Almost Runge-Kutta Methods for Stiff and Non-Stiff Problems

Nicolette Rattenbury

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy, The University of Auckland, 2005.

Abstract

Ordinary differential equations arise frequently in the study of the physical world. Unfortunately many cannot be solved exactly. This is why the ability to solve these equations numerically is important.

Traditionally mathematicians have used one of two classes of methods for numerically solving ordinary differential equations. These are linear multistep methods and Runge–Kutta methods. General linear methods were introduced as a unifying framework for these traditional methods. They have both the multi-stage nature of Runge–Kutta methods as well as the multi-value nature of linear multistep methods. This extremely broad class of methods, besides containing Runge–Kutta and linear multistep methods as special cases, also contains hybrid methods, cyclic composite linear multistep methods and pseudo Runge–Kutta methods.

In this thesis we present a class of methods known as Almost Runge–Kutta methods. This is a special class of general linear methods which retains many of the properties of traditional Runge–Kutta methods, but with some advantages.

Most of this thesis concentrates on explicit methods for non-stiff differential equations, paying particular attention to a special fourth order method which, when implemented in the correct way, behaves like order five. We will also introduce low order diagonally implicit methods for solving stiff differential equations.

Acknowledgements

During the course of my PhD I have been very fortunate to receive the guidance and support of many wonderful people.

My supervisor Prof. John Butcher is an inspiration. His enthusiasm is contagious. It is hard not to be excited about an idea when discussing it with him. Nobody could ask for a more patient, caring and supportive supervisor and friend.

Dr Robert Chan, my co-supervisor, has also been very supportive. He was always available when I wished to discuss my work.

My office mate, Dr Allison Heard, has been a wonderful mentor and friend. The many hours she has spent proof-reading my work has been invaluable. As have the many hours spent poring over the odd cryptic crossword!

Our weekly numerical analysis meetings have also been a great source of support. They have given me the chance to present my work informally and receive feedback. Apart from those I have already mentioned, I would particularly like to thank Dr Will Wright, Dr Shirley Huang, Angela Tsai and Dr Helmut Podhaisky. They have all become good friends as well as supportive colleagues.

Finally I would like to thank my husband, Dr Nicholas Rattenbury. There is a great quotation from the famous Winnie the Pooh that sums up how I feel about him *"If you live to be 100, I want to live to be 100 minus one day, so I never have to live without you"*. I am extremely lucky to have found someone who believes in me as much as he does.

Contents

Abstract **iii**

Acknowledgements **v**

Contents **vii**

List of Tables **xi**

List of Figures **xiii**

1 Introduction **1**

1.1 Ordinary differential equations 2

1.1.1 Existence and uniqueness of solutions 2

1.1.2 Stiff differential equations 3

1.2 Delay differential equations 5

1.3 A brief history of numerical methods 5

2 General linear methods **9**

2.1 Consistency and stability 11

2.1.1 Stability regions 12

2.2 Tree theory 15

2.3 Order 17

2.3.1 Algebraic analysis of order 18

Expansion of the exact solution 19

Elementary weights 22

Expansion of the numerical solution 23

2.4 Examples of general linear methods . . . 23
2.4.1 Runge–Kutta methods . . . 24
2.4.2 Linear multistep methods . . . 24
Adams methods . . . 25
BDF methods . . . 26
2.4.3 DIMSIMs . . . 27
2.4.4 IRKS methods . . . 30

3 Almost Runge–Kutta methods 35
3.1 General form of explicit ARK methods . . . 36
3.2 Order and related conditions . . . 38
3.3 Interpolation . . . 41
3.4 Methods with $s = p$ . . . 42
3.4.1 RK stability . . . 42
3.4.2 Third order methods with three stages . . . 47
Order conditions . . . 47
Derivation of methods . . . 48
Some example methods . . . 50
Interpolation . . . 50
3.4.3 Fourth order methods with four stages . . . 53
Order conditions . . . 53
Derivation of methods . . . 56
Classification of the methods . . . 57
Some example methods . . . 61
Interpolation . . . 62
3.5 Methods with $s = p + 1$ . . . 63
3.5.1 RK-stability . . . 63
3.5.2 Third order methods with four stages . . . 64
Order conditions . . . 65
Derivation of methods . . . 65
Some example methods . . . 68
Interpolation . . . 69
3.5.3 Fourth order method with five stages . . . 71

Order conditions . . . 71
Derivation of methods . . . 72
Some example methods . . . 75
Interpolation . . . 76

4 A special 'fifth' order method 79

4.1 Introduction . . . 79
4.2 Obtaining order 5 performance . . . 83
4.3 Interpolation . . . 87
4.4 Error estimation . . . 88
4.5 Optimising these methods . . . 89
4.5.1 Fifth order error coefficients . . . 90
4.5.2 Sixth order error coefficients . . . 91

5 Stiff ARK methods 99

5.1 Introduction . . . 99
5.2 Order 3 stiff ARK methods . . . 104
Order conditions . . . 106
Derivation of methods . . . 106
Some example methods . . . 107
5.3 Order 4 stiff ARK methods . . . 107
Order conditions . . . 109
Derivation of methods . . . 109
Some example methods . . . 110
5.4 Starting the method . . . 111

6 Numerical Experiments 115

6.1 Non-stiff methods . . . 115
6.1.1 Fixed stepsize . . . 115
6.1.2 Fixed variable stepsize . . . 123
6.1.3 Variable stepsize . . . 129
6.1.4 DDEs . . . 131
6.2 Stiff methods . . . 133

7 Conclusions 135

A Test Problems **139**

A.1 DETest problems 139

A.2 Stiff problems 144

A.2.1 Oregonator 144

A.2.2 HIRES 144

A.2.3 Prothero-Robinson problem 145

A.3 Delay differential equation problems 145

A.3.1 Equation 1.1.6 145

A.3.2 Equation 1.1.10 145

A.3.3 Equation 1.1.12 146

A.3.4 Equation 1.4.1 146

A.3.5 Equation 1.4.6 147

A.3.6 Equation 1.4.9 147

References **149**

Index **155**

List of Tables

2.1 Trees up to order 6. 15
2.2 Number of trees of orders 1 to 10. 16
2.3 Order, density and symmetry of the trees up to order 5. 17
2.4 Elementary differentials for trees up to order 5. 20
2.5 Composition of elementary weight functions up to order 5. 21
2.6 Types of DIMSIMs 28
3.1 Trees up to order 5 omitted due to the simplifying assumptions. 39
4.1 Algebraic analysis of the special 5 stage method. 82
6.1 Comparison of error behaviours for fixed and variable stepsizes for problem A5 using method ARK45. 124
6.2 Comparison of error behaviours for fixed and variable stepsizes for problem B5 using method ARK45. 124
6.3 Comparison of error behaviours for fixed and variable stepsizes for problem C5 using method ARK45. 125
6.4 Comparison of error behaviours for fixed and variable stepsizes for problem D5 using method ARK45. 125
6.5 Comparison of error behaviours for fixed and variable stepsizes for problem E5 using method ARK45. 126
6.6 Comparison of error behaviours for fixed and variable stepsizes for problem A5 using Dormand and Prince. 126
6.7 Comparison of error behaviours for fixed and variable stepsizes for problem B5 using Dormand and Prince. 127
6.8 Comparison of error behaviours for fixed and variable stepsizes for problem C5 using Dormand and Prince. 127

6.9 Comparison of error behaviours for fixed and variable stepsizes for problem D5 using Dormand and Prince. 128
6.10 Comparison of error behaviours for fixed and variable stepsizes for problem E5 using Dormand and Prince. 128

List of Figures

1.1 Implicit Euler and explicit Euler methods applied to differential equation 1.1. 4

2.1 Stability regions for explicit Runge–Kutta and composite Adams–Bashforth methods, for orders 1 to 4. 14

2.2 The order of a general linear method. 18

4.1 The $D1$ problem solved using method (4.1) with 100 equal sized steps. An interpolator has been used to estimate the solution $\frac{1}{3}$ and $\frac{2}{3}$ of the way through each step. 88

4.2 Optimising our special 'fifth' order method. Solving for the free parameters c_2 and c_3. 93

5.1 Error constant for λ in A-stability interval 105

5.2 Values of $R(\infty)$ 105

5.3 Error constant for λ in A-stability interval, where $\lambda_1 = 0.394338$ and $\lambda_2 = 1.28058$. . 108

5.4 Values of $R(\infty)$ in A-stability interval, where $\lambda_1 = 0.394338$ and $\lambda_2 = 1.28058$. . . . 108

6.1 Comparison between RK45, RK56, ARK4, ARK451 and ARK452 using constant stepsize for the class A DETest problems. 118

6.2 Comparison between RK45, RK56, ARK4, ARK451 and ARK452 using constant stepsize for the class B DETest problems. 119

6.3 Comparison between RK45, RK56, ARK4, ARK451 and ARK452 using constant stepsize for the class C DETest problems. 120

6.4 Comparison between RK45, RK56, ARK4, ARK451 and ARK452 using constant stepsize for the class D DETest problems. 121

6.5 Comparison between RK45, RK56, ARK4, ARK451 and ARK452 using constant stepsize for the class E DETest problems. 122

6.6 Comparison between RK56 and ARK45 using variable stepsize for a selection of the DETest problems. 130

6.7 Comparison between RK56 and ARK45 using variable stepsize for a selection of DDE problems. 132
6.8 Comparison between DIARK3, DIARK4, DIRK3 and DIRK4 on a selection of stiff problems. 134

CHAPTER 1

Introduction

> Mathematics is not a careful march down a well-cleared highway, but a journey into a strange wilderness, where the explorers often get lost. Rigour should be a signal to the historian that the maps have been made, and the real explorers have gone elsewhere.
>
> W.S. Anglin

Ordinary differential equations arise frequently in the study of the physical world. Unfortunately many cannot be solved exactly. This is why the ability to obtain accurate numerical approximate solutions is important.

In this chapter we will give a summary of the types of differential equations we are interested in, as well as give a brief background to the numerical methods that have traditionally been used to solve them.

Chapter 2 gives an introduction to general linear methods, which were introduced as a unifying framework for traditional methods. We will also see how much of the theory for traditional methods can be generalised to encompass general linear methods.

In chapter 3 we introduce Almost Runge–Kutta methods. These are a special class of general linear methods which were introduced to retain many of the desirable properties of Runge–Kutta methods, with some of the advantages of linear multi–step methods. This chapter outlines most of the theory of these methods.

Chapter 4 pays particular attention to a family of special fourth order methods which, when implemented in the correct way, behave like order five.

Stiff Almost Runge–Kutta methods are introduced in chapter 5. These methods can be used to solve ordinary differential equations which exhibit the property known as stiffness.

In chapter 6 we give the results from some numerical experiments, where we compare the performance of the methods described in this thesis with traditional Runge–Kutta methods in solving standard test problems.

Finally, chapter 7 gives the conclusions from this study and outlines further work in this area.

1.1 Ordinary differential equations

Ordinary differential equations can be represented in one of two ways. The first is known as non-autonomous form. The ordinary differential equation (ODE) is written as

$$y'(x) = f(x, y(x)).$$

The variable x is called the independent variable and $y(x)$ is the solution to the differential equation. It should be noted that $y(x)$ can be a vector-valued function, going from $\mathbb{R} \to \mathbb{R}^m$, where m is the dimension of the differential equation.

In the second form, $y'(x)$ does not depend directly on x, except as a parameter of $y(x)$. This second form is known as autonomous form and can be written as

$$y'(x) = f(y(x)).$$

In this thesis, we will mainly consider equations in autonomous form. This does not lead to a loss of generality, as any non-autonomous system may be written in autonomous form by adding the equation $x' = 1$ to the system.

If we add the initial condition $y_0 = y(x_0)$ to the system of equations we get the initial value problem (IVP)

$$y'(x) = f(y(x)), \qquad y_0 = y(x_0).$$

1.1.1 Existence and uniqueness of solutions

Before we look at ways to numerically approximate the solution to an initial value problem it is important to consider whether the solution is unique, or even if indeed a solution exists at all. There are many criteria for determining these two considerations, but the most commonly used approach is the Lipschitz condition.

Definition 1.1 *The function $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is said to satisfy a Lipschitz condition in its second variable if there exists a constant L, known as a Lipschitz constant, such that for any $x \in [a, b]$ and Y, $Z \in \mathbb{R}^N$, $||f(x, Y) - f(x, Z)|| \leq L||Y - Z||$.*

This definition is used in the following theorem.

Theorem 1.1 *Consider an initial value problem*

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

where $f : [a, b] \times \mathbb{R}^N \to \mathbb{R}^N$ is continuous in its first variable and satisfies a Lipschitz condition in its second variable. Then there exists a unique solution to this problem.

Proof: A proof of this can be found in many books. See, for example, [6]. ■

1.1.2 Stiff differential equations

There is no agreed formal definition of what stiffness is. Stiff problems can best be recognised from the behaviour they can display when approximated by standard numerical methods. Although the exact solution is extremely stable, the numerical solution can be extremely unstable. Explicit methods cannot be used to solve this type of problem as the bounded stability regions of these methods mean that they have to take excessively small stepsizes, even when the problem being solved is relatively smooth. That is, the stability requirements rather than accuracy requirements drive the sizes of the steps taken. This behaviour is usually observed in problems that have some components that decay much more rapidly than other components.

Due to this behaviour, ordinary differential equations have been divided into stiff and non-stiff problems. Different types of numerical methods are needed for the different problem types. This is a relatively new idea. It was not until 1952 that Curtiss and Hirschfelder [27] realised that different types of methods work better on some classes of problems.

To see the effects of stiffness we will consider the simple initial value problem

$$y' = -100(y - \cos x), \qquad y(0) = 0. \tag{1.1}$$

As we can see in Figure 1.1, when we apply the implicit Euler method the numerical solution follows the exact solution fairly closely, taking only 5 steps. However, if we try to solve the same initial value problem using the explicit Euler method the numerical solution oscillates around the exact solution, even using as many as 75 steps.

With stiff problems, sometimes the Lipschitz condition can be too pessimistic. Instead we consider the idea of a *one-sided* Lipschitz condition.

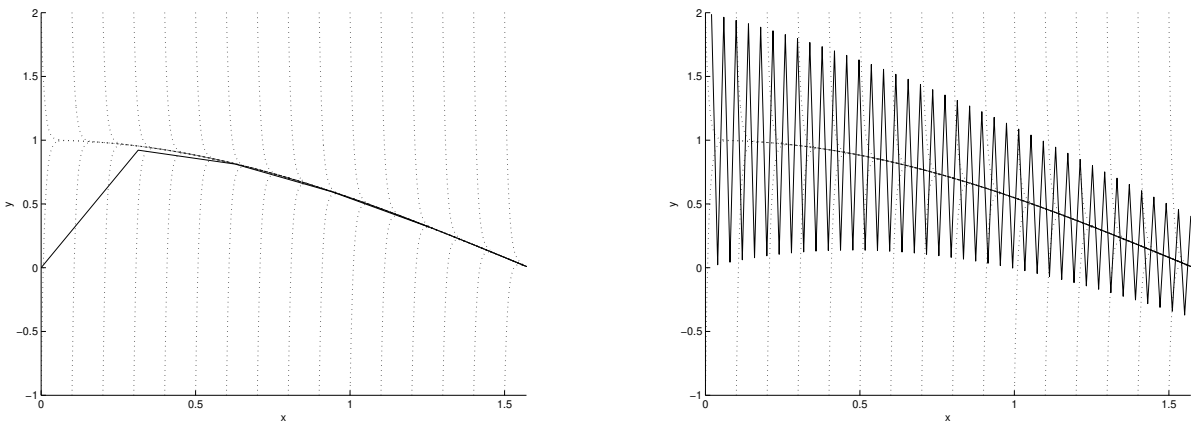


Figure 1.1: Implicit Euler (left) and explicit Euler (right) methods applied to differential equation 1.1.

Definition 1.2 *The function f satisfies a one-sided Lipschitz condition, with one-sided Lipschitz constant l, if for all $x \in [a, b]$ and all $u, v \in \mathbb{R}^N$,*

$$\langle f(x, u) - f(x, v), u - v \rangle \leq l||u - v||^2.$$

It is possible that a problem will have a very large Lipschitz constant, but a manageable one-sided Lipschitz constant. This can help us find realistic growth estimates for the effect of perturbations, as can be seen in the following theorem.

Theorem 1.2 *If f satisfies a one-sided Lipschitz condition with constant l, and y and z are each solutions of*

$$y'(x) = f(x, y(x)),$$

then for all $x \geq x_0$,

$$||y(x) - z(x)|| \leq \exp(l(x - x_0))\, ||y(x_0) - z(x_0)||.$$

Proof: A proof of this can be found in [6]. ■

1.2 Delay differential equations

In many cases ordinary differential equations are not the most natural way to model a physical system. Consider, for example, population growth. This is commonly modelled using the differential equation

$$N'(t) = k\left(1 - \frac{N(t)}{P}\right)N(t), \tag{1.2}$$

where k and P are positive constants. Although this is a reasonable model, it is perhaps more realistic that the rate of change of the population at time t is dependent on the population at some time $t - r$, $r > 0$. This changes equation (1.2) to

$$N'(t) = k\left(1 - \frac{N(t-r)}{P}\right)N(t).$$

This type of equation is known as a delay differential equation (DDE). Delay differential equations depend not only on the solution at time t, but also on the solution at some previous time or times. The general form of a delay differential equation is

$$y'(x) = f(y(x), y(x-\tau_1), ..., y(x-\tau_n)).$$

The terms $\tau_1, ..., \tau_n$ are known as the delays or time lags. The complexity of these delays determines the type of the delay equation. If the delays are constant we have a constant delay differential equation. In the case where $\tau_1, ..., \tau_n$ are dependent on x we have a variable delay differential equation. Finally, if the delays are functions of both x and y the delay differential equation is called state dependent.

One main difference between delay differential equations and ordinary differential equations is that delay differential equations require an initial value function, $\phi(x)$ such that for $x < x_0$, we require $y(x) = \phi(x)$, rather than just an initial value. It turns out that even if $f(y, z)$, $\tau_1(x, y), ..., \tau_n(x, y)$ and $\phi(x)$ are C^∞, the solution $y(x)$ is seldom better than C^0 for $x > x_0$. These discontinuities propagate throughout the interval of integration. Any numerical solver needs a strategy for handling these discontinuities.

1.3 A brief history of numerical methods

The first numerical method for solving ordinary differential equations was devised by Euler in the 1760's and republished in his collected works in 1913 [31]. The idea behind this method is very simple. The interval to be integrated over is divided into sub-intervals of size h_i, where i is the step number. The stepsizes can either be the same, giving us constant stepsize, or of

varying lengths, leading to a variable stepsize implementation. In practice, the stepsizes in a variable stepsize implementation are chosen during the integration process. In each step we take we assume that the value of the derivative does not change much over the step. Euler's method then states that the approximation to the solution at the end of the step is given by

$$y_{n+1} \approx y_n + h_n f(x_n, y_n).$$

When $y_n = y(x_n)$, the Taylor series expansion of this Euler approximation is equal to that for the Taylor series expansion of $y(x_{n+1})$ up to and including terms in the first power of h_n. The method is therefore said to be of order one. There are two natural ways of extending this result to improve the accuracy.

The first generalisation of Euler's method was by Adams and Bashforth [1] in 1883. Their methods use more information from the past to take a step forward. The Adams–Bashforth methods are a special case of a class of methods known as linear multistep methods, which take the form

$$y_n = \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k} + h\left(\beta_0 f(y_n) + \beta_1 f(y_{n-1}) + \cdots + \beta_k f(y_{n-k})\right).$$

In the case of the Adams–Bashforth methods $\alpha_1 = 1$, $\alpha_2, ..., \alpha_k = 0$ and $\beta_0 = 0$. An extension of this idea was developed by Moulton [52] in which $\beta_0 \neq 0$. This gives the methods an implicit structure. Changing the stepsize under this formulation is difficult as the integration coefficients need to be recalculated in each step. In 1962 Nordseick [53] proposed a method which alleviates this problem. The values passed from step to step are the scaled $k+1$ derivatives, including the order zero derivative.

In practice linear multistep methods tend to be implemented as a predictor-corrector pair. An approximation to y_n is predicted using an Adams–Bashforth method and is then corrected using an Adams–Moulton method. This idea was proposed by Milne [50] in 1949. Two advantages of implementing the methods in this way are that the implementation is now explicit in nature and they have a simple type of error estimator known as Milne's device. The scaled difference between the two approximations can be used to approximate the error.

Backward differentiation methods were introduced by Curtiss and Hirshfelder [27] in 1952. For these methods $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. These methods play a special role in the solution of stiff problems, despite not being A-stable for methods of order 3 or above. The most widely used adaptive codes for solving stiff differential equations are based on backward differentiation methods. The first code was written by Gear [36] in 1971, making use of Nordseick representation. For a Nordsieck method of order p, the data imported into step number n consists of

approximations to

$$y(x_{n-1}),\, hy'(x_{n-1}),\, \frac{1}{2!}h^2y''(x_{n-1}),\, \cdots,\, \frac{1}{p!}h^py^{(p)}(x_{n-1}).$$

The output quantities, therefore, approximate

$$y(x_n),\, hy'(x_n),\, \frac{1}{2!}h^2y''(x_n),\, \cdots,\, \frac{1}{p!}h^py^{(p)}(x_n). \tag{1.3}$$

To change the stepsize from h to rh, the quantites in (1.3) are scaled by powers of the scale factor r, giving

$$y(x_n),\, rhy'(x_n),\, \frac{1}{2!}(rh)^2y''(x_n),\, \cdots,\, \frac{1}{p!}(rh)^py^{(p)}(x_n).$$

This is then used as the input to step number $n+1$.

A large proportion of the theory of linear multistep methods was developed by Dahlquist [28].

The other obvious generalisation of Euler's method is to use more derivative values per step. Methods of this type were first devised in 1895 by Runge [61]. Further contributions were made by Heun [40] and Kutta [48]. Kutta completely characterised the family of fourth order methods and developed the first fifth order method. These methods are now known as Runge–Kutta methods and take the form

$$Y_i = y_{n-1} + h\sum_{j=1}^{s} a_{ij}f(x_{n-1}+c_jh, Y_j), \quad i = 1, ..., s \tag{1.4}$$

$$y_n = y_{n-1} + h\sum_{i=1}^{s} b_if(x_{n-1}+c_ih, Y_i), \tag{1.5}$$

where s is the number of internal stages. Many contributions were also made by Nyström who developed special methods for second order differential equations [54]. It was not until the 1950's that methods of order six were developed by Hu$\check{\text{t}}$a [43], [44]. Since then many people have developed methods of higher orders.

Another important development of these methods was the introduction of error estimators, enabling variable stepsize implementation. The first error estimators were developed by Richardson [59] in 1927. These estimators require each step to be repeated using two steps with half the original stepsize. Although effective, this method of error estimation is expensive. The standard approach now used is embedded methods, where a Runge–Kutta method of one order is embedded inside a higher order Runge–Kutta method. The difference between these two approximations can be used to approximate the error. This idea was originally developed by Merson [49] in 1957, but considerable work has also been done in this area by Fehlberg [32], [33], Verner [66] and Dormand and Prince [29].

CHAPTER 2

General linear methods

> Mathematics is like checkers in being suitable for the young, not too difficult, amusing, and without peril to the state.
>
> PLATO

General linear methods were introduced by Butcher [4] as a unifying framework for traditional methods. They have both the multi-stage nature of Runge–Kutta methods as well as the multi-value nature of linear multistep methods.[1] This extremely broad class of methods, besides containing Runge–Kutta and linear multistep methods as special cases, also contains hybrid methods, cyclic composite linear multistep methods and pseudo Runge–Kutta methods.

For compactness of notation we write Y and F for the vector of Y_i and F_i values respectively, where $Y_i \approx y(x_n + c_i h)$ is the approximation at the i-th internal stage and $F_i = f(x, Y_i)$. As with a Runge-Kutta method, the vector $c = [c_1, c_2, \cdots, c_s]^T$, is called the vector of abscissae. For ease of computation it is usually preferred that the stages approximate the solution within the current integration interval i.e. $0 \leq c_i \leq 1$, however this isn't always the case. We also write $y^{[n-1]}$ for the vector of approximations imported into step n and $y^{[n]}$ for the quantities computed in this step and exported for use by the following step. The detailed computation is now based on the formula

$$Y = h(A \otimes I)F + (U \otimes I)y^{[n-1]} \tag{2.1}$$

for the stages, and

$$y^{[n]} = h(B \otimes I)F + (V \otimes I)y^{[n-1]} \tag{2.2}$$

[1] A method is multi-value if it propogates more than one value for each component. In contrast, a method is multi-stage if it utilizes intermediate values on each step to generate the new values to be propogated.

for the output values, where I is the identity matrix of size equal to the differential equation system to be solved. The Kronecker product of two matrices is given by the following definition.

Definition 2.1 *If G is an $m \times n$ matrix and H is a $p \times q$ matrix, then the Kronecker product $G \otimes H$ is the $mp \times nq$ block matrix*

$$G \otimes H = \begin{bmatrix} g_{11}H & \cdots & g_{1n}H \\ \vdots & \ddots & \vdots \\ g_{m1}H & \cdots & g_{mn}H \end{bmatrix}$$

$$= \begin{bmatrix} g_{11}h_{11} & g_{11}h_{12} & \cdots & g_{11}h_{1q} & \cdots & \cdots & g_{1n}h_{11} & g_{1n}h_{12} & \cdots & g_{1n}h_{1q} \\ g_{11}h_{21} & g_{11}h_{22} & \cdots & g_{11}h_{2q} & \cdots & \cdots & g_{1n}h_{21} & g_{1n}h_{22} & \cdots & g_{1n}h_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ g_{11}h_{p1} & g_{11}h_{p2} & \cdots & g_{11}h_{pq} & \cdots & \cdots & g_{1n}h_{p1} & g_{1n}h_{p2} & \cdots & g_{1n}h_{pq} \\ \vdots & \vdots & & \vdots & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \ddots & \vdots & \vdots & & \vdots \\ g_{m1}h_{11} & g_{m1}h_{12} & \cdots & g_{m1}h_{1q} & \cdots & \cdots & g_{mn}h_{11} & g_{mn}h_{12} & \cdots & g_{mn}h_{1q} \\ g_{m1}h_{21} & g_{m1}h_{22} & \cdots & g_{m1}h_{2q} & \cdots & \cdots & g_{mn}h_{21} & g_{mn}h_{22} & \cdots & g_{mn}h_{2q} \\ \vdots & \vdots & \ddots & \vdots & & & \vdots & \vdots & \ddots & \vdots \\ g_{m1}h_{p1} & g_{m1}h_{p2} & \cdots & g_{m1}h_{pq} & \cdots & \cdots & g_{mn}h_{p1} & g_{mn}h_{p2} & \cdots & g_{mn}h_{pq} \end{bmatrix}.$$

With a slight abuse of notation, equations (2.1) and (2.2) are often written in the form

$$\left[\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_s \\ \hline y_1^{[n]} \\ \vdots \\ y_r^{[n]} \end{array}\right] = \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] \left[\begin{array}{c} hf(Y_1) \\ hf(Y_2) \\ \vdots \\ hf(Y_s) \\ \hline y_1^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{array}\right], \tag{2.3}$$

where s is the number of internal stages and r is the number of values passed from step to step.

To begin computation with a general linear method, certain values in addition to the intial values for the ODE are needed. These are determined by special starting methods, such as those detailed in section 2.3.

2.1 Consistency and stability

As with linear multistep methods, a general linear method needs to be consistent and stable in order to give meaningful results.

At the very least we would expect our method to be able to solve the trivial initial value problem $y'(x) = 0,\ y(0) = a$, exactly at the beginning and end of each step. Therefore, we would like to ensure

$$y^{[n-1]} = uy(x_{n-1}) + O(h),$$
$$y^{[n]} = uy(x_n) + O(h),$$

for a vector u, which is called the pre-consistency vector. Applying a general linear method to the problem $y'(x) = 0$ gives

$$Y^{[n]} = Uy^{[n-1]},$$
$$y^{[n]} = Vy^{[n-1]}.$$

This leads to the following definition.

Definition 2.2 *A general linear method is 'preconsistent' if there exists a vector u such that*

$$e = Uu,$$
$$u = Vu,$$

where e is a vector of all ones.

We would also like a method to be able to solve the simple initial value problem $y'(x) = 1$, $y(x_0) = 0$, exactly at the beginning and end of each step. If the quantities being passed from step to step are linear combinations of the solution $y(x)$ and the scaled derivative $y'(x)$ we require

$$Y^{[n]} = ey(x_n) + chy'(x_n) + O(h^2)$$
$$y^{[n-1]} = uy(x_{n-1}) + vhy'(x_{n-1}) + O(h^2)$$
$$y^{[n]} = uy(x_n) + vhy'(x_n) + O(h^2),$$

where the vector v is called the consistency vector. Applying a general linear method to the problem $y'(x) = 1$, $y(x_0) = 0$ gives

$$Y^{[n]} = Aeh + Uy^{[n-1]},$$
$$y^{[n]} = Beh + Vy^{[n-1]}.$$

Using the exact solution $y(x) = x - x_0$ and the equations above leads to the following definition.

Definition 2.3 *A general linear method is 'consistent' if it is preconsistent with preconsistency vector u and there exists a vector v such that*

$$u + v = Be + Vv.$$

Stability is also necessary to obtain meaningful results. Stability guarantees that errors introduced in a step do not grow without bound in subsequent steps. A general linear method is stable if the solution to the trivial differential equation $y'(x) = 0$ is bounded. Applying a general linear method to this differential equation gives

$$y^{[n]} = Vy^{[n-1]} = V^n y^{[0]}.$$

This leads to the following definition.

Definition 2.4 *A general linear method is 'stable' if there exists a constant C such that for all $n = 1, 2, ...,$ $||V^n|| \leq C$.*

As with linear multistep methods, it is known that stability and consistency are necessary and sufficient for convergence of general linear methods. This was shown by Butcher in [4]. A definition of convergence is given here.

Definition 2.5 *A general linear method is 'convergent' if for any initial value problem*

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

subject to the Lipschitz condition $||f(y)-f(z)|| \leq L||y-z||$, there exists a non-zero vector $u \in \mathbb{R}^r$, and a starting procedure $\phi : (0, \infty) \to \mathbb{R}^r$, such that for all $i = 1, 2, ..., r$, $\lim_{h\to 0} \phi_i(h) = u_i y(x_0)$, and such that for any $\bar{x} > x_0$, the sequence of vectors $y^{[n]}$, computed using n steps with stepsize $h = (\bar{x} - x_0)/n$ and using $y^{[0]} = \phi(h)$ in each case converges to $uy(\bar{x})$.

2.1.1 Stability regions

As with Runge–Kutta methods and linear multistep methods, the linear stability of general linear methods is studied by considering the scalar test problem

$$y' = qy.$$

Applying equation (2.3) to this problem gives

$$Y = AhqY + Uy^{[n-1]} \tag{2.4}$$

$$y^{[n]} = BhqY + Vy^{[n-1]} \ . \tag{2.5}$$

Rearranging equation (2.4) and substituting into equation (2.5) gives

$$y^{[n]} = M(hq)y^{[n-1]},$$

where

$$M(z) = V + zB(I - zA)^{-1}U,$$

and $z = hq$. The matrix M is known as the stability matrix of the method.

The stability function of the method is determined by the characteristic polynomial of M, as given in the following definition.

Definition 2.6 *The 'stability function' for a general linear method with stability matrix $M(z)$ is the polynomial $\Phi(w, z)$*

$$\Phi(w, z) = \det(wI - M(z)).$$

The 'stability region' is the subset of the complex plane such that if z is in this subset, then

$$\sup_{n=1}^{\infty} ||M(z)^n|| < \infty. \tag{2.6}$$

The solution to equation (2.6) has a decaying norm, and if z lies in this region, then for this linear problem, the numerical solution obtained by (2.3) decays as well.

The traditional definitions of A-stability and L-stability can be slightly modified to apply to general linear methods.

Definition 2.7 *A general linear method is 'A-stable' if $M(z)$ is power bounded for every z in the left half complex plane.*

Definition 2.8 *A general linear method is 'L-stable' if it is A-stable and $\rho(M(\infty)) = 0$.*

Most other types of stability can also be modified to apply to general linear methods, but this is not required for this work.

The stability function of a general linear method is more complicated than the stability function of a Runge–Kutta method or linear multistep method. One possible way of simplifying

this function is to make it equivalent to the stability function of one of the traditional methods. We would like the stability region to take up as much of the left half complex plane as possible, hence giving good stability properties.

If we compare the stability regions of different methods it becomes apparent that the number of stages has the greatest effect on the size of the stability region. To make the comparison between Runge–Kutta methods and linear multistep methods fair we should use the stability region of s compositions of the linear multistep method, where s is the number of stages of the Runge–Kutta method. This composition gives a linear multistep method with s stages.

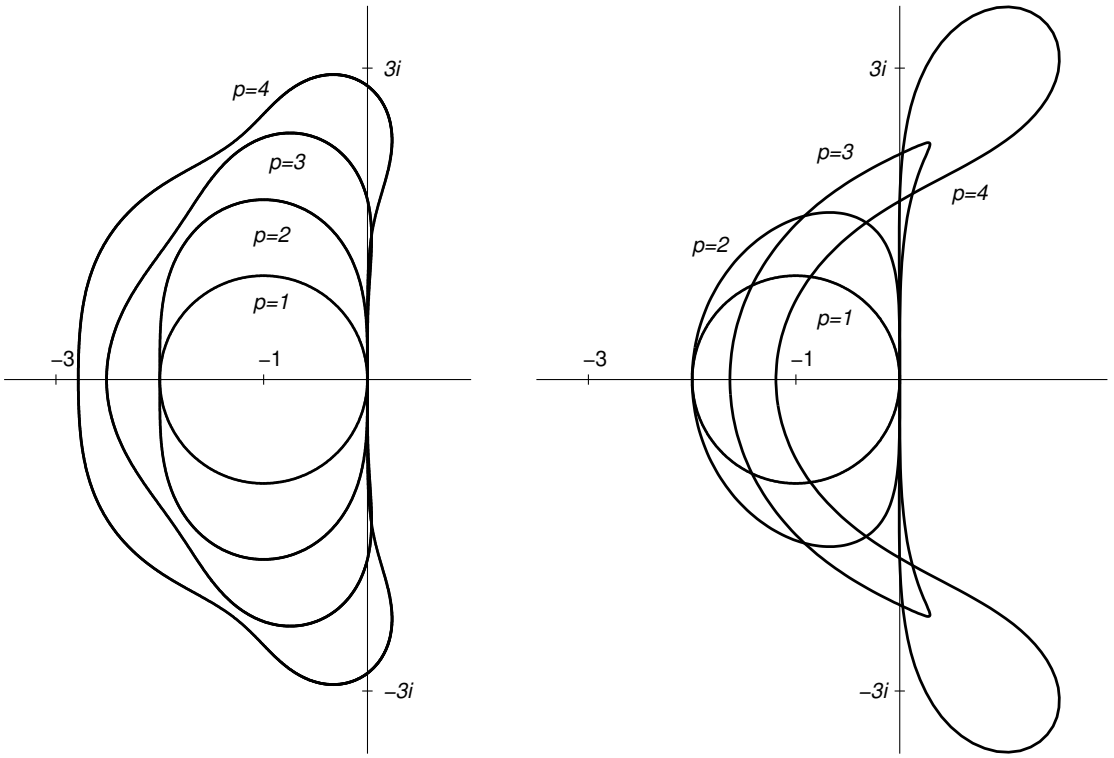


Figure 2.1: Stability regions for explicit Runge–Kutta (left) and composite Adams–Bashforth methods (right), for orders 1 to 4.

The stability regions of explicit Runge–Kutta methods and composite Adams–Bashforth methods of orders 1 to 4 are shown in Figure 2.1. It is clear from the figure that Runge–Kutta methods have the more desirable stability properties. This leads to the following definition.

Definition 2.9 *If a general linear method has a stability function which takes the special form*

$$\Phi(w, z) = \det(wI - M(z)) = w^{r-1}(w - R(z)),$$

where $R(z)$ is the stability region of a Runge–Kutta method, then the method is said to have Runge–Kutta stability.

t_0 $\emptyset$	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}
t_{20}	t_{21}	t_{22}	t_{23}	t_{24}	t_{25}	t_{26}	t_{27}	t_{28}	t_{29}
t_{30}	t_{31}	t_{32}	t_{33}	t_{34}	t_{35}	t_{36}	t_{37}		

Table 2.1: Trees up to order 6.

Trees up to order 6. Each vertex is denoted by a dot. The order of a tree is equal to the number of vertices.

This is equivalent to the stability matrix having only one non-zero eigenvalue, which is $R(z)$.

2.2 Tree theory

For a convenient development of the order of a method, we need to introduce some basic tree theory. This theory will be used in the next section, and throughout the rest of this thesis.

A tree is a rooted graph which contains no circuits. The symbol τ is used to represent the tree with only one vertex. All rooted trees can be represented using τ and the operation $[t_1, ..., t_m]$. This operation takes the roots of the trees $t_1, \ldots, t_m$ and joins them to a new root. This is known as grafting.

We first need to introduce some definitions. The order of a tree is a measure of how big the tree is.

Order	1	2	3	4	5	6	7	8	9	10
Number of trees	1	1	2	4	9	20	48	115	286	719
Cumulative total	1	2	4	8	17	37	85	200	486	1205

Table 2.2: Number of trees of orders 1 to 10.

Definition 2.10 *The order of the tree t is defined by*

$$r(t) = \begin{cases} 1, & \text{if } t = \tau \\ 1 + r(t_1) + \cdots + r(t_m), & \text{if } t = [t_1, ..., t_m] \end{cases}$$

In other words, the order of a tree is the number of vertices the tree has. The trees up to order 6 can be seen in Table 2.1. In Table 2.2 the number of trees of each order up to order ten are given, along with the number of trees of order less than or equal to that order. We see that the number of trees increases quickly.

The height of a tree is $k-1$, where k is the number of vertices in the longest path beginning with the root.

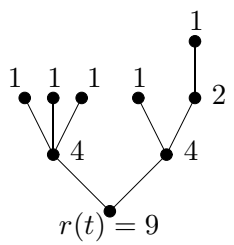
The density of a tree is a measure of 'non-bushyness'. The higher the density the less bushy the tree is.

Definition 2.11 *The density of the tree $t = [t_1, ..., t_m]$ is defined by*

$$\gamma(t) = \begin{cases} 1, & \text{if } t = \tau \\ r(t)\gamma(t_1)\gamma(t_2)\cdots\gamma(t_m), & \text{if } t = [t_1, ..., t_m] \end{cases}$$

A simple way of finding the density of a tree is to attach to each vertex a number that is equal to the number of vertices above it plus one. The density is then equal to the product of the numbers attached to the vertices.

Example: The tree represented by $[[\tau^3], [\tau, [\tau]]]$



t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$r(t)$	1	2	3	3	4	4	4	4	5	5	5	5	5	5	5	5	5
$\gamma(t)$	1	2	3	6	4	8	12	24	5	10	15	30	20	20	40	60	120
$\sigma(t)$	1	1	2	1	6	1	2	1	24	2	2	1	2	6	1	2	1

Table 2.3: Order, density and symmetry of the trees up to order 5.

$$\begin{aligned}\gamma(t) &= 9 \times 4 \times 4 \times 2 \\ &= 288\end{aligned}$$

■

A bushy tree is defined to be a tree of height one, which therefore has a density of $r(t)$. A tall tree is defined to be a tree of height $r(t)-1$, which therefore has a density of $r(t)!$. Examples of bushy trees are t_2, t_3, t_5, t_9, and t_{18}. Examples of tall trees are t_2, t_4, t_8 and t_{17}.

The symmetry of a tree is the order of the automorphism group of t. The mapping of a tree onto itself is a mapping that preserves the root and the tree structure. It is a measure of how symmetric the tree is.

Definition 2.12 *The symmetry of the tree $t = [t_1^{n_1}, ..., t_m^{n_m}]$, where $t_1, .., t_m$ are all distinct is defined by*

$$\sigma(t) = \begin{cases} 1, & \text{if } t = \tau \\ n_1!\, n_2! \ldots n_m! \sigma(t_1)^{n_1} \ldots \sigma(t_m)^{n_m}, & \text{if } t = [t_1^{n_1}, ..., t_m^{n_m}] \end{cases}$$

A high value of σ indicates a highly symmetric tree.

The order, density and symmetry of trees up to order 5 can be found in Table 2.3.

2.3 Order

As many general linear methods are multi-value methods they require a starting procedure to obtain an initial vector, $y^{[0]}$, from the initial value y_0. If we let $\overline{Y}_1, \ldots, \overline{Y}_{\overline{s}}$ be the internal stages, the starting procedure can be defined as

$$\begin{aligned}\overline{Y} &= hS_{11} f(\overline{Y}) + S_{12} y_0 \\ y^{[0]} &= hS_{21} f(\overline{Y}) + S_{22} y_0.\end{aligned}$$

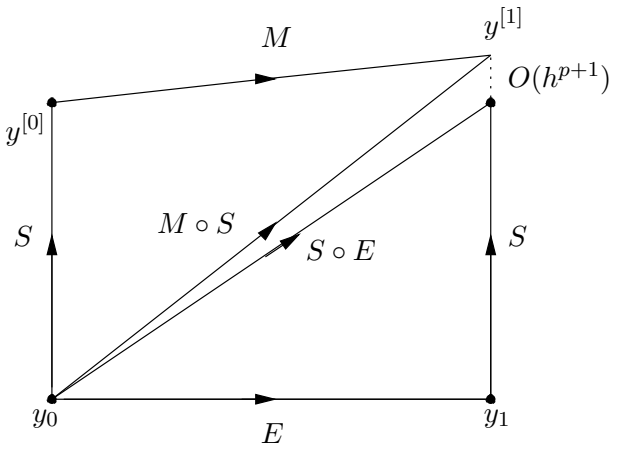


Figure 2.2: The order of a general linear method.

This can be written as the $(\overline{s}+r)\times(\overline{s}+1)$ partitioned tableau:

$$\left[\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & S_{22} \end{array}\right],$$

where $\overline{s}$ is the number of internal stages of the starting procedure and r is the number of initial approximations required. For preconsistency it is required that $S_{22} = u$ and $S_{12} = \overline{e}$, where $\overline{e}$ is the vector of length $\overline{s}$, with each component equal to 1.

If a method is of order p it is generally the case that each of the r components of $y^{[0]}$ will be of order at least p.

The order of a method can now be defined in relation to a starting method. If the starting method, S, is applied to a problem, followed by one step of the method M the result is $M \circ S$. The exact solution shifted forward one step is represented by the shift operator E. If it were possible to take one step forward in time using E then apply the starting method the result would be $S \circ E$. As we can see in Figure 2.2, a method is of order p if the difference between these two approximations is $O(h^{p+1})$. In general, the first component of the solution vector is an approximation to $y(x_n)$. This means it is only the first component that is required to be $O(h^{p+1})$ to have a method of order p.

2.3.1 Algebraic analysis of order

As with traditional methods, to determine the order of a general linear method we compare the Taylor series expansions of the exact and numerical solutions.

Expansion of the exact solution

The Taylor series expansion of the exact solution is given by

$$y(x+h) = y(x) + y'(x) + \frac{h^2 y''(x)}{2!} + \frac{h^3 y'''(x)}{3!} + \cdots ,$$

where $y'(x) = f(y(x))$.

Using the chain rule to evaluate each term gives

$$y(x+h) = y(x) + f(y(x)) + \frac{h^2 f'(y(x))(f(y(x)))}{2!} + \frac{h^3}{3!}\left(f''(y(x))(f(y(x)), f(y(x))) + f'(y(x))(f'(y(x))(f(y(x))))\right) + \cdots$$

Each of these individual terms were named elementary differentials by Butcher [3]. There is a direct relationship between elementary differentials and trees, leading to the following definition.

Definition 2.13 *For any $t \in T$, the elementary differential, $F(t)$, for a function f is defined by:*

$$F(t)(y(x)) = \begin{cases} y(x), & \text{if } t = \emptyset, \\ f(y(x)), & \text{if } t = \tau \\ f^{(m)}(F(t_1), F(t_2), \ldots, F(t_m))(y(x)), & \text{if } t = [t_1, t_2, \ldots, t_m]. \end{cases}$$

Each elementary differential can easily be found uniquely from its associated rooted tree where each vertex is associated with the nth derivative of f, where n is the number of children that vertex has. The elementary differentials for trees up to order 5 are shown in Table 2.4.

The nth derivative of $y(x)$ can be found by taking a linear combination of the elementary differentials of the trees of order n. This leads to the following theorem.

Theorem 2.1 *If $y(x)$ is n times differentiable then*

$$y^{(n)}(x) = \sum_{r(t)=n} \epsilon(t) F(t)(y(x)),$$

where $\epsilon(t)$ is the number of ways of labelling a tree such that if (i, j) is a labelled edge, then $i < j$. The value of $\epsilon(t)$ is

$$\epsilon(t) = \frac{r(t)!}{\gamma(t)\sigma(t)}.$$

Proof: A proof of this can be found in [6]. ■

	t	$F(t)(y)$
t_1		f
t_2		$f'f$
t_3		$f''(f,f)$
t_4		$f'f'f$
t_5		$f'''(f,f,f)$
t_6		$f''(f,f'f)$
t_7		$f'f''(f,f)$
t_8		$f'f'f'f$
t_9		$f^{(4)}(f,f,f,f)$
t_{10}		$f'''(f,f,f'f)$
t_{11}		$f''(f,f''(f,f))$
t_{12}		$f''(f,f'f'f)$
t_{13}		$f''(f'f,f'f)$
t_{14}		$f'(f'''(f,f,f))$
t_{15}		$f'f''(f,f'f)$
t_{16}		$f'f'f''(f,f)$
t_{17}		$f'f'f'f'f$

Table 2.4: Elementary differentials for trees up to order 5.

t_0	$\emptyset$	$\beta(t_0)$
t_1	.	$\alpha(t_1)\beta(t_0) + \beta(t_1)$
t_2	ı	$\alpha(t_2)\beta(t_0) + \alpha(t_1)\beta(t_1) + \beta(t_2)$
t_3	ν	$\alpha(t_3)\beta(t_0) + \alpha(t_1)^2\beta(t_1) + 2\alpha(t_1)\beta(t_2) + \beta(t_3)$
t_4	ı	$\alpha(t_4)\beta(t_0) + \alpha(t_2)\beta(t_1) + \alpha(t_1)\beta(t_2) + \beta(t_4)$
t_5	ψ	$\alpha(t_5)\beta(t_0) + \alpha(t_1)^3\beta(t_1) + 3\alpha(t_1)^2\beta(t_2) + 3\alpha(t_1)\beta(t_3) + \beta(t_5)$
t_6	ȷ	$\alpha(t_6)\beta(t_0) + \alpha(t_1)\alpha(t_2)\beta(t_1) + \alpha(t_2)\beta(t_2) + \alpha(t_1)^2\beta(t_2) + \alpha(t_1)\beta(t_3) + \alpha(t_1)\beta(t_4) + \beta(t_6)$
t_7	ϒ	$\alpha(t_7)\beta(t_0) + \alpha(t_3)\beta(t_1) + \alpha(t_1)^2\beta(t_2) + 2\alpha(t_1)\beta(t_4) + \beta(t_7)$
t_8	ı	$\alpha(t_8)\beta(t_0) + \alpha(t_4)\beta(t_1) + \alpha(t_2)\beta(t_2) + \alpha(t_1)\beta(t_4) + \beta(t_8)$
t_9	ψ	$\alpha(t_9)\beta(t_0) + \alpha(t_1)^4\beta(t_1) + 4\alpha(t_1)^3\beta(t_2) + 6\alpha(t_1)^2\beta(t_3) + 4\alpha(t_1)\beta(t_5) + \beta(t_9)$
t_{10}	ψ	$\alpha(t_{10})\beta(t_0) + \alpha(t_1)^2\alpha(t_2)\beta(t_1) + (2\alpha(t_1)\alpha(t_2) + \alpha(t_1)^3)\beta(t_2) + (\alpha(t_2) + \alpha(t_1)^2)\beta(t_3) + \alpha(t_1)^2\beta(t_4) + 2\alpha(t_1)\beta(t_6) + \alpha(t_1)\beta(t_5) + \beta(t_{10})$
t_{11}	ν	$\alpha(t_{11})\beta(t_0) + \alpha(t_1)\alpha(t_3)\beta(t_1) + \alpha(t_3)\beta(t_2) + \alpha(t_1)^3\beta(t_2) + \alpha(t_1)^2\beta(t_3) + 2\alpha(t_1)^2\beta(t_4) + \alpha(t_1)\beta(t_7) + 2\alpha(t_1)\beta(t_6) + \beta(t_{11})$
t_{12}	ȷ	$\alpha(t_{12})\beta(t_0) + \alpha(t_1)\alpha(t_4)\beta(t_1) + \alpha(t_4)\beta(t_2) + \alpha(t_1)\alpha(t_2)\beta(t_2) + \alpha(t_2)\beta(t_3) + \alpha(t_1)^2\beta(t_4) + \alpha(t_1)\beta(t_6) + \alpha(t_1)\beta(t_8) + \beta(t_{12})$
t_{13}	ψ	$\alpha(t_{13})\beta(t_0) + \alpha(t_2)^2\beta(t_1) + 2\alpha(t_1)\alpha(t_2)\beta(t_2) + 2\alpha(t_2)\beta(t_4) + \alpha(t_1)^2\beta(t_3) + 2\alpha(t_1)\beta(t_6) + \beta(t_{13})$
t_{14}	ϒ	$\alpha(t_{14})\beta(t_0) + \alpha(t_5)\beta(t_1) + \alpha(t_1)^3\beta(t_2) + 3\alpha(t_1)^2\beta(t_4) + 3\alpha(t_1)\beta(t_7) + \beta(t_{14})$
t_{15}	ϒ	$\alpha(t_{15})\beta(t_0) + \alpha(t_6)\beta(t_1) + \alpha(t_1)\alpha(t_2)\beta(t_2) + \alpha(t_2)\beta(t_4) + \alpha(t_1)^2\beta(t_4) + \alpha(t_1)\beta(t_7) + \alpha(t_1)\beta(t_8) + \beta(t_{15})$
t_{16}	ϒ	$\alpha(t_{16})\beta(t_0) + \alpha(t_7)\beta(t_1) + \alpha(t_3)\beta(t_1) + \alpha(t_1)^2 + \beta(t_4) + 2\alpha(t_1)\beta(t_8) + \beta(t_{16})$
t_{17}	ı	$\alpha(t_{17})\beta(t_0) + \alpha(t_8)\beta(t_1) + \alpha(t_4)\beta(t_2) + \alpha(t_2)\beta(t_4) + \alpha(t_1)\beta(t_8) + \beta(t_{17})$

Table 2.5: Composition of elementary weight functions up to order 5.

Elementary weights

Before we look at the Taylor expansion of the numerical approximation we need several definitions.

An elementary weight function is a mapping from trees to the real numbers. There are two special elementary weight functions which we are interested in. The first of these is the ith derivative operator.

Definition 2.14 *Let D_i be the ith derivative operator. Then for $i \in \mathbb{N}$*

$$D_i(t) = \begin{cases} \dfrac{i!}{\gamma(t)}, & \text{if } r(t) = i \\ 0, & \text{if } r(t) \neq i. \end{cases}$$

Provided that $y(x)$ is sufficiently smooth in the neighbourhood of x, the ith derivative operator maps $y(x)$ to $h_i y^{(i)}(x)$. The most common derivative operator we will be using is D_1, which we will simplify to D. From the above definition we obtain

$$D(t) = \begin{cases} 1, & \text{if } t = \tau \\ 0, & \text{if } t \neq \tau. \end{cases}$$

The second elementary weight function of special interest is

$$E^{(n)}(t) = \frac{n^{r(t)}}{\gamma(t)}.$$

This corresponds to the exact solution of the differential equation, as represented by the Picard iteration scheme. In the case $\epsilon = 1$ we get the exact elementary weight function

$$E(t) = \frac{1}{\gamma(t)}, \qquad \text{for all } t \in T. \tag{2.7}$$

The reverse exact elementary weight function is also useful. This is given by

$$E^{-1}(t) = \frac{(-1)^{r(t)}}{\gamma(t)}, \qquad \text{for all } t \in T.$$

The final definition we need before we can continue is the composition of two elementary weight functions.

Definition 2.15 *The composition rule for elementary weight functions, α and β, is given by*

$$(\alpha\beta)(t) = \beta(\emptyset)\alpha(t) + \beta(t) + \sum_{u<t} \beta(u)\alpha(t\backslash u), \qquad \forall t \in T, \tag{2.8}$$

where $u < t$ denotes any proper subtree u sharing the same root with the tree t, and $t\backslash u$ denotes the remainder of the tree t after deleting the subtree u from it. We will let $\alpha(t\backslash u)$ be the product of α of the trees that make up $t\backslash u$.

This rule defines the output when the output of one elementary weight function is used as input to another elementary weight function. It was first published by Butcher [5].

This rule greatly simplifies in the case where the second operator is the ith derivative operator, giving

$$(\alpha D_i)(t) = \begin{cases} 0, & \text{if } r(t) < i \\ \dfrac{i!}{\gamma(t)}, & \text{if } r(t) = i \\ \displaystyle\sum_{u<t, r(u)=i} \frac{i!}{\gamma(u)} \alpha(t \backslash u), & \text{if } r(t) > i. \end{cases}$$

In the case of the first derivative operator, where $i = 1$, this simplifies even further to

$$(\alpha D)(t) = \begin{cases} 0, & \text{if } t = \emptyset, \\ 1, & \text{if } t = \tau, \\ \alpha(t_1) \cdots \alpha(t_m), & \text{if } t = [t_1, ..., t_m]. \end{cases}$$

Expansion of the numerical solution

Let $\xi(t)$ and $\eta(t)$ be elementary weight functions representing the internal stages and the input approximations respectively. We can now write

$$\xi(t) = A(\xi D)(t) + U\eta(t). \tag{2.9}$$

The output approximation can then be found from

$$B(\xi D)(t) + V\eta(t).$$

Assuming the method is of order p, this will correspond to $E\eta(t)$ within $O(h^{p+1})$. We can therefore write

$$E\eta(t) = B(\xi D)(t) + V\eta(t). \tag{2.10}$$

Assuming the first output solution is an approximation to $y(x_n)$, the method is said to be of order p if the first component of equation (2.10) is equal to $E(t)$ for all t such that $r(t) \leq p$.

The functions given in equations (2.9) and (2.10) are said to be the generating functions of the method.

2.4 Examples of general linear methods

As noted above, this class of methods is a large one. It includes the traditional methods such as Runge–Kutta methods and linear multistep methods, along with methods that have been

developed within the general linear methods framework, such as DIMSIMs and IRKS methods. Here we comment briefly on some of these methods.

2.4.1 Runge–Kutta methods

Runge–Kutta methods are very simple to rewrite as general linear methods. The A matrix of the general linear method is the same as the A matrix of the Runge–Kutta method. The B matrix is b^T, where b is the vector of weights of the Runge–Kutta method. Assuming the input vector is an approximation to $y(x_{n-1})$, the U matrix is e, a vector of 1's. The V matrix consists only of the number 1. This can be written as

$$M = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1s} & 1 \\ a_{12} & a_{22} & \cdots & a_{2s} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{s1} & a_{s2} & \cdots & a_{ss} & 1 \\ \hline b_1 & b_2 & \cdots & b_s & 1 \end{array}\right].$$

For example, we could rewrite the classical fourth order Runge–Kutta method with tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

as the general linear method

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ \hline \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 1 \end{array}\right].$$

2.4.2 Linear multistep methods

Linear multistep methods have a multi-value nature. The general form of the methods is

$$y_n = \sum_{i=1}^{k} \alpha_i y_{n-i} + h \sum_{i=0}^{k} \beta_i f(y_{n-i}).$$

If β_0 is equal to 0 the method is called explicit. This means the current approximation depends only on approximations to the solution and approximations to the derivative from the past. If $\beta_o \neq 0$ the method is called implicit because the current approximation depends on the derivative at the current time-step.

Adams methods

The most common linear multistep methods used for solving non-stiff differential equations are Adams methods. For these methods $\alpha_1 = 1$ and $\alpha_i = 0$ for $i > 1$. Therefore they take the form

$$y_n = y_{n-1} + h\sum_{i=0}^{k} \beta_i f(y_{n-i}).$$

Explicit methods of this type are called Adams–Bashforth methods. Implicit methods are known as Adams–Moulton methods.

If we were to write this as a general linear method, the input vector is

$$y^{[n-1]} = \begin{bmatrix} y(x_{n-1}) \\ hy'(x_{n-1}) \\ hy'(x_{n-2}) \\ \vdots \\ hy'(x_{n-k}) \end{bmatrix},$$

where $r = k + 1$. This means we can write the method as

$$\left[\begin{array}{c} Y_1 \\ \hline y_n \\ hf(Y_1) \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ \vdots \\ hf(y_{n-k+1}) \end{array}\right] = \left[\begin{array}{c|cccccc} \beta_0 & 1 & \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \\ \hline \beta_0 & 1 & \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{array}\right] \left[\begin{array}{c} hf(Y_1) \\ \hline y_{n-1} \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ hf(y_{n-3}) \\ \vdots \\ hf(y_{n-k}) \end{array}\right]$$

Although Adams–Moulton methods are implicit, they are only ever used to solve non-stiff problems, due to their small stability regions. They are usually used as part of a predictor-corrector pair. That is, an Adams–Bashforth method is used to predict an approximation and then the Adams–Moulton method is used to correct the approximation. They are used in either a (PEC) or $(PECE)$ scheme, where P stands for predict, E stands for evaluate and C stands for correct. In equation form this can be written as

$$y_n^* = y_{n-1} + h\sum_{i=1}^{k} \beta_i^* f(y_{n-i}),$$

$$y_n = y_{n-1} + h\beta_0 f(y_n^*) + h\sum_{i=1}^{k} \beta_i f(y_{n-i}).$$

A PEC method can be represented as the following general linear method (GLM)

$$\left[\begin{array}{c} Y_1 \\ \hline y_n \\ hf(y_n) \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ \vdots \\ hf(y_{n-k+1}) \end{array}\right] = \left[\begin{array}{c|cccccc} 0 & 1 & \beta_1^* & \beta_2^* & \cdots & \beta_{k-1}^* & \beta_k^* \\ \hline \beta_0 & 1 & \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{array}\right] \left[\begin{array}{c} hf(Y_1) \\ \hline y_{n-1} \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ hf(y_{n-3}) \\ \vdots \\ hf(y_{n-k}) \end{array}\right],$$

whereas a $PECE$ method can be represented as

$$\left[\begin{array}{c} Y_1 \\ Y_2 \\ \hline y_n \\ hf(y_n) \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ \vdots \\ hf(y_{n-k+1}) \end{array}\right] = \left[\begin{array}{cc|cccccc} 0 & 0 & 1 & \beta_1^* & \beta_2^* & \cdots & \beta_{k-1}^* & \beta_k^* \\ \beta_0 & 0 & 1 & \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \\ \hline \beta_0 & 0 & 1 & \beta_1 & \beta_2 & \cdots & \beta_{k-1} & \beta_k \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{array}\right] \left[\begin{array}{c} hf(Y_1) \\ hf(Y_2) \\ \hline y_{n-1} \\ hf(y_{n-1}) \\ hf(y_{n-2}) \\ hf(y_{n-3}) \\ \vdots \\ hf(y_{n-k}) \end{array}\right].$$

BDF methods

Backward differentiation (BDF) methods were the first numerical methods to be proposed for stiff problems. They were introduced in 1952 by Curtiss and Hirschfelder [27] to overcome the difficulties encountered in using Adams methods to solve stiff problems due to their lack of stability. Since Gear's 1971 book [35], they have been widely used to solve stiff problems.

For BDF methods all the β's are zero except β_0, meaning the approximated solution depends on only one derivative value, which is evaluated at the current step. The updated approximation is given by

$$y_n = \sum_{i=1}^{k} \alpha_i y_{n-i} + h\beta_0 f(y_n).$$

It is well-known that the BDF methods of order 7 and above are unstable (see, for example, [35]). Furthermore, only methods with $k = 1$ and $k = 2$ are A-stable. For orders higher than this the stability region becomes increasing inappropriate for solving stiff problems. The methods of orders 1 to 6 are given here.

$$
\begin{aligned}
k = 1: &\quad y_n = y_{n-1} + hf(y_n) \\
k = 2: &\quad y_n = \tfrac{4}{3}y_{n-1} - \tfrac{1}{3}y_{n-2} + \tfrac{2}{3}hf(y_n) \\
k = 3: &\quad y_n = \tfrac{18}{11}y_{n-1} - \tfrac{9}{11}y_{n-2} + \tfrac{2}{11}y_{n-3} + \tfrac{6}{11}hf(y_n) \\
k = 4: &\quad y_n = \tfrac{48}{25}y_{n-1} - \tfrac{36}{25}y_{n-2} + \tfrac{16}{25}y_{n-3} - \tfrac{3}{25}y_{n-4} + \tfrac{12}{25}hf(y_n) \\
k = 5: &\quad y_n = \tfrac{300}{137}y_{n-1} - \tfrac{300}{137}y_{n-2} + \tfrac{200}{137}y_{n-3} - \tfrac{75}{137}y_{n-4} + \tfrac{12}{137}y_{n-5} + \tfrac{60}{137} \\
k = 6: &\quad y_n = \tfrac{120}{49}y_{n-1} - \tfrac{150}{49}y_{n-2} + \tfrac{400}{147}y_{n-3} - \tfrac{75}{49}y_{n-4} + \tfrac{24}{49}y_{n-5} - \tfrac{10}{147} + \tfrac{20}{49}hf(y_n)
\end{aligned}
$$

In general linear form these can be represented as

$$
\left[\begin{array}{c} Y_1 \\ \hline y_n \\ y_{n-1} \\ y_{n-2} \\ y_{n-3} \\ \vdots \\ y_{n-k+1} \end{array}\right] = \left[\begin{array}{c|cccccc} \beta_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{k-1} & \alpha_k \\ \hline \beta_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{k-1} & \alpha_k \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{array}\right] \left[\begin{array}{c} hF(Y_1) \\ \hline y_{n-1} \\ y_{n-2} \\ y_{n-3} \\ y_{n-4} \\ \vdots \\ y_{n-k} \end{array}\right].
$$

2.4.3 DIMSIMs

Diagonally implicit multistage integration methods (DIMSIMs), are a special class of general linear methods which were first introduced by Butcher [7]. These methods were designed to be an extension to diagonally implicit Runge–Kutta methods, retaining the high order of the traditional methods, but increasing the stage order. To be a DIMSIM the method must have several desirable properties. These are:

- The matrix A should be lower triangular, with constant diagonals to lower the cost of solving the stage-value equations.
- The matrix V should be rank one to ensure zero stability.
- The quantities approximated by incoming and outgoing data should be related to the exact solution by a weighted Taylor series.

Type	A	Application	Architecture
1	$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ a_{21} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & a_{s3} & \cdots & 0 \end{bmatrix}$	Non-stiff	Sequential
2	$\begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ a_{21} & \lambda & 0 & \cdots & 0 \\ a_{31} & a_{32} & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & a_{s3} & \cdots & \lambda \end{bmatrix}$	Stiff	Sequential
3	$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$	Non-stiff	Parallel
4	$\begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}$	Stiff	Parallel

Table 2.6: Types of DIMSIMs

- The order of the stages should be close to, if not equal to, the overall order of the method.

There are four different types of DIMSIMs. The type of the method is determined by the structure of the A matrix, depending on whether the intended use of the method is for stiff or non-stiff problems and whether the intended architecture is sequential or parallel. The types of methods can be found in Table 2.6.

As has been mentioned, we require the incoming and outgoing values found in a step to be approximations to a weighted Taylor series. This means we require the incoming approximations

to be given by

$$y_i^{[n-1]} = \alpha_{i0} y(x_{n-1}) + \alpha_{i1} h y'(x_{n-1}) + \cdots + \alpha_{ip} h^p y^{(p)}(x_{n-1}) + O(h^{p+1}), \tag{2.11}$$

and the outgoing approximations by

$$y_i^{[n]} = \alpha_{i0} y(x_n) + \alpha_{i1} h y'(x_n) + \cdots + \alpha_{ip} h^p y^{(p)}(x_n) + O(h^{p+1}). \tag{2.12}$$

If equations (2.11) and (2.12) are true for some choice of the matrix

$$\begin{bmatrix} \alpha_{10} & \alpha_{11} & \cdots & \alpha_{1p} \\ \alpha_{20} & \alpha_{21} & \cdots & \alpha_{2p} \\ \vdots & \vdots & & \vdots \\ \alpha_{r0} & \alpha_{r1} & \cdots & \alpha_{rp} \end{bmatrix},$$

then this implies the method is of order at least p.

When the stage order is equal to the order of the method the order conditions greatly simplify, leaving only

$$\begin{aligned} \exp(cz) &= zA \exp(cz) + Uw(z) + O(h^{p+1}), \\ \exp(z) w(z) &= zB \exp(cz) + Vw(z) + O(h^{p+1}), \end{aligned}$$

where

$$w(z) = \begin{bmatrix} \alpha_{10} + \alpha_{11} z + \cdots + \alpha_{1p} z^p \\ \alpha_{20} + \alpha_{21} z + \cdots + \alpha_{2p} z^p \\ \vdots \\ \alpha_{r0} + \alpha_{r1} z + \cdots + \alpha_{rp} z^p \end{bmatrix}$$

and

$$\exp(cz) = \begin{bmatrix} \exp(c_1 z) \\ \exp(c_2 z) \\ \vdots \\ \exp(c_s) z \end{bmatrix}.$$

Most of the work on this class of methods has focused on methods with $p = q = r = s$ as the number of free parameters these methods have is the same as the number of equations required to ensure RK-stability.

If one assumes that $U = I$, the matrix B can be found in terms of A and V by

$$B = B_0 - AB_1 - VB_2 + VA,$$

where the (i,j)th element of the matrices B_0, B_1 and B_2 is given by

$$B_0 : \int_0^{1+c_i} l_j(t)\,dt,$$
$$B_1 : l_j(c_i+1),$$
$$B_2 : \int_0^{c_i} l_j(t)\,dt,$$

where $l_j(x)$ is the Lagrange interpolation basis polynomial given by

$$l_j(x) = \prod_{\substack{k=1 \\ k\neq j}}^{r} \frac{x-c_k}{c_j-c_k}.$$

Two simple examples are given here. Both of these methods have had their free parameters chosen to ensure RK-stability. The first is a method of type 1, with $c=[0,1]$:

$$M = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ \hline \frac{5}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & -\frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{array}\right].$$

The second method is of type 4:

$$M = \left[\begin{array}{cc|cc} \frac{3-\sqrt{3}}{2} & 0 & 1 & 0 \\ 0 & \frac{3-\sqrt{3}}{2} & 0 & 1 \\ \hline \frac{18-11\sqrt{3}}{4} & -\frac{12+7\sqrt{3}}{4} & \frac{3-2\sqrt{3}}{2} & -\frac{1+2\sqrt{3}}{2} \\ \frac{22-13\sqrt{3}}{4} & -\frac{12+9\sqrt{3}}{4} & \frac{3-2\sqrt{3}}{2} & -\frac{1+2\sqrt{3}}{2} \end{array}\right].$$

2.4.4 IRKS methods

Methods with inherent Runge–Kutta stability (IRKS) have been extensively studied by Butcher and Wright [22], [23], [67], [68]. These methods were introduced to concentrate on general linear methods with Runge–Kutta stability. RK-stability is a difficult condition to impose in the general case, but it is possible to find an inter-relation between the matrices which ensures the method has this property. While the conditions for IRKS are sufficient to ensure RK-stability, they are not necessary.

In the rest of this section we will write '$\equiv$' to denote the equivalence relation between two matrices that deems two matrices to be equivalent if and only if they are identical except for the first row.

Definition 2.16 *A general linear method satisfying $Ve_1 = e_1$ has inherent Runge–Kutta stability if*

$$BA \equiv XB, \tag{2.13}$$

$$BU \equiv XV + VX, \tag{2.14}$$

where X is some matrix and

$$\det(wI - V) = w^p(w - 1).$$

If the method is in Nordsieck form and the stage order is equal to the order of the method, the most general matrix X satisfying equations (2.13) and (2.14) is a doubly companion matrix of the form

$$\begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{p-1} & -\alpha_p & -\alpha_{p+1} - \beta_{p+1} \\ 1 & 0 & 0 & \cdots & 0 & 0 & -\beta_p \\ 0 & 1 & 0 & \cdots & 0 & 0 & -\beta_{p-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\beta_3 \\ 0 & 0 & 0 & \cdots & 1 & 0 & -\beta_2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -\beta_1 \end{bmatrix}.$$

A direct consequence of a method having IRKS is that the eigenvalues of the stability matrix will all be zero, except one, which will be equal to the truncated exponential series. This can be written as

$$\sigma(V + zB(I - zA)^{-1}U) = \{R(z), 0\},$$

where $R(z)$ is the stability function of a Runge–Kutta method and is equal to $\exp(z) + O(z^{p+1})$.

In general, these methods are formulated in Nordsieck form, with the stage order equal to the order and the number of values being passed from step to step equal to $p + 1$. Having the stage order equal to the order of the method greatly simplifies the order conditions. If we let

$$Z = [1,\ z,\ z^2,\ ...,\ z^p]^T,$$

where z is a complex variable, then the order conditions can be written as

$$\exp(cz) = zA\exp(cz) + UZ + O(z^{p+1}),$$

$$\exp(z)Z = zB\exp(cz) + VZ + O(z^{p+1}).$$

This makes the derivation of the methods relatively easy as U and V are completely defined by A, B and the abscissae vector c by

$$U = C - ACK,$$

$$V = E - BCK,$$

where C is the Vandermonde matrix

$$C = \left[e,\, c,\, \frac{c^2}{2!}, \cdots,\, \frac{c^p}{p!}\right],$$

and E is the Toeplitz matrix given by

$$\begin{bmatrix} 1 & \frac{1}{1!} & \frac{1}{2!} & \cdots & \frac{1}{(p-2)!} & \frac{1}{(p-1)!} & \frac{1}{p!} \\ 0 & 1 & \frac{1}{1!} & \cdots & \frac{1}{(p-3)!} & \frac{1}{(p-2)!} & \frac{1}{(p-1)!} \\ 0 & 0 & 1 & \cdots & \frac{1}{(p-4)!} & \frac{1}{(p-3)!} & \frac{1}{(p-2)!} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \frac{1}{1!} & \frac{1}{2!} \\ 0 & 0 & 0 & \cdots & 0 & 1 & \frac{1}{1!} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}.$$

Methods for both stiff and non-stiff problems are known to high order. Two simple examples are given here. The first is an explicit method of order 2, for which $c = [\frac{1}{3}, \frac{2}{3}, 1]$:

$$M = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\ \frac{1}{2} & 0 & 0 & 1 & \frac{1}{6} & \frac{1}{18} \\ 0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\ \hline 0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & -3 & 2 & 0 & -2 & 0 \end{array}\right].$$

The second method is diagonally implicit method of order 2 with $c = [\frac{1}{4}, \frac{1}{2}, 1]$:

$$M = \left[\begin{array}{ccc|ccc} \frac{1}{4} & 0 & 0 & 1 & 0 & -\frac{1}{32} \\ \frac{1}{6} & \frac{1}{4} & 0 & 1 & \frac{1}{12} & -\frac{1}{24} \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{12} & -\frac{1}{24} \\ \hline \frac{1}{6} & \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{12} & -\frac{1}{24} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -2 & 2 & 0 & 0 & 0 \end{array}\right].$$

It should be noted that DESIRE (Diagonally Extended Singly Implicit Runge–Kutta Effective order) [16] and ESIRK (Effective order Singly Implicit Runge–Kutta) methods [15] are special cases of IRKS methods.

CHAPTER 3

Almost Runge–Kutta methods

> Never be afraid to try something new. Remember amateurs built the Ark – professionals built the Titanic.
>
> ANON

Almost Runge–Kutta (ARK) methods are a special class of general linear methods. They were introduced by Butcher in 1997 [10]. The idea of these methods is to retain the multi-stage nature of Runge–Kutta methods, but allow more than one value to be passed from step to step. This gives the methods a multi-value character.

Of the three input and output values in ARK methods, one approximates the solution value and the other two approximate the scaled first and second derivatives respectively. To make it easy to start the methods, the second derivative is required to be accurate only to within $O(h^3)$, where h is the stepsize. The method has inbuilt "annihilation conditions" to ensure this low order does not adversely affect the solution value. These extra input values enable us to obtain stage order two. Traditional explicit Runge–Kutta methods are only able to obtain stage order one.[1] The advantage of this higher stage order is that we are able to interpolate or obtain an error estimate at little extra cost.

[1] A stage is of order q if $Y_i = y(x_0 + hc_i) + O(h^{q+1})$. A method is said to have stage order q if each of the stages is of order q.

The general form of ARK methods is

$$\left[\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_s \\ \hline y_1^{[n]} \\ y_2^{[n]} \\ y_3^{[n]} \end{array}\right] = \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] \left[\begin{array}{c} hF(Y_1) \\ hF(Y_2) \\ \vdots \\ hF(Y_s) \\ \hline y_1^{[n-1]} \\ y_2^{[n-1]} \\ y_3^{[n-1]} \end{array}\right],$$

where s is the number of internal stages. For an order p method the three output values are

$$\begin{aligned} y_1^{[n]} &= y(x_n) + O(h^{p+1}), \\ y_2^{[n]} &= hy'(x_n) + O(h^{p+1}), \\ y_3^{[n]} &= h^2 y''(x_n) + O(h^3). \end{aligned}$$

The coefficients of the method are chosen in a careful way to ensure the simple stability properties of Runge–Kutta methods are retained.

In this chapter we will concentrate on methods where A is strictly lower triangular, and hence the method is explicit, but most of the theory will carry over to implicit methods.

3.1 General form of explicit ARK methods

The general form of an explicit ARK method is

$$\left[\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_{s-1} \\ Y_s \\ \hline y_1^{[n]} \\ y_2^{[n]} \\ y_3^{[n]} \end{array}\right] = \left[\begin{array}{cccccc|ccc} 0 & 0 & 0 & \cdots & 0 & 0 & & & \\ a_{21} & 0 & 0 & \cdots & 0 & 0 & & & \\ a_{31} & a_{32} & 0 & \cdots & 0 & 0 & e & c - Ae & \frac{c^2}{2} - Ac \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & & \\ a_{s-1,1} & a_{s-1,2} & a_{s-1,3} & \cdots & 0 & 0 & & & \\ b_1 & b_2 & b_3 & \cdots & b_{s-1} & 0 & & & \\ \hline b_1 & b_2 & b_3 & \cdots & b_{s-1} & 0 & 1 & b_0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & \cdots & \beta_{s-1} & \beta_s & 0 & \beta_0 & 0 \end{array}\right] \left[\begin{array}{c} hF(Y_1) \\ hF(Y_2) \\ hF(Y_3) \\ \vdots \\ hF(Y_{s-1}) \\ hF(Y_s) \\ \hline y_1^{[n-1]} \\ y_2^{[n-1]} \\ y_3^{[n-1]} \end{array}\right].$$

As with a traditional Runge–Kutta method, b is a vector of length s representing the weights and c is a vector of length s representing the positions at which the function f is evaluated. The vector e is of length s, consisting entirely of ones.

The form of the U matrix is to ensure the stage order of the method is 2. To show this is true, we look at a Taylor series expansion of the internal stages. The internal stages of the method are given by:

$$Y_i = \sum_{j=1}^{i-1} a_{ij} h F(Y_i) + u_{i1} y_1^{[0]} + u_{i2} y_2^{[0]} + u_{i3} y_3^{[0]}. \tag{3.1}$$

To have stage order two we require $Y_i = y(x_0 + hc_i) + O(h^3)$. If we also make the substitutions $y_0 = y_1^{[0]}$, $hy_0' = y_2^{[0]}$ and $h^2 y_0'' + O(h^3) = y_3[0]$, we obtain

$$y(x_0 + hc_i) + O(h^3) = u_{i1} y_0 + u_{i2} h y_0' + u_{i3} h^2 y_0'' + h \sum_{j=1}^{i-1} a_{ij}\, y'(x_0 + hc_j) + O(h^3). \tag{3.2}$$

If we carry out a Taylor series expansion on both sides of equation (3.2) and equate the coefficients in y_0 we find:

$$u_{i1} y_0 = y_0, \quad \text{so that} \quad u_{i1} = 1.$$

Equating the coefficients in y_0' we find:

$$hc_i y_0' = u_{i2} h y_0' + h \sum_{j=1}^{i-1} a_{ij} y_0', \quad \text{so that} \quad u_{i2} = c_i - \sum_{j=1}^{i-1} a_{ij}.$$

Finally, equating the coefficients in y_0'' we find:

$$\frac{h^2 c_i^2}{2} y_0'' = u_{i3} h^2 y_0'' + h^2 \sum_{j=1}^{i-1} a_{ij} c_j y_0'', \quad \text{so that} \quad u_{i3} = \frac{c_i^2}{2} - \sum_{j=1}^{i-1} a_{ij} c_j.$$

We wish the final internal stage to give us the same quantity that is to be exported as the first outgoing approximation. This implies that the first row of the B matrix is the same as the last row of the A matrix, and the first row of the V matrix is the same as the last row of the U matrix. It is also implies that we always have $c_s = 1$.

We also wish the second outgoing approximation to be h times the derivative of the final stage. This implies the second row of the B and V matrices consists of zeros, with the exception of a 1 in the $(2, s)$ position of B.

The use of an ARK method is very similar to that of a Runge–Kutta method. The main difference is that we are now passing three pieces of information between steps. The first two starting values are $y(x_0)$ and $hf(y(x_0))$ respectively. The third starting value is obtained by

taking a single Euler step forward and taking the difference between the derivatives at these two points. The starting vector is therefore

$$\Big[y(x_0), hf(y(x_0)), hf\Big(y(x_0) + hf(y(x_0))\Big) - hf(y(x_0))\Big].$$

This choice of starting method was chosen for its simplicity, but it is adequate, at least for low order methods. The method for computing the three starting approximations can be written in the form of the generalized Runge–Kutta tableau

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{array}, \tag{3.3}$$

where the zero in the first column of the last two rows indicates the fact that the term y_{n-1} is absent from the output approximation. This can be interpreted in the same way as a Runge–Kutta method, but with three output approximations.

Changing the stepsize poses no problem as we can simply scale the vector in the same way we would scale a Nordsieck vector. If we set $r = h_j/h_{j-1}$ then the y vector needs to be scaled by $[1, r, r^2]$.

3.2 Order and related conditions

The order conditions for the first output approximation can be written down using the standard rooted-tree approach that is used for Runge–Kutta methods. The additional structure of ARK methods means that fewer order conditions are required than for traditional Runge–Kutta methods. This is because having a stage order of 2 makes some of the order conditions redundant. The trees that can be omitted are those that would be omitted for a Runge–Kutta method if the $C(2)$[2] condition is assumed; i.e. trees that contain a vertex from which only a single outgoing arc is joined to another vertex, which in turn is joined to a terminal vertex.

For the higher order methods it is also convenient to assume the $D(1)$ condition, that is

$$\sum_{i=1}^{s} b_i a_{ij} = b_j(1 - c_j), \qquad j = 1, \ldots, s. \tag{3.4}$$

[2]The $C(2)$ condition assumes

$$\sum_{j=1}^{s} a_{ij} c_j = \frac{c_i^2}{2},$$

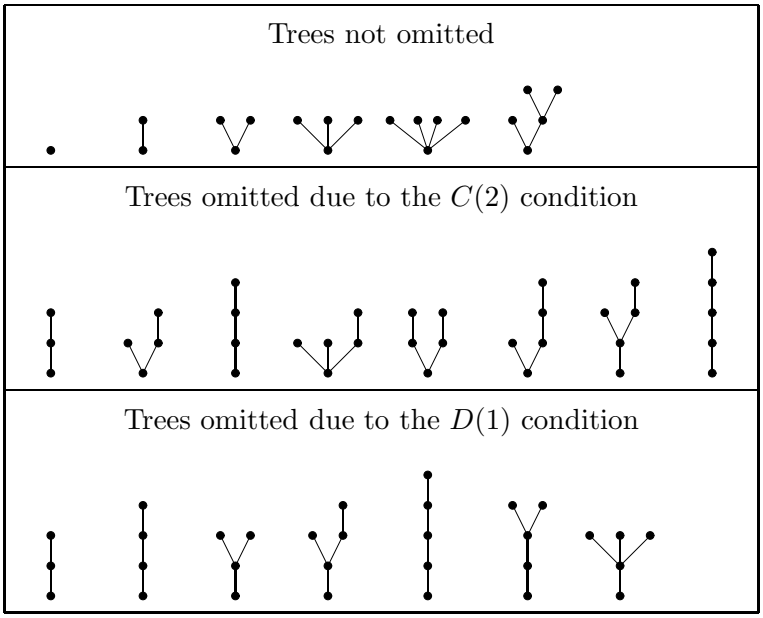


Table 3.1: Trees up to order 5 omitted due to the simplifying assumptions.

This enables us to also omit the trees that have only a single arc branching from the root. As can be seen in Table 3.1 these simplifying assumptions greatly decrease the number of order conditions that need to be considered.

Unfortunately, due to the fact that the third input approximation is accurate only to order 2, some of the conditions that we have just omitted are now restored. This is so that the errors in the third approximation do not combine to give low order error terms in the first or second output approximations. The conditions that ensure the errors in the third input approximation have no major effect on our first output approximation are called "annihilation conditions".

An alternative way of looking at the order conditions is to consider the generating functions given in Section 2.3.

We will use a slightly different notation than in the general case. Let $\xi(t)$, $\alpha(t)$ and $\eta(t)$ be elementary weight functions associated with the internal stages, the first output approximation and the third output approximation respectively. Using the special form of ARK method, (2.9) can be written as

$$\xi(t) = \mathbf{1} + (c - Ae)D(t) + (\tfrac{1}{2}c^2 - Ac)\eta(t) + A(\xi D)(t), \tag{3.5}$$

Similarly, the first and and third components of (2.10) can be written respetively as

$$\alpha(t) = \mathbf{1} + b_0 D(t) + b^T(\xi D)(t), \tag{3.6}$$

$$(E\eta)(t) = \beta_0 D(t) + \beta^T(\xi D)(t), \tag{3.7}$$

where **1** denotes the unit elementary weight function of ones, which maps $y(x)$ to $y(x)$. As the second output approximation is the derivative of the first output approximation, this does not need to be considered seperately.

The order conditions are found by setting $\alpha(t) = \frac{1}{\gamma(t)}$ for all trees of order up to and including p. Due to the stage order we notice that many of these conditions turn out to be equivalent, leaving the same number of conditions as the alternative approach.

The annihilation conditions are needed to ensure the low order of the third input approximation does not have an adverse affect on the first and second output approximations. It is to be used mainly to increase the stage order to two. The annihilation conditions are found by setting to zero the coefficients of any terms in α involving η, for trees of order $\leq p$. For example, α of the tree t_7 is given by

$$\alpha(t_7) = b^T(\tfrac{1}{2}c^2 - Ac)\eta(t_3) + b^T Ac^2.$$

For a method of order four or above, an annihilation condition is

$$\begin{aligned} b^T(\tfrac{1}{2}c^2 - Ac) &= 0 \\ \text{or } b^T Ac &= \frac{1}{6}. \end{aligned}$$

This ensures the third input approximation does not affect the low order terms in the first output approximation.

To ensure the third output value approximates $h^2y''(x_{n+1})$ to within $O(h^3)$ it is necessary to require that

$$\beta^T e + \beta_0 = 0, \tag{3.8}$$

$$\beta^T c = 1. \tag{3.9}$$

This can be verified by carrying out a Taylor series expansion of the third output approximation. The third output approximation is given by

$$y_3^{[1]} = \beta_0 y_2^{[0]} + \sum_{i=1}^{s} \beta_i h F(Y_i).$$

To be of order two, we require $y_3^{[1]} = h^2y''(x_0 + h) + O(h^3)$. If we also make the substitutions $y_0 = y_1^{[0]}$, $hy_0' = y_2^{[0]}$ and $F(Y_i) = y'(x_0 + hc_i) + O(h^3)$, we obtain

$$h^2y''(x_0 + h) + O(h^3) = \beta_0 hy'(x_0) + \sum_{i=1}^{s} \beta_i hy'(x_0 + hc_i) + O(h^3).$$

If we carry out a Taylor series expansion on both sides of this equation we find

$$h^2 y''(x_0) + O(h^3) = \beta_0 h y'(x_0) + h \sum_{i=1}^{s} \beta_i (y'(x_0) + h c_i y''(x_0)) + O(h^3).$$

Equating the coefficients of $y'(x_0)$ gives

$$0 = \beta_0 h + h \sum_{i=1}^{s} \beta_i \quad \text{implying that} \quad \beta^T e + \beta_0 = 0.$$

Equating the coefficients of $y''(x_0)$ gives

$$h^2 = h^2 \sum_{i=1}^{s} \beta_i c_i \quad \text{implying that} \quad \beta^T c = 1.$$

The last constraint that is placed on the coefficients is that the method has RK-stability. This will be discussed in detail in later sections.

For ease of analysis, the above conditions are sorted into two classes, α conditions and β conditions. The α conditions are order conditions that are found from $\alpha(t) = 1/\gamma(t)$, subject to the condition that the stage order is 2, along with the annihilation conditions. They have the same form as corresponding order conditions for Runge–Kutta methods, except that some of the conditions are omitted. They contain entries that occur in matrix A, and the vectors b and c. The β conditions are the remaining conditions, that is $\beta^T e + \beta_0 = 0$ and $\beta^T c = 1$ and the conditions required for RK-stability. They include one or more occurance of β_s.

A list of conditions required for $s = p$ and $s = p + 1$, for methods of orders 3 and 4 are outlined in subsequent sections.

3.3 Interpolation

One of the major advantages of ARK methods is the possibility of a cheap interpolator due to the stage order. Unfortunately it is not possible to obtain an interpolator of the same order as the method but it is possible to obtain an interpolator one order lower than the method. This should be satisfactory for most practical applications.

To interpolate at point $x_n + \xi h$, in a step from x_n to $x_n + h$ we need to find a vector $\tilde{b}(\xi)$ such that some modified order conditions are satisfied. That is, we want to choose polynomial coefficients of degree $p - 1$ so that

$$y(x_n + \xi h) = y_{n-1} + h \sum_{i=1}^{s} \tilde{b}_i f(Y_i) \tag{3.10}$$

is exact when $y(x)$ is a polynomial of degree $p-1$. These conditions are dependent on the order of the method, but are roughly equivalent to taking the standard order conditions for a method one order less and multiplying the right hand side by ξ^r, where r is the order of the tree. Once $\tilde{b}$ has been found, an approximation to the solution at the point $x_n + \xi h$ can be obtained from (3.10).

For consistency any free parameters that remain need to be chosen in such a way that $\tilde{b}^T = b^T$ when $\xi = 1$. We will also try to ensure that the bushy tree of the same order as the method is satisfied.

Further details will be given in each of the individual cases.

3.4 Methods with $s = p$

In this section we look at methods which have the same number of stages as the order of the method. Methods with this property are considered as we wish to minimise computation costs, and it is not possible to satisfy all the order conditions for $s < p$. We will concentrate on third and fourth order methods.

3.4.1 RK stability

As stated in section 2.1, the stability matrix of a general linear method is given by

$$M(z) = V + zB(I - zA)^{-1}U.$$

A method is said to have RK stability if all the eigenvalues of the matrix M are zero, except one which is equal to R(z), the stability region of a Runge Kutta method. For an explicit method $R(z)$ is given by:

$$R(z) = \exp_s(z) = \sum_{i=0}^{s} \frac{z^i}{i!}.$$

As the trace of a matrix is equal to the sum of the eigenvalues, for a method to have RK stability we require

$$\mathrm{Tr}(V + zB(I - zA)^{-1}U) = 1 + z + \frac{z^2}{2} + \cdots + \frac{z^s}{s!}.$$

If we carry out a Taylor series expansion on the left-hand side of this equation and equate the coefficients, this implies

$$\mathrm{Tr}(BA^{i-1}U) = \frac{1}{i!}, \qquad i = 1, \ldots, s. \tag{3.11}$$

Theorem 3.1 *An ARK method of order p with p stages has RK-stability if and only if*

$$\beta^T(I + \beta_s A) = \beta_s e_s^T, \tag{3.12}$$

$$\left(1 + \tfrac{1}{2}\beta_s c_1\right) b^T A^{s-2} c = \frac{1}{s!}, \tag{3.13}$$

$$c_1 = -\frac{2\exp_s(-\beta_s)}{\beta_s \exp_{s-1}(-\beta_s)}, \tag{3.14}$$

where $e_s^T = [0, 0, \cdots, 0, 1]$ *and has s components and*

$$\exp_n(x) = 1 + \frac{x^2!}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}.$$

Proof: *(only if)* From equation (3.11), with $i = 1$ we find:

$$\begin{aligned} \mathrm{Tr}(BU) &= b^T e + e_s^T(c - Ae) + \beta^T\left(\tfrac{1}{2}c^2 - Ac\right) \\ 1 &= b^T e + 1 - b^T e + \beta^T\left(\tfrac{1}{2}c^2 - Ac\right) \\ \implies & \beta^T\left(\tfrac{1}{2}c^2 - Ac\right) = 0. \end{aligned} \tag{3.15}$$

From the generating functions, it can be shown that $b^T A^{i-2} c = 1/i!$, for $1 < i < s$, are order conditions. Using this information, for $1 < i < s$, equation (3.11) can be written as

$$\begin{aligned} \mathrm{Tr}(BA^{i-1}U) &= b^T A^{i-1} e + e_s^T A^{i-1}(c - Ae) + \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ \frac{1}{i!} &= b^T A^{i-1} e + b^T A^{i-2} c - b^T A^{i-1} e + \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ &= b^T A^{i-2} c + \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ &= \frac{1}{i!} + \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right). \\ \implies & \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) = 0, \qquad i = 2, \ldots, s-1. \end{aligned} \tag{3.16}$$

When $i = s$ we can no longer assume $b^T A^{i-2} c = 1/i!$, however since A is strictly lower triangular we know that $A^s = 0$. Due to the form of A, we also find $\beta^T A^{s-1} c^2 = \beta_s c_1 b^T A^{s-2} c$. Equation (3.11) now gives

$$\begin{aligned} \mathrm{Tr}(BA^{s-1}U) &= b^T A^{s-1} e + e_s^T A^{s-1}(c - Ae) + \beta^T A^{s-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ \frac{1}{s!} &= b^T A^{s-2} c + \beta^T A^{s-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ &= b^T A^{s-2} c + \tfrac{1}{2}\beta^T A^{s-1} c^2 \end{aligned}$$

$$\implies \left(1+\tfrac{1}{2}\beta_s c_1\right) b^T A^{s-2} c = \frac{1}{s!}. \tag{3.17}$$

Note that this is the same as equation (3.13).

Let

$$v^T = \beta_s e_s^T - \beta^T(I+\theta A), \tag{3.18}$$

where θ is chosen so that $v_{s-1} = 0$. Using equations (3.15) and (3.16) and the fact that $\frac{1}{2}b^T A^{i-2}c^2 = b^T A^{i-1}c, \quad i = 2, ..., s-2$ we find

$$\begin{aligned} v^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) &= (\beta_s e_s^T - \beta^T(I+\theta A))A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) \\ &= \beta_s b^T A^{i-2}(\tfrac{1}{2}c^2 - Ac) - \beta^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) - \theta\beta^T A^i\left(\tfrac{1}{2}c^2 - Ac\right) \\ &= \tfrac{1}{2}\beta_s b^T A^{i-2}c^2 - \beta_s b^T A^{i-1}c = 0, \end{aligned}$$

and it follows that

$$v^T A^{i-1}\left(\tfrac{1}{2}c^2 - Ac\right) = 0, \qquad i = 1, \ldots, s-2. \tag{3.19}$$

Since $A^{i-1}\left(\frac{1}{2}c^2 - Ac\right) \neq 0$, this implies that $v^T = 0$. We then know that equation (3.19) also holds for $i = s-1$. This gives, in turn

$$\begin{aligned} v^T A^{s-2}\left(\tfrac{1}{2}c^2 - Ac\right) &= 0, \\ (\beta_s e_s^T - \beta^T(I+\theta A))A^{s-2}\left(\tfrac{1}{2}c^2 - Ac\right) &= 0, \\ \tfrac{1}{2}\beta_s b^T A^{s-3}c^2 - \beta_s b^T A^{s-2}c - \beta^T A^{s-2}\left(\tfrac{1}{2}c^2 - Ac\right) - \beta^T\theta A^{s-1}\left(\tfrac{1}{2}c^2 - Ac\right) &= 0. \end{aligned}$$

This can be greatly simplified using $b^T A^{s-3}c^2 = \frac{2}{s!}$, $\beta^T A^{s-2}(\frac{1}{2}c^2 - Ac) = 0$ and $A^s = 0$. This gives, in turn

$$\begin{aligned} \tfrac{1}{2}\beta_s\tfrac{2}{s!} - \beta_s b^T A^{s-2}c - \tfrac{1}{2}\beta^T\theta A^{s-1}c^2 &= 0, \\ \beta_s b^T A^{s-2}c + \tfrac{1}{2}\beta^T\theta A^{s-1}c^2 &= \frac{\beta_s}{s!}, \\ \left(\beta_s + \tfrac{1}{2}\theta\beta_s c_1\right) b^T A^{s-2}c &= \frac{\beta_s}{s!}. \end{aligned}$$

If we compare this equation with equation (3.13) we can see that $\theta = \beta_s$. Substituting this into equation (3.18) we obtain equation (3.12).

Substituting $\beta^T = \beta_s e_s^T(I+\beta_s A)^{-1}$ into the condition $\beta^T c = 1$ gives

$$\begin{aligned} 1 &= \beta_s\left(1 + \sum_{i=1}^{s-1}(-\beta_s)^i e_s^T A^i c\right) \\ &= \beta_s\left(1 + \sum_{i=1}^{s-1}(-\beta_s)^i b^T A^{i-1} c\right). \end{aligned}$$

Using equation (3.13) and the order conditions this gives, in turn

$$\begin{aligned}
1 &= \beta_s \left(1 + \frac{(-\beta_s)^{s-1}}{s!\left(1+\frac{1}{2}c_1\beta_s\right)} + \sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!}\right),\\
\beta_s - 1 + \beta_s\sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!} &= \frac{-\beta_s(-\beta_s)^{s-1}}{s!\left(1+\frac{1}{2}c_1\beta_s\right)},\\
s! + \frac{1}{2}s!c_1\beta_s &= \frac{(-\beta_s)^s}{\beta_s - 1 + \beta_s\sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!}},\\
\frac{1}{2}s!c_1\beta_s &= \frac{(-\beta_s)^s - s!\left(\beta_s - 1 + \beta_s\sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!}\right)}{\beta_s - 1 + \beta_s\sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!}},\\
\frac{1}{2}c_1\beta_s &= \frac{\frac{(-\beta_s)^s}{s!} + \left(1 - \beta_s + \sum_{i=1}^{s-2}\frac{(-\beta_s)^{i+1}}{(i+1)!}\right)}{-\left(1 - \beta_s + \sum_{i=1}^{s-2}\frac{(-\beta_s)^{i+1}}{(i+1)!}\right)},\\
&= \frac{\exp_s(-\beta_s)}{-\exp_{s-1}(-\beta_s)},\\
c_1 &= -\frac{2\exp_s(-\beta_s)}{\beta_s\exp_{s-1}(-\beta_s)}.
\end{aligned}$$

(if) First we need to show that the third output approximation is of order 2. i.e. that $\beta^T c = 1$. To do this, rearrange (3.12) and substitute into this equation, to give

$$1 - \beta^T c = 1 - \beta_s(I + \beta_s A)^{-1}c, \tag{3.20}$$

$$= 1 - \beta_s\left(\sum_{i=0}^{s}(-\beta_s)^i e_s^T A^i c\right), \tag{3.21}$$

$$= 1 - \beta_s\left(1 + \sum_{i=1}^{s}(-\beta_s)^i b^T A^{i-1}c\right). \tag{3.22}$$

Due to the form of the b^T vector and the A matrix, $b^T A^{s-1}c = 0$. Using this information, and (3.13), the above can be written as

$$1 - \beta^T c = 1 - \beta_s\left(1 + \sum_{i=1}^{s-2}\frac{(-\beta_s)^i}{(i+1)!} + \frac{(-\beta_s)^{s-1}}{s!(1+\frac{1}{2}\beta_s c_1)}\right), \tag{3.23}$$

$$= \exp_{s-1}(-\beta_s) + \frac{(-\beta_s)^s}{s!(1+\frac{1}{2}\beta_s c_1)}. \tag{3.24}$$

Using (3.14), this can be written as

$$1 - \beta^T c = \exp_{s-1}(-\beta_s) + \frac{(-\beta_s)^s}{s!}\left(1 + \frac{1}{2}\beta_s\left(\frac{-2\exp_s(-\beta_s)}{\beta_s \exp_{s-1}(-\beta_s)}\right)\right)^{-1}, \tag{3.25}$$

$$= \exp_{s-1}(-\beta_s) + \frac{(-\beta_s)^s}{s!}\left(1 - \frac{\exp_s(-\beta_s)}{\exp_{s-1}(-\beta_s)}\right)^{-1}, \tag{3.26}$$

$$= \exp_{s-1}(-\beta_s) + (\exp_s(-\beta_s) - \exp_{s-1}(-\beta_s))\left(1 - \frac{\exp_s(-\beta_s)}{\exp_{s-1}(-\beta_s)}\right)^{-1}, \tag{3.27}$$

$$= 0. \tag{3.28}$$

Next we need to show that the matrix has one non-zero eigenvalue, which is $R(z)$. From (3.13), we have $\mathrm{tr}(BA^{k-1}U) = 1/k!$ for $k = s$. From (3.12) this holds in turn for $k = s-1, s-2, \ldots, 1$. This implies that the trace of the stability matrix is equal to $R(z)$. To show that two of the eigenvalues are zero we will write the stability matrix in the form

$$M(z) = M_0 + zM_1 + z^2M_2 + \cdots + z^sM_s.$$

A similar matrix $N(z) = (I - ze_2e_1^T)M(z)(I + ze_2e_1^T)$ is defined, and similarly expanded, to give

$$N(z) = N_0 + zN_1 + z^2N_2 + \cdots + z^sN_s + z^{s+1}N_{s+1}.$$

It can be shown that

$$N_0 = \begin{bmatrix} 1 & b_0 & 0 \\ 0 & 0 & 0 \\ 0 & \beta_0 & 0 \end{bmatrix},$$

$$N_k = \begin{bmatrix} b^TA^{k-2}c & b^TA^{k-1}(c - Ae) & b^TA^{k-1}(\frac{1}{2}c^2 - Ac) \\ 0 & 0 & 0 \\ \beta^TA^{k-2}c & \beta^TA^{k-1}(c - Ae) & \beta^TA^{k-1}(\frac{1}{2}c^2 - Ac) \end{bmatrix}, \quad k = 1, 2, \ldots, s+1.$$

This means that the second row of $N(z)$ is zero. The second row and column, therefore, can be deleted without altering the set of non-zero eigenvalues. Denote this modified matrix as $\widetilde{N}(z)$, which can be written as

$$\widetilde{N}(z) = \widetilde{N}_0 + z\widetilde{N}_1 + z^2\widetilde{N}_2 + \cdots + z^s\widetilde{N}_s + z^{s+1}\widetilde{N}_{s+1},$$

so that

$$\widetilde{N}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\widetilde{N}_1 = \begin{bmatrix} 1 & 0 \\ 0 & \beta^T(\frac{1}{2}c^2 - Ac) \end{bmatrix}.$$

$$\widetilde{N}_k = \begin{bmatrix} b^T A^{k-2} c & b^T A^{k-1}(\frac{1}{2}c^2 - Ac) \\ \beta^T A^{k-2} c & \beta^T A^{k-1}(\frac{1}{2}c^2 - Ac) \end{bmatrix}, \quad k = 2, 3, \ldots, s+1.$$

Using (3.12), $\widetilde{N}(z)$ can now be written in the form

$$\widetilde{N}(z) = \begin{bmatrix} (1 + \frac{z}{\beta_s})\beta^T (I - zA)^{-1} c & (1 + \frac{z}{\beta_s})\beta^T (I - zA)^{-1}(\frac{1}{2}c^2 - Ac) \\ z^2 \beta^T (I - zA)^{-1} c & z^2 \beta^T (I - zA)^{-1}(\frac{1}{2}c^2 - Ac) \end{bmatrix}.$$

As the second row is a scalar multiple of the first row, this has zero determinant. ■

3.4.2 Third order methods with three stages

A third order method with three stages takes the form:

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & c_1 & \frac{1}{2}c_1^2 \\ a_{21} & 0 & 0 & 1 & c_2 - a_{21} & \frac{1}{2}c_2^2 - a_{21}c_1 \\ b_1 & b_2 & 0 & 1 & b_0 & 0 \\ \hline b_1 & b_2 & 0 & 1 & b_0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & 0 & \beta_0 & 0 \end{array}\right].$$

Order conditions

The order conditions for a third order method are:

$$b_0 + b^T e = 1, \tag{3.29}$$

$$b^T c = \frac{1}{2}, \tag{3.30}$$

$$b^T c^2 = \frac{1}{3}. \tag{3.31}$$

There are no annihilation conditions for this low order. This is because $\alpha(t_i)$, does not have any terms involving $\eta(t)$, for trees of order 3 or less.

As we saw in Theorem 3.1, there are several equations that need to be satisfied for the method to have RK-stability. For third order, three stage methods the equations are:

$$\beta^T (I + \beta_3 A) = \beta_3 e_3^T, \tag{3.32}$$

$$\left(1 + \tfrac{1}{2}\beta_3 c_1\right) b^T A c = \frac{1}{6}, \tag{3.33}$$

$$c_1 = -\frac{2(1 - \beta_3 + \frac{1}{2}\beta_3^2 - \frac{1}{6}\beta_3^3)}{\beta_3(1 - \beta_3 + \frac{1}{2}\beta_3^2)}. \tag{3.34}$$

For the third output approximation to be of order 2, we also require the method to satisfy

$$\beta^T e + \beta_0 = 0, \tag{3.35}$$

$$\beta^T c = 1. \tag{3.36}$$

Condition (3.36) is actually satisfied by ensuring (3.34), but is mentioned here for completeness. The reason for this can be seen in the proof of Theorem 3.1. This condition will be omitted from further discussions about order, unless needed explicitly for the derivation of the methods.

Derivation of methods

The derivation of methods with $s = p$ is very similar to traditional Runge–Kutta methods. The free parameters are β_s and the nodes c_2, c_3, $\cdots$, c_{s-1}. The value of c_1 can be determined from the conditions for Runge–Kutta stability. The b^T vector can be found from the quadrature conditions, and then the entries in the A matrix can be determined from the rest of the conditions. Finally, the entries of β can be found from one of the conditions for Runge–Kutta stability. Here we outline the case for $s = 3$.

For third order methods we have two free parameters. It is easiest if we take these to be β_3 and c_2.

Once β_3 has been chosen we can find c_1 from

$$c_1 = -\frac{2\left(1 - \beta_3 + \frac{1}{2}\beta_3^2 - \frac{1}{6}\beta_3^3\right)}{\beta_3\left(1 - \beta_3 + \frac{1}{2}\beta_3^2\right)}. \tag{3.37}$$

From equations (3.30) and (3.31) we find

$$b_1 c_1 + b_2 c_2 = \frac{1}{2}, \tag{3.38}$$

$$b_1 c_1^2 + b_2 c_2^2 = \frac{1}{3}. \tag{3.39}$$

Rearranging equation (3.38) to make $b_1 c_1$ the subject, and substituting into equation (3.39) we find

$$\begin{aligned} \frac{1}{2}c_1 - b_2 c_1 c_2 + b_2 c_2^2 &= \frac{1}{3} \\ \implies b_2(c_2^2 - c_1 c_2) &= \frac{1}{3} - \frac{1}{2}c_1 \\ \implies b_2 &= \frac{2 - 3c_1}{6c_2(c_2 - c_1)}. \end{aligned} \tag{3.40}$$

Substituting equation (3.40) back into equation (3.38) we find

$$\begin{aligned} b_1 c_1 &= \frac{1}{2} - \frac{2 - 3c_1}{6(c_2 - c_1)} \\ &= \frac{3c_2 - 2}{6(c_2 - c_1)} \\ \implies b_1 &= \frac{3c_2 - 2}{6c_1(c_2 - c_1)}. \end{aligned} \tag{3.41}$$

It is possible to find methods such that $c_1 = c_2$, however b^T needs to be calculated slightly differently.

Equation (3.33) gives

$$\begin{aligned} \left(1 + \tfrac{1}{2}\beta_3 c_1\right) b_2 a_{21} c_1 &= \frac{1}{6} \\ \implies a_{21} &= \frac{1}{3b_2 c_1 (2 + \beta_3 c_1)}. \end{aligned} \tag{3.42}$$

From equation (3.29) we find

$$b_0 = 1 - b_1 - b_2. \tag{3.43}$$

We can then find the β vector from equation (3.32). Evaluating both sides of this equation we find

$$(\beta_1 + \beta_2\beta_3 a_{21} + \beta_3^2 b_1, \beta_2 + \beta_3^2 b_2, \beta_3) = (0, 0, \beta_3).$$

This implies

$$\beta_1 + \beta_2\beta_3 a_{21} + \beta_3^2 b_1 = 0 \tag{3.44}$$

and

$$\beta_2 + \beta_3^2 b_2 = 0. \tag{3.45}$$

Equation (3.45) gives

$$\beta_2 = -b_2\beta_3^2. \tag{3.46}$$

From equation (3.44) we find

$$\begin{aligned} \beta_1 &= -\beta_2\beta_3 a_{21} - \beta_3^2 b_1 \\ &= b_2\beta_3^3 a_{21} - \beta_3^2 b_1. \end{aligned} \tag{3.47}$$

Finally, we can calculate β_0 from equation (3.35) which gives

$$\beta_0 = -\beta_1 - \beta_2 - \beta_3. \tag{3.48}$$

In summary, once β_3 and c_2 have been chosen and c_1 has been calculated from equation (3.37) the remaining coefficients of a third order method can be found from

$$
\begin{array}{rclcrcl}
b_1 &=& \dfrac{3c_2-2}{6c_1(c_2-c_1)}, & & \beta_2 &=& -b_2\beta_3^2, \\
b_2 &=& \dfrac{2-3c_1}{6c_2(c_2-c_1)}, & & \beta_1 &=& b_2\beta_3^3 a_{21} - \beta_3^2 b_1, \\
b_0 &=& 1-b_1-b_2, & & \beta_0 &=& -\beta_1-\beta_2-\beta_3, \\
a_{21} &=& \dfrac{1}{3b_2c_1(c_1\beta_3+2)}. & & & &
\end{array}
$$

Some example methods

One particularly nice value for β_3 is 2, since this gives $c_1 = \frac{1}{3}$. This choice of β together with a convenient choice for c_2 gives some especially simple methods. Two of these are given below. In the first method $c^T = [\frac{1}{3}, 1, 1]$ and in the second $c^T = [\frac{1}{3}, \frac{2}{3}, 1]$.

$$
\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{3}{2} & 0 & 0 & 1 & -\frac{1}{2} & 0 \\
\frac{3}{4} & \frac{1}{4} & 0 & 1 & 0 & 0 \\
\hline
\frac{3}{4} & \frac{1}{4} & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & -1 & 2 & 0 & -1 & 0
\end{array}\right], \qquad
\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{1}{2} & 0 & 0 & 1 & \frac{1}{6} & \frac{1}{18} \\
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
\hline
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
3 & -3 & 2 & 0 & -2 & 0
\end{array}\right],
$$

Two more examples are given below. For the first $c^T = [\frac{17}{30}, \frac{2}{3}, 1]$ and for the second $c^T = [\frac{8}{15}, \frac{2}{3}, 1]$.

$$
\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{17}{30} & \frac{289}{1800} \\
\frac{25}{136} & 0 & 0 & 1 & \frac{197}{408} & \frac{17}{144} \\
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
\hline
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\frac{150}{17} & -12 & 4 & 0 & -\frac{14}{17} & 0
\end{array}\right], \qquad
\left[\begin{array}{ccc|ccc}
0 & 0 & 0 & 1 & \frac{8}{15} & \frac{32}{225} \\
\frac{25}{108} & 0 & 0 & 1 & \frac{47}{108} & \frac{8}{81} \\
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
\hline
0 & \frac{3}{4} & 0 & 1 & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
\frac{75}{16} & -\frac{27}{4} & 3 & 0 & -\frac{15}{16} & 0
\end{array}\right],
$$

Interpolation

To find a second order interpolator for a third order method we require $\tilde{b}^T$ to satisfy the following requirements

$$\tilde{b}_0 + \tilde{b}^T e = \xi, \tag{3.49}$$

$$\tilde{b}^T c = \frac{\xi^2}{2}, \tag{3.50}$$

$$\tilde{b}^T c^2 = \frac{\xi^3}{3}, \tag{3.51}$$

where we are trying to find an approximation at $x_{n-1} + \xi h$.

We also require $\tilde{b}_1(1) = b_1$, $\tilde{b}_2(1) = b_2$ and $\tilde{b}_3(1) = 0$, to make the interpolator consistent. If possible, we would also like $\tilde{b}_1'(1) = 0$ and $\tilde{b}_2'(1) = 0$.

Example: We will try to find an interpolator for the following method, where $c^T = [\frac{8}{15}, 1, 1]$,

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 1 & \frac{8}{15} & \frac{32}{225} \\ \frac{175}{144} & 0 & 0 & 1 & -\frac{31}{144} & -\frac{4}{27} \\ \frac{75}{112} & \frac{1}{7} & 0 & 1 & \frac{3}{16} & 0 \\ \hline \frac{75}{112} & \frac{1}{7} & 0 & 1 & \frac{3}{16} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{75}{56} & -\frac{9}{7} & 3 & 0 & -\frac{3}{8} & 0 \end{array}\right]$$

From equations (3.49), (3.50) and (3.51) we find

$$\tilde{b}_0 + \tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = \xi, \tag{3.52}$$

$$\frac{8}{15}\tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = \frac{1}{2}\xi^2, \tag{3.53}$$

$$\frac{64}{225}\tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = \frac{1}{3}\xi^3. \tag{3.54}$$

Subtracting equation (3.54) from equation (3.53) gives

$$\frac{56}{225}\tilde{b}_1 = \frac{\xi^2}{6}(3 - 2\xi), \quad \text{so that} \quad \tilde{b}_1 = \frac{75}{112}\xi^2(3 - 2\xi).$$

If we substitute this back into equation (3.53) we obtain

$$\begin{aligned} \tilde{b}_2 + \tilde{b}_3 &= \frac{1}{2}\xi^2 - \frac{5}{14}\xi^2(3 - 2\xi) \\ &= \frac{\xi^2(5\xi - 4)}{7}. \end{aligned}$$

From this we will choose $\tilde{b}_2$ and $\tilde{b}_3$ to be

$$\begin{aligned} \tilde{b}_2 &= \frac{\xi^2((5 - \nu)\xi + \mu - 4)}{7}, \\ \tilde{b}_3 &= \frac{\xi^2(\nu\xi - \mu)}{7}, \end{aligned}$$

for some μ and ν.

We want $\tilde{b}_3$ to vanish at $\xi = 1$. This implies $\mu = \nu$. We also wish to have $\tilde{b}_2'(1) = 0$.

$$\begin{aligned} \tilde{b}_2 &= \frac{1}{7}\left((\nu - 4)\xi^2 + (5 - \nu)\xi^3\right), \\ \tilde{b}_2' &= \frac{1}{7}\left(2(\nu - 4)\xi + 3(5 - \nu)\xi^2\right). \end{aligned}$$

For this to vanish at $\xi = 1$ we require

$$2(\nu - 4) + 3(5 - \nu) = 0, \quad \text{so that} \quad \nu = 7.$$

From this we find

$$\begin{aligned}
\tilde{b}_2 &= \frac{\xi^2}{7}(3 - 2\xi), \\
\tilde{b}_3 &= \xi^2(\xi - 1), \\
\tilde{b}_0 &= \xi - \frac{23}{16}\xi^2 - \frac{5}{8}\xi^3.
\end{aligned}$$

■

In general we find

$$\begin{aligned}
\tilde{b}_1 &= b_1\xi^2(3 - 2\xi), \\
\tilde{b}_2 &= b_2\xi^2(3 - 2\xi), \\
\tilde{b}_3 &= \xi^2(\xi - 1), \\
\tilde{b}_0 &= \xi + \xi^2(1 - 3b_1 - 3b_2) + \xi^3(2b_1 + 2b_2 - 1).
\end{aligned}$$

An approximation at $x_{n-1} + \xi h$ can then be found from

$$y(x_{n-1} + \xi h) \approx y(x_{n-1}) + \tilde{b}_1 hf(Y_1) + \tilde{b}_2 hf(Y_2) + \tilde{b}_3 hf(Y_3) + \tilde{b}_0 hy'(x_{n-1}).$$

For third order equations there is a simpler way of approaching the interpolation problem. Since we only require an interpolant of order 3, we could use a Hermite interpolation to a function φ through two points, x_n and x_{n-1}.

The formula for Hermite interpolation through two points is given by

$$H(x) = \varphi(x_0)H_{1,0}(x) + \varphi(x_1)H_{1,1}(x) + \varphi'(x_0)\widehat{H}_{1,0}(x) + \varphi'(x_1)\widehat{H}_{1,1}(x),$$

where

$$\begin{aligned}
H_{1,0}(x) &= [1 - 2(x - x_0)L'_{10}(x_0)]L^2_{10}(x) \\
&= \left[1 - 2\frac{(x - x_0)}{(x_0 - x_1)}\right]\frac{(x - x_1)^2}{(x_0 - x_1)^2} \\
&= (1 + 2x)(x - 1)^2,
\end{aligned}$$

$$\begin{aligned}
H_{1,1}(x) &= [1 - 2(x - x_1)L'_{11}(x_1)]L^2_{11}(x) \\
&= \left[1 - 2\frac{(x - x_1)}{(x_1 - x_0)}\right]\frac{(x - x_0)^2}{(x_1 - x_0)^2} \\
&= (3 - 2x)x^2,
\end{aligned}$$

$$\begin{aligned}\widehat{H}_{1,0}(x) &= (x - x_0)L_{10}^2(x) \\ &= (x - x_0)\frac{(x - x_1)^2}{(x_0 - x_1)^2} \\ &= x(x - 1)^2,\end{aligned}$$

$$\begin{aligned}\widehat{H}_{1,1}(x) &= (x - x_1)L_{11}^2(x) \\ &= (x - x_1)\frac{(x - x_0)^2}{(x_1 - x_0)^2} \\ &= (x - 1)x^2.\end{aligned}$$

Hence an approximation at $x_{n-1} + \xi h$ can be found from

$$y(x_{n-1} + \xi h) \approx (1 - \xi)^2(1 + 2\xi)y_1^{[n-1]} + \xi(1 - \xi)^2 y_2^{[n-1]} + \xi^2(3 - 2\xi)y_1^{[n]} + \xi^2(\xi - 1)y_2^{[n]}.$$

If we rewrite this in terms of the incoming approximations and the stage derivatives we get the same interpolation formula as above.

Unfortunately, this straight forward derivation of an interpolation formula does not apply to higher orders.

3.4.3 Fourth order methods with four stages

The general form of a fourth order, four stage ARK method is

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & & & \\ a_{21} & 0 & 0 & 0 & e & c - Ae & \frac{1}{2}c^2 - Ac \\ a_{31} & a_{32} & 0 & 0 & & & \\ b_1 & b_2 & b_3 & 0 & & & \\ \hline b_1 & b_2 & b_3 & 0 & 1 & b_0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 0 & \beta_0 & 0 \end{array}\right]. \tag{3.55}$$

Order conditions

The conditions to ensure that a method of this form has the correct order and stability properties are given in Theorem 3.2.

Theorem 3.2 *A four stage method of the form* (3.55)*, with* $c = [c_1, c_2, c_3, 1]$*, has order four and is RK–stable if*

$$b_0 + b^T e = 1, \tag{3.56}$$

$$b^T c = \frac{1}{2}, \tag{3.57}$$

$$b^T c^2 = \frac{1}{3}, \tag{3.58}$$

$$b^T c^3 = \frac{1}{4}, \tag{3.59}$$

$$b^T A c = \frac{1}{6}, \tag{3.60}$$

$$b^T A c^2 = \frac{1}{12}, \tag{3.61}$$

$$\beta^T e + \beta_0 = 0, \tag{3.62}$$

$$\beta^T (I + \beta_4 A) = \beta_4 e_4^T, \tag{3.63}$$

$$c_1 = -\frac{2\exp_4(-\beta_4)}{\beta_4 \exp_3(-\beta_4)}, \tag{3.64}$$

$$(1 + \tfrac{1}{2}\beta_4 c_1) b^T A^2 c = \frac{1}{4!}. \tag{3.65}$$

Proof: A full proof of this can be found in [11], but we outline the reasoning for it here.

First, using equation (3.5), we calculate the generating function of the internal stages for the trees up to and including order 3. Recall that $\eta(t_1) = 0$ and $\eta(t_2) = 1$. For ease of notation we write $\eta_j = \eta(t_j)$.

$$\begin{aligned}
\xi(t_1) &= (c - Ae) + Ae, \\
&= c,
\end{aligned}$$

$$\begin{aligned}
\xi(t_2) &= (\frac{1}{2}c^2 - Ac) + A(\xi(t_1)), \\
&= \frac{1}{2}c^2,
\end{aligned}$$

$$\begin{aligned}
\xi(t_3) &= (\frac{1}{2}c^2 - Ac)\eta_3 + A(\xi(t_1)\xi(t_1)), \\
&= (\frac{1}{2}c^2 - Ac)\eta_3 + Ac^2,
\end{aligned}$$

$$\begin{aligned}
\xi(t_4) &= (\frac{1}{2}c^2 - Ac)\eta_4 + A(\xi(t_2)), \\
&= (\frac{1}{2}c^2 - Ac)\eta_4 + \frac{1}{2}Ac^2.
\end{aligned}$$

Next, using equation (3.6), we calculate the generating function of the first output approximation for the trees up to and including order 4.

$$\begin{aligned}
\alpha(t_1) &= b_0 + b^T e, \\
\alpha(t_2) &= b^T(\xi(t_1)), \\
&= b^T c, \\
\alpha(t_3) &= b^T(\xi(t_1)\xi(t_1)), \\
&= b^T c^2, \\
\alpha(t_4) &= b^T(\xi(t_2)), \\
&= \frac{1}{2} b^T c^2, \\
\alpha(t_5) &= b^T(\xi(t_1)\xi(t_1)\xi(t_1)), \\
&= b^T c^3, \\
\alpha(t_6) &= b^T(\xi(t_1)\xi(t_2)), \\
&= \frac{1}{2} b^T c^3, \\
\alpha(t_7) &= b^T(\xi(t_3)), \\
&= b^T(\frac{1}{2} c^2 - Ac)\eta_3 + b^T A c^2, \\
\alpha(t_8) &= b^T(\xi(t_4)), \\
&= b^T(\frac{1}{2} c^2 - Ac)\eta_4 + \frac{1}{2} b^T A c^2.
\end{aligned}$$

We do not wish the third output approximation to have any effect on the first output approximation up to order 4. For this to be the case we require that all terms containing non zero values of η have zero coefficients, so that $b^T(\frac{1}{2}c^2 - Ac) = 0$, to ensure trees t_7 and t_8 are not affected. This gives us condition (3.60). For the method to be order 4 we require that $\alpha(t_i) = E(t_i)$, where the latter values can be computed using equation (2.7). This gives us conditions (3.56)–(3.59) and (3.61). Also, we require the third output approximation to be of order 2. We need conditions (3.62) to ensure this is true. Finally, conditions (3.63)–(3.65) are the conditions given in Theorem 3.1 for Runge–Kutta stability. ■

Derivation of methods

These methods have three free parameters. For ease of calculations, we will take these to be c_2, c_3 and β_4. Once these parameters have been chosen the method can be uniquely determined from Theorem 3.2. First, c_1 can be calculated from condition (3.64). Then the b^T vector can be found from the quadrature conditions (3.57), (3.58) and (3.59), giving

$$b_1 = \frac{3 - 4c_2 - 4c_3 + 6c_2c_3}{12c_1(c_1 - c_2)(c_1 - c_3)},$$

$$b_2 = \frac{3 - 4c_1 - 4c_3 + 6c_1c_3}{12c_2(c_2 - c_1)(c_2 - c_3)},$$

$$b_3 = \frac{3 - 4c_1 - 4c_2 + 6c_1c_2}{12c_3(c_3 - c_1)(c_3 - c_2)}.$$

After finding b_0 from condition (3.56), a_{32} can be found from a linear combination of conditions (3.60) and (3.61). Provided $c_1 \neq c_2$, we find

$$a_{32} = \frac{1 - 2c_1}{12b_3c_2(c_2 - c_1)}.$$

Next we can find a_{21} and a_{31} from conditions (3.65) and (3.60) respectively:

$$a_{21} = \frac{1}{24b_3a_{32}c_1(1 + \frac{1}{2}\beta_4c_1)},$$

$$a_{31} = \frac{\frac{1}{6} - b_3a_{32}c_2 - b_2a_{21}c_1}{b_3c_1}.$$

Finally, β^T can be found from condition (3.63).

In summary, once β_4, c_2 and c_3 have been chosen and c_1 has been calculated from equation (3.64) the remaining coefficients of a fourth order method can be found from

$$\begin{array}{lcllcl}
b_1 &=& \dfrac{3 - 4c_2 - 4c_3 + 6c_2c_3}{12c_1(c_1 - c_2)(c_1 - c_3)}, & a_{32} &=& \dfrac{1 - 2c_1}{12b_3c_2(c_2 - c_1)}, \\
b_2 &=& \dfrac{3 - 4c_1 - 4c_3 + 6c_1c_3}{12c_2(c_2 - c_1)(c_2 - c_3)}, & a_{21} &=& \dfrac{1}{24b_3a_{32}c_1(1 + \frac{1}{2}\beta_4c_1)}, \\
b_3 &=& \dfrac{3 - 4c_1 - 4c_2 + 6c_1c_2}{12c_3(c_3 - c_1)(c_3 - c_2)}, & a_{31} &=& \dfrac{\frac{1}{6} - b_3a_{32}c_2 - b_2a_{21}c_1}{b_3c_1}, \\
b_0 &=& 1 - b_1 - b_2 - b_3, & \beta^T &=& \beta_4e_4^T(I + \beta_4A)^{-1}, \\
\beta_0 &=& -\beta_1 - \beta_2 - \beta_3 - \beta_4. & & &
\end{array}$$

Classification of the methods

In [11] Butcher identified several special cases based on a possible confluence between the c values. Due to the complicated relationship between β_4 and c_1 it is convenient to find combinations of these parameters which result in reasonably simple numbers. Although a reasonable number of simple pairs are known, for this we are interested in possible confluent cases, so will consider the two choices $\beta_4 = 2$, $c_1 = 1$ and $\beta_4 = \hat{\beta}_4 = 2.625816818958466716$, $c_1 = \frac{1}{2}$.

The seven special cases are given below.

Case 1: $c^T = [\frac{1}{2}, \frac{1}{2}, 1, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{8} \\ \frac{2}{a_{32}(4+\hat{\beta}_4)} & 0 & 0 & 0 & 1 & \frac{1}{2} - \frac{2}{a_{32}(4+\hat{\beta}_4)} & \frac{1}{8} - \frac{1}{a_{32}(4+\hat{\beta}_4)} \\ 2 - a_{32} - \frac{12b_2}{a_{32}(4+\hat{\beta}_4)} & a_{32} & 0 & 0 & 1 & \frac{12b_2}{a_{32}(4+\hat{\beta}_4)} - 1 & \frac{6b_2}{a_{32}(4+\hat{\beta}_4)} - \frac{1}{2} \\ \frac{2}{3} - b_2 & b_2 & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ \hline \frac{2}{3} - b_2 & b_2 & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \frac{\hat{\beta}_4{}^2(a_{32}-6b_2)}{6} & -\frac{\hat{\beta}_4{}^2}{6} & \hat{\beta}_4 & 0 & \frac{-24\hat{\beta}_4+14\hat{\beta}_4{}^2-3\hat{\beta}_4{}^3}{24+6\hat{\beta}_4} & 0 \end{array}\right],$$

with $\beta_1 = \frac{2\hat{\beta}_4{}^2(\hat{\beta}_4-4)}{3(\hat{\beta}_4+4)} + \hat{\beta}_4{}^2(b_2 - \frac{1}{6}a_{32}\hat{\beta}_4)$.

Case 2: $c^T = [1, \frac{1}{2}, 1, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\ -\frac{1}{24b_3} & \frac{1}{3b_3} & 0 & 0 & 1 & 1 - \frac{7}{24b_3} & \frac{1}{2} - \frac{1}{8b_3} \\ \frac{1}{6} - b_3 & \frac{2}{3} & b_3 & 0 & 1 & \frac{1}{6} & 0 \\ \hline \frac{1}{6} - b_3 & \frac{2}{3} & b_3 & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 4b_3 - 1 & 0 & -4b_3 & 2 & 0 & -1 & 0 \end{array}\right].$$

Case 3: $c^T = [\frac{1}{2}, 0, 1, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{8} \\ \frac{2}{a_{32}(4+\hat{\beta}_4)} & 0 & 0 & 0 & 1 & -\frac{2}{a_{32}(4+\hat{\beta}_4)} & -\frac{1}{a_{32}(4+\hat{\beta}_4)} \\ 2-\frac{12b_2}{a_{32}(4+\hat{\beta}_4)} & a_{32} & 0 & 0 & 1 & u_{32} & \frac{6b_2}{a_{32}(4+\hat{\beta}_4)}-\frac{1}{2} \\ \frac{2}{3} & b_2 & \frac{1}{6} & 0 & 1 & \frac{1}{6}-b_2 & 0 \\ \hline \frac{2}{3} & b_2 & \frac{1}{6} & 0 & 1 & \frac{1}{6}-b_2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & -\frac{\hat{\beta}_4{}^2}{6} & \hat{\beta}_4 & 0 & \beta_0 & 0 \end{array}\right],$$

with

$$\beta_1 = \frac{2\hat{\beta}_4{}^3 - 8\hat{\beta}_4{}^2}{12 + 3\hat{\beta}_4},$$

$$\beta_2 = \frac{a_{32}\hat{\beta}_4{}^3 - 6b_2\hat{\beta}_4{}^2}{6},$$

$$u_{32} = \frac{12b_2}{a_{32}(4+\hat{\beta}_4)} - 1 - a_{32},$$

$$\beta_0 = b_2\hat{\beta}_4{}^2 - \frac{1}{6}a_{32}\hat{\beta}_4{}^3 - \frac{24\hat{\beta}_4 - 14\hat{\beta}_4{}^2 + 3\hat{\beta}_4{}^3}{24 + 6\hat{\beta}_4}.$$

Case 4: $c^T = [1, \frac{1}{2}, 0, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\ -\frac{1}{24b_3} & \frac{1}{3b_3} & 0 & 0 & 1 & -\frac{7}{24b_3} & -\frac{1}{8b_3} \\ \frac{1}{6} & \frac{2}{3} & b_3 & 0 & 1 & \frac{1}{6}-b_3 & 0 \\ \hline \frac{1}{6} & \frac{2}{3} & b_3 & 0 & 1 & \frac{1}{6}-b_3 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & -4b_3 & 2 & 0 & 4b_3-1 & 0 \end{array}\right].$$

Case 5: $c^T = [c_1, \frac{1}{2}, 1, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & c_1 & \frac{1}{2}c_1^2 \\ \frac{1}{4(2c_1+\beta_4 c_1^2)} & 0 & 0 & 0 & 1 & \frac{1}{2}-\frac{1}{8c_1+4\beta_4 c_1^2} & \frac{\beta_4 c_1}{16+8\beta_4 c_1} \\ -\frac{1}{2c_1+\beta_4 c_1^2} & 2 & 0 & 0 & 1 & \frac{1}{2c_1+\beta_4 c_1^2}-1 & -\frac{\beta_4 c_1}{4+2\beta_4 c_1} \\ 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ \hline 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{\beta_4^4}{24c_1+12\beta_4 c_1^2} & \frac{\beta_4^3-2\beta_4^2}{3} & -\frac{\beta_4^2}{6} & \beta_4 & 0 & \beta_0 & 0 \end{array}\right],$$

where $\beta_0 = \dfrac{5}{6}\beta_4^2 - \dfrac{1}{3}\beta_4^3 + \dfrac{\beta_4^4}{24c_1 + 12\beta_4 c_1^2} - \beta_4$, and c_1 can be found from equation (3.64).

Case 6: $c^T = [c_1, 1, \frac{1}{2}, 1]$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & c_1 & \frac{1}{2}c_1^2 \\ \frac{c_1-1}{c_1(2c_1-1)(2+\beta_4 c_1)} & 0 & 0 & 0 & 1 & 1-\frac{c_1-1}{c_1(2c_1-1)(2+\beta_4 c_1)} & \frac{c_1(2+\beta_4(2c_1-1))}{(2c_1-1)(2+\beta_4 c_1)} \\ \frac{c_1(\beta_4(1-2c_1)-2)}{8(c_1-1)(2c_1-1)(2+\beta_4 c_1)} & \frac{2c_1-1}{8(c_1-1)} & 0 & 0 & 1 & \frac{1}{4}+\frac{-2+\beta_4-2\beta_4 c_1}{8(2c_1-1)(2+\beta_4 c_1)} & \frac{c_1(-2+\beta_4-2\beta_4 c_1)}{8(2c_1-1)(2+\beta_4 c_1)} \\ 0 & \frac{1}{6} & \frac{2}{3} & 0 & 1 & \frac{1}{6} & 0 \\ \hline 0 & \frac{1}{6} & \frac{2}{3} & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{-2\beta_4^3+\beta_4^4(1-2c_1)}{12c_1(c_1-1)(2+\beta_4 c_1)} & \beta_2 & -\frac{2}{3}\beta_4^2 & \beta_4 & 0 & \beta_0 & 0 \end{array}\right],$$

with

$$\beta_0 = \frac{\beta_4^4}{24c_1 + 12\beta_4 c_1^2} - \beta_4 + \frac{5}{6}\beta_4^3 - \frac{\beta_4^3}{12c_1},$$

$$\beta_2 = \frac{1}{6}(\beta_4^3 - \beta_4^2) + \frac{\beta_4^3}{12(c_1 - 1)}.$$

Case 7: $c^T = [c_1, c_2, c_3, 1]$

$$\left[\begin{array}{cccc|ccc}
0 & 0 & 0 & 0 & 1 & c_1 & \frac{c_1^2}{2} \\
\frac{(c_2-c_1)c_2}{c_1(1-2c_1)(2+\beta_4 c_1)} & 0 & 0 & 0 & 1 & u_{22} & u_{23} \\
a_{31} & \frac{(1-2c_1)c_3(c_3-c_1)(c_3-c_2)}{(c_2-c_1)c_2(3-4c_1-4c_2+6c_1c_2)} & 0 & 0 & 1 & u_{32} & u_{33} \\
\frac{3-4c_2-4c_3+6c_2c_3}{12c_1(c_2-c_1)(c_3-c_1)} & \frac{-3+4c_1+4c_3-6c_1c_3}{12c_2(c_2-c_1)(c_3-c_2)} & \frac{3-4c_1-4c_2+6c_1c_2}{12c_3(c_3-c_1)(c_3-c_2)} & 0 & 1 & b_0 & 0 \\
\hline
\frac{3-4c_2-4c_3+6c_2c_3}{12c_1(c_2-c_1)(c_3-c_1)} & \frac{-3+4c_1+4c_3-6c_1c_3}{12c_2(c_2-c_1)(c_3-c_2)} & \frac{3-4c_1-4c_2+6c_1c_2}{12c_3(c_3-c_1)(c_3-c_2)} & 0 & 1 & b_0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\beta_1 & \beta_2 & \frac{\beta_4^2(-3+4c_1+4c_2-6c_1c_2)}{12c_3(c_3-c_1)(c_3-c_2)} & \beta_4 & 0 & \beta_0 & 0
\end{array}\right],$$

with

$$u_{22} = \frac{-3c_1c_2 + 4c_1^2c_2 - \beta_4 c_1^2 c_2 + 2\beta_4 c_1^3 c_2 + c_2^2}{c_1(-1+2c_1)(2+\beta_4 c_1)},$$

$$u_{23} = \frac{-2c_1c_2 + 4c_1c_2^2 - \beta_4 c_1 c_2^2 + 2\beta_4 c_1^2 c_2^2}{2(-1+2c_1)(2+\beta_4 c_1)},$$

$$a_{31} = \frac{1}{b_3}\left(\frac{2c_2-1}{12c_1(c_2-c_1)} - b_2 a_{21}\right),$$

$$u_{32} = c_3 - a_{31} - a_{32},$$

$$u_{33} = \frac{1}{2}c_3^2 - a_{31}c_1 - a_{32}c_2,$$

$$b_0 = \frac{-3 + 4c_1 + 4c_2 - 6c_1c_2 + 4c_3 - 6c_1c_3 - 6c_2c_3 + 12c_1c_2c_3}{12c_1c_2c_3},$$

$$\beta_2 = \frac{3\beta_4^2 - 4\beta_4^2 c_1 - \beta_4^3 c_2 + 2\beta_4^3 c_1 c_2 - 4\beta_4^2 c_3 + \beta_4^3 c_3 + 6\beta_4^2 c_1 c_3 - 2\beta_4^3 c_1 c_3}{12(c_1c_2^2 - c_2^3 - c_1c_2c_3 + c_2^2c_3)},$$

$$\beta_1 = \frac{1 - \beta_2 c_2 - \beta_3 c_3 - \beta_4}{c_1},$$

$$\beta_0 = -\beta_1 - \beta_2 - \beta_3 - \beta_4.$$

Some example methods

Specific examples of case 4 and case 5 are given below. In the case 4 example $c^T = [1, \frac{1}{2}, 0, 1]$ and $b_3 = 1$. In the case 5 example $c^T = [\frac{11}{24}, \frac{1}{2}, 1, 1]$ and $\beta_4 = 3$.

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\ -\frac{1}{24} & \frac{1}{3} & 0 & 0 & 1 & -\frac{7}{24} & -\frac{1}{8} \\ \frac{1}{6} & \frac{2}{3} & 1 & 0 & 1 & -\frac{5}{6} & 0 \\ \hline \frac{1}{6} & \frac{2}{3} & 1 & 0 & 1 & -\frac{5}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & -4 & 2 & 0 & 3 & 0 \end{array}\right]. \tag{3.66}$$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & \frac{11}{24} & \frac{121}{1152} \\ \frac{16}{99} & 0 & 0 & 0 & 1 & \frac{67}{198} & \frac{11}{216} \\ -\frac{64}{99} & 2 & 0 & 0 & 1 & -\frac{35}{99} & -\frac{11}{54} \\ 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ \hline 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{48}{11} & 3 & -\frac{3}{2} & 3 & 0 & -\frac{3}{22} & 0 \end{array}\right]. \tag{3.67}$$

The last two methods are related to the $\frac{3}{8}$-quadrature formula. The c vectors are $c^T = [1, \frac{1}{3}, \frac{2}{3}, 1]$ and $c^T = [1, \frac{2}{3}, \frac{1}{3}, 1]$ respectively. They are examples of case 7.

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{18} & 0 & 0 & 0 & 1 & \frac{5}{18} & 0 \\ \frac{1}{18} & 1 & 0 & 0 & 1 & -\frac{7}{18} & -\frac{1}{6} \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & 0 & 1 & \frac{1}{8} & 0 \\ \hline \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & 0 & 1 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{3}{2} & -\frac{3}{2} & 2 & 0 & -\frac{3}{2} & 0 \end{array}\right]. \tag{3.68}$$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{18} & 0 & 0 & 0 & 1 & \frac{11}{18} & \frac{1}{6} \\ -\frac{5}{18} & 1 & 0 & 0 & 1 & -\frac{7}{18} & -\frac{1}{3} \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & 0 & 1 & \frac{1}{8} & 0 \\ \hline \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & 0 & 1 & \frac{1}{8} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{3}{2} & \frac{3}{2} & -\frac{3}{2} & 2 & 0 & -\frac{1}{2} & 0 \end{array}\right]. \tag{3.69}$$

Interpolation

To find a third order interpolator we need a vector $\tilde{b}^T$ that satisfies the following conditions

$$\tilde{b}^T c = \frac{\xi^2}{2}, \tag{3.70}$$

$$\tilde{b}^T c^2 = \frac{\xi^3}{3}, \tag{3.71}$$

$$\tilde{b}^T c^3 = \frac{\xi^4}{4}, \tag{3.72}$$

$$\tilde{b}^T Ac = \frac{\xi^3}{6}, \tag{3.73}$$

$$\hat{b_0} + \hat{b_1} + \hat{b_2} + \hat{b_3} + \hat{b_4} = \xi. \tag{3.74}$$

There are no free parameters, but it transpires that $\tilde{b} = b$ when $\xi = 1$ automatically. It should be noted that it is not always possible to find a suitable third order interpolator. For example, it is not possible for method (3.66). For method (3.67) the coefficients are

$$\tilde{b}_1 = \frac{3456}{143}(\xi^2 - 2\xi^3 + \xi^4),$$

$$\tilde{b}_2 = -\frac{2}{3}(33\xi^2 - 70\xi^3 + 36\xi^4),$$

$$\tilde{b}_3 = -\frac{1}{78}(-543\xi^2 + 1034\xi^3 - 504\xi^4),$$

$$\tilde{b}_4 = -\frac{1}{13}(85\xi^2 - 157\xi^3 + 72\xi^4),$$

$$\tilde{b}_0 = \frac{1}{66}\xi(66 - 171\xi + 188\xi^2 - 72\xi^3).$$

The coefficients for method (3.68) are

$$\tilde{b}_1 = -\frac{1}{8}(6\xi^2 - 16\xi^3 + 9\xi^4),$$

$$\tilde{b}_2 = \frac{3}{8}(12\xi^2 - 20\xi^3 + 9\xi^4),$$

$$\tilde{b}_3 = -\frac{3}{8}(6\xi^2 - 16\xi^3 + 9\xi^4),$$

$$\tilde{b}_4 = \frac{1}{4}(5\xi^2 - 14\xi^3 + 9\xi^4),$$

$$\tilde{b}_0 = \xi - \frac{11}{4}\xi^2 + 3\xi^3 - \frac{9}{8}\xi^4.$$

Finally, the coefficients for method (3.69) are

$$\tilde{b}_1 = \frac{1}{8}(30\xi^2 - 56\xi^3 + 27\xi^4),$$

$$\tilde{b}_2 = -\frac{3}{8}(6\xi^2 - 16\xi^3 + 9\xi^4),$$

$$\tilde{b}_3 = \frac{3}{8}(12\xi^2 - 20\xi^3 + 9\xi^4),$$

$$\tilde{b}_4 = -\frac{1}{4}(13\xi^2 - 22\xi^3 + 9\xi^4),$$

$$\tilde{b}_0 = \xi - \frac{11}{4}\xi^2 + 3\xi^3 - \frac{9}{8}\xi^4.$$

3.5 Methods with $s = p + 1$

As with traditional Runge–Kutta methods we can achieve enhanced performance if we have more stages than are required for the order. We will concentrate on the case where we have one more stage than is required.

3.5.1 RK-stability

As for the case where $s = p$ we require the stability matrix to have only one non zero eigenvalue, which is equal to $R(z)$. In the case of $s = p + 1$ we gain a free parameter K, which gives us

some control over the stability region. This is due to the fact that $R(z)$ is now given by

$$\begin{aligned} R(z) &= \exp_{s-1}(z) + Kz^s \\ &= 1 + z + \frac{z}{2} + \cdots + \frac{z^{s-1}}{(s-1)!} + Kz^s. \end{aligned}$$

3.5.2 Third order methods with four stages

A third order method with four stages takes the form

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & & & \\ a_{21} & 0 & 0 & 0 & e & c - Ae & \frac{1}{2}c^2 - Ac \\ a_{31} & a_{32} & 0 & 0 & & & \\ b_1 & b_2 & b_3 & 0 & & & \\ \hline b_1 & b_2 & b_3 & 0 & 1 & b_0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & 0 & \beta_0 & 0 \end{array}\right].$$

For a method of this form with RK stability, the stability function has the form

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + Kz^4. \tag{3.75}$$

Due to the number of free parameters we have a lot more freedom than with the three stage methods. We have some control over the stability region due to the fact that K is at our disposal. We also have control over the error in the bushy tree.

Order conditions

For a four stage method to be of order three it needs to satisfy the following conditions

$$b^T c = \frac{1}{2}, \tag{3.76}$$

$$b^T c^2 = \frac{1}{3}, \tag{3.77}$$

$$b_0 = 1 - b^T e, \tag{3.78}$$

$$\beta_0 = -\beta^T e, \tag{3.79}$$

$$\beta_4 e_4^T (I + \theta A) = \beta^T (I + \phi A + \beta_4 \theta A^2), \tag{3.80}$$

$$K\left(\frac{1}{2}\beta_4 c_1 \theta \alpha_3 - \alpha_4\right) = \left(1 + \frac{1}{2}\beta_4 c_1\right)\left(1 + \alpha_1 + \frac{\alpha_2}{2} + \frac{\alpha_3}{6}\right), \tag{3.81}$$

where the values of α_i are determined by expanding

$$\frac{1 + (\phi - \beta_4)z}{1 + \phi z + \beta_4 \theta z^2} = \sum_{i=0}^{\infty} \alpha_i z^i,$$

$$b^T A c - \frac{1}{6} = \theta(b^T A^2 c - K), \tag{3.82}$$

$$\beta_4\left(\frac{1}{2} b^T A c^2 - K\right) = (\beta_4 - \phi)(b^T A^2 c - K). \tag{3.83}$$

Derivation of methods

We have a lot more free parameters now than we did with the 3-stage methods. We will take these parameters to be c_1, c_2, c_3, β_4, ϕ, θ and L, where $L - \frac{1}{4}$ is the error coefficient corresponding to the bushy tree.

First we need to calculate the coefficients of the b vector. We can do this from the quadrature

conditions (3.76) and (3.77) and the additional condition $b^T c^3 = L$, which give

$$b_1 = \frac{-2c_2 - 2c_3 + 3c_2c_3 + 6L}{6c_1(c_1 - c_2)(c_1 - c_3)},$$

$$b_2 = \frac{-2c_1 - 2c_3 + 3c_1c_3 + 6L}{6c_2(c_2 - c_1)(c_2 - c_3)},$$

$$b_3 = \frac{-2c_1 - 2c_2 + 3c_1c_2 + 6L}{6c_3(c_3 - c_1)(c_3 - c_2)},$$

and then equation (3.78) gives a value for b_0. It is possible to find methods in which two of the c coefficients are equal; however b^T needs to be calculated slightly differently in this case.

Next we can calculate K from equation (3.81). From this we find

$$K = \frac{\left(1 + \frac{1}{2}\beta_4 c_1\right)\left(1 + \alpha_1 + \frac{\alpha_2}{2} + \frac{\alpha_3}{6}\right)}{\frac{1}{2}\beta_4 c_1 \theta \alpha_3 - \alpha_4}.$$

We can also find K from the stability function. Using the same argument as we used in Theorem 3.1, we obtain, in turn

$$\begin{aligned}
\mathrm{Tr}(BA^3U) &= K,\\
b^T A^3 e + e_4^T A^3(c - Ae) + \beta^T A^3\left(\tfrac{1}{2}c^2 - Ac\right) &= K,\\
b^T A^2 c + \frac{1}{2}\beta^T A^3 c^2 &= K,\\
\left(1 + \tfrac{1}{2}\beta_4 c_1\right) b^T A^2 c &= K. \qquad (3.84)
\end{aligned}$$

For ease of computation we will define three more variables. These are

$$K_1 = b^T A^2 c,$$

$$K_2 = b^T A c,$$

$$K_3 = b^T A c^2.$$

From equation (3.84) we find

$$K_1 = \frac{K}{1 + \frac{1}{2}c_1\beta_4}. \qquad (3.85)$$

From equation (3.82) we obtain

$$K_2 = \frac{1}{6} + \theta(K_1 - K). \qquad (3.86)$$

Finally, from equation (3.83) we find

$$K_3 = 2K - 2\left(\frac{\phi}{\beta_4} - 1\right)(K_1 - K). \tag{3.87}$$

To find an expression for a_{21} we will take combinations of K_1, K_2 and K_3. As A is strictly lower triangular we have

$$a_{21}c_1(b^T Ac^2 - c_1 b^T Ac) = c_2(c_2 - c_1)b^T A^2 c.$$

Rearranging and solving for a_{21} gives

$$a_{21} = \frac{c_2(c_2 - c_1)K_1}{c_1(K_3 - c_1K_2)}. \tag{3.88}$$

Next we can find a_{32} from equation (3.85). We obtain, in turn

$$b^T A^2 c = K_1,$$

$$b_3 a_{32} a_{21} c_1 = K_1,$$

$$a_{32} = \frac{K_1}{b_3 a_{21} c_1}.$$

Now we can find a_{31} from equation (3.87). In turn, we find

$$b^T Ac^2 = K_3,$$

$$b_2 a_{21} c_1^2 + a_{31} b_3 c_1^2 + b_3 a_{32} c_2^2 = K_3,$$

$$a_{31} = \frac{K_3 - b_2 a_{21} c_1^2 - b_3 a_{32} c_2^2}{b_3 c_1^2}.$$

The last coefficients left to find are those of the β vector. These can be found simply from equation (3.80).

In summary, once the parameters c_1, c_2, c_3, β_4, ϕ, θ and L have been chosen, the remaining parameters can be found from:

$$
\begin{aligned}
b_1 &= \tfrac{-2c_2-2c_3+3c_2c_3+6L}{6c_1(c_1-c_2)(c_1-c_3)}, & a_{21} &= \tfrac{c_2(c_2-c_1)K_1}{c_1K_3-c_1^2K_2},\\
b_2 &= \tfrac{-2c_1-2c_3+3c_1c_3+6L}{6c_2(c_2-c_1)(c_2-c_3)}, & a_{32} &= \tfrac{K_1}{b_3a_{21}c_1},\\
b_3 &= \tfrac{-2c_1-2c_2+3c_1c_2+6L}{6c_3(c_3-c_1)(c_3-c_2)}, & a_{31} &= \tfrac{K_3-b_2a_{21}c_1^2-b_3a_{32}c_2^2}{b_3c_1^2},\\
b_0 &= 1-b_1-b_2-b_3, & \beta^T &= \beta_4 e_4^T(I+\theta A)(I+\phi A+\beta_4\theta A^2)^{-1},\\
K &= \left(1+\tfrac{1}{2}\beta_4c_1\right)b^TA^2c, & \beta_0 &= -\beta^Te,\\
K_1 &= \tfrac{K}{1+\frac{1}{2}c_1\beta_4},\\
K_2 &= \tfrac{1}{6}+\theta(K_1-K),\\
K_3 &= 2K-2\left(\tfrac{\phi}{\beta_4}-1\right)(K_1-K),
\end{aligned}
$$

Some example methods

It is difficult to determine which combination of parameters gives the best method. One possible choice is to have $\phi = \beta_4 + \theta$. If we substitute this into equation (3.80) we obtain, in turn

$$\beta_4 e_4^T(I+\theta A) = \beta^T(I+\phi A+\beta_4\theta A^2),$$

$$\beta_4 e_4^T(I+\theta A) = \beta^T(I+\beta_4 A)(I+\theta A),$$

$$\beta_4 e_4^T = \beta^T(I+\beta_4 A).$$

This is the same as equation (3.32), with the subscript 3 replaced by 4. This choice of parameters greatly simplifies the method.

Below are two examples with $\phi = \beta_4+\theta$. In the first example $c^T = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1]$, $\beta_4 = 2$, $\phi = 3$, $\theta = 1$, $K = \frac{5}{216}$ and $L = \frac{1}{5}$. In the second example $c^T = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1]$, $\beta_4 = 2$, $\phi = 3$, $\theta = 1$, $K = \frac{5}{216}$ and $L = \frac{1}{4}$.

$$
\left[\begin{array}{cccc|ccc}
0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32}\\
\frac{8}{9} & 0 & 0 & 0 & 1 & -\frac{7}{18} & -\frac{7}{72}\\
-\frac{29}{6} & \frac{5}{8} & 0 & 0 & 1 & \frac{119}{24} & \frac{113}{96}\\
-\frac{14}{15} & \frac{19}{15} & \frac{2}{15} & 0 & 1 & \frac{8}{15} & 0\\
\hline
-\frac{14}{15} & \frac{19}{15} & \frac{2}{15} & 0 & 1 & \frac{8}{15} & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 0\\
\frac{32}{5} & -\frac{22}{5} & -\frac{8}{15} & 2 & 0 & -\frac{52}{15} & 0
\end{array}\right]. \tag{3.89}
$$

$$\left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} \\ \frac{8}{9} & 0 & 0 & 0 & 1 & -\frac{7}{18} & -\frac{7}{72} \\ \frac{7}{6} & \frac{1}{8} & 0 & 0 & 1 & -\frac{13}{24} & -\frac{7}{96} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 1 & 0 & 0 \\ \hline \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & -\frac{8}{3} & 2 & 0 & -\frac{4}{3} & 0 \end{array}\right]. \tag{3.90}$$

Interpolation

The conditions that need to be solved to find an interpolator for a third order four stage method are the same as those for a three stage method. These are given by conditions (3.49) – (3.51). However we now have two free parameters rather than one. We will choose one of these parameters in a similar manner to the case $s = p = 3$ such that $\tilde{b} = b$ at $\xi = 1$ and, if possible, $\tilde{b}' = 0$ at $\xi = 1$. The second parameter we will fix by choosing $\tilde{b}_4(1) = b_4 = 0$.

We will derive the interpolation coefficients for method (3.90) here. First, the conditions to be satisfied are

$$\frac{1}{4}\tilde{b}_1 + \frac{1}{2}\tilde{b}_2 + \frac{3}{4}\tilde{b}_3 = \frac{\xi^2}{2}, \tag{3.91}$$

$$\frac{1}{16}\tilde{b}_1 + \frac{1}{4}\tilde{b}_2 + \frac{9}{16}\tilde{b}_3 = \frac{\xi^3}{3}, \tag{3.92}$$

$$\tilde{b}_0 + \tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 = \xi. \tag{3.93}$$

Subtracting two times equation (3.92) from equation (3.91) leaves us with

$$\tilde{b}_1 - 3\tilde{b}_3 = 4\xi^2 - \frac{16}{3}\xi^3.$$

From this we choose $\tilde{b}_1$ and $\tilde{b}_2$ to be

$$\tilde{b}_1 = 4\xi^2 u - \frac{16}{3}\xi^3 v,$$

$$\tilde{b}_3 = -\frac{4}{3}\xi^2(1-u) + \frac{16}{9}\xi^3(1-v),$$

for some u and v. For consistency we require $\tilde{b}_1 = b_1$ and $\tilde{b}_2 = b_2$ at $\xi = 1$. Both of these conditions simplify to

$$4u - \frac{16}{3}v = \frac{2}{3}. \tag{3.94}$$

We would also like to be able to ensure $\tilde{b}_1' = 0$ at $\xi = 1$. Rearranging equation (3.94) and substituting into our equation for $\tilde{b}_1$ gives us

$$\tilde{b}_1 = 4\xi^2\left(\frac{1}{6} + \frac{4}{3}v\right) - \frac{16}{3}\xi^3 v.$$

Finding the derivative of this at $\xi = 1$ and setting equal to 0 gives $v = \frac{1}{4}$ and hence $u = \frac{1}{2}$ from equation (3.94). We now have the following expressions for $\tilde{b}_1$ and $\tilde{b}_3$

$$\tilde{b}_1 = 2\xi^2 - \frac{4}{3}\xi^3,$$

$$\tilde{b}_3 = -\frac{2}{3}\xi^2 + \frac{4}{3}\xi^3.$$

Substituting back into equation (3.91) gives us the following expression for $\tilde{b}_2$

$$\tilde{b}_2 = \xi^2 - \frac{4}{3}\xi^3.$$

Finally, equation (3.93) gives us an expression for $\tilde{b}_0$

$$\tilde{b}_0 = \frac{\xi}{3}(3 - 7\xi + 4\xi^2).$$

Following a similar procedure for the method given in (3.89) the coefficients are found to be

$$\tilde{b}_1 = -\frac{14}{5}\xi^2 + \frac{28}{15}\xi^3,$$

$$\tilde{b}_2 = \frac{\xi^2}{15}(87 - 68\xi),$$

$$\tilde{b}_3 = -\frac{34}{15}\xi^2 + \frac{36}{15}\xi^3,$$

$$\tilde{b}_4 = 0,$$

$$\tilde{b}_0 = \frac{\xi}{15}(15 - 11\xi + 4\xi^2).$$

3.5.3 Fourth order method with five stages

A fourth order five stage ARK method takes the form

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & & & \\ a_{21} & 0 & 0 & 0 & 0 & & & \\ a_{31} & a_{32} & 0 & 0 & 0 & e & c - Ae & \frac{1}{2}c^2 - Ac \\ a_{41} & a_{42} & a_{43} & 0 & 0 & & & \\ b_1 & b_2 & b_3 & b_4 & 0 & & & \\ \hline b_1 & b_2 & b_3 & b_4 & 0 & 1 & b_0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & 0 & \beta_0 & 0 \end{array}\right],$$

with stability function

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + Kz^5$$

Order conditions

For an ARK method with five stages to have order four it needs to satisfy the conditions

$$b_0 = 1 - b^T e, \tag{3.95}$$

$$b^T c = \frac{1}{2}, \tag{3.96}$$

$$b^T c^2 = \frac{1}{3}, \tag{3.97}$$

$$b^T c^3 = \frac{1}{4}, \tag{3.98}$$

$$b^T Ac = \frac{1}{6}, \tag{3.99}$$

$$b^T Ac^2 = \frac{1}{12}, \tag{3.100}$$

$$\beta_0 = -\beta^T e, \tag{3.101}$$

$$\beta^T c = 1, \tag{3.102}$$

$$\beta^T Ac = \frac{\theta\beta_5 + \beta_5 - \phi}{\theta\beta_5}, \tag{3.103}$$

$$\beta_5 e_5^T (I + \theta A) = \beta^T (I + \phi A + \beta_5 \theta A^2), \tag{3.104}$$

$$K\left(\frac{1}{2}\beta_5 c_1 \theta\alpha_4 - \alpha_5\right) = \left(1 + \frac{1}{2}\beta_5 c_1\right)\left(1 + \alpha_1 + \frac{\alpha_2}{2} + \frac{\alpha_3}{6} + \frac{\alpha_4}{24}\right), \tag{3.105}$$

where the values of α_i are determined by expanding

$$\frac{1 + (\phi - \beta_5)z}{1 + \phi z + \beta_5\theta z^2} = \sum_{i=0}^{\infty} \alpha_i z^i,$$

$$b^T A^2 c - \frac{1}{24} = \theta(b^T A^3 c - K), \tag{3.106}$$

$$\beta_5\left(\frac{1}{2} b^T A^2 c^2 - K\right) = (\beta_5 - \phi)(b^T A^3 c - K). \tag{3.107}$$

Derivation of methods

As there are so many free parameters it would be desirable to make these parameters the ones which we would most like to have control over. We will change the derivation slightly to allow us to make K a free parameter. This makes the derivation more complicated than in the third order case, but gives us more control over the stability function. The two parameters θ and β_5 appear together most of the time. We will create a new parameter $\mu = \theta\beta_5$. Our free parameters are now c_1, c_2, c_3, c_4, ϕ, K, a_{43} and L, where $L - \frac{1}{5}$ is the error in the bushy tree.

- First we need to calculate μ. This can be done by solving equation (3.105). There are only a relatively small number of choices for c_1, K and ϕ which give real, rational values for μ. Some aesthetically pleasing choices are

$$c_1 = \frac{1}{5}, \qquad \phi = 2, \qquad K = \frac{1}{120}, \qquad \mu = \frac{4}{3} \tag{3.108}$$

$$c_1 = \frac{1}{4}, \qquad \phi = 4, \qquad K = \frac{1}{120}, \qquad \mu = 8 \tag{3.109}$$

$$c_1 = \frac{1}{3}, \qquad \phi = 3, \qquad K = \frac{1}{120}, \qquad \mu = 6 \tag{3.110}$$

$$c_1 = \frac{1}{3}, \qquad \phi = 6, \qquad K = \frac{1}{120}, \qquad \mu = 12 \tag{3.111}$$

- The b^T vector can be found from equations (3.96)-(3.98) and $b^T c^4 = L$. Solving gives

$$b_1 = \frac{3c_2 + 3c_3 + 3c_4 - 4c_2c_3 - 4c_2c_4 - 4c_3c_4 + 6c_2c_3c_4 - 12L}{12c_1(c_2 - c_1)(c_1 - c_3)(c_1 - c_4)},$$

$$b_2 = \frac{3c_1 + 3c_3 + 3c_4 - 4c_1c_3 - 4c_1c_4 - 4c_3c_4 + 6c_1c_3c_4 - 12L}{12c_2(c_1 - c_2)(c_2 - c_3)(c_2 - c_4)},$$

$$b_3 = \frac{3c_1 + 3c_2 + 3c_4 - 4c_1c_2 - 4c_1c_4 - 4c_2c_4 + 6c_1c_2c_4 - 12L}{12c_3(c_1 - c_3)(c_3 - c_2)(c_3 - c_4)},$$

$$b_4 = \frac{3c_1 + 3c_2 + 3c_3 - 4c_1c_2 - 4c_1c_3 - 4c_2c_3 + 6c_1c_2c_3 - 12L}{12c_4(c_1 - c_4)(c_2 - c_4)(c_3 - c_4)}.$$

- Next we need to calculate θ. From equation (3.106) we have

$$\theta = \frac{\frac{1}{24} - b^T A^2 c}{K - b^T A^3 c}. \tag{3.112}$$

In order to evaluate this we need to know the values of $b^T A^2 c$ and $b^T A^3 c$. These can be found by rearranging equation (3.104) and substituting into equations (3.102) and (3.103). First, equation (3.102) gives

$$\beta_5 e_5^T (I + \theta A)(I + \phi A + \mu A^2)^{-1} c = 1.$$

A series expansion of the above gives

$$e_5^T (\gamma_0 + \gamma_1 A + \gamma_2 A^2 + \gamma_3 A^3 + \gamma_4 A^4) c - 1 = 0,$$

where γ_i is given by

$$\frac{\beta_5 + \mu z}{1 + \phi z + \mu z^2} = \sum_{i=0}^{\infty} \gamma_i z^i.$$

Using the appropriate order conditions this reduces to

$$\gamma_0 + \frac{\gamma_1}{2} + \frac{\gamma_2}{6} + \gamma_3 b^T A^2 c + \gamma_4 b^T A^3 c - 1 = 0. \tag{3.113}$$

Next, equation (3.103) gives

$$\beta_5 e_5^T (I + \theta A)(I + \phi A + \mu A^2)^{-1} A c = \frac{\mu + \beta_5 - \phi}{\mu}.$$

A series expansion and simplification using the order conditions gives

$$\frac{\gamma_0}{2} + \frac{\gamma_1}{6} + \gamma_2 b^T A^2 c + \gamma_3 b^T A^3 c - \left(\frac{\mu + \beta_5 - \phi}{\mu} \right) = 0. \tag{3.114}$$

Simultaneously solving equations (3.113) and (3.114) gives us values for $b^T A^2 c$ and $b^T A^3 c$, which can be substituted into equation (3.112) to find θ.

- Calculate $\beta_5 = \dfrac{\mu}{\theta}$.

- We now introduce three new temporary variables to aid the calculations. These are

$$K_1 = b^T A^3 c,$$

$$K_2 = b^T A^2 c,$$

$$K_3 = b^T A^2 c^2.$$

An expression for K_1 can be found by looking at the stability matrix. As we have already seen, we require the trace to be equal to $R(z)$. This implies we have

$$\begin{aligned} K &= \mathrm{Tr}(BA^4U) \\ &= b^T A^4 e + e_5^T A^4 (c - Ae) + \beta^T A^4 (\tfrac{1}{2}c^2 - Ac) \\ &= b^T A^3 c + \tfrac{1}{2}\beta^T A^4 c^2 \\ &= (1 + \tfrac{1}{2}\beta_5 c_1) b^T A^3 c, \end{aligned}$$

giving

$$K_1 = \frac{K}{1 + \frac{c_1\beta_5}{2}}. \tag{3.115}$$

An expression for K_2 can be found by rearranging equation (3.106), giving

$$K_2 = \frac{1}{24} + \theta(K_1 - K). \tag{3.116}$$

Similarly, an expression can be found for K_3 by rearranging equation (3.107), giving

$$K_3 = 2K - 2\left(\frac{\phi}{\beta_5} - 1\right)(K_1 - K). \tag{3.117}$$

- To find an expression for a_{21} we will take combinations of K_1, K_2 and K_3. As A is strictly lower triangular we have

$$a_{21} c_1 (b^T A^2 c^2 - c_1 b^T A^2 c) = c_2 (c_2 - c_1) b^T A^3 c.$$

Rearranging and solving for a_{21} gives

$$a_{21} = \frac{c_2(c_2 - c_1)K_1}{c_1(K_3 - c_1 K_2)}. \tag{3.118}$$

- We will use a similar technique to calculate a_{32} from

$$b^T A^2 c^2 - c_1 b^T A^2 c = b_4 a_{43} a_{32} c_2 (c_2 - c_1),$$

giving

$$a_{32} = \frac{K_3 - c_1 K_2}{b_4 a_{43} c_2 (c_2 - c_1)}. \tag{3.119}$$

- To find a_{31}, solve $b^T A^2 c = K_2$, giving

$$a_{31} = \frac{K_2 - a_{21} a_{32} b_3 c_1 - a_{21} a_{42} b_4 c_1 - a_{32} a_{43} b_4 c_2}{a_{43} b_4 c_1}.$$

- An expression for a_{42} can be found from solving the linear combination of equations (3.99) and (3.100). i.e. by solving the equation

$$b^T A c (c - c_1) = \frac{1}{12} - \frac{c_1}{6},$$

for a_{42}.

- To find a_{41}, solve equation (3.99).

- The β^T vector can be found by rearranging equation (3.104) giving

$$\beta^T = \beta_5 e_5^T (I + \theta A)(I + \phi A + \beta_5 \theta A^2)^{-1}.$$

- The U matrix can be found by simply forming the matrix $\left[e, c - Ae, \frac{1}{2}c^2 - Ac\right]$.

- Similarly, B can be found by augmenting the vectors b^T, e_5^T and β^T.

- The only non-constant elements of V are b_0 and β_0 which can be found from equations (3.95) and (3.101) respectively.

Some example methods

We present here two example methods. They have both been chosen for their relatively simple tableaux. Although neither of them have been optimised, we have chosen $L = \frac{1}{5}$ and $K = \frac{1}{120}$ for both methods, ensuring zero error for both the bushy tree and the tall tree.

In this first method the remaining free parameters have been chosen to be $c = [\frac{1}{3}, \frac{1}{2}, \frac{3}{4}, 1, 1]^T$, $\phi = 3$ and $a_{43} = 1$.

$$\left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\ \frac{3}{16} & 0 & 0 & 0 & 0 & 1 & \frac{5}{16} & \frac{1}{16} \\ -\frac{75}{64} & \frac{5}{3} & 0 & 0 & 0 & 1 & \frac{49}{192} & -\frac{31}{192} \\ \frac{45}{4} & -\frac{113}{12} & 1 & 0 & 0 & 1 & -\frac{11}{6} & \frac{17}{24} \\ \frac{27}{50} & -\frac{2}{15} & \frac{32}{75} & \frac{1}{15} & 0 & 1 & \frac{1}{10} & 0 \\ \hline \frac{27}{50} & -\frac{2}{15} & \frac{32}{75} & \frac{1}{15} & 0 & 1 & \frac{1}{10} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{468}{125} & \frac{42}{25} & \frac{8}{125} & \frac{4}{25} & \frac{6}{5} & 0 & \frac{16}{25} & 0 \end{array}\right]. \tag{3.120}$$

In the second method the remaining free parameters have been chosen to be $c = [\frac{1}{5}, \frac{1}{2}, \frac{3}{4}, 1, 1]^T$, $\phi = 2$ and $a_{43} = 1$.

$$\left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{5} & \frac{1}{50} \\ \frac{75}{248} & 0 & 0 & 0 & 0 & 1 & \frac{49}{248} & \frac{2}{31} \\ \frac{60375}{992} & -\frac{93}{4} & 0 & 0 & 0 & 1 & -\frac{36567}{992} & -\frac{33}{124} \\ -\frac{476125}{2046} & \frac{2987}{33} & \frac{4}{3} & 0 & 0 & 1 & \frac{290249}{2046} & \frac{535}{682} \\ \frac{125}{396} & \frac{2}{9} & \frac{32}{99} & \frac{1}{12} & 0 & 1 & \frac{1}{18} & 0 \\ \hline \frac{125}{396} & \frac{2}{9} & \frac{32}{99} & \frac{1}{12} & 0 & 1 & \frac{1}{18} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{203200}{891} & -\frac{31438}{405} & \frac{17032}{4455} & \frac{28}{15} & -\frac{158}{15} & 1 & -\frac{58964}{405} & 0 \end{array}\right]. \tag{3.121}$$

Interpolation

The equations that need to be solved to find a third order interpolator are those given in equations (3.70)–(3.74). We now have one free parameter. We will choose this such that an extra fourth order condition is satisfied. That is

$$\tilde{b}Ac^2 = \frac{\xi^4}{12}. \tag{3.122}$$

We can still also satisfy the consistency condition that $\tilde{b}^T = b^T$ at $\xi = 1$.

For the method given in (3.120) the coefficients are

$$\tilde{b}_1 = \frac{27}{1150}\xi^2(382 - 672\xi + 313\xi^2),$$

$$\tilde{b}_2 = -\frac{2}{345}\xi^2(1257 - 2422\xi + 1188\xi^2),$$

$$\tilde{b}_3 = \frac{16}{1725}\xi^2(189 - 194\xi + 51\xi^2),$$

$$\tilde{b}_4 = \frac{1}{345}\xi^2(117 - 142\xi + 48\xi^2),$$

$$\tilde{b}_5 = -\frac{1}{2}\xi^2(1 - \xi^2),$$

$$\tilde{b}_0 = \frac{1}{230}\xi(230 - 753\xi + 908\xi^2 - 362\xi^3).$$

The coefficients for method (3.121) are

$$\tilde{b}_1 = \frac{125}{1964952}\xi^2(83749 - 147650\xi + 68863\xi^2),$$

$$\tilde{b}_2 = -\frac{2}{22329}\xi^2(16762 - 43448\xi + 24205\xi^2),$$

$$\tilde{b}_3 = \frac{32}{245619}\xi^2(938 + 8048\xi - 6505\xi^2),$$

$$\tilde{b}_4 = \frac{\xi^2}{59544}(1733 + 16382\xi - 13153\xi^2),$$

$$\tilde{b}_5 = \frac{\xi^2}{3308}(213 - 3734\xi + 3521\xi^2),$$

$$\tilde{b}_0 = \frac{\xi}{44658}(44658 - 180508\xi + 236966\xi^2 - 98635\xi^3).$$

CHAPTER 4

A special 'fifth' order method

Mathematics consists of proving the most obvious thing in the least obvious way.

George Polyá

4.1 Introduction

As is always the case when a method has many free parameters, it is difficult to know what choice of parameters is going to give optimal performance. In the course of optimising fourth order methods with five stages a special method was found that had zero error coefficients for the fifth order trees. The values of the free parameters for this method are $c = [\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 1]^T$, $L = \frac{1}{5}$, $K = \frac{1}{120}$, $\phi = 4$ and $a_{43} = \frac{8}{7}$. The defining matrices of the method are

$$\left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} \\ \frac{2}{5} & 0 & 0 & 0 & 0 & 1 & \frac{1}{10} & \frac{1}{40} \\ \frac{27}{160} & \frac{75}{128} & 0 & 0 & 0 & 1 & -\frac{3}{640} & -\frac{69}{1280} \\ \frac{69}{35} & -\frac{51}{28} & \frac{8}{7} & 0 & 0 & 1 & -\frac{41}{140} & \frac{17}{280} \\ \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\ \hline \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{1352}{225} & \frac{34}{15} & -\frac{256}{75} & -\frac{196}{225} & \frac{24}{5} & 0 & \frac{242}{75} & 0 \end{array}\right], \tag{4.1}$$

with stability matrix

$$M = \begin{bmatrix} 1+\frac{83}{90}z+\frac{19}{45}z^2+\frac{3}{32}z^3+\frac{1}{48}z^4 & \frac{7}{90}+\frac{7}{90}z+\frac{7}{96}z^2+\frac{1}{64}z^3+\frac{1}{192}z^4 & \frac{1}{192}z^2+\frac{1}{384}z^3+\frac{1}{1536}z^4 \\ z+\frac{83}{90}z^2+\frac{19}{45}z^3+\frac{3}{32}z^4+\frac{1}{48}z^5 & \frac{7}{90}z+\frac{7}{90}z^2+\frac{7}{96}z^3+\frac{1}{64}z^4+\frac{1}{192}z^5 & \frac{1}{192}z^3+\frac{1}{384}z^4+\frac{1}{1536}z^5 \\ -\frac{242}{75}z+\frac{367}{225}z^2+\frac{111}{100}z^3+\frac{13}{60}z^4+\frac{1}{10}z^5 & \frac{242}{75}-\frac{142}{225}z-\frac{1}{100}z^2+\frac{5}{24}z^3+\frac{1}{60}z^4+\frac{1}{40}z^5 & \frac{1}{192}z^4+\frac{1}{320}z^5 \end{bmatrix}.$$

The eigenvalues of M are $\{1+z+\frac{1}{2}z^2+\frac{1}{6}z^3+\frac{1}{24}z^4+\frac{1}{120}z^5, 0, 0\}$ which is consistent with the RK stability of the method. For this method, the stability function satisfies $R(z) = \exp(z) + O(z^6)$. This property, along with the observation that $b^T c^4 = \frac{1}{5}$, just as for a fifth order Runge–Kutta method, suggests the possibility that we may be able to obtain an order enhancement.

To understand the behaviour of the method, we consider what happens when the exact values $y(x_{n-1})$ and $hy'(x_{n-1})$ are used as incoming approximations to $y_1^{[n-1]}$ and $y_2^{[n-1]}$ respectively with an approximation, accurate only to within $O(h^3)$, of $h^2y''(x_{n-1})$ used for $y_3^{(n-1)}$. We want to carry out the analysis only to within $O(h^6)$ so we need only use trees up to order 5. Because we wish to carry out formal Taylor expansions about x_{n-1}, we represent the first two incoming approximations by $\mathbf{1}$ and D respectively. The third input approximation is represented by a mapping η and we will write $\eta(t_i) = \eta_i$, where i takes values 0 to 17. Because this input quantity approximates $h^2y''(x_{n-1})$ up to h^2 terms, we assume that $\eta_0 = 0$, $\eta_1 = 0$ and $\eta_2 = 1$. We are now in a position to calculate the tree mappings corresponding to the stages, stage derivatives and output approximations. We denote the mappings representing the stages by ξ_i, with $i = 1, 2, \ldots, 5$. From Equation (3.5), we can calculate these in sequence, together with $\xi_i D$ using

$$\xi_i = \mathbf{1} + u_{i2}D + u_{i3}\eta + \sum_{j<i} a_{ij}\xi_j D, \quad i = 1, 2, \ldots, 5.$$

Because ξ_5 also corresponds to the first output approximation and $\xi_5 D$ to the second output approximation, it is sufficient in assessing the accuracy of these approximations to compare ξ_5 to E up to trees of the required order. To assess the quality of the third output approximation we calculate

$$\widehat{\eta} = E^{-1}\Big(\beta_0 D + \sum_{i=1}^{5} \beta_i \xi_i D\Big),$$

where $\beta_0 = v_{32}$ and $\beta_i = b_{3i}$, $i = 1, 2, \ldots, 5$. The factor E^{-1} is introduced because we wish

to carry out the Taylor expansion for this output approximation about x_n, rather than about x_{n-1}.

The results of these calculations, where we show only the essential details, are presented in Table 4.1.

The effect we observe, in which the values of η_i, for $i \geq 3$, do not enter into fourth order terms in the last two columns of Table 4.1, is the result of the so called annihilation conditions we have imposed on the method.

Since the order is determined by terms ξ_i of the penultimate column of this table, which are equal to reciprocals of $\gamma(t)$, we see that order 4 is assured, even if we start the method off with the standard crude approximation to $h^2y''(x_0)$ as would correspond to $(\mathbf{1}+D)D-D$ (the difference between the derivative found after one step of the Euler method and the derivative computed at x_0, as given in (3.3)).

Moreover, we see that if $\eta_3 = \frac{1}{5}$ and $\eta_4 = \frac{1}{10}$ all entries of the penultimate column are equal to reciprocals of $\gamma(t)$, so that order 5 may be achieved for this choice, at least for fixed stepsizes. For example, to obtain a starting value $y_3^{[0]} = h^2y''(x_0) + O(h^3)$, whose leading coefficients in the expansion would have values $\frac{1}{5}$ and $\frac{1}{10}$ in the two terms of order three, we might use the generalized tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{10} & \frac{1}{10} \\ \hline 0 & -5 & 5 \end{array} .$$

This can be interpreted as a standard Runge–Kutta methods, with a coefficient of 0 for the $y(x_n)$ component. That is, once the stage values have been calculated as usual, the output approximation can be found from

$$y_3^{[0]} = h(-5F_1 + 5F_2).$$

The coefficients of this tableau can be found by solving the modified order conditions

$$b^T = e,$$

$$b^T c = 1,$$

$$b^T c^2 = \frac{1}{5},$$

$$b^T Ac = \frac{1}{10}.$$

i	$(\xi_1 D)(t_i)$	$(\xi_2 D)(t_i)$	$(\xi_3 D)(t_i)$	$(\xi_4 D)(t_i)$	$(\xi_5 D)(t_i)$	$\xi_5(t_i)$	$\widehat{\eta}(t_i)$
0	0	0	0	0	0	1	0
1	1	1	1	1	1	1	0
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	1	$\frac{1}{2}$	1
3	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{9}{16}$	1	1	$\frac{1}{3}$	$\frac{1}{5}$
4	$\frac{1}{32}$	$\frac{1}{8}$	$\frac{9}{32}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{10}$
5	$\frac{1}{64}$	$\frac{1}{8}$	$\frac{27}{64}$	1	1	$\frac{1}{4}$	$-\frac{553}{600}$
6	$\frac{1}{128}$	$\frac{1}{16}$	$\frac{27}{128}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$	$-\frac{553}{1200}$
7	$\frac{\eta_3}{32}$	$\frac{1}{40}+\frac{\eta_3}{40}$	$\frac{201}{1280}-\frac{69\eta_3}{1280}$	$\frac{87}{280}+\frac{17\eta_3}{280}$	$\frac{1}{3}$	$\frac{1}{12}$	$-\frac{7}{20}$
8	$\frac{\eta_4}{32}$	$\frac{1}{80}+\frac{\eta_4}{40}$	$\frac{201}{2560}-\frac{69\eta_4}{1280}$	$\frac{87}{560}+\frac{17\eta_4}{280}$	$\frac{1}{6}$	$\frac{1}{24}$	$-\frac{7}{40}$
9	$\frac{1}{256}$	$\frac{1}{16}$	$\frac{81}{256}$	1	1	$\frac{1}{5}$	$\frac{1163}{800}$
10	$\frac{1}{512}$	$\frac{1}{32}$	$\frac{81}{512}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{10}$	$\frac{1163}{1600}$
11	$\frac{\eta_3}{128}$	$\frac{1}{80}+\frac{\eta_3}{80}$	$\frac{603}{5120}-\frac{207\eta_3}{5120}$	$\frac{87}{280}+\frac{17\eta_3}{280}$	$\frac{1}{3}$	$\frac{13}{192}-\frac{\eta_3}{192}$	$\frac{247}{500}+\frac{133\eta_3}{2000}$
12	$\frac{\eta_4}{128}$	$\frac{1}{160}+\frac{\eta_4}{80}$	$\frac{603}{10240}-\frac{207\eta_4}{5120}$	$\frac{87}{560}+\frac{17\eta_4}{280}$	$\frac{1}{6}$	$\frac{13}{384}-\frac{\eta_4}{192}$	$\frac{247}{1000}+\frac{133\eta_4}{2000}$
13	$\frac{1}{1024}$	$\frac{1}{64}$	$\frac{81}{1024}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{20}$	$\frac{1163}{3200}$
14	$\frac{\eta_5}{32}$	$\frac{1}{160}+\frac{\eta_5}{40}$	$\frac{777}{10240}-\frac{69\eta_5}{1280}$	$\frac{639}{2240}+\frac{17\eta_5}{280}$	$\frac{1}{4}$	$\frac{1}{20}$	$\frac{137}{300}$
15	$\frac{\eta_6}{32}$	$\frac{1}{320}+\frac{\eta_6}{40}$	$\frac{777}{20480}-\frac{69\eta_6}{1280}$	$\frac{639}{4480}+\frac{17\eta_6}{280}$	$\frac{1}{8}$	$\frac{1}{40}$	$\frac{137}{600}$
16	$\frac{\eta_7}{32}$	$\frac{\eta_3}{80}+\frac{\eta_7}{40}$	$\frac{15}{1024}+\frac{51\eta_3}{2560}-\frac{69\eta_7}{1280}$	$\frac{15}{112}-\frac{51\eta_3}{1120}+\frac{17\eta_7}{280}$	$\frac{1}{12}$	$\frac{1}{64}+\frac{\eta_3}{192}$	$\frac{3}{20}$
17	$\frac{\eta_8}{32}$	$\frac{\eta_4}{80}+\frac{\eta_8}{40}$	$\frac{15}{2048}+\frac{51\eta_4}{2560}-\frac{69\eta_8}{1280}$	$\frac{15}{224}-\frac{51\eta_4}{1120}+\frac{17\eta_8}{280}$	$\frac{1}{24}$	$\frac{1}{128}+\frac{\eta_4}{192}$	$\frac{3}{40}$

Table 4.1: Algebraic analysis of the special 5 stage method.

Even if we start the computation in the naive way based on the generalized tableau (3.3), after a single step, the output value corresponding to $\widehat{\eta}$ will have the correct values of η_3 and η_4 and, from this point onwards, the method will maintain fifth order behaviour, at least with constant stepsize.

In the next section, we will see that it is possible to obtain fifth order accuracy even if the stepsize varies from step to step.

4.2 Obtaining order 5 performance

Although the method has order 5 behaviour for fixed stepsize, some sort of adjustment is necessary to extend this behaviour to variable h. This is typical of any multistep method but our aim is to ensure that the additional cost involved with changing stepsize is minimal. To maintain only order 4 behaviour, changing stepsize according to the Nordsieck technique is quite satisfactory. Let $\overline{h} = \rho h$ denote the stepsize to be used in step number $n+1$, after step n has been completed with stepsize h. The output quantities from step n are approximations to $y(x_n)$, $hy'(x_n)$ and $h^2y''(x_n)$, respectively. Since the first two of these are accurate to within $O(h^6)$, it will be satisfactory to adjust these, as input to the next step, by leaving the first unchanged and scaling the second by the stepsize ratio ρ. Hence, the second input component will become an approximation to $\rho \cdot hy'(x_n) = \overline{h}y'(x_n)$. Adjusting the third component by a factor ρ^2 will not be an adequate correction, because this component consists of several terms which we can write as

$$h^2y''(x_n) + h^3\frac{1}{10}y^{(3)}(x_n) + h^4\Phi(x_n) + O(h^5),$$

where $\Phi(x_n)$ is a linear combination of elementary differentials whose coefficients can be found from the last column of Table 4.1. When we adjust $y_3^{(n)}$ for input to step number $n+1$ by multiplying by ρ^2, we obtain

$$\rho^2\left(h^2y''(x_n) + h^3\frac{1}{10}y^{(3)}(x_n) + h^4\Phi(x_n) + O(h^5)\right) = \overline{h}^2y''(x_n) + \frac{\overline{h}^3}{\rho}\frac{1}{10}y^{(3)}(x_n) + \frac{\overline{h}^4}{\rho^2}\Phi(x_n) + O(h^5),$$

which will not give the correct result at the end of this step unless, somehow, the factor ρ^{-1} can be removed from the $\overline{h}^3$ term.

We propose to do this by replacing the third output approximation, by two approximations which will approximate

$$h^2y''(x_n) + \theta h^4\Phi(x_n)$$

and

$$h^3\frac{1}{10}y^{(3)}(x_n)+(1-\theta)h^4\Phi(x_n)$$

respectively. These will be scaled by ρ^2 and ρ^3 respectively so that their sum will be

$$\overline{h}^2y''(x_n)+\overline{h}^3\frac{1}{10}y^{(3)}(x_n)+\overline{h}^4\Phi(x_n)\left(\frac{\theta}{\rho^2}+\frac{1-\theta}{\rho}\right)+O(h^5).$$

The quality of this as an approximation, compared with what would have been received as input to step number $n+1$ if the stepsize had been constant with value $\overline{h}$, is determined by how close $\theta/\rho^2+(1-\theta)/\rho$ is to 1 for $\rho\approx 1$. A suitable value for θ is $\theta=-1$ because

$$\frac{\theta}{\rho^2}+\frac{1-\theta}{\rho}-1=-\frac{(\rho-1)((\theta+1)+(\rho-1))}{\rho^2}.$$

We now partition the vector $\left[\begin{array}{cccccc}\beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5\end{array}\right]^T$ in the form

$$\beta^T=\widehat{\beta}^T+\overline{\beta}^T+\widetilde{\beta}^T.$$

The three components are chosen so that, if the β values in the last row of B and the value of v_{32}, are replaced by the values in these components, then the elements up to order 4 corresponding to the $\widehat{\eta}$ column of Table 4.1 are respectively

$$\left[\begin{array}{c}0\\0\\1\\0\\0\\0\\0\\0\\0\\0\end{array}\right],\quad \left[\begin{array}{c}0\\0\\0\\\frac{1}{5}\\\frac{1}{10}\\0\\0\\0\\0\end{array}\right],\quad \left[\begin{array}{c}0\\0\\0\\0\\0\\-\frac{553}{600}\\-\frac{553}{1200}\\-\frac{7}{20}\\-\frac{7}{40}\end{array}\right], \tag{4.2}$$

which are independent of the values of η_3 and η_4.

We now look at the conditions that make this possible. We start by looking at the conditions for $\widehat{\beta}^T$. From (3.7) we obtain an expression for $(E\widehat{\eta})(t_i)$. Using the composition rule, an

alternative expression can be found. Equating these two expressions, and solving for $\widehat{\eta}(t_i)$ leads to the following conditions

$$\widehat{\eta}(t_1) = \widehat{\beta}_0 + \widehat{\beta}^T e - E(t_1)\widehat{\eta}(\emptyset), \tag{4.3}$$

$$\widehat{\eta}(t_2) = \widehat{\beta}^T c - E(t_2)\widehat{\eta}(\emptyset) - E(t_1)\widehat{\eta}(t_1), \tag{4.4}$$

$$\widehat{\eta}(t_3) = \widehat{\beta}^T c^2 - E(t_3)\widehat{\eta}(\emptyset) - E(t_1)E(t_1)\widehat{\eta}(t_1) - 2E(t_1)\widehat{\eta}(t_2), \tag{4.5}$$

$$\widehat{\eta}(t_4) = \frac{1}{2}\widehat{\beta}^T c^2 - E(t_4)\widehat{\eta}(\emptyset) - E(t_1)\widehat{\eta}(t_2) - E(t_2)\widehat{\eta}(t_1), \tag{4.6}$$

$$\widehat{\eta}(t_5) = \widehat{\beta}^T c^3 - E(t_5)\widehat{\eta}(\emptyset) - E(t_1)E(t_1)E(t_1)\widehat{\eta}(t_1) - 3E(t_1)E(t_1)\widehat{\eta}(t_2) - 3E(t_1)\widehat{\eta}(t_3), \tag{4.7}$$

$$\widehat{\eta}(t_6) = \frac{1}{2}\widehat{\beta}^T c^3 - E(t_6)\widehat{\eta}(\emptyset) - E(t_1)E(t_2)\widehat{\eta}(t_1) - E(t_2)\widehat{\eta}(t_2) - E(t_1)E(t_1)\widehat{\eta}(t_2), \tag{4.8}$$

$$- E(t_1)\widehat{\eta}(t_4) - E(t_1)\widehat{\eta}(t_3), \tag{4.9}$$

$$\widehat{\eta}(t_7) = \widehat{\beta}^T Ac^2 - E(t_7)\widehat{\eta}(\emptyset) - E(t_3)\widehat{\eta}(t_1) - 2E(t_1)\widehat{\eta}(t_4) - E(t_1)E(t_1)\widehat{\eta}(t_2), \tag{4.10}$$

$$\widehat{\eta}(t_8) = \frac{1}{2}\widehat{\beta}^T Ac^2 - E(t_8)\widehat{\eta}(\emptyset) - E(t_4)\widehat{\eta}(t_1) - E(t_2)\widehat{\eta}(t_2) - E(t_1)\widehat{\eta}(t_4). \tag{4.11}$$

As we wish $\widehat{\eta}$ to be equal to the first column of (4.2) this leads to the following conditions

$$\widehat{\beta}_0 + \widehat{\beta}^T e = 0,$$

$$\widehat{\beta}^T c = 1,$$

$$\widehat{\beta}^T c^2 = 2,$$

$$\frac{1}{2}\widehat{\beta}^T c^2 = 1,$$

$$\widehat{\beta}^T c^3 = 3,$$

$$\frac{1}{2}\widehat{\beta}^T c^3 = \frac{3}{2},$$

$$\widehat{\beta}^T Ac^2 = 1,$$

$$\frac{1}{2}\widehat{\beta}^T Ac^2 = \frac{1}{2}.$$

A similar analysis for the vector $\overline{\beta}^T$, using the second column vector of (4.2), leads to the following conditions

$$\overline{\beta}_0 + \overline{\beta}^T e = 0,$$

$$\overline{\beta}^T c = 0,$$

$$\overline{\beta}^T c^2 = \frac{1}{5},$$

$$\frac{1}{2}\overline{\beta}^T c^2 = \frac{1}{10},$$

$$\overline{\beta}^T c^3 = \frac{3}{5},$$

$$\frac{1}{2}\overline{\beta}^T c^3 = \frac{3}{10},$$

$$\overline{\beta}^T Ac^2 = \frac{1}{5},$$

$$\frac{1}{2}\overline{\beta}^T Ac^2 = \frac{1}{10}.$$

The final vector, $\widetilde{\beta}^T$, can simply be found from

$$\widetilde{\beta}^T = \beta^T - \widehat{\beta}^T - \overline{\beta}^T.$$

One possible solution to the above equations is

$$\begin{bmatrix} \frac{242}{75} \\ -\frac{1352}{225} \\ \frac{34}{15} \\ -\frac{256}{75} \\ -\frac{196}{225} \\ \frac{24}{5} \end{bmatrix} = \begin{bmatrix} -\frac{7}{45} \\ -\frac{32}{45} \\ \frac{76}{15} \\ -\frac{512}{45} \\ -\frac{532}{45} \\ 19 \end{bmatrix} + \begin{bmatrix} -\frac{38}{75} \\ \frac{32}{75} \\ \frac{84}{25} \\ -\frac{448}{75} \\ -\frac{518}{75} \\ \frac{48}{5} \end{bmatrix} + \begin{bmatrix} \frac{35}{9} \\ -\frac{1288}{225} \\ -\frac{154}{25} \\ \frac{3136}{225} \\ \frac{4018}{225} \\ -\frac{119}{5} \end{bmatrix}.$$

Putting these ideas together, we present a tableau for a variable stepsize version of our main method. Note that in the U and V matrices, the factor ρ is the ratio of the stepsize in the current step to that in the previous step. The third and fourth row of the B matrix are found

from $\widehat{\beta}^T + \theta\widetilde{\beta}^T$ and $\overline{\beta}^T + (1-\theta)\widetilde{\beta}^T$ respectively, where we have chosen $\theta = -1$.

$$\left[\begin{array}{ccccc|cccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4}\rho & \frac{1}{32}\rho^2 & \frac{1}{32}\rho^3 \\ \frac{2}{5} & 0 & 0 & 0 & 0 & 1 & \frac{1}{10}\rho & \frac{1}{40}\rho^2 & \frac{1}{40}\rho^3 \\ \frac{27}{160} & \frac{75}{128} & 0 & 0 & 0 & 1 & -\frac{3}{640}\rho & -\frac{69}{1280}\rho^2 & -\frac{69}{1280}\rho^3 \\ \frac{69}{35} & -\frac{51}{28} & \frac{8}{7} & 0 & 0 & 1 & -\frac{41}{140}\rho & \frac{17}{280}\rho^2 & \frac{17}{280}\rho^3 \\ \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90}\rho & 0 & 0 \\ \hline \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90} & 0 & 1 & \frac{7}{90}\rho & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{376}{75} & \frac{842}{75} & -\frac{5696}{225} & -\frac{742}{25} & \frac{214}{5} & 0 & -\frac{182}{45}\rho & 0 & 0 \\ -\frac{496}{45} & -\frac{224}{25} & \frac{4928}{225} & \frac{6482}{225} & -38 & 0 & \frac{1636}{225}\rho & 0 & 0 \end{array}\right]. \quad (4.12)$$

4.3 Interpolation

A third order interpolator can be found in exactly the same manner as for other fourth order methods with five stages, i.e., solving equations (3.70)–(3.74) and (3.122). This leads to the following coefficients for the interpolator.

$$\begin{aligned} \tilde{b}_1(\xi) &= \tfrac{16}{45}\xi^2(5 - 6\xi + 2\xi^2), \\ \tilde{b}_2(\xi) &= \tfrac{2}{15}\xi^2(25 - 46\xi + 22\xi^2), \\ \tilde{b}_3(\xi) &= -\tfrac{16}{45}\xi^2(10 - 24\xi + 13\xi^2), \\ \tilde{b}_4(\xi) &= -\tfrac{7}{90}\xi^2(70 - 144\xi + 73\xi^2), \\ \tilde{b}_5(\xi) &= \tfrac{\xi^2}{2}(13 - 28\xi + 15\xi^2), \\ \tilde{b}_0(\xi) &= \tfrac{\xi}{90}(90 - 235\xi + 228\xi^2 - 76\xi^3). \end{aligned}$$

To verify this interpolator experimentally, we have solved the D1 problem using 100 equal sized steps. The D1 problem is part of the DETest test set. Details can be found in section

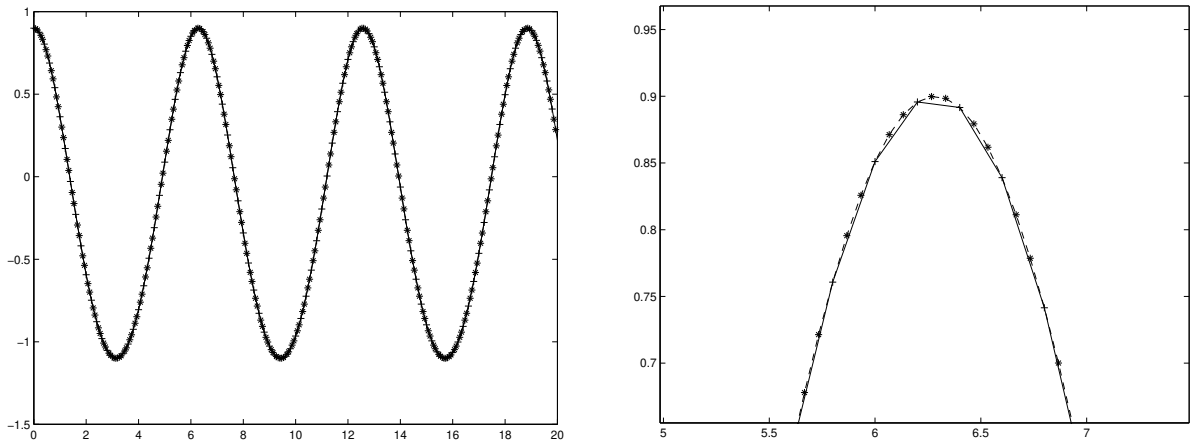


Figure 4.1: The $D1$ problem solved using method (4.1) with 100 equal sized steps. The solution points are represented with plus symbols (+). An interpolator has been used to estimate the solution $\frac{1}{3}$ and $\frac{2}{3}$ of the way through each step. These points are represented with asterisks ($*$). The figure on the right is an enlargement of one of the turning points, as this is where the solution is changing most rapidly. The dashed line is obtained from a cubic spline interpolation through the solution points.

A.1. We have used the interpolator to find the solution $\frac{1}{3}$ and $\frac{2}{3}$ of the way through each step. The results from this are presented in Figure 4.1. As we can see, the interpolator is giving very reliable results.

4.4 Error estimation

There are several possible ways of estimating the error in ARK methods. The technique outlined here is a two-step approximation to zero, as proposed by Butcher and Chan [14]. An approximation to the error is calculated at the end of a step, which is $O(h^5)$. The difference of this quantity over two steps gives an estimate of the local error which is $O(h^6)$. This will not give a good approximation to the local truncation error itself, but will give an asymptotically correct approximation.

Let $\widehat{d} = d_0 y_2^{[n]} + \sum_{i=1}^{5} d_i h F_i$, where $d = [d_1, d_2, ..., d_5]$, be our $O(h^5)$ approximation to the error in each step. To determine the values of $d_0, d_1, ..., d_5$ we return to the generating functions introduced in section 3.2. Let $\delta(t)$ be a mapping from the trees to the real numbers representing the error in a step. Then

$$\delta(t) = d_0 + d^T(\xi D)(t),$$

where $\xi(t)$ is a mapping from trees to the real numbers representing the internal stages. To obtain an approximation to the error that is $O(h^5)$ we need to ensure $\delta(t) = 0$ for all trees up to and including order 4. As with the order conditions, we find the number of conditions is reduced due to the stage order of the method. This leaves us with one free parameter. We will use this to normalise the results by requiring that $\delta(t_9) = \frac{1}{5}$. The conditions on d_0 and d^T are

$$d^T + d_0 = 0, \tag{4.13}$$

$$d^T c = 0, \tag{4.14}$$

$$d^T c^2 = 0, \tag{4.15}$$

$$d^T c^3 = 0, \tag{4.16}$$

$$d^T \left(Ac^2 - \frac{1}{\beta_5}(2Ac - c^2) \right) = 0, \tag{4.17}$$

$$d^T c^4 = \frac{1}{5}. \tag{4.18}$$

Solving these equations for the method given in (4.1), leads to the following coefficients

$$d = \left[-\frac{128}{15}, \frac{64}{5}, -\frac{128}{15}, \frac{532}{15}, -\frac{100}{3} \right], \qquad d_0 = \frac{32}{15}. \tag{4.19}$$

Assuming a constant stepsize over two steps, a higher order approximation to the error can be found by the difference in $\widehat{d}$ over these two steps. Keeping the stepsize constant over two steps is not an unreasonable restriction. An estimate of the error can then be found, helping to determine the stepsize for the next two steps.

An alternative approach is to alter condition (4.17) to incorporate the stepsize change into it so that the error can be estimated after each step. This condition now becomes

$$d^T \left(Ac^2 - \frac{1}{r\beta_5}(2Ac - c^2) \right) = 0,$$

where r is the ratio between the current and previous stepsizes. Solving this new system of equations leads to the same solution as before, with d_4 and d_5 replaced with

$$d_4 = \frac{28(385 - 556r)}{15(-85 + 76r)}, \qquad d_5 = \frac{300(-3 + 4r)}{-85 + 76r}.$$

4.5 Optimising these methods

This special method was discovered almost by chance. It is natural to ask if other methods with this special property exist, and if they do are we able to select an optimal method from among

them. In order to do this we need to decide on our definition of optimal. We have already seen that it is possible to ensure the errors in the fifth order trees are zero. We will define an optimal method to be one that minimises the norm of the vector of the error coefficients of the sixth order trees.

4.5.1 Fifth order error coefficients

To look for methods with this special property, we first need to ensure that the fifth order error coefficients are zero. There are 9 trees of order five, however due to the stage order of the methods many of the error coefficients are a scalar multiple of another. As stated in Section 3.2, the trees that are omitted are those that would be omitted if the $C(2)$ condition were assumed for a Runge–Kutta method. The independent error coefficients are therefore

$$\epsilon_9 = \frac{1}{5} - b^T c^4, \tag{4.20}$$

$$\begin{aligned}\epsilon_{11} &= \frac{1}{15} - b^T c(\frac{1}{2}c^2 - Ac)\eta(t_3) - b^T CAc^2 \\ &= \frac{1}{15} - b^T c(\frac{1}{2}c^2 - Ac)(\beta^T c^2 - 2) - b^T CAc^2,\end{aligned} \tag{4.21}$$

$$\epsilon_{14} = \frac{1}{20} - b^T Ac^3, \tag{4.22}$$

$$\begin{aligned}\epsilon_{16} &= \frac{1}{60} - \frac{1}{2}b^T A(\frac{1}{2}c^2 - Ac)\eta(t_4) - \frac{1}{4}b^T A^2 c^2 \\ &= \frac{1}{60} - \frac{1}{2}b^T A(\frac{1}{2}c^2 - Ac)(\tfrac{1}{4}\beta^T c^2 - 1) - \tfrac{1}{4}b^T A^2 c^2.\end{aligned} \tag{4.23}$$

If we assume the $D(1)$ condition, as we have whilst deriving these methods, then condition (4.21) can be ignored. To see why this is true we start by expanding equation (4.21), giving

$$\epsilon_{11} = \frac{1}{15} - (\tfrac{1}{2}b^T c^3 - b^T Ac + b^T A^2 c)(\beta^T c^2 - 2) - b^T Ac^2 + b^T A^2 c^2.$$

Using the standard order conditions along with conditions (3.106) and (3.107) this can be simplified to

$$\begin{aligned}\epsilon_{11} &= \frac{1}{15} - (\frac{1}{8} - \frac{1}{6} + \frac{1}{24} + \theta(b^T A^3 c - K))(\beta^T c^2 - 2) - \frac{1}{12} + \frac{2(\beta_5 - \phi)}{\beta_5}(b^T A^3 c - K) + 2K \\ &= -\frac{1}{60} - \theta(b^T A^3 c - K)(\beta^T c^2 - 2) + \frac{2(\beta_5 - \phi)}{\beta_5}(b^T A^3 c - K) + 2K.\end{aligned} \tag{4.24}$$

To simplify further we will make use of one of the conditions that is satisfied in order to have RK-stability, i.e.

$$\beta^T(\tfrac{1}{2}c^2 - Ac) = 0.$$

From this condition and condition (3.103) it is easy to see that

$$\beta^T c^2 - 2 = \frac{2(\beta_5 - \phi)}{\theta\beta_5}.$$

Substituting this into equation (4.24) shows that $\epsilon_{11} = -\frac{1}{60} + 2K$, which equals 0 if we choose $K = \frac{1}{120}$.

The equation $\epsilon_9 = 0$ can easily be satisfied by choosing parameter L to be $\frac{1}{5}$. The error in tree (t_{16}) is automatically 0, as we have chosen $K = \frac{1}{120}$ to ensure $\epsilon(t_{11}) = 0$. Once c_1, c_2, c_3 and c_4 have been chosen there is a unique choice of a_{43} that ensures $\epsilon_{14} = 0$ (and hence $\epsilon_{11} = 0$ if we have assumed the $D(1)$ condition).

4.5.2 Sixth order error coefficients

Once the fifth order conditions are satisfied we can concentrate on the sixth order conditions. Assuming the $D(1)$ condition, the independent error coefficients are

$$\begin{aligned}
\epsilon(t_{18}) &= b^T\xi(t_1)^5 - \frac{1}{6} \\
&= b^T c^5 - \frac{1}{6}, \\
\epsilon(t_{20}) &= b^T\xi(t_1)\xi(t_1)\xi(t_3) - \frac{1}{18} \\
&= b^T c^2\left((\frac{1}{2}c^2 - Ac)\eta(t_3) + Ac^2\right) - \frac{1}{18} \\
&= b^T\left((\frac{1}{2}c^2 - Ac)(\beta^T c^2 - 2) + Ac^2\right)c^2 - \frac{1}{18},
\end{aligned}$$

$$\epsilon(t_{23}) = b^T\xi(t_1)\xi(t_5) - \frac{1}{24}$$

$$= b^T c\left((\frac{1}{2}c^2 - Ac)\eta(t_5) + Ac^3\right) - \frac{1}{24}$$

$$= b^T c\left((\frac{1}{2}c^2 - Ac)(\beta^T c^3 - 3\beta^T c^2 - 9) + Ac^3\right) - \frac{1}{24},$$

$$\epsilon(t_{25}) = b^T\xi(t_1)\xi(t_7) - \frac{1}{72}$$

$$= b^T c\left((\frac{1}{2}c^2 - Ac)\eta(t_7) + A(\frac{1}{2}c^2 - Ac)\eta(t_3) + A^2c^2\right) - \frac{1}{72}$$

$$= b^T c\left((\frac{1}{2}c^2 - Ac)(\beta^T(\frac{1}{2}c^2 - Ac)(\beta^T c^2 - 2) + \beta^T Ac^2) + \right.$$

$$\left. A(\frac{1}{2}c^2 - Ac)(\beta^T c^2 - 2) + A^2c^2\right) - \frac{1}{72}.$$

The free parameters we have left are c_1, c_2, c_3 and ϕ. First c_1 and ϕ need to be chosen in such a way that an appropriate value of $\mu = \theta\beta_5$ is found from equation (3.105). After many numerical searches the values found which give usable values for μ are given in (3.108) – (3.111).

Next c_2 and c_3 need to be chosen to ensure ϵ_{18} is small. We have chosen to require $\epsilon_{18} = 0$. This is done by simultaneously solving

$$b^T c = \frac{1}{2}, \qquad b^T c^2 = \frac{1}{3}, \qquad b^T c^3 = \frac{1}{4},$$
$$b^T c^4 = \frac{1}{5}, \qquad b^T c^5 = \frac{1}{6},$$

for an unknown b^T.

If we use the choice of parameters given in (3.108), where $c_1 = \frac{1}{5}$, we find the following relationship between c_2 and c_3

$$c_3 = \frac{-7 + 10c_2}{-10 + 15c_2}.$$

Using the parameters in (3.109), where $c_1 = \frac{1}{4}$, gives the relationship

$$c_3 = \frac{-5 + 7c_2}{-7 + 10c_2}.$$

Finally, using the parameters in either (3.110) or (3.111), where $c_1 = \frac{1}{3}$, the relationship is

$$c_3 = \frac{-3 + 4c_2}{-4 + 5c_2}.$$

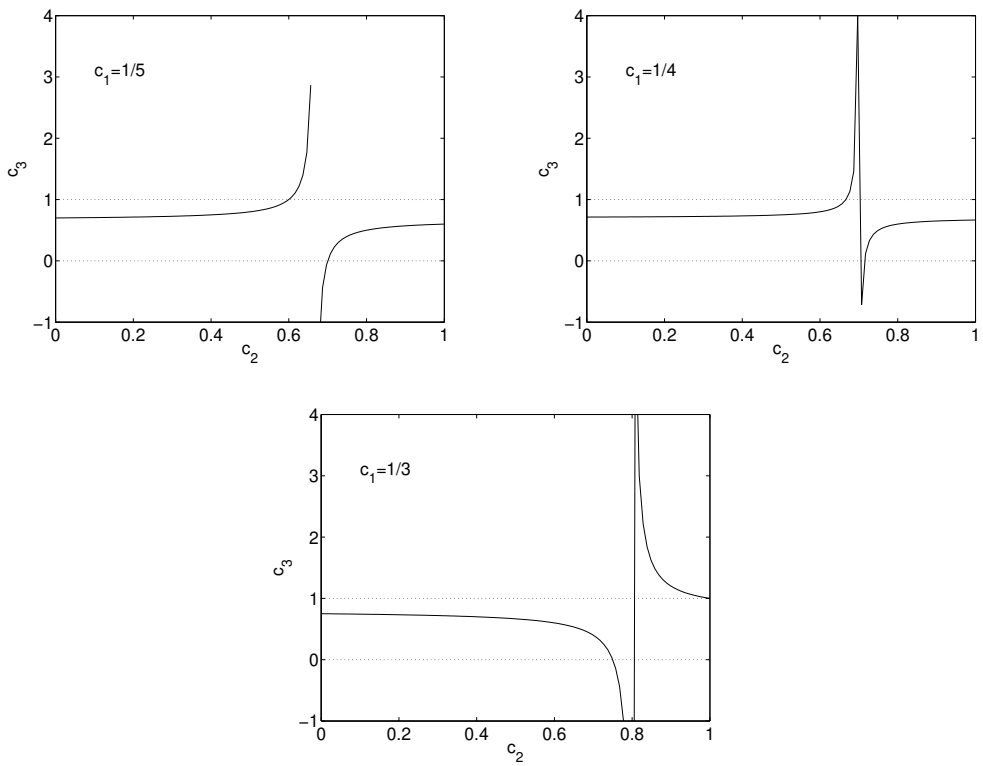


Figure 4.2: Optimising our special 'fifth' order method. Solving for the free parameters c_2 and c_3. Clockwise from top left: $c_1 = \frac{1}{5}$, $c_1 = \frac{1}{4}$ and $c_1 = \frac{1}{3}$.

In order to make the implementation as simple as possible we wish both c_2 and c_3 to lie between 0 and 1. The values of c_2 and c_3 are plotted in Figure 4.2 for each of the above values of c_1.

We are left with only one free parameter, c_2, to minimise the remaining three error coefficients. Due to the complex nature of these equations the only way to optimise them is by performing numerical searches.

A numerical search was performed for each of the four sets of parameters given in (3.108) – (3.111). For the first set of parameters the optimal choice of the free parameters was found to be

$$c = \left[\frac{1}{5},\ \frac{1}{20},\ \frac{26}{37},\ 1,\ 1\right], \qquad \mu = \frac{4}{3}, \qquad \phi = 2.$$

The method defined by this set of parameters is

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{5} & \frac{1}{50} \\
-\frac{15}{992} & 0 & 0 & 0 & 0 & 1 & \frac{323}{4960} & \frac{53}{12400} \\
-\frac{353443935}{3748322} & \frac{724094280}{1874161} & 0 & 0 & 0 & 1 & -\frac{1092110669}{3748322} & -\frac{795197}{3748322} \\
\frac{119973785}{250263} & -\frac{2531594960}{1299753} & \frac{783399298}{523800459} & 0 & 0 & 1 & \frac{530619764}{361491} & \frac{17767}{18538} \\
\frac{7625}{13392} & -\frac{32000}{82593} & \frac{69343957}{154162008} & \frac{299}{3344} & 0 & 1 & \frac{29}{104} & 0 \\
\hline
\frac{7625}{13392} & -\frac{32000}{82593} & \frac{69343957}{154162008} & \frac{299}{3344} & 0 & 1 & \frac{29}{104} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{92996}{243} & \frac{1217291984}{743337} & \frac{166800329}{27972945} & \frac{2093}{1045} & -\frac{158}{15} & 0 & -\frac{2197852}{1755} & 0
\end{array}\right].$$

The 2-norm of the vector of 20 sixth order coefficients is 1.22439 . The error coefficients for the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = \frac{69419}{1032300}, \qquad \epsilon_{23} = -\frac{249829}{324000}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

Although this method optimises the norm of the errors, the numerators and denominators of the coefficients are rather larger than we would prefer. Instead we propose a method with only a slightly larger norm, but whose coefficients require fewer digits in their representations. The free parameters for the method are

$$c = \left[\frac{1}{5},\ \frac{1}{3},\ \frac{11}{15},\ 1,\ 1\right], \qquad \mu = \frac{4}{3}, \qquad \phi = 2.$$

The defining matrices are

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{5} & \frac{1}{50} \\
\frac{25}{279} & 0 & 0 & 0 & 0 & 1 & \frac{68}{279} & \frac{7}{186} \\
\frac{38830}{279} & -\frac{2046}{25} & 0 & 0 & 0 & 1 & -\frac{394801}{6975} & -\frac{1331}{4650} \\
-\frac{660625}{961} & \frac{12627}{31} & \frac{450}{341} & 0 & 0 & 1 & \frac{2957689}{10571} & \frac{2393}{1922} \\
\frac{125}{768} & \frac{9}{32} & \frac{1125}{2816} & \frac{31}{384} & 0 & 1 & \frac{5}{66} & 0 \\
\hline
\frac{125}{768} & \frac{9}{32} & \frac{1125}{2816} & \frac{31}{384} & 0 & 1 & \frac{5}{66} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\frac{23365}{48} & -\frac{8261}{30} & \frac{2995}{528} & \frac{217}{120} & -\frac{158}{15} & 0 & -\frac{34378}{165} & 0
\end{array}\right].$$

The 2-norm of the sixth order coefficients is 1.28536, only slightly larger than for the previous method. The error coefficients for the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = \frac{893}{11160}, \qquad \epsilon_{23} = -\frac{5239}{6480}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

For the second set of parameters the optimal choice of the free parameters was found to be

$$c = \left[\frac{1}{4},\ \frac{19}{20},\ \frac{33}{50},\ 1,\ 1\right], \qquad \mu = 8, \qquad \phi = 4.$$

The defining matrices of the method are

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} \\
\frac{266}{125} & 0 & 0 & 0 & 0 & 1 & -\frac{589}{500} & -\frac{323}{4000} \\
\frac{11896929}{17500000} & \frac{39237}{532000} & 0 & 0 & 0 & 1 & -\frac{555621}{5937500} & -\frac{111111}{5000000} \\
\frac{245787}{11480} & \frac{38055}{30856} & -\frac{106250}{13079} & 0 & 0 & 1 & -\frac{14128}{1045} & -\frac{53}{80} \\
\frac{976}{2583} & \frac{2000}{11571} & \frac{781250}{2001087} & -\frac{5}{306} & 0 & 1 & \frac{283}{3762} & 0 \\
\hline
\frac{976}{2583} & \frac{2000}{11571} & \frac{781250}{2001087} & -\frac{5}{306} & 0 & 1 & \frac{283}{3762} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{24098}{12915} & -\frac{14510}{11571} & -\frac{2350000}{667029} & \frac{28}{153} & \frac{24}{5} & 0 & \frac{5204}{3135} & 0
\end{array}\right]. \tag{4.25}$$

The norm of the sixth order error coefficients is 0.102186. The error coefficients of the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = -\frac{49}{90000}, \qquad \epsilon_{23} = -\frac{3719}{57600}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

As with the first method, the coefficients for this method are rather unattractive. We propose a method with only a slightly larger norm, but with more pleasing coefficients. The free parameters for the method are

$$c = \left[\frac{1}{4},\ \frac{3}{4},\ \frac{1}{2},\ 1,\ 1\right], \qquad \mu = 8, \qquad \phi = 4.$$

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{32} \\
\frac{6}{5} & 0 & 0 & 0 & 0 & 1 & -\frac{9}{20} & -\frac{3}{160} \\
-\frac{31}{160} & \frac{25}{96} & 0 & 0 & 0 & 1 & \frac{13}{30} & -\frac{7}{320} \\
-\frac{69}{40} & \frac{39}{56} & \frac{6}{7} & 0 & 0 & 1 & \frac{41}{35} & -\frac{11}{560} \\
\frac{16}{45} & \frac{16}{45} & \frac{2}{15} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\
\hline
\frac{16}{45} & \frac{16}{45} & \frac{2}{15} & \frac{7}{90} & 0 & 1 & \frac{7}{90} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{602}{225} & -\frac{518}{225} & -\frac{16}{15} & -\frac{196}{225} & \frac{24}{5} & 0 & \frac{476}{225} & 0
\end{array}\right]$$

The norm of the vector of sixth order error coefficients is 0.10496. The error coefficients of the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = -\frac{29}{18000}, \qquad \epsilon_{23} = -\frac{1273}{19200}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

For the third set of parameters the optimal choice for the free parameters was found to be

$$c = \left[\frac{1}{3},\ \frac{8}{11},\ \frac{1}{4},\ 1,\ 1\right], \qquad \mu = 6, \qquad \phi = 3.$$

The method defined by these parameters is

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
\frac{78}{121} & 0 & 0 & 0 & 0 & 1 & \frac{10}{121} & \frac{6}{121} \\
-\frac{1257}{6656} & \frac{4235}{26624} & 0 & 0 & 0 & 1 & \frac{573}{2048} & -\frac{11}{512} \\
-\frac{3063}{2314} & \frac{26983}{32396} & \frac{1536}{623} & 0 & 0 & 1 & -\frac{347}{356} & -\frac{25}{89} \\
\frac{81}{520} & \frac{161051}{393120} & \frac{256}{945} & \frac{89}{1080} & 0 & 1 & \frac{13}{160} & 0 \\
\hline
\frac{81}{520} & \frac{161051}{393120} & \frac{256}{945} & \frac{89}{1080} & 0 & 1 & \frac{13}{160} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{99}{650} & \frac{7139}{23400} & -\frac{512}{225} & \frac{89}{450} & \frac{6}{5} & 0 & \frac{29}{40} & 0
\end{array}\right].$$

The norm of the vector of sixth order error coefficients is 0.120674. The error coefficients of the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = -\frac{1}{200}, \qquad \epsilon_{23} = -\frac{3013}{39600}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

Another method, with only a slightly larger norm, but more pleasing coefficients is defined by the free parameters

$$c = \left[\frac{1}{3},\ \frac{2}{3},\ \frac{1}{2},\ 1,\ 1\right], \qquad \mu = 6, \qquad \phi = 3.$$

The defining matrices are

$$\left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 1 & \frac{1}{6} & \frac{1}{18} \\ \frac{11}{32} & -\frac{5}{32} & 0 & 0 & 0 & 1 & \frac{5}{16} & \frac{11}{96} \\ \frac{71}{22} & \frac{17}{11} & -\frac{32}{11} & 0 & 0 & 1 & -\frac{19}{22} & -\frac{5}{33} \\ \frac{27}{40} & \frac{27}{40} & -\frac{8}{15} & \frac{11}{120} & 0 & 1 & \frac{11}{120} & 0 \\ \hline \frac{27}{40} & \frac{27}{40} & -\frac{8}{15} & \frac{11}{120} & 0 & 1 & \frac{11}{120} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{213}{50} & -\frac{21}{50} & \frac{64}{25} & \frac{11}{50} & \frac{6}{5} & 0 & \frac{7}{10} & 0 \end{array}\right].$$

The norm of the vector of sixth order error coefficients is 0.12145. The error coefficients of the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = \frac{1}{1800}, \qquad \epsilon_{23} = -\frac{829}{10800}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

For the final set of parameters the optimal choice of the free parameters was found to be

$$c = \left[\frac{1}{3},\ \frac{1}{20},\ \frac{56}{75},\ 1,\ 1\right], \qquad \mu = 12, \qquad \phi = 6.$$

The defining matrices for the method are

$$\left[\begin{array}{ccccc|ccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\ -\frac{51}{800} & 0 & 0 & 0 & 0 & 1 & \frac{91}{800} & \frac{9}{400} \\ \frac{3410806}{2390625} & -\frac{1451296}{860625} & 0 & 0 & 0 & 1 & \frac{1273538}{1265625} & -\frac{142394}{1265625} \\ -\frac{2513067}{696694} & \frac{1288200}{123607} & \frac{8015625}{6311228} & 0 & 0 & 1 & -\frac{131123}{18508} & \frac{154}{661} \\ \frac{1647}{4216} & \frac{32000}{202521} & \frac{10546875}{27574624} & \frac{661}{8664} & 0 & 1 & -\frac{5}{672} & 0 \\ \hline \frac{1647}{4216} & \frac{32000}{202521} & \frac{10546875}{27574624} & \frac{661}{8664} & 0 & 1 & -\frac{5}{672} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{48519}{2635} & \frac{4398160}{67507} & -\frac{7340625}{3446828} & -\frac{5288}{1805} & \frac{42}{5} & 0 & -\frac{7011}{140} & 0 \end{array}\right].$$

The norm of the vector of sixth order error coefficients is 0.140703. The error coefficients of the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = -\frac{17}{13500}, \qquad \epsilon_{23} = -\frac{1921}{21600}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

Another method, with only a slightly larger norm, but more pleasing coefficients is defined by the free parameters

$$c = \left[\frac{1}{3},\ \frac{1}{5},\ \frac{11}{15},\ 1,\ 1\right], \qquad \mu = 12, \qquad \phi = 6.$$

The defining matrices of the method are

$$\left[\begin{array}{ccccc|ccc}
0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{3} & \frac{1}{18} \\
-\frac{3}{25} & 0 & 0 & 0 & 0 & 1 & \frac{8}{25} & \frac{3}{50} \\
\frac{121}{75} & -\frac{22}{27} & 0 & 0 & 0 & 1 & -\frac{44}{675} & -\frac{143}{1350} \\
-\frac{168}{31} & \frac{175}{31} & \frac{450}{341} & 0 & 0 & 1 & -\frac{6}{11} & \frac{13}{62} \\
\frac{9}{32} & \frac{125}{768} & \frac{1125}{2816} & \frac{31}{384} & 0 & 1 & \frac{5}{66} & 0 \\
\hline
\frac{9}{32} & \frac{125}{768} & \frac{1125}{2816} & \frac{31}{384} & 0 & 1 & \frac{5}{66} & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-\frac{162}{5} & \frac{305}{8} & -\frac{135}{88} & -\frac{31}{10} & \frac{42}{5} & 0 & -\frac{522}{55} & 0
\end{array}\right].$$

The norm of the vector of sixth order error coefficients is 0.141363. The error coefficients for the four distinct trees we are interested in are

$$\epsilon_{18} = 0, \qquad \epsilon_{20} = -\frac{7}{5400}, \qquad \epsilon_{23} = -\frac{193}{2160}, \qquad \epsilon_{25} = -\frac{1}{360}.$$

It is easy to see that the method given in (4.25) is the overall optimal method among those examined in detail.

CHAPTER 5

Stiff ARK methods

"Obvious" is the most dangerous word in mathematics.

Eric Temple Bell

The methods that have been considered so far have been explicit methods for non-stiff differential equations. It is natural to ask how well these methods extend to implicit methods for stiff differential equations. As the stage order of ARK methods is restricted to two we will only consider methods of low order, with $s = p$, as it is likely they will suffer from some order reduction. Order reduction is when the stiffness of a problem causes the method to decrease to the order of the stages, rather than the expected order of the method.

5.1 Introduction

We will also only consider diagonally implicit methods to ensure computational costs are kept as low as possible. The A matrix for a diagonally implicit method is lower triangular, with a single eigenvalue, λ. The general form of a diagonally implicit ARK method is

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{c|ccc} A & e & c-Ae & \frac{1}{2}c^2-Ac \\ \hline b^T & 1 & b_0 & 0 \\ e_s^T & 0 & 0 & 0 \\ \beta^T & 0 & \beta_0 & 0 \end{array}\right] = \left[\begin{array}{c|ccc} \widehat{A}+\lambda I & e & c-Ae & \frac{1}{2}c^2-Ac \\ \hline \widehat{b}^T+\lambda e_s^T & 1 & b_0 & 0 \\ e_s^T & 0 & 0 & 0 \\ \beta^T & 0 & \beta_0 & 0 \end{array}\right], \quad (5.1)$$

where $b^T = e_s^T A$, $e_1^T V = e_s^T U$ and where $\widehat{A}$ is strictly lower triangular. We will always assume $c_s = 1$. Recall that $e^T = [1, 1, ..., 1]$ and $e_s^T = [0, 0, \ldots, 0, 1]$.

The property of RK stability implies that the stability matrix for the method has only a single non-zero eigenvalue. To simplify the analysis, we will reformulate the method so that, of the three quantities $y_1^{[n]}$, $y_2^{[n]}$ and $y_3^{[n]}$ passed from step to step, only $y_1^{[n]}$ and $y_3^{[n]}$ appear in the formulation. This is a straightforward change in the interpretation of the method because, for the differential equation $y'(x) = f(x, y(x))$, $y_2^{[n]} = hf(x_n, y_1^{[n]})$. This means that we can compute a quantity equal to $y_2^{[n-1]}$ in step number n by artificially inserting an additional stage into the method. In a practical implementation of the method, this would never be done this way but is introduced here purely to aid the analysis.

In this alternative formulation of the method, the defining matrices become

$$\left[\begin{array}{c|c} \widetilde{A} & \widetilde{U} \\ \hline \widetilde{B} & \widetilde{V} \end{array}\right] = \left[\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ c - Ae & A & e & \frac{1}{2}c^2 - Ac \\ \hline b_0 & b^T & 1 & 0 \\ \beta_0 & \beta^T & 0 & 0 \end{array}\right] \tag{5.2}$$

and the stability matrix for this modified method is given by

$$\widetilde{M}(z) = \left[\begin{array}{cc} 1 & 0 \\ 0 & 0 \end{array}\right] + z \left[\begin{array}{cc} b_0 & b^T \\ \beta_0 & \beta^T \end{array}\right] \left[\begin{array}{cc} 1 & 0 \\ -z(c - Ae) & I - zA \end{array}\right]^{-1} \left[\begin{array}{cc} 1 & 0 \\ e & \frac{1}{2}c^2 - Ac \end{array}\right].$$

To simplify this expression, we use the matrices T and T^{-1} given by

$$T = \left[\begin{array}{cc} 1 & 0 \\ e & I \end{array}\right], \qquad T^{-1} = \left[\begin{array}{cc} 1 & 0 \\ -e & I \end{array}\right]$$

to transform the various factors in the last term as follows

$$\left[\begin{array}{cc} b_0 & b^T \\ \beta_0 & \beta^T \end{array}\right] T = \left[\begin{array}{cc} 1 & b^T \\ 0 & \beta^T \end{array}\right], \qquad T^{-1} \left[\begin{array}{cc} 1 & 0 \\ e & \frac{1}{2}c^2 - Ac \end{array}\right] = \left[\begin{array}{cc} 1 & 0 \\ 0 & \frac{1}{2}c^2 - Ac \end{array}\right]$$

and

$$T^{-1} \left[\begin{array}{cc} 1 & 0 \\ -z(c - Ae) & I - zA \end{array}\right]^{-1} T = T^{-1} \left[\begin{array}{cc} 1 & 0 \\ z(I - zA)^{-1}(c - Ae) & (I - zA)^{-1} \end{array}\right] T,$$

$$= \left[\begin{array}{cc} 1 & 0 \\ z(I - zA)^{-1}c & (I - zA)^{-1} \end{array}\right],$$

where we have used the consistency conditions $b_0 + b^T e = 1$ and $\beta_0 + \beta^T e = 0$. The stability

matrix can now be written in the form

$$\widetilde{M}(z) = \widetilde{V} + z\widetilde{B}TT^{-1}\left[\begin{array}{cc} 1 & 0 \\ -z(c-Ae) & I-zA \end{array}\right]^{-1} TT^{-1}\widetilde{U},$$

$$= \left[\begin{array}{cc} 1+z+z^2b^T(I-zA)^{-1}c & zb^T(I-zA)^{-1}(\frac{1}{2}c^2-Ac) \\ z^2\beta^T(I-zA)^{-1}c & z\beta^T(I-zA)^{-1}(\frac{1}{2}c^2-Ac) \end{array}\right].$$

To further simplify this we will make use of $b^T = e_s^TA$ and $e_s^Tc = 1$. The $(1,1)$ element can be written as

$$\begin{aligned} 1+z+z^2e_s^TA(I+zA+z^2A^2+z^3A^3+\cdots)c &= 1+ze_s^Tc+ze_s^T(zA+z^2A^2+z^3A^3+\cdots)c \\ &= 1+ze_s^T(I-zA)^{-1}c. \end{aligned}$$

A similar simplification can be made for the $(1,2)$ element. The stability matrix can now be written as

$$\widetilde{M}(z) = \left[\begin{array}{cc} 1+ze_s^T(I-zA)^{-1}c & e_s^T(I-zA)^{-1}(\frac{1}{2}c^2-Ac) \\ z^2\beta^T(I-zA)^{-1}c & z\beta^T(I-zA)^{-1}(\frac{1}{2}c^2-Ac) \end{array}\right].$$

It will be convenient to make the substitution $z = \widehat{z}/(1+\lambda\widehat{z})$ to obtain the matrix

$$\widehat{M}(\widehat{z}) = \left[\begin{array}{cc} \widehat{M}_{11}(\widehat{z}) & \widehat{M}_{12}(\widehat{z}) \\ \widehat{M}_{21}(\widehat{z}) & \widehat{M}_{22}(\widehat{z}) \end{array}\right] = \left[\begin{array}{cc} 1+\widehat{z}e_s^T(I-\widehat{z}\widehat{A})^{-1}c & \widehat{e}_s^T(I-\widehat{z}\widehat{A})^{-1}(\frac{1}{2}c^2-\widehat{A}c-\lambda c) \\ \widehat{z}^2\beta^T(I-\widehat{z}\widehat{A})^{-1}c & \widehat{z}\beta^T(I-\widehat{z}\widehat{A})^{-1}(\frac{1}{2}c^2-\widehat{A}c-\lambda c) \end{array}\right].$$

Just as the requirements of RK stability will be satisfied if $M(z)$ has a single non-zero eigenvalue equal to $\exp(z)+O(z^{s+1})$, we have the alternative criterion that $\widehat{M}(\widehat{z})$ has only a single non-zero eigenvalue equal to $\exp(\widehat{z}/(1+\lambda\widehat{z}))+O(\widehat{z}^{s+1})$.

Write

$$E^{\lambda}(\widehat{z}) = \exp\left(\frac{\widehat{z}}{1+\lambda\widehat{z}}\right) = 1+\alpha_1\widehat{z}+\alpha_2\widehat{z}^2+\cdots,$$

and write the truncated series as

$$E_s^{\lambda}(\widehat{z}) = 1+\alpha_1\widehat{z}+\alpha_2\widehat{z}^2+\cdots+\alpha_s\widehat{z}^s.$$

A characteristic property of an ARK method is that the third output approximation is accurate only to within $O(h^3)$ and that the coefficients of the method are chosen so that errors of this magnitude in $y_3^{[n-1]}$ do not affect the order s accuracy of $y_1^{[n]}$ computed in step number n. Consider the special case of the differential equation $y'(x) = qy(x)$ and suppose that the input to step number n consists of the quantities

$$y_1^{[n-1]} = 1, \qquad y_2^{[n-1]} = hq, \qquad y_3^{[n-1]} = h^2q^2+h^3q^3\epsilon.$$

The quantity computed as the value of $y_1^{[n]}$ is $e_s^T Y$, where the stage vector Y is given by

$$Y = e + hqAY + hq(c - Ae) + (h^2q^2 + \epsilon h^3 q^3)(\tfrac{1}{2}c^2 - Ac).$$

Solve for Y and evaluate the contribution to $e_s^T Y$ from the term involving the ϵ factor. This contribution is

$$\epsilon h^3 q^3 e_s^T (I - hqA)^{-1}(\tfrac{1}{2}c^2 - Ac)$$

and must be $O(h^{s+1})$ for the order not to be disturbed by this perturbation. Write $z = hq$ and divide by z^3. We see that

$$e_s^T (I - zA)^{-1}(\tfrac{1}{2}c^2 - Ac) = O(z^{s-2}).$$

This means that the $(1,2)$ element of $M(z)$ is $O(z^{s-2})$ which is equivalent to stating that $\widehat{M}_{12}(\widehat{z}) = O(\widehat{z}^{s-2})$. Because $\widehat{A}^s = 0$, $\widehat{M}_{12}(\widehat{z})$ consists of *exactly* two terms and can be written in the form

$$\widehat{M}_{12}(\widehat{z}) = \widehat{z}^{s-2} e_s^T \widehat{A}^{s-2}(\tfrac{1}{2}c^2 - \widehat{A}c - \lambda c) + \widehat{z}^{s-1} e_s^T \widehat{A}^{s-1}(\tfrac{1}{2}c^2 - \widehat{A}c - \lambda c).$$

Now consider the $(2,1)$ element of the transformed stability matrix. As $\beta^T c = 1$ this equals

$$\widehat{M}_{21}(\widehat{z}) = \widehat{z}^2 + O(\widehat{z}^3).$$

To ensure RK stability, $\widehat{M}(\widehat{z})$ must have a zero eigenvalue and is therefore singular. We can write

$$\widehat{M}_{11}(\widehat{z})\widehat{M}_{22}(\widehat{z}) = \widehat{M}_{12}(\widehat{z})\widehat{M}_{21}(\widehat{z}). \tag{5.3}$$

Since

$$\widehat{M}_{11} = 1 + \widehat{z} + O(\widehat{z}^2),$$

it follows that $\widehat{M}_{22}(\widehat{z})$ consists of a single term. This term is equal to

$$\widehat{M}_{22}(\widehat{z}) = \beta^T \widehat{A}^{s-1}(\tfrac{1}{2}c^2 - \widehat{A} - \lambda c)\widehat{z}^s$$

and furthermore, the coefficient of $\widehat{z}^s$ in $\widehat{M}_{22}(\widehat{z})$ is identical to the coefficient of $\widehat{z}^{s-2}$ in $\widehat{M}_{12}(\widehat{z})$. Write P as the product of the elements in the first subdiagonal of $\widehat{A}$. This is the only non-zero element in $\widehat{A}^{s-1}$ and is in the $(s,1)$ position. We can now write the second column of $\widehat{M}(\widehat{z})$ in full. We have

$$\widehat{M}_{12}(\widehat{z}) = \widehat{z}^{s-2}\beta^T \widehat{A}^{s-1}(\tfrac{1}{2}c^2 - \widehat{A}c - \lambda c) + \widehat{z}^{s-1} e_s^T \widehat{A}^{s-1}(\tfrac{1}{2}c^2 - \widehat{A}c - \lambda c),$$

$$= (\beta_s + \widehat{z})Pc_1(\tfrac{1}{2}c_1 - \lambda)\widehat{z}^{s-2},$$

and

$$\widehat{M}_{22}(\widehat{z}) = \beta_s P c_1 (\tfrac{1}{2} c_1 - \lambda) \widehat{z}^s.$$

We will always assume that $\beta_s P c_1 (\frac{1}{2} c_1 - \lambda) \neq 0$ so that the method will be a genuine multivalue method.

We can now evaluate the coefficient of $\widehat{z}^s$ in $\widehat{M}_{11}(\widehat{z})$ in two different ways. As the sum of the eigenvalues of a matrix is equal to the trace, we have

$$\widehat{M}_{11} + \widehat{M}_{22} = E_s^{\lambda}(\widehat{z}).$$

The coefficient of $\widehat{z}^s$ in $\widehat{M}_{11}$ is therefore $\alpha_s - \beta_s P c_1 (\frac{1}{2} c_1 - \lambda)$, where α_s is the coefficient of $\widehat{z}^s$ in $E^{\lambda}(\widehat{z})$. It is also equal to $P c_1$ by evaluation of $e_s^T \widehat{A}^{s-1} c$.

We can now conclude that

$$P c_1 = \frac{\alpha_s}{1 + \beta_s (\frac{1}{2} c_1 - \lambda)}$$

and because of equation (5.3), we also conclude that $\widehat{M}_{11}(\widehat{z})$ has a factor $\widehat{z} + \beta_s$.

We can now summarise the main conclusions of this section.

Theorem 5.1 *For any s stage order s diagonally implicit ARK method such that* $\beta_s P c_1 (\frac{1}{2} c_1 - \lambda) \neq 0$,

$$c_1 = 2\lambda - \frac{2 E_s^{\lambda}(-\beta_s)}{\beta_s E_{s-1}^{\lambda}(-\beta_s)}. \tag{5.4}$$

Furthermore,

$$\beta^T (I + \beta_s A - \beta_s \lambda I) = \beta_s e_s^T, \tag{5.5}$$

and

$$(1 + \tfrac{1}{2} c_1 \beta_s - \lambda \beta_s)(b^T A^{s-2} c + \alpha_s - \tfrac{1}{s!}) = \alpha_s. \tag{5.6}$$

Proof: To prove equation (5.4), evaluate $\widehat{M}_{11}(\widehat{z})$ in the form

$$\widehat{M}_{11}(\widehat{z}) = 1 + \alpha_1 \widehat{z} + \alpha_2 \widehat{z}^2 + \cdots + \alpha_{s-1} \widehat{z}^{s-1} + \frac{\alpha_s}{1 + \beta_s (\frac{1}{2} c_1 - \lambda)} \widehat{z}^s = \frac{E_s^{\lambda}(\widehat{z}) + \beta_s (\frac{1}{2} c_1 - \lambda) E_{s-1}^{\lambda}(\widehat{z})}{1 + \beta_s (\frac{1}{2} c_1 - \lambda)}$$

Because $\widehat{z} + \beta_s$ is a factor of this polynomial,

$$E_s^{\lambda}(-\beta_s) + \beta_s (\tfrac{1}{2} c_1 - \lambda) E_{s-1}^{\lambda}(-\beta_s) = 0$$

and equation (5.4) follows.

To prove equation (5.5), define $v^T = \beta^T(I + \beta_s\widehat{A}) - \beta_s e_s^T = \beta^T(I + \beta_s A - \beta_s\lambda I) - \beta_s e_s^T$ and verify that $v^T\widehat{A}^{i-1}c(c - 2\lambda) = 0$ for all $i = 1, 2, \ldots, s$, using the known values of $\widehat{M}_{12}(\widehat{z})$ and $\widehat{M}_{22}(\widehat{z})$. Hence, $v^T = 0$.

To prove equation (5.6) we need to look at the trace of $\widehat{M}$. This should be equal to $E_s^\lambda(\widehat{z})$. Equating the coefficients of $\widehat{z}^s$ gives

$$\widehat{b}^T\widehat{A}^{s-2}c + \beta^T\widehat{A}^{s-1}(\tfrac{1}{2}c^2 - \widehat{A}c - \lambda c) = \alpha_s.$$

As $\widehat{A}$ is strictly lower triangular, this can be rewritten as

$$\widehat{b}^T\widehat{A}^{s-2}c + \tfrac{1}{2}\beta_s c_1\widehat{b}^T\widehat{A}^{s-2}c - \lambda\beta_s\widehat{b}^T\widehat{A}^{s-2}c = \alpha_s.$$

Rearranging and rewriting in terms of b^T and A gives equation (5.6). ■

5.2 Order 3 stiff ARK methods

We will investigate the possibility of ARK methods with 3 stages and order 3. The tableau defining the methods we are seeking is

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|ccc} \lambda & 0 & 0 & & & \\ a_{21} & \lambda & 0 & e & c - Ae & \frac{1}{2}c^2 - Ac \\ b_1 & b_2 & \lambda & & & \\ \hline b_1 & b_2 & \lambda & 1 & b_0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 & 0 & \beta_0 & 0 \end{array}\right]. \tag{5.7}$$

The stability function for an s-stage diagonally implicit Runge–Kutta method with order $p = s$ is

$$R(z) = \frac{N(z)}{(1 - \lambda z)^s} = \exp z - \sigma(\lambda)z^{s+1} + O(z^{s+2}), \tag{5.8}$$

where $\sigma(\lambda)$ is the error constant.

Since $N(z)$ has degree 3, we can use equation (5.8) to evaluate $N(z)$ and σ. These are

$$N(z) = 1 + (1 - 3\lambda)z + (\tfrac{1}{2} - 3\lambda + 3\lambda^2)z^2 + (\tfrac{1}{6} - \tfrac{3}{2}\lambda + 3\lambda^2 - \lambda^3)z^3, \tag{5.9}$$

$$\sigma = \tfrac{1}{24} - \tfrac{1}{2}\lambda + \tfrac{3}{2}\lambda^2 - \lambda^3. \tag{5.10}$$

We are interested only in A-stable methods. Because, for $\lambda > 0$, $R(z)$ is analytic in the left half-plane, it is necessary only to require that $|R(z)| \leq 1$ for $|Re(z)| \leq 0$. By the maximum

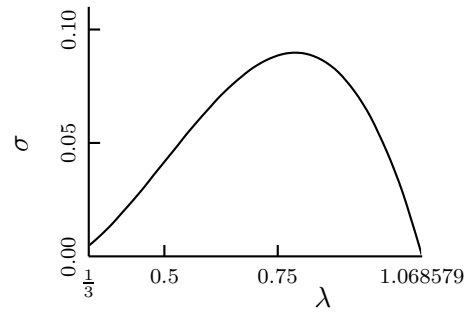


Figure 5.1: Error constant for λ in A-stability interval

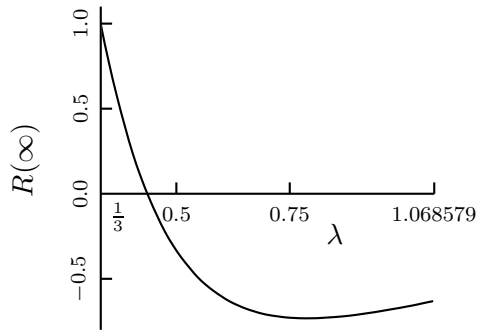


Figure 5.2: Values of $R(\infty)$

modulus principle, this is equivalent to

$$|(1-\lambda z)|^6 - |N(z)|^2 \geq 0,$$

for $z = iy$. This is the so-called E-polynomial and in this case becomes

$$E(y) = y^4(\tfrac{1}{12} - \lambda + 3\lambda - 2\lambda^3) + y^6(-\tfrac{1}{36} + \tfrac{1}{2}\lambda - \tfrac{13}{4}\lambda^2 + \tfrac{28}{3}\lambda^3 - 12\lambda^4 + 6\lambda^5)$$

which is non-negative for all real y as long as the coefficients of y^4 and y^6 are non-negative. It is found that necessary and sufficient conditions for this are

$$\lambda \in [\tfrac{1}{3}, \lambda^*] \quad \text{where} \quad \lambda^* \approx 1.068579.$$

As a guide to the selection of suitable values of λ in this interval, the error constant σ is plotted in Figure 5.1 and the value of $R(\infty)$ is shown in Figure 5.2.

Order conditions

The conditions to ensure the first output approximation is of order three and the third output approximation is of order two are:

$$b_0 + b^T e = 1, \tag{5.11}$$

$$b^T c = \frac{1}{2}, \tag{5.12}$$

$$b^T c^2 = \frac{1}{3}, \tag{5.13}$$

$$\beta^T e + \beta_0 = 0, \tag{5.14}$$

$$\beta^T c = 1. \tag{5.15}$$

From Theorem 5.1, the conditions to ensure the method has the correct stability function are

$$\beta^T (I + \beta_3 A - \lambda \beta_3 I) = \beta_3 e_3^T, \tag{5.16}$$

$$(1 + \tfrac{1}{2} c_1 \beta_3 - \lambda \beta_3)(b^T A c + \lambda^2 - \lambda) = \tfrac{1}{6} + \lambda^2 - \lambda, \tag{5.17}$$

$$c_1 = 2\lambda - \frac{2E_3^\lambda(-\beta_3)}{\beta_3 E_2^\lambda(-\beta_3)}. \tag{5.18}$$

Derivation of methods

The derivation of these methods is very simple. Once the free parameters, λ, β_3 and c_2, have been chosen the method can be uniquely determined from conditions (5.11) to (5.18).

First c_1 can be determined from equation (5.18). Then b_1 and b_2 can be determined from equations (5.12) and (5.13), giving

$$b_1 = \frac{\frac{1}{3} - \lambda + (\lambda - \frac{1}{2}) c_2}{c_1^2 - c_1 c_2},$$

$$b_2 = \frac{\frac{1}{3} - \lambda + (\lambda - \frac{1}{2}) c_1}{c_2^2 - c_1 c_2}.$$

Next, b_0 can be found from equation (5.11). The only remaining term of the A matrix, a_{21} can be found by solving equation (5.17), then the β^T vector can be found from equation (5.16). Finally, β_0 can be found from equation (5.14).

Some example methods

Two example methods are given here. This first method has been chosen to minimise the error coefficients. The c values have been specially chosen to give zero error for the bushy tree. We have chosen $\lambda = \frac{1}{3}$ as this is a simple fraction and comes close to minimising $\sigma(\lambda)$.

$$c = \begin{bmatrix} 1 \\ \frac{1}{2} \\ 1 \end{bmatrix}, \qquad \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|ccc} \frac{1}{3} & 0 & 0 & 1 & \frac{2}{3} & \frac{1}{6} \\ -\frac{1}{16} & \frac{1}{3} & 0 & 1 & \frac{11}{48} & \frac{1}{48} \\ -\frac{1}{6} & \frac{2}{3} & \frac{1}{3} & 1 & \frac{1}{6} & 0 \\ \hline -\frac{1}{6} & \frac{2}{3} & \frac{1}{3} & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & -\frac{8}{3} & 2 & 0 & \frac{1}{3} & 0 \end{array}\right]. \tag{5.19}$$

In this next method λ has been chosen to obtain close to L-stability. To obtain L-stability we would require $\lambda = 0.435867$. To find pleasing coefficient matrices for the method we have chosen $\lambda = \frac{2}{5}$.

$$c = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{2} \\ 1 \end{bmatrix}, \qquad \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|ccc} \frac{2}{5} & 0 & 0 & 1 & \frac{4}{15} & -\frac{2}{45} \\ -\frac{11}{144} & \frac{2}{5} & 0 & 1 & \frac{127}{720} & -\frac{13}{540} \\ -\frac{21}{20} & \frac{8}{5} & \frac{2}{5} & 1 & \frac{1}{20} & 0 \\ \hline -\frac{21}{20} & \frac{8}{5} & \frac{2}{5} & 1 & \frac{1}{20} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{39}{20} & -\frac{18}{5} & \frac{3}{2} & 0 & \frac{3}{20} & 0 \end{array}\right]. \tag{5.20}$$

5.3 Order 4 stiff ARK methods

In this section we will extend the analysis to stiff ARK methods with order four. As for the order three case, we will consider only methods for which A has the diagonally implicit structure with constant λ on the diagonal. We will consider only fourth order methods with exactly four stages. Our first consideration is the choice of λ, where we will wish to find a balance between stability and accuracy. The polynomial $N(z)$ appearing in equation (5.8) and the error constant σ are now given by

$$N(z) = 1+(1-4\lambda)z+(\tfrac{1}{2}-4\lambda+6\lambda^2)z^2+(\tfrac{1}{6}-2\lambda+6\lambda^2-4\lambda^3)z^3+(\tfrac{1}{24}-\tfrac{2}{3}\lambda+3\lambda^2-4\lambda^3+\lambda^4)z^4,$$

$$\sigma = \tfrac{1}{120}-\tfrac{1}{6}\lambda+\lambda^2-2\lambda^3+\lambda^4.$$

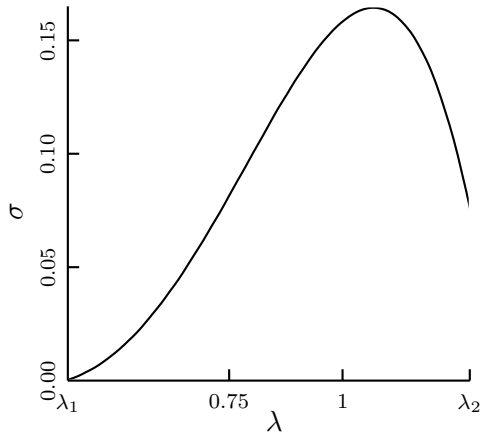


Figure 5.3: Error constant for λ in A-stability interval, where $\lambda_1 = 0.394338$ and $\lambda_2 = 1.28058$.

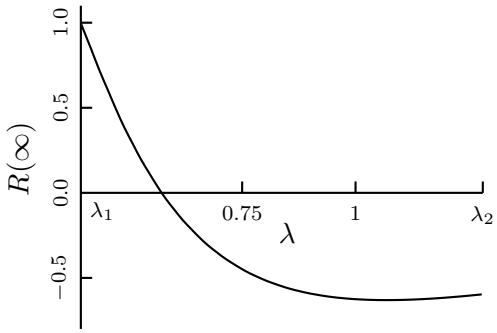


Figure 5.4: Values of $R(\infty)$ in A-stability interval, where $\lambda_1 = 0.394338$ and $\lambda_2 = 1.28058$.

As for third order methods, we are interested only in A-stable methods and we analyse this using the E-polynomial, which in this case is found to be

$$E(y) = (\tfrac{1}{72} - \tfrac{1}{3}\lambda + \tfrac{17}{6}\lambda^2 - \tfrac{32}{3}\lambda^3 + 17\lambda^4 - 8\lambda^5)y^6 + (-\tfrac{1}{576} + \tfrac{1}{18}\lambda - \tfrac{25}{36}\lambda^2 + \tfrac{13}{3}\lambda^3 - \tfrac{173}{12}\lambda^4 + \tfrac{76}{3}\lambda^5 - 22\lambda^6 + 8\lambda^7)y^8.$$

Using the maximum modulus principle, we see that A-stability is equivalent to $E(y) \geq 0$ for all real y and this is found to be the case if and only if λ lies in an interval $[\lambda_1, \lambda_2]$, where

$$\lambda_1 \approx 0.394338, \quad \lambda_2 \approx 1.28058.$$

Values of σ and $R(\infty)$ are shown for $\lambda \in [\lambda_1, \lambda_2]$ in Figure 5.3 and Figure 5.4, respectively.

Order conditions

The conditions for the first output approximation to be of order 4 and the third output approximation to be of order 2 are:

$$b_0 + b^T e = 1, \tag{5.21}$$

$$b^T c = \frac{1}{2}, \tag{5.22}$$

$$b^T c^2 = \frac{1}{3}, \tag{5.23}$$

$$b^T c^3 = \frac{1}{4}, \tag{5.24}$$

$$b^T A c = \frac{1}{6}, \tag{5.25}$$

$$b^T A c^2 = \frac{1}{12}, \tag{5.26}$$

$$\beta^T e + \beta_0 = 0, \tag{5.27}$$

$$\beta^T c = 1. \tag{5.28}$$

From Theorem 5.1, the conditions for stability are:

$$\beta^T (I + \beta_4 A - \lambda \beta_4 I) = \beta_4 e_4^T, \tag{5.29}$$

$$(1 + \tfrac{1}{2} c_1 \beta_4 - \lambda \beta_4)(b^T A^2 c - \tfrac{\lambda}{2} + \tfrac{3}{2}\lambda^2 - \lambda^3) = \tfrac{1}{24} - \tfrac{\lambda}{2} + \tfrac{3}{2}\lambda^2 - \lambda^3, \tag{5.30}$$

$$c_1 = 2\lambda - \frac{2E_4^\lambda(-\beta_4)}{\beta_4 E_3^\lambda(-\beta_4)}. \tag{5.31}$$

Derivation of methods

The derivation of fourth order methods is also reasonably simple. The only difficulty lies in choosing the parameters λ and β_4 such that c_1 lies in the interval $[0, 1]$ and the method is A-stable.

First c_1 can be found from equation (5.31). Then b_1, b_2 and b_3 can be found from equations

(5.22) – (5.24), giving

$$b_1 = \frac{\frac{1}{4} - \lambda + (\lambda - \frac{1}{3})c_2 + (\lambda - \frac{1}{3})c_3 + (\frac{1}{2} - \lambda)c_2c_3}{c_1(c_1 - c_2)(c_1 - c_3)},$$

$$b_2 = \frac{\frac{1}{4} - \lambda + (\lambda - \frac{1}{3})c_1 + (\lambda - \frac{1}{3})c_3 + (\frac{1}{2} - \lambda)c_1c_3}{c_2(c_2 - c_1)(c_2 - c_3)},$$

$$b_3 = \frac{\frac{1}{4} - \lambda + (\lambda - \frac{1}{3})c_1 + (\lambda - \frac{1}{3})c_2 + (\frac{1}{2} - \lambda)c_1c_2}{c_3(c_3 - c_1)(c_3 - c_2)}.$$

Next, b_0 can be found from equation (5.21). To find the element a_{32} we solve a liner combination of equations (5.25) and (5.26). That is, we solve

$$b^T Ac^2 - c_1 b^T Ac = \frac{1}{6} - \frac{c_1}{6}$$

for a_{32}, giving

$$a_{32} = \frac{192b_3\lambda + 25(24\lambda^2 + 24b_2c_2(3c_2 - 1) - 1)}{300b_3c_2(1 - 3c_2)}.$$

Element a_{21} can be found by solving equation (5.30) and a_{31} can then be found from equation (5.25), giving

$$a_{31} = \frac{1}{10b_3}(5 - 30\lambda^2 - 60b_2c_2\lambda - 32b_3\lambda - 20b_1\lambda - 30a_{32}b_3c_2 - 10a_{21}b_2).$$

The vector β^T can now be found from equation (5.29) and β_0 can be found from equation (5.27).

Some example methods

Two example methods are given here. In the first method we have chosen $\lambda = \frac{1}{2}$ as this means we obtain a reasonably small error constant. In the second method we have chosen $\lambda = \frac{3}{5}$ as this gives us a small value for $|R(\infty)|$.

$$c = \begin{bmatrix} \frac{21}{22} \\ \frac{1}{3} \\ \frac{2}{3} \\ 1 \end{bmatrix}, \quad \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|ccc} \frac{1}{2} & 0 & 0 & 0 & 1 & \frac{5}{11} & -\frac{21}{968} \\ -\frac{4961}{42336} & \frac{1}{2} & 0 & 0 & 1 & -\frac{2095}{42336} & \frac{1}{1344} \\ \frac{223003}{1518804} & -\frac{57}{82} & \frac{1}{2} & 0 & 1 & \frac{26485}{37044} & -\frac{23}{1176} \\ -\frac{2662}{5453} & \frac{12}{41} & \frac{21}{38} & \frac{1}{2} & 1 & \frac{1}{7} & 0 \\ \hline -\frac{2662}{5453} & \frac{12}{41} & \frac{21}{38} & \frac{1}{2} & 1 & \frac{1}{7} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{123178}{147231} & -\frac{176}{123} & -\frac{56}{57} & \frac{4}{3} & 0 & \frac{46}{189} & 0 \end{array}\right] \tag{5.32}$$

$$c = \left[\frac{911}{1146},\ \frac{1}{3},\ \frac{2}{3},\ 1\right]^T,$$

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|ccc} \frac{3}{5} & 0 & 0 & 0 & 1 & \frac{1117}{5730} & -\frac{2114431}{13133160} \\ -\frac{38596898}{214854795} & \frac{3}{5} & 0 & 0 & 1 & -\frac{18697714}{214854795} & -\frac{2321}{1415070} \\ \frac{642101935076}{12312970210125} & -\frac{85603}{343850} & \frac{3}{5} & 0 & 1 & \frac{12265149017}{46551872250} & -\frac{20886679}{153299250} \\ -\frac{264779098}{118070155} & -\frac{273}{1058} & \frac{130}{49} & \frac{3}{5} & 1 & \frac{451}{1822} & 0 \\ \hline -\frac{264779098}{118070155} & -\frac{273}{1058} & \frac{130}{49} & \frac{3}{5} & 1 & \frac{451}{1822} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{598434324}{118070155} & -\frac{8721}{5290} & -\frac{585}{98} & \frac{3}{2} & 0 & \frac{9561}{9110} & 0 \end{array}\right]. \tag{5.33}$$

5.4 Starting the method

For explicit ARK methods, it is possible to start the numerical process using a starter of the form

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right] = \left[\begin{array}{cc|c} 0 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 0 & 0 \\ -1 & 1 & 0 \end{array}\right].$$

From an initial value $y_0 = y(x_0)$, this preliminary step computes in turn

$$\begin{aligned} Y_1 &= y_0, \\ hF_1 &= hf(x_0, Y_1) && = hy'(x_0), \\ Y_2 &= y_0 + hF_1 && = y(x_0 + h) + O(h^2), \\ hF_2 &= hf(x_0 + h, Y_2) && = hy'(x_0 + h) + O(h^3), \\ y_1^{[0]} &= y_0, \\ y_2^{[0]} &= hF_1 && = hy'(x_0), \\ y_3^{[0]} &= hF_2 - hF_1 && = h^2 y''(x_0) + O(h^3). \end{aligned}$$

Because the method requires input to the first step of approximations to $y(x_0)$, $hy'(x_0)$ and $h^2y''(x_0)$, with the last of these accurate to within $O(h^3)$, this simple process is perfectly adequate.

However, for a stiff ARK method we will need to avoid computing $hf(x_0, y_0)$, except as the solution to an implicit equation of the form

$$Y = \lambda h f(X, Y) + C,$$

where C is a known quantity. Hence, we consider starting procedures of the form

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right],$$

where $\widehat{A}$ has a diagonally implicit structure, with the diagonal element λ equal to that of the main method, and where $\widehat{U} = \mathbf{1}$ and $\widehat{V} = e_1$. It is advisable to advance the solution a single step in carrying out the starting process, so that we actually compute approximations at the point $x_1 = x_0 + h$, as follows

$$y_1^{[1]} \approx y(x_1),$$

$$y_2^{[1]} \approx hy'(x_1),$$

$$y_3^{[1]} \approx h^2 y''(x_1).$$

We will examine in detail the construction of a starting method appropriate for the three stage third order method (5.19).

As part of the design of this method, $\beta^T c = 1$ appears as an order condition and corresponds to the requirement that, for the starting method, $e_3^T \widehat{B}\widehat{c} = 1$, which, together with $\beta^T \mathbf{1} + v_{31} = 0$, corresponding to $e_3^T \widehat{B}\mathbf{1} = 0$, is exactly the condition that $y_3^{[1]} = h^2 y''(x_1) + O(h^3)$. Note also that $\beta^T c^2 = \frac{5}{3}$ and $\beta^T Ac = \frac{5}{6}$. Although corresponding assumptions are not strictly necessary for the starting method, we will assume as additional requirements that

$$e_3^T \widehat{B}\widehat{c}^2 = \frac{5}{3}, \qquad e_3^T \widehat{A}\widehat{c} = \frac{5}{6}.$$

We also assume that $\lambda = \frac{1}{3}$ for the starting method, as for the main method.

To obtain order 3 and to satisfy these additional constraints, four stages are necessary and we will also assume that $c_4 = 1$ so that we can then aim for a method for which

$$e_1^T \widehat{B} = e_4^T \widehat{A}, \quad e_2^T \widehat{B} = e_4^T.$$

Although the starting method requires more stages than the ARK method used for propogation, because it is used only to start the method, this does not add substantially to the cost of using a method.

It now transpires that the starting method is fully determined once suitable values of $\widehat{c}_2$ and $\widehat{c}_3$ have been determined. Because values of the coefficients of the method are very sensitive to values of these free abscissae, we will choose values that give reasonably small values of the magnitudes of these coefficients. Suitable choices are

$$\widehat{c}_2 = \frac{2}{3}, \qquad \widehat{c}_3 = 0.$$

From these considerations, the following starting method has been found:

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right] = \left[\begin{array}{cccc|c} \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 1 \\ -\frac{5}{3} & \frac{4}{3} & \frac{1}{3} & 0 & 1 \\ 1 & -\frac{1}{4} & -\frac{1}{12} & \frac{1}{3} & 1 \\ \hline 1 & -\frac{1}{4} & -\frac{1}{12} & \frac{1}{3} & 1 \\ 0 & 0 & 0 & 1 & 0 \\ -2 & -1 & \frac{2}{3} & \frac{7}{3} & 0 \end{array}\right].$$

Even though the method is used only for the starting step it is interesting to note that the first component (which approximates $y(x_0 + h)$) is A-stable with stability function

$$\frac{1 - \frac{1}{3}z - \frac{1}{6}z^2 + \frac{1}{54}z^3}{(1 - \frac{1}{3}z)^4}.$$

For the method (5.20), a suitable starting method is found to be

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right] = \left[\begin{array}{cccc|c} \frac{2}{5} & 0 & 0 & 0 & 1 \\ \frac{1}{10} & \frac{2}{5} & 0 & 0 & 1 \\ -\frac{9}{11} & \frac{78}{55} & \frac{2}{5} & 0 & 1 \\ \frac{25}{18} & -\frac{2}{3} & -\frac{11}{90} & \frac{2}{5} & 1 \\ \hline \frac{25}{18} & -\frac{2}{3} & -\frac{11}{90} & \frac{2}{5} & 1 \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{5}{9} & -\frac{4}{3} & \frac{979}{1395} & \frac{184}{155} & 0 \end{array}\right].$$

Again it is found that the Runge–Kutta method which generates the value of $y_1^{[1]}$ is A-stable, with stability function

$$\frac{1 - \frac{3}{5}z - \frac{7}{50}z^2 + \frac{53}{750}z^3}{(1 - \frac{2}{5}z)^4}.$$

Although finding starting methods for the fourth order methods is complicated it can be done in a similar way to the third order methods. For example, a suitable starting method for

(5.32) is found to be

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right] = \left[\begin{array}{ccccc|c} \frac{1}{2} & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{10} & \frac{1}{2} & 0 & 0 & 0 & 1 \\ -2 & \frac{5}{2} & \frac{1}{2} & 0 & 0 & 1 \\ -\frac{25}{2} & \frac{35}{2} & -\frac{7}{2} & \frac{1}{2} & 0 & 1 \\ \frac{10}{3} & -\frac{125}{42} & \frac{1}{6} & -\frac{1}{42} & \frac{1}{2} & 1 \\ \hline \frac{10}{3} & -\frac{125}{42} & \frac{1}{6} & -\frac{1}{42} & \frac{1}{2} & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ -\frac{262}{75} & 2 & \frac{11}{50} & \frac{4}{75} & \frac{61}{50} & 0 \end{array}\right].$$

For the method (5.33), a suitable starting method is found to be

$$\left[\begin{array}{cc} \widehat{A} & \widehat{U} \\ \widehat{B} & \widehat{V} \end{array}\right] = \left[\begin{array}{ccccc|c} \frac{3}{5} & 0 & 0 & 0 & 0 & 1 \\ -\frac{348}{637} & \frac{3}{5} & 0 & 0 & 0 & 1 \\ -\frac{68311}{178350} & -\frac{3029}{178350} & \frac{3}{5} & 0 & 0 & 1 \\ -\frac{567619159}{2934888450} & -\frac{50172957737}{105204462900} & \frac{2736300189}{3144041420} & \frac{3}{5} & 0 & 1 \\ \frac{448231}{208800} & \frac{73148383399}{115642000800} & -\frac{134603}{167760} & -\frac{674687}{427860} & \frac{3}{5} & 1 \\ \hline \frac{448231}{208800} & \frac{73148383399}{115642000800} & -\frac{134603}{167760} & -\frac{674687}{427860} & \frac{3}{5} & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ -\frac{7192100307338731}{1418153128942500} & \frac{9782934214337011}{1418153128942500} & -\frac{22410947189722}{2445091601625} & \frac{22410947189722}{2445091601625} & -\frac{496328334674}{271676844625} & 0 \end{array}\right].$$

CHAPTER 6

Numerical Experiments

I think there's a world market for maybe five computers.

Thomas Watson, chairman of IBM (1943)

This chapter presents the results from a variety of numerical experiments which verify that ARK methods are competitive methods for solving both ordinary differential equations and delay differential equations. We also wish to confirm that the special 'fifth' order methods discussed in Chapter 4 behave like fifth order methods in practice, both for fixed and variable stepsize implementations.

6.1 Non-stiff methods

The explicit methods discussed in Chapters 3 and 4 will be compared against existing methods using fixed stepsize, fixed variable stepsize and variable stepsize implementations.

6.1.1 Fixed stepsize

In code development there are many choices that need to be made apart from the basic method to be used. By comparing ARK methods against existing methods using fixed stepsize implementation it is possible to compare the methods themselves and not any design choices that have been made.

The problems that will be used for this comparison are the DETest problem set [42]. For convenience they are listed in section A.1.

We wish to compare both fourth order methods with four stages and our special 'fifth' order methods against existing methods. The methods we will use for these comparisons are ARK4, a fourth order, four stage ARK method, the tableau of which is

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{cccc|ccc} 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} \\ \frac{1}{16} & 0 & 0 & 0 & 1 & \frac{7}{16} & \frac{1}{16} \\ -\frac{1}{4} & 2 & 0 & 0 & 1 & -\frac{3}{4} & -\frac{1}{4} \\ 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ \hline 0 & \frac{2}{3} & \frac{1}{6} & 0 & 1 & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & -\frac{2}{3} & 2 & 0 & -1 & 0 \end{array}\right],$$

where $c = [1, \frac{1}{2}, 1, 1]$; ARK451, the original 'fifth' order method, the tableau of which is given in equation (4.1); ARK452, the optimised 'fifth' order method, the tableau of which is given in equation (4.25); RK45, a fourth order, five stage Runge–Kutta method, the tableau of which is

$$\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{3} & \frac{1}{3} & & & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & & & \\ \frac{1}{2} & \frac{1}{8} & 0 & \frac{3}{8} & & & \\ 1 & \frac{1}{2} & 0 & -\frac{3}{2} & 2 & & \\ \hline & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6} \end{array}$$

and RK56, the popular Dormand and Prince method [29] which is a fifth order, seven stage

Runge–Kutta method, the tableau of which is

$$
\begin{array}{c|ccccccc}
0 & & & & & & & \\
\frac{1}{5} & \frac{1}{5} & & & & & & \\
\frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\
\frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\
\frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\
1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\
\hline
 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\
 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \\
\hline
 & -\frac{71}{57600} & 0 & \frac{71}{16695} & -\frac{71}{1920} & \frac{17253}{339200} & -\frac{22}{525} & \frac{1}{40}
\end{array}
\quad . \tag{6.1}
$$

To ensure a fair comparison between the methods, the method with 4 stages has been implemented with a stepsize of $\frac{4}{5}h$ and the method with 7 stages has been implemented with a stepsize of $\frac{7}{5}h$. The values of h used is problem dependent, as some problems required smaller h than others to obtain reasonable accuracy. The results of these experiments are given in Figures 6.1 - 6.5.

We can see that these results are very promising for our special 'fifth' order methods. Not only are they behaving like fifth order, but in many cases they are competitive with the Dormand and Prince method. There is very little difference between the accuracy of the two special methods, although the optimised method does perform slightly better on some problems.

The fourth order ARK method also performs well, although unfortunately not always as well as the fourth order Runge–Kutta method. This method was chosen for its simple coefficients. It is hoped that if these experiments were repeated with an optimised method the results might be more competitive.

It should be noted that for some of the results, a flattening-out is observed when h is small. This is due to round-off error.

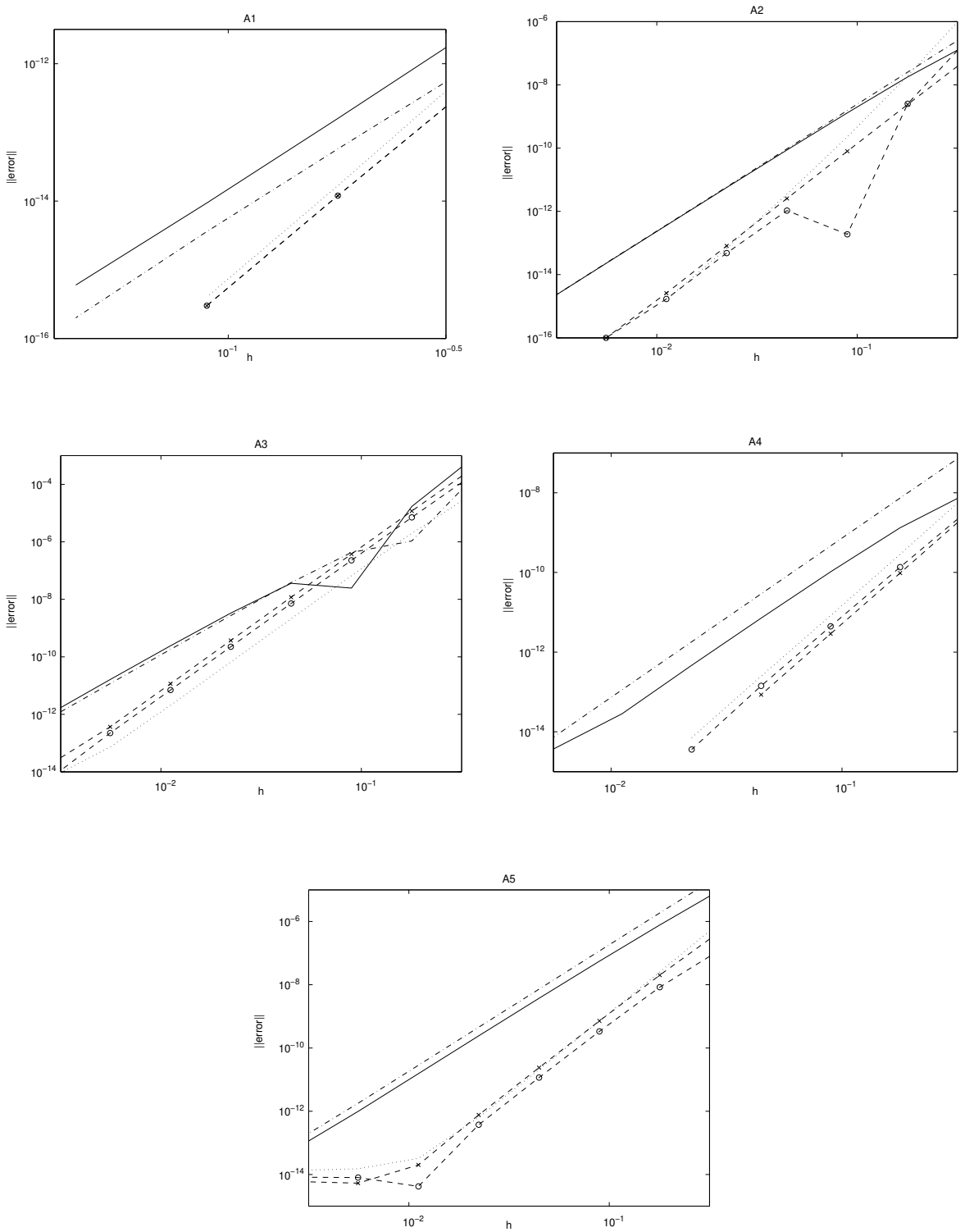


Figure 6.1: Comparison between RK45 ($-\cdot$), RK56 ($\cdots$), ARK4 (—), ARK451 (x) and ARK452 ($\circ$) using constant stepsize for the class A DETest problems.

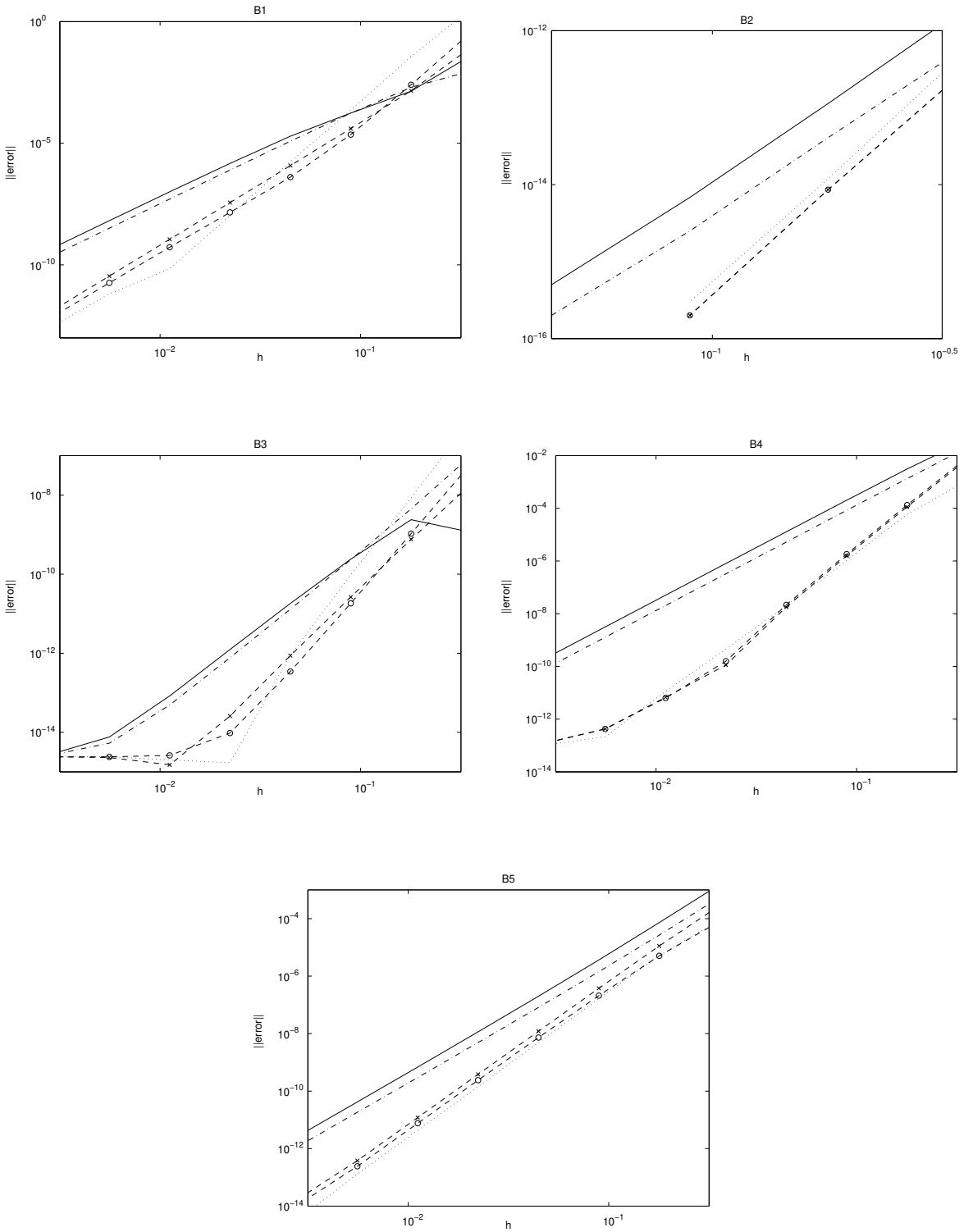


Figure 6.2: Comparison between RK45 ($-\cdot$), RK56 ($\cdots$), ARK4 (—), ARK451 (x) and ARK452 ($\circ$) using constant stepsize for the class B DETest problems.

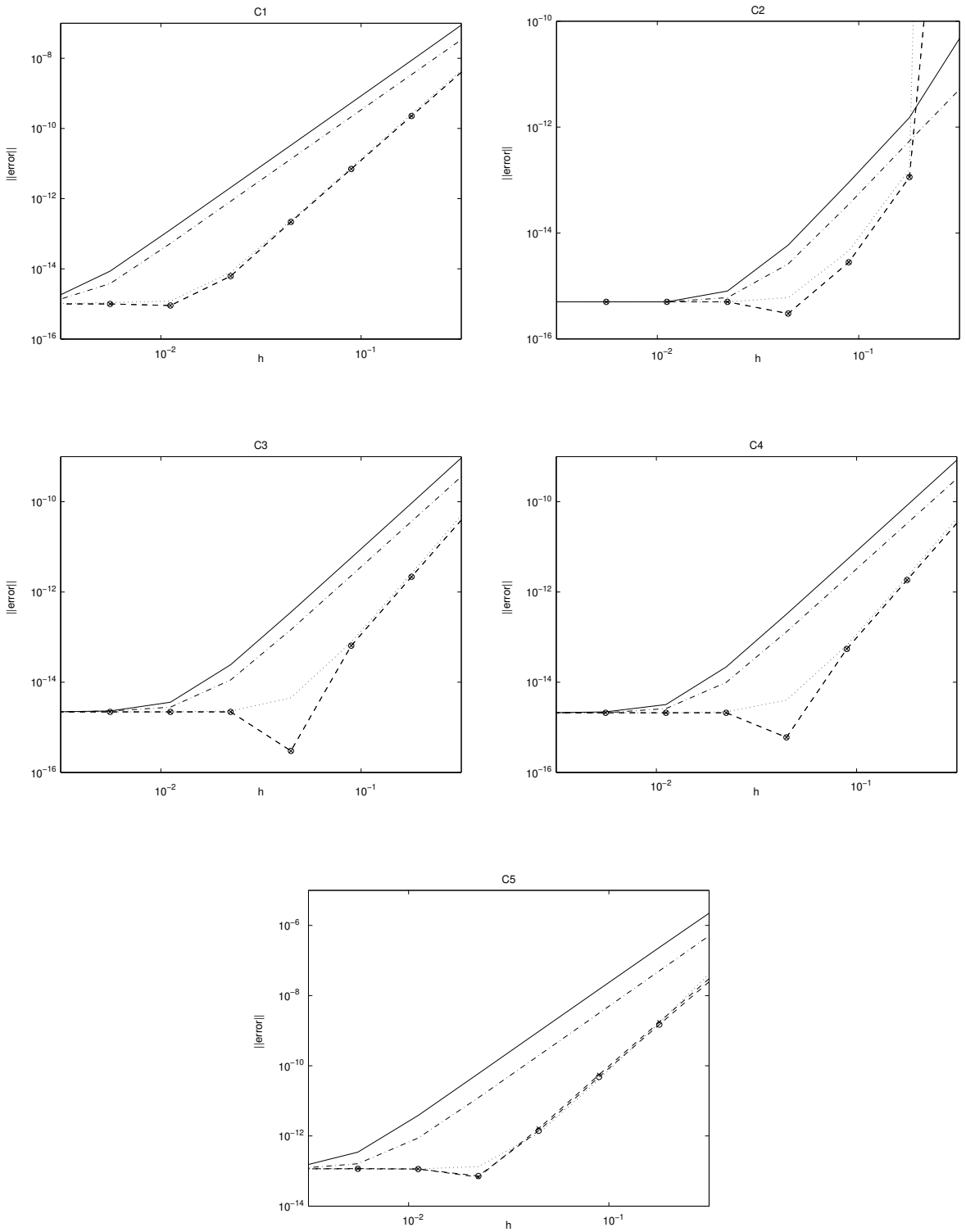


Figure 6.3: Comparison between RK45 ($-\cdot$), RK56 ($\cdots$), ARK4 (—), ARK451 (x) and ARK452 ($\circ$) using constant stepsize for the class C DETest problems.

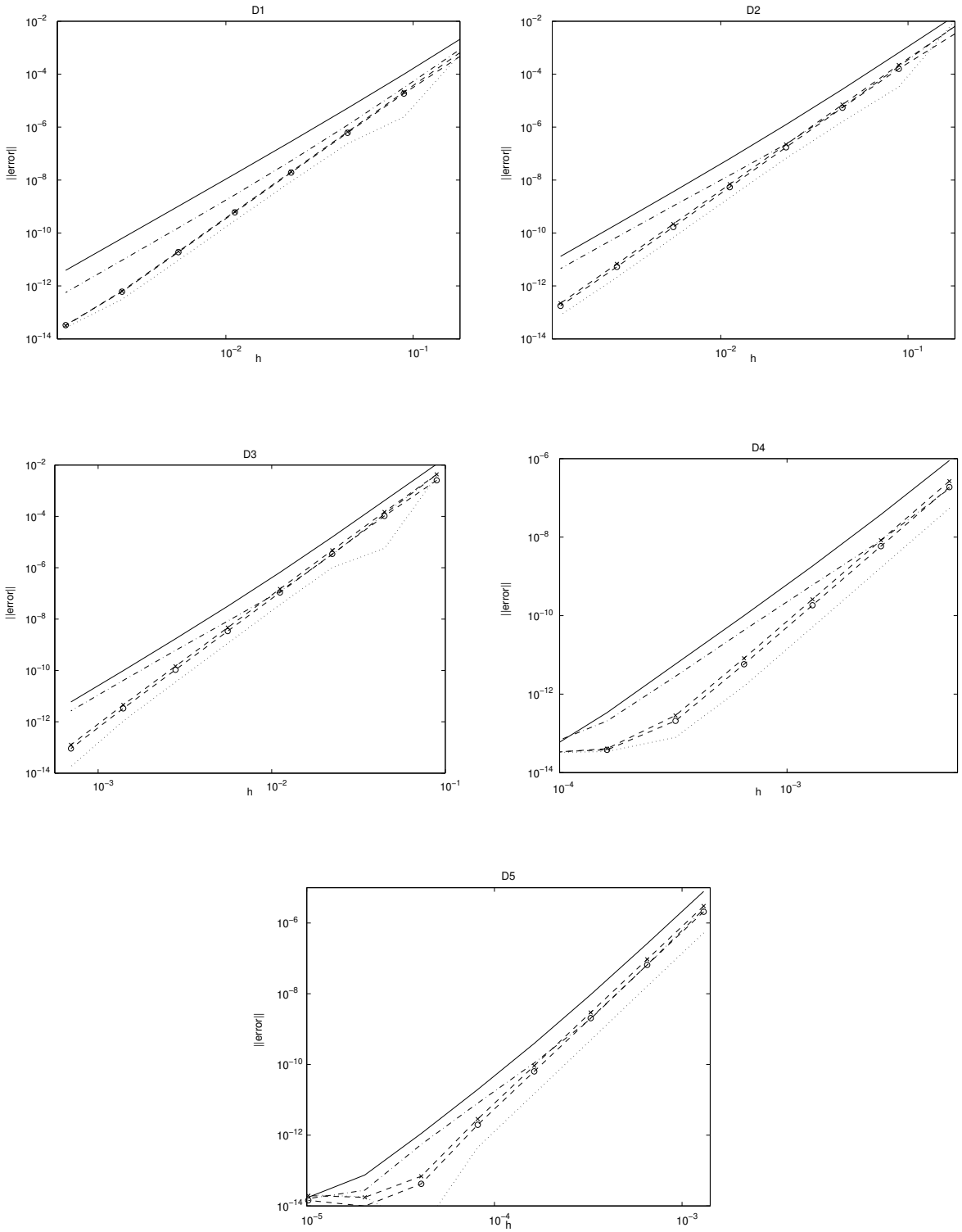


Figure 6.4: Comparison between RK45 ($-\cdot$), RK56 ($\cdots$), ARK4 (—), ARK451 (x) and ARK452 ($\circ$) using constant stepsize for the class D DETest problems.

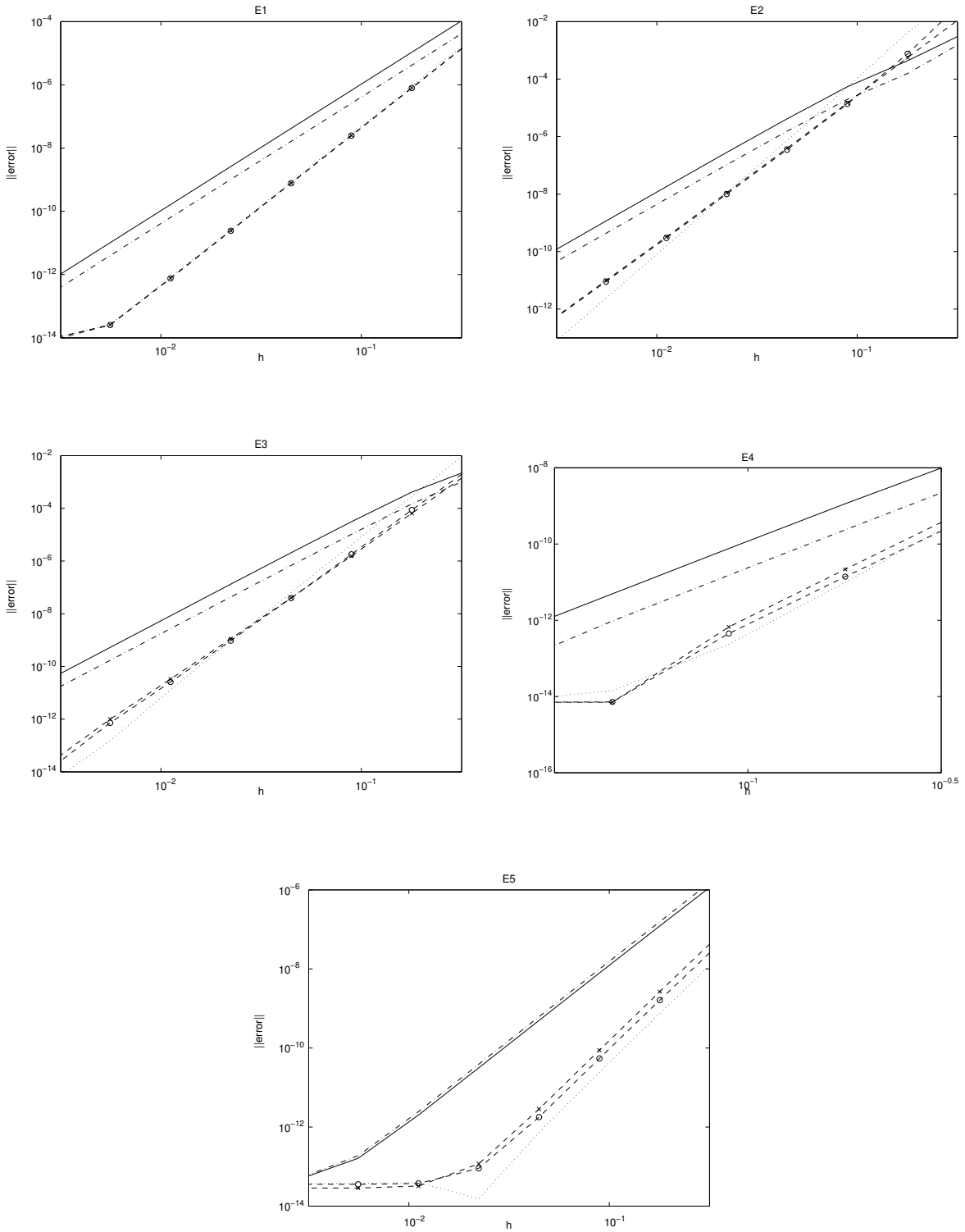


Figure 6.5: Comparison between RK45 ($-\cdot$), RK56 ($\cdots$), ARK4 (—), ARK451 (x) and ARK452 ($\circ$) using constant stepsize for the class E DETest problems.

6.1.2 Fixed variable stepsize

In this section we will examine the effect of stepsize changes on the accuracy of the methods. To do this, experiments were carried out using a scheme in which a predetermined sequence of stepsizes was imposed. For each sequence of 5 steps, stepsizes in the ratios $1 : r : r^2 : r : 1$ were used, where r is a parameter. The expected performance of a fifth order method, assuming asymptotic behaviour under such a scheme, would be that the global truncation errors would be scaled up by a factor

$$F(r) = \left(\frac{2 + 2r + r^2}{5}\right)^{-6} \left(\frac{2 + 2r^6 + r^{12}}{5}\right), \tag{6.2}$$

assuming that the total number of steps actually carried out is independent of r.

To see why this is true we examine the error that would have been generated using a fixed stepsize. The size of each step would have been $\left(\frac{2+2r+r^2}{5}\right) h$, so the error on each step would have been

$$Ch^6 \left(\frac{2 + 2r + r^2}{5}\right)^6.$$

Using the fixed variable scheme the total error generated over 5 steps is $2Ch^6 + 2Ch^6r^6 + h^6r^{12}$, giving an average error per step of

$$Ch^6 \left(\frac{2 + 2r^6 + r^{12}}{5}\right).$$

Dividing the above two expressions gives equation (6.2). For a stepsize ratio of $r = 1.5$ we would expect the error to grow by a factor of 3.3253 and for $r = 2$ we would expect a factor of 13.2063. However, this prediction is somewhat optimistic because it ignores the possibility that additional errors may have been introduced by the very process of adjusting the data between one step and the next.

Two methods have been compared. These are the special 'fifth' order ARK method given in (4.1) and the fifth order Runge–Kutta method given in (6.1). For the two methods, global errors have been computer for n steps. The ARK method has been implemented in the manner suggested in Chapter 4 to ensure fifth order behaviour. Only the results for the last problem of each class of the DETest problems are presented here, for berevity. The remainder of the problems produced similar results. The results for the ARK methods are given in Tables 6.1 - 6.5. The results for the Runge–Kutta method are given in Tables 6.6 - 6.10. The top grid of each table gives the ratio between the errors when the stepsize has been doubled. We would expect this to be about 32 for a fifth order method. The lower grid of each table gives the deterioration factor. We would expect this to be approximately $F(r)$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
240	1.7183×10^{-10}	30.58	6.0015×10^{-10}	32.46	2.1213×10^{-9}	32.73
480	5.6189×10^{-12}	32.84	1.8491×10^{-11}	33.08	6.4819×10^{-11}	32.96
960	1.711×10^{-13}	50.32	5.59×10^{-13}	57.04	1.9665×10^{-12}	38.71
1920	3.4×10^{-15}		9.8×10^{-15}		5.08×10^{-14}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
240	1.7183×10^{-10}	3.49	6.0015×10^{-10}	12.35	2.1213×10^{-9}
480	5.6189×10^{-12}	3.29	1.8491×10^{-11}	11.54	6.4819×10^{-11}
960	1.711×10^{-13}	3.27	5.59×10^{-13}	11.49	1.9665×10^{-12}
1920	3.4×10^{-15}	2.88	9.8×10^{-15}	14.94	5.08×10^{-14}

Table 6.1: Comparison of error behaviours for fixed and variable stepsizes for problem A5 using method ARK45. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
480	2.8246×10^{-9}	31.71	8.5225×10^{-9}	31.61	2.9871×10^{-8}	31.47
960	8.9064×10^{-11}	31.82	2.6959×10^{-10}	31.81	9.4911×10^{-10}	31.76
1920	2.7988×10^{-12}	30.62	8.4747×10^{-12}	31.54	2.9886×10^{-11}	31.76
3840	9.14×10^{-14}		2.687×10^{-13}		9.411×10^{-13}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
480	2.8246×10^{-9}	3.02	8.5225×10^{-9}	10.58	2.9871×10^{-8}
960	8.9064×10^{-11}	3.03	2.6959×10^{-10}	10.66	9.4911×10^{-10}
1920	2.7988×10^{-12}	3.03	8.4747×10^{-12}	10.68	2.9886×10^{-11}
3840	9.14×10^{-14}	2.94	2.687×10^{-13}	10.30	9.411×10^{-13}

Table 6.2: Comparison of error behaviours for fixed and variable stepsizes for problem B5 using method ARK45. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
60	1.3022×10^{-8}	31.61	3.9008×10^{-8}	31.33	1.3675×10^{-7}	31.15
120	4.1199×10^{-10}	32.05	1.2452×10^{-9}	31.77	4.3899×10^{-9}	31.62
240	1.2854×10^{-11}	43.11	3.9199×10^{-11}	34.84	1.3882×10^{-10}	32.60
480	2.982×10^{-13}		1.1251×10^{-12}		4.2586×10^{-12}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
60	1.3022×10^{-8}	3.00	3.9008×10^{-8}	10.50	1.3675×10^{-7}
120	4.1199×10^{-10}	3.02	1.2452×10^{-9}	10.66	4.3899×10^{-9}
240	1.2854×10^{-11}	3.05	3.9199×10^{-11}	10.80	1.3882×10^{-10}
480	2.982×10^{-13}	3.77	1.1251×10^{-12}	14.28	4.2586×10^{-12}

Table 6.3: Comparison of error behaviours for fixed and variable stepsizes for problem C5 using method ARK45. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
3840	9.5847×10^{-4}	30.76	2.9954×10^{-3}	30.89	1.0156×10^{-2}	29.75
7680	3.1159×10^{-5}	31.71	9.6976×10^{-5}	32.00	3.4141×10^{-4}	31.80
15360	9.8247×10^{-7}	31.94	3.0301×10^{-6}	32.16	1.0735×10^{-5}	32.14
30720	3.0756×10^{-8}		9.4225×10^{-8}		3.3398×10^{-7}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
3840	9.5847×10^{-4}	3.13	2.9954×10^{-3}	10.60	1.0156×10^{-2}
7680	3.1159×10^{-5}	3.11	9.6976×10^{-5}	10.96	3.4141×10^{-4}
15360	9.8247×10^{-7}	3.08	3.0301×10^{-6}	10.93	1.0735×10^{-5}
30720	3.0756×10^{-8}	3.06	9.4225×10^{-8}	10.86	3.3398×10^{-7}

Table 6.4: Comparison of error behaviours for fixed and variable stepsizes for problem D5 using method ARK45. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
60	1.874×10^{-8}	29.50	4.9849×10^{-8}	27.28	1.6662×10^{-7}	26.64
120	6.3521×10^{-10}	30.65	1.8274×10^{-9}	29.84	6.2546×10^{-9}	29.51
240	2.0723×10^{-11}	29.91	6.1248×10^{-11}	30.50	2.1197×10^{-10}	30.69
480	6.928×10^{-13}		2.008×10^{-12}		6.9068×10^{-12}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
60	1.874×10^{-8}	2.66	4.9849×10^{-8}	8.89	1.6662×10^{-7}
120	6.3521×10^{-10}	2.88	1.8274×10^{-9}	9.85	6.2546×10^{-9}
240	2.0723×10^{-11}	2.96	6.1248×10^{-11}	10.23	2.1197×10^{-10}
480	6.928×10^{-13}	2.90	2.008×10^{-12}	9.97	6.9068×10^{-12}

Table 6.5: Comparison of error behaviours for fixed and variable stepsizes for problem E5 using method ARK45. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
240	8.9316×10^{-11}	33.34	2.8857×10^{-10}	32.38	1.1882×10^{-9}	32.89
480	2.6793×10^{-12}	28.35	8.9109×10^{-12}	31.37	3.613×10^{-11}	32.61
960	9.45×10^{-14}	5.83	2.841×10^{-13}	14.06	1.1078×10^{-12}	23.62
1920	1.62×10^{-14}		2.02×10^{-14}		4.69×10^{-14}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
240	8.9316×10^{-11}	3.23	2.8857×10^{-10}	13.30	1.1882×10^{-9}
480	2.6793×10^{-12}	3.33	8.9109×10^{-12}	13.48	3.613×10^{-11}
960	9.45×10^{-14}	3.01	2.841×10^{-13}	11.72	1.1078×10^{-12}
1920	1.62×10^{-14}	1.25	2.02×10^{-14}	2.90	4.69×10^{-14}

Table 6.6: Comparison of error behaviours for fixed and variable stepsizes for problem A5 using Dormand and Prince. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
480	6.1763×10^{-10}	33.25	2.12×10^{-9}	33.68	8.6542×10^{-9}	33.98
960	1.8576×10^{-11}	32.77	6.2941×10^{-11}	33.02	2.547×10^{-10}	33.25
1920	5.668×10^{-13}	34.99	1.9063×10^{-12}	33.92	7.6608×10^{-12}	32.91
3840	1.62×10^{-14}		5.62×10^{-14}		2.328×10^{-13}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
480	6.1763×10^{-10}	3.43	2.12×10^{-9}	14.01	8.6542×10^{-9}
960	1.8576×10^{-11}	3.39	6.2941×10^{-11}	13.71	2.547×10^{-10}
1920	5.668×10^{-13}	3.36	1.9063×10^{-12}	13.52	7.6608×10^{-12}
3840	1.62×10^{-14}	3.47	5.62×10^{-14}	14.37	2.328×10^{-13}

Table 6.7: Comparison of error behaviours for fixed and variable stepsizes for problem B5 using Dormand and Prince. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
60	8.3002×10^{-9}	43.89	3.5367×10^{-8}	48.18	1.7603×10^{-7}	51.81
120	1.8909×10^{-10}	37.81	7.3405×10^{-10}	40.91	3.3974×10^{-9}	43.73
240	5.0009×10^{-12}	22.74	1.7945×10^{-11}	32.28	7.7685×10^{-11}	37.06
480	2.199×10^{-13}		5.559×10^{-13}		2.0961×10^{-12}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
60	8.3002×10^{-9}	4.26	3.5367×10^{-8}	21.21	1.7603×10^{-7}
120	1.8909×10^{-10}	3.88	7.3405×10^{-10}	17.97	3.3974×10^{-9}
240	5.0009×10^{-12}	3.59	1.7945×10^{-11}	15.53	7.7685×10^{-11}
480	2.199×10^{-13}	2.53	5.559×10^{-13}	9.53	2.0961×10^{-12}

Table 6.8: Comparison of error behaviours for fixed and variable stepsizes for problem C5 using Dormand and Prince. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
3840	2.5141×10^{-5}	8.68	2.0741×10^{-4}	27.92	3.0414×10^{-3}	234.99
7680	2.8966×10^{-6}	29.59	7.4281×10^{-6}	23.10	1.2942×10^{-5}	10.74
15360	9.7901×10^{-8}	33.02	3.2153×10^{-7}	31.97	1.205×10^{-6}	29.83
30720	2.9652×10^{-9}		1.0058×10^{-8}		4.039×10^{-8}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
3840	2.5141×10^{-5}	8.25	2.0741×10^{-4}	120.97	3.0414×10^{-3}
7680	2.8966×10^{-6}	2.56	7.4281×10^{-6}	4.47	1.2942×10^{-5}
15360	9.7901×10^{-8}	3.28	3.2153×10^{-7}	12.31	1.205×10^{-6}
30720	2.9652×10^{-9}	3.39	1.0058×10^{-8}	13.62	4.039×10^{-8}

Table 6.9: Comparison of error behaviours for fixed and variable stepsizes for problem D5 using Dormand and Prince. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

n	Error ($r = 1$)	Ratio	Error ($r = 1.5$)	Ratio	Error ($r = 2$)	Ratio
60	3.0722×10^{-9}	30.39	8.7498×10^{-9}	27.44	3.2743×10^{-8}	26.59
120	1.0109×10^{-10}	31.57	3.1882×10^{-10}	30.25	1.2314×10^{-9}	29.74
240	3.2026×10^{-12}	51.00	1.054×10^{-11}	34.84	4.14×10^{-11}	31.93
480	6.28×10^{-14}		3.025×10^{-13}		1.2965×10^{-12}	

n	Error ($r = 1$)	Deterioration factor	Error ($r = 1.5$)	Deterioration factor	Error ($r = 2$)
60	3.0722×10^{-9}	2.85	8.7498×10^{-9}	10.66	3.2743×10^{-8}
120	1.0109×10^{-10}	3.15	3.1882×10^{-10}	12.18	1.2314×10^{-9}
240	3.2026×10^{-12}	3.29	1.054×10^{-11}	12.93	4.14×10^{-11}
480	6.28×10^{-14}	4.82	3.025×10^{-13}	20.64	1.2965×10^{-12}

Table 6.10: Comparison of error behaviours for fixed and variable stepsizes for problem E5 using Dormand and Prince. We expect the ratio to be about 32. The theoretical values for the deterioration factor are $F(1.5) = 3.3253$ and $F(2) = 13.2063$.

As we can see from the tables, both methods produce a ratio of approximately 32 when the stepsize is doubled. This indicates that our variable stepsize implementation is maintaining the correct order. The few ratios that diverge significantly from 32 are due to round off error in the smaller step-sizes.

The deterioration factors for the Runge–Kutta method are very close to the theoretical values predicted by $F(r)$, however the deterioration factors for the ARK method are much better than predicted. For $r = 1.5$ this factor is approximately 3, for $r = 2$ this factor is approximately 11. This can be interpreted to mean that changing stepsize does not add additional errors to the computation, but rather that there can even be a cancellation of some of the accumulated truncation error under stepsize change.

6.1.3 Variable stepsize

In practice, unless it is required by the problem, most ordinary differential equations are solved using a variable stepsize code. This is because parts of the solution can be very smooth, hence a large stepsize is appropriate, while other parts can change rapidly, requiring a small stepsize. It is therefore important to see how well our methods compare in a variable stepsize implementation.

We have compared our special 'fifth' order ARK method given in (4.1) with the Dormand and Prince method given in (6.1). They were tested on the DETest problem set, with many different tolerances, $tol = 10^{-i}, i = 3, .., 12$. The results for the final problem in each class are plotted in Figure 6.6. The number of function evaluations has been plotted against the error. Function evaluations have been chosen as it was felt that `etime` is an unreliable measure of work done on shared computers, and `flops`, while giving the same information as function evaluations, are more difficult to measure.

As we can see, these results are promising, although the Dormand and Prince method gives slightly better results. This is possibly, in part, due to the manner in which the error estimator for the ARK method was implemented. The stepsize was kept constant over two steps, and then the error estimated at the end of two steps. This has two disadvantages. First, if the error is too large in a step, two steps need to be repeated. Also, it is possible that keeping the stepsize constant over two steps is restricting the growth rate of the stepsize, thereby requiring more steps to finish the integration.

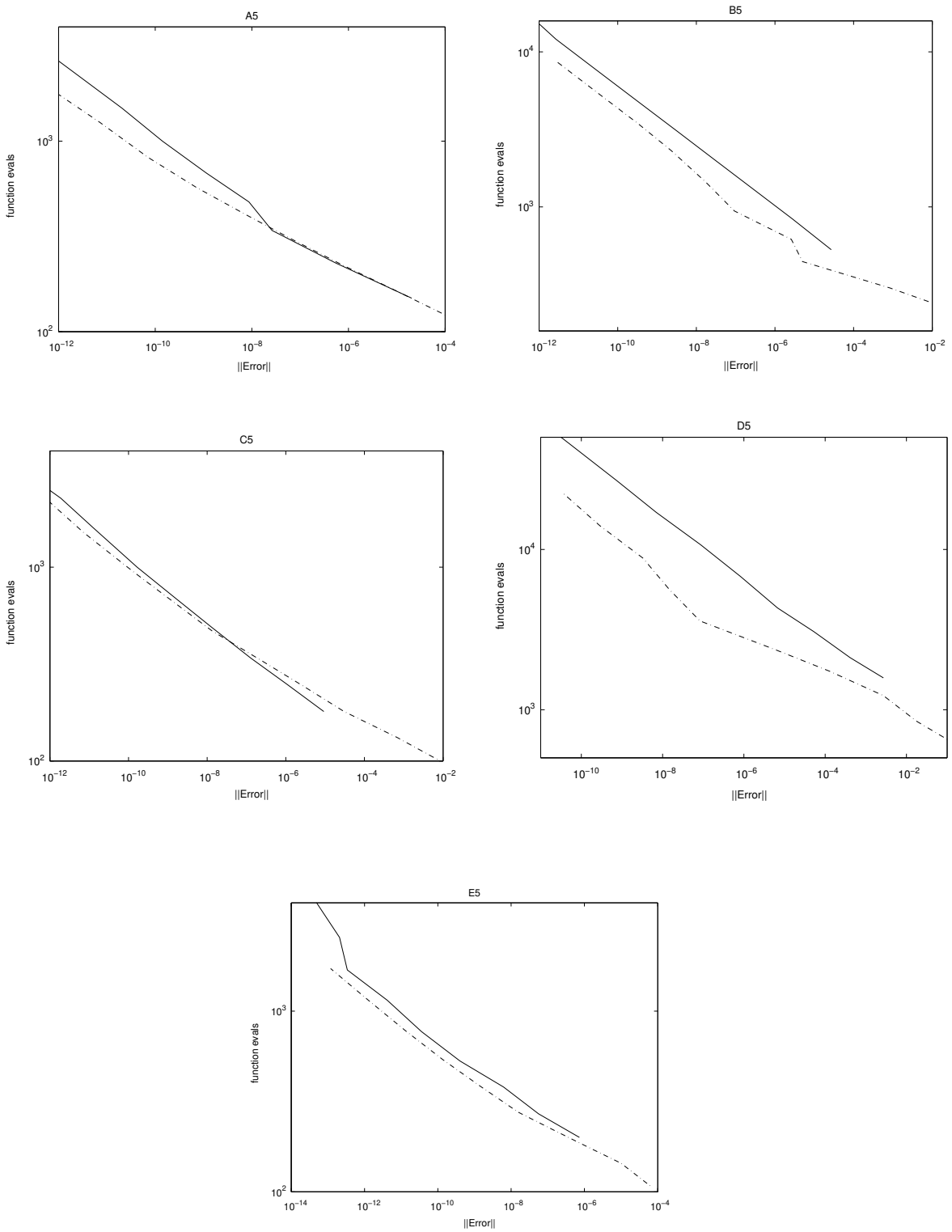


Figure 6.6: Comparison between RK56 (−·) and ARK45 (—) using variable stepsize for a selection of the DETest problems.

6.1.4 DDEs

Solving DDEs requires the use of a good quality interpolator in order to calculate the lag term. One of the main advantages of ARK methods is the ability to interpolate without any additional function evaluations being required. This set of experiments allows us to test not only the method itself, but the interpolator as well.

Again, we have compared our special 'fifth' order ARK method given in (4.1) with the Dormand and Prince method given in (6.1). The interpolator used for the ARK method in these experiments is discussed in section 4.3. The interpolator used for the Dormand and Prince method has been taken from [55]. The coefficients are

$$\tilde{b}_1(\xi) = \xi - \xi^2\frac{1337}{480} + \frac{1039}{360}\xi^3 - \xi^4\frac{1163}{1152},$$

$$\tilde{b}_2(\xi) = 0,$$

$$\tilde{b}_3(\xi) = \xi^2\frac{4216}{1113} - \xi^3\frac{18728}{3339} + \xi^4\frac{7580}{3339},$$

$$\tilde{b}_4(\xi) = \xi^2 - \frac{27}{16} + \xi^3\frac{9}{2} - \xi^4\frac{415}{192},$$

$$\tilde{b}_5(\xi) = -\xi^2\frac{2187}{8480} + \xi^3\frac{2673}{2120} - \xi^4\frac{8991}{6784},$$

$$\tilde{b}_6(\xi) = \xi^2\frac{33}{35} - \xi^3\frac{319}{105} + \xi^4\frac{187}{84},$$

$$\tilde{b}_7(\xi) = 0.$$

The methods were tested on a variety of delay differential equations using variable stepsize, with varying tolerances, $tol = 10^{-i}, i = 3, ..., 13$. The details of these equations are given in section A.3. The results from these experiments are given in Figure 6.7.

The results are very favourable. The ARK method has delivered results which are better than the Dormand and Prince method on three out of the six problems, and worse results on only one. We can also see, for this method, that the order has not deteriorated, and it is still giving order 5 performance.

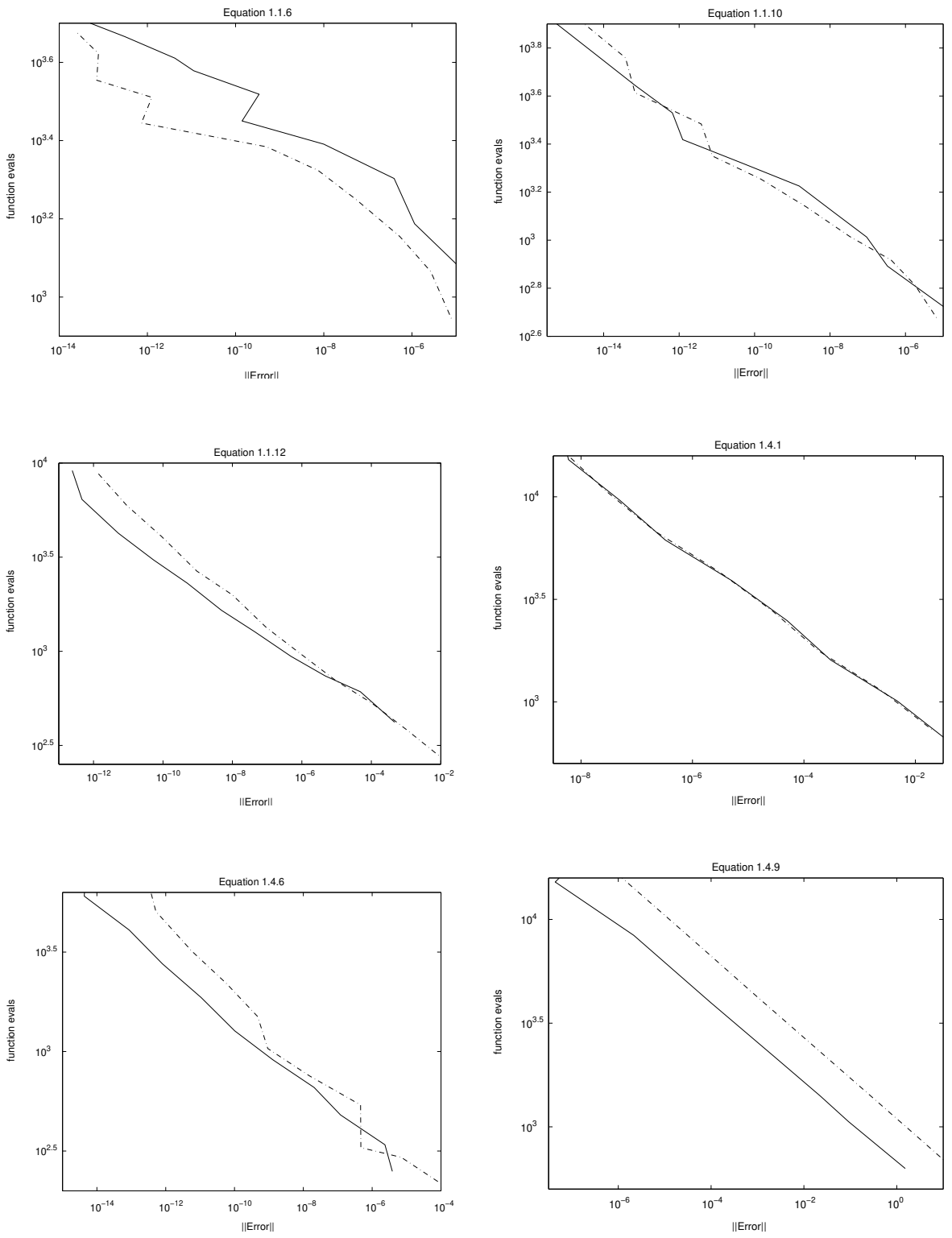


Figure 6.7: Comparison between RK56 (−·) and ARK45 (—) using variable stepsize for a selection of DDE problems.

6.2 Stiff methods

Stiff methods are much more difficult to implement than many non-stiff methods as they are, by necessity, implicit. Newton iterations are now needed to solve for the stage values. As with the explicit methods, we are interested in comparing the stiff methods using fixed stepsize. This allows us to compare the basic methods, and not any design choices or error estimators. To do this we compared the best of the third and fourth order stiff ARK methods (DIARK), i.e. (5.20) and (5.33), and third and fourth order diagonally implicit Runge–Kutta methods (DIRK). The tableaux for the DIRK methods are

$$\begin{array}{c|ccc} \lambda & \lambda & & \\ \frac{1}{2}(1+\lambda) & \frac{1}{2}(1-\lambda) & \lambda & \\ 1 & \frac{1}{4}(-6\lambda^2+16\lambda-1) & \frac{1}{4}(6\lambda^2-20\lambda+5) & \lambda \\ \hline & \frac{1}{4}(-6\lambda^2+16\lambda-1) & \frac{1}{4}(6\lambda^2-20\lambda+5) & \lambda \end{array},$$

where $\lambda = 0.4358665215$, and

$$\begin{array}{c|ccccc} \frac{1}{4} & \frac{1}{4} & & & & \\ \frac{3}{4} & \frac{1}{2} & \frac{1}{4} & & & \\ \frac{11}{20} & \frac{17}{50} & -\frac{1}{25} & \frac{1}{4} & & \\ \frac{1}{2} & \frac{371}{1360} & -\frac{137}{2720} & \frac{15}{544} & \frac{1}{4} & \\ 1 & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4} \\ \hline & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4} \end{array}.$$

The problems used for this comparison are the Oregonator problem, Prothero-Robinson problem and the HIRES problem. Readers unfamiliar with these problems are referred to A.2. The results from these experiments are presented in Figure 6.8.

There is little difference between the performances of the ARK and DIRK methods for the Oregonator and HIRES problems. However, for the Prothero-Robinson problem we see a definite order reduction for all methods. The Prothero-Robinson problem is very stiff. Due to this the order of the method has decreased to the stage order. This gives the ARK methods a big advantage. The DIRK methods are only giving order 1 performance, but the ARK methods are giving order 2 performance. This can be seen from the slope of the graphs. For small stepsizes

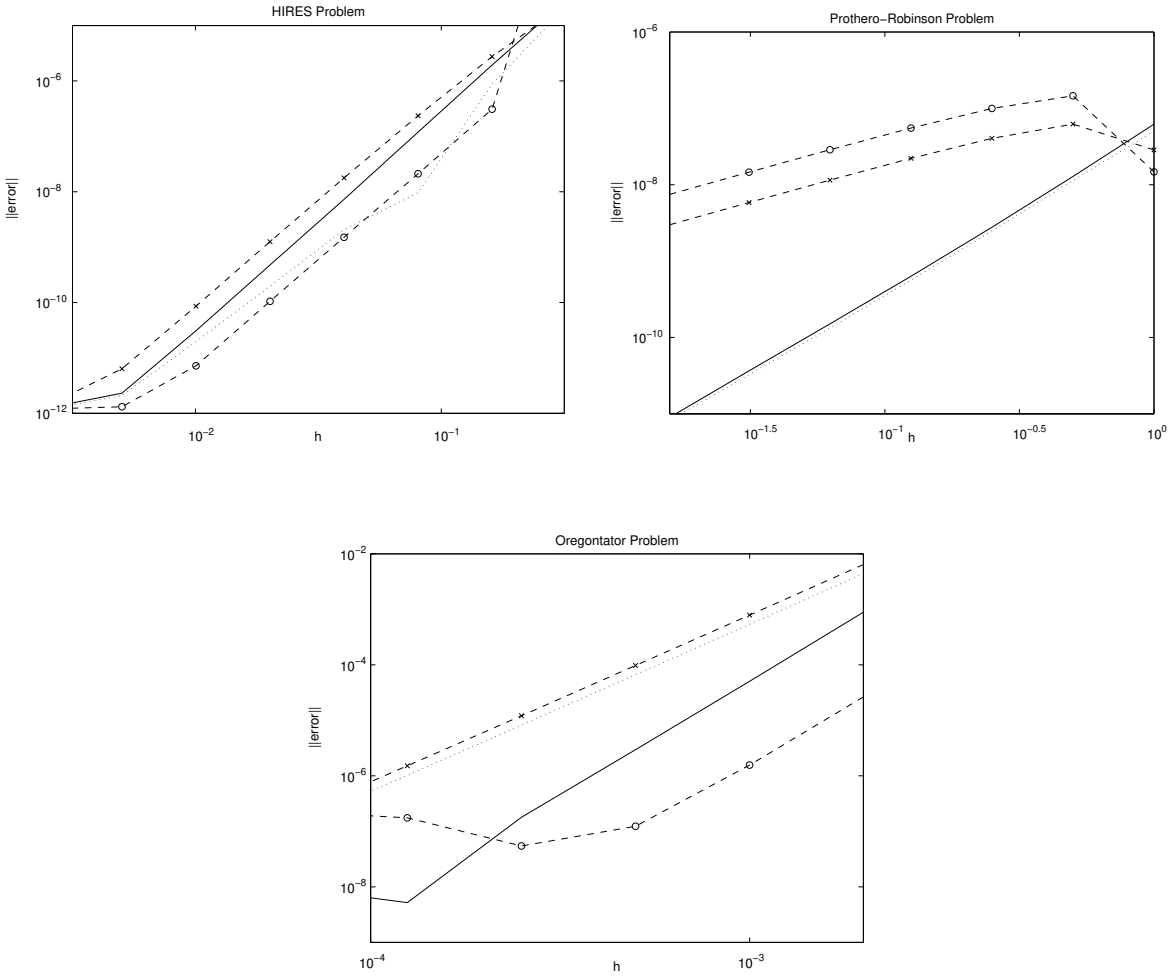


Figure 6.8: Comparison between DIARK3 ($\cdots$), DIARK4 (—), DIRK3 (x) and DIRK4 (o) on a selection of stiff problems.

this particularly makes a big difference. This order reduction reinforces our original motivation to explore only low order diagonally implicit ARK methods.

CHAPTER 7

Conclusions

A Mathematician is a machine for turning coffee into theorems.

Paul Erdös

ARK methods are a special class of general linear methods which retain many of the properties of traditional Runge-Kutta methods, but with some advantages. The main aim of this thesis was to explore ARK methods and to see how these methods compare with traditional Runge-Kutta methods.

The multi-value nature of ARK methods allows a stage order of 2. The advantage of this is that we are able to interpolate or obtain an error estimate at little extra cost. For higher orders, this also means we are able to obtain methods with less stages than for a traditional Runge-Kutta method. The details of third and fourth order methods have been extensively explored. The majority of these ideas will carry forward naturally to higher orders.

A special class of 'fifth' order methods has been explored. This special class of methods is fourth order with five stages, but have had their free parameters chosen in such a way to ensure we have zero error coefficients for the fifth order trees. Although the method exhibits fifth order behaviour for fixed stepsize, the method experiences a reduction in order for variable h, as the fifth order annihilation conditions have not been satisfied. We have shown it is possible, by implementing the method in the correct way, to retain order five behaviour, even for variable h. Numerical optimisation has led to a specific choice of the free parameters which gives competitive performance.

Low order ARK methods for solving stiff problems have been considered. We have chosen to consider only methods which are diagonally implicit to ensure computational costs are kept

as low as possible. We have also restricted ourselves to low order methods as it is likely they will suffer from some order reduction. Methods of order 3 and 4 have been explored, along with the corresponding starting methods.

In general, the results from our numerical experiments are promising. Our special 'fifth' order methods have retained their order for both fixed and variable stepsize. We have also seen that the deterioration in the error when changing stepsize is much better than theoretically predicted. For fixed stepsize, the special ARK method compares well against Dormand and Prince. Unfortunately this is not the case for variable stepsize. This is possibly owing, in part, to the manner in which the error estimator for the ARK method was implemented. The stepsize was kept constant over two steps, and then the error estimated at the end of two steps. This has two disadvantages. First, if the error is too large in a step, two steps need to be repeated. Also, it is possible that keeping the stepsize constant over two steps is restricting the growth rate of the stepsize, meaning it requires more steps to finish the integration.

Unfortunately our fourth order method did not compare as favourably for the fixed stepsize experiments. Although it gave good results, the performance of the fourth order, five stage Runge-Kutta method was better on most problems. The ARK method was chosen for its simple coefficients. It is hoped that if these experiments were repeated with an optimised method the results might be more competitive.

As expected, the ARK methods performed very well on DDEs. It was expected that the higher stage order, and hence ability to interpolate cheaply, would make these methods particularly suited for solving this type of problem. When our 'fifth' order ARK method was compared with Dormand and Prince, the ARK method gave better performance on the large majority of problems used for testing. Using an interpolator which has an order less than that of the method does not appear to have adversely affected its performance.

Our diagonally implicit ARK methods also compared very favourably with traditional DIRK methods. For moderately stiff problems there is little difference between the performance of ARK and DIRK methods. For very stiff problems the ARK methods perform much better. As is expected, on very stiff problems we experience an order reduction to the stage order of the method. This means the DIRK methods only give order 1 performance, but the ARK methods give order 2 performance.

There are still many ideas we would like to explore. The most obvious of these is the extension of explicit ARK methods to higher orders. A stage order of 2 means we can find ARK methods with less stages than a traditional Runge-Kutta method of the same order.

There are still many improvements that could be made to the code developed for this thesis. One possible change is the way in which the error is estimated. It is probable that estimating the error after each step, rather than every two steps, will lead to more efficient code. The stepsize controller we have implemented is the traditional dead-beat controller. Although we have implemented this in such a way as to try and limit the number of unnecessary rejections, it is thought that the proportional integral controller, which has been successfully used by Runge-Kutta methods, will lead to more stable stepsize control. Given how successful our results were for DDEs, we would like to extend the code written to solve DDEs to work for problems with variable delay and state dependent delays. We would like to extend the code developed for solving stiff differential equations to allow variable stepsize.

APPENDIX A

Test Problems

> The mathematician may be compared to a designer of garments, who is utterly oblivious of the creatures whom his garments may fit. To be sure, his art originated in the necessity for clothing such creatures, but this was long ago; to this day a shape will occasionally appear which will fit into the garment as if the garment had been made for it. Then there is no end of surprise and delight.
>
> DANTZIG

The test problems used in this thesis are well known problems used for testing initial value problem solvers. These have been divided into non-stiff and stiff problems. The non-stiff problems come from the DETest problem set [42]. The stiff problems come from a variety of sources. A selection of DDE problems are also included. These come from [57].

A.1 DETest problems

Class A: Single equations

A1: (the negative exponential).

$$y' = -y, \qquad y(0) = 1 \tag{A.1}$$

(solution: $y = Ce^{-x}, C = 1$)

A2: (a special case of the Riccati equation).

$$y' = -\frac{y^3}{2}, \qquad y(0) = 1 \tag{A.2}$$

(solution: $y = 1/\sqrt{x + C}, C = 1$)

A3: (an oscillatory problem).

$$y' = y \cos x, \qquad y(0) = 1 \tag{A.3}$$

(solution: $y = Ce^{\sin x}, C = 1$)

A4: (a logistic curve).

$$y' = \frac{y}{4}\left(1 - \frac{y}{20}\right), \qquad y(0) = 1 \tag{A.4}$$

(solution: $y = \dfrac{20}{1 + 19Ce^{-x^4}}, C = 1$)

A5: (a spiral curve).

$$y' = \frac{y - x}{y + x}, \qquad y(0) = 4 \tag{A.5}$$

(solution in polar co-ordinates: $r = Ce^{-\theta}, C = 4e^{\pi/2}$)

Class B: Small systems

B1: (the growth of two conflicting populations).

$$\begin{aligned} y_1' &= 2(y_1 - y_1 y_2), & y_1(0) &= 1, \\ y_2' &= -(y_2 - y_1 y_2), & y_2(0) &= 3. \end{aligned} \tag{A.6}$$

B2: (a linear chemical reaction).

$$\begin{aligned} y_1' &= -y_1 + y_2, & y_1(0) &= 2, \\ y_2' &= y_1 - 2y_2 + y_3, & y_2(0) &= 0, \\ y_3' &= y_2 - y_3, & y_3(0) &= 1. \end{aligned} \tag{A.7}$$

B3: (a nonlinear chemical reaction).

$$\begin{aligned} y_1' &= -y_1, & y_1(0) &= 1, \\ y_2' &= y_1 - y_2^2, & y_2(0) &= 0, \\ y_3' &= y_2^2, & y_3(0) &= 0. \end{aligned} \tag{A.8}$$

B4: (the integral surface of a torus).

$$\begin{aligned} y_1' &= -y_2 - \frac{y_1 y_3}{\sqrt{y_1^2 + y_2^2}}, & y_1(0) &= 3, \\ y_2' &= \frac{y_1 - y_2 y_3}{\sqrt{y_1^2 + y_2^2}}, & y_2(0) &= 0, \\ y_3' &= \frac{y_1}{\sqrt{y_1^2 + y_2^2}}, & y_3(0) &= 0. \end{aligned} \tag{A.9}$$

B5: (Euler equations of motion for a rigid body without external forces).

$$\begin{aligned} y_1' &= y_2 y_3, & y_1(0) &= 0, \\ y_2' &= -y_1 y_3, & y_2(0) &= 1, \\ y_3' &= -.51 y_1 y_2, & y_3(0) &= 1. \end{aligned} \tag{A.10}$$

Class C: Moderate systems

C1: (a radioactive decay chain).

$$\begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_9' \\ y_{10}' \end{bmatrix} = \begin{bmatrix} -1 & & & & & 0 \\ 1 & -1 & & & & \\ & \ddots & \ddots & & & \\ & & & 1 & -1 & \\ & & & & 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \\ y_{10} \end{bmatrix}, \qquad y(0) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{A.11}$$

C2: (another radioactive decay chain).

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \\ \vdots \\ y_9' \\ y_{10}' \end{bmatrix} = \begin{bmatrix} -1 & & & & & & 0 \\ 1 & -2 & & & & & \\ & 2 & -3 & & & & \\ & & \ddots & \ddots & & & \\ & & & & 8 & -9 & \\ & & & & & 9 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_9 \\ y_{10} \end{bmatrix}, \qquad y(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{A.12}$$

C3: (derived from a parabolic differential equation).

$$\begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \\ y_{10} \end{bmatrix}, \qquad y(0) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{A.13}$$

C4: The same as C3 except with 51 equations.

C5: (The five body problem).

The five body problem models the motion of the 5 outer planets around the sun, assumed in this model to contain the four inner planets. The 3 spatial coordinates of the jth body are y_{1j}, y_{2j}, y_{3j} where $j = 1, 2, ..., 5$. Each satisfy the second order differential equation

$$y''_{ij} = k_2 \left(-(m_0 + m_j)\frac{y_{ij}}{r_j^3} + \sum_{\substack{k=1 \\ k \neq j}}^{5} m_k \left[\frac{y_{ik} - y_{ij}}{d_{jk}^3} - \frac{y_{ik}}{r_k^3} \right] \right), \tag{A.14}$$

where

$$r_j^2 = \sum_{i=1}^{3} y_{ij}^2 \qquad \text{and} \qquad d_{kj}^2 = \sum_{i=1}^{3} (y_{ik} - y_{ij})^2, \qquad k, j = 1, ..., 5.$$

and the physical constants are

$$\begin{aligned}
k_2 &= 2.95912208286 \quad \text{(gravitational constant)} \\
m_0 &= 1.00000597682 \quad \text{(mass of the sun and the 4 inner planets)} \\
m_1 &= 0.00095478610 \quad \text{(Jupiter)} \\
m_2 &= 0.00028558373 \quad \text{(Saturn)} \\
m_3 &= 0.00004372731 \quad \text{(Uranus)} \\
m_4 &= 0.00005177591 \quad \text{(Neptune)} \\
m_5 &= 0.00000277777 \quad \text{(Pluto).}
\end{aligned}$$

When this system of equations is rewritten using first order differential equations the dependent vector has 30 components with initial values

$$\begin{aligned}
y_{11} &= 3.4294741518, & y_{21} &= 3.3538695971, & y_{31} &= 1.3549401715, \\
y'_{11} &= -0.5571605704, & y'_{21} &= 0.5056967832, & y'_{31} &= 0.2305785439, \\
y_{12} &= 6.6414554255, & y_{22} &= 5.9715695787, & y_{32} &= 2.1823149972, \\
y'_{12} &= -0.4155707763, & y'_{22} &= 0.3656827228, & y'_{32} &= 0.1691432132, \\
y_{13} &= 11.2630437207, & y_{23} &= 14.6952576794, & y_{33} &= 6.2796052506, \\
y'_{13} &= -0.3253256691, & y'_{23} &= 0.1897060219, & y'_{33} &= 0.0877265322, \\
y_{14} &= -30.1552268759, & y_{24} &= 1.6569996640, & y_{34} &= 1.4378575272, \\
y'_{14} &= -0.0240476254, & y'_{24} &= -0.2876595326, & y'_{34} &= -0.1172195431, \\
y_{15} &= -21.1238353380, & y_{25} &= 28.4465098142, & y_{35} &= 15.3882659679, \\
y'_{15} &= -0.1768607531, & y'_{25} &= -0.2163934530, & y'_{35} &= -0.0148647893.
\end{aligned}$$

Class D: Orbit equations

$$\begin{aligned} y_1' &= y_3, & y_1(0) &= 1-\epsilon, \\ y_2' &= y_4, & y_2(0) &= 0, \\ y_3' &= \frac{-y_1}{(y_1^2+y_2^2)^{\frac{3}{2}}}, & y_3(0) &= 0, \\ y_4' &= \frac{-y_2}{(y_1^2+y_2^2)^{\frac{3}{2}}}, & y_4(0) &= \sqrt{\frac{1+\epsilon}{1-\epsilon}}. \end{aligned} \tag{A.15}$$

D1: Equation (A.15) with $\epsilon = .1$.

D2: Equation (A.15) with $\epsilon = .3$.

D3: Equation (A.15) with $\epsilon = .5$.

D4: Equation (A.15) with $\epsilon = .7$.

D5: Equation (A.15) with $\epsilon = .9$.

Class E: Higher order equations

E1: (derived from Bessel's equation of order $\frac{1}{2}$ with the origin shifted one unit to the left)

$$\begin{aligned} y_1' &= y_2, & y_1(0) &= J_{\frac{1}{2}}(1) = .6713967071418030, \\ y_2' &= y_1\left(\frac{1}{4(x+1)^2}-1\right) - \frac{y_2}{x+1}, & y_2(0) &= J_{\frac{1}{2}}'(1) = .09540051444747446. \end{aligned} \tag{A.16}$$

E2: (derived from Van der Pol's equation).

$$\begin{aligned} y_1' &= y_2, & y_1(0) &= 2, \\ y_2' &= (1-y_1^2)y_2 - y_1, & y_2(0) &= 0. \end{aligned} \tag{A.17}$$

E3: (derived from Duffing's equation)

$$\begin{aligned} y_1' &= y_2, & y_1(0) &= 0, \\ y_2' &= \frac{y_1^3}{6} - y_1 + 2\sin(2.78535x), & y_2(0) &= 0. \end{aligned} \tag{A.18}$$

E4: (derived from the falling body equation)

$$\begin{aligned} y_1' &= y_2, & y_1(0) &= 30, \\ y_2' &= .032 - .4y_2^2, & y_2(0) &= 0. \end{aligned} \tag{A.19}$$

E5: (derived from a linear pursuit equation)

$$\begin{aligned} y_1' &= y_2, & y_1(0) &= 0, \\ y_2' &= \frac{\sqrt{1+y_2^2}}{(25-x)}, & y_2(0) &= 0. \end{aligned} \tag{A.20}$$

A.2 Stiff problems

A.2.1 Oregonator

The 'Oregantor' is the chemical reaction between $HBrO_2$, Br^- and $Ce(IV)$ [34]. The system of equations is

$$\begin{aligned} y_1' &= 77.27(y_2 + y_1(1 - 8.375 \times 10^{-6} y_1 - y_2)), \\ y_2' &= \frac{1}{77.27}(y_3 - (1 + y_1)y_2), \\ y_3' &= 0.161(y_1 - y_3), \end{aligned}$$

with initial condition $y(0) = (1, 2, 3)$.

A.2.2 HIRES

This problem was first proposed by Schäfer [62] in 1975. It originates from plant physiology and describes how light is involved in morphogenesis. More specifically, it explains the 'High Irradiance Responses' (HIRES) of photomorphogenesis on the basis of phytochrome, by means of a chemical reaction involving eight reactants.

The initial value problem is given by

$$\frac{dy}{dt} = f(y) \qquad y(0) = y_0,$$

where

$$f(y) = \begin{pmatrix} -1.71y_1 + 0.43y_2 + 8.32y_3 + 0.0007 \\ 1.71y_1 - 8.75y_2 \\ -10.03y_3 + 0.43y_4 + 0.035y_5 \\ 8.32y_2 + 1.71y_3 - 1.12y_4 \\ -1.745y_5 + 0.43y_6 + 0.43y_7 \\ -280y_6y_8 + 0.69y_4 + 1.71y_5 - 0.43y_6 + 0.69y_7 \\ 280y_6y_8 - 1.81y_7 \\ -280y_6y_8 + 1.81y_7 \end{pmatrix} \tag{A.21}$$

and

$$y_0 = (1, 0, 0, 0, 0, 0, 0, 0.0057)^T.$$

A.2.3 Prothero-Robinson problem

The Prothero and Robinson problem [58] takes the form

$$y'(x) = L(y - \phi(x)) + \phi'(x), \quad y_0 = y(x_0) = \phi(x_0),$$

where $\mathrm{Re}(L) \leq 0$. It has the exact solution $y(x) = \phi(x)$. We choose $\phi(x) = \sin(x)$ and $L = -10^6$, which makes the problem stiff.

A.3 Delay differential equation problems

These problems have been taken from [57]. The numbering of the equations is the same as in this paper.

A.3.1 Equation 1.1.6

$$\begin{aligned} y'(t) &= -y(t-1) + y(t-2) - y(t-3)y(t-4), \quad t \geq 0, \\ Y(t) &= 1, \qquad\qquad\qquad\qquad\qquad\qquad\quad t < 0, \\ Y(0) &= 0. \end{aligned} \tag{A.22}$$

The analytical solution is

$$y(t) = \begin{cases} -t, & 0 \leq t \leq 1, \\ \frac{1}{2}t^2 - t - \frac{1}{2}, & 1 \leq t \leq 2, \\ -\frac{1}{6}t^3 + \frac{1}{2}t^2 - \frac{7}{6}, & 2 \leq t \leq 3, \\ \frac{1}{24}t^4 - \frac{1}{6}t^3 - \frac{1}{4}t^2 + t - \frac{19}{24}, & 3 \leq t \leq 4, \\ -\frac{1}{120}t^5 + \frac{1}{6}t^4 - \frac{5}{3}t^3 + \frac{109}{12}t^2 - 24t + \frac{2689}{120}, & 4 \leq t \leq 5. \end{cases}$$

This problem originally comes from [56]. It has a zeroth-order discontinuity at $t = 0$ and an n-th order discontinuity at $t = \{4n-3, 4n-2, 4n-1, 4n\}$ for $n \geq 1$. The equation is linear up to $t = 4$ and non-linear beyond.

A.3.2 Equation 1.1.10

$$\begin{aligned} y'(t) &= y(t-\pi)y(t), \qquad\qquad t \geq 0, \\ Y(t) &= \begin{cases} 0 & t < -\frac{\pi}{2}, \\ -2 & -\frac{\pi}{2} \leq t < 0, \\ -1 & t = 0. \end{cases} \end{aligned} \tag{A.23}$$

The analytical solution is

$$y(t) = \begin{cases} -1, & 0 \le t \le \frac{\pi}{2}, \\ -\exp(\pi - 2t), & \frac{\pi}{2} \le t \le \pi, \\ -\exp(-t), & \pi \le t \le \frac{3\pi}{2}, \\ -\exp\left(-\frac{3}{2}\pi + \frac{1}{2}(\exp(3\pi - 2t) - 1)\right), & \frac{3\pi}{2} \le t \le 6. \end{cases}$$

This problem originally comes from [64]. There is an n-th order discontinuity at $t = \left\{ \frac{(2n-1)\pi}{2}, n\pi \right\}$. The problem also has a discontinuous initial function.

A.3.3 Equation 1.1.12

$$\begin{aligned} y'(t) &= y(t) + y(t-1), & t \ge 0, \\ Y(t) &= \begin{cases} 0, & -1 \le t < -\frac{1}{3}, \\ 1, & -\frac{1}{3} \le t \le 0. \end{cases} \end{aligned} \tag{A.24}$$

The analytical solution is

$$y(t) = \begin{cases} \exp(t), & 0 \le t \le \frac{2}{3}, \\ c_1 \exp(t) - 1, & \frac{2}{3} \le t \le 1, \\ t\exp(t-1) + c_2 \exp(t), & 1 \le t \le \frac{5}{2}, \\ 1 + c_1 t \exp(t-1) + c_3 \exp(t), & \frac{5}{3} \le t \le 2, \\ (\frac{1}{2}t^2 - t)\exp(t-2) + c_2 t \exp(t-1) + c_4 \exp(t), & 2 \le t \le \frac{8}{3}, \end{cases}$$

where $c_1 = 1 + \exp(-\frac{2}{3})$, $c_2 = c_1 - 2e^{-1}$, $c_3 = \frac{5}{3}e^{-1}(1 - c_1) + c_2 - \exp(-\frac{5}{3})$ and $c_4 = e^{-2} + c_3 + 2(c_1 - c_2)e^{-1}$.

This problem originally comes from [45]. It is a version of the linear stability DDE test equation, but with a discontinuous initial function. It has an $(n+1)$-st order discontinuity at $t = n, n + \frac{2}{3}$.

A.3.4 Equation 1.4.1

$$\begin{aligned} y_1'(t) &= y_1(t-1) + y_2(t), & t \ge 0, \\ y_2'(t) &= y_1(t) - y_1(t-1), & t \ge 0, \\ Y_1(t) &= e^t, & t \le 0, \\ Y_2(0) &= 1 - e^{-1}. \end{aligned} \tag{A.25}$$

The analytical solution is

$$\begin{aligned} y_1(t) &= e^t, & t \ge 0, \\ y_2(t) &= e^t - \exp(t-1), & t \ge 0. \end{aligned}$$

This problem originally comes from [56]. This system is equivalent to solving a scalar integro-differential equation.

A.3.5 Equation 1.4.6

$$\begin{aligned} y_1'(t) &= y_2(t), & t \geq 0, \\ y_2'(t) &= -y_1(t) - y_2(t-1), & t \geq 0, \\ Y(t) &= [0, \sin(2\pi t)]^T, & t \leq 0. \end{aligned} \tag{A.26}$$

The analytical solution is

$$y_1(t) = \begin{cases} \frac{1}{4\pi^2-1}(\sin(2\pi t) - 2\pi \sin(t)), & 0 \leq t \leq 1, \\ \frac{2\pi}{4\pi^2-1}(\frac{1}{2}(t+1)\sin(t-1) - \sin(t) + \frac{1}{4\pi^2-1}(\cos(2\pi t) - \cos(t-1))), & 1 \leq t \leq 2, \end{cases}$$

$$y_2(t) = \begin{cases} \frac{2\pi}{4\pi^2-1}(2\cos(\pi t)^2 - \cos(t) - 1), & 0 \leq t \leq 1, \\ \frac{\pi}{4\pi^2-1}(\sin(t-1) + (t+1)\cos(t-1) - 2\cos(t) + \frac{2\sin(t-1) - 4\pi \sin(2\pi t)}{4\pi^2-1}), & 1 \leq t \leq 2. \end{cases}$$

This system is equivalent to the second-order scalar DDE which appears in [2]. It has a first order discontinuity at $t = 0$ and a $(n+2)$-nd order discontinuity at $t = n$ in $y_1(t)$ for $n \geq 1$. It also has a first-order discontinuity at $t = 0$ and a $(n+1)$-st order discontinuity at $t = n$ in $y_2(t)$ for $n \geq 1$.

A.3.6 Equation 1.4.9

$$\begin{aligned} y_1'(t) &= y_3(t), & t \geq 0, \\ y_2'(t) &= y_4(t), & t \geq 0, \\ y_3'(t) &= -2m y_2(t) + (1+m^2)(-1)^m y_1(t-\pi), & t \geq 0, \\ y_4'(t) &= -2m y_1(t) + (1+m^2)(-1)^m y_2(t-\pi), & t \geq 0, \\ Y_1(t) &= \sin(t)\cos(mt), & t \leq 0, \\ Y_2(t) &= \cos(t)\sin(mt), & t \leq 0, \\ Y_3(t) &= \cos(t)\cos(mt) - m\sin(t)\sin(mt), & t \leq 0, \\ Y_4(t) &= m\cos(t)\cos(mt) - \sin(t)\sin(mt), & t \leq 0, \end{aligned} \tag{A.27}$$

where we have chosen $m = 2$. The analytical solution is

$$\begin{aligned} y_1(t) &= \sin(t)\cos(mt), & t \geq 0, \\ y_2(t) &= \cos(t)\sin(mt), & t \geq 0, \\ y_3(t) &= \cos(t)\cos(mt) - m\sin(t)\sin(mt), & t \geq 0, \\ y_4(t) &= m\cos(t)\cos(mt) - \sin(t)\sin(mt), & t \geq 0. \end{aligned}$$

This system of equations originally appears in [45]. The analytical solution is a continuation of the initial function.

References

> The simplest schoolboy is now familiar with facts for which Archimedes would have sacrificed his life.
>
> Ernest Renan

[1] F. Bashforth and J. C. Adams, *An attempt to Test the Theories of Capillary Action by Comparing the Theoretical and Measured Forms of Drops of Fluid, with an Explanation of the Method of Integration Employed in Constructing the Tables which Give the Theoretical Forms of Such Drops*, Cambridge University Press, Cambridge (1883).

[2] H. T. Banks and F. Kappel, *Spline Approximations for Functional Diff. Eqns.*, J. Differential Equations **34** (1979), 496–522.

[3] J. C. Butcher, *Coefficients for the study of Runge–Kutta integration processes*, J. Aust. Math. Soc. **3** (1963), 185–201.

[4] J. C. Butcher, *On the convergence of numerical solutions of ordinary differential equations*, Math. Comp. **20** (1966), 1–10.

[5] J. C. Butcher, *An algebraic theory of integration methods*, Math. Comp. **26** (1972), 79–106.

[6] J. C. Butcher, *The numerical analysis of ordinary differential equations:* Runge–Kutta and general linear methods, John Wiley & Sons, Chichester, New York, 1987.

[7] J. C. Butcher, *Diagonally implicit multistage integration methods*, Appl. Numer. Math. **11** (1993), 347–363.

[8] J. C. Butcher, *An introduction to DIMSIMs*, Comput. Appl. Math. **14** (1995), 59–72.

[9] J. C. Butcher, *On fifth order Runge–Kutta methods*, BIT **35**, (1995), 202–209.

[10] J. C. Butcher, *An introduction to "Almost Runge–Kutta" methods*, Appl. Numer. Math., **24** (1997), 331–342.

[11] J. C. Butcher, *ARK methods up to order five*, Numer. Algorithms, **17** (1998), 193–221.

[12] J. C. Butcher, *Order and effective order*, Appl. Numer. Math., **28** (1998), 179–191.

[13] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, J. Wiley, Chichester, 2003.

[14] J. C. Butcher and T. M. H. Chan, *Multi-step zero approximations for stepsize control*, Appl. Numer. Math., **34** (2000), 167–177.

[15] J. C. Butcher and D. J. L. Chen, *ESIRK methods and variable stepsize*, Appl. Numer. Math., **28** (1998), 193–207.

[16] J. C. Butcher and M. Diamantakis, *DESIRE: diagonally extended singly implicit Runge–Kutta effective order methods*, Numer. Algorithms **17** (1998), 121–145.

[17] J. C. Butcher and N. Moir, *Experiments with a new fifth order method*, Numer. Algorithms, **33** (2003), 137–151.

[18] J. C. Butcher and Z. Jackiewicz, *Diagonally implicit general linear methods for ordinary differential equations*, BIT **33** (1993), 452–472.

[19] J. C. Butcher and Z. Jackiewicz, *Construction of diagonally implicit general linear methods of type 1 and 2 for ordinary differential equations*, Appl. Numer. Math., **21** (1996), 385–415.

[20] J. C. Butcher and N. Rattenbury, *ARK methods for stiff problems*, Appl. Numer. Math., **53** (2005), 165–181.

[21] J. C. Butcher and G. Wanner, *Runge–Kutta methods: some historical notes*, Appl. Numer. Math. **22** (1996), 113–151.

[22] J. C. Butcher and W. M. Wright, *A transformation relating explicit and diagonally-implicit general linear methods*, Appl. Numer. Math. **44** (2003), 313–327.

[23] J. C. Butcher and W. M. Wright, *The construction of practical general linear methods*, BIT **43** (2003), 695–721.

[24] G. D. Byrne and R. J. Lambert, *Pseudo Runge–Kutta methods involving two points*, J. Assoc. Comput. Mach. **13** (1966), 114–123.

[25] T. M. Chan, *Algebraic structures for the analysis of numerical methods*, Ph.D thesis, The University of Auckland, Department of Mathematics, 1998.

[26] P. Chartier, *The Potential of Parallel Multi-Value Methods for the Simulation of Large Real-life Problems*, CWI Quart., **11(1)** (1998), 7-32.

[27] C. F. Curtiss and J. O. Hirschfelder, *Integration of stiff equations,* Proc. Nat. Acad. Sci., **38** (1952), 235-243.

[28] G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand., **4** (1956), 33–53.

[29] J. R. Dormand and P. J. Prince, *A family of embedded Runge–Kutta formulae*, J. Comput. Appl. Math., **6** (1980), 19–26.

[30] R. D. Driver, *Ordinary and Delay Differential Equations*, Springer-Verlag, New York (1977).

[31] L. Euler, *De integratione aequationum differentialium per approximationem*, In Opera Omnia, 1st series, Vol. 11, Institutiones Calculi Integralis, Teubner, Leipzig and Berlin, 424–434, (1913).

[32] E. Fehlberg, *Classical fifth, sixth, seventh and eighth order Runge–Kutta formulas with stepsize control*, NASA TR R-287, (1968).

[33] E. Fehlberg, *Klassiche Runge–Kutta-Formeln fünfter und siebenter Ordnung mit Schrittweiten-Kontrolle*, Computing, **4** (1969), 93–106.

[34] J. R. Field and R. M. Noyes, *Oscillations in chemical systems. IV. Limit cycle behaviour in a model of a real chemical reaction*, J. Chem. Physics, **60** (1974), 1877-1884.

[35] C. W. Gear, *Numerical initial value problems in ordinary differential equations*, Prentice-Hall, (1971).

[36] C. W. Gear, *The Automatic Integration of Ordinary Differential Equations*, Commun. of ACM, **14** (1971), 176–179.

[37] C. W. Gear, *Runge–Kutta starters for multistep methods*, ACM. Trans. Math. Software, **6** (1980), 263–279.

[38] E. Hairer, S. P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I: Non-stiff problems*, Springer-Verlag, (2000).

[39] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Equations*, Springer-Verlag, (1991).

[40] K. Heun, *Neue Methoden zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Z. Math. Phys., **45** (1900), 23–38.

[41] R. Hügel, *Numerischer Vergleich von Programmen zur Lösung von Delay Gleichungen*, 5/85 N, Westfalische Wilhelms-Universitat, Münxter, West Germany (1985).

[42] T. E. Hull, W. H. Enright, B. M. Fellen and A. E. Sedgwick, *Comparing numerical methods for ordinary differential equations*, SIAM J. Numer. Anal, **9** (1972), 603–637.

[43] A. Huťa, *Une amélioration de la méthode de Runge–Kutta-Nyström pour la résolution numérique des équations différentielles du premier ordre*, Acta. Fac. Nat. Univ. Comenian. Math., **1** (1956), 201–224.

[44] A. Huťa, *Contribution à la formule de sixième ordre dans la méthode de Runge–Kutta-Nyström*, Acta Fac. Nat. Univ. Comenian. Math., **2** (1957), 21–24.

[45] K. Ito, H. T. Tran and A. Manitius, *A Fully-Discrete Spectral Method for Delay Diff. Eqns.*, SIAM J. Numer. Anal. **28** (1991), 1121–1140.

[46] Z. Jackiewicz, R. Renaut and A. Feldstein, *Two-step Runge–Kutta methods*, SIAM J. Numer. Anal. **28** (1991), 1165–1182.

[47] U. Kirchgraber, *Multistep Methods are Essentially One-step Methods*, Numer. Math., **48** (1986), 85–90.

[48] W. Kutta, *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Z. Math. Phys., **46** (1901), 435–453.

[49] R. H. Merson, *An operational method for the study of integration processes*, Proc. Symp. Data Processing, (1957), 1–25.

[50] W. E. Milne, *A note on the numerical integration of differential equations*, J. Res. Nat. Bur. Stand., **43** (1949), 537–542.

[51] N. Moir, *ARK methods: some recent developments*, J. Comput. Appl. Math., to appear.

[52] F. R. Moulton, *New Methods in Exterior Ballistics*, University of Chicago Press (1926).

[53] A. Nordseick, *On numerical integration of ordinary differential equations*, Math. Comp., **16** (1962), 22–49.

[54] E. J. Nyström, *Über die numerische Integration von Differentialgleichungen*, Acta Coc. Sci. Fenn., **50** (1925), 1–54.

[55] B. Owren and M. Zennaro, *Derivation of efficient, continuous, explicit Runge–Kutta methods*, SIAM J. Sci. Statist. Comput., **13** (1992), 1488–1501.

[56] C. A. H. Paul, *Concerning Explicit Runge–Kutta Techniques for Delay Diff. Eqns.*, MSc Thesis, Math. Dept., Manchester University (1989).

[57] C. A. H. Paul, *A Test Set of Functional Differential Equations*, Numerical Analysis Report No. 243, The University of Manchester (1994).

[58] A. Prothero and A. Robinson, *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*, Math. Comp., **28** (1974), 145–162.

[59] L. F. Richardson, *The deferred approach to the limit*, Philos. Trans. Roy. Soc. London, **Ser. A**, 299–361 (1927).

[60] H. H. Robertson, *The solution of a set of reaction rate equations*, In Numerical Analysis, An Introduction, J. Walsh (Ed.), Academ. Press, (1966), 178–182.

[61] C. Runge, *Über die numerische Auflösung von Differentialgleichungen*, Math. Ann., **46** (1895), 167–178.

[62] E. Schäfer, *A new approach to explain the 'high irradiance responses' of photomorphogenesis on the basis of phytochrome*, J. Math. Biol., **2** (1975), 41–56.

[63] D. Stoffer, *General linear methods: connection to one step methods and invariant curves*, Numer. Math., **64** (1993), 395–407.

[64] L. Tavernini, *CTMS User Guide*, Math. Div., Comp. Sci. and Systems Design, Univ. of Texas at San Antonio, Texas (1987).

[65] B. van der Pol, *On relaxation-oscillations*, Philos. Mag. Ser. 7, **2** (1926), 978–992.

[66] J. H. Verner, *Explicit Runge–Kutta methods with estimates of the local truncation error*, SIAM J. Numer. Anal., **15** (1978), 772–790.

[67] W. M. Wright, *General linear methods with inherent Runge–Kutta stability*, PhD thesis, The University of Auckland, Department of Mathematics, 2002.

[68] W. M. Wright, *Explicit general linear methods with inherent Runge–Kutta stability*, Numer. Algorithms, **31** (2002), 381–399.

Index

A-stability, 13
Adams methods, 6, 24
Anglin, W. S., 1
annihilation conditions, 36, 37

BDF methods, 6, 25
Butcher, J. C., 29

$C(2)$ condition, 36
composition of elementary weight functions, 21
consistency, 11
convergence, 12

$D(1)$ condition, 36
Dantzig, 133
DDEs, *see* delay differential equations
delay differential equations, 5, 125
density of a tree, 14
derivation of methods, 43, 51, 60, 67, 100, 103
derivative operator, 21
DETest, 109, 117, 123, 133
DIMSIMs, 26
Dormand and Prince, 111, 123, 125

E-polynomial, 98, 101
elementary differentials, 18
elementary weights, 19
Erdös, P., 129
error estimators, 7, 82
Euler's method, 6
existence and uniqueness, 2

general linear methods, 9
generating functions, 22, 37

Hermite interpolation, 47
HIRES, 138

Inherent Runge–Kutta stability, 29
initial value problem, 2
interpolation, 39, 46, 57, 63, 70, 81, 125
    Hermite, 47
IRKS methods, 29

L-stability, 13
linear multistep methods, 6
    Adams methods, 6, 24
    as general linear methods, 23
    BDF methods, 6, 25
Lipschitz condition, 2
    one-sided, 4

optimisation, 83
order, 16
order conditions, 36, 43, 49, 60, 66, 99, 103
order of a tree, 14
ordinary differential equations, 2
Oregonator, 138

Plato, 9
Polyá, G., 73
population growth, 5
preconsistency, 10
Prothero-Robinson problem, 139

Renan, Ernest, 143
RK stability, 38, 39, 58
Runge–Kutta methods, 7
    as general linear methods, 23

simplifying assumptions, 36
stability, 11
    A-stability, 13
    L-stability, 13
stability function, 12
stability matrix, 12
stability region, 12
starting procedures, 16, 35, 105
stiff ARK methods, 93
stiff differential equations, 3, 93
symmetry of a tree, 16

trees, 14
    density, 14
    elementary differentials, 18
    elementary weights, 19
    order, 14
    symmetry, 16

variable stepsize, 36, 80, 123

Watson, Thomas, 109
Wright, W. M., 29