{4} & -\frac{567}{1024} & \frac{987}{8192} & \frac{435}{8192} & \frac{729}{8192} & \frac{21141}{40960} & \frac{21}{40} & 0 \\
\hline
 & & \frac{29}{360} & \frac{29}{360} & \frac{27}{200} & \frac{27}{200} & \frac{64}{225} & \frac{64}{225}
\end{array} . \tag{6.29}$$

The application of stage order six leads to 6 order conditions for 6^{th} order which are $b^T(\theta)e = \theta$, $b^T(\theta)c = \frac{\theta^2}{2}$, $b^T(\theta)c^2 = \frac{\theta^3}{3}$, $b^T(\theta)c^3 = \frac{\theta^4}{4}$, $b^T(\theta)c^4 = \frac{\theta^5}{5}$, $b^T(\theta)c^5 = \frac{\theta^6}{6}$. There is 1 order condition for order 7 which is $b^T(\theta)c^6 = \frac{\theta^7}{7}$.

Thus, this CGMIRK will already have the ACDC property because there is only one unsatisfied continuous order condition for order seven. After embedding the GMIRK666 (6.29), we require the scheme to satisfy the 6 sixth order continuous quadrature order conditions.

This will require the scheme to have only six stages because it has to satisfy only the six sixth order continuous quadrature order conditions. We then obtain the ACDC CGMIRK666 scheme with the GMIRK666 embedded. The resulting scheme will have the same coefficients as the discrete method presented in the tableau (6.29). The continuous weight polynomials are:

$$b_1(\theta) = -\frac{1}{360}\theta(-360 + 1950\theta - 5240\theta^2 + 7365\theta^3 - 5184\theta^4 + 1440\theta^5),$$

$$b_2(\theta) = \frac{1}{360}\theta^2(180 - 1180\theta + 3045\theta^2 - 3456\theta^3 + 1440\theta^4),$$

$$b_3(\theta) = -\frac{27}{200}\theta^2(180 - 940\theta + 1815\theta^2 - 1536\theta^3 + 480\theta^4),$$

$$b_4(\theta) = \frac{27}{200}\theta^2(90 - 560\theta + 1335\theta^2 - 1344\theta^3 + 480\theta^4),$$

$$b_5(\theta) = \frac{64}{225}\theta^2(90 - 410\theta + 735\theta^2 - 594\theta^3 + 180\theta^4),$$

$$b_6(\theta) = -\frac{64}{225}\theta^2(30 - 190\theta + 465\theta^2 - 486\theta^3 + 180\theta^4).$$

We can predict the shape of the defect for this scheme (CGMIRK666) and the location of the maximum defect for each subinterval. The defect will be the derivative of the unsatisfied order condition ($b^T(\theta)c^6 = \frac{\theta^7}{7}$); see Figure 6.12. The maximum defect of each subinterval will occur where the second derivative is zero. It will be at $\theta = 0.07286456136$ or $\theta = 0.9271354386$.

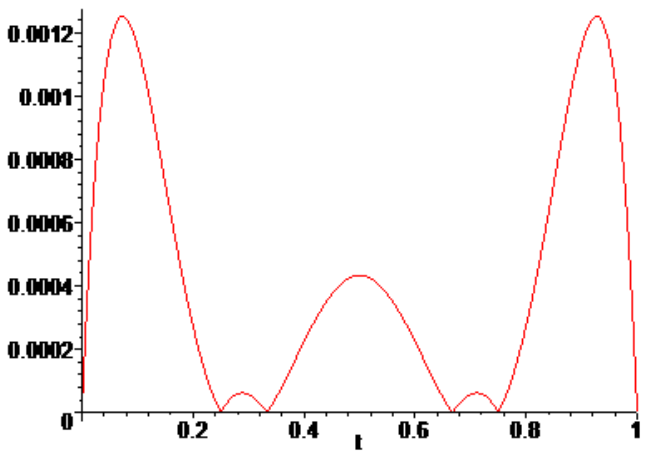

Figure 6.12: A plot of the absolute derivative of the lone unsatisfied 7^{th} order condition for CGMIRK666.

6.6 Comparing 6^{th} Order Schemes

Table 6.1 shows the number of stages for the different 6^{th} order ACDC schemes discussed in this chapter. Implementations for 6^{th} order ACDC CMIRK and ACDC CGMIRK schemes is left for future work.

Table 6.1: A comparison of the number of stages for the 6^{th} Order Hermite-Birkhoff interpolant, the 6^{th} Order ACDC CMIRK schemes, and the 6^{th} Order ACDC CGMIRK schemes.

Scheme	Number of stages
6^{th} order Hermite-Birkhoff interpolant	12
ACDC CMIRK1363	13
ACDC CMIRK1163	11
ACDC CGMIRK1064	10
ACDC CGMIRK765	7
ACDC CGMIRK666	6

We note that while the CGMIRK schemes use fewer stages, the computation of these stages is more expensive because some of the stages are defined implicitly in terms of each other and thus the solution of a non-linear system to determine these stages is required.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we discussed methods that involve defect control for solving BVODEs. We reviewed some of the background literature for standard CMIRK schemes and Hermite-Birkhoff interpolants. Then, numerical results were presented showing plots of the absolute scaled defect to allow observation of the location of the maximum defect for each subinterval for standard CMIRK schemes and Hermite-Birkhoff (H-B) interpolants on BVODEs. These results showed that standard CMIRK schemes, for orders 4, 5 and 6, do not, in general, lead to defects for which the location of the maximum defect on each subinterval and for each problem is the same. This means that asymptotically correct defect control (ACDC) is not possible. These results also showed that H-B interpolants can be constructed that do have the ACDC property.

After that, ACDC CMIRK methods of orders 4 and 5 were developed by requiring the methods to satisfy all but one of the order conditions for one higher order. This means that the method will have a defect for which the leading order term in the defect expansion will be a multiple of a single polynomial corresponding to the

derivative of the lone unsatisfied order condition. We then provided numerical experiments that show that these ACDC CMIRK methods indeed have the ACDC property. We also compared the efficiency of the order 4 and 5 H-B interpolants and ACDC CMIRK methods. Finally, we investigated standard CMIRK, H-B, ACDC CMIRK, and ACDC CGMIRK methods for 6^{th} order.

7.2 Future Work

The main direction of future work is to implement the ACDC CMIRK methods within the BVP-SOLVER-2 [4] software package. A second direction for future work is to optimize the ACDC CMIRK schemes. The idea in choosing the coefficients optimally is to try to make the factor in the local error expansion that depends on those coefficients as small as possible. A third direction for future work is to investigate the reason why when we choose the free coefficients to be the same in the ACDC CMIRK543-I and ACDC CMIRK543-II schemes we obtain the same scheme. A similar question to consider would be why when we choose the free coefficients to different in the ACDC CGMIRK765-I and ACDC CGMIRK765-II schemes we obtain different schemes, but with the same shape for the defect. A fourth direction for future work is to develop Scilab implementations for the ACDC CMIRK1363, ACDC CMIRK1163, and ACDC CGMIRK schemes, so that we can conduct numerical testing of these schemes.

Bibliography

1. U. Ascher, *Collocation for two-point boundary value problems revisited*, SIAM J. Numer. Anal. 23 (1986) 596-609.

2. U.M. Ascher, R.M.M. Mattheij, R.D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Classics in Applied Mathematics Series, SIAM, Philadelphia, 1995.

3. U.M. Ascher and L.P. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.

4. J. Boisvert, P.H. Muir, R.J. Spiteri, *A Runge-Kutta BVODE Solver with Global Error and Defect Control*, ACM Trans. Math. Softw. 39 (2013), Art. 11.

5. J.C. Butcher, *Implicit Runge-Kutta processes*, Math. Comp. 18 (1964) 50-64.

6. J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, Wiley, Chichester, 1987.

7. K. Burrage, F.H. Chipman and P.H. Muir, *Order results for mono-implicit Runge-Kutta methods*, SIAM J. Numer. Anal. 31 (1994) 876-891.

8. J.R. Cash, *On the numerical integration of nonlinear two-point boundary value problems using iterated deferred corrections, Part 1 : A survey and comparison of some one-step formulae*, Comput. Math. Appl. 12 (1986) 1029-1048.

9. J.R. Cash and D.R. Moore, *A high order method for the numerical solution of two-point boundary value problems*, BIT 20 (1980), 44-52.

10. J.R. Cash and A. Singhal, *High order methods for the numerical solution of two-point boundary value problems*, BIT 22 (1982) 184-199.

11. J.R. Cash and A. Singhal, *Mono-implicit Runge-Kutta formulae for the numerical integration of stiff differential systems*, IMA J. Numer. Anal. 2 (1982), 211-227.

12. J.R. Cash and M. H. Wright, *A deferred correction method for nonlinear two-point boundary value problems: implementation and numerical evaluation*, SIAM J. Sci. Statist. Comput. 12 (1991) 971-989.

13. F. Dow, *Generalized Mono-Implicit Runge-Kutta Methods for Stiff Ordinary Differential Equations*, M.Sc. Thesis, Saint Mary's University, Dept. of Mathematics and Computing Science. 2017.

14. A. Ellis, *Asymptotically Correct Defect Control Software for Boundary Value Ordinary Differential Equations*, M.Sc. Thesis, Saint Mary's University, Dept. of Mathematics and Computing Science. 2014.

15. W.H. Enright,*Continuous numerical methods for ODEs with defect control*, J. Comput. Appl. Math. 125 (2000), 159-170.

16. W.H. Enright, *The relative efficiency of alternative defect control schemes for high-order continuous Runge-Kutta formulas*, SIAM J. Numer. Anal. 30 (1993), 1419-1445.

17. W.H. Enright and W.B. Hayes, *Robust and reliable defect control for Runge-Kutta methods*, ACM Trans. Math. Softw. 33 (2007), 1-19.

18. W.H. Enright, K.R. Jackson, S.P. Norsett, P.G. Thomsen, *Interpolants for Runge-Kutta formulas*, ACM Trans. Math. Softw. 12 (1986), 193-218.

19. W.H. Enright and P.H. Muir, *Efficient classes of Runge-Kutta methods for two-point boundary value problems*, Comp. 37 (1986) 315-334.

20. W.H. Enright and P.H. Muir, *New interpolants for asymptotically correct defect control of BVODES*, Numer. Alg. 53 (2010), 219-238.

21. W.H. Enright and P.H. Muir, *Runge-Kutta software with defect control for boundary value ODEs*, SIAM J. Sci. Comput. 17 (1996), 479-497.

22. S. Gupta, *An adaptive boundary value Runge-Kutta solver for first order boundary value problems*, SIAM J. Numer. Anal. 22 (1985), 114-126.

23. D. J. Higham, *Robust defect control with Runge-Kutta schemes*, SIAM J. Numer. Anal. 26, (1989) 1175-1183.

24. H. B. Keller, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Academic Press, New York, 1975.

25. J. Kierzenka and L.F. Shampine, *A BVP Solver based on residual control and the MATLAB PSE*, ACM Trans. Math. Softw. 27 (2001), 299-316.

26. J. Kierzenka and L.F. Shampine, *A BVP Solver that controls residual and error*, J. Numer. Anal. Ind. Appl. Math. 3 (2008), 27-41.

27. J.D. Lambert, *Numerical Methods for Ordinary Differential Equations.* John Wiley, New York, 1991.

28. P.H. Muir, *Optimal discrete and continuous mono-implicit Runge-Kutta schemes for BVODEs*, Adv. Comput. Math. 10 (1999) 135-167.

29. P.H Muir and B. Owren, *Order barriers and characterizations for continuous mono-implicit Runge-Kutta schemes*, Math. Comp. 61 (1993), 675-699.

30. P.H. Muir and W.H. Enright, *Relationships among some classes of implicit Runge-Kutta methods and their stability functions*, BIT 27 (1987) 403-423.

31. H. H. Robertson, *The solution of a set of reaction rate equations*, in Numerical Analysis: An Introduction, J. Walsh, ed., Academic Press, 1966, 178-182.

32. R. D. Russell and J. Christiansen, *Adaptive mesh selection strategies for solving boundary value problems*, SIAM J. Numer. Anal. 14, (1978) 59-80.

33. L. F. Shampine, *Interpolation for Runge-Kutta methods*, SIAM J. Numer. Anal. 22, (1985) 1014-1027.

34. L. F. Shampine, *Solving ODEs and DDEs with Residual Control*, Appl. Numer. Math. 52, (2005) 113-117.

35. L.F. Shampine, R.C. Allen and S. Pruess, *Fundamentals of Numerical Computing.* John Wiley, New York, 1997.

36. L. F. Shampine, I. Gladwell, and S. Thompson, *Solving ODEs with Matlab*, Cambridge Univ. Press, New York, 2003.

37. L.F. Shampine and P.H. Muir, *Estimating conditioning of BVPs for ODEs*, Math. Comput. 40 (2004) 1309-1321.

38. L.F. Shampine, P.H. Muir, and H. Xu, *A user-friendly Fortran BVP solver*, J. Numer. Anal. Ind. Appl. Math. 1 (2006) 201-217.

39. L.F. Shampine, J. Kierzenka, and M.W. Reichelt, *Solving boundary value problems for ordinary differential equations in MATLAB with bvp4c*, http://www.mathworks.com, 2000.

40. J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York-Heidelberg, 1980.

41. D. Tan and Z. Chen, *On a general formula of fourth order Runge-Kutta method*, J. Math. Sci. Math. Educ. 7, (2012) 1-10.

42. W.M.G. Van Bokhoven, *Efficient higher order implicit one-step methods for integration of stiff differential equations*, BIT 20 (1980) 34-43.

43. D. Voss and P.H. Muir, *Mono-implicit Runge-Kutta schemes for the parallel solution of initial value problems*, J. Comp. Appl. Math. 102 (1999) 235-252.

44. R. Weiss, *The application of implicit Runge-Kutta and collocation methods to boundary value problems*, Math. Comp. 28 (1974) 449-464.

45. The Scilab Consortium. Scilab. http://www.scilab.org.