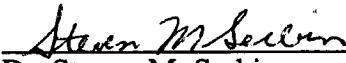
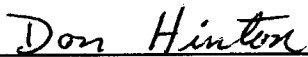


To the Graduate Council:

I am submitting herewith a thesis written by Pamela Stanfield Lancaster entitled "A Study of Continuous Explicit and Diagonally Implicit Runge-Kutta Methods." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Mathematics.


Dr. Steven M. Serbin,
Major Professor

We have read this thesis
and recommend its acceptance:





Accepted for the Council:


Associate Vice Chancellor
and Dean of The Graduate School

A Study of Continuous Explicit and Diagonally Implicit Runge-Kutta Methods

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Pamela Stanfield Lancaster
December, 1994

DEDICATION

This thesis is dedicated to my parents
William and Anna Faye Stanfield,
who supported me in every way and made it
possible to fulfill my dream of attending the
University of Kentucky.

ACKNOWLEDGMENTS

I would like to thank my major professor, Dr. Steve Serbin, for his guidance and support. I would also like to thank the thesis committee members, Dr. Charles Collins and Dr. Don Hinton, who gave of their time and scholarship. I want to express my gratitude to my husband Mark, who gave me his love and encouragement, and who also did a very beautiful job typing this thesis.

ABSTRACT

Runge-Kutta methods provide a popular way to solve ordinary differential equations. However, traditional schemes only provide solutions at a discrete set of points. In this thesis, continuous Runge-Kutta methods for first order ordinary differential equations are considered. The developments of some continuous explicit Runge-Kutta formulas by various researchers are presented in the first chapter. They impose different conditions on their schemes, but all make use of what are called order conditions. In the second chapter, two-stage and three-stage continuous diagonally implicit methods are developed. With the two-stage continuous method, second order may be attained with third order at the mesh points. For the particular third-order continuous method presented, third order may be attained with fourth order at the mesh points. As revealed in the conclusion of the thesis, a three-stage continuous diagonally implicit Runge-Kutta method that is globally third order and fourth order at the mesh points will not satisfy certain stability requirements.

TABLE OF CONTENTS

Chapter 1: Continuous Explicit Runge-Kutta Methods	
1.1 Introduction	1
1.2 Development of Order Conditions for CERKs	4
1.3 Sarafyan's Development of CERKs	8
1.4 Derr, Outlaw, and Sarafyan's Development of CERKs	11
1.5 Owren and Zennaro's Development of CERKs	14
Chapter 2: Continuous Diagonally Implicit Runge-Kutta Methods	
2.1 Introduction	16
2.2 Two-Stage Continuous MDIRK method	16
2.3 Particular Two-Stage Discrete MDIRKs	19
2.4 Stability of Two-Stage MDIRKs	21
2.5 Three-Stage Continuous MDIRK Methods	24
2.6 A Family of Continuous Three-Stage MDIRKs	29
Bibliography	32
Vita	34

Chapter 1

Continuous Explicit Runge-Kutta Methods

1.1 Introduction

There are a number of methods for explicitly finding solutions for first order initial-value problems of the form:

$$y' = f(x, y(x)), \quad y(x_0) = y_0 \quad (1.1.1)$$

where $f: \mathbf{R} \times \mathbf{R}^m \rightarrow \mathbf{R}^m$ determines m possible nonlinear differential equations and y_0 is the m vector of initial values. However, there are many situations in which none of these methods can be applied. Therefore we must turn to numerical methods in order to find an approximate solution. It is well-known that we could use Taylor series expansion so that:

$$\left. \begin{aligned} y'' &= f_x + f_y y' = f_x + f_y f \\ y''' &= f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f) \\ &\vdots \end{aligned} \right\} \quad (1.1.2)$$

and the solution for this method is, in principle, generated by:

$$y(x_0 + h) = y(x_0) + \sum_{n=1}^{\infty} \frac{h^n}{n!} y^{(n)}(x_0, y_0)$$

where h is the step size. Taylor approximation methods can be generated by truncating this expansion.

Obviously the formulas in (1.1.2) become quite complicated for most problems. That is why numerical methods have evolved from the Taylor series method: Euler's method and the Midpoint method just to name two. At the turn of the century, C. Runge and K. Heun constructed improved methods based on Euler's method. It was W. Kutta who soon afterwards formulated the general scheme of what is now called a Runge-Kutta

method [HNW]. There are several ways of expressing these Runge-Kutta methods, but we will use the following definition:

Definition 1: Let $s \in \mathbb{Z}^+$ represent the number of stages, and let a_{ij}, b_j, c_i be real coefficients, where $i, j = 1, 2, 3, \dots, s$. Then the method

$$\left. \begin{aligned} k_i &= f\left(x_n + c_i h, y_n + h \sum_{j=1}^s a_{ij} k_j\right), \quad i = 1, 2, 3, \dots, s \\ y_{n+1} &= y_n + h \sum_{i=1}^s b_i k_i \end{aligned} \right\} \quad (1.1.3)$$

is called an s -stage Runge-Kutta Method for (1.1.1).

It is important to note that some authors opt to use the notation (x_0, y_0) to denote the starting point for *each* step in the process (as opposed to (x_n, y_n)), so that the approximation at the next point $x_0 + h$ is y_1 . A particular Runge-Kutta method is typically specified by the tableau of coefficients:

c_1	a_{11}	a_{12}	$\cdots$	$a_{1,s-1}$	a_{1s}
c_2	a_{21}	$\cdots$	$\cdots$	$\cdots$	a_{2s}
c_3	$\vdots$				$\vdots$
$\vdots$	$\vdots$				$\vdots$
c_s	a_{s1}	$\cdots$	$\cdots$	$a_{s,s-1}$	a_{ss}
	b_1	b_2	$\cdots$	b_{s-1}	b_s

Tableau I.I.I

This tableau is called the Butcher array, after John Butcher [HNW].

An important aspect of any Runge-Kutta method is that of order. We will use the following definition as found in [HNW].

Definition 2: A Runge-Kutta method has *order* p if for sufficiently smooth problems as in (1.1.1),

$$\|y(x_0 + h) - y_1\| \leq Kh^{p+1}$$

i.e., if the Taylor series for the exact solution $y(x_0 + h)$ and for y_1 coincide up to and including the term h^p .

As mentioned earlier, Runge-Kutta methods actually evolved from Taylor series methods. Notice that in the definition of order, we are comparing the Runge-Kutta approximation y_1 to the exact Taylor series solution $y(x_0 + h)$. The advantage of the Runge-Kutta methods is that we can obtain a high order method while eliminating the need to compute and evaluate the derivatives of $f(x, y)$. If in Definition 1 we require that $a_{ij} = 0$ for all $i \geq j$, then the method is called an *s-stage explicit Runge-Kutta method*. The benefit of using an explicit method is that we need not attempt to solve the k_i 's simultaneously. Using an explicit method will reduce the complexity of the problem, thus making these methods particularly attractive.

Traditionally, Runge-Kutta methods only approximated solutions at a discrete set of points called *mesh* points. Notice that for the method described in Definition 1 this is the case. However, sometimes the need arises for dense output, and using a conventional discrete Runge-Kutta method is inconvenient. If we slightly modify Definition 1 so that the b_j 's are no longer scalar coefficients, but instead polynomials in terms of θ , then we have what is called a *continuous Runge-Kutta method*. Here we are attempting to produce approximations to the solution at a continuum of points $x_0 + \theta h$, where $0 \leq \theta \leq 1$.

Therefore the form of the solution would be

$$\tilde{y}(x_0 + \theta h) = y_0 + h(b_1(\theta)k_1 + b_2(\theta)k_2 + \cdots + b_s(\theta)k_s).$$

The point of departure of this thesis is the study of various classes of continuous Runge-Kutta methods. Among a vast and growing literature on the topic, we have chosen to study explicit and diagonally implicit methods. The focus of Chapter 1 is the development of continuous explicit Runge-Kutta (CERK) methods. In Chapter 2, we will look at diagonally implicit Runge-Kutta (DIRK) methods.

As in many other areas of study, the development of CERK methods is relatively new and has had several recent developments. It was ten years ago that Diran Sarafyan's paper on continuous approximations was published [Sar]. As we shall see in Section 1.3 of this

thesis, he actually developed his method by comparing the Taylor series expansion of $y(x_0 + \theta h)$ to the Runge-Kutta approximation $\tilde{y}(x_0 + \theta h)$. He does develop a solid CERK method, but in the process removes one of the major benefits of using a Runge-Kutta method: elimination of the computation of high order derivatives of $f(x, y)$. However, papers published within the past few years have shown that all of these computations are not necessary in order to develop a CERK method. In Section 1.4 and Section 1.5, we will look at developments of methods for which this is the case. These methods present necessary *order conditions* for Runge-Kutta schemes produced with much less messy computations. It was John Butcher who was first credited with discovering a relationship between the elementary differentials used in the Taylor series expansion and certain rooted labeled trees [HNW]. The main difficulty in the derivation of the order conditions is to understand the correspondence of the formulas to these trees. This relationship, once understood, allows one to write down the conditions of order somewhat easily.

The next section (Section 1.2) displays the author's initial attempts to understand the use of these rooted trees to produce the order conditions for a continuous explicit Runge-Kutta method.

1.2 Development of Order Conditions for CERKs

We now want to look at development of continuous explicit Runge-Kutta (CERK) methods. For an s -stage CERK to have order p , we require the following order conditions:

$$\sum_{j=1}^s b_j(\theta) \Phi_j(t) = \frac{\theta^{\rho(t)}}{\gamma(t)} \quad (1.2.1)$$

for all trees t such that $\rho(t) \leq p$, where $\Phi_j(t)$ is the j th elementary weight of the tree t ,

Table I: Trees and elementary differentials up to order 4

$\rho(t)$	t	graph	$\gamma(t)$	$\alpha(t)$	$F^J(t)(y)$	$\Phi_j(t)$
0	$\emptyset$	$\emptyset$	1	1	y^J	
1	τ	$\bullet j$	1	1	f^J	1
2	t_{21}		2	1	$\sum_K f_K^J f^K$	$\sum_k a_{jk}$
3	t_{31}		3	1	$\sum_{K,L} f_{KL}^J f^K f^L$	$\sum_{k,l} a_{jk} a_{jl}$
	t_{32}		6	1	$\sum_{K,L} f_K^J f_L^K f^L$	$\sum_{k,l} a_{jk} a_{kl}$
4	t_{41}		4	1	$\sum_{K,L,M} f_{KLM}^J f^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{jl} a_{jm}$
	t_{42}		8	3	$\sum_{K,L,M} f_{KM}^J f_L^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{jm}$
	t_{43}		12	1	$\sum_{K,L,M} f_K^J f_{LM}^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{km}$
	t_{44}		24	1	$\sum_{K,L,M} f_K^J f_L^K f_M^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{lm}$

Source: E. Hairer, S. P. Nørsett, and G. Wanner (1987): *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag, New York, p. 147.

$\rho(t)$ is the order of t , and $\gamma(t)$ is a coefficient depending on the tree t . Table I contains information necessary to compute (1.2.1) of order up to four. It is obvious that the degree of the polynomial $b_j(\theta)$ should be at least as large as p , the order of the method.

However, choosing the degree of $b_j(\theta)$ to be larger than p can lead to approximate solutions whose derivatives are unbounded as $h \rightarrow 0$. Therefore, the degree of the polynomial $b_j(\theta)$ should always be chosen to be equal to the order of the method [OZ1].

We first attempt to construct a three-stage CERK of order three. This method is of the form:

c_1			
c_2	a_{21}		
c_3	a_{31}	a_{32}	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$

Tableau I.II.I

where blank entries are taken to be zero, $b_1(\theta) = b_{10} + b_{11}\theta + b_{12}\theta^2 + b_{13}\theta^3$, $b_2(\theta) = b_{20} + b_{21}\theta + b_{22}\theta^2 + b_{23}\theta^3$, and $b_3(\theta) = b_{30} + b_{31}\theta + b_{32}\theta^2 + b_{33}\theta^3$. For this to be order three, we must satisfy the following order conditions based on trees of order less than or equal to three (The symbol to the left of the equation denotes the tree that gives rise to the particular condition.):

$$\tau: \quad b_1(\theta) + b_2(\theta) + b_3(\theta) = \theta \quad (1.2.2)$$

$$t_{21}: \quad a_{21}b_2(\theta) + (a_{31} + a_{32})b_3(\theta) = \frac{1}{2}\theta^2 \quad (1.2.3)$$

$$t_{31}: \quad a_{21}^2b_2(\theta) + (a_{31}^2 + 2a_{31}a_{32} + a_{32}^2)b_3(\theta) = \frac{1}{3}\theta^3 \quad (1.2.4)$$

$$t_{32}: \quad a_{21}a_{32}b_3(\theta) = \frac{1}{6}\theta^3 \quad (1.2.5)$$

We now use the above in order to solve for the coefficients of the b_j 's. The equation (1.2.5) implies $b_{30} = b_{31} = b_{32} = 0$, $b_{33}a_{32}a_{21} = \frac{1}{6}$, $a_{21} \neq 0$, and $a_{32} \neq 0$. By (1.2.3), we need $a_{21}b_{22} + (a_{31} + a_{32})b_{32} = \frac{1}{2} \Rightarrow a_{21}b_{22} = \frac{1}{2} \Rightarrow a_{21} \neq 0$ and $b_{22} \neq 0$. By (1.2.4), we need $a_{21}^2b_{22} + (a_{31}^2 + 2a_{31}a_{32} + a_{32}^2)b_{32} = 0 \Rightarrow a_{21}^2b_{22} = 0 \Rightarrow a_{21} = 0$ or $b_{22} = 0$. However, this contradicts $a_{21} \neq 0$ and $b_{22} \neq 0$. Therefore, we cannot satisfy these conditions.

Consequently, there are no three-stage third order CERK methods.

Owren and Zennaro [OZ1] have proven that a CERK method must have four stages to be of order three. Now we try to develop the general form for such a method. For this scheme the order conditions are:

$$b_1(\theta) + b_2(\theta) + b_3(\theta) + b_4(\theta) = \theta$$

$$a_{21}b_2(\theta) + (a_{31} + a_{32})b_3(\theta) + (a_{41} + a_{42} + a_{43})b_4(\theta) = \frac{1}{2}\theta^2$$

$$\begin{aligned}
& a_{21}^2 b_2(\theta) + (a_{31}^2 + 2a_{31}a_{32} + a_{32}^2) b_3(\theta) + \\
& (a_{41}^2 + a_{42}^2 + a_{43}^2 + 2a_{41}a_{42} + 2a_{41}a_{43} + 2a_{42}a_{43}) b_4(\theta) = \frac{1}{3} \theta^3 \\
& a_{21}a_{32} b_3(\theta) + (a_{42}a_{21} + a_{43}a_{31} + a_{43}a_{32}) b_4(\theta) = \frac{1}{6} \theta^3
\end{aligned}$$

At this point we make use of what are called *simplifying assumptions*, which produce sufficient, but not necessary conditions. The idea of using simplifying assumptions is common throughout the study of Runge-Kutta methods. Here we employ merely the two equations $c_i = \sum_{j=1}^s a_{ij}$ and $\frac{1}{2}c_i^2 = \sum_{j=1}^s a_{ij}c_j$; however, there are several other possible simplifying conditions. By using these two simplifying conditions, the order conditions become:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & c_2 & c_3 & c_4 \\ 0 & c_2^2 & c_3^2 & c_4^2 \\ 0 & 0 & \frac{1}{2}c_3^2 & \frac{1}{2}c_4^2 \end{bmatrix} \begin{bmatrix} b_1(\theta) \\ b_2(\theta) \\ b_3(\theta) \\ b_4(\theta) \end{bmatrix} = \begin{bmatrix} \theta \\ \frac{1}{2}\theta^2 \\ \frac{1}{3}\theta^3 \\ \frac{1}{6}\theta^3 \end{bmatrix}$$

Through tedious but straightforward computations, we obtain:

$$b_1(\theta) = \frac{\frac{1}{3}\theta^3 - \frac{1}{2}(c_3 + c_4)\theta^2 + c_3c_4\theta}{c_3c_4} \quad (1.2.6)$$

$$b_2(\theta) = 0 \quad (1.2.7)$$

$$b_3(\theta) = \frac{\frac{1}{3}\theta^3 - \frac{1}{2}c_4\theta^2}{c_3(c_3 - c_4)} \quad (1.2.8)$$

$$b_4(\theta) = \frac{-\frac{1}{3}\theta^3 + \frac{1}{2}c_3\theta^2}{c_4(c_4 - c_3)} \quad (1.2.9)$$

where $c_3 \neq 0$, $c_4 \neq 0$, and $c_3 \neq c_4$ are assumptions. This is just one version of the $b_j(\theta)$'s for a continuous four-stage, third order explicit Runge-Kutta method.

1.3 Sarafyan's Development of CERKs

We first want to focus our attention on the CERK methods developed by Sarafyan [Sar]. He initially sets out to develop a fifth order *discrete* Runge-Kutta method and reminds us that it is well-known that methods of this form require six stages. For this six-stage method he requires that the equations $c_i = \sum_{j=1}^s a_{ij}$ hold. The order conditions required for fifth order are then presented to us. These are found by comparing the Taylor series expansion of $y(x_0 + \theta h)$ to $\tilde{y}(x_0 + \theta h)$, the Runge-Kutta approximation. (At this point we assume $\theta = 1$ for the discrete method, thus the notation would be $y_1 = \tilde{y}(x_0 + h)$.) By equating these two expressions and matching like terms, one obtains a nonlinear system of sixteen equations involving twenty-one unknowns. The equations that he derived are exactly the same as what he would have obtained using (1.2.1) with $\theta = 1$. So that he may obtain a solution to these sixteen order conditions, he introduces two simplifying assumptions:

$$\sum_{j=1}^s a_{ij} c_j = \frac{1}{2} c_i^2 \quad i = 1, 2, \dots, s \quad (1.3.1)$$

$$\sum_{j=1}^s a_{ij} c_j^2 = \frac{1}{3} c_i^3 \quad i = 1, 2, \dots, s \quad (1.3.2)$$

One may realize that he considers (1.3.1) and (1.3.2) to contribute only four new equations each since in an explicit Runge-Kutta method $c_1 = 0$. At this point, there are twenty-four equations in twenty-one unknowns. However, by using these simplifying assumptions, seven of the sixteen original equations may be eliminated, and it is also determined that $b_2 = 0$.

As expected, the solution of this system of equations is complex and time-consuming, but by his efforts we are presented with the solution:

0						
$\frac{1}{6}$	$\frac{1}{6}$					
$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{16}$				
$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{3}{4}$	1			
$\frac{3}{4}$	$\frac{3}{16}$	0	0	$\frac{9}{16}$		
1	$-\frac{4}{7}$	$\frac{3}{7}$	$\frac{12}{7}$	$-\frac{12}{7}$	$\frac{8}{7}$	
	$\frac{7}{90}$	0	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$

Tableau I.III.I

This is not the only solution set that satisfies the order conditions and simplifying assumptions given earlier. (Recall that there were seventeen equations with twenty unknowns.) However, the size and simplicity of the constants makes this particular method attractive. Sarafyan states that extensive tests have revealed that this formula is among the most effective six-stage discrete explicit Runge-Kutta methods of order five.

Next Sarafyan develops his six-stage CERK method. There is no explanation given as to why a fifth order CERK method cannot be developed from a six-stage method, but it is stated that this is not possible. The next most appropriate task is to obtain a fourth order CERK. Again he obtains the order conditions by comparing $y(x_0 + \theta h)$ and $\tilde{y}(x_0 + \theta h)$ of like terms up to the fourth power. (Note that $0 < \theta \leq 1$ since we are considering the continuous case.) This time we are presented with eight order conditions. Again, the simplifying assumptions (1.3.1) and (1.3.2) are followed and $b_2(\theta) = 0$ results.

Sarafyan uses these conditions and the known values of the a_{ij} 's and c_i 's to obtain

$$b_1(\theta) = -\frac{10}{9}\theta^4 + \frac{142}{45}\theta^3 - \frac{89}{30}\theta^2 + \theta$$

$$b_2(\theta) = 0$$

$$b_3(\theta) = \frac{16}{9}\theta^4 - \frac{208}{45}\theta^3 + \frac{16}{5}\theta^2$$

$$b_4(\theta) = \frac{4}{3}\theta^4 - \frac{12}{5}\theta^3 + \frac{6}{5}\theta^2$$

$$b_5(\theta) = -\frac{32}{9}\theta^4 + \frac{272}{45}\theta^3 - \frac{32}{15}\theta^2$$

$$b_6(\theta) = \frac{14}{9}\theta^4 - \frac{98}{45}\theta^3 + \frac{7}{10}\theta^2$$

These values of the $b_j(\theta)$'s paired with the a_{ij} 's and c_i 's of tableau I.III.I give us a six-stage CERK of order four with fifth order at the mesh points.

He next derives a third order four stage CERK. Similarly, $y(x_0 + \theta h)$ and $\tilde{y}(x_0 + \theta h)$ are compared up to third degree, and from this we obtain four order conditions. Again, the simplifying assumption (1.3.1) is assumed and $b_2(\theta) = 0$ results. Sarafyan still assumes the same a_{ij} 's and c_i 's from tableau I.III.I, and we obtain

$$b_1(\theta) = \frac{8}{3}\theta^3 - 3\theta^2 + \theta$$

$$b_2(\theta) = 0$$

$$b_3(\theta) = -\frac{16}{3}\theta^3 + 4\theta^2$$

$$b_4(\theta) = \frac{8}{3}\theta^3 - \theta^2$$

One may notice that these are the same $b_j(\theta)$'s that would be found by employing equations (1.2.6) - (1.2.9).

Following the same method as before, it is possible to obtain both two-stage and three-stage CERKs of order two. Still the same a_{ij} 's and c_i 's from tableau I.III.I are used. For the three-stage method, we have $b_1(\theta) = -2\theta^2$, $b_2(\theta) = 0$, and $b_3(\theta) = 2\theta^2 - \theta$. In this case, it is not required that $b_2(\theta) = 0$, but this was Sarafyan's choice. For the two-stage method, $b_1(\theta) = -3\theta^2 - \theta$, and $b_2(\theta) = 3\theta^2$ are obtained.

Sarafyan states that typically $0 < \theta \leq 1$, but that excellent numerical results have been obtained when $-0.2 \leq \theta \leq 1.2$. However, outside these particular bounds, the accuracy is noticeably decreased.

1.4 Derr, Outlaw, and Sarafyan's Development of CERKs

In the work by Derr, Outlaw, and Sarafyan [DOS], they develop a five-stage CERK of order three that is fourth order at the mesh points. For clarity, the form of the five stages are listed below:

$$k_1 = f(x_0, y_0)$$

$$k_2 = f(x_0 + c_2 h, y_0 + h a_{21} k_1)$$

$$k_3 = f(x_0 + c_3 h, y_0 + h(a_{31} k_1 + a_{32} k_2))$$

$$k_4 = f(x_0 + c_4 h, y_0 + h(a_{41} k_1 + a_{42} k_2 + a_{43} k_3))$$

$$k_5 = f(x_0 + c_5 h, y_0 + h(a_{51} k_1 + a_{52} k_2 + a_{53} k_3 + a_{54} k_4))$$

The first four stages (k_1 , k_2 , k_3 , and k_4) are needed to develop the fourth order discrete method and the additional fifth stage is needed to create the embedded third order CERK. The solution for this method is of the form

$$\tilde{y}(x_0 + \theta h) = y_0 + h \sum_{j=1}^5 b_j(\theta) k_j$$

where $b_j(\theta) = b_{j0} + b_{j1}\theta + b_{j2}\theta^2 + b_{j3}\theta^3$ for $0 \leq \theta \leq 1$. Since it is desired that the approximation $\tilde{y}(x_0 + \theta h) = y_0$ when $\theta = 0$, it is necessary that each $b_{j0} = 0$.

For this construction of a CERK method, what distinguishes this work is the desire that the last stage of one step be the first stage of the next step. This reduces the number of function evaluations. In order to accomplish this, it is required that $c_5 = 1$, $a_{5j} = b_j(1)$, and $b_5(1) = 0$. To see why this is so, realize that we are insisting that $k_5 = f(x_0 + h, \tilde{y}(x_0 + h))$. They also wish for the third order CERK to be C^1 . For this condition to be satisfied, it is necessary that $h \sum_{i=1}^5 b'_i(1) k_i = k_5$. Forcing the previous

equation to hold allows us to obtain the correct slope of the solution $\tilde{y}(x_0 + \theta h)$ at the mesh point.

In addition to the previous conditions, they also make use of the simplifying assumption $c_i = \sum_{j=1}^s a_{ij}$. At this point, Derr, Outlaw, and Sarafyan consider the order conditions. They remind us that the equations for third order must be satisfied for arbitrary $0 \leq \theta \leq 1$, so that

$$b_1(\theta) + b_2(\theta) + b_3(\theta) + b_4(\theta) + b_5(\theta) - \theta = 0 \quad (1.4.1)$$

$$c_2 b_2(\theta) + c_3 b_3(\theta) + c_4 b_4(\theta) + c_5 b_5(\theta) - \frac{1}{2} \theta^2 = 0 \quad (1.4.2)$$

$$c_2^2 b_2(\theta) + c_3^2 b_3(\theta) + c_4^2 b_4(\theta) + c_5^2 b_5(\theta) - \frac{1}{3} \theta^3 = 0 \quad (1.4.3)$$

$$\begin{aligned} c_2 a_{32} b_3(\theta) + (c_2 a_{42} + c_3 a_{43}) b_4(\theta) + \\ (c_2 a_{52} + c_3 a_{53} + c_4 a_{54}) b_5(\theta) - \frac{1}{6} \theta^3 = 0 \end{aligned} \quad (1.4.4)$$

The equations for fourth order must be satisfied for $\theta = 1$, so that

$$c_2^3 b_2(1) + c_3^3 b_3(1) + c_4^3 b_4(1) + c_5^3 b_5(1) - \frac{1}{4} = 0$$

$$\begin{aligned} c_3(a_{32}c_3)b_3(1) + c_4(a_{42}c_2 + a_{43}c_3)b_4(1) + \\ c_5(a_{52}c_2 + a_{53}c_3 + a_{54}c_4)b_5(1) - \frac{1}{8} = 0 \end{aligned}$$

$$\begin{aligned} c_2^2 a_{32} b_3(1) + (c_2^2 a_{42} + c_3^2 a_{43}) b_4(1) + \\ (c_2^2 a_{52} + c_3^2 a_{53} + c_4^2 a_{54}) b_5(1) - \frac{1}{12} = 0 \end{aligned}$$

$$(c_2 a_{32} a_{43}) b_4(1) + (c_2 a_{53} a_{32} + a_{54} [a_{42} c_2 + a_{43} c_3]) b_5(1) - \frac{1}{24} = 0$$

It is interesting to note that they do not make use of the simplifying assumptions (1.3.1) and (1.3.2) in this scheme.

They now use all of the previous information to solve for the c_i 's, a_{ij} 's, and the coefficients of the b_j 's simultaneously. The equations for third order are now differentiated three times with respect to θ , and then θ is set to zero in each of the

resulting equations. This gives us twelve linear equations (three for each of the order conditions) in terms of the coefficients of the b_j 's. If the remaining coefficients (recall $b_{j0} = 0$ for all j) are arranged in the order $\{b_{11}, b_{21}, b_{31}, b_{41}, b_{51}, b_{12}, b_{22}, \dots, b_{52}, b_{13}, \dots, b_{53}\}$, then the 12×15 coefficient matrix repeats a 4×5 block on its diagonal. The block augmented by three "right-hand side" vectors is:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & c_2 & c_3 & c_4 & c_5 & 0 & \frac{1}{2} & 0 \\ 0 & c_2^2 & c_3^2 & c_4^2 & c_5^2 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & c_2 a_{32} & c_2 a_{42} + c_3 a_{43} & c_2 a_{52} + c_3 a_{53} + c_4 a_{54} & 0 & 0 & \frac{1}{6} \end{bmatrix}$$

Using the above as well as taking the fourth order conditions and previous equations into consideration, Derr, Outlaw, and Sarafyan are able to solve for all coefficients with three free parameters left: c_2 , c_3 , and b_{53} . For simplicity, they chose $c_2 = \frac{1}{3}$, $c_3 = \frac{2}{3}$, and $b_{53} = 1$. By choosing these values, they obtained the tableau

0					
$\frac{1}{3}$	$\frac{1}{3}$				
$\frac{2}{3}$	$-\frac{1}{3}$	1			
1	1	-1	1		
1	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$	$b_5(\theta)$

Tableau I.IV.I

with the coefficients $b_j(\theta)$ given by

$$b_1(\theta) = \frac{3}{4}\theta^3 - \frac{13}{8}\theta^2 + \theta$$

$$b_2(\theta) = -\frac{3}{4}\theta^3 + \frac{9}{8}\theta^2$$

$$b_3(\theta) = -\frac{3}{4}\theta^3 + \frac{9}{8}\theta^2$$

$$b_4(\theta) = -\frac{1}{4}\theta^3 + \frac{8}{3}\theta^2$$

$$b_5(\theta) = \theta^3 - \theta^2$$

This is a third order C^1 CERK that is of order four at the mesh points. When $\theta = 1$, then we arrive at a well-known formula of Kutta [DOS].

Employing the same methods, Derr, Outlaw, and Sarafyan were also able to construct a ten-stage fifth order C^1 CERK that is of order six at the mesh points.

1.5 Owren and Zennaro's Development of CERKs

In two papers by Owren and Zennaro, we are presented with the clever way of stating the order conditions for continuous Runge-Kutta methods ([OZ1], [OZ2]). As we have seen earlier, equation (1.2.1) from Section 1.2 exhibits Owren and Zennaro's concise method for producing order conditions. By using this formula they have eliminated the laborious task of computing the differentials. (In order to see just how large a task this is, refer to [Sar].)

In the construction of their CERK method, Owren and Zennaro not only make use of (1.2.1), but also require $c_i = \sum_{j=1}^s a_{ij}$ and $b_j(0) = 0$. The last condition, as seen in the previous section, is needed for continuity. They also want to produce a C^1 CERK. To accomplish this, they use the last stage of one step as the first stage of the next step. Therefore they impose what they call the *stage reuse conditions*: $c_s = 1$ and $a_{sj} = b_j(1)$. They use the above conditions to produce several CERK methods with varying stage sizes. Their four-stage CERK of order three is presented here:

0				
$\frac{12}{23}$	$\frac{12}{23}$			
$\frac{4}{5}$	$-\frac{68}{375}$	$\frac{368}{375}$		
1	$\frac{31}{144}$	$\frac{529}{1152}$	$\frac{125}{384}$	
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$	$b_4(\theta)$

Tableau I.V.I

with the coefficients $b_j(\theta)$ given by

$$b_1(\theta) = \frac{41}{72} \theta^3 - \frac{65}{48} \theta^2 + \theta$$

1 4

$$b_2(\theta) = -\frac{529}{576}\theta^3 + \frac{529}{384}\theta^2$$

$$b_3(\theta) = -\frac{125}{192}\theta^3 + \frac{125}{128}\theta^2$$

$$b_4(\theta) = \theta^3 - \theta^2$$

They also produce a six-stage CERK of order four, and an eight-stage CERK of order five.

There are a few items of interest that distinguish Owren and Zennaro's CERK methods from others that we have considered previously. As stated earlier, the six-stage CERK has order four, but it does not attain any higher order when $\theta = 1$; they use *continuous* order conditions exclusively to construct their method. This is in contrast to what was done in methods considered in the previous two sections. Owren and Zennaro choose their free parameters not for simplicity, but instead so that the method is *optimal* (i.e. reduced error in the estimate of y_1 in some appropriate sense).

When reflecting on the methods presented in this chapter, it is interesting to note what assumptions and conditions are employed. All have made use of $c_i = \sum_{j=1}^s a_{ij}$, but not always the simplifying assumptions (1.3.1) and (1.3.2). Two of the methods utilize the stage reuse conditions, while the other does not. Also notice that in [Sar] no assumptions are made about the $b_j(\theta)$'s, while in the other two methods not only are we given the general form of the polynomials, but restrictions are placed on them. Most importantly, we make note of the development of the continuous order conditions for each of the methods and realize that the process can be simplified by using the conditions of Owren and Zennaro.

Chapter 2

Continuous Diagonally Implicit Runge-Kutta Methods

2.1 Introduction

Fewer researchers appear to be concerned with continuous Runge-Kutta methods which are not explicit. One explanation is that the nonlinear algebraic equations involved are, in general, coupled and thus much more expensive to solve. One remedy is to use methods for which $a_{ij} = 0$ for $j > i$ and at least one $a_{ii} \neq 0$. These schemes are called multiply diagonal implicit Runge-Kutta (MDIRK) methods.

In this chapter we present the development of some continuous MDIRKs and their related discrete methods. In Section 2.2 we show in a very detailed manner the development of a two-stage continuous MDIRK. As it turns out, this method is closely related to the known family of discrete two-stage MDIRKs of order three.

We also present the development of a three-stage continuous MDIRK of order three. We develop a particular three-stage family and then investigate what happens when we apply the continuous order conditions to a known family of three-stage MDIRKs.

2.2 Two-Stage Continuous MDIRK Method

Let us first focus our attention on the two-stage MDIRK method:

a_{11}	a_{11}	
$a_{21} + a_{22}$	a_{21}	a_{22}
	$b_1(\theta)$	$b_2(\theta)$

Tableau II.II.I

where $b_1(\theta) = b_{10} + b_{11}\theta + b_{12}\theta^2$ and $b_2(\theta) = b_{20} + b_{21}\theta + b_{22}\theta^2$. Notice that for this method we are requiring that $c_i = \sum_{j=1}^s a_{ij}$.

In order to guarantee that this is a method of order two, we must impose the conditions $b_1(\theta) + b_2(\theta) = \theta$ and $a_{11}b_1(\theta) + (a_{21} + a_{22})b_2(\theta) = \frac{1}{2}\theta^2$. These two order condition require that the following equations be satisfied:

$$b_{10} + b_{20} = 0, \quad a_{11}b_{10} + (a_{21} + a_{22})b_{20} = 0$$

$$b_{11} + b_{21} = 1, \quad a_{11}b_{11} + (a_{21} + a_{22})b_{21} = 0$$

$$b_{12} + b_{22} = 0, \quad a_{11}b_{12} + (a_{21} + a_{22})b_{22} = \frac{1}{2}$$

For convenience, use the notation $c_2 = a_{21} + a_{22}$ (whether or not we attach any particular meaning to c_2), so that now we are solving the following system of equations in order to obtain values for the b_j 's:

$$\begin{bmatrix} 1 & 1 \\ a_{11} & c_2 \end{bmatrix} \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (2.2.1)$$

$$\begin{bmatrix} 1 & 1 \\ a_{11} & c_2 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (2.2.2)$$

$$\begin{bmatrix} 1 & 1 \\ a_{11} & c_2 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} \quad (2.2.3)$$

Solving these, we obtain $b_{10} = 0$, $b_{20} = 0$, $b_{11} = 1 + \frac{a_{11}}{c_2 - a_{11}}$, $b_{21} = \frac{-a_{11}}{c_2 - a_{11}}$, $b_{12} = \frac{-\frac{1}{2}}{c_2 - a_{11}}$,

and $b_{22} = \frac{\frac{1}{2}}{c_2 - a_{11}}$, so that $b_1(\theta) = \left(\frac{a_{11}}{c_2 - a_{11}} \right) \theta + \left(\frac{-\frac{1}{2}}{c_2 - a_{11}} \right) \theta^2$ and

$b_2(\theta) = \left(\frac{-a_{11}}{c_2 - a_{11}} \right) \theta + \left(\frac{\frac{1}{2}}{c_2 - a_{11}} \right) \theta^2$. Since we do not want $c_2 = a_{11}$, let us investigate

whether this is a possibility. Let us assume $c_2 = a_{11}$. Then (2.2.3) implies $b_{12} + b_{22} = 1$ and $c_2(b_{12} + b_{22}) = \frac{1}{2}$, so that $c_2 = \frac{1}{2}$. If we consider (2.2.2), then this implies $b_{11} + b_{21} = 1$ and $c_2(b_{11} + b_{21}) = 0$, so that $c_2 = 0$. However, this contradicts $c_2 = \frac{1}{2}$; therefore $a_{11} \neq c_2$.

At this point, we have a family of continuous MDIRK methods of order two. If we wish to be more restrictive and have third order at the mesh points, then we also need the order conditions:

$$a_{11}^2 b_1(1) + c_2^2 b_2(1) = \frac{1}{3} \quad (2.2.4)$$

$$a_{11}^2 b_1(1) + (a_{21} a_{11} + a_{22} c_2) b_2(1) = \frac{1}{6} \quad (2.2.5)$$

We now want to use these additional equations to get all variables in terms of a_{11} .

Let us first solve for a_{21} . Subtracting (2.2.5) from (2.2.4) leaves us with

$$\begin{aligned} b_2(1) [c_2^2 - a_{21} a_{11} - a_{22} c_2] &= \frac{1}{6} \Leftrightarrow \left(\frac{a_{11} - \frac{1}{2}}{a_{11} - c_2} \right) [c_2^2 - a_{21} a_{11} - a_{22} c_2] = \frac{1}{6} \Leftrightarrow \\ (a_{11} - \frac{1}{2}) [c_2^2 - a_{21} a_{11} - c_2(c_2 - a_{21})] &= \frac{1}{6} (a_{11} - c_2) \Leftrightarrow (a_{11} - \frac{1}{2}) [a_{21} c_2 - a_{21} a_{11}] = \frac{1}{6} (a_{11} - c_2) \Leftrightarrow \\ (a_{11} - \frac{1}{2}) a_{21} [c_2 - a_{11}] &= \frac{1}{6} (a_{11} - c_2) \Leftrightarrow (a_{11} - \frac{1}{2}) a_{21} = -\frac{1}{6} \Leftrightarrow a_{21} = -\frac{1}{3 - 6a_{11}}, \text{ provided} \end{aligned}$$

$$a_{11} \neq c_2.$$

Now we solve for c_2 in terms of a_{11} . Notice that if the expressions for $b_1(1)$ and $b_2(1)$ are put into (2.2.4), we obtain $\left(\frac{\frac{1}{2} - c_2}{a_{11} - c_2} \right) a_{11}^2 + \left(\frac{a_{11} - \frac{1}{2}}{a_{11} - c_2} \right) c_2^2 = \frac{1}{3} \Leftrightarrow$

$$\begin{aligned} (\frac{1}{2} - c_2) a_{11}^2 + (a_{11} - \frac{1}{2}) c_2^2 &= \frac{1}{3} (a_{11} - c_2) \Leftrightarrow (\frac{1}{2} a_{11}^2 - \frac{1}{2} c_2^2) + (-c_2 a_{11}^2 + a_{11} c_2^2) = \frac{1}{3} (a_{11} - c_2) \Leftrightarrow \\ \frac{1}{2} (a_{11} - c_2) (a_{11} + c_2) - a_{11} c_2 (a_{11} - c_2) &= \frac{1}{3} (a_{11} - c_2) \Leftrightarrow \frac{1}{2} (a_{11} + c_2) - a_{11} c_2 = \frac{1}{3} \Leftrightarrow \\ c_2 &= \frac{2 - 3a_{11}}{3 - 6a_{11}}, \text{ provided } a_{11} \neq c_2. \end{aligned}$$

With these values of a_{21} and c_2 , the general tableau for the family of two-stage MDIRK methods parameterized in terms of a_{11} is:

a_{11}	a_{11}		
$\frac{2 - 3a_{11}}{3 - 6a_{11}}$	$\frac{1}{3 - 6a_{11}}$	$\frac{1 - 3a_{11}}{3 - 6a_{11}}$	$(a_{11} \neq \frac{1}{2})$
	$\frac{\frac{1}{2}}{6a_{11}^2 - 6a_{11} + 2}$	$\frac{6a_{11}^2 - 6a_{11} + \frac{3}{2}}{6a_{11}^2 - 6a_{11} + 2}$	

Tableau II.II.II

Employing this scheme, we can elect to use these values of the a_{ij} 's and c_i 's, which in turn

$$\text{give us } b_1(\theta) = \left(\frac{-6a_{11}^2 + 3a_{11}}{6a_{11}^2 - 6a_{11} + 2} \right) \theta + \left(\frac{3a_{11} - \frac{3}{2}}{6a_{11}^2 - 6a_{11} + 2} \right) \theta^2 \text{ and}$$

$$b_2(\theta) = \left(\frac{6a_{11}^2 - 3a_{11}}{6a_{11}^2 - 6a_{11} + 2} \right) \theta + \left(\frac{-3a_{11} + \frac{3}{2}}{6a_{11}^2 - 6a_{11} + 2} \right) \theta^2 \text{ for the continuous case. This is the form}$$

of the two-stage MDIRK of order two that is in addition order three at the mesh points.

In a similar work, Crouzeix [Crz] developed the following third order discrete Runge-Kutta scheme:

τ_1	τ_1	
τ_2	$2\tau_2 - 1$	$1 - \tau_2$
	$\frac{\alpha^2}{1 + \alpha^2}$	$\frac{1}{1 + \alpha^2}$

Tableau II.II.III

where $\tau_1 = \frac{1}{2} + \frac{1}{2\alpha\sqrt{3}}$ and $\tau_2 = \frac{1}{2} - \frac{\alpha}{2\sqrt{3}}$. Notice that this is actually the same as the discrete scheme developed earlier if one identifies $\alpha = \frac{1}{(a_{11} - \frac{1}{2})2\sqrt{3}}$. In Crouzeix's work, he does not reveal what approach he used to obtain this method, but he does state that this is the only third order MDIRK. Referring to the earlier computations that were used to derive tableau II.II.II, then one sees that we have proven this to be the case.

2.3 Particular Two-stage Discrete MDIRKs

Now that we have the general form for the family of third order MDIRK methods, let us turn our attention to some particular schemes. If one first chooses $a_{11} = 0$, then we obtain the tableau:

0	0	
$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	$\frac{1}{4}$	$\frac{3}{4}$

Tableau II.III.I

This is the Hammer-Hollingsworth method as found in [HNW]. If one writes this scheme as $k_1 = f(x_n, y_n)$, $k_2 = f(x_n + \frac{2}{3}h, y_n + \frac{h}{3}(k_1 + k_2))$, and $y_{n+1} = y_n + \frac{h}{4}(k_1 + 3k_2)$, then one can immediately see the benefit to using this method. The value of k_1 is simply found by evaluating the given function at (x_n, y_n) , a known value. Knowing this value explicitly

simplifies later calculations. This scheme is easier to implement than other two-stage MDIRKs, and one still has the benefit of a third order method.

Now look at the case $a_{11} = \frac{1}{3}$. The tableau becomes:

$\frac{1}{3}$	$\frac{1}{3}$	0
1	1	0
		$\frac{3}{4}$
		$\frac{1}{4}$

Tableau II.III.II

This scheme can equivalently be written as $k_1 = f(x_n + \frac{h}{3}, y_n + \frac{h}{3}k_1)$, $k_2 = f(x_n + h, y_n + hk_1)$, and $y_{n+1} = y_n + \frac{h}{4}(3k_1 + k_2)$. Notice that the benefit of this particular method is that k_2 can be determined explicitly after k_1 has been found. Again this will simplify the calculations when the method is applied to a specific problem.

Most certainly, there are other two-stage MDIRKs. Two examples are the Lobatto IIIA-method:

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
		$\frac{1}{2}$
		$\frac{1}{2}$

Tableau II.III.III

and the Lobatto IIIB-method:

0	$\frac{1}{2}$	0
1	$\frac{1}{2}$	0
		$\frac{1}{2}$
		$\frac{1}{2}$

Tableau II.III.IV

Notice that these also have the obvious benefits of well-placed zeros in the tableau. The drawback of these methods is that they are only of order two [DV].

For DIRK schemes, if we require $a_{11} = a_{22}$, then we obtain a particular subclass of methods called singly diagonal implicit Runge-Kutta (SDIRK) methods. Earlier we developed the two-stage MDIRK of order three. However, the two-stage SDIRK of order three is given by:

γ	γ	
$1-\gamma$	$1-2\gamma$	γ
	$\frac{1}{2}$	$\frac{1}{2}$

Tableau II.III.V

provided that $\gamma = \frac{1}{2} \pm \frac{1}{6}\sqrt{3}$. The benefit of using a SDIRK as opposed to an MDIRK is that the usual implementation of the scheme is more efficient computation-wise [DV].

All of the schemes that we have considered so far in Chapter 2 have been two-stage diagonally implicit Runge-Kutta methods. We examined these particular methods because they are easier to compute than the fully implicit Runge-Kutta methods. Instead of being forced to solve the k_i 's as a system of equations, they can be solved sequentially.

However, the author does want to draw attention to the Hammer-Hollingsworth method of two stages:

$\frac{1}{2} - \frac{1}{6}\sqrt{3}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{1}{6}\sqrt{3}$
$\frac{1}{2} + \frac{1}{6}\sqrt{3}$	$\frac{1}{4} + \frac{1}{6}\sqrt{3}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Tableau II.III.VI

This is the only two-stage IRK method of order four [HNW]. Implementing this scheme would be more involved, but we do get higher order than with any other two-stage method. Special transformations are possible which can significantly improve the computational requirements, though.

2.4 Stability of Two-Stage MDIRKs

One important aspect of any Runge-Kutta method is that of stability. The simple equation

$$y'(x) = \lambda y(x), \quad \lambda \in \mathbb{C}, \quad x \geq 0, \quad y(0) = y_0 \quad (2.4.1)$$

has long been acknowledged as a respectable model to predict stability of numerical methods. Applying any Runge-Kutta method to this equation gives the result $y_{n+1} = R(z)y_n$ for $z = h\lambda$, where $R: \mathbb{C} \rightarrow \mathbb{C}$ is a rational function with real coefficients. The function R is called the *stability function* of the method. The stability function of a Runge-Kutta method is given by

$$R(z) = \frac{\det(I - zA + zeb^T)}{\det(I - zA)} = \frac{P(z)}{Q(z)}$$

as found in [HW], where e is a vector each of whose components is 1. The method is called *absolutely stable* at $z \in \mathbb{C}$ if $|R(z)| \leq 1$ for the particular z . Consequently, the *absolute stability region* is the set of all points $z \in \mathbb{C}$ for which $|R(z)| \leq 1$ [DV].

There are several ways to define stability. We will give attention to a few of those various ways here. A Runge-Kutta method is called *B-stable* if the contractivity condition $\langle f(x, y) - f(x, z), y - z \rangle \leq 0$ implies for all $h \geq 0$, $\|y_1 - \hat{y}_1\| \leq \|y_0 - \hat{y}_0\|$. Here, y_1 and $\hat{y}_1$ are the numerical approximations after one step starting with initial values y_0 and $\hat{y}_0$, respectively [HW]. If the left half-plane $\operatorname{Re}(z) \leq 0$ is contained in the absolute stability region, then the method is called *A-stable* [DV]. Also, a consequence is that a Runge-Kutta method is A-stable if and only if $|R(iy)| \leq 1$ for all real y , and $R(z)$ is analytic for $\operatorname{Re}(z) < 0$. The condition $|R(iy)| \leq 1$ alone means stability on the imaginary axis and implies *I-stability*. I-stability is equivalent to the fact that the polynomial

$$E(y) = |Q(iy)|^2 - |P(iy)|^2 = Q(iy)Q(-iy) - P(iy)P(-iy) \quad (2.4.2)$$

satisfies $E(y) \geq 0$ for all $y \in \mathbb{R}$ [HW]. If a method is B-stable, then it is A-stable, and if it is A-stable, then it is also I-stable.

In order to obtain some stability results, we will make use of the following propositions found in [HW].

Proposition 1: If $E(y)$ is an even polynomial of degree $\leq \max(2(\deg P, \deg Q))$ and $R(z)$ is an approximation of order p , then $E(y) = O(y^{p+1})$ for $y \rightarrow 0$.

Proposition 2: A rational function $R(z)$ of order $p \geq 2(\deg Q) - 2$ is I-stable if and only if $|R(\infty)| \leq 1$.

We now investigate I-stability for the method in tableau II.II.II. The stability polynomial for this method is:

$$R(z) = \frac{P(z)}{Q(z)} = \frac{1 + Dz + Fz^2}{1 + Gz + Hz^2}$$

where $D = \frac{(4 - 12a_{11} + 12a_{11}^2)}{(6 - 12a_{11})}$, $F = \frac{(1 - 4a_{11} + 6a_{11}^2)}{(6 - 12a_{11})}$, $G = \frac{(\frac{1}{3} - 2a_{11})}{(-1 + 2a_{11})}$, and

$H = \frac{(-\frac{1}{3}a_{11} + a_{11}^2)}{(-1 + 2a_{11})}$. Using these polynomials for $P(z)$ and $Q(z)$, we now determine

$E(y)$. By (2.4.2), $E(y) = (H^2 - F^2)y^4 + (G^2 - D^2 + 2(F - H))y^2$, a fourth degree polynomial. By Proposition 1 with $p = 3$, the coefficient of y^2 must vanish, and thus we need only look at the coefficient $(H^2 - F^2)$ to determine if $E(y) \geq 0$. Now we need to determine for what values of a_{11} the leading coefficient $(H^2 - F^2)$ is non-negative. In our case, $H^2 - F^2 = \frac{(2a_{11} - 1)(12a_{11}^2 - 6a_{11} + 1)}{36(1 - 4a_{11} + 4a_{11}^2)}$, and this is positive for $a_{11} \geq \frac{1}{2}$. However,

one may recall that $a_{11} \neq \frac{1}{2}$ in tableau II.II.II. Therefore, the method described in tableau II.II.II is I-stable for $a_{11} > \frac{1}{2}$.

If we take advantage of Proposition 2, then we need only check if $\lim_{z \rightarrow \infty} |R(z)| < 1$, or equivalently, if $-1 \leq \frac{F}{H} \leq 1$. This is satisfied if and only if $-1 \leq \frac{1 - 4a_{11} + 6a_{11}^2}{2a_{11} - 6a_{11}^2} \leq 1$. As expected, this conditions is fulfilled if $a_{11} \geq \frac{1}{2}$. Thus we obtain the same result; the method in tableau II.II.II is I-stable for $a_{11} > \frac{1}{2}$.

Now that we know the method is I-stable for $a_{11} > \frac{1}{2}$, we need only show that $R(z)$ is analytic for $\text{Re}(z) < 0$ in order to obtain A-stability. We want to check that there are no

poles in the left-hand plane. So we determine for what values of z we have $Q(z) = 0$.

Since $Q(z) = (1 - a_{11}z) \left(1 - z \left(\frac{1 - 3a_{11}}{3 - 6a_{11}} \right) \right)$, then we see that $Q(z) = 0$ for $z = \frac{1}{a_{11}}$ and

$z = \frac{3 - 6a_{11}}{1 - 3a_{11}}$. Since we have already determined that $a_{11} > \frac{1}{2}$, it is obvious that our

singularities are in the right half-plane. Therefore, the method in tableau II.II.II is

A-stable for $a_{11} > \frac{1}{2}$.

Crouzeix stated that for the method described in tableau II.II.III, which is equivalent to that in tableau II.II.II, we have B-stability for $\alpha > 0$ [Crz].

2.5 Three-Stage Continuous MDIRK Methods

Earlier we looked at two-stage MDIRKs; now we turn our attention to three-stage MDIRKs. The general form of continuous three-stage MDIRKs is

c_1	a_{11}		
c_2	a_{21}	a_{22}	
c_3	a_{31}	a_{32}	a_{33}
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$

Tableau II.V.I

where $b_1(\theta) = b_{10} + b_{11}\theta + b_{12}\theta^2 + b_{13}\theta^3$, $b_2(\theta) = b_{20} + b_{21}\theta + b_{22}\theta^2 + b_{23}\theta^3$, and $b_3(\theta) = b_{30} + b_{31}\theta + b_{32}\theta^2 + b_{33}\theta^3$. This method will be third order if the order conditions for a three-stage MDIRK are satisfied:

$$b_1(\theta) + b_2(\theta) + b_3(\theta) = \theta \quad (2.5.1)$$

$$c_1 b_1(\theta) + c_2 b_2(\theta) + c_3 b_3(\theta) = \frac{1}{2} \theta^2 \quad (2.5.2)$$

$$c_1^2 b_1(\theta) + c_2^2 b_2(\theta) + c_3^2 b_3(\theta) = \frac{1}{3} \theta^3 \quad (2.5.3)$$

$$a_{11} c_1 b_1(\theta) + (a_{21} c_1 + a_{22} c_2) b_2(\theta) + (a_{31} c_1 + a_{32} c_2 + a_{33} c_3) b_3(\theta) = \frac{1}{6} \theta^3 \quad (2.5.4)$$

Here, we assume that the simplifying assumption $c_i = \sum_{j=1}^s a_{ij}$ holds. If we also make use the simplifying assumption $\sum_{j=1}^s a_{ij} c_j = \frac{1}{2} c_i^2$, then solving for the $b_j(\theta)$'s is a matter of solving

$$\begin{bmatrix} 1 & 1 & 1 \\ c_1 & c_2 & c_3 \\ c_1^2 & c_2^2 & c_3^2 \\ \frac{1}{2} c_1^2 & \frac{1}{2} c_2^2 & \frac{1}{2} c_3^2 \end{bmatrix} \begin{bmatrix} b_1(\theta) \\ b_2(\theta) \\ b_3(\theta) \end{bmatrix} = \begin{bmatrix} \theta \\ \frac{1}{2} \theta^2 \\ \frac{1}{3} \theta^3 \\ \frac{1}{6} \theta^3 \end{bmatrix}$$

Notice the benefit of employing this simplifying assumption. The last equation is a multiple of the third, so that instead of trying to solve four equations with three unknowns, we may reduce the problem to solving three equations with three unknowns.

Through some tedious but straightforward computations, we obtain

$$b_1(\theta) = \frac{\frac{1}{3} \theta^3 - \frac{1}{2} \theta^2 (c_3 + c_2) + \theta c_3 c_2}{(c_3 - c_1)(c_2 - c_1)} \quad (2.5.5)$$

$$b_2(\theta) = \frac{-\frac{1}{3} \theta^3 + \frac{1}{2} \theta^2 (c_3 + c_1) - \theta c_1 c_3}{(c_3 - c_2)(c_2 - c_1)} \quad (2.5.6)$$

$$b_3(\theta) = \frac{\frac{1}{3} \theta^3 - \frac{1}{2} \theta^2 (c_2 + c_1) - \theta c_1 c_2}{(c_3 - c_1)(c_3 - c_2)} \quad (2.5.7)$$

where $c_1 \neq c_2 \neq c_3 \neq c_1$. Notice the great increase in the complexity of the calculations for the $b_j(\theta)$'s just by progressing from two stages to three. It is interesting to note that we could have obtained the same equations of the $b_j(\theta)$'s by integrating the fundamental

Lagrange polynomials:

$$b_1'(\theta) = \frac{(\theta - c_3)(\theta - c_2)}{(c_3 - c_1)(c_2 - c_1)}$$

$$b'_2(\theta) = \frac{(\theta - c_3)(\theta - c_1)}{(c_3 - c_2)(c_1 - c_2)}$$

$$b'_3(\theta) = \frac{(\theta - c_2)(\theta - c_1)}{(c_2 - c_3)(c_1 - c_3)}$$

using $b_j(0) = 0$ for $j = 1, 2, 3$. In the case of distinct c_i 's, this choice can be understood via the connection between implicit Runge-Kutta methods and a process known as collocation, provided we satisfy the simplifying assumption $c_i = \sum_{j=1}^s a_{ij}$ [HNW].

Obviously, this is a much simpler way of computing (2.5.5) - (2.5.7) than the method we used earlier.

For reasons of symmetry, we now want to look at a particular case where $c_1 = a_{11}$, $c_2 = \frac{1}{2}$, and $c_3 = 1 - a_{11}$. For these values, $b_1(\theta) = \frac{\frac{1}{3}\theta^3 - (\frac{3}{4} - \frac{1}{2}a_{11})\theta^2 + \frac{1}{2}(1 - a_{11})\theta}{2w^2}$, $b_2(\theta) = \frac{-\frac{2}{3}\theta^3 + \theta^2 - 2a_{11}(1 - a_{11})\theta}{2w^2}$, and $b_3(\theta) = \frac{\frac{1}{3}\theta^3 - (\frac{1}{2}a_{11} + \frac{1}{4})\theta^2 + \frac{1}{2}a_{11}\theta}{2w^2}$, where $w = \frac{1}{2} - a_{11}$ and $a_{11} \neq \frac{1}{2}$.

Since we are using (2.5.5), (2.5.6), and (2.5.7) for the $b_j(\theta)$'s, we know that the particular case we are considering is third order. However, now we want make use of the condition (2.5.4) in order to solve for some of the a_{ij} coefficients. In this case, condition (2.5.4) becomes $Ab_1(\theta) + Bb_2(\theta) + Cb_3(\theta) = \frac{1}{6}\theta^3$, where $A = a_{11}^2$, $B = a_{21}a_{11} + \frac{1}{2}a_{22}$, and $C = a_{31}a_{11} + \frac{1}{2}a_{32} + a_{33}(1 - a_{11})$. This condition is equivalent to the following:

$$\left(\frac{1}{6w^2}\right)A + \left(\frac{-2}{6w^2}\right)B + \left(\frac{1}{6w^2}\right)C = \frac{1}{6} \quad (2.5.8)$$

$$-\left(\frac{3}{4} - \frac{1}{2}a_{11}\right)A + B - \left(\frac{1}{2}a_{11} + \frac{1}{4}\right)C = 0 \quad (2.5.9)$$

$$\frac{1}{2}(1 - a_{11})A - 2a_{11}(1 - a_{11})B + \frac{1}{2}a_{11}C = 0 \quad (2.5.10)$$

By utilizing *Mathematica*, we are able to reduce the algebraic problem to the following three equations:

$$(-1 + 2a_{11})(-1 - 4a_{21} + 4a_{31} + 2a_{11} + 2a_{32}) = 0 \quad (2.5.11)$$

$$(-1 + 2a_{11})(-4a_{21} + 2a_{31} + 4a_{31}a_{11} + a_{32} + 2a_{11}a_{32}) = 0 \quad (2.5.12)$$

$$a_{11}(-1 + 2a_{11})(-4a_{21} + 2a_{31} + 4a_{21}a_{11} + a_{32}) = 0 \quad (2.5.13)$$

Notice that $a_{11} = \frac{1}{2}$ would satisfy all of the equations, but we have already eliminated that as a possibility. Let us assume $a_{11} \neq 0$. Then from (2.5.9) it must be that

$$-4a_{21} + 2a_{31} + 4a_{31}a_{11} + a_{32} + 2a_{11}a_{32} = 0 \text{ and from (2.5.10) that}$$

$$-4a_{21} + 2a_{31} + 4a_{21}a_{11} + a_{32} = 0. \text{ Subtracting these two equations results in}$$

$$(4a_{31} + 2a_{32} - 4a_{21})a_{11} = 0, \text{ so that } 4a_{31} + 2a_{32} - 4a_{21} = 0. \text{ From (2.5.8) we have}$$

$$4a_{31} + 2a_{32} - 4a_{21} = 1 - 2a_{11}. \text{ Combining this with the previous equations implies}$$

$$1 - 2a_{11} = 0, \text{ so that } a_{11} = \frac{1}{2}. \text{ However, we know that this is not possible. Therefore, } a_{11}$$

must equal zero. We now use the equations (2.5.11), (2.5.12), (2.5.13), the fact $a_{11} = 0$, and the convention $c_i = \sum_{j=1}^s a_{ij}$ in order to solve for the other coefficients. We obtain the

continuous third order MDIRK:

0	0		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
1	$\frac{1-a_{32}}{2}$	a_{32}	$\frac{1-a_{32}}{2}$
	$b_1(\theta)$	$b_2(\theta)$	$b_3(\theta)$

Tableau II.V.II

where $b_1(\theta) = \frac{2}{3}\theta^3 - \frac{3}{4}\theta^2 + \frac{1}{2}\theta$, $b_2(\theta) = -\frac{4}{3}\theta^3 + 2\theta^2$, and $b_3(\theta) = \frac{2}{3}\theta^3 - \frac{1}{2}\theta^2$.

If we wish for this to be fourth order at the mesh points, then we need to satisfy the order conditions [HNW]:

$$\sum_i b_i(1)c_i^3 = \frac{1}{4} \quad (2.5.14)$$

$$\sum_{i,j} b_i(1)c_i a_{ij} c_j = \frac{1}{8} \quad (2.5.15)$$

$$\sum_{i,j} b_i(1)a_{ij}c_j^2 = \frac{1}{2} \quad (2.5.16)$$

$$\sum_{i,j,k} b_i(1)a_{ij}a_{jk}c_k = \frac{1}{24} \quad (2.5.17)$$

As one may verify, (2.5.14) and (2.5.15) are satisfied for any value of a_{32} . As for (2.5.16) and (2.5.17), they require that $a_{32} = 1$. Therefore the underlying discrete fourth order MDIRK method is given by:

0	0		
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	
1	0	1	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Tableau II.V.III

This is actually Butcher's Lobatto formula of order four as found in [HNW].

Crouzeix developed a more general family of three-stage *discrete* MDIRKs of order four. The tableau for his method parameterized in terms of α_1 and α_2 is

τ_1	τ_1		
τ_2	$\tau_2 - \alpha$	α	
τ_3	$\tau_3 - \beta - \gamma$	β	γ
	b_1	b_2	b_3

Tableau II.V.IV

where $\tau_1 = \frac{1+\alpha_1}{2}$, $\tau_2 = \frac{1+\alpha_2}{2}$, $\tau_3 = \frac{1+\alpha_3}{2}$, $\alpha = \frac{\alpha_3 - \alpha_1 + 3\alpha_1\alpha_3}{2(1+3\alpha_1\alpha_3)}$,

$\beta = \frac{(\alpha_1 - \alpha_3)(1+2\alpha_1)}{2(1+3\alpha_1\alpha_2)(\alpha_2 - \alpha_1)}$, $\gamma = \frac{1-\alpha_3}{2}$, $b_1 = \frac{1+3\alpha_2\alpha_3}{3(\alpha_1 - \alpha_2)(\alpha_1 - \alpha_3)}$,

$b_2 = \frac{1+3\alpha_1\alpha_3}{3(\alpha_2 - \alpha_1)(\alpha_2 - \alpha_3)}$, $b_3 = \frac{1+3\alpha_1\alpha_2}{3(\alpha_3 - \alpha_1)(\alpha_3 - \alpha_2)}$, $3\alpha_1\alpha_2\alpha_3 + \alpha_1 + \alpha_2 + \alpha_3 = 0$, and

$\alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_1$. One may notice that the discrete MDIRK method in tableau II.V.III is just one member of the family found in tableau II.V.IV with $\alpha_1 = -1$, $\alpha_2 = 0$, and $\alpha_3 = 1$.

In tableau II.V.IV, Crouzeix presented us with the only scheme of three-stage MDIRK methods of order four. As a special case, we have the following three-stage SDIRK methods of order four given by

γ	γ	γ	
$\frac{1}{2}$	$\frac{1}{2} - \gamma$	$1 - 4\gamma$	γ
$1 - \gamma$	2γ		
	$\frac{1}{24(\frac{1}{2} - \gamma)^2}$	$1 - \frac{1}{12(\frac{1}{2} - \gamma)^2}$	$\frac{1}{24(\frac{1}{2} - \gamma)^2}$

Tableau II.V.V

where γ is a zero of the polynomial $24x^3 - 36x^2 + 12x - 1$, namely

$\gamma = \gamma_1 = \frac{1}{2} + \frac{1}{3}\sqrt{3}\cos(\frac{\pi}{18})$ or $\gamma = \gamma_{2,3} = \frac{1}{2} - \frac{1}{6}\sqrt{3}\cos(\frac{\pi}{18}) \pm \frac{1}{2}\sin(\frac{\pi}{18})$, which is a well-known scheme studied in the literature (e.g. see [DV]).

2.6 A Family of Continuous Three-Stage MDIRKs

In Section 2.5, we were presented with Crouzeix's family of three-stage MDIRKs of order four. Let us now take this general discrete scheme and impose equations (2.5.1) - (2.5.4) (the continuous order conditions for MDIRKs of order three) on the method described in tableau II.V.IV. It can be verified easily by utilizing *Mathematica* that (2.5.1) - (2.5.3) are satisfied with the $b_j(\theta)$'s given by (2.5.5) - (2.5.7); this reduces the problem to satisfying (2.5.4). In this case, we choose not to make use of the simplifying assumption $\frac{1}{2}c_i^2 = \sum_{j=1}^s a_{ij}c_j$. In order to satisfy (2.5.4), we must actually satisfy three conditions: the sum of the coefficients of θ must equal zero, the sum of the coefficients of θ^2 must equal zero, and the sum of the coefficients of θ^3 must equal $\frac{1}{6}$. Again, because of the complexity of the problem, we employ *Mathematica* to help us solve the problem. By doing this we find that we must satisfy the conditions:

$$\frac{12\alpha_1^2 + 12\alpha_1 + 4}{36\alpha_1^2 - 12} = \frac{1}{6}$$

$$\frac{-3(\alpha_1 + 1)^2}{12\alpha_1^2 - 4} = 0$$

$$\frac{(\alpha_1 + 1)^2}{12\alpha_1^2 - 4} = 0$$

All of these equations are satisfied if, and only if, $\alpha_1 = -1$. Therefore it is necessary and sufficient that $\alpha_1 = -1$ to obtain a third order continuous MDIRK from tableau II.V.IV. However, it is necessary that α_1 be greater than zero for the method to be A-stable [Crz], so A-stability is not possible.

Even with $\alpha_1 = -1$, we still have the free parameter α_2 . Several ways to utilize this parameter come to mind. For example, we could use this parameter and attempt to minimize the error constant. In order to do this, we must first realize that we want the polynomial $R(z) = \frac{P(z)}{Q(z)}$ to approximate e^z as closely as possible. (To see why we desire

this, refer to (2.4.1).) So we compute $Q(z)e^z - P(z)$ and attempt to obtain the highest order possible. For this problem we need only expand e^z through the terms

$1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4$ and compute $P(z)$ and $Q(z)$, so that $P(z) = A + Bz + Cz^2 + Dz^3$ and $Q(z) = A + Ez + Fz^2$, where $A = -4 + 12\alpha_2$, $B = -3 + 6\alpha_2 - 3\alpha_2^2$, $C = -1 + \alpha_2 - 2\alpha_2^2$, $D = -\frac{1}{6} - \frac{1}{2}\alpha_2^2$, $E = 1 + 6\alpha_2 - 3\alpha_2^2$, and $F = \alpha_2 + \alpha_2^2$. When we compute $Q(z)e^z - P(z)$, all terms of degree less than five are eliminated, so that we now know that the error is $O(z^5)$. The coefficient of z^5 , which is our error constant, is $\frac{1}{120} + \frac{2}{120}\alpha_2 + \frac{5}{120}\alpha_2^2$. We

initially try to eliminate this term, but there are no real roots of the coefficient of z^5 . When we attempt to minimize this value, it is found that $\alpha_2 = -\frac{1}{3}$ is the optimal choice, and the value of the error constant is $\frac{1}{130}$. Another option which might be explored would be the unconventional choice of complex parameters to eliminate the term $\frac{1}{120} + \frac{2}{120}\alpha_2 + \frac{5}{120}\alpha_2^2$. A further thought might be to choose α_2 to somehow expand the region of absolute stability to its fullest capacity.

As stated earlier, the study of continuous Runge-Kutta methods is an active field of research. In this paper we have attempted to display our efforts to obtain a better understanding of these methods. We have researched several papers in order to become well-versed in the knowledge of these processes. The focus of this paper has been on explicit and diagonally implicit continuous Runge-Kutta methods for first order ordinary differential equations, but one is able to find in various journals continuous methods for second order ordinary differential equations. The author is aware of these methods, but only for explicit schemes. The author has not encountered formulas for continuous *fully implicit* Runge-Kutta methods. The information in this thesis could be a starting point for developing these schemes.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [Crz] M. Crouzeix (1975): *Sur L'approximation des équations différentielles opérationnelles linéaires par de méthodes de Runge-Kutta*, Thèse, Univ. Paris.
- [DV] K. Dekker and J. G. Verwer (1984): *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, Elsevier Science Publishers B. V., Amsterdam.
- [DOS] L. Derr, C. Outlaw, and D. Sarafyan (1993): *A New Method for Derivation of Continuous Runge-Kutta Formulas*, Computers Math. Applic., Vol. 26, No. 3, pp. 7-13.
- [EJT] W. Enright, K. Jackson, S. Nørsett, and P. Thomsen (1986): Interpolants for Runge-Kutta Formulas, AMC Transactions on Mathematical Software, Vol. 12, No. 3, pp 193-218.
- [HNW] E. Hairer, S. P. Nørsett, and G. Wanner (1987): *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag, New York.
- [HW] E. Hairer and G. Wanner (1991): *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer-Verlag, New York.
- [OZ1] B. Owren and M. Zennaro (1991): *Order Barriers for Continuous Explicit Runge-Kutta Methods*, Mathematics of Computation, Vol. 56, No. 194, pp. 645-661.
- [OZ2] B. Owren and M. Zennaro (1992): *Derivation of Efficient, Continuous, Explicit Runge-Kutta Methods*, SIAM J. Sci. Stat. Comput., Vol. 13, No. 6, pp. 1488-1501.
- [PT] G. Papageorgiou and Ch. Tsitouras (1989): *Scaled Runge-Kutta-Nyström Methods for the Second Order Differential Equation*, Intern. J. Computer Math., Vol. 28, pp. 139-150.
- [Sar] D. Sarafyan (1984): *Continuous Approximate Solution of Ordinary Differential Equations and Their Systems*, Computers Math. Applic., Vol. 10, No. 2, pp. 139-159.
- [Sha] L. Shampine (1985): *Interpolation for Runge-Kutta Methods*, SIAM J. Numer. Anal., Vol. 22, No. 5, pp. 1014-1027.
- [Sim] T. Simos (1993): *Runge-Kutta-Nyström Interpolants for the Numerical Integration of Special Second-Order Periodic Initial-Value Problems*, Computers Math. Applic., Vol 26, No. 12, pp. 7-15.

VITA

Pamela Lynn Stanfield Lancaster was born in Owensboro, Kentucky on March 4, 1969. She graduated from Roane County High School in Kingston, Tennessee in June 1987. The following August she entered the University of Kentucky. As an undergraduate, she worked one summer for Martin Marietta in Oak Ridge, Tennessee. In May 1991, she received the degree of Bachelor of Science in Mathematics with high distinction from the University of Kentucky. In August 1991, she entered the University of Tennessee as a graduate student and was given a fellowship by the university. During her tenure as a graduate student, she also worked as a graduate teaching associate for the Department of Mathematics. She received her Master of Science degree in Mathematics from the University of Tennessee in December 1994.