Accurate and Efficient Methods for Differential Systems with Special Structures

Gulshad Imran

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Applied Mathematics,
The University of Auckland, 2014.

Abstract

The numerical solution of differential equations is important in many branches of science, and accurate and stable algorithms are needed. The derivation and analysis of these methods is quite sophisticated and uses intricate mathematical tools such as **B**-series and Lie Algebras and Lie Groups. Our aim is to develop new numerical methods, and new methods of analysis, for Hamiltonian problems and other problems for which good qualitative behaviour is essential.

The starting point in this thesis is a new class of effective order five methods based on the algebraic structure associated with Runge–Kutta methods. This leads us to another class of symplectic effective order methods which are designed for the solution of differential equations with quadratic invariants.

G-symplectic general linear methods, for which parasitic growth factors are zero, are a possible alternative to symplectic Runge–Kutta methods. They have a similar ability to preserve quadratic invariants over extended time intervals. They have lower implementation costs and are designed to approximately preserve the quadratic invariants of the Hamiltonian systems. In this thesis, G-symplectic general linear methods are investigated theoretically and computationally. As for Runge–Kutta methods, there is an interaction between order conditions and the symplectic conditions, resulting in significant simplifications. To evaluate the order conditions for the G-symplectic general linear methods and the possible extension to any order is presented in this thesis.

Finally the construction of sixth order G-symplectic general linear method with the additional properties of time reversal symmetry and freedom from parasitism is presented.

Dedication

To my loving husband, Dr.Imran K. Khan

and

my lovely daughters, Aliza and Shanzay

Acknowledgements

All praises are for Almighty Allah, the most Merciful and most Benevolent, Whose blessings are always upon me.

First of all, I would like to express gratitude and respect for my supervisor, Emeritus Professor John Butcher, who taught me how to do good mathematics. He taught me how to think about research problems, how to write precisely and concisely and even helped me in correcting my punctuation and grammatical mistakes. His outstanding contributions to Numerical Methods have always been a source of ideas and inspiration for me. I must want to appreciate him for his easy availability and prompt feedback. He was also kind enough to support me financially via Marsden Fund and to organise many illuminating trips to top class conferences and introduced me to his circle of outstanding researchers.

I would extend my special thanks to Prof Rod Gover for being my official supervisor during my last year of research studies. I also wish to thank my co-supervisor Dr. Robert Chan for his encouragement and always made himself available for discussions whenever I wished. My special thanks go to Dr. Allison Heard for her valuable suggestions on my initial draft.

I would also like to thank the members of the Numerical Runge–Kutta group in Auckland especially; Dr. Robert Chan, Dr. Shixiao Wang, Dr Allison Heard, Dr. Annie Gorgey, Dr. Angela Tsai, Dr. Yousaf Habib, Noorhelyna Razali and Saghir Ahmad, for their insight and friendship.

I am grateful to the Department of Mathematics for providing me a wonderful research environment and for helping me to attend national and international conferences through PReSS Account and PBRF grant. I would also like to thank administrative staff of the Mathematics Department especially Adina Nagy, Olita Moala, Lynda Pitcaithly, Min-Ah Lee and Alexandra Dumitrescu for helping me to sort out various administrative tasks efficiently and timely.

Finally, my family has been an endless source of love, affection, support and motivation for me. Dear Ami and Abu I have no words to thank you for all your love, prayers and encouragement. Dear Abu, it is because of your wonderful lessons in mathematics in my childhood that enabled me to come this far. I would also like to thank my sisters for their sincere love. I am deeply indebted to my loving and caring husband Dr. Imran Khaliq. I thank you for all your love and constant support, for keeping me sane over the past few months, and for taking care of Shanzay when she was just few days old. Thank you for accompanying me on this adventure.

A very special thanks go to my lovely daughters Aliza and Shanzay for their love and patience. Each word that I have written represents time spent away from them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Ordinary differential equations (ODEs) are used to describe motion and change in many fields of mathematics and sciences. To model physical systems like infectious diseases in biology, weather modelling in geophysics, stock trends and the market equilibrium prices changes in economics, population growth and radioactive decay, all are studied through their mathematical models. We need to solve differential equations to predict the development in a natural process for these models.

Usually the solution to a problem is not defined by a differential equation alone but additional information is given in the form of an *initial value*. Thus an *initial value problem* takes the form

$$y' = f(x, y), \qquad y(x_0) = y_0. \tag{1.1}$$

In this initial value problem(IVP), x is refereed to as the independent variable and y as the dependent variable. Generally, we are interested in *autonomous* IVPs of the form

$$y' = f(y(x)), \qquad y(x_0) = y_0. \tag{1.2}$$

In a high dimensional setting, (1.1) and (1.2) are not different because (1.1) can be rewritten in the form of (1.2) by introducing an additional component of the y vector which is always equals to x. Physical systems are modelled using initial value systems. We want to deal with problems for which the solution to initial value problem exists and is unique. We also want to focus on problems for which the solution depends smoothly on the initial data. The existence, uniqueness and smoothness are concerned with the continuity of the function f. We want the function f to satisfy the Lipschitz condition [19].

Definition 1.0.1. *The function f satisfies a Lipschitz condition (f is Lipschitz continuous) if for some constant L,*

$$\|f(\mu) - f(\nu)\| \leq L\|\mu - \nu\|.$$

The constant L is known as the Lipschitz constant.

Theorem 1.0.2. *If f satisfies a Lipschitz condition with constant L then the initial value problem*

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

possesses a unique solution on any interval $[x_0, \hat{x}]$, furthermore the solution depends continuously on y_0.

We assumed that the function $f(y(x))$ satisfies the Lipschitz condition, and therefore a unique solution exists to the initial value problems. In [33] a special phenomenon for initial value problems (IVPs) was discovered by Curtiss and Hirschfelder, which is now known as *stiffness*. The concept of stiffness became recognised as a significant factor in numerical computation because of strange results arising in some physical problems. Since then differential equations have been divided into stiff and non-stiff problems.

Numerical methods can be characterised as one-step methods, multistep methods and multistage methods. The famous method of Euler was the first one in the family of one step methods. The basic idea of generating this method is very simple. In a short period of time, x_0 to x_1, how much should we expect y to change? We assume that at the beginning of the time step, $f(x_0, y_0)$ is the velocity(rate of increase). Thus the change in position will be equal to

$$y_1 = y_0 + (x_1 - x_0)f(x_0, y_0).$$

We keep on doing this over many steps to get the sequence of approximations y_2, y_3, ... Thus at the n^{th} step we have

$$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1}),$$

where $x_n = x_0 + nh$ and h is the step size. It is first order only and therefore has very low accuracy. This idea of Euler is the starting point for understanding all numerical methods for differential equations.

Now, for higher order one–step methods, it is possible to get better accuracy by

evaluating the function f more than once in the integration interval. This gives the idea of so called, *Runge–Kutta methods*. The first methods of Runge [65], Heun [51] and Kutta [54] generalise the classical Euler method by allowing multiple function evaluation in each time step.

$$Y_i = y_{n-1} + \sum_{j=1}^{s} a_{ij} h f(Y_j), \qquad i = 1, 2, \cdots, s,$$
$$y_n = y_{n-1} + \sum_{i=1}^{s} b_i h f(Y_i),$$

where (c_i, a_{ij}, b_i) are referred to coefficients of the method. The Y_i are the internal stages and are approximations to $y(x_{n-1} + c_i h)$. In [44] Gill pointed out the way to a full analysis of the order conditions by comparing the numerical solution with the exact solution for any autonomous initial value problem. Butcher [8] proposed that the numerical solution could be presented by using Taylor series in terms of what are known as *elementary differentials*. The number of order conditions for a specific method can be reduced by using simplifying assumptions. These assumptions consist of a set of conditions, which when obeyed by the method, reduce the number of conditions to attain a certain order. This idea was presented in [8].

In 1972, Butcher [13] presented a very important paper that was on the structure of algebraic analysis of integration methods. Later Hairer and Wanner [49] studied the concepts developed in [13] and named it as **B**-series with a slight difference in presentation. The algebraic concepts of Runge–Kutta methods were developed in [13] and [49].

Implicit Runge–Kutta methods based on Gauss, Radau and Lobatto quadratures were introduced by Butcher in [9]. Gauss methods have the highest possible order for s stages and are A-stable. Whereas the Radau and Lobatto methods are of order $2s-1$ and $2s-2$ respectively and are not A-stable. The family of A-stable Radau IA, Radau IIA, Lobatto IIIA, Lobatto IIIB and Lobatto IIIC methods were introduced by Ehle [40] in 1969. The Radau methods are very successful for solving differential algebraic equations and stiff problems. All these methods use simplifying assumptions to reduce the order conditions. But both methods are very expensive to implement and both can suffer from order reduction. However for an s stage implicit method the most expensive part of the calculation is solving stages and stage derivatives, which are sN dimensional vectors. This means that the Newton integration schemes require $O(s^3N^3)+O(s^2N^2)$ operations at every integration step. This cost increases rapidly as s and N increases.

Diagonally implicit Runge–Kutta methods were introduced in [1] for lowering the computational cost of implicit methods. In these methods coefficient matrix A has a lower triangular structure and therefore the non linear system of equations can be solved sequentially. These are the least expensive methods because the total cost reduces to $O(sN^3) + O(sN^2)$, but they have low stage order.

The concept of singularly implicit Runge–Kutta methods was developed by Nørsett in [63], and the coefficient matrix of these methods has only a single eigenvalue. These methods also have low cost for many problems and have high stage order. In [4] Burrage introduced error estimates for these methods. In 1979 Butcher [15] introduced the similarity transformation and showed that the matrix A can be transformed to Jorden canonical form to reduce the computational cost. In 1980, Burrage, Butcher and Chipman in [7] developed the variable stepsize variable order code. But the transformation increases the computational cost to $O(s^2N)$ at every integration step and this extra cost makes these methods less competitive.

The surprising idea of *effective order* was introduced by Butcher [12], where a Runge–Kutta method α is said to have effective order p if there exists another method β such that $\beta\alpha\beta^{-1}$ has order p. Thus by introducing the starting method, the main method α has more freedom. Effective order was generalised to effective order singly implicit methods by Butcher and Chartier [20].

Linear multistep methods use the idea that after the first few steps have been taken, it becomes possible to use information computed in past steps to get more accurate answers. A general form of a k-step linear multistep method is

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

In the special case when $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = \ldots = \alpha_k = 0$, we obtain Adams–Bashforth methods. Adams–Moulton methods are obtained if instead we assume that $\beta_0 \neq 0$ to attain order $k + 1$. These methods were developed by Moulton [61]. Other special types of linear multisteps methods were developed by Nyström in [64] and Milne in [59]. Adams–Moulton methods are always more accurate than Adams–Bashforth methods of the same order and to obtain the same accuracy as Adams–Bashforth methods, Adams–Moulton methods require less past information.

While implementing Adams methods in a recursive way, some serious difficulties arise. These involve finding the right stepsize, obtaining the starting values, estimating the error, changing step size and order and deciding when to change order. Using Predictor Corrector (PC) methods [60], was found to be most successful in the implementation of Adams methods. This means that first finding n approximations

to $y_n^* \approx y(x_n)$ by using p-step Adams–Bashforth methods of order p, then calculating $f_n^* = f(x_n, y_n^*)$, and finally correcting y_n using the order p Adams-Moulton $(p-1)$-step method but with f_n replaced by f_n^*. These methods are often expressed as PEC or $PECE$ methods. For progress in numerical computation, it is required that codes of modern differential equations adapt their behaviour automatically. The central idea is to estimate the new error that is introduced in every time step. This can be achieved in Predictor Corrector (PC) methods, and is known as *Milne's device*.

General linear methods were introduced in [11] as a unifying framework for studying basic questions of consistency, stability and convergence for a wide range of traditional methods. Classical Runge–Kutta methods and linear multistep methods are two traditional classes of numerical methods which have always been studied separately. But general linear methods combine the flavours of these two extreme classes of numerical methods such as hybrid, generalised multistep and modified multistep methods were developed by Gear [43], Gragg and Stetter [45] and Butcher [10] respectively.

General linear methods are both multistage and multivalue and are designed for the numerical solution of differential equations. Formulation of general linear methods was introduced by Burrage and Butcher [6], and this is now the standard way of representing these methods. A general linear method has a collection of r input approximations to step number n and at the end of the step it produces r output approximations. During the calculation s stage values are evaluated together with s stage derivative approximations. That is

$$y^{[n-1]} = \begin{bmatrix} y_1^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{bmatrix}, \quad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_s \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ \vdots \\ F_s \end{bmatrix}$$

These are related by the equations of the form

$$\begin{aligned} Y &= h(A \otimes I)F + (U \otimes I)y^{[n-1]}, \qquad F_i = f(Y_i) \\ y^{[n]} &= h(B \otimes I)F + (V \otimes I)y^{[n-1]}, \end{aligned}$$

where the stage value vector Y and the stage derivative vector F are each composed of s subvectors related by $F_i = f(Y_i)$, $i = 1, 2, \ldots, s$. The four coefficient matrices

(A, U, B, V) characterise a specific general linear method.

$$Y_i = h\sum_{j=1}^{s} a_{ij}F_j + \sum_{j=1}^{r} u_{ij}y_j^{[n-1]}, \qquad i = 1, 2, \ldots, s,$$
$$y_i^{[n]} = h\sum_{j=1}^{s} b_{ij}F_j + \sum_{j=1}^{r} v_{ij}y_j^{[n-1]}, \qquad i = 1, 2, \ldots, r.$$

To define the order of a general linear method, usually we require a *starting procedure*, which computes the possible collection of r inputs. These are the values the method is designed to approximate. Once these are available, we calculate stages and hence the output vectors.

From the detailed discussion above, it is well understood that methods with good quantitative behaviour are always desirable. For instance, the coefficient matrix A should be lower triangular to save the cost of the diagonally implicit Runge–Kutta methods. For stiff problems it is necessary that the stage order should be equal to overall order to get the high accuracy. However the qualitative behaviour is also important, and this is another question which is also explored in this thesis too.
Particular physical systems have properties which needs to be preserved. Although the traditional numerical methods do not preserve these properties, we are interested in getting numerical methods that respect qualitative and geometric features of the numerical ODEs, together with the quantitative properties. These numerical methods are known as *canonical methods or structure preserving methods* and they are best known for long term integration. Here we summarise some of the qualitative properties that a numerical method could preserve.

Many physical systems have invariants such as conservation of energy and conservation of momentum. Usually a differential equation and the physical laws they obey determine suitable geometric integrators. For instance, we use Lie group methods [52] for differential equations of the type that evolves on Lie groups. Symmetry preserving methods [47] can be use for the differential equations of the type that respect symmetry. In addition to conservation of energy and angular momentum, Hamiltonian systems also possess what is known as the symplectic property. By symplecticity we mean that the variational equations preserve quadratic invariants. The integrators that have the symplectic property are known as symplectic integrators. Below is a detailed discussion on Hamiltonian systems.

1.1 Importance of Hamiltonian systems

The theory of modern scientific computing is based on the numerical solution of differential equations, for instance, Newton, Euler, Laplace, Lagrange, Maxwell, Navier-Stokes and so on. The most prominent is the Newton's equation of motion and the numerical solution of these equations. On the other hand, starting from Euler to modern computer, there has been a great development in scientific literature of numerical methods of differential equations. With this development, it is understood that all physical systems whether they are quantum, relativistic or classical can be modelled by a Hamiltonian system.

A Hamiltonian system with Hamiltonian function is a special type of ordinary differential equation, which is defined in an even dimensional phase space. By definition it is given by

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \qquad \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \qquad i = 1, \cdots, n, \tag{1.3}$$

where $(p_i, q_i) = \big((p_1, p_2, \cdots, p_n), (q_1, q_2, \cdots, q_n)\big)$ are usually called generalised coordinates and momenta, respectively. The integer n is called the *number of degree of freedom* and (p_i, q_i) constitute the coordinates of *phase space*. The Hamiltonian H usually refers to total energy of the mechanical system and has a special separable structure which is given by

$$H(p, q) = T(p) + V(q),$$

where T corresponds to kinetic energy and V represents the potential energy. If we write $y = (p, q)$, then (1.3) takes the form

$$\frac{dy}{dt} = J^{-1}\nabla H. \tag{1.4}$$

In (1.4), $\nabla = (\frac{\partial}{\partial p_1}, \ldots \frac{\partial}{\partial p_n}, \frac{\partial}{\partial q_1}, \ldots, \frac{\partial}{\partial q_n})$ is the gradient operator and J is the skew symmetric matrix having zeros on the diagonal and units on the off-diagonal, that is

$$J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{1.5}$$

Example 1.1.1. *The Harmonic oscillator*

This is a simple example of a Hamiltonian system with one degree of freedom, in which the motion of a unit mass attached to a spring is described by the differential equations

$$q' = p, \qquad p' = -q. \tag{1.6}$$

The total energy of the Hamiltonian system is given by

$$H = \frac{1}{2}p^2 + \frac{1}{2}q^2.$$

Example 1.1.2. *The Simple pendulum*

We consider a simple pendulum problem, which describes the motion in such a way that the mass of the bob, the length of the rod and acceleration due to gravity all are of unit length. Then the total energy H is given as,

$$\begin{aligned} H &= \text{kinetic energy } + \text{potential energy} \\ &= \frac{p^2}{2} - \cos(q), \end{aligned}$$

whereas, the equations of motion are

$$p' = -\sin(q), \qquad q' = p.$$

1.1.1 The phase flow of a Hamiltonian system

An important feature of the Hamiltonian system is the *symplecticness* of their phase flow, that is the total area is preserved with one degree of freedom. Let us consider that ϕ the phase flow is a transformation mapping of the phase space into itself that is $\phi : (p(0), q(0)) \longmapsto (p(t), q(t))$ or in other words, $\phi((p(0), q(0))) = (p(t), q(t))$ There are many ways to check the symplecticity of a Hamiltonian system, some of which are listed below:

Preservation of area via Jacobians: Let us consider the transformation map ϕ, such that $\phi((p(0), q(0))) = (p'(t), q'(t))$. According to standard theory for transforming variables into integrals, the map ϕ is symplectic if and only if the Jacobian determinant is identity: for all (p, q), we have

$$\frac{\partial p' \partial q'}{\partial p \partial q} - \frac{\partial p' \partial q'}{\partial q \partial p} = I \tag{1.7}$$

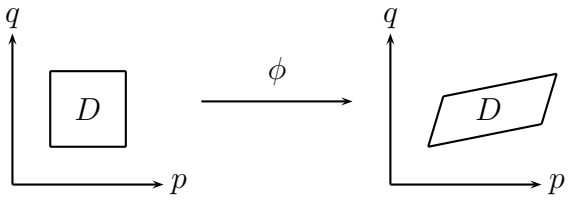


Figure 1.1: Symplectiness in one degree of freedom case

where ϕ and J are given as

$$\phi = \begin{bmatrix} \frac{\partial p'}{\partial p} & \frac{\partial p'}{\partial q} \\ \frac{\partial q'}{\partial p} & \frac{\partial q'}{\partial q} \end{bmatrix}, \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

Thus,

$$\phi^T J \phi = \begin{bmatrix} \frac{\partial p'\partial q'}{\partial p\partial p} - \frac{\partial p'\partial q'}{\partial p\partial p} & \frac{\partial p'\partial q'}{\partial p\partial q} - \frac{\partial p'\partial q'}{\partial q\partial p} \\ \frac{\partial p'\partial q'}{\partial q\partial p} - \frac{\partial p'\partial q'}{\partial p\partial q} & \frac{\partial p'\partial q'}{\partial q\partial q} - \frac{\partial p'\partial q'}{\partial q\partial q} \end{bmatrix} = J$$

Preservation of area via differential forms: We consider differentials dp' and dq' of the components (p', q'). These are actually differential 1-forms, that is

$$\begin{aligned} dp' &= \tfrac{\partial p'}{\partial p}dp + \tfrac{\partial p'}{\partial q}dq, \\ dq' &= \tfrac{\partial q'}{\partial p}dp + \tfrac{\partial q'}{\partial q}dq. \end{aligned}$$

Now, these two dp' and dq' differential 1-forms will provide a differential 2-forms $dp' \wedge dq'$ with the help of exterior product $\wedge$, that is

$$dp' \wedge dq' = \frac{\partial p'}{\partial p}\frac{\partial q'}{\partial p}dp \wedge dp + \frac{\partial p'}{\partial p}\frac{\partial q'}{\partial q}dp \wedge dq + \frac{\partial p'}{\partial q}\frac{\partial q'}{\partial p}dq \wedge dp + \frac{\partial p'}{\partial q}\frac{\partial q'}{\partial q}dq \wedge dq. \tag{1.8}$$

The exterior product is skew symmetric and in particular we have, $dp \wedge dq = -dq \wedge dp$ and $dp \wedge dp = 0 = dq \wedge dq$. Thus (1.8), becomes

$$dp' \wedge dq' = \left(\frac{\partial p'\partial q'}{\partial p\partial q} - \frac{\partial p'\partial q'}{\partial q\partial p}\right) dp \wedge dq \tag{1.9}$$

Therefore by using (1.7) the conservation of area is equivalent to

$$dp' \wedge dq' = dp \wedge dq.$$

1.1.2 Invariants

We consider the possible existence of invariants $\mathcal{I}$ for the initial value problem

$$y'(x) = f(y(x)), \quad x \in \mathbb{R} \qquad y(x_0) = y_0 \in \mathbb{R}^N, \qquad f : \mathbb{R}^N \to \mathbb{R}^N. \tag{1.10}$$

Let $\mathcal{I} : \mathbb{R}^N \to \mathbb{R}$ is called the first integral of (1.10) if $\mathcal{I}'(y)f(y) = 0$.
In particular we are interested in problems having invariants which are quadratic in nature, that is

$$\mathcal{I}(y) = y^{\intercal} Q y,$$

where Q is a real symmetric $N \times N$ matrix.

We consider the Kepler problem, which describes the motion of a planet revolving around the sun which is fixed at the origin. The equations of motion are given by

$$\begin{aligned} y_1' &= -\frac{y_3}{(y_3^2 + y_4^2)^{\frac{3}{2}}}, \\ y_2' &= -\frac{y_4}{(y_3^2 + y_4^2)^{\frac{3}{2}}}, \\ y_3' &= y_1, \\ y_4' &= y_2. \end{aligned}$$

The system has two invariants namely, the *Hamiltonian* and *angular momentum*. These are given by

$$\begin{aligned} H &= \tfrac{1}{2}(y_1^2 + y_2^2) - \frac{1}{\sqrt{y_3^2 + y_y^2}}, \\ A &= y_3 y_2 - y_4 y_1. \end{aligned}$$

1.2 Hamiltonian numerical methods

The Hamiltonian system is one of the most important systems among all dynamical systems, and has broad applications which includes material mechanics, celestial mechanics, partial differential equations, structural biology, superconductivity, and semiconductivity. While doing the numerical integration of these systems, it is absolutely necessary to preserve the Hamiltonian structure which includes symplectic

behaviour of the phase flow, conservation of energy and momentum and conservation of linear and quadratic invariants. So based on this principal we are interested in those numerical methods which respect these properties of Hamiltonian systems. Such methods do exist and are called *symplectic or canonical methods*.

Historically symplectic one–step numerical methods appeared first in the pioneering work of de Vogelaere [76], Ruth [66], and Feng [42]. Sanz–Serna [67], Suris [73] and Lasagni [55], discovered that some implicit Runge–Kutta methods achieve symplecticity through an appropriate choice of coefficients. An interesting feature about these methods was that the order conditions can be expressed in terms of unrooted trees rather than rooted trees.

A Runge–Kutta method is symplectic or canonical if its coefficients satisfy the following relation

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0.$$

We observe the coefficient matrix $M = b_i a_{ij} + b_j a_{ji} - b_i b_j$ is closely connected to the non-linear stability of a Runge–Kutta method. In [5], Burrage and Butcher discovered that the coefficient matrix M is analysed to test that whether the underlying Runge –Kutta method is algebraically stable or not.

Symplectic numerical methods are used for long term integration of Hamiltonian systems. Moreover these methods also preserve the quadratic invariants exclusively.

Not all numerical methods are symplectic, in 1993 Tang proved that the multistep methods are not symplectic because they require more than one initial condition to start the integration process and hence cannot have a transformation map on phase space. Furthermore, Butcher and Hewitt [25], discovered that the multi value methods are not symplectic unless they pass only a single value from a present state to a future state.

General linear methods are multivalue methods and are not practical symplectic methods unless they reduce to Runge–Kutta methods. However, it is possible to achieve the general solution of the symplectic behaviour by making use of the non-linear stability matrix of a general linear method.

$$\widetilde{M} = \begin{bmatrix} DA + A^T D - B^T G B & DU - B^T G V \\ U^T D - V^T G B & G - V^T G V \end{bmatrix}. \tag{1.11}$$

Just as symplectic Runge–Kutta method, the algebraic stability matrix, $M = 0$, so it was quite interesting to analyse the similar behaviour for a general linear method for which $\widetilde{M} = 0$. This gives rise to the notion of *G-symplecticity* introduced by

Butcher in [19] and the first G-symplectic general linear method was discovered. This was a two-stage order four method based on Gaussian quadrature, and is given below.

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{1.12}$$

G-symplectic methods, for which parasitic growth factors are zero, are a possible alternative to symplectic Runge–Kutta methods. They have a similar ability to preserve quadratic invariants over extended time intervals and have lower implementation costs.

1.3 Framework of the thesis

The central aim of this thesis is to develop different theoretical techniques for the study of differential systems. In recent years much interest has developed in structure preserving numerical methods for problems possessing quadratic invariants. The **B**-series approach will be use in deriving such numerical methods in purely algebraic way. The main benefit of using the **B**-series approach is the efficient derivation of order conditions.

The thesis can be summarised as follows:

The basic introduction of numerical methods is given in Chapter 2. We demonstrate the new fifth effective order method in Chapter 3, which breaks the Butcher barrier by weakening the classical order conditions. These methods are based on the $C(2)$ simplifying assumptions. By doing this we are able to reduce the set of seventeen order conditions reduced to only nine.

The effective order leads us to Chapter 4, where the flavours of effective order are combine with symplecticity for the derivation of a new class of symplectic implicit Runge–Kutta methods with effective order four. The beauty of these methods is that these have real eigenvalues which make it possible to implement the method efficiently. The results for this chapter are also presented in [27].

In Chapter 5, we will investigate the order conditions for G-symplectic general linear methods. Some remarkable facts will be discussed regarding the effect of symplectic conditions applied to order conditions not only for the Runge–Kutta methods but

also for general linear methods. A special case of order four methods will be discussed in detail. These methods are specially designed for the long term integration of Hamiltonian problems where it is desirable to close conservation of energy, angular momentum and symplectic behaviour. The results presented in this chapter are available in [28].

In Chapter 6, a sixth order general linear method will be presented with the properties such as G-symplecticity, time reversal symmetry and freedom from parasitism. These methods will be constructed with five stages and four output values. For the matrices B and V complex numbers will be considered such that the second and third rows of the respective matrices will be the complex conjugates. In addition to this the diagonal matrix V will have the eigenvalues on the unit circle. The results derived in this chapter are presented in [29].

Chapter 2

Analysis of numerical methods

Numerical methods are widely used for approximating the exact solution of ordinary differential equations and are classified as one–step, multistep and multivalue methods. These methods are necessary for analysing the behaviour of differential systems. A brief description of traditional numerical methods is given along with recent developments and an overview of the algebraic approach to such methods is studied in this chapter.

2.1 Runge–Kutta methods

Rung–Kutta methods are characterised as one step methods for the solution of initial value problems. An s-stage Runge–Kutta method is defined by

$$\begin{aligned} Y_i &= y_{n-1} + \sum_{j=1}^{s} a_{ij} h f(Y_j), \qquad i = 1, 2, \cdots, s, \\ y_n &= y_{n-1} + \sum_{i=1}^{s} b_i h f(Y_i), \end{aligned} \tag{2.1}$$

where Y_i are internal stages, and y_n are the update values. Alternatively, Runge–Kutta methods can be represented by using the Butcher tableau,

$$R = \begin{array}{c|c} c & A \\ \hline & b^{\intercal} \end{array} = \qquad \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array},$$

we always assume that

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \qquad i = 1, 2, \ldots, s.$$

The structure of the matrix A plays a vital role in the computational cost of the method. Runge–Kutta methods are categorised as follows

- Runge–Kutta methods are *explicit* when the upper triangular components of matrix A are zero that is, $a_{ij} = 0$ for all $i \leq j$. Explicit Runge–Kutta methods requires less computation time and are therefore preferred for the solution of ordinary differential equations. But these methods are not suitable for stiff differential equations because of their bounded regions of absolute stability.
- Runge–Kutta methods are *implicit* when $a_{ij} \neq 0$ for some $i \leq j$. The solution of implicit stage equations can be achieved using a Newton iteration scheme. There are several good reasons to use implicit Runge–Kutta methods, the most important reason being the group structure of these methods. Implicit Runge–Kutta methods are the only hope for the solution of stiff differential equations.

Example 2.1.1.

The classical fourth order Runge–Kutta method is an example of explicit method. The stages and the output vector are given below

$$\begin{aligned} Y_1 &= y_0, \\ Y_2 &= y_0 + \tfrac{1}{2}hF_1, \\ Y_3 &= y_0 + 0hF_1 + \tfrac{1}{2}hF_2, \\ Y_4 &= y_0 + 0hF_1 + 0hF_2 + 1hF_3, \\ y_1 &= y_0 + (\tfrac{1}{6}hF_1 + \tfrac{1}{3}hF_2 + \tfrac{1}{3}hF_3 + \tfrac{1}{6}hF_4). \end{aligned}$$

Where

$$F_1 = f(Y_1),$$
$$F_2 = f(Y_2),$$
$$F_3 = f(Y_3),$$
$$F_4 = f(Y_4).$$

This method is given by the tableau

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

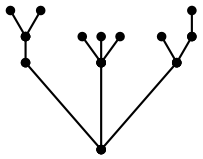
Example 2.1.2.

Famous implicit methods are based on Gauss–Legendre polynomials and known as *Gauss methods*. These methods, introduced by Butcher in [9], have the highest possible order for a given number of stages. The Gauss method with two stages and order four is given below.

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{2.2}$$

2.1.1 Trees and rooted trees

Let V be a set of *vertices* and E a set of *edges* (where every element of E is a pair of members of V). The pair (V, E) defines a graph. If a graph has no circuits and is connected, then it is a tree. If one particular vertex is distinguished from the rest (this vertex is called the root), then (V, E) is a rooted tree. A single vertex is a rooted tree and a graph formed by joining a new root to the original roots of some existing rooted trees is also a rooted tree. We will denote τ for the tree with only one vertex. This can also be written as [ ]. If $t_i, i = 1, 2, \ldots, m$ are (rooted) trees then $[t_1 t_2 \cdots t_m]$ is the tree formed by joining the roots of $t_1, t_2, \ldots, t_m$ to a new root. Let us consider an example of a tree



$$t = \Big[\ [[[\tau\tau]]\ \ [\tau\tau\tau]\ \ [\tau[\tau]]\Big]$$

2.1.2 Functions on trees

- The number of vertices of a tree is the order of a tree, and is denoted by $r(t)$

$$\begin{aligned} r(\tau) &= 1, \\ r(\phi) &= 0, \\ r([t_1 t_2 \cdots t_m]) &= 1 + \sum_{i=1}^{m} r(t_i). \end{aligned}$$

- The symmetry of a tree is the order of automorphism group of t and is denoted by $\sigma(t)$

$$\begin{aligned} \sigma(\tau) &= 1, \\ \sigma([t_1^{s_1} t_2^{s_2} \cdots t_m^{s_m}]) &= \prod_{i=1}^{m} \sigma(t_i)^{s_i} s_i!. \end{aligned}$$

- The density of a tree $\gamma(t)$ is measurement of non-thickness of a tree. This is defined as

$$\begin{aligned} \gamma(\tau) &= 1, \\ \gamma(\phi) &= 0, \\ \gamma([t_1 t_2 \cdots t_m]) &= r([t_1 t_2 \cdots t_m]) \prod_{i=1}^{m} \gamma(t_i). \end{aligned}$$

Furthermore, $\alpha(t)$ is the number of ways of labelling a tree t with an ordered set and $\beta(t)$ is defined as the number of ways of labelling with an unordered set. This was shown in [19] that is

$$\begin{aligned} \alpha(t) &= \frac{r(t)!}{\sigma(t)\gamma(t)}, \\ \beta(t) &= \frac{r(t)!}{\sigma(t)}. \end{aligned}$$

Tree notation with the order of the tree, $r(t)$ the symmetry, $\sigma(t)$ and the density, $\gamma(t)$, are shown together with $\alpha(t)$ and $\beta(t)$ in Table 2.1.

Table 2.1: Functions on trees up to order three

$r(t)$	t	Notation	$\sigma(t)$	$\gamma(t)$	$\alpha(t)$	$\beta(t)$
1		τ	1	1	1	1
2		$[\tau]$	1	2	1	2
3		$[\tau\tau] = [\tau^2]$	2	3	1	3
3		$[[\tau]]$	1	6	1	6

2.1.3 Elementary differentials

To understand the order of Runge–Kutta methods, we need to write the Taylor expansions of the numerical solution at the end of a single time step and then compare this with the Taylor expansion of the exact solution. For this purpose, we need formulae for the second, third, ..., derivatives, and all of these are related to trees.

$$
\begin{aligned}
y'(x) &= \mathbf{f} \\
y''(x) &= \mathbf{f}'\mathbf{f} \\
y'''(x) &= \mathbf{f}''(\mathbf{f},\mathbf{f}) + \mathbf{f}'\mathbf{f}'\mathbf{f} \\
y^{(4)}(x) &= \mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f}) + 3\mathbf{f}''(\mathbf{f},\mathbf{f}'\mathbf{f}) + \mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f}) + \mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}
\end{aligned}
$$

The terms in the formulae for the first, second, third ,... derivatives are known as *elementary differentials* $F(t)(y(x))$. Generally, we represent these elementary differentials in terms of trees recursively. That is,

$$
\begin{aligned}
F(\phi)(y) &= y(x) \\
F(\tau)(y) &= f(y(x)) \\
F([t_1 t_2 \cdots t_m])(y) &= f^{(m)}(y(x))(F(t_1)(y(x)), \ldots, F(t_m)(y(x))).
\end{aligned}
$$

The formal Taylor expansion of the exact solution at $x_0 + h$ is

$$
y(x_0 + h) = y_0 + \sum_{t \in T} \frac{\alpha(t) h^{r(t)}}{r(t)!} F(t)(y_0).
$$

Using the known formula for $\alpha(t)$, we can write this as

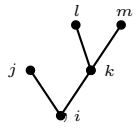
$$y(x_0 + h) = y_0 + \sum_{t \in T} \frac{h^{r(t)}}{\sigma(t)\gamma(t)} F(t)(y_0). \tag{2.3}$$

Our aim will now be to find a corresponding formula for the result computed by one step of a Runge-Kutta method. By comparing these formulae term by term, we will obtain conditions for a specific order of accuracy.

2.1.4 Elementary weights

The elementary weights are used to determine the order of a given Runge–Kutta method. For a tree t this is denoted by $\Phi(t)$.

Let

$\mathbf{t} =$ 

then

$$\Phi(\mathbf{t}) = \sum_{i,j,k,l,m=1}^{s} b_i a_{ij} a_{ik} a_{kl} a_{km}.$$

Simplify by summing over l, m:

$$\Phi(\mathbf{t}) = \sum_{i,j,k=1}^{s} b_i a_{ij} a_{ik} c_k^2.$$

The formal Taylor expansion of the *computed* solution at $x_0 + h$ is

$$y_1 = y_0 + \sum_{t \in T} \frac{\beta(t) h^{r(t)}}{r(t)!} \Phi(t) F(t)(y_0).$$

Using the known formula for $\beta(t)$, we can write this as

$$y_1 = y_0 + \sum_{t \in T} \frac{h^{r(t)}}{\sigma(t)} \Phi(t) F(t)(y_0) \tag{2.4}$$

Table 2.2: Elementary weights $\Phi(t)$ and elementary differentials $F(t)(y)$ up to order three

$r(t)$	t	$\Phi(t)$	$F(t)(y)$
1		$\sum b_i$	$\mathbf{f}$
2		$\sum b_i c_i$	$\mathbf{f}'\mathbf{f}$
3		$\sum b_i c_i^2$	$\mathbf{f}''(\mathbf{f},\mathbf{f})$
3		$\sum b_i a_{ij} c_j$	$\mathbf{f}'\mathbf{f}'\mathbf{f}$

Table 2.3: Order three conditions

tree	order conditions
	$b_1 + b_2 + b_3 = 1$
	$b_2 c_2 + b_3 c_3 = \frac{1}{2}$
	$b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3}$
	$b_3 a_{32} c_2 = \frac{1}{6}$

To match the two Taylor series (2.3) and (2.4) up to h^p terms we need to ensure that

$$\Phi(t) = \frac{1}{\gamma(t)},$$

for all trees such that

$$r(t) \leq p.$$

These are known as the *order conditions*. Thus for order three Runge–Kutta method, we need to satisfy the following order conditions which are given in Table 2.3.

2.1.5 Simplifying assumptions

The number of order conditions on the coefficients of a Runge–Kutta method increases rapidly as we seek higher order methods. For this reason Butcher [8] introduced simplifying assumptions. By using these assumptions, certain collections of order conditions can be removed from consideration.

1. $E(\kappa, \varkappa) : \sum_{j=1}^{s} \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} c_j^{l-1} = \frac{1}{l(k+l)}, l = 1, 2, ..., \varkappa, \quad k = 1, 2, ..., \kappa.$

2. $B(p) : \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, 2, ..., p.$ This condition states that the method has order p, and the order conditions related to bushy trees hold up to order p.

3. $C(\kappa) : \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, i = 1, 2, ..., s, \quad k = 1, 2, ..., \kappa.$ This condition relates pair of trees t_a and t_b, of the same order but with the slight difference in elementary weight function representation.

4. $D(\varkappa) : \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j(1 - c_j^k)}{k}, j = 1, 2, ..., s, \quad k = 1, 2, ..., \varkappa.$ This condition relates three trees t_a, t_b and t_c. The order condition related to t_a including corresponding elementary weight function $b_i c_i^{k-1} a_{ij}$ holds if there are lower order trees having elementary weights b_j and $b_j c_j^k$.

2.1.6 The use of B-series

We have noted that the Taylor expansions for the approximate solutions can be found and compared with the expansions for the exact solutions to assess accuracy of numerical methods. However the Taylor expansion of the numerical solution (2.4) can be written in terms of a formal series as,

$$\mathbf{B}(a, y_0) = a(\phi) y_0 + \sum_{t \in T} a(t) \frac{F(t)}{\sigma(t)} h^{r(t)} (y(t)), \tag{2.5}$$

where, $a(\phi), a(\tau), a([\tau]), \ldots, a(t), \ldots$, is a sequence of real numbers. The series (2.5) was named as **B**-series, by Hairer and Wanner [49], the name after John Butcher. We will now look at the possible choices of coefficients.

Special characteristics of B

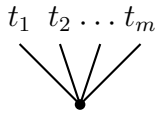
- $\mathbf{B}(\mathbf{1}(t), y_0) = y_0$, represents the identity mapping corresponding to $\mathbf{1}(t)$.
- $y_1 = \mathbf{B}(a(t), y_0), \Longleftrightarrow y_0 = \mathbf{B}(a^{-1}(t), y_1)$, is inverse mapping for $a^{-1}(t)$.
- $\mathbf{B}(a(t), y) + \mathbf{B}(\varphi(t), y) = \mathbf{B}((a + \varphi)(t), y)$, represents sum of two **B**-series.
- $\mathbf{B}(a(t), \mathbf{B}(\varphi(t), y)) = \mathbf{B}(a\varphi(t), y)$, represents the composition of two **B**-series .

We can write E for the flow for unit stepsize h, that is $\mathsf{E}(\phi) = 1$ and $\mathsf{E}(\phi) = \frac{1}{t!}$, for $t \in T$.
For a flow through a distance θh, we write $\mathsf{E}^{(\theta)}(\phi) = 1$ and $\mathsf{E}^{(\theta)}(\phi) = \frac{\theta^{|t|}}{t!}$, for $t \in T$.
Furthermore, the scaled differentiation D to product $hf(y_0) = hF(\tau)(y_0)$, is given by

$$\begin{aligned} \mathsf{D}(\phi) &= 0, \\ \mathsf{D}(\tau) &= 1, \\ \mathsf{D}(t) &= 0, \quad |t| > 1. \end{aligned}$$

B-series for $a\mathsf{D}$ is defined by using the known composition rule, that is, $\mathbf{B}(a\mathsf{D}, y_0) = \mathbf{B}(\mathsf{D}, \mathbf{B}(a, y_0))$. Assuming that $a(\phi) = 1$, it is found that

$$\begin{aligned} a\mathsf{D}(\phi) &= 0, \\ a\mathsf{D}(\tau) &= 1, \\ a\mathsf{D}(t) &= a(t_1)a(t_2)\dots a(t_m), \quad t = [t_1 t_2 \dots t_m]. \end{aligned}$$

$t =$ 

The details on **B**-series are given in Chapter 3.

2.2 Canonical Runge–Kutta methods

Consider the differential equation system

$$y'(x) = f(y(x)), \quad f : \mathbb{R}^N \to \mathbb{R}^N, \quad \langle Y, f(Y) \rangle = 0.$$

If a Runge–Kutta method R conserves quadratic first integral (i.e, $\mathcal{I}(y_1) = \mathcal{I}(y_0)$) whenever $\mathcal{I}(y) = y^T Q y$ is a first integral of (2.2), with symmetric matrix Q, then it is canonical or symplectic. Pioneering development in the study of symplectic Runge–Kutta methods is due to Cooper [31], Sanz–Serna [68], Lasagni [55] and Suris [73]. Their idea is based on features of algebraic stability introduced, in connection with stiff systems, that was introduced by Burrage and Butcher [7] and Crouzeix [32].

Theorem 2.2.1. *A Runge–Kutta method is said be symplectic (canonical), if* $\mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b) - bb^{\intercal} = 0$.

Symplectic numerical methods are an excellent choice for the long–term integration behaviour of Hamiltonian systems because these methods respect the characteristic properties of the underlying systems. Additionally, symplectic methods preserve the quadratic invariants. The matrix $M = \mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b) - bb^{\intercal}$ that has important features for the symplecticity of a Runge–Kutta method also plays an important role in the algebraic stability of that method.

2.2.1 Superfluous and non-superfluous trees and their effect on order conditions

Superfluous trees are those trees where any two adjacent nodes of the tree can be taken as a root and generate identical rooted trees. These trees play an important role in the formulation of order conditions for symplectic Runge–Kutta methods. Order conditions corresponding to superfluous trees are automatically satisfied because of the symplectic condition.

Once the superfluous trees are identified then the rest are named as non-superfluous trees, and these trees contribute one order condition. Table 2.4 shows the superfluous and non-superfluous trees up to order four.

Theorem 2.2.2. *[69] A symplectic Runge–Kutta method has order p if for each non-superfluous tree $\tilde{t}$ with any vertex as a root,*

$$\phi(\rho\widetilde{t}) = \frac{1}{\gamma(\rho\widetilde{t})},$$

where $\rho\widetilde{t}$ represents the rooted tree of $\widetilde{t}$ of order up to p.

Table 2.4: Trees and rooted trees as superfluous and non-superfluous trees

nature	tree	i	t_i
non-superfluous		1	
superfluous		2	
non-superfluous		3	
		4	
non-superfluous		5	
		7	
superfluous		6	
		8	

It is observed that the trees with an even number of vertices are superfluous. Therefore each symplectic Runge–Kutta method has at least order two. Superfluous and non-superfluous trees have an important impact on the order conditions of symplectic Runge–Kutta methods. To investigate this, we take tree having four vertices. We multiply c_i and c_j with $\mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b) - bb^{\intercal} = 0$, and then apply summation over i and j,

$$\begin{aligned} &\sum_{i,j} b_i c_i a_{ij} c_j + \sum_{i,j} b_j c_j a_{ji} c_i - \sum_i b_i c_i \sum_j b_j c_j = 0, \\ \Rightarrow &\sum_{i,j} b_i c_i a_{ij} c_j = \tfrac{1}{8}. \end{aligned} \tag{2.6}$$

This is infact the order condition corresponding to rooted tree  , which belongs to the same underlying superfluous tree . Therefore, order conditions corresponding to superfluous trees are automatically satisfied because of the symplectic condition and hence are no longer needed.

For symplectic Runge–Kutta methods with order ≥ 3, we assume the symplectic

Table 2.5: Number of order conditions.

Order	General Runge–Kutta method	Symplectic Runge–Kutta method
1	1	1
2	2	1
3	4	2
4	8	3
5	17	6

condition and multiply by c_j and take summation over i and j,

$$\sum_{i,j} b_i a_{ij} c_j + \sum_{i,j} b_j c_j a_{ji} - \sum_i b_i \sum_j b_j c_j = 0,$$
$$\Rightarrow (\sum_{i,j} b_i a_{ij} c_j - \tfrac{1}{6}) + (\sum_j b_j c_j^2 - \tfrac{1}{3}) = 0,$$

which is the sum of order conditions corresponding to rooted trees  and  . It is sufficient to impose the order condition as either  or  , because these are corresponds to the same tree . Hence we require only one order condition.

Table 2.5 shows the number of order conditions for general and symplectic Runge–Kutta methods up to order five.

Example 2.2.3. *The two stage order four Gauss method is symplectic.*

We recall that the famous two stage Gauss method is symplectic and achieves order four behaviour.

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} . \tag{2.7}$$

For the Hénon–Heiles problem, the close adherence of the Hamiltonian to its initial value for the two-stage Gauss method with order four can be seen in Figure (2.1), where a total of 100000 steps have been taken with the stepsize 0.01.

The main motivation for choosing implicit Runge–Kutta methods is to solve stiff differential equations. However the Hamiltonian problems with which we are concerned

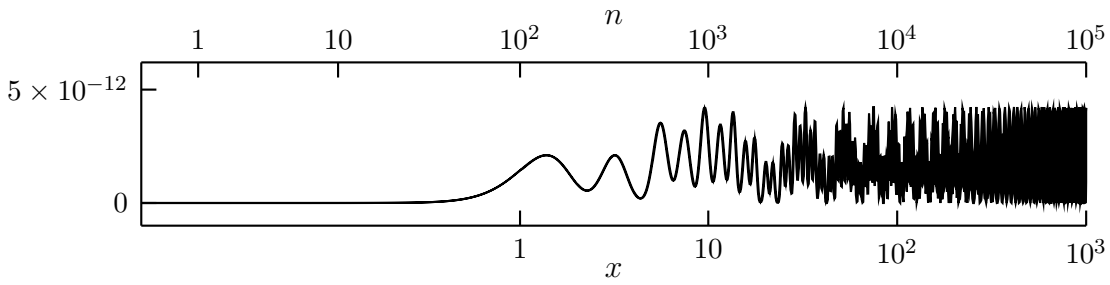


Figure 2.1: The error in energy for Hénon–Heiles problem using the Gauss method with $h = 0.01$

are not necessarily stiff, and to save the implementation cost we restrict ourselves to diagonally implicit methods. These methods respect symplectic behaviour if and only if they have the following tableau

$$\begin{array}{ccccc|} \frac{b_1}{2} & 0 & 0 & \cdots & 0 \\ b_1 & \frac{b_2}{2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_1 & b_2 & b_3 & \cdots & \frac{b_s}{2} \\ \hline b_1 & b_2 & b_3 & \cdots & b_s \end{array} . \tag{2.8}$$

(2.8) is a composition of an implicit midpoint step of length $b_1 h$, and an implicit midpoint step having length $b_2 h$ and so on. Since the implicit midpoint rule is symplectic, it can be proved that (2.8) is also symplectic. Therefore, these methods are as easy to implement as the implicit midpoint rule.

2.3 Linear multistep methods

Linear multistep methods are another class of numerical methods, which uses more past calculations to achieve the accuracy. These methods are defined by using an expression of the form

$$y_n = \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k} + h\big(\beta_0 f(x_n, y_n) + \beta_1 f(x_{n-1}, y_{n-1}) + \cdots + \beta_k f(x_{n-k}, y_{n-k})\big).$$

We always assume that α_k and β_k are not both zero. Note that if $\beta_0 = 0$ the method is *explicit*. It is *implicit* in nature, if $\beta_0 \neq 0$, and for the solution of non-linear stage

equations, it is necessary to solve for y_n. For a given linear multistep method, define polynomials ρ and σ as follows:

$$\begin{aligned}\rho(w) = &\quad w^k - \alpha_1 w^{k-1} - \alpha_2 w^{k-2} - \cdots - \alpha_k,\\ \sigma(w) = &\beta_0 w^k + \beta_1 w^{k-1} + \beta_2 w^{k-2} + \cdots + \beta_k.\end{aligned}$$

The first important methods in this class were Adams–Bashforth methods [2] together with Adams–Moulton methods [61]. In Adams–Bashforth methods, we assume $\beta_0 = 0$, and $\rho(w) = w^k - w^{k-1}$ to obtain order k. For all polynomials ϕ of degree up to $k-1$, we have

$$\beta_1\phi(-1) + \beta_2\phi(-2) + \cdots + \beta_k\phi(-k) = \int_{-1}^{0} \phi(x)dx.$$

For the two step Adams–Bashforth method, we have

$$\begin{aligned}\beta_1 + \beta_2 &= 1,\\ -\beta_1 - 2\beta_2 &= -\tfrac{1}{2},\end{aligned}$$

giving the results $\beta_1 = \frac{3}{2}$ and $\beta_2 = -\frac{1}{2}$. Therefore the method takes the form $y_n = y_{n-1} + \frac{3}{2}hf(y_{n-1}) - \frac{1}{2}hf(y_{n-2})$.

The methods are known as Adams-Moulton methods, if $\beta_0 \neq 0$ to obtain order $k+1$. The Adams-Moulton methods are implicit and more accurate than Adams–Bashforth methods.
The family of *backward difference methods* was introduced by Curtiss and Hirschfelder [33] for the solution of stiff differential equations. These are defined under the following assumptions

1. $\sigma(w) = \beta_0 w^k$

2. α_1, α_2. $\ldots$, α_k and β_0 chosen to give order $p = k$

That is, for all polynomials ϕ of degree up to k, we have $\phi(0) - \alpha_1\phi(-1) - \alpha_2\phi(-2) - \cdots - \alpha_k\phi(-k) = \beta_0\phi'(0)$.

Example 2.3.1. *The backward difference method of degree $k = 1$ satisfies*

$$1 - \alpha_1 = 0,$$
$$0 + \alpha_1 = \beta_0$$

giving the result $\alpha_1 = \beta_0 = 1$.

Example 2.3.2. *The backward difference method of degree $k = 2$ satisfies*

$$1 - \alpha_1 - \alpha_2 = 0,$$
$$0 + \alpha_1 + 2\alpha_2 = \beta_0,$$
$$0 + \alpha_1 + 4\alpha_2 = 0,$$

giving the result $\alpha_1 = \frac{4}{3}$, $\alpha_2 = -\frac{1}{3}$, $\beta_0 = \frac{2}{3}$.

2.3.1 Consistency, stability and convergence

Linear multistep methods are consistent if we get $\rho(1) = 0$, $\rho'(1) = \sigma(1)$ for the solution of simple differential equations.
A linear multistep method is *stable* if all solutions to the difference equation

$$y_n = \alpha_1 y_{n-1} + \cdots + \alpha_k y_{n-k},$$

are bounded as $n \to \infty$. We can obtain bounded solutions if all zeros of ρ lies in the closed unit disc and all *repeated* zeros lie in the *open* unit disc. It has been proved that a stable and consistent linear multistep method is always convergent.
A linear multistep method has order p, if the polynomials ρ and σ satisfy,

$$\rho(\exp(z)) - z\sigma(\exp(z)) = O(z^{p+1}).$$

The *First Dahlquist barrier* states that if a stable linear k step method has order p then

$$p \leq \begin{cases} k+1, & k \text{ odd} \\ k+2, & k \text{ even.} \end{cases}$$

It was later proved by Dahlquist that for a linear multistep method the maximum order consistent with A-stability is two. This is known as the *Second Dahlquist barrier*. The modern theory of linear multistep methods is due to Dahlquist [35].

2.4 General linear methods

General linear methods are the natural generalisation of linear multistep methods and Runge–Kutta methods, and are multistage and multivalue in nature. These methods were introduced by Butcher [11]. General linear methods are defined by

$$\begin{aligned} Y &= h(A \otimes I)f(Y) + (U \otimes I)y^{[n-1]}, \\ y^{[n]} &= h(B \otimes I)f(Y) + (V \otimes I)y^{[n-1]}, \end{aligned} \tag{2.9}$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \quad f(Y) = \begin{bmatrix} f(Y_1) \\ f(Y_2) \\ \vdots \\ f(Y_s) \end{bmatrix}, \quad y^{[n-1]} = \begin{bmatrix} y_1^{[n-1]} \\ y_2^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{bmatrix}, \quad y^{[n]} = \begin{bmatrix} y_1^{[n]} \\ y_2^{[n]} \\ \vdots \\ y_r^{[n]} \end{bmatrix}.$$

In (2.9),

1. I is the identity matrix
2. $A \otimes I$ represents the Kronecker product of the matrix A with the identity matrix
3. h is the stepsize
4. $f(Y)$ is stage derivatives
5. The vector $y^{[n-1]}$ having $r-$components are regarded as input values at the beginning of a step
6. $Y_i \approx y(x_n + c_i h)$ represents the stages and is an approximation to the solution at c_i.

General linear methods can be rewritten with a slight modification as

$$\begin{aligned} Y &= hAf(Y) + Uy^{[n-1]}, \\ y^{[n]} &= hBf(Y) + Vy^{[n-1]}. \end{aligned} \tag{2.10}$$

The elements A, U, V and B are the representatives of a particular general linear methods, and are written in partitioned matrix as

$$M = \left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right]. \tag{2.11}$$

This formulation of general linear method was introduced by Burrage and Butcher in [6] .

2.4.1 Pre–consistency, consistency, stability and convergence

The pre–consistency condition is related to the trivial solution of the ordinary one-dimensional equation $y'(x) = 1$, with initial condition $y(0) = 0$,

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

where u, the *pre–consistency vector*, has a certain value for a certain numerical method. By using a general linear method to solve the differential equation $y'(x) = 1$, the internal stages and the output approximations are given by

$$\begin{aligned} Y = Uy^{[n-1]} &= Uuy(x_{n-1}) + O(h), \\ y^{[n]} = Vy^{[n-1]} &= Vuy(x_{n-1}) + O(h). \end{aligned}$$

Therefore the pre-consistency conditions are $Uu = \mathbf{1}$, $Vu = u$.

The consistency condition is concerned with the solution of the differential equation $y'(x) = 1$, with initial condition $y(0) = 0$. The numerical solution should be exact at both the beginning and the end of the step. Thus the consistency vector v is determined by the equations

$$\begin{aligned} y^{[n-1]} &= uy(x_{n-1}) + vhy'(x_{n-1}) + O(h^2), \\ y^{[n]} &= uy(x_n) + vhy'(x_n) + O(h^2). \end{aligned}$$

Using a general linear method to solve the differential equation $y'(x) = 1$, the internal stages and output vector are represented by the following set of equations

$$Y = A\mathbf{1}h + Uy^{[n-1]} = A\mathbf{1}h + Uuy(x_{n-1}) + hUvy'(x_{n-1}) + O(h^2),$$
$$y^{[n]} = B\mathbf{1}h + Vy^{[n-1]} = B\mathbf{1}h + Vuy(x_{n-1}) + hVvy'(x_{n-1}) + O(h^2).$$

The consistency conditions are therefore $B\mathbf{1} + Vv = u + v$. If the solution of the trivial differential equation $y'(x) = 0$ is bounded, then we are certain of the stability of general linear method. Stability implies that the error in one step do not grow to the later steps. Thus

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

This shows that the stability of method depends on the stability of matrix V. The matrix V is stable if there exists a constant L such that $\|V^n\|_\infty \leq L, n = 1, 2, \cdots$.

Definition 2.4.1. *A general linear method is said to be strictly stable if all the eigenvalues of matrix V lies inside the unit disc except one which is on the boundary.*

A general linear method is *convergent* if, for any choice of initial value problem, there exist a non-zero vector $u \in \mathbb{R}^n$ such that if the starting approximation $y^{[0]}$ approaches to $uy(x_0)$, then $y^{[n]} = uy(x_0 + nh)$ for all n. Stability and consistency are necessary and sufficient conditions for the convergence of a general linear method. This was actually generalised by Butcher [11], and based on the fundamental idea of Dahlquist [34].

2.4.2 GLM order and its analysis using B-series

We require a starting method S_h to obtain the initial value vector $y^{[0]}$ from only a single initial condition $y(x_0) = y_0$. A starting method can be thought of as a Runge–Kutta method with multiple outputs. We consider a $(\bar{s} + r) \times (\bar{s} + \mathbf{1})$, partitioned matrix to represent a starting procedure S_h, given by

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

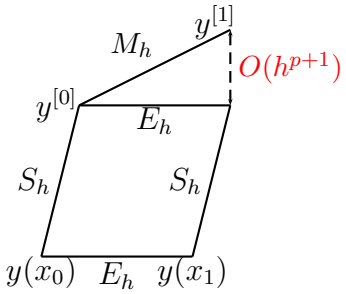


Figure 2.2: Order of accuracy

where $\bar{s}$ represents the number of stages and r are the number of approximations, which needs to be computed. The pre–consistency conditions for the starting method S_h are $S_{22} = u, \quad S_{12} = \mathbf{1}$.

When the starting method S_h is applied to a given initial value problem then the output can be used as the input of the main method M_h. The combined operation of the starting method S_h and the main method M_h is denoted by $M_h o S_h$. Similarly E_h represents the exact solution and $S_h \circ E_h$ represents the result of applying the starting method to the exact solution evaluated after a time step h. The order of accuracy of the general linear method M_h has order p relative to S_h if

$$M_h \circ S_h - S_h \circ E_h = O(h^{p+1}).$$

This is illustrated in Figure 2.2.

In addition to the mapping S_h, a finishing method is required to cancel out the effect of starting method. The mapping S_h can be written in terms of **B**-series as,

$$y^{[0]} = \mathbf{B}(S_h(t), y(x_{n-1})).$$

The vector η represents the **B**-series coefficients for the stage values. That is

$$Y = \mathbf{B}(\eta(t), y(x_{n-1})).$$

The vector ξ represents **B**-series coefficients for the r components of the inputs. That is

$$Y = \mathbf{B}(\xi(t), y(x_{n-1})).$$

We can interpret the relation between one-step method in terms of **B**-series as:

$$\begin{aligned} Y = hAf(Y) + US_h(y(x_0)), \quad & \eta = A(\eta \mathsf{D})(t) + U\xi(t), \\ S_h(y(x_0 + h)) = hBf(Y) + VS_h(y(x_0)), \quad & \mathsf{E}\xi(t) = B(\eta \mathsf{D})(t) + V\xi(t). \end{aligned} \tag{2.12}$$

In $(\eta \mathbf{D})(t)$, $\mathbf{D}$ stands for scalar post-multiplier for the vector $\eta(t)$ and corresponds to evaluating $hf(Y_i)$ for each component of Y. Also $\mathbf{E}\xi(t)$, $\mathbf{E}$ has a role of a scalar pre-multiplier of the vector $\xi(t)$. This corresponds to shifting a unit time-step along the trajectory before calculating the starting operation. This gives an abstract meaning of order to a practical method of interest. These ideas have been widely used to find reliable and efficient methods for the solution of stiff and non–stiff differential systems and Hamiltonian problems.

Chapter 3

Effective order five methods

The concept of effective order was introduced in [12] and a modern presentation is given in [19]. The original idea was aimed at overcoming the order barrier on five-stage methods but it has a wider relevance. Chan [30] showed how effective order can yield more efficient explicit methods.

This chapter is organised as follows. In Section 3.1, we discuss different ways in which Runge–Kutta methods can be thought of as equivalent. It will be shown that equivalence classes of Runge–Kutta methods form a group. This group is homomorphic to the group of mappings from trees to real numbers. By introducing these groups, we can get a better understanding of the order conditions of Runge–Kutta methods.

The concept of effective order and its relative order conditions will be discussed in Section 3.2. From these starting points, we will derive a new fifth effective order method with five stages in Section 3.3. These methods are derived in such a way that they satisfy a certain simplifying assumptions $C(2)$. A detailed discussion about the simplifying assumptions is given in Chapter 2. These assumptions lead us to a direct and simpler derivation of these methods. In Section 3.4, we will present fifth effective order methods satisfying $D(1)$ assumptions only.

3.1 The Runge–Kutta group

Runge–Kutta methods can be regarded as mappings from R^N to R^N, for any initial value problem. If a Runge–Kutta method with s stages is combined with another

Runge–Kutta method with $\widetilde{s}$ stages, then we can get a composed method with $s+\widetilde{s}$ stages. This combined method can be viewed as the product of the original methods. Formulating the product of methods in such a manner is not a best way. Therefore, we need to work on the equivalence classes of Runge–Kutta methods because these equivalence classes help us in constructing the group structure of Runge–Kutta methods rather than their semigroup pattern.
Two Runge–Kutta methods could be considered equivalent in several different ways. For instance,

- If for any autonomous initial value problem, defined by a function $f(x)$ satisfying Lipschitz condition, both methods yield same numerical results for small step sizes.

- This implies a second way of looking at equivalence. It might be possible to delete unrelated or repeated stages from a method to yield a reduced method. So if two methods reduced to the same method, in this respect they regarded as equivalent.

- This yields another way of looking at equivalence of two methods. If for any tree t the elementary weight function $\Phi(t)$ for the first method is similar to the elementary weight function $\xi(t)$ of second method.

Thus, if any two Runge–Kutta methods are equivalent in one of the above mention way, then they are also equivalent in each of the other way. By using equivalence classes of Runge–Kutta methods R and $\widetilde{R}$, we can construct a tableau for the com-

position of these methods. Consider two Runge–Kutta methods R and $\widetilde{R}$,

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|ccccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & & b_s \end{array}$$

$$\begin{array}{c|c} \widetilde{c} & \widetilde{A} \\ \hline & \widetilde{b}^T \end{array} = \begin{array}{c|ccccc} \widetilde{c}_1 & \widetilde{a}_{11} & \widetilde{a}_{12} & \cdots & \widetilde{a}_{1\widetilde{s}} \\ \widetilde{c}_2 & \widetilde{a}_{21} & \widetilde{a}_{22} & \cdots & \widetilde{a}_{2\widetilde{s}} \\ \vdots & \vdots & \vdots & & \vdots \\ \widetilde{c}_{\widetilde{s}} & \widetilde{a}_{\widetilde{s}1} & \widetilde{a}_{\widetilde{s}2} & \cdots & \widetilde{a}_{\widetilde{s}\widetilde{s}} \\ \hline & \widetilde{b}_1 & \widetilde{b}_2 & & \widetilde{b}_{\widetilde{s}} \end{array} .$$

To see how the composed method is constructed, we have written s and $\widetilde{s}$ for the number of stages of these two methods. Let the initial value for the method R be y_0 and y_1 the value after one step of R, Then write y_2 for the output value computed by the second method $\widetilde{R}$, using y_1 for its initial value. We can see that y_2 is the result computed by the composed method.

To see how this works, look at the formulae for the stages and output values:

$$\begin{aligned}
Y_i &= y_0 + h\sum_{j=1}^{s} a_{ij}F_j, \\
y_1 &= y_0 + h\sum_{j=1}^{s} b_j F_j, \\
\widetilde{Y}_i &= y_1 + h\sum_{j=1}^{\widetilde{s}} \widetilde{a}_{ij}\widetilde{F}_j && = y_0 + h\sum_{j=1}^{s} b_j F_j + h\sum_{j=1}^{\widetilde{s}} \widetilde{a}_{ij}\widetilde{F}_j, \\
y_2 &= y_1 + h\sum_{j=1}^{\widetilde{s}} \widetilde{b}_j\widetilde{F}_j && = y_0 + h\sum_{j=1}^{s} b_j F_j + h\sum_{j=1}^{\widetilde{s}} \widetilde{b}_j\widetilde{F}_j.
\end{aligned}$$

Hence, y_2 can be computed from y_0 using a single Runge–Kutta method with $s+\widetilde{s}$ stages given by

$$\begin{array}{c|cccccccc}
c_1 & a_{11} & a_{12} & \cdots & a_{1s} & 0 & 0 & \cdots & 0 \\
c_2 & a_{21} & a_{22} & \cdots & a_{2s} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
c_s & a_{s1} & a_{s2} & \cdots & a_{ss} & 0 & 0 & \cdots & 0 \\
\sum b_i + \widetilde{c}_1 & b_1 & b_2 & \cdots & b_s & \widetilde{a}_{11} & \widetilde{a}_{12} & \cdots & \widetilde{a}_{1\widetilde{s}} \\
\sum b_i + \widetilde{c}_2 & b_1 & b_2 & \cdots & b_s & \widetilde{a}_{21} & \widetilde{a}_{22} & \cdots & \widetilde{a}_{2\widetilde{s}} \\
 & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\sum b_i + \widetilde{c}_{\widetilde{s}} & b_1 & b_2 & \cdots & b_s & \widetilde{a}_{\widetilde{s}1} & \widetilde{a}_{\widetilde{s}2} & \cdots & \widetilde{a}_{\widetilde{s}\widetilde{s}} \\
\hline
 & b_1 & b_2 & \cdots & b_s & \widetilde{b}_1 & \widetilde{b}_2 & \cdots & \widetilde{b}_{\widetilde{s}}
\end{array}.$$

Also for a Runge–Kutta method R, we write $[R]$ for the corresponding equivalence class containing this method.
We now calculate some of the elementary weights for this composed method $R\widetilde{R}$. Let $\widehat{\Phi}(t)$ be the elementary weight function for the composed method $R\widetilde{R}$, and see how $\widehat{\Phi}(t)$ depend on $\Phi(t)$ for the first method and $\widetilde{\Phi}(t)$ for the second method. This is illustrated in Table 3.2.

Now by using equivalence classes, under the composition operation, the semigroup group of Runge–Kutta tableaux, can be used to form a group. Let G denote the set of mappings from trees to real numbers. Let α and β be two members of G. Our objective is now to define a group operation which describes the values of $(\alpha\beta)(t)$, for every tree t.

For a given tree t, let $\digamma(t)$ represent the set of prunings of tree t. A pruning of a tree t is characterised by a subtree t', which shares the same root as t. Let $t\backslash t'$ represent the set of trees formed by joining the parts of tree t left over, when t' is removed from tree t. Finally, $\alpha(t\backslash t')$ is the product over the connected components of individual α factors. This is illustrated in Table 3.3 for a particular tree $t=\mathbf{t}_{11}$. The product $(\alpha\beta)(t)$, is defined as the value of $\alpha(t)$ plus the sum over all subtrees

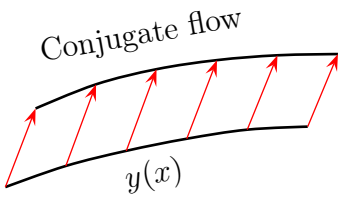


Figure 3.1: Algebraic interpretation of effective order

of the form $\alpha(t\backslash t')\beta(t')$. That is,

$$(\alpha\beta)(t) = \beta(\phi)\alpha(t) + \beta(t) + \sum_{t \prec t'} \alpha(t\backslash t')\beta(t'). \tag{3.1}$$

The product $\alpha\beta(t)$ up to order five trees is given in Table 3.1. The product defined in this way is known to be associative and forms a group.

3.2 Effective order

It is known that order five method is impossible with five stages, but there is something which might be almost as good. Butcher introduced *Effective Order* [12].

Definition 3.2.1. *A Runge–Kutta method R with group element α is of effective order p if there exists another Runge–Kutta method $\widetilde{R}$ with corresponding group element β such that $\beta\alpha\beta^{-1}$ is of order p.*

Algebraic interpretation

The idea of effective order is that *instead of agreeing to within a required order of accuracy to the exact solution, it is required that the computed results agree with a modification of the exact solution.* Therefore the main method will not keep us exactly on the flow, but it will keep us close to the conjugate flow where every point $y(x)$ is mapped. This is illustrated in the Figure 3.1.

Geometric interpretation

An alternative way to explain effective order is that every input value for the effective order method α is perturbed by the method β. So β method offers us some freedom

Table 3.1: $\alpha\beta$ product up to fifth order trees

t_i	$(\alpha\beta)(t_i)$
$\mathbf{t}_1$	$\alpha_1 + \beta_1$
$\mathbf{t}_2$	$\alpha_2 + \alpha_1\beta_1 + \beta_2$
$\mathbf{t}_3$	$\alpha_3 + \alpha_1\beta_1^2 + \alpha_1^2\beta_1 + \beta_3$
$\mathbf{t}_4$	$\alpha_4 + \alpha_2\beta_1 + \alpha_1\beta_2 + \beta_4$
$\mathbf{t}_5$	$\alpha_5 + 3\alpha_1\beta_3 + \frac{3}{2}\alpha_1^2\beta_1^2 + \alpha_1^3\beta_1 + \beta_5$
$\mathbf{t}_6$	$\alpha_6 + \beta_6 + \frac{1}{2}\alpha_1\beta_1^3 + \frac{1}{2}\alpha_1^3\beta_1 + \frac{3}{4}\alpha_1^2\beta_1^2$
$\mathbf{t}_7$	$\alpha_7 + \frac{1}{2}\alpha_1^2\beta_1^2 + \alpha_3\beta_1 + \alpha_1\beta_3 + \beta_7$
$\mathbf{t}_8$	$\alpha_8 + \alpha_4\beta_1 + \alpha_2\beta_2 + \alpha_1\beta_4 + \beta_8$
$\mathbf{t}_9$	$\alpha_9 + 4\alpha_1\beta_5 + 6\alpha_1^2\beta_3 + 2\alpha_1^3\beta_1^2 + \alpha_1^4\beta_1 + \beta_9$
$\mathbf{t}_{10}$	$\alpha_{10} + \beta_{10} + 2\alpha_1\beta_6 + \alpha_1\beta_5 + \frac{1}{2}\alpha_1^2\beta_1^3 + \frac{3}{2}\alpha_1^2\beta_3 + \frac{1}{4}\alpha_1^3\beta_1^2 + \frac{1}{2}\alpha_1^3\beta_1^2 + \frac{1}{4}\alpha_1^4\beta_1^2$
$\mathbf{t}_{11}$	$\alpha_{11} + \alpha_1\beta_7 + \alpha_1\beta_5 + 2\alpha_1^2\beta_3 + \frac{1}{2}\alpha_3\beta_1^2 + \frac{1}{2}\alpha_1^3\beta_1^2 + \alpha_1\alpha_3\beta_1 + \beta_{11}$
$\mathbf{t}_{12}$	$\alpha_{12} + \beta_{12} + \alpha_1\beta_1\beta_3 + \frac{1}{2}\alpha_1^2\beta_1^3 - \frac{1}{2}\alpha_1^2\beta_3 + \frac{1}{2}\alpha_1^3\beta_1^2 - \frac{1}{2}\alpha_3\beta_1^2 + \frac{1}{2}\alpha_1^4\beta_1 - \alpha_1\alpha_3\beta_1$
$\mathbf{t}_{13}$	$\alpha_{13} + \beta_{13} + 2\alpha_1\beta_6 + \frac{1}{2}\alpha_1^2\beta_1^3 + \frac{1}{2}\alpha_1^3\beta_1^2 + \frac{1}{4}\alpha_1^4\beta_1$
$\mathbf{t}_{14}$	$\alpha_{14} + 3\alpha_1\beta_7 + \frac{3}{2}\alpha_1^2\beta_3 + \frac{1}{2}\alpha_1^3\beta_1^2 + \alpha_5\beta_1 + \beta_{14}$
$\mathbf{t}_{15}$	$\alpha_{15} + \alpha_1\beta_8 + \alpha_1\beta_7 + (\alpha_1^2 + \alpha_2)\beta_4 + \alpha_1\alpha_2\beta_2 + \alpha_6\beta_1 + \beta_{15}$
$\mathbf{t}_{16}$	$\alpha_{16} + \alpha_1\beta_7 + \frac{1}{2}\alpha_1^2\beta_3 + \frac{1}{2}\alpha_3\beta_1^2 + \alpha_7\beta_1 + \beta_{16}$
$\mathbf{t}_{17}$	$\alpha_{17} + \alpha_1\beta_8 + \alpha_2\beta_4 + \alpha_4\beta_2 + \alpha_8\beta_1 + \beta_{17}$

Table 3.2: Elementary weights for the composed method

Tree	Elementary weights	
•	$\widehat{\Phi}(\mathbf{t}_1)$	$= \sum b_i + \sum \widetilde{b}_i$
		$= \Phi(\mathbf{t}_1) + \widetilde{\Phi}(\mathbf{t}_1)$
	$\widehat{\Phi}(\mathbf{t}_2)$	$= \sum b_i c_i + \sum \widetilde{b}_i(\sum b_j + \widetilde{c}_i)$
		$= \Phi(\mathbf{t}_2) + \widetilde{\Phi}(\mathbf{t}_1)\Phi(\mathbf{t}_1) + \widetilde{\Phi}(\mathbf{t}_2)$
	$\widehat{\Phi}(\mathbf{t}_3)$	$= \sum b_i c_i^2 + \sum \widetilde{b}_i(\sum b_j + \widetilde{c}_i)^2$
		$= \Phi(\mathbf{t}_3) + \widetilde{\Phi}(\mathbf{t}_1)\Phi(\mathbf{t}_1)^2 + 2\widetilde{\Phi}(\mathbf{t}_2)\Phi(\mathbf{t}_1) + \widetilde{\Phi}(\mathbf{t}_2^2)$

in the order conditions that need to be satisfied by the method α to attain effective order p. So by introducing the starting method, one can have more free parameters to build up one order higher method. At the end of the procedure every output value is returned back to the original flow by using β^{-1}. This concept is shown in the Figure 3.2.

Remark 3.2.2.

We could let $\beta(\tau) = 0$ without loss of generality for finding effective order method α. In this way we can simplify the calculations for effective order method. For the

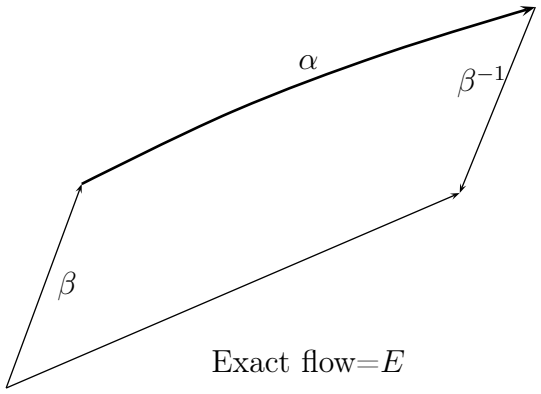


Figure 3.2: Geometric interpretation of effective order

Table 3.3: Example of $\alpha(t\backslash t')$

t'	$t \setminus t'$	$\alpha(t\backslash t')$
		1
		$\alpha(\bullet)$
		$\alpha(\bullet)$
		$\alpha(\bullet)$
		$\alpha(\bullet)^2$
		$\alpha(\bullet)^2$
		$\alpha(\bullet)^2$
		$\alpha(\bullet)^3$
		$\alpha($  $)$
		$\alpha(\bullet)\alpha($  $)$

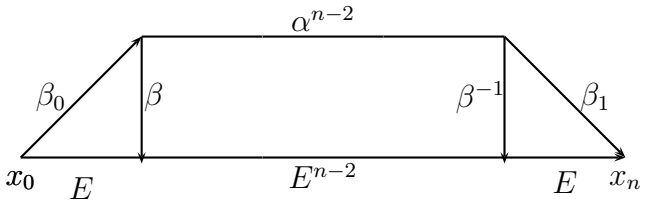


Figure 3.3: Integration for effective order

starting method β_0, we could let $\beta_0(\tau) = 1$, If we make this assumption then the starting method not only perturbs the initial value, but also moves the integration process one step forward. Therefore the starting method β is chosen in such way, that $\beta_0 = E\beta$. Similarly, finishing method can be designed in a similar way so that the finishing method not only take us to the original trajectory but also integrate over stepsize h, thus for finishing method, $\beta_1 = \beta^{-1}E$ provided that $\beta(\tau) = 0$. This can be shown in the Figure 3.3.

3.3 Construction of new fifth effective order method

We are interested in deriving fifth effective order methods with five stages. Originally we have seventeen order conditions for these methods which are given in Table 3.1. Solution of these seventeen non-linear conditions is not easy, therefore, we assume $C(2)$ simplifying assumption up to order five trees. Doing so, we can reduce the number of order conditions to nine.
We select the following trees of this type of explicit method up to order five.

For the nine trees the product $\beta\alpha$ and $E\beta$ are presented in Table 3.4.

$$\alpha_1 = 1, \tag{3.2}$$

$$\alpha_2 = \frac{1}{2}, \tag{3.3}$$

$$\alpha_3 = \frac{1}{3}, \tag{3.4}$$

$$\alpha_5 = \frac{1}{4} + 3\beta_3, \tag{3.5}$$

$$\alpha_7 = \frac{1}{12}, \tag{3.6}$$

$$\alpha_9 = \frac{1}{5} + 6\beta_3 + 4\beta_5, \tag{3.7}$$

$$\alpha_{11} = \frac{1}{15} + \frac{3}{2}\beta_3 + \beta_5 + \beta_7, \tag{3.8}$$

$$\alpha_{14} = \frac{1}{20} + \frac{3}{2}\beta_3 - \beta_5 + 3\beta_7, \tag{3.9}$$

$$\alpha_{16} = \frac{1}{60}. \tag{3.10}$$

Table 3.4: Expressions for $\beta\alpha$ and $E\beta$ for trees up to order five

$(\beta\alpha)(t_i)$	$(E\beta)(t_i)$
$\beta_1 + \alpha_1$	$1 + \beta_1$
$\beta_2 + \alpha_2$	$\frac{1}{2} + \beta_2$
$\beta_3 + \beta_1\alpha_1^2 + \beta_1^2\alpha_1 + \alpha_3$	$\frac{1}{3} + \beta_1 + \beta_1^2 + \beta_3$
$\beta_5 + 3\beta_1\alpha_3 + \frac{3}{2}\beta_1^2\alpha_1^2 + \beta_1^3\alpha_1 + \alpha_5$	$\frac{1}{4} + \beta_1 + \frac{3}{2}\beta_1^2 + 3\beta_3 + \beta_5$
$\beta_7 + \frac{1}{2}\beta_1^2\alpha_1^2 + \beta_3\alpha_1 + \beta_1\alpha_3 + \alpha_7$	$\frac{1}{12} + \frac{1}{3}\beta_1 + \frac{1}{2}\beta_1^2 + \beta_3 + \beta_7$
$\beta_9 + 4\beta_1\alpha_5 + 6\beta_1^2\alpha_3 + 2\beta_1^3\alpha_1^2 + \beta_1^4\alpha_1 + \alpha_9$	$\frac{1}{5} + \beta_1 + 2\beta_1^2 + 6\beta_3 + 4\beta_5 + \beta_9$
$\beta_{11} + \beta_1\alpha_7 + \beta_1\alpha_5 + 2\beta_1^2\alpha_3 + \frac{1}{2}\beta_1^3\alpha_1^2$ $+\frac{1}{2}\beta_3\alpha_1^2 + \beta_1\beta_3\alpha_1 + \alpha_{11}$	$\frac{1}{15} + \frac{1}{3}\beta_1 + \frac{2}{3}\beta_1^2 + 2\beta_3 + \beta_5 + \beta_7 + \beta_{11}$
$\beta_{14} + 3\beta_1\alpha_7 + \frac{3}{2}\beta_1^2\alpha_3 + \frac{1}{2}\beta_1^3\alpha_1^2 + \beta_5\alpha_1 + \alpha_{14}$	$\frac{1}{20} + \frac{1}{4}\beta_1 + \frac{1}{2}\beta_1^2 + \frac{3}{2}\beta_3 + 3\beta_7 + \beta_{14}$
$\beta_{16} + \beta_1\alpha_7 + \frac{1}{2}\beta_1^2\alpha_3 + \frac{1}{2}\beta_3\alpha_1^2 + \beta_7\alpha_1 + \alpha_{16}$	$\frac{1}{60} + \frac{1}{12}\beta_1 + \frac{1}{6}\beta_1^2 + \frac{1}{2}\beta_3 + \beta_7 + \beta_{16}$

Eliminate β values from equations (3.5), (3.7), (3.8) and (3.9) we get

$$\alpha_5 - \alpha_9 + 3\alpha_{11} - \alpha_{14} = \frac{1}{5}. \tag{3.11}$$

The six effective order conditions in terms of the coefficients of a Runge–Kutta method are

$$\begin{aligned}
\sum b_i &= 1, \\
\sum b_i c_i &= \tfrac{1}{2}, \\
\sum b_i c_i^2 &= \frac{1}{3}, \\
\sum b_i a_{ij} c_j^2 &= \frac{1}{12}, \\
\sum b_i a_{ij} a_{jk} c_k^2 &= \frac{1}{60}, \\
\sum b_i c_i^3 - \sum b_i c_i^4 + 3\sum b_i c_i a_{ij} c_j^2 - \sum b_i a_{ij} c_j^3 &= \frac{1}{5}.
\end{aligned}$$

The steps we need to carry out to derive one of the effective order five methods satisfying the above order conditions are as follows:

- Choose c_2, c_3, c_4, noting that $c_1 = 0$ and $c_5 = \frac{4}{5}$,
- $b_2 = 0$ and b_5 as a parameter,
- Choose b_1, b_3, b_4 to satisfy $\sum_{i=1}^{5} b_i c_i^{k-1} = \frac{1}{k}$ for $k = 1, 2, 3$,
- $\sum_{j=2}^{i-1} a_{ij} c_j = \frac{1}{2} c_i^2, \qquad i = 3, 4, 5,$
- $\sum_{j=1}^{i-1} a_{ij} = c_i, \qquad i = 2, 3, 4, 5,$
- $b_2 = \sum_{i=1}^{5} b_i (1 - c_i) a_{i2} = 0,$
- $b_5 a_{54} c_4 (c_4 - c_3) = \frac{1}{12} - \frac{c_3}{6}.$

The general solution, is given by the following coefficients:

$$
\begin{aligned}
a_{21} &= c_2, \\
a_{31} &= c_1 - \frac{c_3^2}{2c_2}, \\
a_{32} &= \frac{c_3^2}{2c_2}, \\
a_{42} &= -\frac{b_3(c_3 - c_5)a_{32}}{b_4(c_4 - c_5)}, \\
a_{43} &= \frac{c_4^2}{2c_3} - a_{42}\frac{c_2}{c_3}, \\
a_{41} &= c_4 - a_{42} - a_{43}, \\
a_{52} &= -\frac{b_3(c_3 - c_4)a_{32}}{b_5(c_5 - c_4)}, \\
a_{53} &= \frac{c_5^2}{2c_3} - a_{54}\frac{c_4}{c_3} - a_{52}\frac{c_2}{c_3}, \\
a_{51} &= c_5 - a_{52} - a_{53} - a_{54}.
\end{aligned}
$$

The order condition on tree $\mathbf{t}_{11} =$  can be simplified by using the following tree combination that is

$$\sum b_i a_{ij} a_{jk} c_k (c_k - c_3) = \frac{1}{60} - \frac{c_3}{24} = 0,$$

which implies that

$$c_3 = \frac{2}{5}.$$

For the order condition on tree $\mathbf{t}_7 =$  , we can solve the following relation

$$\sum b_i a_{ij} c_j (c_j - c_3) = \frac{1}{12} - \frac{c_3}{6}, \qquad i = j = 1 \ldots 5,$$

therefore

$$a_{54} = \frac{1}{60 b_5 c_4 (c_4 - c_3)}.$$

A possible solution to these equations gives the method represented by the tableau

$$\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{3}{5} & \frac{3}{5} & 0 & 0 & 0 & 0 \\
\frac{2}{5} & \frac{4}{15} & \frac{2}{15} & 0 & 0 & 0 \\
\frac{2}{3} & -\frac{2}{27} & -\frac{10}{27} & \frac{10}{9} & 0 & 0 \\
\frac{4}{5} & \frac{28}{75} & \frac{4}{15} & -\frac{1}{5} & \frac{9}{25} & 0 \\
\hline
 & \frac{19}{96} & 0 & \frac{25}{96} & \frac{9}{32} & \frac{25}{96}
\end{array}.$$

A suitable starting method, which does not advance the solution forward but introduces the correct perturbation so that α method faithfully produce this perturbation to within effective order five. We need to construct this method satisfying the following order conditions:

$$\sum b_i = 0, \tag{3.12}$$

$$\sum b_i c_i = 0, \tag{3.13}$$

$$\sum b_i c_i^2 = -\frac{1}{180}, \tag{3.14}$$

$$\sum b_i c_i^3 = \frac{1}{1800}. \tag{3.15}$$

The starting method is given by the tableau,

$$\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
\frac{7}{10} & \frac{7}{10} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{81}{154} & -\frac{2}{77} & 0 & 0 \\
\frac{3}{10} & \frac{129}{350} & \frac{3}{91} & -\frac{33}{325} & 0 \\
\hline
 & -\frac{16}{189} & \frac{5}{56} & -\frac{11}{36} & \frac{65}{216}
\end{array}.$$

Table 3.5: Group elements for effective order five

	t	$E(t)$	$\alpha(t)$	$\beta(t)$
$\mathbf{t}_1$		1	1	0
$\mathbf{t}_2$		$\frac{1}{2}$	$\frac{1}{2}$	0
$\mathbf{t}_3$		$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{180}$
$\mathbf{t}_4$		$\frac{1}{6}$	$\frac{1}{6}$	$-\frac{1}{360}$
$\mathbf{t}_5$		$\frac{1}{4}$	$\frac{7}{30}$	$\frac{1}{1800}$
$\mathbf{t}_6$		$\frac{1}{8}$	$\frac{7}{60}$	$\frac{1}{3600}$
$\mathbf{t}_7$		$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{900}$
$\mathbf{t}_8$		$\frac{1}{24}$	$\frac{1}{24}$	$\frac{1}{1800}$

Finally, the finishing method is chosen in such a way that it cancels out the effect of starting method, and is given by

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{7}{10} & \frac{7}{10} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{81}{154} & -\frac{2}{77} & 0 & 0 \\ \frac{3}{10} & \frac{129}{350} & \frac{3}{91} & -\frac{33}{325} & 0 \\ \hline & \frac{16}{189} & -\frac{5}{56} & \frac{11}{36} & -\frac{65}{216} \end{array}.$$

The starting and finishing methods are exactly the same except the opposite signs in their weight vectors. So the effective order five methods with $C(2)$ assumption come with the flavour of starting and finishing method in such a way that first we perturb the trajectory with the starting method then takes all the steps with the main method and at the end we apply the finishing method to get back to the originally trajectory.

3.3.1 Examples

Below are some examples of effective order five methods with five stages:

Example 3.3.1.

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{2}{3} & \frac{2}{9} & \frac{4}{9} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{37}{121} & -\frac{27}{121} & \frac{91}{363} & 0 & 0 \\ \frac{6}{11} & \frac{639}{2662} & -\frac{486}{1331} & \frac{861}{2662} & \frac{42}{121} & 0 \\ \hline & \frac{1939}{5820} & 0 & \frac{729}{1940} & -\frac{1452}{3395} & \frac{14641}{20370} \end{array}.$$

Example 3.3.2.

0	0	0	0	0	0
$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	0	0	0
$\frac{1}{2}$	0	$-\frac{3}{58}$	$\frac{16}{29}$	0	0
1	$\frac{19}{39}$	$\frac{2}{39}$	$-\frac{40}{39}$	$\frac{58}{39}$	0
	$\frac{23}{210}$	0	$\frac{16}{105}$	$\frac{58}{105}$	$\frac{13}{70}$

Example 3.3.3.

0	0	0	0	0	0
$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	0	0	0
$\frac{1}{2}$	0	$-\frac{11}{446}$	$\frac{117}{223}$	0	0
$\frac{3}{4}$	$\frac{391}{1336}$	$\frac{11}{1336}$	$-\frac{73}{334}$	$\frac{223}{334}$	0
	$\frac{52}{245}$	0	$\frac{22}{735}$	$\frac{223}{735}$	$\frac{334}{735}$

3.4 Effective order five methods with $D(1)$ assumption

To investigate the class of effective order five methods which obey the $D(1)$ simplifying assumption. The number of order conditions corresponding to each tree reduces from seventeen to ten. This assumption disposes of all order conditions except those associated with the following trees

The conditions on the α method together with type one and type two conditions are presented in Table 3.6. We can notice that the type one conditions are independent of any β factor. Therefore, by eliminating β from type two order conditions, we get

Table 3.6: Expressions for order condition up to order five

type one	type two
$\alpha_1 = 1$	$\alpha_9 = \frac{1}{5} + 4\beta_5 + 6\beta_3$
$\alpha_2 = \frac{1}{2}$	$\alpha_{10} = \frac{1}{10} + 2\beta_6 + \beta_5 + \frac{3}{2}\beta_3$
$\alpha_3 = \frac{1}{3}$	$\alpha_{11} = \frac{1}{15} + 2\beta_6 - \beta_5 - \frac{3}{2}\beta_3$
$\alpha_5 = \frac{1}{4}$	$\alpha_{13} = \frac{1}{20} + 2\beta_6$
$\alpha_6 = \frac{1}{8}$	
$\alpha_{12} = \frac{1}{30}$	

$$\tfrac{1}{4}\alpha_9 + \alpha_{13} = \alpha_{10}, \tag{3.16}$$

$$\alpha_{10} + \alpha_{11} - 2\alpha_{13} = \tfrac{1}{15}. \tag{3.17}$$

Thus the eight order conditions in terms of the coefficients of explicit Runge–Kutta method are

$$\begin{aligned}
\sum b_i &= 1,\\
\sum b_i c_i &= \tfrac{1}{2},\\
\sum b_i c_i^2 &= \tfrac{1}{3},\\
\sum b_i c_i^3 &= \tfrac{1}{4},\\
\sum b_i a_{ij} c_j^2 &= \tfrac{1}{8},\\
\sum b_i c_i a_{ij} a_{jk} c_k &= \tfrac{1}{30},\\
\sum b_i (\sum a_{ij} c_j - \tfrac{1}{2} c_i^2)^2 &= 0,\\
\sum b_i c_i^2 a_{ij} c_j + \sum b_i c_i a_{ij} c_j^2 - 2\sum b_i (a_{ij} c_j)^2 &= \tfrac{1}{15}.
\end{aligned}$$

A possible solution to these equations gives the effective order five method having five stages which is represented by the tableau.

$$\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\
\frac{2}{3} & 0 & -\frac{14}{33} & \frac{12}{11} & 0 & 0 \\
1 & \frac{2}{7} & \frac{1}{2} & -\frac{4}{7} & \frac{11}{14} & 0 \\
\hline
 & \frac{11}{100} & 0 & \frac{21}{50} & \frac{33}{100} & \frac{7}{50}
\end{array}.$$

Remark 3.4.1.

In Section 3.4, we investigated methods of effective order five satisfying the $D(1)$ simplifying assumption. While deriving these methods we observed that a subsidiary condition of the $C(2)$ simplifying assumption was also true, that is

$$\sum b_i \big(\sum a_{ij} c_j - \tfrac{1}{2} c_i^2 \big)^2 = 0,$$

therefore we come to conclusion that the class of effective order five methods which obey the $D(1)$ simplifying assumption gives no new generality in deriving such methods.

Example 3.4.2.

We give an example of an effective order five method with five stages, together with starting and finishing methods as β_0 and β_1. Let these three methods satisfy the $C(2)$ and $D(1)$ assumptions up to order five. The methods derived in this example are exactly the same as the effective order methods with starting and finishing methods found in Butcher [12].

The conditions on the α method are

$$\alpha_1 = 1 \tag{3.18}$$
$$\alpha_2 = \frac{1}{2} \tag{3.19}$$
$$\alpha_3 = \frac{1}{3} + 2\beta_2 \tag{3.20}$$
$$\alpha_5 = \frac{1}{4} + 3\beta_2 + 3\beta_3 \tag{3.21}$$
$$\alpha_9 = \frac{1}{5} + 4\beta_2 + 6\beta_3 + 4\beta_5 \tag{3.22}$$

From the condition on tree $\mathbf{t}_{11}$, we have an expression of the form

$$\begin{aligned} ((\beta\alpha) - (E\beta))(\mathbf{t}_{11}) &= 0, \\ \alpha_{11} + \beta_3\alpha_2 - \beta_7 - 2\beta_6 - 2\beta_4 - \beta_3 - \tfrac{1}{15} &= 0, \end{aligned}$$

or

$$\begin{aligned} \alpha_{11} &= \frac{1}{15} + \frac{1}{2}\beta_3 + \beta_3 + \beta_5 + (\beta_3 - \beta_5), \\ &= \frac{1}{15} + \frac{5}{2}\beta_3. \end{aligned}$$

From the equation

$$\frac{1}{120} = \sum b_i a_{ij}^3 c_j = \sum b_i a_{ij} a_{ij}^2 c_j = \sum b_j (1 - c_j) a_{ij}^2 c_j = \sum b_j a_{ij}^2 c_j - \sum b_j a_{ij}^2 c_j^2$$
$$\frac{1}{120} = \frac{1}{24} - \frac{1}{2}(\frac{1}{15} + \frac{5}{2}\beta_3),$$

which implies that $\beta_3 = 0$. So the order condition for effective order five method (together with $C(2)$ and $D(1)$) become

$$\alpha_1 = 1, \qquad \alpha_2 = \frac{1}{2}, \qquad \alpha_3 = \frac{1}{3},$$
$$\alpha_5 = \frac{1}{4}, \qquad \alpha_9 = \frac{1}{5} + 4\beta_5, \qquad \alpha_{11} = \frac{1}{15}.$$

An explicit method α which satisfies the above order conditions is given by

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 \\
\frac{2}{5} & 0 & \frac{2}{5} & 0 & 0 & 0 \\
\frac{1}{2} & \frac{3}{16} & 0 & \frac{5}{16} & 0 & 0 \\
1 & \frac{1}{4} & 0 & -\frac{5}{4} & 2 & 0 \\
\hline
 & \frac{1}{6} & 0 & 0 & \frac{2}{3} & \frac{1}{6}
\end{array}.
$$

For finding the starting and finishing methods, since all these methods satisfy $C(2)$ and $D(1)$ conditions. We have $\beta_0 = E\beta$, $\beta_1 = \beta^{-1}E$ and $\beta(t) = 0$ for all trees of order less than or equal to three. Therefore we have,

$$
\begin{aligned}
\beta_0(\mathbf{t}_1) = \beta_1(\mathbf{t}_1) &= 1, \\
\beta_0(\mathbf{t}_2) = \beta_1(\mathbf{t}_2) &= \frac{1}{2}, \\
\beta_0(\mathbf{t}_3) = \beta_1(\mathbf{t}_3) &= \frac{1}{3},
\end{aligned}
$$

and

$$
\begin{aligned}
\beta_0(\mathbf{t}_5) &= \frac{121}{480}, \\
\beta_1(\mathbf{t}_5) &= \frac{119}{480}.
\end{aligned}
$$

So a suitable starting method, which advances the solution one step forward is given by the tableau

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 \\
\frac{2}{5} & 0 & \frac{2}{5} & 0 & 0 & 0 \\
\frac{3}{4} & \frac{75}{64} & -\frac{9}{4} & \frac{117}{64} & 0 & 0 \\
1 & -\frac{37}{36} & \frac{7}{3} & -\frac{3}{4} & \frac{4}{9} & 0 \\
\hline
 & \frac{19}{144} & 0 & \frac{25}{48} & \frac{2}{9} & \frac{1}{8}
\end{array},
$$

and a finishing method method which moves the solution one step backward is given by the tableau

$$\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 \\
\frac{2}{5} & 0 & \frac{2}{5} & 0 & 0 & 0 \\
\frac{3}{4} & \frac{161}{192} & -\frac{19}{12} & \frac{287}{192} & 0 & 0 \\
1 & -\frac{27}{28} & \frac{19}{7} & -\frac{291}{196} & \frac{36}{49} & 0 \\
\hline
 & \frac{7}{48} & 0 & \frac{475}{1008} & \frac{2}{7} & \frac{7}{72}
\end{array}.$$

Chapter 4

Symplectic effective order methods

Effective order was originally introduced to bypass the fifth order barrier, that is, fifth order methods are impossible with just five stages. A detailed discussion concerning the structure and properties of effective order methods is given in Chapter 3. Effective order also has an application to implicit Runge–Kutta methods for the solution of stiff problems; see[4][7][15]. Application to symplectic integration referred to as *processing* was introduced in [58]. An important property of Hamiltonian system is that they conserve the energy of the mechanical system. Symplecticity is the property of Hamiltonian systems which is explained in Chapter 1.

The aim of this chapter is to derive symplectic Runge–Kutta methods with the order conditions replaced by effective order conditions. For order four, which is our main focus, the interplay between the symplectic conditions and the effective order conditions leads to simple criteria for acceptable methods.

In addition to these conditions, we will also require the eigenvalues of A (the coefficient matrix) in the method tableau to have only real eigenvalues. The advantages of having real eigenvalues enable in the use of transformation matrix T which was introduced by Butcher in [14]. By including transformations into the stage iterations, we can reduce the computational cost for these methods. For large Hamiltonian problems there is a significant gain in efficiency.

The chapter is organised as follows. In Section 4.1 we will discuss symplectic Runge–Kutta methods and properties of their composition group. From these starting points, we will derive a new symplectic effective order method in Section 4.2. Section 4.3 deals with some example methods. Section 4.4 will be concerned with the efficient implementation of the new method.

The results derived in this chapter are presented in [27].

4.1 Symplectic integrators

We are interested in solving autonomous initial value problems

$$y'(x) = f(y(x)), \quad f : \mathbb{R}^N \to \mathbb{R}^N, \tag{4.1}$$

which possess first integrals of the form $\langle y, Qy\rangle$ where Q is a symmetric matrix. For any such problem, we define a Runge–Kutta method R to be *canonical* or *symplectic* if the quadratic first integral is conserved; that is

$$\langle y_1, Qy_1\rangle = \langle y_0, Qy_0\rangle.$$

Pioneering work in the development of canonical Runge–Kutta methods was due to Sanz–Serna [67], Lasagni [55], Cooper [31] and Suris [73], their idea was based on the properties of algebraic stability introduced for stiff problems by Burrage and Butcher [5] and Crouzeix [32].

An application of a Runge–Kutta method (2.1) to solve (4.1) results in,

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} f(Y_j).$$

Since, $\langle Y_i, Qf(Y_i)\rangle = 0$, it follows that

$$\langle y_0, Qf(Y_i)\rangle + h\sum_{j=1}^{s} a_{ij}\langle f(Y_j), Qf(Y_i)\rangle = 0, \qquad i = 1, 2, \ldots, s. \tag{4.2}$$

The output value is $y_1 = y_0 + h\sum_{i=1}^{s} b_i f(Y_i)$, and it follows that

$$\begin{aligned}\langle y_1, Qy_1\rangle = \langle y_0, Qy_0\rangle &+ h\sum_{i=1}^{s} b_i\langle y_0, Qf(Y_i)\rangle \\ &+ h\sum_{j=1}^{s} b_j\langle f(Y_j), Qy_0\rangle + h^2\sum_{i,j=1}^{s} b_i b_j\langle f(Y_i), Qf(Y_j)\rangle. \end{aligned}\tag{4.3}$$

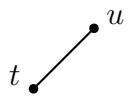
From (4.2) and (4.3), it follows that,

$$\langle y_1, Qy_1\rangle = \langle y_0, Qy_0\rangle,$$

provided that,

$$\mathrm{diag}(b)A + A^{\intercal}\,\mathrm{diag}(b) - bb^{\intercal} = 0.$$

If t, u are rooted trees, then the product is defined by joining the roots and regarding the root of t as also the root of the product. This is illustrated in the diagram



For a symplectic method R, it always holds that

$$\Phi(tu) + \Phi(ut) = \Phi(t)\Phi(u),$$

because if $\Phi(t) = \sum_{i=1}^{s} b_i\chi_i$ and $\Phi(u) = \sum_{i=1}^{s} b_i\psi_i$ then

$$\begin{aligned}
\Phi(tu) + \Phi(ut) &= \sum_{i=1}^{s} b_i\chi_i a_{ij}\psi_j + \sum_{j=1}^{s} b_j\psi_j a_{ji}\chi_i \\
&= \sum_{i,j=1}^{s} (b_i a_{ij} + b_j a_{ji})\chi_i\psi_j \\
&= \sum_{i=1}^{s} b_i\chi_i \sum_{j=1}^{s} b_j\psi_j \\
&= \Phi(t)\Phi(u).
\end{aligned}$$

Remark 4.1.1. *We verify that if both R and $\widetilde{R}$ have the symplectic property then $R\widetilde{R}$ is also symplectic.*

Let us consider two Runge–Kutta methods R and $\widetilde{R}$, with s and $\widetilde{s}$ stages respectively given by

$$R = \begin{array}{c|c} c & A \\ \hline & b^{\intercal} \end{array}, \qquad \widetilde{R} = \begin{array}{c|c} \widetilde{c} & \widetilde{A} \\ \hline & \widetilde{b}^{\intercal} \end{array}.$$

We assume that these two methods R, $\widetilde{R}$ are symplectic such that

$$\begin{aligned}
\mathrm{diag}(b)A + A^{\intercal}\mathrm{diag}(b) - bb^{\intercal} = 0, \\
\mathrm{diag}(\widetilde{b})\widetilde{A} + \widetilde{A}^{\intercal}\mathrm{diag}(\widetilde{b}) - \widetilde{b}\widetilde{b}^{\intercal} = 0.
\end{aligned}$$

The product $R\widetilde{R}$ is defined by first carrying out the calculations in R and then applying method $\widehat{R}$ to the output which is given by

$$R\widetilde{R} = \begin{array}{c|cc} c & A & 0 \\ (b^\intercal \mathbf{1}_n)\mathbf{1}_{\widetilde{n}} + \widetilde{c} & \mathbf{1}_{\widetilde{n}} b^\intercal & \widetilde{A} \\ \hline & b^\intercal & \widetilde{b}^\intercal \end{array},$$

where for n a positive integer, $\mathbf{1}_n$ is the vector in $\mathbb{R}^n$ with every component equal to 1.

Composition $R\widetilde{R}$ of two Runge–Kutta methods is symplectic if,

$$\mathrm{diag}(\bar{b})\bar{A} + \bar{A}^\intercal \mathrm{diag}(\bar{b}) - \bar{b}\bar{b}^\intercal = 0,$$

For this, we calculate the terms one by one,

$$\mathrm{diag}(\bar{b})\bar{A} = \begin{bmatrix} \mathrm{diag}(b)A & 0 \\ \widetilde{b}b^\intercal & \mathrm{diag}(\widetilde{b})\widetilde{A} \end{bmatrix},$$

$$\bar{A}^\intercal \mathrm{diag}(\bar{b}) = \begin{bmatrix} A^\intercal \,\mathrm{diag}(b) & b\widetilde{b}^\intercal \\ 0 & \widetilde{A}^\intercal \,\mathrm{diag}(\widetilde{b}) \end{bmatrix},$$

$$\bar{b}\bar{b}^\intercal = \begin{bmatrix} bb^\intercal & b\widetilde{b}^\intercal \\ \widetilde{b}b^\intercal & \widetilde{b}\widetilde{b}^\intercal \end{bmatrix},$$

$$\mathrm{diag}(\bar{b})\bar{A} + \bar{A}^\intercal \mathrm{diag}(\bar{b}) - \bar{b}\bar{b}^\intercal = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

This implies the symplecticity of $R\widetilde{R}$. □

Remark 4.1.2. *The symplectic* **B***-series form a subgroup.*

Define the group G as the set of mapping from trees to real numbers corresponding to the elementary weight functions for equivalence classes of Runge–Kutta methods. Let α, β be the two members of the group G, we define the product $\alpha\beta$ of two multiplicative mappings of trees to real numbers as

$$\begin{aligned}(\alpha\beta)(tu) = \alpha(tu) &+ \sum_{t'<t,u'<u} \alpha(t\backslash t')\alpha(u\backslash u')\beta(t'u') \\ &+ \alpha(u)\sum_{t'<t} \alpha(t\backslash t')\beta(t').\end{aligned} \tag{4.4}$$

$$\begin{aligned}(\alpha\beta)(ut) = \alpha(ut) &+ \sum_{t'<t,u'<u} \alpha(u\backslash u')\alpha(t\backslash t')\beta(u't') \\ &+ \alpha(t)\sum_{u'<u} \alpha(u\backslash u')\beta(u'),\end{aligned} \tag{4.5}$$

where t', u' are subtrees of t and u such that sharing the same root with t and u respectively. Moreover $t\backslash t'$ and $u\backslash u'$ are the remaining parts of t and u after chopping the subtrees t' and u' respectively.

By adding (4.4) and (4.5) we get

$$\begin{aligned}(\alpha\beta)(tu) + (\alpha\beta)(ut) = \alpha(tu) + \alpha(ut) &+ \\ \sum_{t'<t,u'<u} \alpha(t\backslash t')\alpha(u\backslash u')\beta(t'u') &+ \\ \sum_{t'<t,u'<u} \alpha(u\backslash u')\alpha(t\backslash t')\beta(u't') &+ \\ \alpha(u)\sum_{t'<t} \alpha(t\backslash t')\beta(t') &+ \\ \alpha(t)\sum_{u'<u} \alpha(u\backslash u')\beta(u').\end{aligned} \tag{4.6}$$

For Runge–Kutta methods α and β to be symplectic methods, it always holds that

$$\begin{aligned}\alpha(tu) + \alpha(ut) &= \alpha(t)\alpha(u), \\ \beta(tu) + \beta(ut) &= \beta(t)\beta(u).\end{aligned}$$

Thus (4.6) takes the form

$$(\alpha\beta)(tu) + (\alpha\beta)(ut) = \alpha(t)\alpha(u) + \sum_{t'<t, u'<u} \alpha(t\backslash t')\alpha(u\backslash u')\beta(t')\beta(u') + \alpha(u)\sum_{t'<t}\alpha(t\backslash t')\beta(t') + \alpha(t)\sum_{u'<u}\alpha(u\backslash u')\beta(u'),$$

or

$$(\alpha\beta)(tu) + (\alpha\beta)(ut) = \alpha(t)\alpha(u) + \sum_{t'<t}\alpha(t\backslash t')\beta(t')\sum_{u'<u}\alpha(u\backslash u')\beta(u') + \alpha(u)\sum_{t'<t}\alpha(t\backslash t')\beta(t') + \alpha(t)\sum_{u'<u}\alpha(u\backslash u')\beta(u'),$$

or

$$(\alpha\beta)(tu) + (\alpha\beta)(ut) = \alpha(t)\alpha(u) + \big(\alpha\beta(t) - \alpha(t)\big)\big((\alpha\beta(u) - \alpha(u)\big) \alpha(u)\big(\alpha\beta(t) - \alpha(t)\big) + \alpha(t)\big(\alpha\beta(u) - \alpha(u)\big),$$

This implies that

$$(\alpha\beta)(tu) + (\alpha\beta)(ut) = (\alpha\beta)(t).(\alpha\beta)(u). \quad \square$$

4.2 Derivation of Symeff43 method

We will derive a three stage symplectic Runge–Kutta method for which A has only real eigenvalues. Instead of having full set of order four conditions we require effective order four conditions. By doing so we can reduce the number of order conditions needing to be satisfied.

As we have discussed in Chapter 3, the effective order is built on the assumption that a function from trees to real number is associated with each Runge–Kutta method and is defined from the elementary weights of the method. Here α corresponds to method φ, β to method ϕ and E the exact method to ψ.

If we denote by E the exact trajectory through a unit step size h, then the expression $\alpha(t) = E(t)$ means that for tree t, the corresponding order condition is satisfied. The Runge–Kutta method is said to have *effective order* p, if there exists a corresponding Runge–Kutta method ϕ with element β, such that $(\beta\alpha)(t) = (E\beta)(t)$.

For this, the products $\beta\alpha$ and $E\beta$ for the trees up to fourth order are presented in

i	t_i	$(\beta\alpha)(t_i)$	$(E\beta)(t_i)$
1		α_1	1
2		$\beta_2 + \alpha_2$	$\beta_2 + \frac{1}{2}$
3		$\beta_3 + \alpha_3$	$\beta_3 + 2\beta_2 + \frac{1}{3}$
4		$\beta_4 + \alpha_1\beta_2 + \alpha_4$	$\beta_4 + \beta_2 + \frac{1}{6}$
5		$\beta_5 + \alpha_5$	$\beta_5 + 3\beta_3 + 3\beta_2 + \frac{1}{4}$
6		$\beta_6 + \beta_2\alpha_2 + \alpha_6$	$\beta_6 + \beta_4 + \beta_3 + \frac{3}{2}\beta_2 + \frac{1}{8}$
7		$\beta_7 + \beta_3\alpha_1 + \alpha_7$	$\beta_7 + 2\beta_4 + \beta_2 + \frac{1}{12}$
8		$\beta_8 + \beta_4\alpha_1 + \beta_2\alpha_2 + \alpha_8$	$\beta_8 + \beta_4 + \frac{1}{2}\beta_2 + \frac{1}{24}$

Table 4.1: Expressions for $\beta\alpha$ and $E\beta$ for trees up to order four

Table 4.1. We write $\alpha_i = \alpha(t_i)$, $\beta_i = \beta(t_i)$ and without loss of generality we assume that $\beta_1 = 0$

By equating the expressions in the third and fourth columns in Table 4.1 we therefore find the effective order four conditions as

$$\alpha_1 = 1, \tag{4.7}$$
$$\alpha_2 = \tfrac{1}{2}, \tag{4.8}$$
$$\alpha_3 = \tfrac{1}{3} + 2\beta_2, \tag{4.9}$$
$$\alpha_4 = \tfrac{1}{6}, \tag{4.10}$$
$$\alpha_5 = \tfrac{1}{4} + 3\beta_3 + 3\beta_2, \tag{4.11}$$
$$\alpha_6 = \tfrac{1}{8} + \beta_4 + \beta_3 + \beta_2, \tag{4.12}$$
$$\alpha_7 = \tfrac{1}{12} + 2\beta_4 - \beta_3 + \beta_2, \tag{4.13}$$
$$\alpha_8 = \tfrac{1}{24}, \tag{4.14}$$

Equations (4.7), (4.8), (4.10) and (4.14) involves α values and do not depend on β values. Eliminating the β values from (4.9), (4.11), (4.12) and (4.13), we get

$$\alpha_3 - \alpha_5 + 2\alpha_6 - \alpha_7 = \tfrac{1}{4}. \tag{4.15}$$

The conditions for a method to have effective order 4 can now be written down. They are

$$\begin{aligned} \alpha_1 &= 1, \\ \alpha_2 &= \tfrac{1}{2}, && (4.16) \\ \alpha_4 &= \tfrac{1}{6}, && (4.17) \\ \alpha_8 &= \tfrac{1}{24}, && (4.18) \\ \alpha_3 - \alpha_5 + 2\alpha_6 - \alpha_7 &= \tfrac{1}{4}. && (4.19) \end{aligned}$$

where α_i is the elementary weight for the method corresponding to tree t_i, where

$$\begin{aligned} &\mathbf{t}_1 = \tau, && \mathbf{t}_2 = \tau\tau, && \mathbf{t}_3 = \tau\tau.\tau, && \mathbf{t}_4 = \tau.\tau\tau \\ &\mathbf{t}_5 = (\tau\tau.\tau)\tau, && \mathbf{t}_6 = \tau\tau.\tau\tau, && \mathbf{t}_7 = \tau(\tau\tau.\tau), && \mathbf{t}_8 = \tau(\tau.\tau\tau). \end{aligned}$$

Because the method is symplectic, therefore (4.16) is automatically satisfied, also $\alpha_3 + \alpha_4 = \frac{1}{2}$, enabling (4.10) to be replaced by $\alpha_3 = \frac{1}{3}$. Furthermore, because the rooted trees $\mathbf{t}_6$ and $\mathbf{t}_8$ belong to the same underlying superfluous tree, it follows that $\alpha_6 = \frac{1}{8}$ and $\alpha_8 = \frac{1}{24}$. Also note that $\mathbf{t}_5$ and $\mathbf{t}_7$ are related through a non-superfluous tree so that $\alpha_5 + \alpha_7 = \frac{1}{3}$. Taking into account these simplifications, we see that (4.18) and (4.19) can be deleted because they are automatically satisfied. The conditions for effective order 4 can now be written in terms of the tableau coefficients as

$$\begin{aligned} \sum_{i=1}^{3} b_i &= 1, \\ \sum_{i=1}^{3} b_i c_i^2 &= \tfrac{1}{3}. \end{aligned}$$

Because $\operatorname{diag}(b)A + A^{\mathsf{T}} \operatorname{diag}(b) = bb^{\mathsf{T}}$, it follows that $\operatorname{diag}(b)A - \frac{1}{2}bb^{\mathsf{T}}$ is a skew-symmetric matrix which we will write as $\operatorname{diag}(b)S \operatorname{diag}(b)$. Because $A\mathbf{1} = c$, it follows that S has the form

$$S = c\mathbf{1}^{\mathsf{T}} - \mathbf{1}c^{\mathsf{T}} + t \begin{bmatrix} 0 & b_3 & -b_2, \\ -b_3 & 0 & b_1 \\ b_2 & -b_1 & 0 \end{bmatrix}.$$

The matrix A is found to be

$$\begin{aligned} A &= (\tfrac{1}{2}\mathbf{1}\mathbf{1}^{\mathsf{T}} + S)\,\mathrm{diag}(b) \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2}+c_1-c_2+tb_3 & \frac{1}{2}+c_1-c_3-tb_2 \\ \frac{1}{2}+c_2-c_1-tb_3 & \frac{1}{2} & \frac{1}{2}+c_2-c_3+tb_1 \\ \frac{1}{2}+c_3-c_1+tb_2 & \frac{1}{2}+c_3-c_2-tb_1 & \frac{1}{2} \end{bmatrix} \mathrm{diag}(b) \end{aligned} \tag{4.20}$$

and the characteristic polynomial is found to be

$$w^3 - \tfrac{1}{2}w^2 + \tfrac{1}{12}w - \theta(w - \tfrac{1}{2}), \qquad \theta = -t^2 b_1 b_2 b_3. \tag{4.21}$$

Verification of characteristic polynomial

Let the characteristic polynomial be

$$w^3 - \bar{\alpha}w^2 + \bar{\beta}w - \bar{\gamma}$$

where by (4.20)

$$\begin{aligned} \bar{\alpha} &= a_{11} + a_{22} + a_{33}, \\ \bar{\beta} &= \big(a_{11}a_{22} - a_{12}a_{21}\big) + \big(a_{33}a_{22} - a_{32}a_{23}\big) + \big(a_{11}a_{33} - a_{13}a_{31}\big), \\ \bar{\gamma} &= \det(A). \end{aligned}$$

We check these one by one.

$$\begin{aligned} \bar{\alpha} &= \frac{1}{2}b_1 + \frac{1}{2}b_2 + \frac{1}{2}b_3 \\ &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} \bar{\beta} = b_1b_2(c_1 - c_2 + tb_3)^2 + b_2b_3(c_2 - c_3 + tb_1)^2 + b_3b_1(c_3 - c_1 + tb_2)^2 \\ = b_1b_2(c_1^2 + c_2^2 + t^2b_3^2 - 2c_1c_2 - 2c_2tb_3 + 2c_1tb_3) + \\ b_2b_3(c_2^2 + c_3^2 + t^2b_1^2 - 2c_2c_3 - 2c_3tb_1 + 2c_2tb_1) + \\ b_3b_1(c_3^2 + c_1^2 + t^2b_2^2 - 2c_3c_1 - 2c_1tb_2 + 2c_3tb_2) \end{aligned}$$

$$\bar{\beta} = t^2(b_1b_2b_3)(b_1+b_2+b_3) + t(2c_2b_1b_2b_3 + 2c_1b_1b_2b_3 - 2c_3b_1b_2b_3$$
$$-2c_2b_1b_2b_3 - 2c_1b_1b_2b_3 + 2c_3b_1b_2b_3) +$$
$$b_1b_2(c_1^2+c_2^2-2c_1c_2) + b_2b_3(c_2^2+c_3^2-2c_2c_3) + b_3b_1(c_3^2+c_1^2-2c_3c_1)$$

which becomes

$$\bar{\beta} = t^2(b_1b_2b_3) + b_1c_1^2(b_2+b_3) + b_2c_2^2(b_1+b_3) + b_3c_3^2(b_2+b_1) +$$
$$-2b_1b_2c_1c_2 - 2b_2b_3c_2c_3 - 2b_1b_3c_1c_3$$
$$\bar{\beta} = t^2(b_1b_2b_3) + b_1c_1^2(1-b_1) + b_2c_2^2(1-b_2) + b_3c_3^2(1-b_3) +$$
$$-2b_1b_2c_1c_2 - 2b_2b_3c_2c_3 - 2b_1b_3c_1c_3$$
$$\bar{\beta} = t^2(b_1b_2b_3) + \sum_{i=1}^{3} b_ic_i^2 - \big(\sum_{i}^{3} b_ic_i\big)^2$$
$$\bar{\beta} = t^2(b_1b_2b_3) + \tfrac{1}{12}.$$

Finally,

$$\det(A) = b_1b_2b_3 \det(\widetilde{A})$$

where $\widetilde{A}$ is the matrix in (4.20). We can simplify the matrix $\widetilde{A}$ by subtracting the first column from each of the other columns and then the first row from each of the other rows. Thus the matrix $\widetilde{A}$ takes the form

$$\widetilde{A} = \begin{bmatrix} \frac{1}{2} & c_1-c_2+tb_3 & c_1-c_3-tb_2 \\ c_2-c_1-tb_3 & 0 & t(b_1+b_2+b_3) \\ c_3-c_1+tb_2 & -t(b_1+b_2+b_3) & 0 \end{bmatrix}$$

and the determinant of $\det(\widetilde{A})$ is $\frac{1}{2}t^2$. Therefore $\det(A)$ becomes

$$\det(A) = \frac{1}{2}t^2(b_1b_2b_3).$$

Now consider the possible choice of θ to ensure that there are three real eigenvalues of A. We get a guide to this from Figure 4.1, where $\theta = \theta_0$ is chosen so that the cubic curve is tangential to the straight line shown in the figure. The critical value is

$$\theta_0 = \tfrac{1}{6}2^{2/3} + \tfrac{5}{24}2^{1/3} + \tfrac{1}{4} \approx 0.7770503941. \tag{4.22}$$

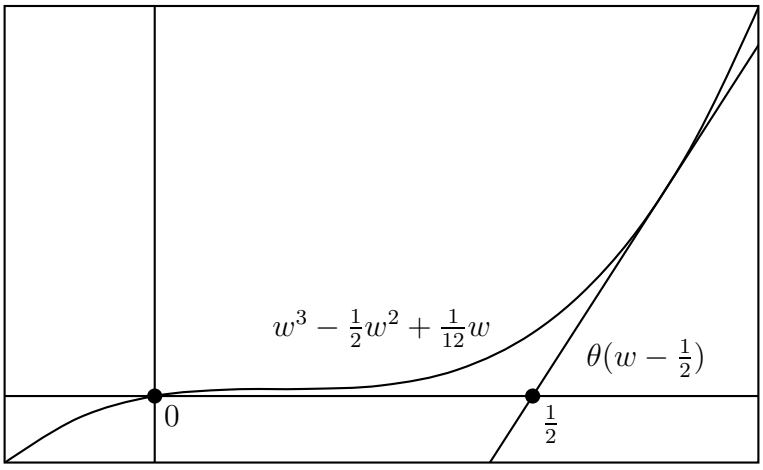


Figure 4.1: Intersection of the curves $y = w^3 - \frac{1}{2}w^2 + \frac{1}{12}w$ and $y = \theta(w - \frac{1}{2})$

For any $\theta > \theta_0$, there will be three distinct zeros of the polynomial (4.21) and we will aim to choose convenient values of θ to obtain a suitable method. Such convenient values can be found by choosing one of the zeros of (4.21) and then evaluating θ to ensure that the cubic equation is satisfied.

We wish to choose the parameters of the method so that all elements of the tableau are rational. This will mean that t, b_1, b_2, b_3 are rational. One way of achieving this is by defining

$$c_1 = c_2 = 0, \quad c_3 = \frac{2}{3}, \quad b_1 = \frac{1}{4+4k}, \quad b_2 = \frac{k}{4+4k}, \quad b_3 = \frac{3}{4}. \tag{4.23}$$

and then choosing k so that t has a rational value. It is found that this is possible only if $-16\theta k/3$ is a perfect square. Write this quantity as u^2 and we find that

$$t = 2u - \frac{32\theta}{3u}, \quad b_1 = \frac{-4\theta}{-16\theta + 3u^2}, \quad b_2 = \frac{3u^2/4}{-16\theta + 3u^2}$$

We construct a single method using (4.23). For $\theta = \frac{19}{8}$, a suitable choice of u is $\frac{19}{9}$. This gives the Runge–Kutta method

$$\begin{array}{c|ccc} 0 & \frac{27}{140} & \frac{76}{105} & -\frac{11}{12} \\ 0 & \frac{171}{70} & -\frac{19}{280} & -\frac{19}{8} \\ \frac{2}{3} & \frac{6}{7} & -\frac{95}{168} & \frac{3}{8} \\ \hline & \frac{27}{70} & -\frac{19}{140} & \frac{3}{4} \end{array} . \tag{4.24}$$

To verify the effective order of the method, compute the elementary weights:

$$\alpha_1 = 1, \quad \alpha_2 = \frac{1}{2}, \quad \alpha_3 = \frac{1}{3}, \quad \alpha_4 = \frac{1}{6}, \quad \alpha_5 = \frac{2}{9}, \quad \alpha_6 = \frac{1}{8}, \quad \alpha_7 = \frac{1}{9}, \quad \alpha_8 = \frac{1}{24},$$

and we see that (4.7), (4.8), (4.10), (4.14) and (4.15), are satisfied. Furthermore, values of the β_i to satisfy (4.9), (4.11), (4.12), and (4.13) are given by

$$\beta_1 = \beta_2 = 0, \quad \beta_3 = -\frac{1}{108}, \quad \beta_4 = \frac{1}{108}.$$

4.2.1 Starting and finishing methods

To obtain a starting method to enable the method (4.24) to exhibit effective order four behaviour, we need to construct a Runge–Kutta method satisfying the order conditions

$$\sum b_i = 0, \tag{4.25}$$

$$\sum b_i c_i = 0, \tag{4.26}$$

$$\sum b_i c_i^2 = -\frac{1}{108}, \tag{4.27}$$

$$\sum b_i a_{ij} c_j = \frac{1}{108}. \tag{4.28}$$

It is possible to construct a method satisfying these conditions with $s = 3$ and at the same time to find a corresponding finishing method. Preliminary experiments have demonstrated the importance of both these methods being symplectic and this has been imposed as a requirement. To build such a starting method with diagonally implicit structure, it is only necessary to choose a tableau of the form

$\frac{1}{2}b_1$	$\frac{1}{2}b_1$		
$b_1 + \frac{1}{2}b_2$	b_1	$\frac{1}{2}b_2$	
$b_1 + b_2 + \frac{1}{2}b_3$	b_1	b_2	$\frac{1}{2}b_3$
	b_1	b_2	b_3

where $b_2 = -b_1 - b_3$, to satisfy (4.25) and (4.26) and

$$b_1(\tfrac{1}{2}b_1)^2 - (b_1 + b_3)(\tfrac{1}{2}b_1 - \tfrac{1}{2}b_3)^2 + b_3(\tfrac{1}{2}b_3)^2 = -\tfrac{1}{108},$$

to satisfy (4.27) and (4.28). This simplifies to $b_1 b_3 (b_1 + b_3) = -\frac{1}{27}$. Let $b_1 + b_3 = t$ so that b_1, b_3 are the roots of the equation

$$x^2 - tx - \frac{1}{27t} = 0.$$

The following solutions are possible:

$$\begin{aligned} t &= \frac{1}{3}, \qquad x = \frac{1 \pm \sqrt{5}}{6}, \\ t &= \frac{2}{3}, \qquad x = \frac{2 \pm \sqrt{6}}{6}. \end{aligned}$$

Using the $t = \frac{1}{3}$ case, we obtain the starting method given by using $\pm = +$ and $\mp = -$ in the following tableau

$$\begin{array}{c|cccc} \frac{\pm 1+\sqrt{5}}{12} & \frac{\pm 1+\sqrt{5}}{12} & & & \\ \frac{\sqrt{5}}{6} & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{6} & & \\ \frac{\mp 1+\sqrt{5}}{12} & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{3} & \frac{\pm 1-\sqrt{5}}{12} & \\ \hline & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{3} & \frac{\pm 1-\sqrt{5}}{6} & \end{array}.$$

To obtain a suitable finishing method, use the signs $\pm = -$ and $\mp = +$ in this tableau. The proposed finishing method exactly undoes the work of the starting method.

4.3 Examples

Example 4.3.1.

$$\begin{array}{c|ccc} 0 & \frac{127}{827} & \frac{16989}{13232} & -\frac{23}{16} \\ 0 & \frac{107315}{14886} & -\frac{189}{6616} & -\frac{517}{72} \\ \frac{3}{4} & \frac{4445}{4962} & -\frac{3997}{6616} & \frac{3}{8} \\ \hline & \frac{254}{827} & -\frac{189}{3308} & \frac{3}{4} \end{array}.$$

The characteristic polynomial is given by

$$P(w) = w^3 - \frac{1}{2}w^2 - \frac{883}{72}w + \frac{889}{144}.$$

Three distinct real eigenvalues are

$$\sigma = \left\{ \frac{7}{2}, \frac{-18 + 17\sqrt{2}}{12}, \frac{-18 - 17\sqrt{2}}{12} \right\}.$$

Example 4.3.2.

$$\begin{array}{c|ccc} 0 & \frac{541}{3950} & \frac{117999}{537200} & -\frac{97}{272} \\ 0 & \frac{1123657}{402900} & -\frac{189}{15800} & -\frac{1133}{408} \\ \frac{3}{4} & \frac{162841}{402900} & -\frac{30219}{268600} & \frac{3}{8} \\ \hline & \frac{541}{1975} & -\frac{189}{7900} & \frac{3}{4} \end{array}.$$

The characteristic polynomial is given by

$$P(w) = w^3 - \frac{1}{2}w^2 - \frac{10205}{13872}w + \frac{3787}{9248}.$$

We get the three distinct real eigenvalues as

$$\sigma = \left\{ \frac{21}{34}, \frac{-12 + 43\sqrt{15}}{204}, \frac{-12 - 43\sqrt{15}}{204} \right\}.$$

Example 4.3.3.

$$\begin{array}{c|ccc} 0 & \frac{17461}{131318} & \frac{11756409}{57254648} & -\frac{295}{872} \\ 0 & \frac{2378589803}{644114790} & -\frac{4185}{525272} & -\frac{144593}{39240} \\ \frac{3}{4} & \frac{16570489}{42940986} & -\frac{5394713}{57254648} & \frac{3}{8} \\ \hline & \frac{17461}{65659} & -\frac{4185}{262636} & \frac{3}{4} \end{array}.$$

The characteristic polynomial is given by

$$P(w) = w^3 - \frac{1}{2}w^2 - \frac{1986949}{2138580}w + \frac{541291}{1069290}.$$

Three distinct real eigenvalues are

$$\sigma = \left\{ \frac{62}{109}, \frac{-225 + 2761\sqrt{5}}{6540}, \frac{-225 - 2761\sqrt{5}}{6540} \right\}.$$

Example 4.3.4.

$$\begin{array}{c|ccc} 0 & \frac{5149}{37682} & \frac{3487965}{15977168} & -\frac{301}{848} \\ 0 & \frac{101924455}{35948628} & -\frac{1755}{150728} & -\frac{10775}{3816} \\ \frac{3}{4} & \frac{4824613}{11982876} & -\frac{886405}{7988584} & \frac{3}{8} \\ \hline & \frac{5149}{18841} & -\frac{1755}{75364} & \frac{3}{4} \end{array}.$$

The characteristic polynomial is given by

$$P(w) = w^3 - \frac{1}{2}w^2 - \frac{300977}{404496}w + \frac{334685}{808992}.$$

Three distinct real eigenvalues are

$$\sigma = \left\{ \frac{65}{106}, \frac{-36 + 127\sqrt{17}}{636}, \frac{-36 - 127\sqrt{17}}{636} \right\}.$$

4.4 Efficient implementation

We will follow the standard approach to the implementation of implicit Runge–Kutta methods, which is to use a modified form of Newton iterations. In carrying out this approach the Jacobian matrix has a key role.

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} & \dots & \frac{\partial f_1}{\partial y_N} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} & \dots & \frac{\partial f_2}{\partial y_N} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_N}{\partial y_1} & \frac{\partial f_N}{\partial y_2} & \dots & \frac{\partial f_N}{\partial y_N} \end{bmatrix}. \tag{4.29}$$

Because the evaluation of the Jacobian matrix J is typically an expensive part of the overall calculation, it is normally evaluated much less than once per step. It is important to avoid unnecessary re-evaluations and at the same time avoid repeated LU factorizations of $I - h\lambda J$.

Furthermore, there is much less need to re-evaluate the Jacobian because it changes

very slowly and the Newton iteration schemes are dependent on J and will maintain their rapid convergence even if we use a past value of J. Although it may be efficient to use the same factorization over many steps, in the experimental implementation adopted in this chapter, a single evaluation and factorization is carried out in each time step.

In each step of the method, it is required to calculate

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i F_i,$$

where the F_i are found from the simultaneous equations

$$Y_i = y_n + h \sum_{j=1}^{s} a_{ij} F_j, \tag{4.30}$$

$$F_i = f(Y_i), \tag{4.31}$$

for $i = 1, 2, \ldots, s$.

To express (4.30) and (4.31) in a compact form, write

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix}, \qquad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_s \end{bmatrix}$$

and the system (4.30) becomes

$$Y - 1 \otimes y_n - h(A \otimes I_m)F = 0. \tag{4.32}$$

We will look at the process of taking a single Newton step making use of the matrix

$$M = I_s \otimes I_N - h(A \otimes J), \tag{4.33}$$

which we assume has been computed in advance. For the single step we are considering, existing approximations Y and F are updated to $Y - \rho$ and $F - \varphi$ respectively

where

$$\rho = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_s \end{bmatrix}, \qquad \varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_s \end{bmatrix}.$$

At the end of an iteration we need the updated values to satisfy the linear system corresponding to (4.32), that is

$$\rho - h(A \otimes I_N)\varphi = Y - 1 \otimes y_n - h(A \otimes I_N)F. \tag{4.34}$$

To approximately satisfy the *nonlinear* conditions given by (4.31) we need to relate ρ_i and φ_i by

$$\varphi_i - J\rho_i = F_i - f(Y_i), \quad i = 1, 2, \ldots, s. \tag{4.35}$$

Denote by $\widehat{f}$ the mapping from $\mathbb{R}^{sN}$ to $\mathbb{R}^{sN}$ so that the N dimensional subvectors in $\widehat{f}(Y)$ are equal to $f(Y_i), i = 1, 2 \ldots, s$. We can then write (4.35) in the compact form

$$\varphi - (I_s \otimes J)\rho = F - \widehat{f}(Y). \tag{4.36}$$

Combining (4.35) with (4.34) we obtain the linear system

$$\begin{bmatrix} I_s \otimes I_N & -h(A \otimes I_N) \\ -I_s \otimes J & I_s \otimes I_N \end{bmatrix} \begin{bmatrix} \rho \\ \varphi \end{bmatrix} = \begin{bmatrix} Y - 1 \otimes y_n - h(A \otimes I_N)F \\ F - \widehat{f}(Y) \end{bmatrix}. \tag{4.37}$$

Eliminate φ to obtain

$$M\rho = G, \tag{4.38}$$

where

$$G = Y - 1 \otimes y_n - h(A \otimes I_N)\widehat{f}(Y).$$

In assessing the computational cost of this scheme, we note that not only is M computed only occasionally, but the LU factorization of M is computed at the same time. In general it is not possible to estimate for how many steps the same value of M should be used before this phase of the computation has to be repeated. It is also not possible to predict how many iterations need to be performed to achieve convergence but a reliable predictor should be used to minimize the overall iteration cost.

Lacking this knowledge about how the iteration scheme will perform, in the general case, leaves us with two questions:
(a) what is the cost of the occasional Jacobian evaluation and the LU factorization of M? and
(b) what is the cost of a single iteration in a single step?

For (a) there is the cost of the actual Jacobian evaluation plus a constant times s^3N^3 for the factorization.

For (b) there is the cost of evaluating f times s plus a constant times s^2N^2 for the back substitutions.
We can lower the cost of (a) and (b) by using the transformation.

4.4.1 Using transformations for cost reduction

It was shown in [21] that how transformations can be used to lower implementation costs for large problems. Although this was originally in the context of stiff problems, it is equally relevant for Hamiltonian problems where a Newton scheme is required to achieve accurate results most efficiently.

Suppose that A has only real distinct eigenvalues and let T be a transformation matrix such that

$$T^{-1}AT = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_s).$$

Define a transformed version of the update vector,

$$\overline{\rho} = (T^{-1} \otimes I_N)\rho,$$

and rewrite (4.38) in transformed form

$$\overline{M}\overline{\rho} = \overline{G}.$$

where

$$\begin{aligned}
\overline{G} &= (T^{-1} \otimes I_N)G, \\
\overline{M} &= (T^{-1} \otimes I_N)M(T \otimes I_N) \\
&= \mathrm{diag}(I_N - h\lambda_1 J, I_N - h\lambda_2 J, \ldots, I_N - h\lambda_s J).
\end{aligned}$$

The algorithm for carrying out the Newton iterations can now be modified taking the transformations into account.

In phase (a) of the computation, carried out as preparation for what is often a large number of steps, instead of calculating the LU factors of the sN dimensional matrix M, s separate LU factorizations of the N dimensional matrices $I - h\lambda J$, $i = 1, 2, \ldots, s$. This results in a considerable saving by a factor of approximately s^2.

In phase (b), the actual Newton updates in each iteration of each step, the solution of (4.38) is replaced by

1. calculation of $\overline{G}$ from G,
2. solution of the linear systems
$$(I - h\lambda J)\overline{\rho}_i = \overline{G}_i, \quad i = 1, 2, \ldots, s,$$
3. calculation of ρ from $\overline{\rho}$.

Although there are additional transformation costs introduced, equal to s^2N, in each of steps 1 and 3, this is, for large N, insignificant compared with the cost of 2 which is now sN^2 compared with s^2N^2 in the untransformed version of the algorithm.

4.4.2 Application to new method

For the symplectic effective order four method (4.24), the eigenvalues are given by

$$\lambda_1 = \frac{3}{2},$$
$$\lambda_2 = \frac{-6 + 5\sqrt{6}}{12},$$
$$\lambda_3 = \frac{-6 - 5\sqrt{6}}{12}$$

As we can notice, all three eigenvalues are real. Now the transformation matrix for the method (4.24) is given by

$$T = \begin{bmatrix} -\frac{101}{9} & \frac{39+35\sqrt{6}}{87} & \frac{39-35\sqrt{6}}{87} \\ -19 & \frac{1227-70\sqrt{6}}{551} & \frac{1227+70\sqrt{6}}{551} \\ 1 & 1 & 1 \end{bmatrix}.$$

This transformation matrix enables us to implement the method very effectively for large Hamiltonian problems.

Chapter 5

Order conditions for G-symplectic general linear methods

General linear methods for the solution of ordinary differential equations are both multivalue and multistage. The order conditions for general linear methods can be stated and analysed using a **B**-series approach and we will make use of this formulation. However, imposing the G-symplectic structure can modify the nature of the order conditions considerably.

In [69] it was discovered that for Runge–Kutta methods, rooted trees belonging to the same tree have equivalent order conditions and if the trees are superfluous they are automatically satisfied and can be ignored. For G-symplectic methods there is a similar close relationship between order conditions related to trees which are identical except for the placement of the root. If the method is G-symplectic the order conditions are interrelated and can be reduced to a smaller set.

This chapter is organised as follows:

In Section 5.1 a brief introduction to G-symplectic general linear methods is given. This is followed, Section 5.2 by a discussion of trees and rooted trees. Because of the order conditions for the non-principal components, there will be a relationship between the coefficients in the non-principal rows of B and the non-principal starting values. This relationship is examined in Section 5.3 which is followed by the core results for this chapter in Section 5.4. In Section 5.5 the special case of order 4 method with $V = \text{diag}(1, -1)$ will be considered in details. Although there are only two order conditions, there are additional constraints on a possible starting method, represented by $\boldsymbol{\xi}$.

The results derived in this chapter are presented in [28].

5.1 G-symplectic general linear methods

General linear methods are multistage and multivalue and are designed to give numerical solutions to the initial value problem

$$y'(x) = f(y(x)), \quad y(x_0) = y_0, \qquad f : \mathbb{R}^N \to \mathbb{R}^N, \quad y_0 \in \mathbb{R}^N. \tag{5.1}$$

At the start of step number n, r vectors $y_i^{[n-1]}, i = 1, 2, \ldots, r$ are used as input and at the end of the step the output is $y_i^{[n]}$, $i = 1, 2, \ldots, r$. Like a Runge–Kutta method, there are s stage values $Y_s, i = 1, 2, \ldots, s$, to be computed as well as the corresponding stage derivatives $F_i = f(Y_l)$, $i = 1, 2, \ldots, s$. These are interrelated by the equations

$$\begin{aligned} Y_i &= h \sum_{j=1}^{s} a_{ij} F_j + \sum_{j=1}^{r} u_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, s \\ y_i^{[n]} &= h \sum_{j=1}^{s} b_{ij} F_j + \sum_{j=1}^{r} v_{ij} y_j^{[n-1]}, \quad i = 1, 2, \ldots, r, \end{aligned} \tag{5.2}$$

where the coefficients $a_{ij}, u_{ij}, b_{ij}, v_{ij}$ comprise the $(s+r) \times (s+r)$ partitioned matrix

$$\left[\begin{array}{cc} A & U \\ B & V \end{array} \right].$$

If we write

$$y^{[n-1]} = \left[\begin{array}{c} y_1^{[n-1]} \\ y_2^{[n-1]} \\ \vdots \\ y_r^{[n-1]} \end{array} \right], \quad y^{[n]} = \left[\begin{array}{c} y_1^{[n]} \\ y_2^{[n]} \\ \vdots \\ y_r^{[n]} \end{array} \right], \quad Y = \left[\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{array} \right], \quad F = \left[\begin{array}{c} F_1 \\ F_2 \\ \vdots \\ F_s \end{array} \right],$$

then (5.2) can be written in a compact form

$$\begin{aligned} Y &= h(A \otimes I_N)F + (U \otimes I_N)y^{[n-1]}, \\ y^{[n]} &= h(B \otimes I_N)F + (V \otimes I_N)y^{[n-1]}. \end{aligned}$$

General linear methods for $r > 1$ cannot preserve quadratic invariants was investigated in [25]. For $r = 1$ general linear methods reduce to classical Runge–Kutta methods and the methods preserve quadratic invariants if the symplectic conditions holds. We seek a similar criteria for general linear methods. We will introduce the G-norm property which is similar to that of nonlinear stability of general linear methods [5]. For this we assume a symmetric matrix $G = g_{ij}$ and define the inner product $\langle , \rangle$

$$\langle y, z \rangle_G = \sum_{i,j=1}^{r} g_{ij} \langle y_i, z_j \rangle,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_r \end{bmatrix}, \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_r \end{bmatrix}.$$

Then the G-norm defined by the inner product $\langle , \rangle_G$ will be $\|y\|_G^2 = \langle y, y \rangle_G$.
We also introduce a diagonal matrix D with diagonal elements d_i and write as

$$\langle y, z \rangle_D = \sum_{i=1}^{r} d_i \langle y_i, z_j \rangle.$$

According to the *B-series* criterion for order presented in [19], a general linear method (A, U, B, V) has order p if there exists $\xi \in G^r$ and $\eta \in G_1^s$ such that

$$\eta = A\eta\mathbf{D} + U\boldsymbol{\xi}, \tag{5.3}$$

$$\mathbf{E}\boldsymbol{\xi} = B\eta\mathbf{D} + V\boldsymbol{\xi}, \tag{5.4}$$

where (5.4) has to hold only for trees of order p or less.

Theorem 5.1.1. *[18] If a method is G-symplectic, then there exists a symmetric matrix G and a diagonal matrix D such that*

$$G = V^* G V, \tag{5.5}$$

$$DU = B^* G V, \tag{5.6}$$

$$DA + A^\intercal D = B^* G B. \tag{5.7}$$

From (5.5) it follows that the eigenvalues of V all lie on the unit circle and we will assume that they are distinct. By recursively evaluating the order conditions (5.3) and (5.4) tree by tree it is possible to obtain the order conditions up to any required order. However, just as for Runge–Kutta methods, the canonical conditions have an influence on the results. Also the results need to be understood relative to the starting method. In other words, the value of $\boldsymbol{\xi}(t)$ needs to be taken into account for $|t| \leq p$. It is found that a generalization of the results in [69] apply in the multi-value case and this will be one of the principal conclusions of the chapter.

Example 5.1.2.

We will consider the G-symplectic general linear method M, which was introduced in [19].

$$M = \left[\begin{array}{cc|cc} \frac{3+\sqrt{3}}{6} & 0 & 1 & -\frac{3+2\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} & \frac{3+\sqrt{3}}{6} & 1 & \frac{3+2\sqrt{3}}{3} \\ \hline \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & -1 \end{array}\right]. \tag{5.8}$$

The method M satisfies the G-symplectic conditions with

$$G = \begin{bmatrix} 1 & 0 \\ 0 & \frac{3+2\sqrt{3}}{3} \end{bmatrix}, \qquad D = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

The starting method is

$$\left[\begin{array}{cc|c} \frac{3+\sqrt{3}}{6} & 0 & 1 \\ 0 & -\frac{3+\sqrt{3}}{6} & 1 \\ \hline 0 & 0 & 1 \\ \frac{\sqrt{3}-1}{8} & -\frac{\sqrt{3}-1}{8} & 0 \end{array}\right].$$

For the approximate conservation of Hamiltonian the general linear method M is applied to simple pendulum which is based on the Hamiltonian

$$H = \frac{p^2}{2} - \cos(q).$$

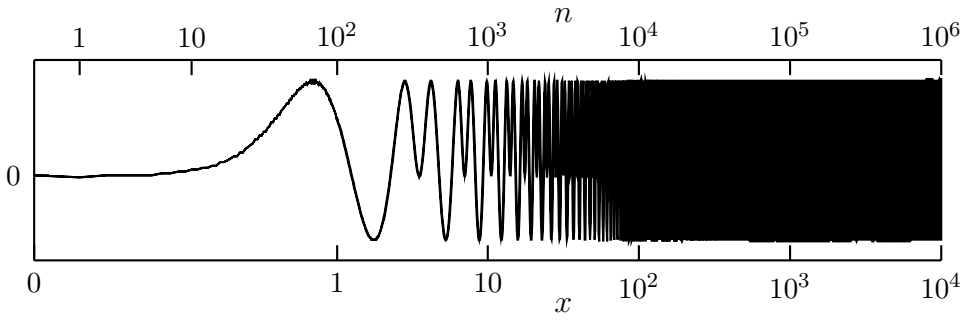


Figure 5.1: The variation in the numerical Hamiltonian for the simple pendulum problem with initial value $y = [0, 1.2]^{\intercal}$, using method M and $h = 0.01$

The initial values are chosen as $p = 0, q = 1.2$ with $h = 0.01$ for the 1000000 steps. Figure 5.1 shows that the Hamiltonian is approximately conserved, this is because the method is G-symplectic. Thus the G-symplecticity is an important property for the general linear methods.

5.2 Trees and rooted trees

The significance of *trees*, as distinct from *rooted trees*, for canonical Runge–Kutta was analyzed by Sanz-Serna and Abia [69].
If t_1, t_2 are arbitrary rooted trees with positive order we can construct an unsymmetrical product $t_1 t_2$ by attaching the roots and designating the root of t_1 as the root of $t_1 t_2$. This will mean that, although $t_1 t_2$ and $t_2 t_1$ are not in general the same rooted tree, they are equivalent in the sense that they have the same underlying tree even though the roots may be different. Following the definitions in [67] we state

Definition 5.2.1. *Two rooted trees are equivalent if they have the same underlying tree.*

We will regard the underlying tree as an equivalence class for the corresponding rooted trees.

Definition 5.2.2. *A tree is superfluous if it contains a rooted tree of the form tt. A rooted tree is superfluous if it belongs to a superfluous tree. A tree (respectively rooted tree) is non-superfluous if it is not superfluous.*

A list of trees and the corresponding rooted trees, showing in particular which of these are superfluous is given in Figure 5.2.

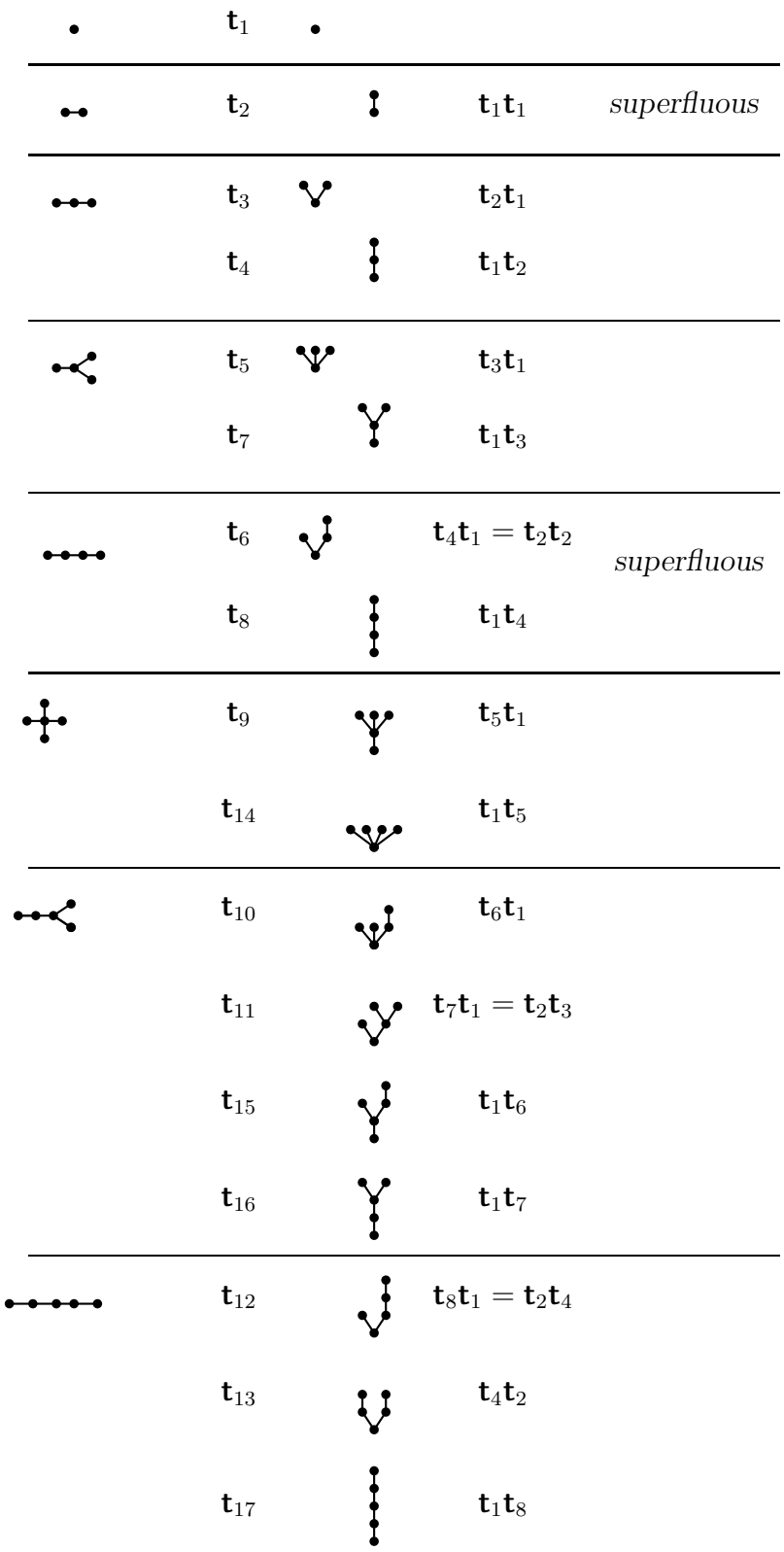


Figure 5.2: Trees and corresponding rooted trees including two superfluous trees

5.3 Structure of canonical general linear methods: General case

By assuming that the eigenvalues of V are distinct, the methods fall into a diagonal case in which V is a diagonal, though possible complex, matrix. The analysis of order is much more convenient and simpler under this assumption but, there is no loss of generality because a simpler transformation makes all elements of the defining matrices real. We will assume that a method has the form

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{cccc|cccc} a_{11} & 0 & \cdots & 0 & 1 & u_{12} & \cdots & u_{1r} \\ a_{21} & a_{22} & \cdots & 0 & 1 & u_{22} & \cdots & u_{2r} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{s1} & a_{s2} & \cdots & a_{ss} & 1 & u_{s2} & \cdots & u_{sr} \\ \hline b_{11} & b_{12} & \cdots & b_{1s} & 1 & 0 & \cdots & 0 \\ b_{21} & b_{22} & \cdots & b_{2s} & 0 & z_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rs} & 0 & 0 & \cdots & z_r \end{array}\right],$$

where V is a diagonal matrix such that

$$V = \mathrm{diag}(z_1, z_2, \ldots, z_r),$$

with $z_1 = 1$.
If r is odd then the remaining z values occur in conjugate pairs and if r is even, one of the z values is -1 and the remainder occur in conjugate pairs.
It follows from (5.5) that $G = \mathrm{diag}(g_1, g_2, \ldots, g_r)$, with $g_i = g_j$ if $z_i = \overline{z}_j$. Without loss of generality, assume that $g_1 = 1$. By consistency combined with (5.6), it follows that

$$Ue_1 = \mathbf{1}, \qquad D = \mathrm{diag}(b), \text{ where } e_1^{\mathsf{T}} B = b^{\mathsf{T}}.$$

This enables us to write

$$B = \begin{bmatrix} b^{\mathsf{T}} \\ \\ \widehat{B} \end{bmatrix}, \qquad U = \begin{bmatrix} \mathbf{1} & \widehat{U} \end{bmatrix}, \qquad D\widehat{U} = \widehat{B}^{*}\widehat{G}\widehat{V}$$

We will denote the *principal component* of $\boldsymbol{\xi}$ by ζ and the remainder of this vector by $\widehat{\xi}$. Hence (5.4) takes the form

$$\mathsf{E}\zeta = b^{\mathsf{T}}\eta\mathbf{D} + \zeta, \tag{5.9}$$

$$\mathsf{E}\widehat{\xi} = \widehat{B}\eta\mathbf{D} + \widehat{V}\widehat{\xi}. \tag{5.10}$$

The conditions associated with (5.10) do not impose conditions on the coefficients of the method and should be thought of as conditions on $\widehat{\xi}$. We can assume that they are always satisfied. Hence, the difficulty of constructing a method of specific order is to choose the coefficients of the method along with ζ so that (5.9) is satisfied.

The procedure we will adopt to derive the order conditions is to construct the vectors η and $\eta\mathbf{D}$ recursively using the unsymmetrical product to express formulae for high order rooted trees in terms of lower order. From $\eta\mathbf{D}(t)$, we will then construct the order condition related to each t. We will also use the canonical conditions to interrelate conditions for t_1t_2 and t_2t_1 as is done in [67].

5.3.1 Parasitism-free methods

Parasitic solutions are the unwanted solutions which occur with the numerical solution of the particular problems. In [24] it was demonstrated that the effect of parasitism can be disastrous to long term integrations of practical problems and that when growth rates are forced to equal zero, this unacceptable behaviour is entirely eliminated. In the case of methods in which $V = \text{diag}(1, z_2, z_3, \ldots, z_r)$, as in this chapter, this means that the diagonal elements of BU are, apart from the $(1,1)$ element, equal to zero. That is

$$\beta^{\mathsf{T}}u = 0.$$

Accordingly this requirement will be built into the design of the method derived in this chapter.

5.3.2 The value of $\widehat{\xi}$

As noted in Section 5.3, the order conditions (5.9), (5.10) apply in the diagonal case. In particular (5.10) are conditions on the non-principal starting method coefficients $\widehat{\xi}$. Consider a particular row of $\widehat{B}$ denoted by β^{T} with the corresponding diagonal element of $\widehat{V}$ denoted by z and the corresponding component of $\widehat{\xi}$ denoted by ξ. Then these quantities satisfy the equation

$$\mathsf{E}\xi = \beta^{\mathsf{T}}(\eta\mathsf{D}) + z\xi. \tag{5.11}$$

Because $\xi(\mathbf{t}_0) = 0$, the operation E is linear acting on a vector made up from $\xi(\mathbf{t}_1)$, $\xi(\mathbf{t}_2)$, ..., with as many terms as there are rooted trees to consider. In the case of order four, this vector and the corresponding vector of $\beta^{\mathsf{T}}(\eta\mathsf{D})$ values for these rooted trees are, shown below, together with the matrix $\widetilde{E}$ representing the linear operator that has been referred to.

$$\xi = \begin{bmatrix} \xi(\mathbf{t}_1) \\ \xi(\mathbf{t}_2) \\ \xi(\mathbf{t}_3) \\ \xi(\mathbf{t}_4) \\ \xi(\mathbf{t}_5) \\ \xi(\mathbf{t}_6) \\ \xi(\mathbf{t}_7) \\ \xi(\mathbf{t}_8) \end{bmatrix}, \qquad \beta^{\mathsf{T}}(\eta\mathsf{D}) = \begin{bmatrix} \beta^{\mathsf{T}}\mathbf{1} \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_1) \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_1)^2 \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_2) \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_1)^3 \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_1)\eta(\mathbf{t}_2) \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_3) \\ \beta^{\mathsf{T}}\eta(\mathbf{t}_4) \end{bmatrix}, \qquad \widetilde{E} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 3 & 3 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{3}{2} & 1 & 1 & 0 & 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 2 & 0 & 0 & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

At least in the case of order four, (5.11) can be rewritten in the form

$$(\widetilde{E} - zI)\xi = \beta^{\mathsf{T}}(\eta\mathsf{D}). \tag{5.12}$$

In the important case $r = 2$, $z = -1$, which is known to be sufficient for the construction of methods up to order four, (5.11) can be written $\xi = (I+\widetilde{E})^{-1}\beta^{\mathsf{T}}(\eta\mathsf{D})$

and, for convenience, we write

$$(I+\widetilde{E})^{-1}=\begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{8} & 0 & -\frac{3}{4} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{16} & 0 & -\frac{1}{4} & -\frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{24} & 0 & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{48} & 0 & 0 & -\frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

5.3.3 The B-series for η and $\eta\mathsf{D}$

The generating function η for the internal stages and their stages derivatives $\eta\mathsf{D}$ plays an important role in the construction of a particular general linear method. For compactness, we will write $(\eta\mathsf{D})(\mathbf{t}_i)=(\eta\mathsf{D})_i$ and $\eta(\mathbf{t}_i)=\eta_i$, $i=1,2,\ldots,8$.

First evaluate η_1 from $(\eta\mathsf{D})_1=\mathbf{1}$:

$$\eta_1=A(\eta\mathsf{D})_1+U\boldsymbol{\xi}_1=A\mathbf{1}+U\boldsymbol{\xi}_1.$$

This will be denoted by c. Values of $(\eta\mathsf{D})_i$ are listed as follows up to order 4:

$$\begin{aligned}
(\eta\mathsf{D})_1&=\mathbf{1},\\
(\eta\mathsf{D})_2&=c,\\
(\eta\mathsf{D})_3&=c^2,\\
(\eta\mathsf{D})_4&=Ac+U\boldsymbol{\xi}_2,\\
(\eta\mathsf{D})_5&=c^3,\\
(\eta\mathsf{D})_6&=cAc+cU\boldsymbol{\xi}_2,\\
(\eta\mathsf{D})_7&=Ac^2+U\boldsymbol{\xi}_3,\\
(\eta\mathsf{D})_8&=A^2c+AU\boldsymbol{\xi}_2+U\boldsymbol{\xi}_4.
\end{aligned}$$

Now the corresponding η_i are

$$\begin{aligned}
\eta_2 &= A(\eta\mathbf{D})_2 + U\boldsymbol{\xi}_2 = Ac + U\boldsymbol{\xi}_2,\\
\eta_3 &= A(\eta\mathbf{D})_3 + U\boldsymbol{\xi}_3 = Ac^2 + U\boldsymbol{\xi}_3,\\
\eta_4 &= A(\eta\mathbf{D})_4 + U\boldsymbol{\xi}_4 = A^2c + AU\boldsymbol{\xi}_2 + U\boldsymbol{\xi}_4,\\
\eta_5 &= A(\eta\mathbf{D})_5 + U\boldsymbol{\xi}_5 = Ac^3 + U\boldsymbol{\xi}_5,\\
\eta_6 &= A(\eta\mathbf{D})_6 + U\boldsymbol{\xi}_6 = cA^2c + AcU\boldsymbol{\xi}_2 + U\boldsymbol{\xi}_6,\\
\eta_7 &= A(\eta\mathbf{D})_7 + U\boldsymbol{\xi}_7 = A^2c^2 + AU\boldsymbol{\xi}_3 + U\boldsymbol{\xi}_7,\\
\eta_8 &= A(\eta\mathbf{D})_8 + U\boldsymbol{\xi}_8 = A^3c + A^2U\boldsymbol{\xi}_2 + AU\boldsymbol{\xi}_4 + U\boldsymbol{\xi}_8.
\end{aligned}$$

5.4 Order conditions for G-symplectic methods

The algorithm for describing order questions for any G-symplectic general linear method will come in three parts:

- The order conditions for the output vector b^{T} can be written in terms of the input vector $\zeta's$ for all trees up to order $\leq p$, that is

$$b^{\mathsf{T}}(\eta\mathbf{D})(t) = (\mathbf{E}\zeta)(t) - \zeta(t), \quad |t| \leq p$$

- Conditions on ζ to ensure the starting method allows order p to be achieved
- Relations between pairs of equivalent trees t_1t_2 and t_2t_1 as a consequences of the canonical conditions:

$$b^{\mathsf{T}}(\eta D)(t_1t_2) + b^{\mathsf{T}}(\eta D)(t_2t_1) = (B(\eta\mathbf{D})(t_1))^* GB(\eta\mathbf{D})(t_2)$$

or, equivalently,

$$\begin{aligned}
(\mathbf{E}\zeta)(t_1t_2) + (\mathbf{E}\zeta)(t_2t_1) - (\mathbf{E}\boldsymbol{\xi})(t_1)^* G(\mathbf{E}\boldsymbol{\xi})(t_2) = \\
\zeta(t_1t_2) + \zeta(t_2t_1) - \boldsymbol{\xi}(t_1)^* G\boldsymbol{\xi}(t_2)
\end{aligned}$$

Theorem 5.4.1. *If for rooted trees t_1 and t_2,*

$$\mathbf{E}\zeta(t_1t_2) + \mathbf{E}\zeta(t_2t_1) - \big(\mathbf{E}\boldsymbol{\xi}(t_1)\big)^* G\mathbf{E}\boldsymbol{\xi}(t_2) = \zeta(t_1t_2) + \zeta(t_2t_1) - \big(\boldsymbol{\xi}(t_1)\big)^* G\boldsymbol{\xi}(t_2) \quad (5.13)$$

then

$$b^{\mathsf{T}}(\eta\mathbf{D})(t_1t_2) - (\mathbf{E}\zeta(t_1t_2) - \zeta(t_1t_2)) + b^{\mathsf{T}}(\eta\mathbf{D})(t_2t_1) - (\mathbf{E}\zeta(t_2t_1) - \zeta(t_2t_1)) = 0.$$

Proof. We first find a convenient expression for rewriting $b^{\mathsf{T}}\eta\mathbf{D}(t_1t_2)$:

$$\begin{aligned}
b^{\mathsf{T}}(\eta\mathbf{D})(t_1t_2) &= b^{\mathsf{T}}(\eta\mathbf{D})(t_1)\eta(t_2)\\
&= b^{\mathsf{T}}(\eta\mathbf{D})(t_1)\big(A(\eta\mathbf{D})(t_2)+U\boldsymbol{\xi}(t_2)\big)\\
&= \big((\eta\mathbf{D})(t_1)\big)^*\big(DA(\eta\mathbf{D})(t_2)\big)+\big((\eta\mathbf{D})(t_1)\big)^*DU\boldsymbol{\xi}(t_2)\\
&= \big((\eta\mathbf{D})(t_1)\big)^*\big(DA(\eta\mathbf{D})(t_2)\big)+\big((\eta\mathbf{D})(t_1)\big)^*B^*GV\boldsymbol{\xi}(t_2)\\
&= \big((\eta\mathbf{D})(t_1)\big)^*\big(DA(\eta\mathbf{D})(t_2)\big)+\big(B(\eta\mathbf{D})(t_1)\big)^*GV\boldsymbol{\xi}(t_2)\\
&= \big((\eta\mathbf{D})(t_1)\big)^*\big(DA(\eta\mathbf{D})(t_2)\big)+\big(\mathbf{E}\boldsymbol{\xi}(t_1)-V\boldsymbol{\xi}(t_1)\big)^*GV\boldsymbol{\xi}(t_2)
\end{aligned}$$

We can write a similar expression with t_1 and t_2 interchanged, that is

$$b^{\mathsf{T}}(\eta\mathbf{D})(t_2t_1)=\big((\eta\mathbf{D})(t_2)\big)^*DA(\eta\mathbf{D})(t_1)+\big(\mathbf{E}\boldsymbol{\xi}(t_2)-V\boldsymbol{\xi}(t_2)\big)^*GV\boldsymbol{\xi}(t_1)$$

By adding expressions $b^{\mathsf{T}}(\eta\mathbf{D})(t_1t_2)$ and $b^{\mathsf{T}}(\eta\mathbf{D})(t_2t_1)$ and using (5.7), we get

$$\begin{aligned}
b^{\mathsf{T}}(\eta\mathbf{D})(t_1t_2)+b^{\mathsf{T}}(\eta\mathbf{D})(t_2t_1) &= \big(B\boldsymbol{\xi}(t_1)\big)^*GB\boldsymbol{\xi}(t_2)+\\
&\big(\mathbf{E}\boldsymbol{\xi}(t_1)-V\boldsymbol{\xi}(t_1)\big)^*GV\boldsymbol{\xi}(t_2))+V\boldsymbol{\xi}(t_1)^*G\big(\mathbf{E}\boldsymbol{\xi}(t_2)-V\boldsymbol{\xi}(t_2)\big).
\end{aligned}$$

Substitute $B\boldsymbol{\xi}(t_i)=\mathbf{E}\boldsymbol{\xi}(t_i)-V\boldsymbol{\xi}(t_i)$, $i=1,2$ and expand to obtain

$$b^{\mathsf{T}}(\eta\mathbf{D})(t_1t_2)+b^{\mathsf{T}}(\eta\mathbf{D})(t_2t_1)=\mathbf{E}\boldsymbol{\xi}(t_1)^*G\mathbf{E}\boldsymbol{\xi}(t_2)-\boldsymbol{\xi}(t_1)^*G\boldsymbol{\xi}(t_2).$$

Use (5.13) and the result follows. □

Remark 5.4.2. *Theorem 5.4.1 is a natural generalization of Theorem 3.1 in [69].*

In [28] the following theorem is proved.

Theorem 5.4.3.

$$\mathbf{E}\zeta(t_1t_2)+\mathbf{E}\zeta(t_2t_1)-\big(\mathbf{E}\boldsymbol{\xi}(t_1)\big)^*G\mathbf{E}\boldsymbol{\xi}(t_2)=\zeta(t_1t_2)+\zeta(t_2t_1)-\big(\boldsymbol{\xi}(t_1)\big)^*G\boldsymbol{\xi}(t_2)$$

holds whenever $|t_1t_2|\le p$ if and only if

$$\zeta(t_1t_2)+\zeta(t_2t_1)=\boldsymbol{\xi}(t_1)^*G\boldsymbol{\xi}(t_2),\qquad |t_1t_2|<p.$$

Theorem 5.4.4. *Consider a G-symplectic general linear method with order at least one. Then the order conditions for a family of equivalent rooted trees of order p are all satisfied if*

1 one of them is satisfied;

2 all lower order conditions are satisfied; and

3 for any pair of rooted trees t_1 and t_2 such that $|t_1| + |t_2| < p$ it holds that

$$\zeta(t_1 t_2) + \zeta(t_2 t_1) = \boldsymbol{\xi}(t_1)^* G \boldsymbol{\xi}(t_2).$$

Furthermore, the order conditions for a family of superfluous rooted trees are all satisfied if 2 and 3 hold.

We can find all fourth order methods for $r = 2$ by using the following result:

Corollary 5.4.5 (Order 4 Theorem)**.** *A canonical general linear method with $V = \mathrm{diag}(1, -1)$, $G = \mathrm{diag}(1, g)$ has order four iff the coefficients of the method satisfy the conditions*

$$\begin{aligned} b^{\mathsf{T}}\mathbf{1} &= 1, \\ b^{\mathsf{T}}c^2 &= \tfrac{1}{3} + g\xi_1^2 \end{aligned}$$

and the starting method (assuming $\zeta_1 = 0$) satisfy

$$\begin{aligned} \zeta_2 &= \tfrac{1}{2} g \xi_1^2, \\ \zeta_3 &= \tfrac{1}{3}(b^{\mathsf{T}}c^3 - \tfrac{1}{4}) - \zeta_2, \\ \zeta_4 &= g\xi_1\xi_2 - \zeta_3. \end{aligned}$$

5.4.1 Conditions for orders one and two

The process of deriving order conditions consists of recursively evaluating $\eta(t)$ from (5.3), and $\widehat{\xi}(t)$ from (5.10). At the same time (5.9) will impose an order condition on the coefficients of the method, related also to $\zeta(t')$ for various t' such that $|t'| < |t|$.

To start this process, write $\boldsymbol{\xi}(\mathbf{t}_0) = e_1$ from consistency and $\eta D(\mathbf{t}_0) = 0$. This gives $\eta(\mathbf{t}_0) = \mathbf{1}$ from (5.3) and the value of $\widehat{\xi}(\mathbf{t}_0) = 0$. Because $\eta D(\mathbf{t}_1) = \eta(\mathbf{t}_0) = \mathbf{1}$, we can find $\eta(\mathbf{t}_1) = A\mathbf{1} + U\boldsymbol{\xi}(\mathbf{t}_1)$. Expand $U\boldsymbol{\xi}(\mathbf{t}_1) = \zeta(\mathbf{t}_1)\mathbf{1} + \widehat{U}\widehat{\xi}(\mathbf{t}_1)$. The value of $\xi(\mathbf{t}_1)$ can be found from (5.11) but a value is needed for $\zeta(\mathbf{t}_1)$. It turns out that this quantity is arbitrary and, for, simplicity, we will set it to zero. We now have a value for $\eta(\mathbf{t}_1)$ and we will denote this by c, as in a common notation used for Runge–Kutta methods. Now use the fact that $\eta D(\mathbf{t}_2) = \eta(\mathbf{t}_1)$ to obtain the order two condition

$$b^{\mathsf{T}}c = \tfrac{1}{2}. \tag{5.14}$$

We will use the canonical conditions to deduce that (5.14) automatically holds. First we note that, because $\boldsymbol{\xi}(\mathbf{t}_0) = e_1$, it follows that

$$\mathsf{E}\boldsymbol{\xi}(\mathbf{t}_1) - \boldsymbol{\xi}(\mathbf{t}_1) = e_1. \tag{5.15}$$

Now, using (5.5), (5.6) and (5.7) as necessary we calculate

$$\begin{aligned}
2b^{\mathsf{T}}c &= \mathbf{1}^{\mathsf{T}}D(A\mathbf{1} + U\boldsymbol{\xi}(\mathbf{t}_1)) + (A\mathbf{1} + U\boldsymbol{\xi}(\mathbf{t}_1))^*D\mathbf{1} \\
&= \mathbf{1}^{\mathsf{T}}(DA + A^{\mathsf{T}}D)\mathbf{1} + \mathbf{1}^{\mathsf{T}}DU\boldsymbol{\xi}(\mathbf{t}_1) + \boldsymbol{\xi}(\mathbf{t}_1)^*U^*D\mathbf{1} \\
&= \mathbf{1}^{\mathsf{T}}B^*GB\mathbf{1} + \mathbf{1}^{\mathsf{T}}B^*GV\boldsymbol{\xi}(\mathbf{t}_1) + \boldsymbol{\xi}(\mathbf{t}_1)^*V^*GB\mathbf{1} \\
&= (E\boldsymbol{\xi}(\mathbf{t}_1) - V\boldsymbol{\xi}(t_1))^*G(E\boldsymbol{\xi}(\mathbf{t}_1) - V\boldsymbol{\xi}(t_1)) + \\
&\qquad (E\boldsymbol{\xi}(\mathbf{t}_1) - V\boldsymbol{\xi}(t_1))^*GV\boldsymbol{\xi}(\mathbf{t}_1) + \boldsymbol{\xi}(\mathbf{t}_1)^*V^*G(E\boldsymbol{\xi}(\mathbf{t}_1) - V\boldsymbol{\xi}(t_1)).
\end{aligned}$$

After some simplification, and making used of (5.15), this is found to equal

$$(\mathsf{E}\boldsymbol{\xi}(\mathbf{t}_1) - \boldsymbol{\xi}(\mathbf{t}_1))^*G(\mathsf{E}\boldsymbol{\xi}(\mathbf{t}_1) - \boldsymbol{\xi}(\mathbf{t}_1)) = e_1^{\mathsf{T}}Ge_1 = 1.$$

5.5 Construction of a three stage, two input value, fourth order method

We construct a three stage G-symplectic general linear method with two input values. This method is assumed to be parasitism-free. For cheap implementation cost, we choose our matrix A to be lower triangular. The structure of the method would be

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} a_{11} & 0 & 0 & u_{01} & u_1 \\ a_{21} & a_{22} & 0 & u_{02} & u_2 \\ a_{31} & a_{32} & a_{33} & u_{03} & u_3 \\ \hline b_1 & b_2 & b_3 & v_{11} & v_{12} \\ \beta_1 & \beta_2 & \beta_3 & v_{21} & v_{22} \end{array}\right],$$

with

$$G = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}, \qquad D = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}.$$

We recall η for the internal stages and $\boldsymbol{\xi}$ for the output vectors such that $\boldsymbol{\xi} = \begin{bmatrix} \zeta \\ \xi \end{bmatrix}$ with $\zeta(\phi) = 1$ and $\xi(\phi) = 0$. To obtain the basic consistency condition, we solve the (5.3) for the empty tree, keeping the fact that $\eta \mathbf{D}(\phi) = 0$, that is

$$\mathbf{E}\zeta(\phi) = v_{11}\zeta(\phi) + v_{12}\xi(\phi), \tag{5.16}$$
$$\mathbf{E}\xi(\phi) = v_{21}\zeta(\phi) + v_{22}\xi(\phi). \tag{5.17}$$

Thus by comparing the above equations we get $v_{11} = 1$ and $v_{21} = 0$, therefore the V matrix takes the form

$$V = \begin{bmatrix} 1 & v_{12} \\ 0 & v_{22} \end{bmatrix}.$$

We want that the V has the diagonal form such that the eigenvalues of V lies on the unit circle. Thus V takes the form

$$V = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Now by solving the (5.4) for the empty tree, we get

$$\eta_1(\phi) = U\boldsymbol{\xi}(\phi),$$
$$\eta_2(\phi) = U\boldsymbol{\xi}(\phi),$$
$$\eta_3(\phi) = U\boldsymbol{\xi}(\phi),$$

this leads us to

$$u_{01} = 1,$$
$$u_{02} = 1,$$
$$u_{03} = 1.$$

Thus the GLM takes the form

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} a_{11} & 0 & 0 & 1 & u_1 \\ a_{21} & a_{22} & 0 & 1 & u_2 \\ a_{31} & a_{32} & a_{33} & 1 & u_3 \\ \hline b_1 & b_2 & b_3 & 1 & 0 \\ \beta_1 & \beta_2 & \beta_3 & 0 & -1 \end{array}\right],$$

From the condition $G = V^T G V$, we can write

$$\begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$\begin{aligned} g_{11} &= g_{11}, \\ g_{12} &= -g_{12}, \\ g_{21} &= -g_{21}, \\ g_{22} &= g_{22}. \end{aligned}$$

By comparing both sides of the equations we get

$$G = \begin{bmatrix} 1 & 0 \\ 0 & g \end{bmatrix},$$

we consider the case when

$$G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Now consider the G-symplectic condition,

$$B^T G V = DU,$$

which becomes

$$\begin{bmatrix} b_1 & \beta_1 \\ b_2 & \beta_2 \\ b_3 & \beta_3 \end{bmatrix} = \begin{bmatrix} d_1 & d_1 u_1 \\ d_2 & d_2 u_2 \\ d_3 & d_3 u_3 \end{bmatrix}.$$

By comparing both sides of the above expression, we get from the respective columns the expressions like

$$b_i = d_i,$$

similarly by comparing the second columns in respective matrices we get

$$\beta_i = d_i u_i.$$
$$u_i = \frac{\beta_i}{b_i}.$$

For simplification we call $x_i = \frac{\beta_i}{b_i}$, so above expression becomes

$$u_i = x_i$$

Thus the structure of general linear method becomes

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} a_{11} & 0 & 0 & 1 & x_1 \\ a_{21} & a_{22} & 0 & 1 & x_2 \\ a_{31} & a_{32} & a_{33} & 1 & x_3 \\ \hline b_1 & b_2 & b_3 & 1 & 0 \\ b_1 x_1 & b_2 x_2 & b_3 x_3 & 0 & -1 \end{array}\right],$$

Now look at the $DA + A^{\intercal}AD = B^{\intercal}GB$ condition. This gives properties of the coefficient matrices because the method is symplectic and are quite different in nature from the order conditions.

$$DA + A^T D = \begin{bmatrix} 2b_1a_{11} & b_2a_{21} & b_3a_{31} \\ b_2a_{21} & 2b_2a_{22} & b_3a_{32} \\ b_3a_{31} & b_3a_{32} & 2b_3a_{33} \end{bmatrix},$$

$$B^T GB = \begin{bmatrix} b_1^2(1-x_1^2) & b_1b_2(1-x_1x_2) & b_1b_3(1-x_1x_3) \\ b_1b_2(1-x_1x_2) & b_2^2(1-x_2^2) & b_2b_3(1-x_2x_3) \\ b_1b_3(1-x_1x_3) & b_2b_3(1-x_2x_3) & b_3^2(1-x_3^2) \end{bmatrix}.$$

Comparing coefficients of $DA + A^T D = B^T GB$ on both sides, we get

$$\begin{aligned}
a_{11} &= \tfrac{1}{2} b_1 (1 - x_1^2), \\
a_{21} &= b_1 (1 - x_1 x_2), \\
a_{22} &= \tfrac{1}{2} b_2 (1 - x_2^2), \\
a_{31} &= b_1 (1 - x_1 x_3), \\
a_{32} &= b_2 (1 - x_2 x_3), \\
a_{33} &= \tfrac{1}{2} b_3 (1 - x_3^2).
\end{aligned}$$

Therefore the coefficient matrix A takes the form

$$A = \begin{bmatrix}
\frac{1}{2} b_1 (1 - x_1^2) & 0 & 0 \\
b_1 (1 - x_1 x_2) & \frac{1}{2} b_2 (1 - x_2^2) & 0 \\
b_1 (1 - x_1 x_3) & b_2 (1 - x_2 x_3) & \frac{1}{2} b_3 (1 - x_3^2)
\end{bmatrix}$$

For simplicity we choose $x_1 = 1$ and $b_1 = 1$. The methods satisfies the order conditions. The first order condition becomes

$$b_2 = -b_3.$$

We are constructing parasitism-free method and therefore we have

$$\begin{aligned}
b_1 x_1^2 + b_2 x_2^2 + b_3 x_3^2 &= 0, \\
b_2 (x_3^2 - x_2^2) &= 1.
\end{aligned}$$

Appropriate choices of the parameters are $x_2 = \frac{5}{3}$ and $x_3 = \frac{1}{3}$, leading to the following general linear method.

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc}
0 & 0 & 0 & 1 & 1 \\
-\frac{2}{3} & \frac{1}{3} & 0 & 1 & \frac{5}{3} \\
\frac{2}{3} & -\frac{1}{6} & \frac{1}{6} & 1 & \frac{1}{3} \\
\hline
1 & -\frac{3}{8} & \frac{3}{8} & 1 & 0 \\
1 & -\frac{5}{8} & \frac{1}{8} & 0 & -1
\end{array}\right], \tag{5.18}$$

with

$$G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{8} & 0 \\ 0 & 0 & \frac{3}{8} \end{bmatrix}, \qquad c = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{12} \\ \frac{3}{4} \end{bmatrix}.$$

The stages and the outputs are defined by

$$\begin{aligned}
Y_1 &= \quad 0hF_1 + 0hF_2 + 0hF_3 + 1y_1^{[n-1]} + 1y_2^{[n-1]} \\
Y_2 &= -\tfrac{2}{3}hF_1 + \tfrac{1}{3}hF_2 + 0hF_3 + 1y_1^{[n-1]} + \tfrac{5}{3}y_2^{[n-1]} \\
Y_3 &= \quad \tfrac{2}{3}hF_1 - \tfrac{1}{6}hF_2 + \tfrac{1}{6}hF_3 + 1y_1^{[n-1]} + \tfrac{1}{3}y_2^{[n-1]} \\
y_1^{[n]} &= \quad 1hF_1 - \tfrac{3}{8}hF_2 + \tfrac{3}{8}hF_3 + 1y_1^{[n-1]} + 0y_2^{[n-1]} \\
y_2^{[n]} &= \quad 1hF_1 - \tfrac{5}{8}hF_2 + \tfrac{1}{8}hF_3 \quad + 0y_1^{[n-1]} - 1y_2^{[n-1]}
\end{aligned}$$

The method does not suffer from parasitism and is G-symplectic.

5.5.1 Example methods of type 4123

We have constructed three methods, each of type 4123. This notation means
- 4 is the order
- 1 is the stage -order
- 2 is the number of inputs
- 3 is the number of stages

In each of these methods, A is lower triangular, and each $c's$ value lies in the interval $[0, 1]$.

Example 5.5.1.

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 1 \\ \frac{19}{30} & -\frac{1}{12} & 0 & 1 & -\frac{5}{6} \\ \frac{133}{330} & -\frac{31}{66} & \frac{7}{12} & 1 & -\frac{1}{6} \\ \hline \frac{19}{55} & -\frac{6}{11} & \frac{6}{5} & 1 & 0 \\ \frac{19}{55} & \frac{5}{11} & -\frac{1}{5} & 0 & -1 \end{array}\right],$$

with

$$G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad D = \begin{bmatrix} \frac{19}{55} & 0 & 0 \\ 0 & -\frac{6}{11} & 0 \\ 0 & 0 & \frac{6}{5} \end{bmatrix}, \qquad c = \begin{bmatrix} \frac{3}{10} \\ \frac{3}{10} \\ \frac{7}{15} \end{bmatrix}.$$

Example 5.5.2.

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 1 \\ -\frac{7}{15} & \frac{3}{8} & 0 & 1 & \frac{6}{5} \\ 1 & -\frac{15}{28} & \frac{1}{8} & 1 & \frac{4}{7} \\ \hline \frac{7}{3} & -\frac{75}{44} & \frac{49}{132} & 1 & 0 \\ \frac{7}{3} & -\frac{45}{22} & \frac{7}{33} & 0 & -1 \end{array}\right],$$

with

$$G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad D = \begin{bmatrix} \frac{7}{3} & 0 & 0 \\ 0 & -\frac{75}{44} & 0 \\ 0 & 0 & \frac{49}{132} \end{bmatrix} \qquad c = \begin{bmatrix} \frac{1}{4} \\ \frac{5}{24} \\ \frac{41}{56} \end{bmatrix}.$$

The above method are G-symplectic methods and does not suffer from parasitic solution.

5.5.2 Algebraic analysis

The above method (5.18) is G-symplectic and of order four. We apply the algebraic analysis of order to the general linear method which we call as method M_h. To

start the procedure the method M_h requires two inputs, whereas we are given just one initial condition with the initial value problem (5.1). We obtain the rest of the initial conditions using the starting procedure say S_h. In addition to this we also introduce a finishing method F_h such that $F_h \circ S_h = id$. A method M_h is said to have order p if

$$M_h \circ S_h - S_h \circ E_h = O(h^{p+1}),$$

where E is the flow of the method.
For the algebraic analysis, we assume that $\boldsymbol{\xi} = [\zeta, \xi]$ is the generating function for the input approximations and let η be the generating function for the internal stages.

$$\eta = A\eta\mathsf{D} + \zeta\mathbf{1} + \xi u \tag{5.19}$$

$$\mathsf{E}\zeta = b^{\mathsf{T}}\eta\mathsf{D} + \zeta \tag{5.20}$$

$$\mathsf{E}\xi = \beta^{\mathsf{T}}\eta\mathsf{D} - \xi \tag{5.21}$$

From the analysis of the method (5.18), we get the values for the generating function up to four order trees such that

	ϕ								
ζ	1	0	$-\frac{1}{32}$	$\frac{5}{864}$	$-\frac{19}{1728}$	ζ_5	ζ_6	ζ_7	ζ_8
ξ	0	$\frac{1}{4}$	$\frac{1}{48}$	$-\frac{47}{576}$	$-\frac{23}{1152}$	$\frac{1}{6912}$	$-\frac{61}{4608}$	$-\frac{23}{2304}$	$-\frac{5}{1536}$

The values $[\zeta_5,\ \zeta_6,\ \zeta_7,\ \zeta_8]$ can be found from the starting method.
Now the details for the generating functions η for the internal stages, $\eta\mathsf{D}$ their stages derivatives and output values $\bar{\boldsymbol{\xi}} = [\bar{\zeta}, \bar{\xi}]$ are given below:
Case 1: For the empty tree ϕ

	η_1	$\eta_1\mathsf{D}$	η_2	$\eta_2\mathsf{D}$	η_3	$\eta_3\mathsf{D}$	$\bar{\zeta}$	$\bar{\xi}$
ϕ	1	0	1	0	1	0	1	0

.

Case 2: For the tree with one vertex $\bullet$,

$$\begin{aligned}
\eta_1\mathbf{D}(\bullet) &= \eta_1(\phi)\\
&= 1,\\
\eta_1(\bullet) &= 1\zeta(\bullet) + u_{12}\xi(\bullet)\\
&= \frac{1}{4},\\
\eta_2\mathbf{D}(\bullet) &= \eta_2(\phi)\\
&= 1,\\
\eta_2(\bullet) &= a_{21}\eta_1\mathbf{D}(\bullet) + a_{22}\eta_2\mathbf{D}(\bullet) + 1\zeta(\bullet) + u_{22}\xi(\bullet)\\
&= \frac{1}{12}
\end{aligned}$$

$$\begin{aligned}
\eta_3\mathbf{D}(\bullet) &= \eta_3(\phi)\\
&= 1,\\
\eta_3(\bullet) &= a_{31}\eta_1\mathbf{D}(\bullet) + a_{32}\eta_2\mathbf{D}(\bullet) + a_{33}\eta_3\mathbf{D}(\bullet) + 1\zeta(\bullet) + u_{32}\xi(\bullet)\\
&= \frac{3}{4}\\
\bar{\zeta}(\bullet) &= b_1\eta_1\mathbf{D}(\bullet) + b_2\eta_2\mathbf{D}(\bullet) + b_3\eta_3\mathbf{D}(\bullet) + \zeta(\bullet)\\
&= 1\\
\bar{\xi}(\bullet) &= \beta_1\eta_1\mathbf{D}(\bullet) + \beta_2\eta_2\mathbf{D}(\bullet) + \beta_3\eta_3\mathbf{D}(\bullet) - \xi(\bullet)\\
&= \frac{1}{4}.
\end{aligned}$$

Case 3: For the tree  ,

$$\begin{aligned}
\eta_1\mathbf{D}(\,\vdots\,) &= \eta_1(\bullet)\\
&= \frac{1}{4}\\
\eta_1(\,\vdots\,) &= 1\zeta(\,\vdots\,) + u_{12}\xi(\,\vdots\,)\\
&= -\frac{1}{96}
\end{aligned}$$

$$\begin{aligned}
\eta_2\mathbf{D}(\,\vdots\,) &= \eta_2(\bullet)\\
&= \frac{1}{12}\\
\eta_2(\,\vdots\,) &= a_{21}\eta_1\mathbf{D}(\,\vdots\,) + a_{22}\eta_2\mathbf{D}(\,\vdots\,) + 1\zeta(\,\vdots\,) + u_{22}\xi(\,\vdots\,)\\
&= -\frac{13}{96}
\end{aligned}$$

$$\eta_3\mathbf{D}(\ \bullet\!\!\bullet\ ) = \eta_3(\bullet)$$
$$= \frac{3}{4}$$
$$\eta_3(\ \bullet\!\!\bullet\ ) = a_{31}\eta_1\mathbf{D}(\ \bullet\!\!\bullet\ ) + a_{32}\eta_2\mathbf{D}(\ \bullet\!\!\bullet\ ) + a_{33}\eta_3\mathbf{D}(\ \bullet\!\!\bullet\ ) + 1\zeta(\ \bullet\!\!\bullet\ ) + u_{32}\xi(\ \bullet\!\!\bullet\ )$$
$$= \frac{73}{288}$$

$$\bar{\zeta}(\ \bullet\!\!\bullet\ ) = b_1\eta_1\mathbf{D}(\ \bullet\!\!\bullet\ ) + b_2\eta_2\mathbf{D}(\ \bullet\!\!\bullet\ ) + b_3\eta_3\mathbf{D}(\ \bullet\!\!\bullet\ ) + \zeta(\ \bullet\!\!\bullet\ )$$
$$= \frac{15}{32}$$
$$\bar{\xi}(\ \bullet\!\!\bullet\ ) = \beta_1\eta_1\mathbf{D}(\ \bullet\!\!\bullet\ ) + \beta_2\eta_2\mathbf{D}(\ \bullet\!\!\bullet\ ) + \beta_3\eta_3\mathbf{D}(\ \bullet\!\!\bullet\ ) - \xi(\ \bullet\!\!\bullet\ )$$
$$= \frac{13}{48}$$

Case 4: For the tree $\bullet\!\!\bullet\!\!\bullet$

$$\eta_1\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) = \eta_1^2(\bullet)$$
$$= \frac{1}{16}$$
$$\eta_1(\bullet\!\!\bullet\!\!\bullet) = 1\zeta(\bullet\!\!\bullet\!\!\bullet) + u_{12}\xi(\bullet\!\!\bullet\!\!\bullet)$$
$$= -\frac{131}{1728}$$
$$\eta_2\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) = \eta_2^2(\bullet)$$
$$= \frac{1}{144}$$
$$\eta_2(\bullet\!\!\bullet\!\!\bullet) = a_{21}\eta_1\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + a_{22}\eta_2\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + 1\zeta(\bullet\!\!\bullet\!\!\bullet) + u_{22}\xi(\bullet\!\!\bullet\!\!\bullet)$$
$$= -\frac{293}{1728}$$

$$\eta_3\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) = \eta_3^2(\bullet)$$
$$= \frac{9}{16}$$
$$\eta_3(\bullet\!\!\bullet\!\!\bullet) = a_{31}\eta_1\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + a_{32}\eta_2\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + a_{33}\eta_3\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + 1\zeta(\bullet\!\!\bullet\!\!\bullet) + u_{32}\xi(\bullet\!\!\bullet\!\!\bullet)$$
$$= \frac{65}{576}$$
$$\bar{\zeta}(\bullet\!\!\bullet\!\!\bullet) = b_1\eta_1\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + b_2\eta_2\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + b_3\eta_3\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + \zeta(\bullet\!\!\bullet\!\!\bullet)$$
$$= \frac{239}{864}$$
$$\bar{\xi}(\bullet\!\!\bullet\!\!\bullet) = \beta_1\eta_1\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + \beta_2\eta_2\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) + \beta_3\eta_3\mathbf{D}(\bullet\!\!\bullet\!\!\bullet) - \xi(\bullet\!\!\bullet\!\!\bullet)$$
$$= \frac{121}{576}$$

Case5: For the tree

$$\eta_1\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) = \eta_1(\,\cdot\!\cdot\,)$$
$$= -\frac{1}{96}$$
$$\eta_1(\,\cdot\!\cdot\!\cdot\,) = 1\zeta(\,\cdot\!\cdot\!\cdot\,) + u_{12}\xi_2(\,\cdot\!\cdot\!\cdot\,)$$
$$= -\frac{107}{3456}$$

$$\eta_2\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) = \eta_2(\,\cdot\!\cdot\,)$$
$$= -\frac{13}{96}$$
$$\eta_2(\,\cdot\!\cdot\!\cdot\,) = a_{21}\eta_1\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + a_{22}\eta_2\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + 1\zeta(\,\cdot\!\cdot\!\cdot\,) + u_{22}\xi(\,\cdot\!\cdot\!\cdot\,)$$
$$= -\frac{95}{1152}$$
$$\eta_3\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) = \eta_3(\,\cdot\!\cdot\,)$$
$$= \frac{73}{288}$$
$$\eta_3(\,\cdot\!\cdot\!\cdot\,) = a_{31}\eta_1\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + a_{32}\eta_2\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + a_{33}\eta_3\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + 1\zeta(\,\cdot\!\cdot\!\cdot\,) + u_{32}\xi(\,\cdot\!\cdot\!\cdot\,)$$
$$= \frac{139}{3456}$$
$$\bar{\zeta}(\,\cdot\!\cdot\!\cdot\,) = b_1\eta_1\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + b_2\eta_2\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + b_3\eta_3\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + \zeta(\,\cdot\!\cdot\!\cdot\,)$$
$$= \frac{215}{1728}$$
$$\bar{\xi}(\,\cdot\!\cdot\!\cdot\,) = \beta_1\eta_1\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + \beta_2\eta_2\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) + \beta_3\eta_3\mathbf{D}(\,\cdot\!\cdot\!\cdot\,) - \xi(\,\cdot\!\cdot\!\cdot\,)$$
$$= \frac{145}{1152}$$

The general linear method (5.18) will be of order four if

$$\mathbf{E}\zeta(t) = \bar{\zeta}(t),$$
$$\mathbf{E}\xi(t) = \bar{\xi}(t)$$

where,

	ϕ	$\bullet$	$\cdot\!\cdot$	$\cdot\!\cdot\!\cdot$	$\cdot\!\cdot\!\cdot\!\cdot$
$\mathbf{E}\zeta$	1	1	$\frac{15}{32}$	$\frac{239}{864}$	$\frac{215}{1728}$
$\bar{\zeta}$	1	1	$\frac{15}{32}$	$\frac{239}{864}$	$\frac{215}{1728}$
$\mathbf{E}\xi$	0	$\frac{1}{4}$	$\frac{13}{48}$	$\frac{121}{576}$	$\frac{145}{1152}$
$\bar{\xi}$	0	$\frac{1}{4}$	$\frac{13}{48}$	$\frac{121}{576}$	$\frac{145}{1152}$

This proves the order four behaviour.

5.5.3 Starting and finishing methods

In order to construct an explicit four stage two outputs method which satisfies

$$\Phi(t_i) = \xi_i, \qquad i = 1, 2, \ldots, 8,$$

$$\widetilde{\phi}(t_i) = \zeta_i, \qquad i = 1, 2, 3, 4.$$

In the case of the first input we get a starting method $y_0 + h\sum_{i=1}^{4} \widetilde{b}_i f(Y_i)$ and for the second input $h\sum_{i=1}^{8} b_i f(Y_i)$. We can find a method to satisfy

$$\begin{aligned}
\sum_i \widetilde{b}_i &= \zeta_1, \\
\sum_i \widetilde{b}_i c_i &= \zeta_2, \\
\sum_i \widetilde{b}_i c_i^2 &= \zeta_3, \\
\sum_{ij} \widetilde{b}_i a_{ij} c_j &= \zeta_4,
\end{aligned} \tag{5.22}$$

and

$$\begin{aligned}
\sum_i b_i &= \xi_1, \\
\sum_i b_i c_i &= \xi_2, \\
\sum_i b_i c_i^2 &= \xi_3, \\
\sum_{ij} b_i a_{ij} c_j &= \xi_4, \\
\sum_i b_i c_i^3 &= \xi_5, \\
\sum_{ij} b_i c_i a_{ij} c_j &= \xi_6, \\
\sum_{ij} b_i a_{ij} c_j^2 &= \xi_7, \\
b_4 a_{43} a_{32} c_2 &= \xi_8
\end{aligned}$$

The structure of the starting method is such that

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ a_{21} & 0 & 0 & 0 & 1 \\ a_{31} & a_{32} & 0 & 0 & 1 \\ a_{41} & a_{42} & a_{43} & 0 & 1 \\ \hline \widetilde{b}_1 & \widetilde{b}_2 & \widetilde{b}_3 & \widetilde{b}_4 & 1 \\ b_1 & b_2 & b_3 & b_4 & 0 \end{array}\right].$$

We can solve for the b_i by linear equations and then the a_{ij} by linear equations as long as the ξ_8 product comes out right. This depends on choosing suitable c_2 and c_4.

We choose $c = [0 \quad \frac{1}{3} \quad \frac{2}{3} \quad -\frac{2119}{3532}]$. The starting method will have one input and two outputs, where the b values can be found by solving the following equations simultaneously

$$\begin{aligned} \sum_i b_i &= \xi_1, \\ \sum_i b_i c_i &= \xi_2, \\ \sum_i b_i c_i^2 &= \xi_3, \\ \sum_i b_i c_i^3 &= \xi_5. \end{aligned}$$

Therefore

$$\left[\begin{array}{c} b_1 \\ b_2 \\ b_3 \\ b_4 \end{array}\right] = \left[\begin{array}{c} \frac{14813}{33904} \\ \frac{2093}{79112} \\ -\frac{19665}{214736} \\ -\frac{137004612013}{1124937000044} \end{array}\right],$$

and the structure of the starting method becomes

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{1}{69} & \frac{15}{23} & 0 & 0 & 1 \\ -\frac{8389502714541}{17536590337664} & -\frac{6431381589347}{26304885506496} & \frac{6468387750253}{52609771012992} & 0 & 1 \\ \hline \widetilde{b}_1 & \widetilde{b}_2 & \widetilde{b}_3 & \widetilde{b}_4 & 1 \\ \frac{14813}{33904} & \frac{2093}{79112} & -\frac{19665}{214736} & -\frac{137004612013}{1124937000044} & 0 \end{array}\right].$$

The values for $[\widetilde{b}_1, \widetilde{b}_2, \widetilde{b}_3, \widetilde{b}_4]$ can be found by solving the set of equations (5.22) with the same A and c values. That is

$$\begin{bmatrix} \widetilde{b}_1 \\ \widetilde{b}_2 \\ \widetilde{b}_3 \\ \widetilde{b}_4 \end{bmatrix} = \begin{bmatrix} -\frac{76717}{813696} \\ \frac{16555}{172608} \\ -\frac{261211}{5153664} \\ \frac{15050336767}{306801000012} \end{bmatrix}.$$

Thus we get the starting method as

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{1}{69} & \frac{15}{23} & 0 & 0 & 1 \\ -\frac{8389502714541}{17536590337664} & -\frac{6431381589347}{26304885506496} & \frac{6468387750253}{52609771012992} & 0 & 1 \\ \hline -\frac{76717}{813696} & \frac{16555}{172608} & -\frac{261211}{5153664} & \frac{15050336767}{306801000012} & 1 \\ \frac{14813}{33904} & \frac{2093}{79112} & -\frac{19665}{214736} & -\frac{137004612013}{1124937000044} & 0 \end{array}\right]. \tag{5.23}$$

We can get the values of remaining $\zeta's$ from the starting method such that

$$\zeta_5 = -\frac{134629}{6103296},$$
$$\zeta_6 = -\frac{89833}{12206592},$$
$$\zeta_7 = -\frac{14189}{6103296},$$
$$\zeta_8 = \frac{5335}{4068864}.$$

Next we need to find a suitable finishing method F_h, that is applied to first output $y_1^{[n]}$ of the starting method. The finishing method F_h, is fourth order Runge–Kutta method based on the following structure

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \check{a}_{21} & 0 & 0 & 0 & 1 \\ \check{a}_{31} & \check{a}_{32} & 0 & 0 & 1 \\ \check{a}_{41} & \check{a}_{42} & \check{a}_{43} & 0 & 1 \\ \hline \check{b}_1 & \check{b}_2 & \check{b}_3 & \check{b}_4 & 1 \end{array}\right].$$

where the finishing method F_h is defined by $F_h(t_i) = \zeta^{-1}(t_i)$ such that $(F_h\zeta)(t_i) = 0$ for all trees up to order four. That is

$$F_h(\mathbf{t}_1) = -\zeta(\mathbf{t}_1),$$
$$F_h(\mathbf{t}_2) = -\zeta(\mathbf{t}_2),$$
$$F_h(\mathbf{t}_3) = -\zeta(\mathbf{t}_3),$$
$$F_h(\mathbf{t}_4) = -\zeta(\mathbf{t}_4),$$
$$F_h(\mathbf{t}_5) = -\zeta(\mathbf{t}_5),$$
$$F_h(\mathbf{t}_6) = \zeta^2(\mathbf{t}_2) - \zeta(\mathbf{t}_6),$$
$$F_h(\mathbf{t}_7) = -\zeta(\mathbf{t}_7),$$
$$F_h(\mathbf{t}_8) = \zeta^2(\mathbf{t}_2) - \zeta(\mathbf{t}_8).$$

We can find method F_h to satisfy

$$\begin{aligned}
\sum \check{b}_i - F_h(\mathbf{t}_1) &= 0,\\
\sum \check{b}_i\check{c}_i - F_h(\mathbf{t}_2) &= 0,\\
\sum \check{b}_i\check{c}_i^2 - F_h(\mathbf{t}_3) &= 0,\\
\sum \check{b}_i\check{a}_{ij}\check{c}_j - F_h(\mathbf{t}_4) &= 0,\\
\sum \check{b}_i\check{c}_i^3 - F_h(\mathbf{t}_5) &= 0,\\
\sum \check{b}_i\check{c}_i\check{a}_{ij}\check{c}_j - F_h(\mathbf{t}_6) &= 0,\\
\sum \check{b}_i\check{a}_{ij}\check{c}_j^2 - F_h(\mathbf{t}_7) &= 0,\\
\check{b}_4\check{a}_{43}\check{a}_{32}\check{c}_2 - F_h(\mathbf{t}_8) &= 0.
\end{aligned}$$

We choose $\check{c} = [0 \quad \frac{1}{3} \quad \frac{2}{3} \quad \frac{175797913}{1126404248}]$, where the $\check{b}$ values can be found by solving the following equations simultaneously

$$\begin{aligned}
\sum \check{b}_i &= F_h(\mathbf{t}_1),\\
\sum \check{b}_i\check{c}_i &= F_h(\mathbf{t}_2),\\
\sum \check{b}_i\check{c}_i^2 &= F_h(\mathbf{t}_3),\\
\sum \check{b}_i\check{c}_i^3 &= F_h(\mathbf{t}_5).
\end{aligned}$$

Thus

$$\begin{bmatrix} \check{b}_1 \\ \check{b}_2 \\ \check{b}_3 \\ \check{b}_4 \end{bmatrix} = \begin{bmatrix} -\frac{39483516397}{33753199296} \\ -\frac{21949176959}{14376252216} \\ \frac{72556835}{310477632} \\ \frac{1258158182698422366360446}{5108560018534306631907 21} \end{bmatrix}.$$

By solving the following equations we can get the values for $\check{a}_{32}, \check{a}_{42}, \check{a}_{43}$. That is

$$\begin{aligned}
\sum \check{b}_i\check{a}_{ij}\check{c}_j &= F_h(\mathbf{t}_4),\\
\sum \check{b}_i\check{c}_i\check{a}_{ij}\check{c}_j &= F_h(\mathbf{t}_6),\\
\sum \check{b}_i\check{a}_{ij}\check{c}_j^2 &= F_h(\mathbf{t}_7).
\end{aligned}$$

Thus the finishing method F_h takes the form

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{73657}{147246} & \frac{8169}{49082} & 0 & 0 & 1 \\ \check{a}_{41} & \check{a}_{42} & \check{a}_{43} & 0 & 1 \\ \hline -\frac{39483516397}{33753199296} & -\frac{21949176959}{14376252216} & \frac{72556835}{310477632} & \frac{125815818269842236636044 6}{5108560018534306631907 21} & 1 \end{array}\right], \qquad (5.24)$$

with

$$\begin{bmatrix} \check{a}_{41} \\ \check{a}_{42} \\ \check{a}_{43} \end{bmatrix} = \begin{bmatrix} \frac{475837122063856242814508098059}{304984139016832209717393151 6928} \\ \frac{227474306762243546589639403 5751}{910377654965244146006418557 803008} \\ -\frac{417897238049501396845449 4687}{170642484529567787442627658 4448} \end{bmatrix}.$$

5.5.4 Order four verification

To verify that the method (5.18) has order four, we will use the Order 4 Theorem.

We have

$$\xi_1 = \tfrac{1}{4}, \quad c = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{12} \\ \frac{3}{4} \end{bmatrix}, \quad \xi_2 = \tfrac{1}{48}.$$

Now we check the necessary condition $b^\intercal c^2 = \frac{1}{3} + g\xi_1^2$. That is

$$b^\intercal c^2 - \tfrac{1}{3} - g\xi_1^2 = \tfrac{13}{48} - \tfrac{1}{3} + \tfrac{1}{16} = 0.$$

Also from the Order 4 Theorem we have

$$\begin{aligned} \zeta_2 - \tfrac{1}{2}g\xi_1^2 = -\tfrac{1}{32} + \tfrac{1}{32} \quad &= 0, \\ \zeta_3 - \tfrac{1}{3}(b^\intercal c^3 - \tfrac{1}{4}) - \zeta_2 = \tfrac{5}{864} - \tfrac{5}{864} \quad &= 0, \\ \zeta_4 - g\xi_1\xi_2 - \zeta_3 = -\tfrac{19}{1728} + \tfrac{19}{1728} &= 0. \end{aligned}$$

Thus the method (5.18) has order four behaviour together with the starting S_h (5.23) and finishing F_h (5.24) methods.

Chapter 6

Extension to six order G-symplectic general linear methods

G-symplectic general linear methods are especially suitable for the long term integration of Hamiltonian problems where it is necessary to closely conserve energy, angular momentum and symplectic behaviour. The expression *canonical* will be used interchangeably with *G-symplectic* when referring to general linear methods. In Chapter 2 and Chapter 5 we have discussed this in detail.

This chapter deals with 6th order method which is G-symplectic and symmetric. In Section 6.1 we will discuss the possible layout for the matrices V and G for the method. This is followed by the Section 6.2, which defines the classes of trees and rooted trees, stage order and time reversal symmetry. In Section 6.3 we will present the design requirements for new method and in Section 6.4, the construction will be presented.

The results derived in this chapter are presented in [29].

6.1 The V and G structures

It is known that if the method is G-symplectic, there exists a symmetric matrix G and a diagonal matrix D such that

$$G = V^* G V, \tag{6.1}$$

$$DU = B^* G V, \tag{6.2}$$

$$DA + A^\intercal D = B^* G B. \tag{6.3}$$

We will consider the case $r = 4$, therefore

$$V = \mathrm{diag}(1, \zeta, -\widehat{\zeta}, -1). \qquad |\zeta| = 1. \quad \mathrm{Im}\ \zeta \neq 0. \tag{6.4}$$

We will use the value $\zeta = i$ in the construction of a practical method. It follows from (6.1) that G is diagonal and, without loss of generality, because we can rescale the method coefficients if necessary, we can assume

$$G = \mathrm{diag}(1, \pm\tfrac{1}{2}, \pm\tfrac{1}{2}, \pm 1). \tag{6.5}$$

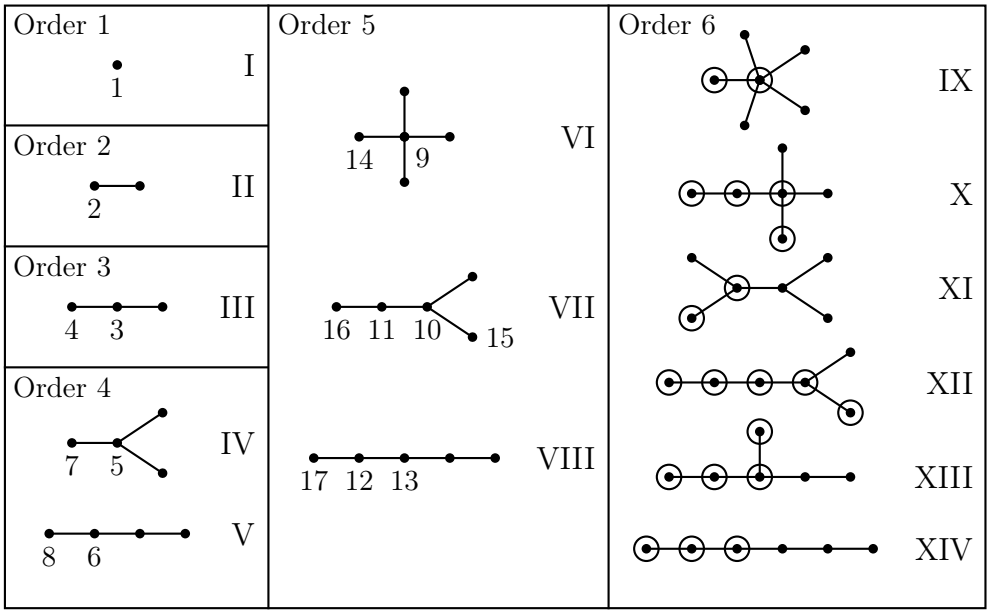
Note that the factors $\frac{1}{2}$ are inserted because this will later simplify notation when we convert to a real formulation of the method. For the method derived in this chapter, particular choices for the $\pm$ signs are made as $G = \mathrm{diag}(1, -\frac{1}{2}, -\frac{1}{2}, 1)$ but it is possible that, for some methods, with a similar design, $G = \mathrm{diag}(1, -\frac{1}{2}, -\frac{1}{2}, -1)$.

It will be convenient to denote the first row of B by $b^\intercal$ and it will then follow from (6.4), (6.5) and the first column of (6.2) that $D = \mathrm{diag}(b)$. The abscissae vector for the method is equal to $\eta(\mathbf{t}_1)$, which for the methods considered in this chapter, is equal to $c = A\mathbf{1}$. The usual quadrature conditions for a Runge–Kutta method are $b^\intercal \mathbf{1} = 1$, $b^\intercal c^{k-1} = \frac{1}{k}$, $k = 2, 3, \ldots, p$ and, for the methods considered here, these conditions will also hold but with the general linear methods interpretation.

6.2 Trees and rooted trees classes

There are 37 rooted trees of order up to six but these break into equivalence classes based on the underlying *unrooted* trees. We will summarize this graph-theoretic starting point to this discussion by showing in Table 6.1 a list of all trees up to order six with an index number attached to some of the vertices. The indexed

Table 6.1: Trees to order six



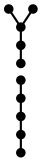
rooted trees, those of order not exceeding five, will be referred to as $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_{17}$. However, the trees of order six will not be allocated a serial number here. Possible sites for the placement of the root for the order six trees are shown with large circles. This will provide for 20 distinct rooted trees of order six. The non-rooted trees will be numbered using Roman numbering I, II, III, IV, ...

To clarify the entries in Table 6.1, Table 6.2 is also presented. This show the *rooted* trees to order five with a reference to the (non-rooted) trees to which they are related.

In the case of canonical methods, the order conditions have been analyzed in [28]. We will use these conditions but simplified by using enhanced stage order.

The main aim, of deriving a method which is G-symplectic, was shown in [28] to simplify the order conditions, and hence simplify the construction of the method. If appropriate assumptions are made concerning the starting method then, as in [69] for Runge–Kutta methods, all rooted trees with the same underlying tree structure become equivalent and satisfying the order condition for one of these rooted trees applies to all members of its equivalent class, assuming that lower order conditions are satisfied. As for Runge–Kutta methods the superfluous trees can be disregarded because each of the trees in such a class are automatically satisfied. In Table 6.1 this applies to trees II, V, XI and XIV.

Table 6.2: Rooted trees to order five

order	trees	number
1		I
2		II
3		III
3		III
4		IV
4		V
4		VI
4		V
5		VI
5		VII
5		VII
5		VIII
5		VIII
5		VI
5		VII
5		VII
5		VIII

6.2.1 Enhanced stage order

To reduce the number of order conditions that remain to be satisfied it will be an advantage to require that a condition similar to the C(2) condition for Runge–Kutta methods is satisfied. For the derivation of the method in this chapter, this will be coupled with further requirements on the B matrix (6.8), (6.9). In this case, the C(2) condition is given by (6.6) and (6.7).

$$A\mathbf{1} = c, \tag{6.6}$$

$$Ac = \tfrac{1}{2}c^2, \tag{6.7}$$

$$B\mathbf{1} = e_1, \tag{6.8}$$

$$Bc = \tfrac{1}{2}e_1. \tag{6.9}$$

To appreciate the effect of these assumptions, we see that, under appropriate choices of the starting method, all rooted trees which contain a single terminal vertex attached to any non-terminal vertex can be eliminated from consideration. Extending this to the non-rooted trees in Table 6.1, we see that trees numbers V, VII, VIII, X, XII, XIII, XIV can be discarded. The same applies to tree III because the order conditions associated with rooted tree numbers 3 and 4 are related in two ways

$$(b^{\mathsf{T}}c^2 - \tfrac{1}{3}) + (b^{\mathsf{T}}Ac - \tfrac{1}{6}) = 0, \tag{6.10}$$

$$(b^{\mathsf{T}}Ac - \tfrac{1}{6}) = \tfrac{1}{2}(b^{\mathsf{T}}c^2 - \tfrac{1}{3}). \tag{6.11}$$

Equation (6.10) is a consequence of the G-symplectic condition and (6.11) follows from the C(2) condition. These two equations together imply that each of the two order conditions related to tree III are satisfied.

6.2.2 Time reversal symmetry

Time-reversal symmetry is as an important attribute of numerical schemes for the long-term integration of mechanical problems. Furthermore the symmetric general linear methods performs well over long time intervals. We can define a general linear method to be symmetric in a similar fashion to a Runge–Kutta method. A general linear method is symmetric if it is equal to its adjoint general linear method, where the adjoint general linear method takes the stepsize with opposite sign. But the symmetry in general linear method is not as simple as for Runge–Kutta method because the output approximations contain the matrix V, which is multiplied by the

input approximations and it could be possible that the inverse matrix V^{-1} is not equal to V. For this reason we introduce an involution matrix L and permutation matrix P such that $L^2 = P$ and $LV^{-1}L = V$.

In particular, because of time reversal symmetry trees with even order can be ignored because the corresponding conditions will be automatically satisfied. Thus referring to Table 6.1, we see that trees IV, IX and XI can be discarded.

6.3 Design requirements for sixth order methods

After abandoning some unpromising lines of investigation it seemed to be appropriate to impose some assumptions about the structure of a possible sixth order method.

Structure of the method

We will assume the method has the following form

$$\left[\begin{array}{cc} A & U \\ B & V \end{array}\right] = \left[\begin{array}{ccccc|cccc} 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{2}i\overline{w}_1 & -\frac{1}{2}iw_1 & -\gamma_1 \\ a_{21} & a_{22} & 0 & 0 & 0 & 1 & \frac{1}{2}i\overline{w}_2 & -\frac{1}{2}iw_2 & -\gamma_2 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 1 & \frac{1}{2}i\overline{w}_3 & -\frac{1}{2}iw_3 & -\gamma_3 \\ a_{41} & a_{42} & a_{43} & a_{44} & 0 & 1 & \frac{1}{2}i\overline{w}_4 & -\frac{1}{2}iw_4 & -\gamma_4 \\ a_{51} & a_{52} & a_{53} & a_{54} & 0 & 1 & \frac{1}{2}i\overline{w}_5 & -\frac{1}{2}iw_5 & -\gamma_5 \\ \hline b_1 & b_2 & b_3 & b_2 & b_1 & 1 & 0 & 0 & 0 \\ b_1w_1 & b_2w_2 & b_3w_3 & b_2w_4 & b_1w_5 & 0 & i & 0 & 0 \\ b_1\overline{w}_1 & b_2\overline{w}_2 & b_3\overline{w}_3 & b_2\overline{w}_4 & b_1\overline{w}_5 & 0 & 0 & -i & 0 \\ b_1\gamma_1 & b_2\gamma_2 & b_3\gamma_3 & b_2\gamma_4 & b_1\gamma_5 & 0 & 0 & 0 & -1 \end{array}\right], \qquad (6.12)$$

where $w_i = \alpha_i + i\beta_i, i = 1, 2, \ldots, 5$.
Symmetry conditions are

$$\begin{aligned}
\alpha_3 &= 0 \\
\alpha_4 &= -\alpha_2, \\
\alpha_5 &= -\alpha_1, \\
\beta_4 &= \beta_2, \\
\beta_5 &= \beta_1, \\
\gamma_4 &= \gamma_2, \\
\gamma_5 &= \gamma_1.
\end{aligned}$$

The values in A are found from

$$a_{ij} = \begin{cases} b_j(1 - \alpha_i\alpha_j - \beta_i\beta_j + \gamma_i\gamma_j), & j < i, \\ \frac{1}{2}b_j(1 - \alpha_i\alpha_j - \beta_i\beta_j + \gamma_i\gamma_j), & j = i, \\ 0, & j > i. \end{cases}$$

in accordance with (6.3).

6.4 Derivation of the method

Choose the abscissae vector as $c = [0, c_2, \frac{1}{2}, 1-c_2, 1]$ and the vector $b = [b_1, b_2, b_3, b_2, b_1]$ such that the following order conditions are satisfied

$$\begin{aligned}
b^\intercal \mathbf{1} &= 1, \\
b^\intercal c^2 &= \tfrac{1}{3}, \\
b^\intercal c^4 &= \tfrac{1}{5}.
\end{aligned}$$

The choice of c_2 must yield a negative coefficient amongst b_1, b_2, b_3 to ensure that the parasitism growth factors can be eliminated.

Define the 5×5 symmetric matrix W with elements

$$w_{ij} = \alpha_i\alpha_j + \beta_i\beta_j - \gamma_i\gamma_j, \quad i, j = 1, 2, \ldots, 5. \tag{6.13}$$

The method defined by the matrix (6.12) will have the required properties if the

following conditions on W are satisfied:

$$w_{11} = 1, \tag{6.14}$$

$$b_1(1 - w_{12}) = \tfrac{1}{2}c_2, \tag{6.15}$$

$$b_2(1 - w_{22}) = \tfrac{1}{2}c_2, \tag{6.16}$$

$$w_{ij} = w_{6-j,6-i}, \tag{6.17}$$

$$(i, j) = (4, 3), (4, 4), (5, 3), (5, 4), (5, 5),$$

$$2b_1 + b_2(2 - c_2) + b_3 w_{33} = 0, \tag{6.18}$$

$$\sum_{j=1}^{i-1} b_j(1 - w_{ij}) + \tfrac{1}{2}b_i(1 - w_{ii}) = c_i, \quad i = 3, 4, 5 \tag{6.19}$$

$$\sum_{j=2}^{i-1} b_j c_j(1 - w_{ij}) + \tfrac{1}{2}b_i c_i(1 - w_{ii}) = \tfrac{1}{2}c_i^2, \quad i = 3, 4, 5 \tag{6.20}$$

Equations (6.14), (6.15) and (6.16), together with (6.19) and (6.20), imply that the C(2) condition is satisfied. Equation (6.17) can be written as

$$\sum_{i=1}^{5} b_i(\alpha_i^2 + \beta_i^2) - \sum_{i=1}^{5} b_i \gamma_i^2 = 0, \tag{6.21}$$

which is implied by the parasitism-free requirement. The condition that the two sums in (6.21) are separately equal to zero will be imposed when W is decomposed into components arising from the α, β and γ terms. But the immediate aim is to evaluate W itself. First evaluate w_{11}, w_{21}, w_{22}, w_{33}, w_{32}, w_{31} from (6.14), (6.15), (6.16), (6.18), (6.20), (6.19) respectively. The values of w_{ij}, $i = 4, 5$, $j = 3, 4, 5$, are then found from (6.17) and finally the first two elements of the last two rows of the symmetric matrix W are found from (6.20), (6.19).

We will now show that the rank of W cannot exceed 3. This follows because $b^\intercal W = b^\intercal \operatorname{diag}(c) W = 0$, which can be verified by detailed calculations. A consequence of

this is

$$\sum_{i=1}^{5} b_i\alpha_i = 0$$
$$\sum_{i=1}^{5} b_i\beta_i = 0$$
$$\sum_{i=1}^{5} b_i\gamma_i = 0$$
$$\sum_{i=1}^{5} b_i c_i\alpha_i = 0$$
$$\sum_{i=1}^{5} b_i c_i\beta_i = 0$$
$$\sum_{i=1}^{5} b_i c_i\gamma_i = 0,$$

which in turn implies (6.8) and (6.9).

A special case

We will choose $c = [0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1]^\intercal$ leading to $b^\intercal = [\frac{11}{120}, \frac{27}{40}, -\frac{8}{15}, \frac{27}{40}, \frac{11}{120}]$ and

$$W = \begin{bmatrix} 1 & -\frac{9}{11} & -\frac{14}{11} & -\frac{83}{297} & -\frac{39}{121} \\ -\frac{9}{11} & \frac{41}{81} & \frac{22}{27} & \frac{209}{729} & -\frac{83}{297} \\ -\frac{14}{11} & \frac{22}{27} & \frac{13}{8} & \frac{22}{27} & -\frac{14}{11} \\ -\frac{83}{297} & \frac{209}{729} & \frac{22}{27} & \frac{41}{81} & -\frac{9}{11} \\ -\frac{39}{121} & -\frac{83}{297} & -\frac{14}{11} & -\frac{9}{11} & 1 \end{bmatrix},$$

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\ \frac{5}{24} & \frac{1}{8} & \frac{1}{6} & 0 & 0 \\ \frac{19}{162} & \frac{13}{27} & -\frac{8}{81} & \frac{1}{6} & 0 \\ \frac{4}{33} & \frac{19}{22} & -\frac{40}{33} & \frac{27}{22} & 0 \end{bmatrix}$$

To recover the vectors α, β, γ from

$$W = \alpha\alpha^{\intercal} + \beta\beta^{\intercal} - \gamma\gamma^{\intercal},$$

form the two symmetric matrices $\widehat{W} = \widehat{T}^{\intercal}W\widehat{T}$, $\widetilde{W} = \widetilde{T}^{\intercal}W\widetilde{T}$, where

$$\widehat{T} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \\ 0 & 0 \\ 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \qquad \widetilde{T} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

It is found that

$$\widehat{W} = \widehat{\alpha}\widehat{\alpha}^{\intercal}, \quad \widetilde{W} = \widetilde{\beta}\widetilde{\beta}^{\intercal} - \widetilde{\gamma}\widetilde{\gamma}^{\intercal},$$

$$\widehat{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \widetilde{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \widetilde{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix},$$

leading to

$$\widehat{W} = \begin{bmatrix} \frac{80}{729} & -\frac{80}{297} \\ -\frac{80}{297} & \frac{80}{121} \end{bmatrix}$$

so that

$$\alpha = \begin{bmatrix} \frac{4\sqrt{5}}{27} \\ -\frac{4\sqrt{5}}{11} \\ 0 \\ \frac{4\sqrt{5}}{11} \\ -\frac{4\sqrt{5}}{27} \end{bmatrix} \quad \text{and} \quad \widetilde{W} = \begin{bmatrix} \frac{41}{121} & -\frac{163}{297} & -\frac{14}{11} \\ -\frac{163}{297} & \frac{289}{729} & \frac{22}{27} \\ -\frac{14}{11} & \frac{22}{27} & \frac{13}{8} \end{bmatrix}.$$

Choose $\widetilde{\gamma}$ by the conditions that $\widetilde{W}+\widetilde{\gamma}\widetilde{\gamma}^{\intercal}$ has rank 1 and that $b_1\widetilde{\gamma}_1^2+b_2\widetilde{\gamma}_2^2+\frac{1}{2}b_3\widetilde{\gamma}_3^2=0$. This gives

$$\beta=\left(\tfrac{65274\sqrt{330}-347009}{1265902}\right)^{-1/2}\begin{bmatrix}\frac{65274\sqrt{330}-347009}{1265902}\\ -\frac{70518\sqrt{330}+318613}{3107214}\\ -\frac{18285\sqrt{330}+162856}{460328}\\ -\frac{70518\sqrt{330}+318613}{3107214}\\ \frac{65274\sqrt{330}-347009}{1265902}\end{bmatrix}, \tag{6.22}$$

$$\gamma=\left(\tfrac{5934\sqrt{330}-70541}{115082}\right)^{-1/2}\begin{bmatrix}\frac{5934\sqrt{330}-70541}{115082}\\ -\frac{23506\sqrt{330}-462231}{1035738}\\ -\frac{18285\sqrt{330}-423016}{460328}\\ -\frac{23506\sqrt{330}-462231}{1035738}\\ \frac{5934\sqrt{330}-70541}{115082}\end{bmatrix}. \tag{6.23}$$

Chapter 7

Numerical experiments

Comparing numerical methods is not an easy task. This is because of the different designs that are used when implementing these numerical methods. This chapter presents numerical simulations for the numerical methods constructed in Chapter 3, Chapter 4 and Chapter 5. The experiments are performed on Hamiltonian problems and problems with quadratic invariants. The main aim is to study the numerical behaviour of these methods.

We have used constant step size for implementation purposes and the Newton modified iteration scheme is used for the implicit stage evaluations. The experiments presented here are of preliminary nature and intend to illustrate the behaviour of various methods.

This chapter is designed as follows:

In Section 7.1 we have presented different numerical methods constructed in Chapters 3, 4 and 5. Section 7.2 is about the numerical problems which are used for the testing of these numerical methods. Section 7.3 deals with the numerical results and discussions for Chapter 3. Section 7.4 deals with the numerical results and discussions for Chapter 4 and finally in Section 7.5, numerical results and comments are given for Chapter 5.

7.1 Numerical methods

7.1.1 Symplectic effective order five: RKEG5

The main method is

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{5} & \frac{3}{5} & 0 & 0 & 0 & 0 \\ \frac{2}{5} & \frac{4}{15} & \frac{2}{15} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{2}{27} & -\frac{10}{27} & \frac{10}{9} & 0 & 0 \\ \frac{4}{5} & \frac{28}{75} & \frac{4}{15} & -\frac{1}{5} & \frac{9}{25} & 0 \\ \hline & \frac{19}{96} & 0 & \frac{25}{96} & \frac{9}{32} & \frac{25}{96} \end{array},$$

starting method is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{7}{10} & \frac{7}{10} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{81}{154} & -\frac{2}{77} & 0 & 0 \\ \frac{3}{10} & \frac{129}{350} & \frac{3}{91} & -\frac{33}{325} & 0 \\ \hline & -\frac{16}{189} & \frac{5}{56} & -\frac{11}{36} & \frac{65}{216} \end{array},$$

the finishing method is

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{7}{10} & \frac{7}{10} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{81}{154} & -\frac{2}{77} & 0 & 0 \\ \frac{3}{10} & \frac{129}{350} & \frac{3}{91} & -\frac{33}{325} & 0 \\ \hline & \frac{16}{189} & -\frac{5}{56} & \frac{11}{36} & -\frac{65}{216} \end{array}.$$

7.1.2 Symplectic effective order four method: Symeff43

The main method which is symplectic and has effective order four is given by the following tableau

$$\begin{array}{c|ccc} 0 & \frac{27}{140} & \frac{76}{105} & -\frac{11}{12} \\ 0 & \frac{171}{70} & -\frac{19}{280} & -\frac{19}{8} \\ \frac{2}{3} & \frac{6}{7} & -\frac{95}{168} & \frac{3}{8} \\ \hline & \frac{27}{70} & -\frac{19}{140} & \frac{3}{4} \end{array}.$$

The starting and finishing methods are

$$\begin{array}{c|ccc} \frac{\pm 1+\sqrt{5}}{12} & \frac{\pm 1+\sqrt{5}}{12} & & \\ \frac{\sqrt{5}}{6} & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{6} & \\ \frac{\mp 1+\sqrt{5}}{12} & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{3} & \frac{\pm 1-\sqrt{5}}{12} \\ \hline & \frac{\pm 1+\sqrt{5}}{6} & \mp\frac{1}{3} & \frac{\pm 1-\sqrt{5}}{6} \end{array},$$

where the starting method is obtained by using $\pm = +$ and $\mp = -$ and the finishing method is taken as by using $\pm = -$ and $\mp = +$ in the above tableau.

7.1.3 G-symplectic general linear method: G4123

The G-symplectic general linear method with three stages and two outputs is given by

$$\left[\begin{array}{c|c} A & U \\ \hline B & V \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 1 & 1 \\ \frac{-2}{3} & \frac{1}{3} & 0 & 1 & \frac{5}{3} \\ \frac{2}{3} & \frac{-1}{6} & \frac{1}{6} & 1 & \frac{1}{3} \\ \hline 1 & -\frac{3}{8} & \frac{3}{8} & 1 & 0 \\ 1 & -\frac{5}{8} & \frac{1}{8} & 0 & -1 \end{array}\right],$$

with,

$$G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{3}{8} & 0 \\ 0 & 0 & \frac{3}{8} \end{bmatrix}, \qquad c = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{12} \\ \frac{3}{4} \end{bmatrix}.$$

The starting method is

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{1}{69} & \frac{15}{23} & 0 & 0 & 1 \\ -\frac{8389502714541}{17536590337664} & -\frac{6431381589347}{26304885506496} & \frac{6468387750253}{52609771012992} & 0 & 1 \\ \hline -\frac{76717}{813696} & \frac{16555}{172608} & -\frac{261211}{5153664} & \frac{15050336767}{306801000012} & 1 \\ \frac{14813}{33904} & \frac{2093}{79112} & -\frac{19665}{214736} & -\frac{137004612013}{1124937000044} & 0 \end{array}\right].$$

The finishing method is

$$\left[\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 & 1 \\ \frac{73657}{147246} & \frac{8169}{49082} & 0 & 0 & 1 \\ \check{a}_{41} & \check{a}_{42} & \check{a}_{43} & 0 & 1 \\ \hline -\frac{39483516397}{33753199296} & -\frac{21949176959}{14376252216} & \frac{72556835}{310477632} & \frac{125815818269842236636 0446}{510856001853430663190721} & 1 \end{array}\right],$$

with

$$\begin{bmatrix} \check{a}_{41} \\ \check{a}_{42} \\ \check{a}_{43} \end{bmatrix} = \begin{bmatrix} \frac{4758371220638562428145080980 59}{3049841390168322097173931516928} \\ \frac{2274743067622435465896394035751}{9103776549652441460064185578030 08} \\ -\frac{4178972380495013968454494687}{1706424845295677874426276584448} \end{bmatrix}.$$

7.1.4 Comparison with the Gauss method

We will compare the performance of our new methods, Symeff43 and G4123, with that of the classical Gauss method with order $p = 4$. This Gauss method is known to be symplectic. The classical Gauss method has the Runge–Kutta tableau:

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}. \tag{7.1}$$

7.1.5 Comparison with the classical fifth order method

To study the performance of the new method RKEG5, we will use a classical fifth order method with six stages. We will use the Kepler problem as a test problem with a constant step size h. The classical fifth order method is given as

$$\begin{array}{c|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{3}{4} & \frac{3}{16} & -\frac{3}{8} & \frac{3}{8} & \frac{9}{16} & 0 & 0 \\ 1 & -\frac{3}{7} & \frac{8}{7} & \frac{6}{7} & -\frac{12}{7} & \frac{8}{7} & 0 \\ \hline & \frac{7}{90} & 0 & \frac{16}{45} & \frac{2}{15} & \frac{16}{45} & \frac{7}{90}. \end{array}$$

7.2 Test problems

7.2.1 The simple pendulum problem

The equations of motion are defined as

$$y_1' = -\sin(y_2), \qquad y_2' = y_1.$$

This is based on the Hamiltonian

$$H = \frac{y_1^2}{2} - \cos(y_2).$$

We will carry out experiments with this problem using initial values $y_2 = 1.2 \approx 0.382\pi$, $y_1 = 0$, as a low amplitude case, and as a $y_2 = 1.76 \approx 0.560\pi$, $y_1 = 0$ as a high amplitude case.

7.2.2 The Kepler problem

The Kepler problem describes the motion of two bodies under mutual gravitational attraction, with equation of motion

$$\begin{aligned} y_1' &= y_3, \\ y_2' &= y_4, \\ y_3' &= \frac{-y_1}{(y_1^2 + y_2^2)^{3/2}}, \\ y_4' &= \frac{-y_2}{(y_1^2 + y_2^2)^{3/2}}. \end{aligned}$$

To obtain an elliptic orbit with eccentricity $e \in [0, 1)$, we use initial values

$$(y_1, y_2, y_3, y_4) = (1 - e, 0, 0, \sqrt{\tfrac{1+e}{1-e}}).$$

The Hamiltonian for this problem is

$$H = \tfrac{1}{2}(y_3^2 + y_4^2) - \frac{1}{\sqrt{y_1^2 + y_2^2}},$$

and this should preserved to within narrow bounds.
The second conserved quantity is angular momentum, which is given by

$$A = y_1 y_4 - y_2 y_3,$$

A series of experiments has been performed using this problem with step size $h = 0.01$ and $e = 0.3$ and $e = 0.5$ respectively.

7.2.3 The Hénon-Heiles problem

Hénon–Heiles approximated the Hamiltonian, and hence the total energy is given by

$$H = \tfrac{1}{2}(y_3^2 + y_4^2) + \tfrac{1}{2}(y_1^2 + y_2^2) + y_1^2 y_2 - \tfrac{1}{3}y_2^3.$$

$$\begin{aligned} y_1' &= y_3, \\ y_2' &= y_4, \\ y_3' &= -y_1 - 2y_1 y_2, \\ y_4' &= -y_2 - y_1^2 + y_2^2. \end{aligned}$$

We choose

$$(y_1, y_2, y_3, y_4) = (0, 0, \sqrt{0.3185}, 0).$$

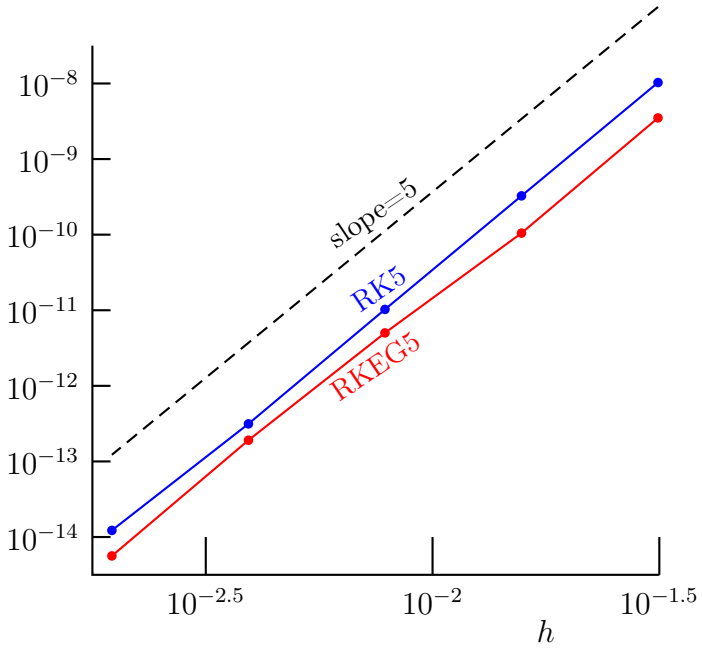


Figure 7.1: Global error for the Kepler problem with $e = 0.3$, $h = \frac{\pi}{n}$

7.3 Results and discussions for RKEG5

h	n	RK5	RKEG5
$\frac{\pi}{100}$	100	-7.9876	-8.4555
$\frac{\pi}{200}$	200	-9.4868	-9.9794
$\frac{\pi}{400}$	400	-10.9888	-11.2999
$\frac{\pi}{800}$	800	-12.5028	-12.7194
$\frac{\pi}{1600}$	1600	-13.9121	-14.2500

Table 7.1: Logrithm (base 10) of global error for the Kepler problem with $e = 0.3$ on $[0,\pi)$.

In order to analyse the outcomes due to effective order, we will compare RKEG5 with RK5. In Figure 7.1, we have compared these methods with stepsize $h = \frac{\pi}{n}$ and for $n = 100 : 1600$ number of steps for the Kepler problem. It is clearly shown that

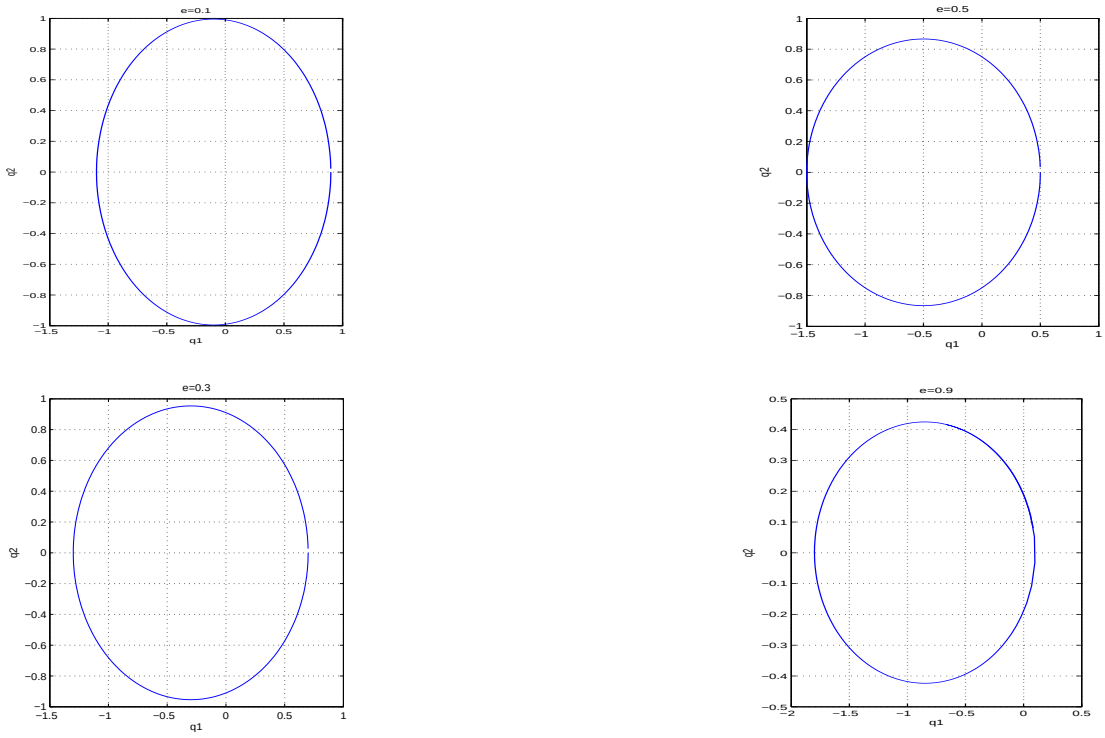


Figure 7.2: Orbit patterns for the Kepler problem with $e = 0.1,\ 0.3,\ 0.5,\ 0.9$ for RKEG5

method RKEG5 performs better than RK5. Moreover the new method have less computational cost compared to RK5. The numerical results for these methods are given in Table 7.1.
The second experiment is to check the orbit pattern by the method RKEG5. We shall consider different eccentricities $e = 0.1,\ 0.3,\ 0.5,\ 0.9$ against the same number of steps $n = [10\ 20\ 40\ 80\ 160\ 320]$.

7.4 Results and discussions for Symeff43

In Figures 7.3 and 7.4 we present results for the simple pendulum problem with initial values $y = [0, 1.2]^\intercal$ and $y = [0, 1.76]^\intercal$ respectively. In each case $h = 0.01$.

For Symeff43 method although the value H is not conserved using a symplectic integrator, we expect H to lie within a narrow band and this is observed in these results. For comparison, the high amplitude case is repeated but using the fourth

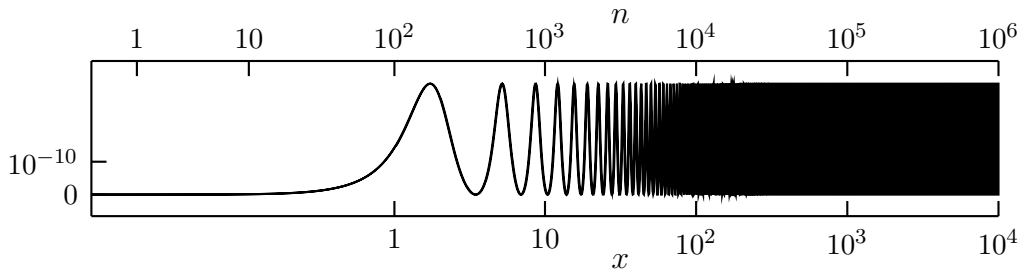


Figure 7.3: The variation in the Hamiltonian for the simple pendulum with initial value $[0, 1.2]^\intercal$ using Symeff43 method with $h = 0.01$

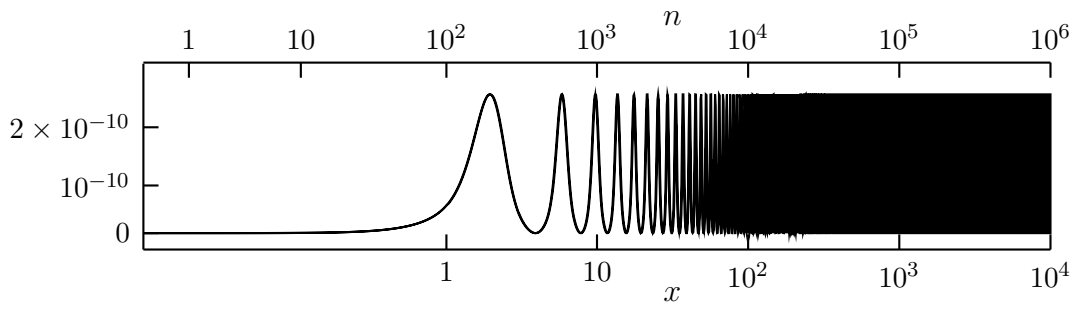


Figure 7.4: The variation in the Hamiltonian for the simple pendulum with initial value $[0, 1.76]^\intercal$ using Symeff43 method with $h = 0.01$

order Gauss method. The Gauss results, shown in Figure 7.5, show significantly narrower bands for the H values.

The second test problem is taken as the Kepler problem, two experiments has been performed using this problem with step size $h = 0.01$ and $e = 0.3$ and $e = 0.5$ respectively.

For this problem, the Symeff43 method preserves angular momentum exactly and this should be observed in simulations as shown in Figures 7.7 and 7.9. The deviations from the initial value are similar to what would be formed through the growth of round-off errors in the individual steps.

For the Hénon-Heiles problem The deviation from the initial value of H is shown in Figure 7.10 for the Symeff43 method.

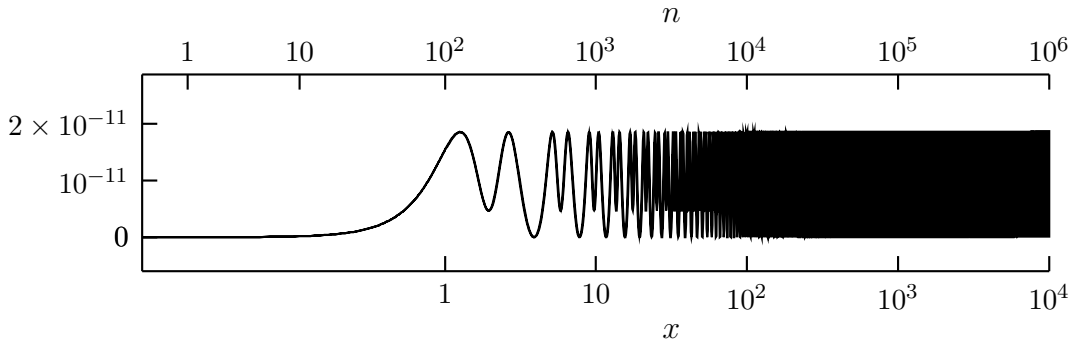


Figure 7.5: The variation in the Hamiltonian for the simple pendulum with initial value $[0, 1.76]^\intercal$ using the Gauss method with $h = 0.01$

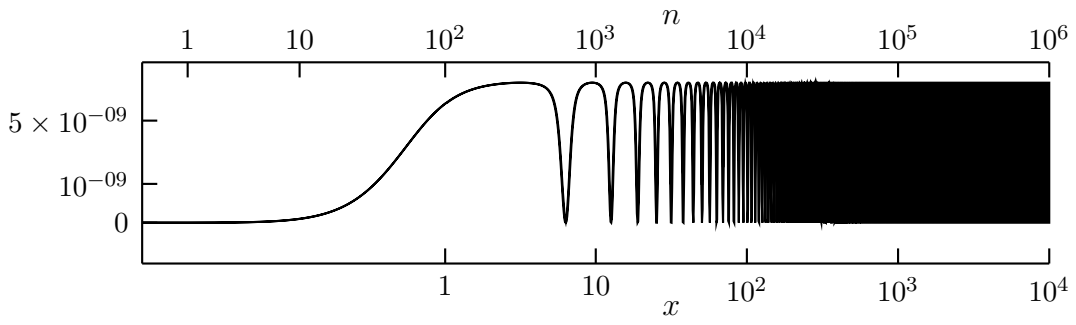


Figure 7.6: The variation in the Hamiltonian for the Kepler problem with initial value $e = 0.3$ using Symeff43 method with $h = 0.01$

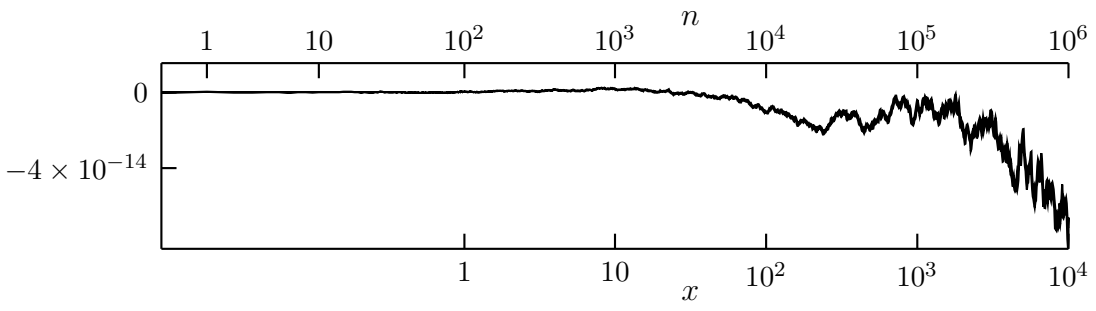


Figure 7.7: The variation in the angular momentum for the Kepler problem with initial value $e = 0.3$ using Symeff43 method with $h = 0.01$

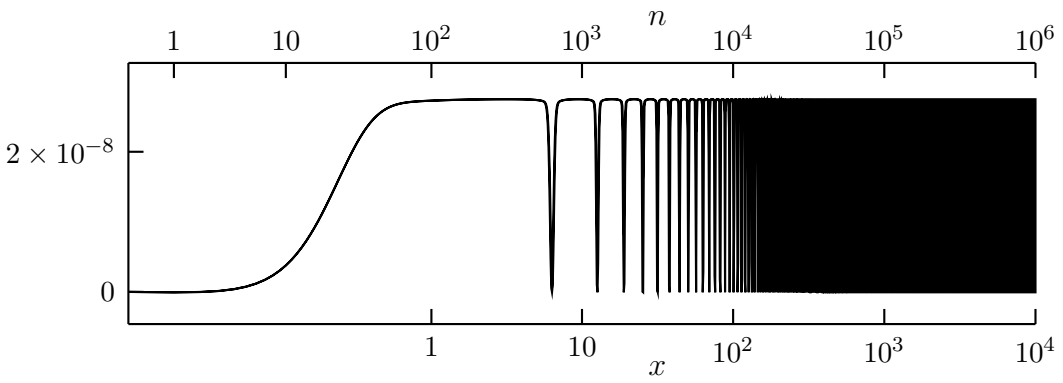


Figure 7.8: The variation in the Hamiltonian for the Kepler problem with initial value $e = 0.5$ using Symeff43 method with $h = 0.01$

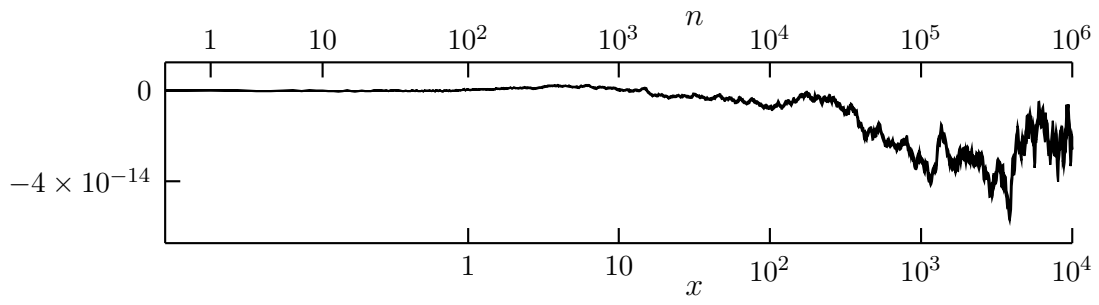


Figure 7.9: The variation in the angular momentum for the Kepler problem with initial value $e = 0.5$ using Symeff43 method with $h = 0.01$

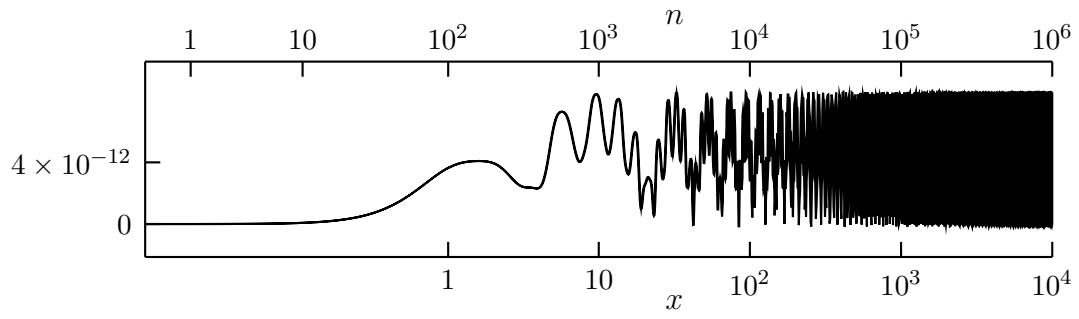


Figure 7.10: The Hénon–Heiles problem solved by the Symeff43 method with $h = 0.01$

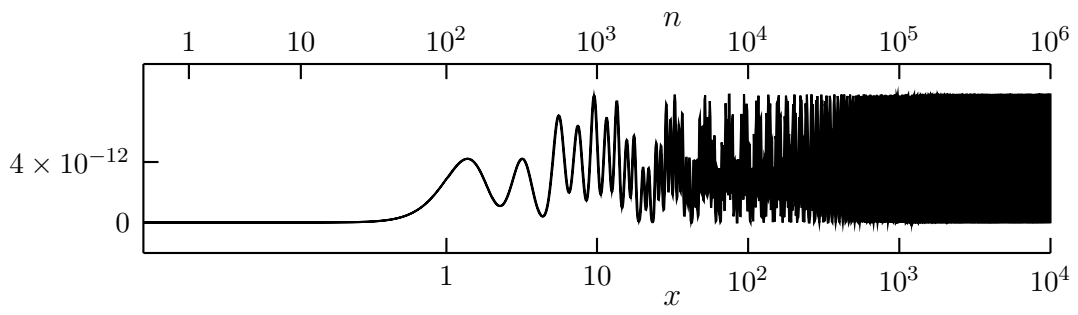


Figure 7.11: The Hénon–Heiles problem solved by the Gauss method with $h = 0.01$

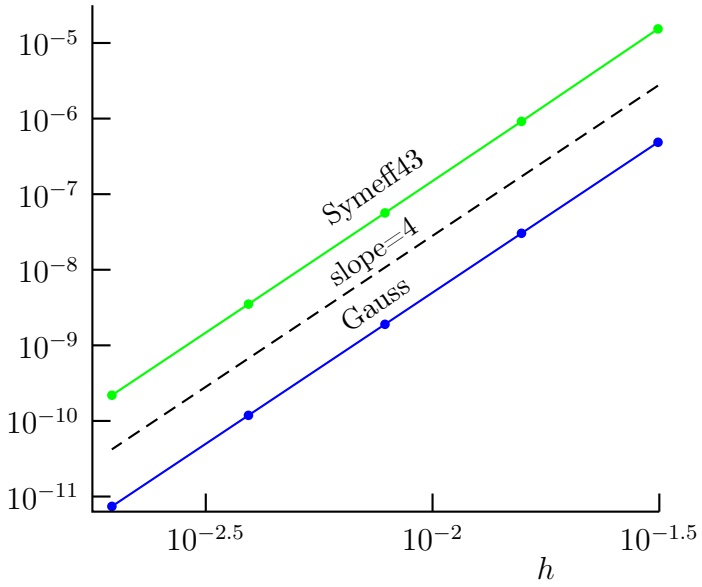


Figure 7.12: Global error for the Kepler problem with $e = 0.3$

Our comparisons start with a numerical experiment based on Figure 7.11 but with the Gauss method used instead of the Symeff43 . This case is presented because it is typical of many similar calculations. We observe that the ability of the new method to adhere to almost constant Hamiltonian values is inferior to the Gauss method.

A second comparison, based on the Kepler problem, aims to compare accuracy of the two methods, and, at the same time to verify the order behaviour. The results displayed in Figure 7.12 confirm that the two methods both show the same fourth order behaviour, with the Gauss method achieving greater accuracy for this problem.

Finally, we remark that, Symeff43 has definite advantages, our experiments have all been carried out with small problems, but if N were very large, the implementation costs would become of overwhelming significance. There are three components of the cost in a Newton type of stage evaluation. These are

(a) the occasional cost of evaluating a current value of the Jacobian matrix

(b) factorization of matrices of the form $I_s \otimes I_N - hA \otimes J$ in preparation for Newton iterations and

(c) the actual substitutions to evaluate the Newton updates in each iteration.

We will assume that the frequency, and therefore the cost of (a) is similar for the two methods. As far as (b) is concerned, we have little choice for the Gauss method, in which $s = 2$, than to do a full factorization of a $2N \times 2N$ matrix with a cost equal to a small factor times $8N^3$ for the new method for which $s = 3$, if we use transformations as we recommend, this is reduced to $3N^3$, or only N^3 if parallelism is available. For (c) the Gauss method would incur a cost of $4N^2$ in each iteration and only $3N^2$ for the new method, when implemented sequentially; reduced further to N^2 in a parallel computer.

7.5 Results and discussions for G4123

To illustrate the behaviour of G4123 method, first we consider the simple pendulum problem. The results are shown in Figures 7.13 and 7.14 respectively for the amplitudes 1.2 and 1.76 for G4123 method.

For G4123 method we can observe that the Hamiltonian is approximately conserved. This is typical behaviour of parasitism-free G-symplectic general linear methods. For comparison, again the high amplitude case is repeated by using the Gauss method. It is clearly observed that the performance of the G4123 method is comparable with the Gauss method.

Next we consider the Kepler problem and test the efficiency of this method for $e = 0.3$ and $e = 0.5$ respectively.

Again for this problem, the G4123 method performs very well. This method has the ability to closely conserve the Hamiltonian and angular momentum and is comparable to what we obtained using the implicit Runge–Kutta method based on Gaussian quadrature. This can be seen in Figures 7.15–7.18.

Lastly we take the Hénon-Heiles problem and observed in Figure 7.19 that the Hamiltonian varies from its initial values for G4123 method.

Thus the G4123 is much more efficient because the Gauss method is fully implicit while the A matrix of the G4123 general linear method is lower triangular with only two non-zero diagonal elements.

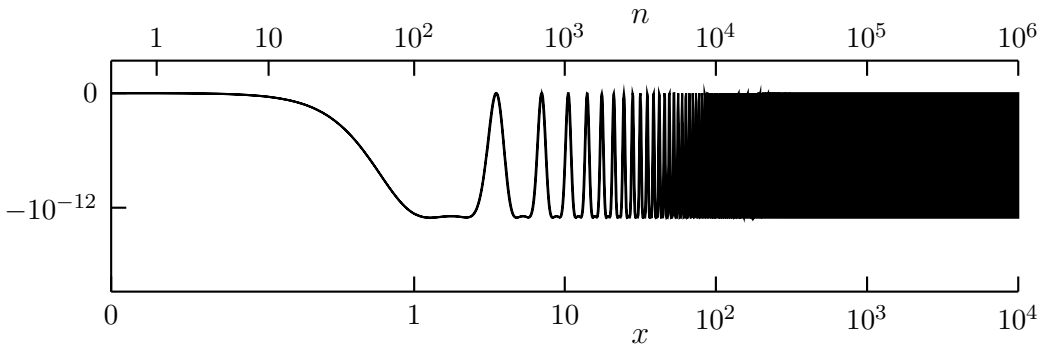


Figure 7.13: The Pendulum problem solved by the G4123 method with $h = 0.01$ and $y = [0, 1.2]^\intercal$

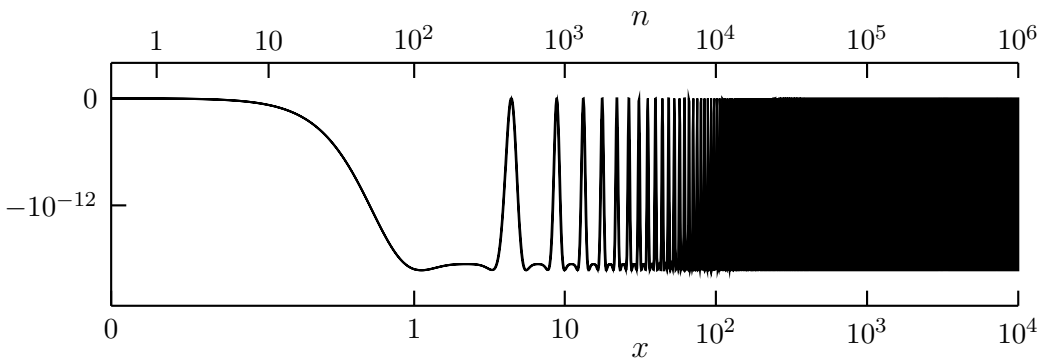


Figure 7.14: The Pendulum problem solved by the G4123 method with $h = 0.01$ and $y = [0, 1.76]^\intercal$

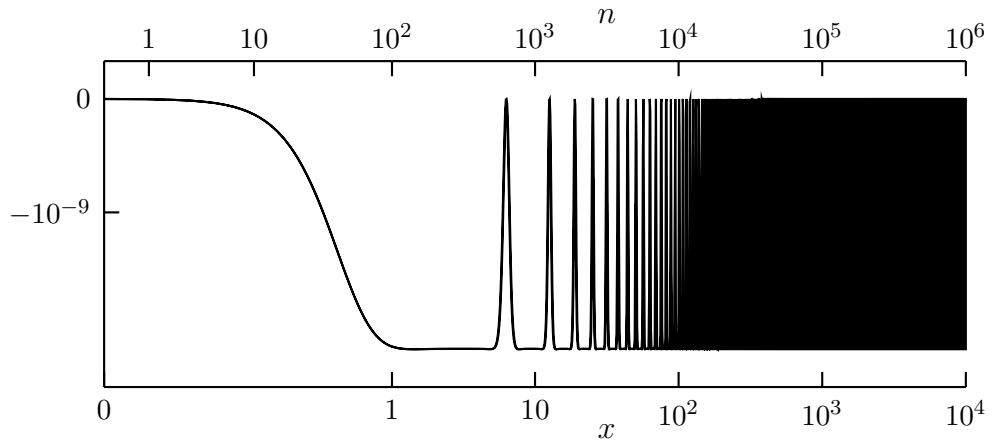


Figure 7.15: The variation in the Hamiltonian for the Kepler problem with $h = 0.01$ and $e = 0.3$ solved by the G4123 method

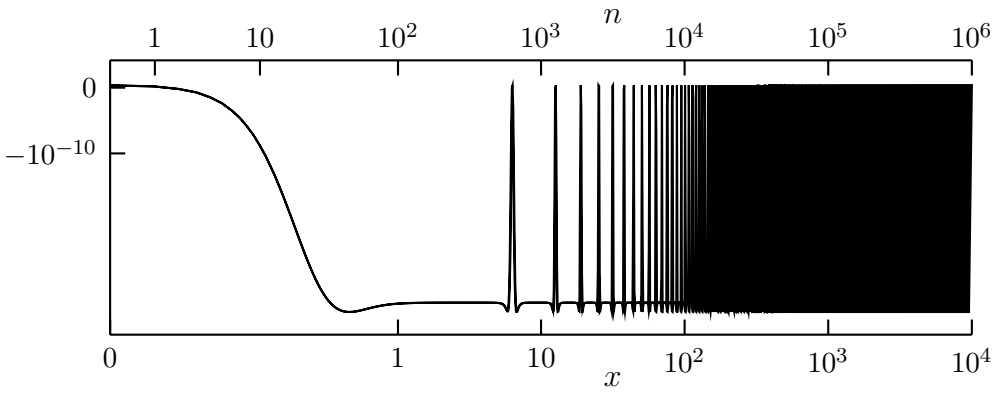


Figure 7.16: The variation in the Hamiltonian for the Kepler problem with $e = 0.5$, using G4123 with $h = 0.01$

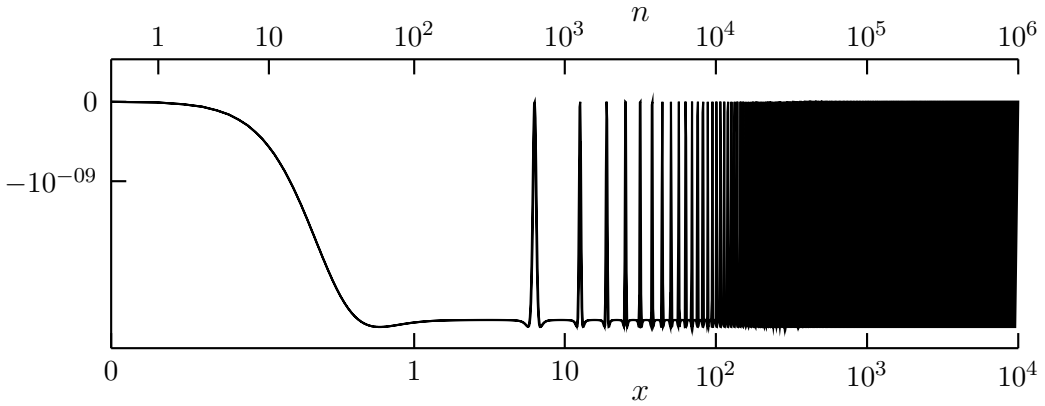


Figure 7.17: Variation of the angular momentum for the Kepler problem solved by the G4123 method with $h = 0.01$ and $e = 0.5$

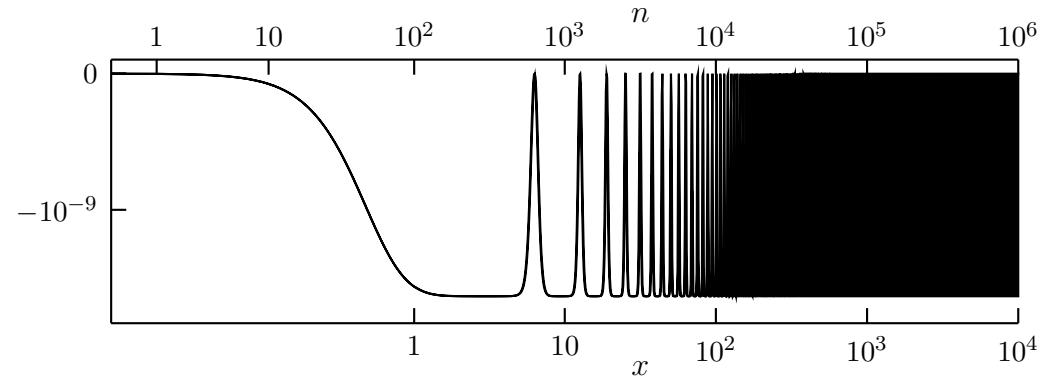


Figure 7.18: Variation of the angular momentum for the Kepler problem solved by the G4123 method with $h = 0.01$ and $e = 0.3$

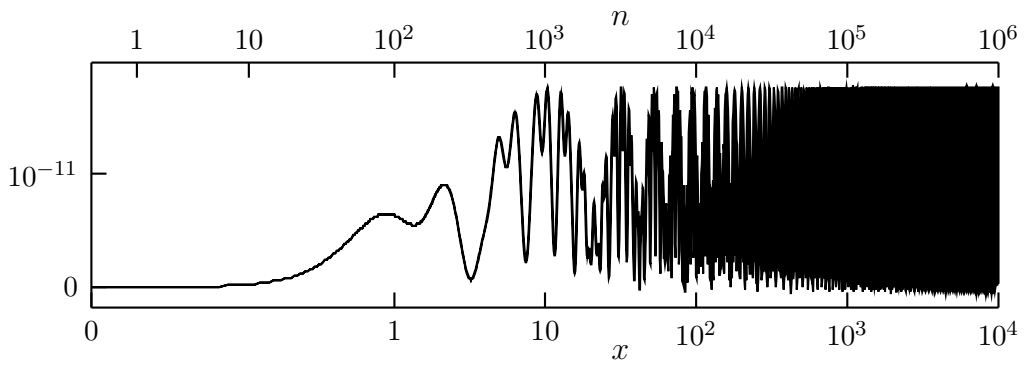


Figure 7.19: The Hénon–Heiles problem solved by the G4123 method with $h = 0.01$

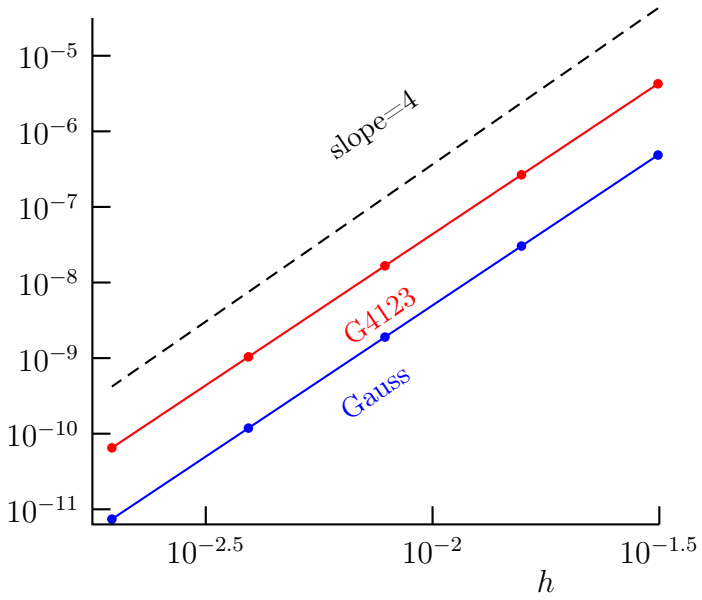


Figure 7.20: Global error for the Kepler problem with $e = 0.3$

We aim to compare the accuracy of the two methods, and at the same time want to verify the order four behaviour. The results given in Figure 7.20 confirms that the two methods G4123 and Gauss both have the same order behaviour with the Gauss method slightly greater accuracy for the Kepler problem. Also the numerical results are given for both methods in Table 7.2.

h	Gauss	G4123
$\frac{\pi}{100}$	-6.3136	-5.3711
$\frac{\pi}{200}$	-7.5176	-6.5750
$\frac{\pi}{400}$	-8.7216	-7.7791
$\frac{\pi}{800}$	-9.9259	-8.9832
$\frac{\pi}{1600}$	-11.1303	-10.1873

Table 7.2: Global error for the Kepler problem with e = 0.3 over π.

Chapter 8

Conclusions and future work

The main aim of this thesis was to build different theoretical approaches to accurate and efficient numerical methods for studying differential equation systems. Moreover these numerical techniques were for the long term integration of conservative systems. These are the questions which were explored in this thesis with an emphasis on canonical Runge–Kutta methods and more generally on G-symplectic general linear methods with high order.

The thesis provides an understanding of many different ideas. In Chapter 1 a review of traditional methods for the numerical solution of ordinary differential equations was given. Also an introduction to Hamiltonian systems and their underlying properties for conservative problems was provided. In Chapter 2 this was extended to a detailed study of numerical methods such as Runge–Kutta methods, linear multistep methods and general linear methods.

In Chapter 3 we have provided an introduction to effective order being an important concept with many applications. We have derived a fifth effective order explicit Runge–Kutta method satisfying the $C(2)$ assumption. With this assumption we have reduced to a smaller set of order conditions for the construction of such methods. It was also observed that the $D(1)$ assumption gives no new generality to the construction of fifth effective order methods.

In Chapter 4 we have constructed a three stage symplectic Runge–Kutta method with effective order four. The new method not only have effective order four but are also symplectic. Thus, it shared the advantages of effective order methods and symplectic methods. Moreover, the coefficient matrix of the new method has only real eigenvalues and this made it possible to implement the method efficiently. Numerical

experiments proved that in terms of the global error behaviour the new method was comparable to the Gauss method. Experiments with some model problems proved that the method was able to preserve quadratic invariants over extended time intervals and to closely adhere to energy invariance for Hamiltonians problems. It was argued that for large N the cost of the Gauss method scales up more rapidly than this new method.

In Chapter 5 we have investigated the order conditions for G-symplectic general linear methods using a **B**-series approach. It was observed that the G-symplectic structure considerably modified the nature of order conditions, and also discovered that for G-symplectic methods rooted trees belonging to same tree had equivalent order conditions and if the tree were superfluous, they were automatically satisfied and were ignored in a similar way as for Runge–Kutta methods. Another observation was that the evaluation of the order conditions for G–symplectic general linear methods tree by tree it was possible to obtain order conditions up to any required order.

Moreover, a special case of order four methods with $V = \text{diag}(1, -1)$ was considered in detail. It was found that although there were only two order conditions for such methods, there were additional constraints on a possible starting method. But this was not a serious handicap because in a constant step size implementation, a complicated starting method does not represent a computational overhead. Furthermore, the numerical test proved that these methods provide qualitatively correct numerical results over extended time.

In Chapter 6 a sixth order G-symplectic general linear method was constructed with no parasitism. This had five stages, two of these were explicit and the remaining three were diagonally implicit. This method had four output values. Complex number were chosen for matrices B and V such that the second and third rows of matrices B and V were complex conjugates of each other. Moreover, for V to have eigenvalues on the unit circle, we took $V = \text{diag}(1, \zeta, -\widehat{\zeta}, -1), \quad |\zeta| = 1, \quad \text{Im } \zeta \neq 0$ and we used value $\zeta = i$ in the construction of a practical method. Also $G = \text{diag}(1, -\frac{1}{2}, -\frac{1}{2}, 1)$ where the factors $\frac{1}{2}$ were inserted because to simplify notation when we converted to a real formulation of the method. We also used $C(2)$ conditions to reduce the number of order conditions.

There are several open questions we would like to explore. The construction of a sixth order G-symplectic general linear method was a complicated task and, although a 6245 methods has been constructed. It is not yet known whether or not 6th order methods with $s = 4$, might do exist, but without the property of symmetry. It

is also not yet known how well these methods will perform with practical problems.

It is very interesting to investigate eighth order general linear methods with the desired features of G-symplecticity and time reversal symmetry, which are free from parasitism. These methods are highly desirable for long term integration of Hamiltonian systems.

We would like to investigate methods for problems in which $y' = f(y)$ and also $y'' = g(y)$ are given have the potential to be more efficient for stiff problems. It would be interesting to know if second derivative methods have any useful role in long term integration for mechanical problems.

Bibliography

[1] R. Alexander, Diagonally implicit Runge–Kutta methods for stiff ODEs, *SIAM J. Numer. Anal.*, 14 (1977), pp 1006–1021.

[2] 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*, 1883.

[3] G. Benettin and A. Giorgilli, On the Hamiltonian interpolation of near to identity symplectic mappings with application to symplectic integration algorithms, *J. Statis. Phys.*, 74 (1994), pp 1117–1143.

[4] K. Burrage, A special family of Runge–Kutta methods for solving stiff differential equations, *BIT Numer. Math.*, 18 (1978), pp 22–41.

[5] K. Burrage and J. C. Butcher, Stability criteria for Implicit Runge–Kutta Methods, *SIAM J. Numer. Anal.*, 16 (1979), pp 46–57.

[6] K. Burrage and J. C. Butcher, Non-linear stability of general class of differential equation methods, *BIT Numer. Math.*, 20 (1980), pp 185–203.

[7] K. Burrage, J. C. Butcher and F. H. Chipman, An implementation of singly-implicit Runge–Kutta methods, *BIT Numer. Math.*, 20 (1980), pp 326–340.

[8] J. C. Butcher, Coefficients for the study of Runge–Kutta integration processes, *J. Austral. Math. Soc.*, 3 (1963), pp 185–201.

[9] J. C. Butcher, Implicit Runge–Kutta processes, *Math. Comp.*, 18 (1964), pp 50–64.

[10] J. C. Butcher, A modified multistep method for numerical integration of ordi nary differential equations, *J Assoc. Comput. Mach.*, 12 (1965), pp 124–135.

[11] J. C. Butcher, On the convergence of numerical solution to ordinary differential equations. *Math. Comp.*, 20 (1966), pp 1–10.

[12] J. C. Butcher, The effective order of Runge–Kutta methods, *Lecture Notes in Math.*, 109 (1969), pp 133–139.

[13] J. C. Butcher, An algebraic theory of integration methods, *Math. Comp*, 26 (1972), pp 79–106.

[14] J. C. Butcher, *On the implementation of implicit Runge–Kutta methods*, *BIT Numer. Math.*, 16 (1976), pp 237–240.

[15] J. C. Butcher, A transformed implicit Runge–Kutta method, *J. ACM.*, 26 (1979), pp 731–738.

[16] J. C. Butcher, Diagonally–implicit multi–stage integration methods, *Appl. Numer. Math.*, 11 (1993), pp 347–363.

[17] J. C. Butcher, Order and effective order, *Appl. Numer. Math.*, 28 (1998), pp 179–191.

[18] J. C. Butcher, General linear methods, *Acta Numer.*, 15 (2006), pp 157–256.

[19] J. C. Butcher, Numerical Methods for Ordinary Differential Equations, *Wiley, second edition*, 2008.

[20] J. C. Butcher and P. Chartier, A generalization of singly-implicit Runge–Kutta methods, *Appl. Numer. Math.*, 24 (1997), pp 343–350.

[21] J. C. Butcher and P. Chartier, The effective order of singly-implicit Runge–Kutta methods, *Numer. Algorithms*, 20 (1999), pp 269–284.

[22] J. C. Butcher and D. J. L. Chen, Effective order Diagonally Extended Singly-implicit Runge–Kutta methods for stiff differential equations, *unpublished work.*

[23] J. C. Butcher and M. T. Diamantakis, DESIRE: diagonally extended singly implicit Runge–Kutta effective order methods, *Numer. Algorithms.*, 17 (1998), pp 121–145.

[24] J. C. Butcher,Y. Habib, A. T. Hill and T. J. T. Norton, The Control of Parasitism in G-Symplectic Methods, *SIAM J. Numer. Anal.*, in press.

[25] J. C. Butcher and L. L. Hewitt, The existence of symplectic general linear methods, *Numer. Algorithms.*, 51 (2009), pp 77–84.

[26] J. C. Butcher, A. T. Hill and T. J. T. Norton, Symmetric general linear methods, in preparation.

[27] J. C. Butcher and G. Imran, Symplectic effective order methods, *Numer. Algorithms.*, (2014), pp 1–19.

[28] J. C. Butcher and G. Imran, Order conditions for G-symplectic methods, Submitted.

[29] J. C. Butcher and G. Imran, A G-symplectic method with order 6, in preparation.

[30] T. M. H. Chan, Algebraic structures for the analysis of numerical methods, PhD. Thesis, University of Auckland, NZ, 1998.

[31] G. J. Cooper, Stability of Runge–Kutta methods for trajectory problems, *IMA J. Numer. Anal.*, 7 (1987), pp 1–13.

[32] M. Crouzeix, Sur la B-stabilité des méthodes de Runge–Kutta, *Numer. Math.*, 32 (1979), pp 75–82.

[33] C. F. Curtiss and J. O. Hirschfelder, Integration of stiff equations, *Proc. Nat. Acad. Sci.*, 38 (1952), pp 235–243.

[34] G. Dahlquist, Convergence and stability in the numerical integration of ordinary differential equations, *Math. Scand.*, 4 (1956), pp 33–53.

[35] G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, *Trans. Royal Inst. Techn. Stockholm, Sweden.*, 130, 1959.

[36] G. Dahlquist, A special stability property for linear multistep methods, *BIT Numer. Math.*, 3 (1963), pp 27–43.

[37] G. Dahlquist, Error analysis of a class of methods for stiff nonlinear initial value problems, *Numerical Analysis, Dundee, Lecture Notes in Mathematics*, 506 (1976), pp 60–74.

[38] G. Dahlquist, G-stability is equivalent to A-stability, *BIT Numer. Math.*, 18 (1978), pp 384–401.

[39] G. Dahlquist, On one-leg multistep methods, *SIAM J. Numer. Anal.*, 20 (1983) pp 1130–1138.

[40] B. L. Ehle, On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems, *Report 2010 (1969), University of Waterloo, Ontario, Canada*.

[41] T. Eirola and J. M. Sanz-Serna, Conservation of integrals and symplectic structure in the integration of differential equations by multistep methods, *Numer. Math.*, 61 (1992), pp 281–290.

[42] K. Feng, On difference schemes and symplectic geometry, *Proc.* 5^{th} *Intern. Symposium on Differential geometry & differential equations*, Beijing(1985), pp 42–58.

[43] C. W. Gear, Hybrid methods for initial value problems in ordinary differential equations, *SIAM J Numer. Anal.*, 2 (1965), pp 69–86.

[44] S. Gill, A process for the step by step integration of differential equations in an automatic digital computing machine, *Proc. Cambridge Philos. Soc.*, 47 (1951), pp 515–534.

[45] W. B. Gragg, and H. J. Stetter, Generalized multistep predictor-corrector methods, *J. Assoc. Comput. Mach.*, 11 (1964), pp 188–209.

[46] Y. Habib, Long-term behaviour of G-symplectic methods, PhD. Thesis, University of Auckland, NZ, 2010.

[47] E. Hairer, C. Lubich and G. Wanner, Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, Springer, second edition, 2005.

[48] E. Hairer, Conjugate–symplecticity of linear multistep methods, *J. Comput. Math.*, 26 (2008), pp 657–659.

[49] E. Hairer and G. Wanner, On the Butcher group and general multivalue methods, *Computing (Arch. Elektron. Rechnen)*, 13 (1974), pp 1–15.

[50] E. Hairer, C. Lubich and G. Wanner, Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, Springer, first edition, 2003.

[51] K. Heun, Neue Methoden zur approximativen Integration Differentialgleichungen einer unabhäneigen Veränderlichen, *Z. Math. Phys.*, 45 (1900), pp 23–38.

[52] A. Iserles, H. Z. Munthe-Kaas S.P. Nørsett, and A. Zanna, Lie-group methods, *Acta Numer.*, 9 (2000), pp 215–365.

[53] U. Kirchgraber, Multistep methods are essentially one-step methods, *Numer. Math.*, 48 (1986), pp 85–90.

[54] W. Kutta, Beitrag zur näherungsweisen Integration totaler Differentialgleichungen, *Z. Math. Phys.*, 46 (1901), pp 435–453.

[55] F. M. Lasagni, Canonical Runge–Kutta methods, *ZAMP.*, 39 (1988), pp 952–953.

[56] W. Liniger, A criterion for A-Stability of Linear Multistep Integration Formulae, *Computing (Arch. Elektron. Rechnen).*, 3 (1968), pp 280–285.

[57] W. Liniger, G. Dahlquist and O. Nevanlinna, Stability of two-step methods for variable integration steps, *SIAM J. Numer. Anal.*, 20 (1983), pp 1071–1085.

[58] M. Lopez-Marcos, J. M. Sanz-Serna, and R. D. Skeel, Cheap Enhancement of Symplectic Integrators, *D. F. Griffiths and G. A. Watson editors, Numerical Analysis.*, (1996), pp 107–122.

[59] W. E. Milne, Numerical integration of ordinary differential equations, *Amer. Math. Monthly.*, 33 (1926), pp 455–460.

[60] W. E. Milne, A note on the numerical integration of differential equations, *J. Research Nat. Bur. Standards.*, 43 (1949), pp 537–542.

[61] F. R. Moulton, New methods in Exterior Ballistics, University of Chicago, 1926.

[62] H. Z. Munthe-Kaas K. Engø, A. Marthinsen, Diffman User's Guide Version 2.0, Department of Informatics, University of Bergen, Norway, 2000.

[63] S.P. Nørsett, Runge–Kutta methods with a multiple real eigenvalue only, *BIT Numer. Math.*, 16 (1976), pp 388–393.

[64] E. J. Nyström, Über die numerische Integration von Differentialgleichungen, *Acta Soc. Sci. Fennicae.*, 50 (1925), pp 1–54.

[65] C. Runge, Über die numerische Auflösung von Differentialgleichungen , *Math Ann.* 48 (1895), pp 167–178.

[66] R. D. Ruth, A canonical integration technique, *IEEE Trans. Nuclear Science.*, 30 (1983), pp 2669–2671.

[67] J. M. Sanz-Serna. Runge–Kutta schemes for Hamiltonian systems. *BIT Numer. Math.*, 28 (1988), pp 877–883.

[68] J. M. Sanz-Serna, Symplectic integrators for Hamiltonian problems, *Acta Numer.*, 1 (1992), pp 243–286.

[69] J. M. Sanz-Serna and L. Abia, Order conditions for canonical Runge-Kutta schemes. *SIAM J Numer. Anal.*, 28 (1991), pp 1081–1096.

[70] J. M. Sanz-Serna and M. P. Calvo, Numerical Hamiltonian Problems, Chapman and Hall, first edition, 1994.

[71] H. J. Stetter, Local estimation of the global discretization error, *SIAM J Numer. Anal.*, 8 (1971), pp 512–523.

[72] D. Stoffer, General linear methods: connection to one step methods and invariant curves, *Numer. Math.*, 64 (1993), pp 395–408.

[73] Y. B. Suris, Preservation of symplectic structure in the numerical solution of Hamiltonian systems, Numerical Solution of Differential Equations (S. S. Filippov, ed.) *Akad. Nauk SSSR, (In Russian).*, (1988), pp 148–160.

[74] Y. B. Suris, Canonical transformations generated by methods of Runge–Kutta type for the numerical integration of the system $x'' = \frac{\partial u}{\partial x}$, *Zh. Vychisl. Mat. Fiz. (In Russian).*, 29 (1989), pp 202–211.

[75] Y. F. Tang, The symplecticity of multistep methods, *Computers Math. Applic.*, 25 (1993), pp 83–90.

[76] R. de. Vogelaere, Methods of integration which preserve the contact transformation property of the Hamiltonian equations, Report No. 4, Dept. Math, Univ of Notre Dame, Notre Dame, Ind. 1956.

[77] Hans Van de Vyver, An embedded exponentially fitted Runge-Kutta-Nyström method for the numerical solution of orbital problems, *New Astronomy.*, 11 (2006), pp 577–587.

[78] D. S. Watanabe and Q. M. Sheikh, One-leg Formulas for Stiff Ordinary Differential Equations, *SIAM J. Sci. and Stat. Comput.*, 5 (1984), pp 489–496.

[79] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, first edition, 1965.