The Algebraic Structure of B-Series

A thesis presented in total fulfillment of the requirements for the degree of

Master of Science

in

Mathematics

at Massey University, Palmerston North

New Zealand

James Benn

2010

Abstract

Runge-Kutta methods are some of the most widely used numerical integrators for approximating the solution of an ordinary differential equation (ODE). These methods form a subset of a larger class of numerical integrators called B-series methods. B-Series methods are expressed in terms of rooted trees, a type of combinatorial graph, which are related to the vector field of the ODE that is to be solved. Therefore, the conditions for B-series methods to preserve important properties of the solution of an ODE, such as symplecticity and energy-preservation, may be expressed in terms of rooted trees. Certain linear combinations of rooted trees give conditions for a B-series to be Energy-preserving while other linear combinations give conditions for a B-series to be Hamiltonian. B-series methods may be conjugate (by another B-series) to an Energy-preserving or an Hamiltonian B-series. Such B-series methods are called conjugate-to-Energy preserving and conjugate-to-Hamiltonian, respectivley. The conditions for a B-series to be conjugate-to-Energy preserving or conjugate-to-Hamiltonian may also be expressed in terms of rooted trees.

The rooted trees form a vector space over the Real numbers. This thesis explores the algebraic structure of this vector space and its natural energy-preserving, Hamiltonian, conjugate-to-Energy preserving and conjugate-to-Hamitlonian subspaces and dual subspaces.

The first part of this thesis reviews important concepts of numerical integrators and introduces the general Runge-Kutta methods. B-series methods, along with rooted trees, are then introduced in the context of Runge-Kutta methods. The theory of rooted trees is developed and the conditions for a B-series to be Hamiltonian or have first integral are given and discussed. In the final chapter we interpret the conditions in the context of vector spaces and explore the algebraic structure of, and the relationships between, the natural vector subspaces and dual spaces.

"Do you like Phil Collins? I've been a big Genesis fan ever since the release of their 1980 album, Duke. Before that, I really didn't understand any of their work. Too artsy, too intellectual. It was on Duke where, uh, Phil Collins' presence became more apparent. I think Invisible Touch was the group's undisputed masterpiece. It's an epic meditation on intangibility. At the same time, it deepens and enriches the meaning of the preceding three albums. Christy, take off your robe."

-Brett Easton Ellis

"Wu-Tang Clan ain't nuttin to f' wit"

-The RZA

"Possibilities of sweetness on technicolor beaches had been trickling through my spine for some time..."

-Vladimir Nabokov

"I have to return some videotapes."

-Brett Easton Ellis

Acknowledgments

Ladies and Gentlemen of the jury, if you've made it this far then you're doing well. First and foremost I would like to thank my parents, Ken and Cheryl, for their support over the past six months. Thanks for the dinners, the room and the heat pump. I couldn't have done it without the heat pump.

I would also like to thank my distinguished supervisor, Prof. Robert McLachlan (aka Rob Dogg, Rob G, the DE Hunter), for his bottomless well of ideas and knowledge. Thank you for the opportunity to attend COCO2010 and pushing me to present work. Thanks for your confidence and encouragement over the past two years and making it an enjoyable experience.

To Matt Perlmutter and Stephen Marsland, thank you for your support and encouragement. It hasn't gone unnoticed.

Thank you to Elena Celledoni, Hans Munthe-Kaas, Brynjulf Owren, Reinout Quispel and Will Wright for their insight, input and assistance while at COCO2010. Big thanks.

A big thank you to Laila, Dave, Caitlin, Jake (aka MC Sleepy Conrad), Jeremy (aka DJ Tender Loins), Ross (aka Fudgey C), Brad and Isabel. Thank you for always reminding me that I am a nerd and will never be accepted by mainstream society.

A very honorable mention goes out to the RZA, the GZA, the Ol Dirty BZA, U-God, Chef, the Ghostface Killah and Meth, Rebel I soldier for the foreclosure don't forget about the Masta, yo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Suppose that x is an unknown function of a variable t. If we know $\frac{dx}{dt}$ (or x'), then we can recover x, up to an additive constant, by integrating x':

$$x(t) = \int x'(t).dt + c = F(t) + c.$$

If we don't know x' but we know x'', then we can still recover x, but now we have to integrate twice and there will be two arbitrary constants:

$$x'(t) = \int x''(t).dt + C_1 = G(t) + C_1$$

$$x(t) = \int \left[G(t) + C_1\right].dt + C_2 = F(t) + C_1 t + C_2.$$

It often happens in mathematics and in applications to other fields that we don't know x', we don't know x'', we don't know any of the derivatives explicitly, but we do have an equation that relates x to one or more of its derivatives. An equation that relates an unknown function to one or more of its derivatives is called an *ordinary differential equation*. From some differential equations we can recover x completely and describe its action explicitly as a function of t (up to one or more arbitrary constants). More frequently, we cannot recover x completely, but we can obtain an equation in t and x which is satisfied by x and involves none of the derivatives of x. Such an equation, carrying one or more arbitrary constants, represents a family of curves called *integral curves* (*solution curves*) of the differential equation.

If an equation contains only ordinary derivatives of one or more dependent variables with respect to a single independent variable, the equation is said to be an *ordinary differential equation* (ODE). An equation involving partial derivatives of one or more dependent variables of two or

more independent variables is said to be a *partial differential equation.*

The *order* of a differential equation is the order of the highest derivative that appears in the equation. We can express an n-th order differential equation in one dependent variable as

$$F(t, x, x', \ldots, x^{(n)}) = 0$$

An n-th order ordinary differential equation is said to be *linear* if F is linear in $x, x', \ldots, x^{(n)}$. If F is not linear then the differential equation is said to be $non - linear$.

It also happens in applications that a single differential equation is insufficient and we in fact need a set of differential equations to describe some phenomena. This set of differential equations is called a system of differential equations.

$$\begin{aligned}
\tfrac{dx_1}{dt} &= g_1(t, x_1, x_2, \ldots, x_n) \\
\tfrac{dx_2}{dt} &= g_2(t, x_1, x_2, \ldots, x_n) \\
&\vdots \\
\tfrac{dx_n}{dt} &= g_n(t, x_1, x_2, \ldots, x_n)
\end{aligned}$$

When g_1, g_2,..., g_n are linear in the variables x_1,x_2, ..., x_n the system is said to be a *linear system*, otherwise the system is non-linear. A solution to the system is a set of functions $(x_1(t), x_2(t), ..., x_n(t))$ which satisfy all equations in the system. We may write the system above in the form

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$$

where $\mathbf{x}$ represents a point in the *Phase Space* (the space in which all possible states of the system may be represented, where each possible state is represented by a unique point in the Phase Space) and the vector valued function $\mathbf{f}(\mathbf{x})$ represents a vector-field which, at any point in the phase space, prescribes the velocity (direction and speed) of the solution $\mathbf{x}(t)$ that passes through that point (Hairer, Lubich and Wanner(2002)).

The $flow$ map, φ_t, of the system is a mapping which, to any point $\mathbf{x}_0$ in the phase space, associates the value $\mathbf{x}(t)$ of the solution with initial value $\mathbf{x}(0) = \mathbf{x}_0$. This map is thus defined by

$$\varphi_t(\mathbf{x}_0) = \mathbf{x}(t) \quad if \quad \mathbf{x}(0) = \mathbf{x}_0.$$

(Hairer, Lubich and Wanner (2002)).

Finally, there are differential equations (and systems of differential equations) from which we can extract no explicit solutions and no integral curves. Such equations have to be approached by other methods, in particular, numerical methods.

1.1 Outline of Thesis

It is with these numerical methods, mentioned above, that this thesis is primarily concerned. In the next chapter I will summarize some basic numerical methods and attempt to illustrate their usefulness as well as their short comings with some numerical experiments. General properties of these methods will be described and commented upon. In chapter 3 I will describe the more general Runge-Kutta methods; how they are derived and the different forms they may take, i.e. implicit or explicit Runge-Kutta schemes. In chapter 4 we enter the great jungle of trees and use the concepts of rooted trees and B-series, pioneered by J. C. Butcher in the years 1963-72, to analyze Runge-Kutta numerical integrators. In this chapter I will develop and extend the concepts of rooted trees and B-series, as in [10], and show how they are used to analyze Rung-Kutta methods. This will involve studying properties of B-series and compositions of B-series which allow us to consider numerical integrators that are conjugate to other numerical integrators in chapter 6. Chapter 5 deals with Backward error analysis and uses the developed tools of B-series and rooted trees to analyze how well a numerical Integrator preserves the qualitative behavior of an ODE. In particular, this chapter will study the conditions under which a B-series (and hence its corresponding numerical integrator) is Symplectic or Energy-preserving. Examples are given to illustrate these conditions. In chapter 6 the conditions for a B-series to be Symplectic or Energy-preserving are studied purely in terms of rooted trees. That is, certain linear combinations of rooted trees determine whether a B-series is Symplectic and certain linear combinations determine whether a B-series preserves the energy of a system. These linear combinations form vector subspaces of the vector space of rooted trees and it is the goal of this chapter to understand the algebraic structure of these vector spaces. Section 6.1 gives an overview of the vector spaces of Energy-preserving trees and Hamiltonian trees along with some basic results (already given in [5]) which are consequences of the study in chapter 5 and the results given in [14]. Section 6.2 extends the ideas of the previous section to conjugate B-series and gives an overview of the results given in [5] along with some new results on the space of Energy-preserving and Conjugate-to-Hamiltonian trees ($\mathcal{T}_H \cap \mathcal{T}_{\bar{\Omega}}$) and the maps involved in the construction of this space. Lemmas 56 and 57 and Theorems 59 and 60 of Section 6.2 are original to this thesis. Section 6.3 gives two constructions of the annihilator of the conjugate-to-Energy preserving space of trees, one following directly from the results given in chapter 5 (which were originally given in [9]), and the other, Theorem 63, a new efficient algorithm based only on the structure of the Energy-preserving space of trees and the annihilator of the Hamiltonian trees. The result of section 6.3 is original to this thesis. Finally, in section 6.4 I extend the properties of the maps introduced in section 6.2 and use these maps to study the relationships and links between the different vector-spaces of rooted trees which was due to an idea of Elena Celledoni. These relationships are interpreted and commented upon. The results of this section are all original to this thesis.

Chapter 2

Basic Numerical Methods

As mentioned before, many of the differential equations that arise in the study of physical phenomena cannot be solved exactly. Therefore, we need other methods to describe solutions or approximate them. So-called *qualitative* methods which are used to describe the nature and behavior of solutions may be employed but they shan't be discussed here, there is a wealth of literature describing these methods in detail. Rather, we shall focus on basic numerical methods for approximating solutions and study their accuracy.

2.1 The Lipschitz Condition

Our main goal is to approximate the solution of an ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ (where $\dot{\mathbf{x}}$ is as defined in section 1.1) with some initial data $(t_0, \mathbf{x}(t_0))$. Here $\mathbf{f}$ is a sufficiently well behaved function that maps $[t_0, \infty) \times \mathbb{R}^n$ to $\mathbb{R}^n$ and the initial data $\mathbf{x}(t_0) \in \mathbb{R}^n$ is a given vector. Now, a "well behaved" function could mean a whole range of things but at the very least we insist on $\mathbf{f}$ obeying, in a given vector norm $\|.\|$, the *Lipschitz Condition*. That is, $\mathbf{f}$ is Lipschitz in $\mathbf{x}$ (or Lipschitz continuous) on an open set $U \subseteq [t_0, \infty) \times \mathbb{R}^n$ if there exists a $K > 0$ such that

$$\|\mathbf{f}(\mathbf{x}, t) - \mathbf{f}(\mathbf{y}, t)\| \leq K \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Here $K > 0$ is a real constant that is independent of the choice of $\mathbf{x}$ and $\mathbf{y}$. It is called the Lipschitz constant. This condition implies that $\mathbf{f}$ is continuous with respect to $\mathbf{x}$, but not all continuous functions are Lipschitz (consider $f(x) = x^{\frac{1}{3}}$). If $\frac{\partial f_i}{\partial x_i}$ exists and is bounded in U, then $\mathbf{f}$ is Lipschitz with $K = \sup_{\mathbf{x} \in U} \left\| \frac{\partial f_i}{\partial x_i} \right\|$.

Subject to the Lipschitz condition it is possible to prove that the system of ODEs has a unique solution in the interval $(t_0 - \varepsilon, t_0 + \varepsilon)$, for some $\varepsilon > 0$. Thus, throughout this thesis we shall assume that the function $\mathbf{f}$ is Lipschitz and hence the system of differential equations will have a solution.

2.2 The Explicit Euler Method

Let us imagine we have a differential equation, $\frac{dx}{dt} = f(t,x)$, which cannot be solved analytically. Since by definition $\frac{dx}{dt} = \lim_{h\to 0} \frac{x(t+h)-x(t)}{h}$, the simplest and most obvious approach to solving $\frac{dx}{dt} = f(t,x)$ is to approximate it by

$$\frac{x(t+h)-x(t)}{h} \approx f(t,x)$$

where h is a small but non-zero step size.Thus, given some initial data $(t_0, x(t_0))$, we are able to approximate $x(t_0+h)$ by $x(t_0) + hf(t_0, x(t_0))$:

$$x(t_0+h) \approx y(t_0) + hf(t_0, x(t_0)).$$

Using this estimate for $x(t_0+h)$, we can go on to estimate $x(t_0+2h)$, $x(t_0+3h)$, $x(t_0+4h)$, etc. Letting $t_n = t_0 + nh$, $x_0 = x(t_0)$, where h is the time-step, and x_n is the numerical estimate of the exact solution $x(t_n), n = 0, 1, ...$ we obtain the iterative scheme

$$x_{n+1} = x_n + hf(t_n, x_n).$$

This method of approximating solution curves is called Euler's Method. Of course this scheme easily generalizes to systems of ordinary differential equations

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\mathbf{f}(t_n, \mathbf{x}_n).$$

Euler's method is by far the simplest of all iterative schemes but also of profound importance. It provides the foundation for the numerical analysis of differential equations and their solutions. In fact, the more complicated and involved methods such as Runge Kutta methods are simply generalizations of the concepts provided by the Euler Method.

As an example, consider the differential equation $\frac{dx}{dt} = 2t(1+x^2)$, with $x(0) = 0$, which has the solution $x(t) = tan(t^2)$. Euler's method was used to approximate the solution with a step size of $h = 0.1$, $h = 0.05$, and $h = 0.01$ and the graphs of these approximations superimposed on the actual solution are shown in figure 2.1.

The initial condition being, by definition, exact, so is our approximation at t_0. As we step further and further away from our initial condition, our approximation deviates further and further from the exact solution of our differential equation. However, the goal of a numerical solution is not to avoid errors altogether, after all, the reason we use numerical methods is because we do not know the exact solution! In every numerical method we will always incur errors but our goal is to understand why these errors occur and ensure that they do not grow and accumulate beyond a

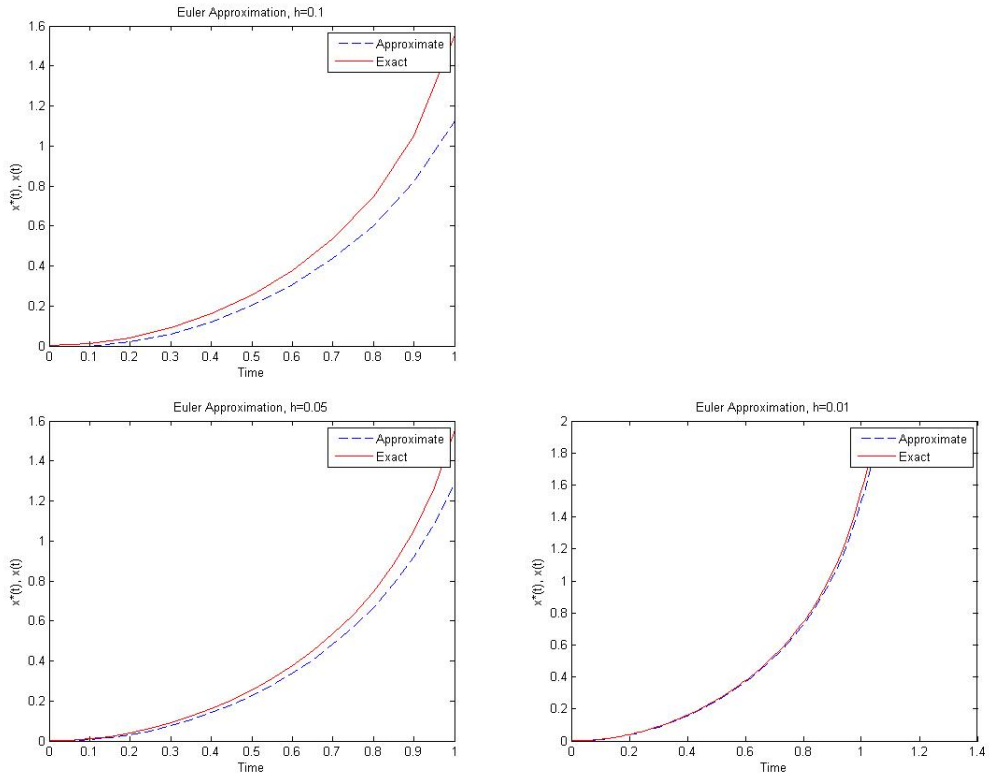


Figure 2.1: Approximate solution curves to $\frac{dx}{dt} = 2t(1 + x^2)$ using Euler's Method with step sizes of $h = 0.1$, $h = 0.05$, $h = 0.01$.

reasonable level. Thus, the real question of interest is how well does Euler's Method approximate solutions of ODEs? Figure 2.1 suggests that as the step size h becomes smaller and smaller, our approximate solution approaches the actual solution. Let us formulate this idea with a little more formalism.

Suppose we want to approximate the solution of an ODE on the interval $[t_0, t_0 + \epsilon]$ with some numerical method involving time steps. It doesn't even need to be Euler's method! We now cover this interval by an equidistant grid (in the case of Euler's method the spacing of the grid is determined by the time step h) and use our numerical method to approximate a solution to the given ODE. The question is whether, as the time step becomes smaller and smaller (and hence the grid spacing becomes smaller and smaller), the numerical solution tends to the exact solution of the ODE. Letting $\mathbf{x}_n = \mathbf{x}_{n,h}$, where h is the time step in the numerical method, for $n = 0, 1, ... \lfloor \frac{\epsilon}{h} \rfloor$, we say a numerical method is *convergent* if, for every ODE with a Lipschitz function $\mathbf{f}$ and every $\epsilon > 0$, it is true that

$$\lim_{h \to 0} \max_{n=0,1,... \lfloor \frac{\epsilon}{h} \rfloor} \|\mathbf{x}_{n,h} - \mathbf{x}(t_n)\| = 0$$

where $\lfloor a \rfloor$ is the integer part of $a \in \mathbb{R}$. That is, for every Lipschitz function, the numerical solution approaches the actual solution of an ODE as the grid becomes finer and finer.

It should be noted that convergence is a very necessary attribute of any numerical method. After all, what good is a numerical method if it is not guaranteed to approach the actual solution. There are a number of attributes a good numerical method should have, which will be discussed in the following sections, but convergence is the main ingredient for a useful numerical method. The following theorem confirms our previous suspicions that Euler's Method is indeed convergent. The proof is given by Arieh Iserles and assumes $\mathbf{f}$ is analytic, although it is enough to assume that $\mathbf{f}$ is only continuously differentiable. We follow that proof here.

Theorem 1. *Euler's Method is convergent.*

Proof. Let $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ be an ODE where $\mathbf{f}$ is analytic. Given $h > 0$ and $\mathbf{x}_n = \mathbf{x}_{n,h}$, $n = 0, 1, ..., \lfloor \frac{\epsilon}{h} \rfloor$, let $\mathbf{e}_{n,h} = \mathbf{x}_{n,h} - \mathbf{x}(t_n)$ be the numerical error. We wish to prove $\lim_{h \to 0} \max_{n=0,1,... \lfloor \frac{\epsilon}{h} \rfloor} \|\mathbf{e}_{n,h}\| = 0$. By Taylor's theorem,

$$\mathbf{x}(t_{n+1}) = \mathbf{x}(t_n + h) = \mathbf{x}(t_n) + h\dot{\mathbf{x}}(t_n) + \mathcal{O}(h^2) = \mathbf{x}(t_n) + h\mathbf{f}(t_n, \mathbf{x}(t_n)) + \mathcal{O}(h^2),$$

and, $\mathbf{x}$ being continuously differentiable, the $\mathcal{O}(h^2)$ term can be bounded (in a given norm) uniformly for all $h > 0$ and $n \leq \lfloor \frac{\epsilon}{h} \rfloor$ by a term of the form ch^2, where $c > 0$ is a constant. Subtracting the above equation from Euler's Method, we obtain

$$\mathbf{e}_{n+1,h} = \mathbf{e}_{n,h} + h[\mathbf{f}(t_n, \mathbf{x}(t_n) + \mathbf{e}_{n,h}) - \mathbf{f}(t_n, \mathbf{x}(t_n))] + \mathcal{O}(h^2).$$

By the triangle inequality, the Lipschitz condition, and the bound on $\mathcal{O}(h^2)$ that

$$\|\mathbf{e}_{n+1,h}\| \leq \|\mathbf{e}_{n,h}\| + h\,\|\mathbf{f}(t_n, \mathbf{x}(t_n) + \mathbf{e}_{n,h}) - \mathbf{f}(t_n, \mathbf{x}(t_n))\| + ch^2$$

$$\leq (1 + hK)\,\|\mathbf{e}_{n,h}\| + ch^2, \quad n = 0, 1, ..., \left\lfloor \frac{\epsilon}{h} \right\rfloor - 1.$$

The remainder of the proof is done by induction on n. We claim that $\|\mathbf{e}_{n,h}\| \leq \frac{c}{K}h[(1 + hK)^n - 1], \quad n = 0, 1, ...$ It is certainly true that $\|\mathbf{e}_{0,h}\| \leq 0$ since our initial condition is, by definition, exact. Assume the claim is true up to n. Then

$$\|\mathbf{e}_{n+1,h}\| \leq (1 + hK)\,\|\mathbf{e}_{n,h}\| + ch^2 = (1 + hK)\frac{c}{K}h[(1 + hK)^n - 1] + ch^2 = \frac{c}{K}h[(1 + hK)^{n+1} - 1]$$

and the claim is proved. The constant hK is positive and therefore $1 + hK < e^{hK}$, and we deduce that $(1+hK)^n < e^{nhK}$. The index n may range from 0 up to $\lfloor \frac{\epsilon}{h} \rfloor$ and hence $(1+hK)^n < e^{\lfloor \frac{\epsilon}{h} \rfloor hK} \leq e^{\epsilon K}$. Thus we obtain the inequality

$$\|\mathbf{e}_{n,h}\| \leq \frac{c}{K}(e^{\epsilon K} - 1)h, \quad n = 0, 1, ... \left\lfloor \frac{\epsilon}{h} \right\rfloor.$$

Since $\frac{c}{K}(e^{\epsilon K} - 1)$ is independent of h, it is clear that

$$\lim_{h \to 0, 0 \leq nh \leq \epsilon} \|\mathbf{e}_{n,h}\| = 0.$$

□

We have in fact done more than just show that Euler's Method is convergent, we have also found an upper bound on the error. It is always true that the error is bounded above by $\frac{c}{K}(e^{\epsilon K} - 1)h$, but this bound has very little practical value. The problem is not in our proof but in the unresponsiveness of the Lipschitz constant. An obvious example is the ODE $\dot{x} = -50x$, $x(0) = 1$. The Lipschitz constant is $K = 50$, and with $x(t) = e^{-50t}$, $c = K^2$ we derive the upper bound of $50h(e^{50\epsilon} - 1)$.With $\epsilon = 1$ we have

$$|x_n - x(nh)| \leq 2.59 \times 10^{23} h.$$

But we can show that $x_n = (1 - 100h)^n$ and obtain the exact expression

$$|x_n - x(nh)| = \left|(1 - 100h)^n - e^{-50nh}\right|$$

which is much smaller by several orders of magnitude! Therefore, the bound provided by the proof does not offer any useful information about the accumulated errors in using Euler's Method. We are able to obtain a more useful bound on the local error (the error incurred after each step, or

local truncation error). That is, if we assume $\mathbf{x}_n$ is exact, then $\mathbf{x}_{n+1}$ will contain a local truncation error. We use Taylor's theorem with remainder. If a function $\mathbf{x}(t)$ has $k+1$ derivatives that are continuous on an open interval containing t_n and $t_n + h$, then, letting $t_{n+1} = t_n + h$ and expanding $\mathbf{x}(t_{n+1})$in a Taylor series we obtain

$$\mathbf{x}(t_{n+1}) = \mathbf{x}(t_n) + h\dot{\mathbf{x}}(t_n) + \frac{h^2}{2}\ddot{\mathbf{x}}(t_n) + ... + \frac{h^{k+1}}{(k+1)!}\mathbf{x}^{(k+1)}(c)$$

where c is a point in the open interval. With $k = 1$ we obtain

$$\mathbf{x}(t_{n+1}) = \mathbf{x}(t_n) + h\mathbf{f}(t_n, \mathbf{x}_n) + \frac{h^2}{2}\ddot{\mathbf{x}}(c)$$

or

$$\mathbf{x}(t_{n+1}) = \mathbf{x}_n + h\mathbf{f}(t_n, \mathbf{x}_n) + \frac{h^2}{2}\ddot{\mathbf{x}}(c)$$

since $\mathbf{x}_n$ is exact. The first two terms are simply Euler's Method, thus we have

$$\mathbf{x}(t_{n+1}) = \mathbf{x}_{n+1} + \frac{h^2}{2}\ddot{\mathbf{x}}(c)$$

and the local truncation error in $\mathbf{x}_{n+1}$ is

$$\frac{h^2}{2}\ddot{\mathbf{x}}(c), \ \textit{where} \quad t_n < c < t_{n+1}.$$

The value of c is usually unknown and so the *exact* error cannot be calculated, but an upper bound on the error can be calculated. This upper bound is

$$\max_{t_n < t^* < t_{n+1}} |\dot{\mathbf{x}}(t^*)| \, \frac{h^2}{2}$$

Euler's method is of *order* one. If we write Euler's method in the form $\mathbf{x}_{n+1} - [\mathbf{x}_n - h\mathbf{f}(t_n, \mathbf{x}_n)] = 0$, replace $\mathbf{x}_i$ by the exact solution $\mathbf{x}(t_j)$, $j = n, n+1$ and expand into a Taylor series about $t_j = t_0 + nh$, we obtain

$$\mathbf{x}(t_{n+1}) - [\mathbf{x}(t_n) - h\mathbf{f}(t_n, \mathbf{x}(t_n))] = [\mathbf{x}(t_n) + h\dot{\mathbf{x}}(t_n) + \mathcal{O}(h^2)] - [\mathbf{x}(t_n) + h\dot{\mathbf{x}}(t_n)] = \mathcal{O}(h^2).$$

If we are given an arbitrary time-stepping method

$$\mathbf{x}_{n+1} = \mathcal{X}_n(\mathbf{f}, h, \mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_n), \quad n = 0, 1, ...$$

for an ODE, we say it is of *order* k if

$$\mathbf{x}(t_{n+1}) - \mathcal{X}_n(\mathbf{f}, h, \mathbf{x}(t_0), \mathbf{x}(t_1), ..., \mathbf{x}(t_n)) = \mathcal{O}(h^{k+1})$$

for every analytic function $\mathbf{f}$ and $n = 0, 1, ...$ The order of a numerical method measures the change in error of a numerical solution as the step size is decreased, but the information provided is only about the method's local behavior - moving from t_n to t_{n+1}. That is, as we move from t_n to t_{n+1}, for sufficiently small $h > 0$, we are incurring an error of $\mathcal{O}(h^2)$. As before, define $\mathbf{e}_{n,h}$ to be the total error in the numerical calculation after completing n steps. Then $\mathbf{e}_{n,h}$ is of order h^k, denoted by $\mathcal{O}(h^k)$, if there exists a constant C and a positive integer k such that $|\mathbf{e}_{n,h}| \leq Ch^k$ for sufficiently small h. In general, if $\mathbf{e}_{n,h}$ in a numerical method is of order h^k and h is halved, the new error is approximately $C(\frac{h}{2})^k = C\frac{h^k}{2^k}$. That is, the error is reduced by a factor of $\frac{1}{2^k}$. Theorem 1 shows us that the total error committed by Euler's method is proportional to h. This is a serious limitation because decreasing the value of h increases the computation time and produces only a marginal improvement in accuracy. We would prefer "higher order" methods for which the error in a step behaves as $\mathcal{O}(h^{p+1})$ and the total error after n steps behaves as $\mathcal{O}(h^p)$ for $p \geq 2$.

2.3 The Implicit Euler Method

In many practical problems Euler's Method is not particularly useful. One particular class of ODEs in which Euler's method (and other methods) is not useful is the class of "stiff" ODEs. A quick look at stability is needed first. As an example, consider the differential equation

$$x'' - 10x' - 11x = 0$$

with initial conditions $x(0) = 1, x'(0) = -1$. The solution of this ODE is $x(t) = e^{-t}$. Suppose now we change the initial conditions by a small amount $\varepsilon > 0$ so that the initial conditions are now $x(0) = 1 + \varepsilon, x'(0) = -1$. Now the solution to the ODE with the new initial conditions is

$$x(t) = (1 + \frac{11}{12}\varepsilon)e^{-t} + \frac{\varepsilon}{12}e^{11t}.$$

Therefore, no matter how small $\varepsilon > 0$ is, the second term in the new solution causes the solution to approach infinity as t approaches infinity. The solution $x(t) = e^{-t}$ of the original ODE is *unstable*. That is, arbitrarily small changes in the initial conditions produce arbitrarily large changes in the solution as $t \to \infty$. It is extremely difficult to calculate the solutions of these types of ODEs numerically since round off and truncation error have the same effect as changing the initial conditions which causes the solutions to diverge. ODEs that are said to be "stiff" are

ODEs that exhibit unstable behavior if the time step in the numerical approximation is too large. Consider the ODE

$$x' = -100x + 100, \quad x(0) = x_0$$

The exact solution is given by

$$x(t) = (x_0 - 1)e^{-100t} + 1.$$

If we change the initial conditions from $x(0) = x_0$ to $x(0) = x_0 + \varepsilon$ then the solution becomes

$$x(t) = (x_0 + \varepsilon - 1)e^{-100t} + 1$$

and we see that the solution is stable. Let us apply Euler's method to the problem, then we get

$$x_{n+1} = x_n + h(-100x_n + 100) = (1 - 100h)x_n + 100h.$$

We may solve this equation recursively to obtain

$$x_n = (x_0 - 1)(1 - 100h)^n + 1.$$

Suppose that $x_0 = 2$. Then the original solution to the ODE becomes

$$x(t) = e^{-100t} + 1$$

and the recursion equation becomes

$$x_n = (1 - 100h)^n + 1.$$

Now, $x(t)$ decreases very rapidly from its initial condition to its limiting value of 1. Therefore, we would expect to require a small step size h to compute the solution accurately. However, for $t > 0.1$, the solution varies very slowly and is essentially equal to 1 with almost negligible variation, thus we should expect to be able to obtain sufficient accuracy with Euler's method using a relatively large value for h. But, we also see from the recursion equation that if $h > 0.02$, then $|1 - 100h| > 1$ and the approximation $x_n$grows rapidly with each step and shows unstable behavior. The quantity $(1 - 100h)^n$ is an approximation to e^{-100t}, and we know from Taylor's theorem that it a good approximation for small h but rapidly becomes a rubbish approximation as h becomes as large as 0.02. Even though the exponential term contributes nearly nothing to the solution for small t, Euler's method requires us to calculate the numerical approximation sufficiently accurately to ensure the stability of the solution. This is often ineffective and expensive in terms of computing time. The example illustrates the essence of the problem of stiffness.

The general approach to the problem of stiffness is to use implicit methods. For the ODE $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$, the method

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\mathbf{f}(t_{n+1}, \mathbf{x}_{n+1})$$

is known as the *backward Euler method*. It has the same form as Euler's method except that $\mathbf{f}$ is evaluated at $(t_{n+1}, \mathbf{x}_{n+1})$ rather than at $(t_n, \mathbf{x}_n)$; hence the method is *implicit*. The backward Euler method is implicit which means we need to solve an, often non-linear, equation for x_{n+1}. As an example let $\frac{dx}{dt} = xcosx$. To implement the backward Euler method we will need to solve the non-linear equation $x_{n+1} - x_{n+1}cos(x_{n+1}) = x_n$ for x_{n+1} at any given time-step. A suitable root finding technique such as the Newton-Raphson method can be used for this purpose. This is evidently much more computationally expensive than Euler's method but we use implicit methods because they are stable. If we apply the backward Euler Method to the ODE $x' = -100x + 100$, $x(0) = x_0$, we obtain

$$x_{n+1} = x_n + h(-100x_{n+1} + 100)$$

which is easily rearranged to give

$$x_{n+1} = \frac{(x_n + 100h)}{(1 + 100h)}.$$

The actual solution of this recursive equation is

$$x_n = (x_0 - 1)(1 + 100h)^{-n} + 1.$$

For the initial condition $x(0) = 2$, the solution becomes $x_n = (1 + 100h)^{-n} + 1$ and we see there is no unstable behavior regardless of the magnitude of h. The backward Euler method is based on the Taylor series $x_n \approx x(t_{n+1} - h) = x(t_{n+1}) - hx'(t_{n+1}) + \mathcal{O}(h^2)$ and a similar argument used in section 2.2 shows that the backward Euler method is of order 1. Moreover, the backward Euler method is convergent.

Theorem 2. *The backward Euler method is convergent.*

I shall omit the proof here as it is almost identical to the proof of the convergence of Euler's method. We also deduce that the local and global errors in the backward Euler method behave in much the same way as the local and global errors of Euler's method.

2.4 The Implicit Mid-Point Rule

Taking the mean of $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ in the argument of $\mathbf{f}$, we obtain the *implicit midpoint rule*

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\mathbf{f}(t_n + \frac{h}{2}, \frac{\mathbf{x}_n + \mathbf{x}_{n+1}}{2}).$$

The approximation $\mathbf{x}_{n+1}$ is obtained implicitly from evaluating $\mathbf{f}$ at the midpoint of $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ and the midpoint of t_n and t_{n+1}. This acts as a smoothing mechanism as we are approximating the ODE, i.e. the slope of the solution, at the end points $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ and the midpoint of $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$. Similarly for the time step endpoints t_n and t_{n+1} and their midpoint. The implicit midpoint rule is similar to the backward Euler method in that we need to solve a system of equations in order to determine $\mathbf{x}_{n+1}$.

To obtain the order of the implicit midpoint rule we substitute the exact solution,

$$\mathbf{x}(t_{n+1}) - [\mathbf{x}(t_n) + h\mathbf{f}(t_n + \frac{h}{2}, \frac{\mathbf{x}(t_n) + \mathbf{x}(t_{n+1})}{2}]$$

$$= [\mathbf{x}(t_n) + h\mathbf{x}'(t_n) + \frac{h^2}{2}\mathbf{x}''(t_n) + \mathcal{O}(h^3)] - [\mathbf{x}(t_n) + h(\mathbf{x}'(t_n) + \frac{h}{2}\mathbf{x}''(t_n) + \mathcal{O}(h^2))] = \mathcal{O}(h^3).$$

Therefore, the implicit midpoint rule is of order 2. A simple generalization of Theorem 1 will tell us that the implicit midpoint rule is a convergent method and that the global error behaves as $\mathcal{O}(h^2)$.

Theorem 3. *The implicit midpoint rule is convergent with error bound* $\|\mathbf{e}_{n,h}\| \leq \frac{c}{K} e^{(\frac{\epsilon K}{1-\frac{hK}{2}})} h^2$.

The bound obtained is, once again, of no practical value. However, what is established is that the global error of the implicit midpoint rule behaves as $\mathcal{O}(h^2)$, a marked improvement on the Euler method and the backward Euler method. This is to be expected of a 2nd order method whose convergence has been established. The implicit midpoint rule is a stable method (although we have not demonstrated this) and has an acceptable accuracy; therefore, it can be implemented using a larger step size, thus saving on computation time. However, this computational saving is offset by the need to solve a system of equations to determine $\mathbf{x}_{n+1}$. With all numerical methods there is a trade off.

2.5 The Symplectic Euler Method

Suppose that we have a system of ODEs

$$\dot{x} = f(x, y)$$

$$\dot{y} = g(x, y).$$

We may consider what is called a *partitioned* Euler method

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

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

which is a combination of the explicit Euler method and the backward Euler method. The x-variable is treated by the implicit Euler method and the y-variable is treated by the explicit Euler method. To illustrate the convergence of the method, let $z_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ so that the method becomes

$$z_{n+1} = z_n + h\mathbf{M}(x_{n+1}, y_{n+1}),$$

where $\mathbf{M}(x_{n+1}, y_{n+1}) = \begin{pmatrix} f(x_{n+1}, y_n) \\ g(x_{n+1}, y_{n+1}) \end{pmatrix}$. Letting $n = 0$ (for illustrative purposes) we have $z(t_0 + h) = z(t_0) + hz'(t_0) + \mathcal{O}(h^2)$ so that $\mathbf{e}_{1,h} = z_1 - z(t_0 + h) = h(\mathbf{M}(x_1, y_0) - \mathbf{M}(x_1, y_0)) + \mathcal{O}(h^2)$. By the triangle inequality, Lipschitz condition and big oh notation we have that $\|\mathbf{e}_{1,h}\| \leq hK \|\mathbf{e}_{1,h}(x)\| + ch^2$, where $\|\mathbf{e}_{1,h}(x)\|$ is the error in the x variable. Following an identical procedure we have $\|\mathbf{e}_{1,h}(x)\| \leq \frac{h^2 K}{(1-hK)}$. Since we are considering the limit as h tends to zero we may assume the $h < \frac{1}{k}$ so that the inequality is positive. It is clear that $\|\mathbf{e}_{1,h}\|$ tends to zero as h tends to zero. As in the proof of Theorem 1, we may continue the procedure inductively and find that $\|\mathbf{e}_{n,h}\| \to 0$ as h tends to zero.

The local and global errors behave the same way as the local and global errors of the explicit and implicit Euler methods. That is, the partitioned Euler method is a first order method whose local errors behave as $\mathcal{O}(h^2)$ and whose global errors behave as $\mathcal{O}(h)$. We call it the *Symplectic Euler Method*.

2.6 Numerical Experiments

Using the numerical methods discussed, we shall perform experiments on a few standard problems. The accuracy and qualitative features of the methods will be compared and evaluated.

2.6.1 The Lotka-Volterra Model

The Lotka-Volterra model describes the interaction between a predator species and a prey species. Let $x(t)$ and $y(t)$ denote the predator and the prey populations, respectively, at time t; then, a plausible model of interaction between the two species is given by

$$\dot{x} = x(\alpha y - \beta)$$

$$\dot{y} = y(\gamma - \delta x),$$

where α, β, γ, and δ are positive constants. For concreteness, suppose $\alpha = \gamma = \delta = 1$ and $\beta = 2$. We solve the system by first obtaining an expression for $\frac{dy}{dx}$, then integrating the expression using separation of variable.

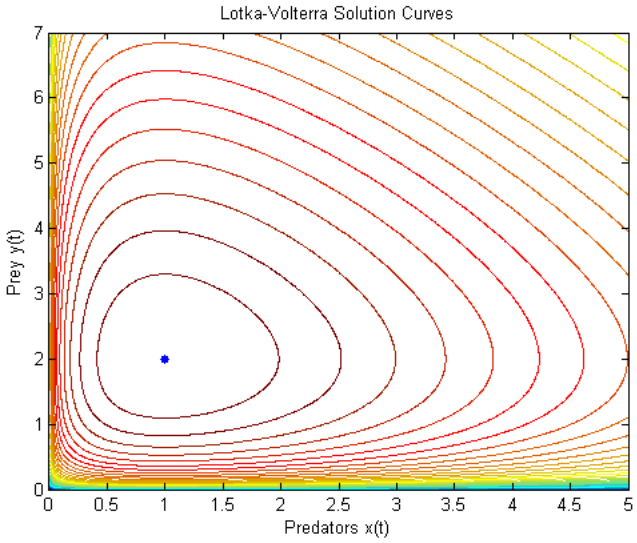


Figure 2.2: Level curves of the Lotka-Volterra Equation.

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{y(1-x)}{x(y-2)}$$

$$\int \frac{y-2}{y} dy = \int \frac{1-x}{x} dx$$

$$y - 2ln(y) = ln(x) - x + C$$

$$ln(x) - x + 2ln(y) - y = C$$

where C is a constant of integration. This equation holds for all t and every solution of the system lies on a level curve of $ln(x) - x + 2ln(y) - y = C$. Some of these level curves are drawn in figure 2.2. Since the level curves are closed, all solutions of the system are periodic.

Figure 2.3 displays numerical approximations of the Lotka-Volterra solution curves using the Euler method, backward Euler method, implicit midpoint rule and the symplectic Euler method. Only the implicit midpoint rule and symplectic Euler method have the correct qualitative behavior. Both the Euler method and backward Euler method solutions spiral outwardly. The Lotka-Volterra model is not a stiff system; every solution curve is periodic and therefore stable. This tells us that the problems in using the Euler method and backward Euler method are not problems with stability but rather problems with the preservation of some geometric or physical property.

The implicit midpoint rule is more accurate than the symplectic Euler method; this is to be expected of a second order method.

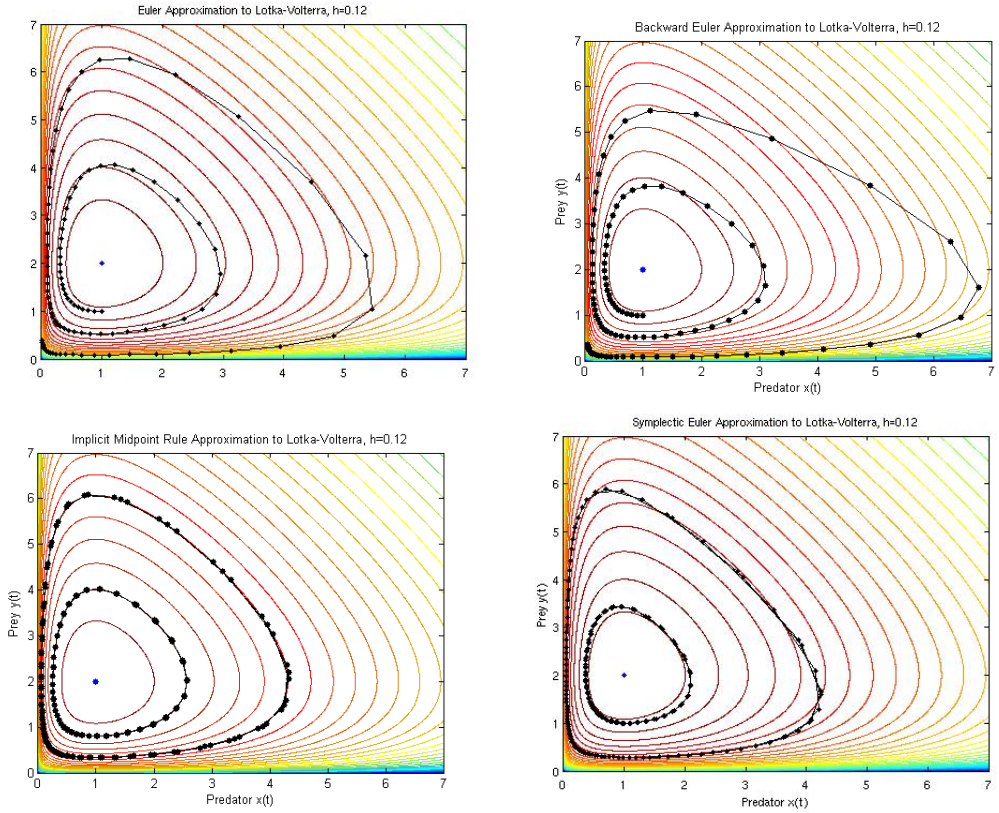


Figure 2.3: Solutions of the Lotka-Volterra equations with step size h=0.12; initial conditions (,) for the Euler method, (1,1) for the backward Euler method, (2,1) for the implicit midpoint rule and (,) for the symplectic Euler method.

2.6.2 The Pendulum

Hamiltonian problems are a very important class of problems, particularly in analytic mechanics and many other problems of motion. These problems are of the form

$$\dot{p} = -H_q(p,q) \quad \dot{q} = H_p = (p,q),$$

where the *Hamiltonian* $H(p_1, p_2, ..., p_d, q_1, q_2, ..., q_d)$ represents the total energy; q_i are the position coordinates and p_i the momenta for $i = 1, 2, ..., d$, with the d the number of degrees of freedom; H_p and H_q are the vectors of partial derivatives. The total energy is always conserved. That is, along solution curves of the Hamiltonian system,

$$H(p(t), q(t)) = Const.$$

$H_p = \nabla_p H = (\frac{\partial H}{\partial p})^T$ and $H_q = \nabla_q H = (\frac{\partial H}{\partial q})^T$ are the column vectors of partial derivatives. Differentiating the Hamiltonian along solution curves $(p(t), q(t))$,

$$\frac{dH}{dt} = \frac{\partial H}{\partial p}.\frac{dp}{dt} + \frac{\partial H}{\partial q}.\frac{dq}{dt} = \frac{\partial H}{\partial p}.(-\frac{\partial H}{\partial q})^T + \frac{\partial H}{\partial q}.(\frac{\partial H}{\partial p})^T = 0.$$

The Hamiltonian is said to be a *first integral* of the system.

The mathematics of a pendulum are quite complicated, but we can simplify the situation with a few assumptions in order to solve the equations of motion analytically. A simple pendulum works on the following assumption: the rod on which the bob swings is mass-less, inextensible, remains taut and is of length 1, motion occurs in a 2-dimensional plane, the joint on which the pivots is frictionless and air resistance is non-existent. The simple pendulum is a Hamiltonian system with one degree of freedom having the Hamiltonian

$$H(p,q) = \frac{1}{2}p^2 - \cos q$$

so that the equations of motion become

$$\dot{p} = -\sin q \quad \dot{q} = p.$$

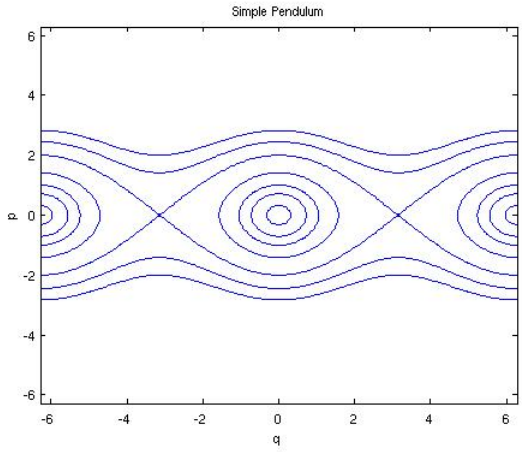


Figure 2.4: Level Curves of the Pendulum

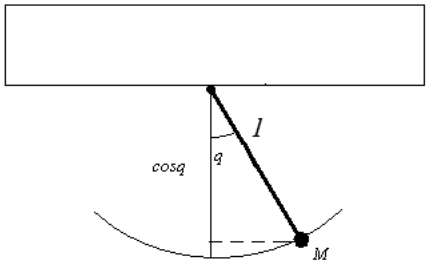


Figure 2.4 displays the solution curves of the Pendulum. As was the case for the Lotka-Volterra model, $H(p(t), q(t)) = Const.$ for all time t. All the level curves are periodic, although this may not be obvious at first. The variable q is 2π periodic and thus natural to consider as a variable on the circle S^1. Hence the true phase space of points (p, q) becomes the cylinder $\mathbb{R} \times S^1$. If we identify the points $q = -\pi$ and $q = \pi$ and glue the lines $q = -\pi$ and $q = \pi$ together, we have created the cylinder. Now the curves that appear not to close up, drawn in figure 2.4, closes up on the cylinder. These curves tell us that if we give the pendulum enough energy it will continue to rotate 360 degrees on its pivot forever.

We apply the Euler method, Backward Euler method, Implicit Midpoint Rule and Symplectic Euler method to the pendulum. Figure shows the results of this experiment. As in the case of the Lotka-Volterra model we observe that the Euler method and backward Euler method display the wrong qualitative behavior; the Euler method spirals outward and the backward Euler method spirals inward. These methods are not conserving the energy in the system. The implicit midpoint rule and the symplectic Euler method both display the correct qualitative behavior, thus conserving the energy in the system. As before, the implicit midpoint rule is more accurate than the symplectic Euler method.

The graphs in figure 2.6 display how each numerical method affects the energy in the system. The

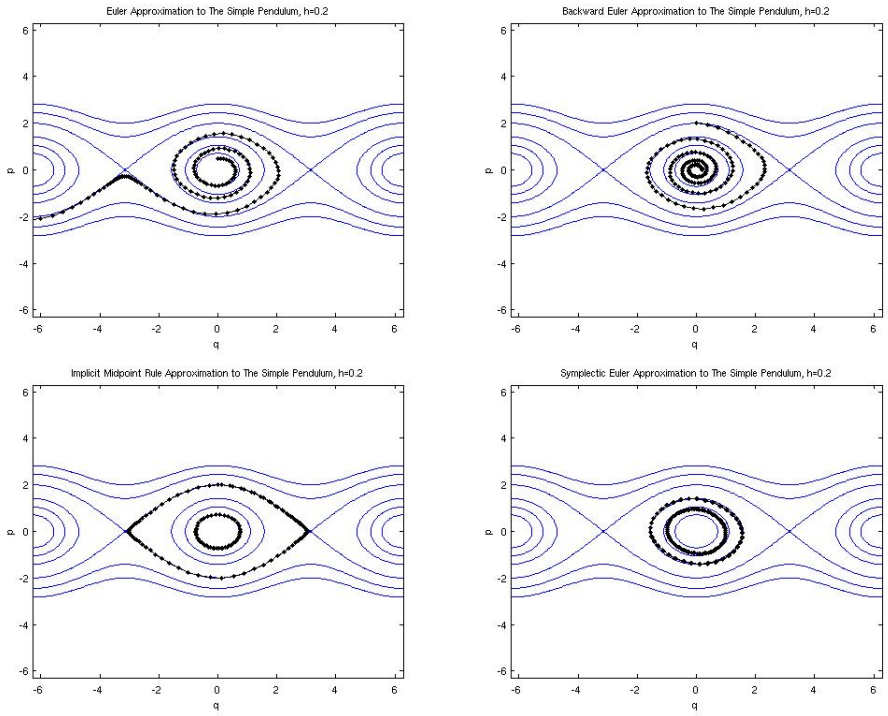


Figure 2.5: Solutions of the Pendulum equations; step sizes h=0.2.

symplectic Euler method and the Implicit Midpoint rule both show long time energy conservation. The implicit midpoint rule still outperforms the symplectic Euler method as there are less deviations from the actual energy in the system. The main observation is that in both the Symplectic Euler method and Implicit midpoint rule, the error in the total energy is small and bounded. The error in the energy of the Euler method grows linearly with time, whilst the energy in using the backward Euler method decays exponentially from its initial value.

We have learned from these experiments that we need to use methods that are of a high order and preserve some physical or geometric property of the system. The numerical methods that achieve this are called Geometric numerical Integrators. They are of great importance when modeling any physical system because they preserve the important physical features of the system that classical numerical methods, such as Euler's method, cannot.

2.7 Symplectic Transformations and Symplectic Integrators

Before advancing to the next chapter it will be worthwhile briefly defining Symplectic transformations and Symplectic Integrators which will be mentioned in Chapter 5.

The first property of Hamiltonian systems is that the Hamiltonian $H(p, q)$ is a *first integral* of the system. A second property is the *symplecticity* of the flow map. We first have the following definition from Hairer, Lubich and Wanner (2002).

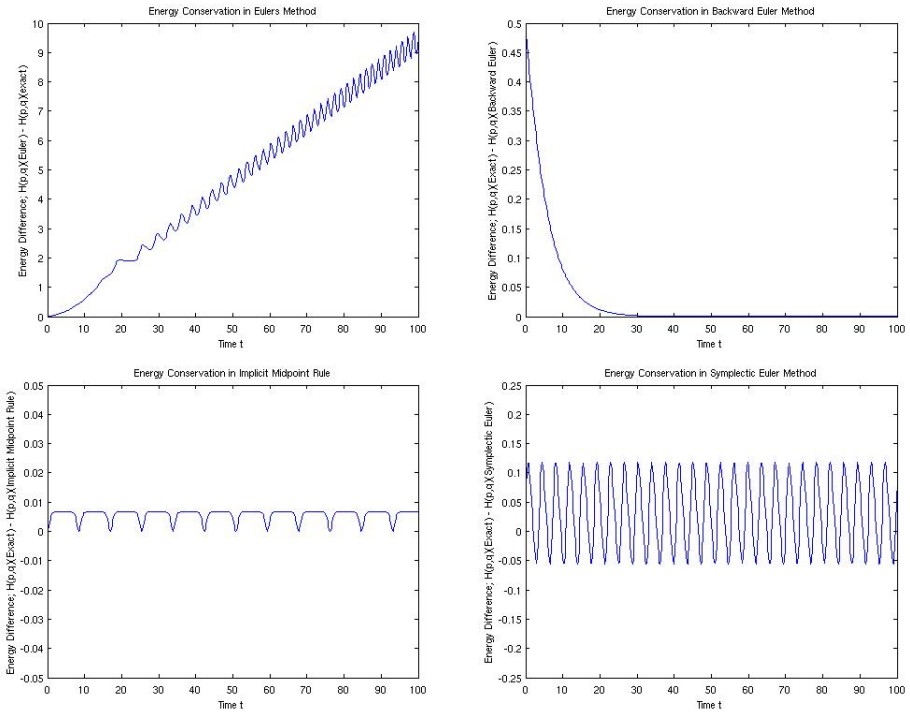


Figure 2.6: Energy Conservation of the numerical methods applied to the Pendulum system.

Definition 4. A linear mapping $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called *symplectic* if

$$A^T J A = J$$

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

A non-linear map $\Phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic if its linearization $\Phi'(x)$ is symplectic for all x.

In the case $d = 1$, symplecticity of a linear mapping A is equivalent to $area - preservation$. That is, if P is a parallelogram in $\mathbb{R}^2$ then $A(P)$ has the same area as P. In the general case ($d > 1$), symplecticity means that the sum of the oriented areas of the projections of P is the same as that for the transformed parallelogram $A(P)$.

Now we have the following Theorem due to Poincare.

Theorem 5. *(Poincare 1899). Let $H(p, q)$ be a twice continuously differentiable function on $U \subset \mathbb{R}^{2d}$. Then, for each fixed t, the flow φ_t is a symplectic transformation wherever it is defined.*

The natural thing to do is to consider numerical integrators that are symplectic.

Definition 6. A numerical one-step method is called *symplectic* if the one-step map $\mathbf{x}_1 = \Phi_h(\mathbf{x})$ is symplectic whenever the method is applied to a smooth Hamiltonian system.

Two examples of such integrators are the so-called *symplectic Euler method* of section 2.5 and

the *implicit midpoint rule* of section 2.4. For proofs of the symplecticity of these methods see Hairer, Lubich and Wanner, IV.3 (2002).

Chapter 3

Runge-Kutta Methods

3.1 Gaussian Quadrature

The exact solution of the differential equation

$$x' = f(t), \quad t \geq t_0, \quad x(t_0) = x_0,$$

is given by $x_0 + \int_{t_0}^{t} f(\tau)d\tau$. Often, the ODEs we wish to solve are of the form

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), \quad t \geq t_0, \quad \mathbf{x}(t_0) = \mathbf{x}_0.$$

Notice that the right hand side of the system is now a function of $\mathbf{x}$ and t. The integral that needs to be computed to solve the system is now decidedly non-trivial and we need to employ the techniques of computing integrals numerically. This is the logic behind *Runge Kutta* methods. Before discussing these methods it will be worthwhile paying some attention to the numerical calculation of integrals.

The standard practice in calculating integrals numerically is to replace the integral with a finite sum, a procedure known as *quadrature*. Specifically, let ω be a non-negative function acting in the interval (a, b), such that

$$0 < \int_a^b \omega(\tau)d\tau < \infty, \quad \left| \int_a^b \tau^j \omega(\tau)d\tau \right| < \infty, \ j = 1, 2, ...;$$

ω is called the *weight function*. We approximate as follows

$$\int_a^b f(\tau)\omega(\tau)d\tau \approx \sum_{j=1}^{\nu} b_j f(c_j),$$

where the numbers $b_1, b_2, ..., b_\nu$ and $c_1, c_2, ..., c_\nu$, which are independent of f (but generally depend upon ω, a and b), are called the quadrature *weights* and *nodes*, respectively. We do not require a and b to be bounded, although we must have $a < b$.

The obvious question to now ask is, how good is the approximation? It can be easily shown, using the Peano Kernel Theorem, that a quadrature formula is of order p if

$$\left| \int_a^b f(\tau)\omega(t)d\tau - \sum_{j=1}^{\nu} b_j f(c_j) \right| \leq c \max_{a \leq t \leq b} \left| f^{(p)}(t) \right|,$$

or any function f, and $c > 0$ a constant. We can even go one better with the following Lemma from Iserles's (1996).

Lemma 7. *Given any distinct set of nodes $c_1, c_2, ..., c_\nu$, it is possible to find a unique set of weights $b_1, b_2, ..., b_\nu$ such that the quadrature formula, $\int_a^b f(\tau)\omega(\tau)d\tau \approx \sum_{j=1}^{\nu} b_j f(c_j)$, is of order $p \geq \nu$.*

3.2 Explicit Runge-Kutta Schemes

We can now extend the quadrature formula to the ODE, $\mathbf{x}' = \mathbf{f}(t, \mathbf{x}), t \geq t_0, \mathbf{x}(t_0) = \mathbf{x}_0$, by integrating from t_n to $t_{n+1} = t_n + h$:

$$\mathbf{x}(t_{n+1}) = \mathbf{x}(t_n) + \int_{t_n}^{t_{n+1}} \mathbf{f}(\tau, \mathbf{x}(\tau))d\tau = \mathbf{x}(t_n) + h \int_0^1 \mathbf{f}(t_n + h\tau, \mathbf{x}(t_n + h\tau))d\tau,$$

and then replacing the second integral by a quadrature. The outcome of this is

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h \sum_{j=1}^{\nu} b_j \mathbf{f}(t_n + c_j h, \mathbf{x}(t_n + c_j h)), \quad n = 0, 1, ...$$

But there is one problem with this. We do not know the value of $\mathbf{x}$ at the nodes $t_n + c_1 h, t_n + c_2 h, ..., t_n + c_\nu h$. We need to resort to an approximation.

We denote our approximation of $\mathbf{x}(t_n + c_j h))$ by k_j, $j = 1, 2, ..., \nu$. We first let $c_1 = 0$, since then the approximation is already provided by the previous step of the numerical method, $k_1 = \mathbf{f}(t_n, k_1)$. The idea behind the *explicit Runge−Kutta* method is to express each $k_j, j = 2, 3, ..., \nu$, by updating $\mathbf{x}_n$ with a linear combination of $\mathbf{f}(t_n, k_1), \mathbf{f}(t_n + hc_2, k_2), ..., \mathbf{f}(t_n + c_{j-1}h, k_{j-1})$. More specifically,

$$k_1 = \mathbf{x}_n$$

$$k_2 = \mathbf{f}(t_n + c_2 h, \mathbf{x}_n + h a_{2,1} \mathbf{f}(t_n, k_1))$$

$$k_3 = \mathbf{f}(t_n + c_3 h, \mathbf{x}_n + h a_{3,1} \mathbf{f}(t_n, k_1) + h a_{3,2} \mathbf{f}(t_n + c_2 h, k_2))$$

$$\vdots$$

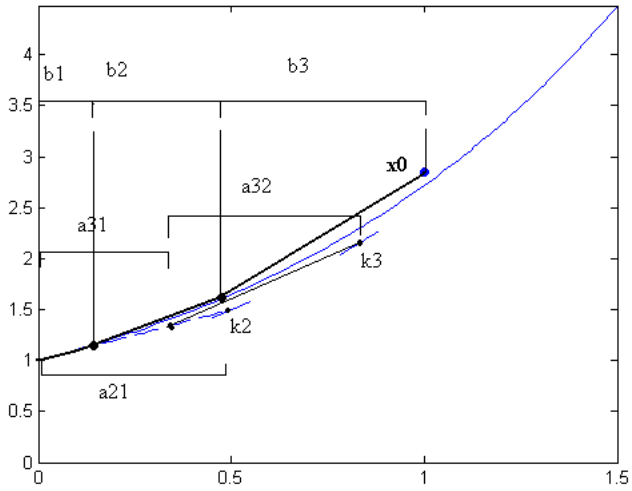


Figure 3.1: Geometric depiction of the Runge-Kutta method; h=1.

$$k_\nu = \mathbf{f}(t_n + c_\nu h, \mathbf{x}_n + h\sum_{i=1}^{\nu-1} a_{\nu,i}\mathbf{f}(t_n + c_i h, k_i))$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\sum_{j=1}^{\nu} b_j k_j.$$

The matrix $A = (a_{j,i})_{j,i=1,2,\ldots,\nu}$, where missing elements are defined to be zero, is called the *RK matrix*, while

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_\nu \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_\nu \end{bmatrix}$$

are the *RK weights* and *RK nodes* respectively. We say that the above method has ν *stages*.

This method has a nice geometric interpretation. We compute several polygonal lines, each starting at $\mathbf{x}_0$ and assuming the various slopes at the points k_j on portions of the integral interval, which are proportional to the given constants a_{ji}; at the final point of each polygon evaluate a new slope k_i. The last of these polygons, with constants b_i, determines the numerical solution $\mathbf{x}_1$. Figure 3.1 shows this. The first polygonal line extends from k_1 to k_2 in an interval proportional to $a_{2,1}$ (since we have taken $h = 1$) by an Euler step. The next polygon is a line with the same slope as at k_1, of length proportional to an amount $a_{3,1}$, followed by a line with the same slope as at k_2, of length proportional to an amount $a_{3,2}$, this gives us k_3. The final polygon is given by a line with the same slope as at k_1, of length proportional to an amount b_1, followed by a line with the same slope as at k_2, of length proportional to an amount b_2, with the same slope as at k_3, of length proportional to an amount b_3, thus giving us $\mathbf{x}_0$.

How should we choose the RK matrix, i.e. the $a_{j,i}$s. The most obvious way is expand the method into a Taylor series about $(t_n, \mathbf{x}_n)$, but the usefulness of this idea is very limited. As an example consider the simplest case, $\nu = 2$. Assuming the vector function is differentiable, we have

$$\mathbf{f}(t_n + c_2 h, k_2) = \mathbf{f}(t_n + c_2 h, \mathbf{x}_n + a_{2,1} h \mathbf{f}(t_n, \mathbf{x}_n))$$

$$= \mathbf{f}(t_n, \mathbf{x}_n) + h[c_2 \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial t} + a_{2,1} \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial t} \mathbf{f}(t_n, \mathbf{x}_n)] + \mathcal{O}(h^2);$$

therefore the final step in the Runge-Kutta method (the expression for $\mathbf{x}_{n+1}$) becomes

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h(b_1 + b_2)\mathbf{f}(t_n, \mathbf{x}_n) + h^2 b_2 [c_2 \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial t} + a_{2,1} \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial \mathbf{x}} \mathbf{f}(t_n, \mathbf{x}_n)] + \mathcal{O}(h^3).$$

We need to compare this expansion to the Taylor expansion of the exact solution about the same point $(t_n, \mathbf{x}_n)$.The first derivative is provided by the ODE and we can obtain $\mathbf{x}''$ from this expression by differentiating the ODE with respect to t;

$$\mathbf{x}'' = \frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial t} + \frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x}).$$

Denoting the exact solution at t_{n+1}, subject to the initial conditions $\mathbf{x}_n$ at t_n, by $\tilde{\mathbf{x}}$, we obtain

$$\tilde{\mathbf{x}}(t_{n+1}) = \mathbf{x}_n + h\mathbf{f}(t_n, \mathbf{x}_n) + \frac{1}{2} h^2 [\frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial t} + \frac{\partial \mathbf{f}(t, \mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(t, \mathbf{x})] + \mathcal{O}(h^3).$$

Now

$$\mathbf{x}_{n+1} - \tilde{\mathbf{x}}(t_{n+1}) =$$

$$= h(t_n, \mathbf{x}_n)(1 - (b_1 + b_2)) + h^2 \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial t}(b_2 c_2 - \frac{1}{2}) + h^2 \frac{\partial \mathbf{f}(t_n, \mathbf{x}_n)}{\partial \mathbf{x}} \mathbf{f}(t_n, \mathbf{x}_n)(b_2 a_{2,1} - \frac{1}{2}) + \mathcal{O}(h^3)$$

$$= \mathcal{O}(h^3)$$

if

$$b_1 + b_2 = 1, \quad b_2 c_2 = \frac{1}{2}, \quad a_{2,1} = c_2.$$

These are the conditions for the order of the method to be greater than or equal to 2.

These conditions do not define a 2-stage Runge-Kutta method uniquely. There are a number of popular choices for these coefficients shown in table 3.1 below. Since Butcher's work, the coefficients are usually displayed in the RK *tableau*, which is of the form

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

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

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

Table 3.1: RK Tableau's displaying some popular coefficient choices for a 2-stage Runge-Kutta method

c	A
	$\mathbf{b}^T$

.

We may obtain the conditions for third order schemes by performing another Taylor expansion for $\nu = 3$. This requires an incredible amount of effort, focus, speed and agility - which happen to be all the things I lack. Therefore, I shan't perform this feat but rather give the order conditions instead. We notice from the last expansion that we required $b_1 + b_2 = 1$, and in general if we wish our method to be of order 1 the first conditions are

$$\sum_i b_i = 1$$

$$\sum_i a_{ji} = c_j.$$

For order 2 we require the above conditions, along with

$$\sum_i b_i c_i = \tfrac{1}{2}$$

$$\sum_i a_{ji} = c_j \sum_i b_i c_i^2 = \tfrac{1}{3}.$$

In addition to the conditions just given, for order 3 we require

$$\sum_{i,j} b_j a_{j,i} c_j = \frac{1}{6}.$$

For higher orders, the analysis of the order conditions becomes complicated. It even happens that ν-stage Runge-Kutta methods of order ν exist only for $\nu \leq 4$, although, fourth order schemes are not beyond the powers of the Taylor expansion. Fortunately, there are substantially more powerful methods of deriving and analyzing the order conditions of Runge-Kutta methods than Taylor series expansions. However, this method will be deferred to the next chapter.

Some instances of third-order three stage Runge-Kutta methods are important enough to bear an individual name, for example the *classical* Runge-Kutta method and the *Nystrom* method shown in table 3.2.

In fact, the implicit midpoint rule that we used in the last chapter is a second order Runge-Kutta method. Its tableaux is also shown in table 3.2. By setting $b_1 = 1$ and $b_2 = 0$ we recover Euler's method!

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

0			
$\frac{2}{3}$	$\frac{2}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$

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

Table 3.2: Left: The classical Runge-Kutta method, Centre Left: The Nystrom Method, Centre Right: Implicit Midpoint Rule

The fourth order, four stage method is given by

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h[\frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6}]$$

with

$$k_1 = \mathbf{f}(t_n, \mathbf{x}_n)$$

$$k_2 = \mathbf{f}(t_n + \frac{h}{2}, \mathbf{x}_n + \frac{h}{2}k_1)$$

$$k_3 = \mathbf{f}(t_n + \frac{h}{2}, \mathbf{x}_n + \frac{h}{2}k_2)$$

$$k_4 = \mathbf{f}(t_n + h, \mathbf{x}_n + hk_3).$$

Its Butcher tableaux has the form

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

.

The local truncation error for this method is $\mathcal{O}(h^5)$.

3.3 Implicit Runge-Kutta Schemes

Just like the implicit numerical methods described in chapter 2, we can define an implicit Runge-Kutta method by allowing the vector functions to depend on each other in a more general way. Consider the scheme

$$k_j = \mathbf{f}(t_n + c_j h, \mathbf{x}_n + h\sum_{i=1}^{\nu} a_{j,i}\mathbf{f}(t_n + c_i h, k_i)), \quad j = 1, 2, ..., \nu$$

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h\sum_{j=1}^{\nu} b_j k_j$$

In this case, the matrix $A = (a_{j,i})_{j,i=1,2,...,\nu}$ is an arbitrary matrix, whereas with explicit Runge-Kutta methods the matrix was strictly lower triangular. In the implicit case, the slopes k_i can no longer be computed explicitly, and may not even necessarily exist! But we are assured by the

implicit function theorem that, for sufficiently small h, the nonlinear Implicit Runge-Kutta scheme, for the values $k_1, k_2, ..., k_\nu$, has a locally unique solution close to $k_j \approx \mathbf{f}(t_0, \mathbf{x}_0)$.

To check the order conditions of the implicit Runge-Kutta method we can once again expand the method into a Taylor series. This leads to the same order conditions obtained for explicit Runge-Kutta methods.

Chapter 4

Butcher's Order Conditions for Runge-Kutta Methods

4.1 Runge-Kutta Order Conditions

In this section we derive the order conditions of Runge-Kutta methods by comparing the Taylor series of the exact solution of $\mathbf{x}' = \mathbf{f}(\mathbf{x})$ with that of the numerical solution. Note that the system of ODE's we are considering does not depend on time explicitly. This system is called an *autonomous* system of ODE's. The computations of the Taylor series are greatly simplified by considering autonomous systems and by the use of rooted trees (defined later). One can see that it is reasonable to only consider autonomous systems because any system of the form $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ can be brought into the autonomous form by appending the equation $\dot{t} = 1$.

Although the theory of the order conditions of Runge-Kutta methods was developed by Butcher in the years 1963-72, this chapter will follow the four-step development given by Hairer, Lubich and Wanner (2002) where all definitions, theorems and proofs are provided by the aforementioned authors.

4.1.1 Derivation of the Order Conditions

Step 1. We compute the higher derivatives of the solution $\mathbf{x}$ to $\mathbf{x}' = \mathbf{f}(\mathbf{x})$ at the initial point t_0. For this we have

$$\mathbf{x}^{(k)} = (\mathbf{f}(\mathbf{x}))^{(k-1)}$$

where the right-hand-side is computed using the chain rule, product rule, the symmetry of partial derivatives and the notation $\mathbf{f}'(\mathbf{x})$ for the derivative as a linear map (i.e. the Jacobian),

$\mathbf{f}''(\mathbf{x})$ the second derivative as a bilinear map and similarly for higher derivatives. To make this idea clear consider the second derivative of $\mathbf{x}$ in terms of components. The ith component of the second derivative of $\mathbf{x}$ is

$$\begin{aligned}\tfrac{d^2}{dt^2}x^i(t) &= \tfrac{d}{dt}f^i(\mathbf{x}(t)) \\ &= \textstyle\sum_j \tfrac{\partial}{\partial x^j}f^i(\mathbf{x}(t))\tfrac{d}{dt}x^j(t) \\ &= \textstyle\sum_j \tfrac{\partial}{\partial x^j}f^i(\mathbf{x}(t))f^j(\mathbf{x}(t))\end{aligned}$$

or in matrix notation

$$\frac{d^2}{dt^2}\begin{bmatrix} x^1(t) \\ x^2(t) \\ \vdots \\ x^n(t) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x^1}f^1(\mathbf{x}(t)) & \frac{\partial}{\partial x^2}f^1(\mathbf{x}(t)) & \cdots & \frac{\partial}{\partial x^n}f^1(\mathbf{x}(t)) \\ \frac{\partial}{\partial x^1}f^2(\mathbf{x}(t)) & \frac{\partial}{\partial x^2}f^2(\mathbf{x}(t)) & \cdots & \frac{\partial}{\partial x^n}f^2(\mathbf{x}(t)) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x^1}f^n(\mathbf{x}(t)) & \frac{\partial}{\partial x^2}f^n(\mathbf{x}(t)) & \cdots & \frac{\partial}{\partial x^n}f^n(\mathbf{x}(t)) \end{bmatrix} \begin{bmatrix} \frac{d}{dt}x^1(t) \\ \frac{d}{dt}x^2(t) \\ \vdots \\ \frac{d}{dt}x^n(t) \end{bmatrix}.$$

By associating the left-hand-side with the notation $\mathbf{x}''$, the Jacobian matrix with $\mathbf{f}'(\mathbf{x})$ and the vector of time derivatives of $\mathbf{x}$ with $\mathbf{x}'$, the computations of the derivatives become greatly simplified. Using this notation we obtain

$$\begin{aligned}\mathbf{x}' &= \mathbf{f}(\mathbf{x}) \\ \mathbf{x}'' &= \mathbf{f}'(\mathbf{x})\mathbf{x}' \\ \mathbf{x}^{(3)} &= \mathbf{f}''(\mathbf{x})(\mathbf{x}',\mathbf{x}') + \mathbf{f}'(\mathbf{x})\mathbf{x}'' \\ \mathbf{x}^{(4)} &= \mathbf{f}^{(3)}(\mathbf{x})(\mathbf{x}',\mathbf{x}',\mathbf{x}') + 3\mathbf{f}''(\mathbf{x})(\mathbf{x}'',\mathbf{x}') + \mathbf{f}'(\mathbf{x})\mathbf{x}^{(3)} \\ \mathbf{x}^{(5)} &= \mathbf{f}^{(4)}(\mathbf{x})(\mathbf{x}',\mathbf{x}',\mathbf{x}',\mathbf{x}') + 6\mathbf{f}^{(3)}(\mathbf{x})(\mathbf{x}'',\mathbf{x}',\mathbf{x}') + 4\mathbf{f}''(\mathbf{x})(\mathbf{x}^{(3)},\mathbf{x}') + 3\mathbf{f}''(\mathbf{x})(\mathbf{x}'',\mathbf{x}'') + \mathbf{f}'(\mathbf{x})\mathbf{x}^{(4)}\end{aligned}$$

etc.

Step 2. The above procedure has given a recursive relationship between the derivatives of $\mathbf{x}$. We can substitute the computed derivatives $\mathbf{x}', \mathbf{x}'', ...$ into the right-hand-side of the higher order derivatives. The result of doing this is shown for the first few formulas with the arguments $(\mathbf{x})$ suppressed.

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

The terms that appear in each formula, denoted by $F(\tau)$, will be called the *elementary differentials*. We represent each of them by a suitable graph (rooted tree) τ, but first we need a few definitions of graphs, their properties and rooted trees.

Definition 8. (Graphs) A graph G is an ordered triple $(V(G), E(G), \psi_G)$ consisting of a non-empty set, $V(G)$, of vertices, a set $E(G)$, disjoint from $V(G)$, of edges and an incidence function ψ_G that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G.

Definition 9. (Walks) A Walk in a graph G is a finite, non-null sequence $W = v_0 e_1 v_1 e_2 v_2 ... e_k v_k$, whose terms alternately vertices and edges, such that for $1 \leq i \leq k$, the ends of e_i are v_{i-1} and v_i. The vertices v_0 and v_k are called the Origin and Terminus of W, respectively. The integer k is the length of the walk, W. A walk, W, is closed if it has positive length and the Origin and Terminus are the same.

Definition 10. (Trails and Paths) If the edges $e_1, e_2, ..., e_k$ in a walk, W, are all distinct, then W is called a Trail. If, in addition, the vertices $v_0, v_1, ..., v_k$ are all distinct, then W is called a path.

Definition 11. (Connected Graph) A graph G is Connected if for any two vertices $u, v \in V(G)$ there exists a Path such that u is the Origin (Terminus) and v is the Terminus (Origin).

Definition 12. (Cycles and Acyclic Graphs) A closed Trail is called a Cycle. A Graph without any closed Trails (Cycles) is called Acyclic.

Definition 13. (Rooted Trees) A rooted tree is a Connected, Acyclic graph where one vertex is specified as the root.

After a laborious definition of Rooted trees we may now continue with the development of the order conditions of Runge-Kutta methods. As mentioned before, we may represent each elementary differential by a rooted tree. This is done as follows:

Each $\mathbf{f}$ becomes a vertex, each $\mathbf{f}'$ becomes a vertex with one edge pointing upwards, and a kth derivative $\mathbf{f}^{(k)}$ becomes a vertex with k edges pointing upwards. The arguments of the k-linear mapping $\mathbf{f}^{(k)}(\mathbf{x})$ correspond to trees that are attached on the upper ends of these edges. As an example I shall construct the rooted tree corresponding to the elementary differential $\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f})$.

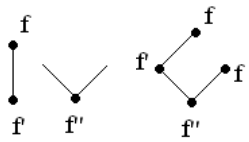


Figure 4.1: Construction of the rooted tree corresponding to $\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f})$.

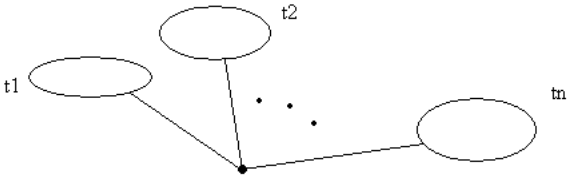


Figure 4.2: Rooted tree obtained recursively from $\tau_1, ..., \tau_m$.

This will be a multi-step process as we will need to also construct the rooted tree corresponding to $\mathbf{f}'\mathbf{f} = \mathbf{f}'(\mathbf{f})$, the steps are shown in figure 4.1. The first derivative $\mathbf{f}'$ corresponds to a vertex with one edge pointing upwards, then $\mathbf{f}$ corresponds to a single vertex which is attached to the top of this edge. Now $\mathbf{f}''$ corresponds to a vertex with two edges pointing upwards; the rooted tree we have just constructed is then attached to the top of one of these edges and there is one argument remaining, namely $\mathbf{f}$, which is attached to the top of the remaining edge as a single vertex. We can construct all rooted trees corresponding to the elementary differentials in this recursive manner.

Definition 14. (Trees). The set of (rooted) trees T is recursively defined as follows:

1. the graph $\bullet$ with only one vertex (called the root) belongs to T;
2. if $\tau_1, ..., \tau_n \in T$, then the graph obtained by grafting the roots of $\tau_1, ..., \tau_n$ to a new vertex also belongs to T (shown in figure 4.2). It is denoted by $\tau = [\tau_1, ..., \tau_n]$, and the new vertex (denoted by []) is the root of τ.

We also denote by $|\tau|$ the *order* of τ (the number of vertices), and by $\alpha(\tau)$ the coefficients appearing in the formulas for $\mathbf{x}', \mathbf{x}'', \mathbf{x}^{(3)}, ...$ It is important to note that some of the trees among $\tau_1, ..., \tau_m$ may be equal and that τ does not depend on the ordering of $\tau_1, ..., \tau_m$. That is, we do not distinguish between $[[\bullet], \bullet]$ and $[\bullet, [\bullet]]$.

Essentially, we can build any tree from the single node tree, $\bullet$, via this recursive definition. For instance, we can start with $\bullet$ and build a new tree $[\bullet]$, which is obtained by joining two single node trees by an edge. From these two trees we can obtain a number of new trees; $[\bullet, \bullet]$, $[[\bullet]]$, $[[\bullet], \bullet]$, $[[\bullet], [\bullet]]$ etc.

Definition 15. (Elementary Differentials). For a tree $\tau \in T$ the *elementary differential* is a mapping $F(\tau) : \mathbb{R}^n \to \mathbb{R}^n$, defined recursively by $F(\bullet)(\mathbf{x}) = f(\mathbf{x})$ and

$$F(\tau)(\mathbf{x}) = f^{(m)}(\mathbf{x})(F(\tau_1)(\mathbf{x}), ...F(\tau_m)(\mathbf{x})) \qquad for\ \tau = [\tau_1, ..., \tau_m].$$

With these definitions and the expressions for the derivatives of $\mathbf{x}(t)$, we obtain:

Theorem 16. (Hairer, Lubich and Wanner (2002)). *The kth derivative of the exact solution is given by*

$$\mathbf{x}^{(k)}(t_0) = \sum_{|\tau|=k} \alpha(\tau) F(\tau)(\mathbf{x}_0),$$

where $\alpha(\tau)$ are positive integer coefficients.

We also note that the coefficient $\alpha(\tau)$ is the number of monotonic labellings of the tree τ. As an example, consider how the tree  is obtained:

$\mathbf{x}^{(3)} = \mathbf{f}''(\mathbf{f},\mathbf{f}) + \mathbf{f}'\mathbf{f}'\mathbf{f}$ corresponds to the expression $\mathbf{x}^{(3)} =$  , and the vertices have been labeled monotonically (each vertex attached to the top of an upward pointing edge has a label strictly greater than the vertex connected directly below). Note also that the labellings shown are the only possible monotonic labellings of these trees. In differentiating $\mathbf{x}^{(3)}$, we obtain

$$[\mathbf{f}^{(3)}(\mathbf{f},\mathbf{f},\mathbf{f}) + 2\mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f})] + [\mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f}) + \mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f}) + \mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}].$$

In the first bracket of the above expression we obtain 2  . This comes from the labeled tree  and in differentiating the arguments in the expression $\mathbf{f}''(\mathbf{f},\mathbf{f})$ which is equivalent to adding a vertex in the following two ways:  and  . In the second bracket we obtain  which comes from the tree  and differentiating the first $\mathbf{f}'$ in th expression $\mathbf{f}'\mathbf{f}'\mathbf{f}$. This is equivalent to adding a vertex to  to obtain  , which is the only other possible monotonic labeling of  . The monotonicity of the labeling is arising because of matrix multiplication which is non-commutative. For example, we only obtain one monotonic labeling of  from differentiating the expression $\mathbf{f}'\mathbf{f}'\mathbf{f}$ because by the product rule differentiation of the first $\mathbf{f}'$ does not yield the same tree as differentiation of the second $\mathbf{f}'$ because matrix multiplication is not commutative. Therefore, the ordering of the differentiation matters (only in a single term such as $\mathbf{f}'\mathbf{f}'\mathbf{f}$ of course!) and dictates that the ordering of the trees is monotonic. Therefore, the coefficient $\alpha(\tau)$ is the number of monotonic labellings of the tree τ.

Step 3. We now turn to the numerical solution of the ODE, namely the Runge-Kutta method, which, by putting $hk_i = g_i$, we write as

$$g_i = h\mathbf{f}(u_i)$$

and

$$u_i = \mathbf{x}_0 + \sum_j a_{ij} g_j \qquad \mathbf{x}_1 = \mathbf{x}_0 + \sum_i b_i g_i,$$

where u_i, g_i and $\mathbf{x}_1$ are functions of h. Here we have used the more general implicit Runge-Kutta method. Substituting the expression for u_i into the argument of $\mathbf{f}$ in the expression for g_i will yield the form of Runge-Kutta methods given in chapter 3. We can compute the derivatives of g_i in the same way as we computed the derivatives of the exact solution in step 2. Differentiating with respect to h and using the product rule yields

$$\begin{aligned}
&\dot{g}_i = \mathbf{f}(u_i) + h\mathbf{f}'(u_i)\dot{u}_i \\
&\ddot{g}_i = 2\mathbf{f}'(u_i)\dot{u}_i + h(\mathbf{f}''(u_i)(\dot{u}_i, \dot{u}_i) + \mathbf{f}'(u_i)\ddot{u}_i) \\
&g_i^{(3)} = 3(\mathbf{f}''(u_i)(\dot{u}_i, \dot{u}_i) + \mathbf{f}'(u_i)\ddot{u}_i) + h(\mathbf{f}^{(3)}(u_i)(\dot{u}_i, \dot{u}_i, \dot{u}_i) + 3\mathbf{f}''(u_i)(\ddot{u}_i, \dot{u}_i) + \mathbf{f}'(u_i)u_i^{(3)})
\end{aligned}$$

and one is easily convinced of the general formula

$$g_i^{(k)} = h(\mathbf{f}(u_i))^{(k)} + k(\mathbf{f}(u_i))^{(k-1)}.$$

We are interested in the limit as h tends to zero, after-all, Runge-Kutta methods are convergent and approach the exact solution as h tends to zero. The general formula, for $h = 0$, gives

$$g_i^{(k)} = k(\mathbf{f}(u_i))^{(k-1)},$$

which is exactly the same as the expression for the derivatives of $\mathbf{x}$ given in the first step except with $\mathbf{x}$ replaced with u_i and with an extra factor k. Consequently,

$$\begin{aligned}
&\dot{g}_i = \mathbf{f}(\mathbf{x}_0) \\
&\ddot{g}_i = 2\mathbf{f}'(\mathbf{x}_0)\dot{u}_i \\
&g_i^{(3)} = 3(\mathbf{f}''(\mathbf{x}_0)(\dot{u}_i, \dot{u}_i) + \mathbf{f}'(\mathbf{x}_0)\ddot{u}_i) \\
&g_i^{(4)} = 4(\mathbf{f}^{(3)}(\mathbf{x}_0)(\dot{u}_i, \dot{u}_i, \dot{u}_i) + 3\mathbf{f}''(\mathbf{x}_0)(\ddot{u}_i, \dot{u}_i) + \mathbf{f}'(\mathbf{x}_0)u_i^{(3)}) \\
&g_i^{(5)} = 5(\mathbf{f}^{(4)}(\mathbf{x}_0)(\dot{u}_i, \dot{u}_i, \dot{u}_i, \dot{u}_i) + 6\mathbf{f}^{(3)}(\mathbf{x}_0)(\ddot{u}_i, \dot{u}_i, \dot{u}_i) + 4\mathbf{f}''(\mathbf{x}_0)(u_i^{(3)}, \dot{u}_i) + 3\mathbf{f}''(\mathbf{x}_0)(\ddot{u}_i, \ddot{u}_i) + \mathbf{f}'(\mathbf{x}_0)u_i^{(4)}) \\
&\text{etc.}
\end{aligned}$$

The argument of $\mathbf{f}$ appears because we are evaluating the derivatives of g_i at $h = 0$, therefore $u_i = \mathbf{x}_0$ in the argument of $\mathbf{f}$.

Step 4. Once again, we can substitute the computed derivatives $\dot{g}_i$, $\ddot{g}_i$,... into the expressions

on the right-hand-side of the higher order derivatives. This will give the next higher derivative of g_i, and, using

$$u_i{}^{(k)} = \sum_j a_{ij} g_j^{(k)},$$

which follows from the Runge-Kutta method defined in step 3, we can find the derivatives of u_i. This process begins as

$$\dot{g}_i = 1 \cdot \mathbf{f} \qquad \dot{u}_i = (\textstyle\sum_j a_{ij}) \cdot \mathbf{f}$$

$$\ddot{g}_i = (1 \cdot 2)(\textstyle\sum_j a_{ij})\mathbf{f}'\mathbf{f} \qquad \ddot{u}_i = (1 \cdot 2)(\textstyle\sum_{j,k} a_{ij} a_{jk})\mathbf{f}'\mathbf{f}$$

and so on. If we compare these expressions with expressions derived for the exact solution in step 2, we see that the results are exactly the same, apart from the extra factors. We denote the integer factors $1, 1 \cdot 2, ...$ by $\gamma(\tau)$, the factors containing the a_{ij}'s in the expression for $g_i^{(k)}$ by $\mathbf{g}_i(\tau)$ and the factors containing the a_{ij}'s in the expression for $u_i^{(k)}$ by $\mathbf{u}_i(\tau)$. Continuing the above process inductively we obtain, in contrast to Theorem 16,

$$g_i^{(k)}|_{h=0} = \textstyle\sum_{|\tau|=k} \gamma(\tau) \cdot \mathbf{g}_i(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0)$$

$$u_i^{(k)}|_{h=0} = \textstyle\sum_{|\tau|=k} \gamma(\tau) \cdot \mathbf{u}_i(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0),$$

where $\alpha(\tau)$ and $F(\tau)$ are the same quantities as in Theorem 16. We can see that $\alpha(\tau)$ and $F(\tau)$ are the same quantities as in Theorem 16 by continuing the insertion process of the derivatives of $u_i^{(k)}$ into the right-hand side of the derivatives of g_i. For example, if $\dot{u}_i$ and $\ddot{u}_i$ are inserted into $3\mathbf{f}''(\ddot{u}_i, \dot{u}_i)$ we obtain

$$3\mathbf{f}''((1 \cdot 2)(\textstyle\sum_{j,k} a_{ij} a_{jk})\mathbf{f}'\mathbf{f}, (\textstyle\sum_j a_{ij})\mathbf{f})$$

$$= [(1 \cdot 2)(\textstyle\sum_{j,k} a_{ij} a_{jk})(\textstyle\sum_j a_{ij})] 3\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f})$$

and the $3\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f})$ corresponds to the quantities $\alpha(\tau)$ and $F(\tau)$ defined in Theorem 13. The elementary differential $\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f})$ corresponds to the tree $[[\bullet], \bullet]$; the factors containing the a_{ij}'s in the above expression correspond to $\mathbf{u}_i([\bullet])$ and $\mathbf{u}_i(\bullet)$, and the integer factors correspond to $\gamma([\bullet])$ and $\gamma(\bullet)$. Therefore, in the expression for $g_i^{(4)}|_{h=0}$, the term corresponding to $3\mathbf{f}''(\ddot{u}_i, \dot{u}_i)$ will have the form

$$(1 \cdot 2 \cdot 4)[\mathbf{u}_i([\bullet])\mathbf{u}_i(\bullet)] 3\mathbf{f}''(\mathbf{f}'\mathbf{f}, \mathbf{f}).$$

Firstly, we can see that $\mathbf{g}_i([[\bullet], \bullet]) = \mathbf{u}_i([\bullet])\mathbf{u}_i(\bullet)$. For a general tree $\tau = [\tau_1, ..., \tau_m]$ this will be

$$\mathbf{g}_i(\tau) = \mathbf{u}_i(\tau_1) \cdot ... \cdot \mathbf{u}_i(\tau_m).$$

Secondly, the integer factor corresponds to $4 \cdot \gamma([\bullet])\gamma(\bullet)$. The factor of 4 appears because of the order of the tree corresponding to the elementary differential. For a general tree $\tau = [\tau_1, ..., \tau_m]$, the γ's will receive the additional factor $k = |\tau|$ (which comes from the derivatives of g_i), thus we have in general

$$\gamma(\tau) = |\tau|\, \gamma(\tau_1) \cdot ... \cdot \gamma(\tau_m).$$

Using ${u_i}^{(k)} = \sum_j a_{ij} g_j^{(k)}$ and the expressions for the derivatives of u_i and g_i we obtain

$$\sum_{|\tau|=k} \gamma(\tau) \cdot \mathbf{u}_i(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0) = \sum_j a_{ij} \sum_{|\tau|=k} \gamma(\tau) \cdot \mathbf{g}_j(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0).$$

after equating equivalent trees and canceling the γ and α factors we see that

$$\mathbf{u}_i(\tau) = \sum_j a_{ij} \mathbf{g}_j(\tau) = \sum_j a_{ij} \mathbf{u}_j(\tau_1) \cdot ... \cdot \mathbf{u}_j(\tau_m).$$

This formula can be used repeatedly, as long as some of the trees $\tau_1, ..., \tau_m$ are of order> 1. Finally, we use the expression $\mathbf{x}_1 = \mathbf{x}_0 + \sum_i b_i g_i$ from the Runge-Kutta method. We have that

$$\mathbf{x}_1^{(k)}|_{h=0} = \textstyle\sum_i b_i g_i^{(k)} = \sum_i b_i \sum_{|\tau|=k} \gamma(\tau) \cdot \mathbf{g}_i(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0)$$

$$= \textstyle\sum_{|\tau|=k} \gamma(\tau) \cdot (\sum_i b_i \mathbf{g}_i(\tau)) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0).$$

We denote the term $\sum_i b_i \mathbf{g}_j(\tau)$ by $\phi(\tau)$ and call the *elementary weights*. These results are summarized in the following theorem.

Theorem 17. (Hairer, Lubich and Wanner (2002)). *The derivatives of the numerical solution of a Runge-Kutta method, for h=0, are given by*

$$\mathbf{x}_1^{(k)}|_{h=0} = \sum_{|\tau|=k} \gamma(\tau) \cdot \phi(\tau) \cdot \alpha(\tau) F(\tau)(\mathbf{x}_0),$$

where $\alpha(\tau)$ and $F(\tau)$ are the same as in Theorem 16, the coefficients $\gamma(\tau)$ satisfy $\gamma(\bullet) = 1$ and $\gamma(\tau) = |\tau|\, \gamma(\tau_1) \cdot ... \cdot \gamma(\tau_m)$. The elementary weights $\phi(\tau)$ are obtained from the tree τ as follows: attach to every vertex a summation letter (i to the root), then $\phi(\tau)$ is the sum, over all summation indices, of a product composed of b_i, and factors a_{jk} for each vertex j directly connected with 'k' by an upwards directed edge.

Proof. For a general tree $\tau = [\tau_1, ..., \tau_m]$ we have that

$$\phi(\tau) = \sum_i b_i \mathbf{g}_i(\tau) = \sum_i b_i \mathbf{u}_i(\tau_1) \cdot ... \cdot \mathbf{u}_i(\tau_m)$$

$$= \sum_i b_i (\sum_j a_{ij} \mathbf{g}_j(\tau_1)) \cdot ... \cdot (\sum_s a_{is} \mathbf{g}_s(\tau_m))$$

$$= \sum_{i,j,...,s} b_i a_{ij} \cdot ... \cdot a_{is} \mathbf{g}_j(\tau_1) \cdot ... \cdot \mathbf{g}_s(\tau_m).$$

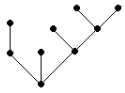
This is equivalent to attaching the summation index i to the root of τ and the summation indices $j, ..., s$ to all vertices that are directly connected to the root by an upward edge and summing over the coefficients b_i and factors a_{ij}. But the τ_k's may also be of order> 1, therefore we may repeat the above procedure on $\mathbf{g}_j(\tau_1) \cdot ... \cdot \mathbf{g}_s(\tau_m)$ and obtain an identical interpretation. Continuing this procedure until the argument in each $\mathbf{g}_l$ is the single node tree we find that the elementary weight $\phi(\tau)$ is the collection of $\sum_i b_i$ and all $\sum_j a_{ij}$. This is equivalent to attaching to every vertex a summation letter ('i' to the root), then summing over all summation indices a product composed of b_i, and factors a_{jk} for each vertex j directly connected with k by an upwards directed edge.□

Theorem 18. (Hairer, Lubich and Wanner (2002)). *The Runge-Kutta method has order p if and only if*

$$\phi(\tau) = \frac{1}{\gamma(\tau)} \qquad for\ |\tau| \leq p.$$

Proof. The sufficiency follows from a comparison of Theorem 17 with Theorem 16. The necessity follows from the linear independence of the elementary differentials.

Example 19. ([10]) For the following tree of order 9 we can calculate the elementary weight, $\phi(\tau)$, using $\gamma(\tau)$ as follows



Using the recursive definition of $\gamma(\tau)$ we have $\gamma(\tau) = |\tau|\,\gamma(\tau_1)\gamma(\tau_2)\gamma(\tau_3) = 9 \cdot \gamma(\tau_1)\gamma(\tau_2)\gamma(\tau_3)$, where

$\tau_1 =$, $\tau_2 = \bullet$, $\tau_3 =$  . Then $\gamma(\tau_1) = |\tau_1| \cdot \gamma(\bullet) = 2 \cdot 1 = 2$, $\gamma(\tau_2) = \gamma(\bullet) = 1$ and $\gamma(\tau_3) = |\tau_3| \cdot \gamma(\tau_{1a})\gamma(\tau_{2a}) = 5 \cdot \gamma(\tau_{1a})\gamma(\tau_{2a})$, where

$\tau_{1a} = \bullet$ and $\tau_{2a} =$  . Hence $\gamma(\tau_{1a}) = \gamma(\bullet) = 1$ and $\gamma(\tau_{2a}) = |\tau_{2a}|\,\gamma(\bullet)\gamma(\bullet) = 3 \cdot 1 \cdot 1 = 3$, thus $\gamma(\tau_3) = 5 \cdot 3$. Therefore $\gamma(\tau) = 9 \cdot 2 \cdot 5 \cdot 3$ and we require $\phi(\tau) = \frac{1}{270}$.

The quantities $\phi(\tau)$ and $\gamma(\tau)$ for all trees up to order 4 are given in Table 4.1. This also verifies the formulas for the order conditions of Runge-Kutta methods stated in the last chapter.

$\|\tau\|$	τ	Graph	$\alpha(\tau)$	F(τ)	$\gamma(\tau)$	$\phi(\tau)$	$\sigma(\tau)$
1	$\bullet$		1	f	1	$\sum_i b_i$	1
2	$[\bullet]$		1	$f'f$	2	$\sum_{ij} b_i a_{ij}$	1
3	$[\bullet,\bullet]$		1	$f''(f,f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$	2
3	$[[\bullet]]$		1	$f'f'f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$	1
4	$[\bullet,\bullet,\bullet]$		1	$f'''(f,f,f)$	4	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{il}$	6
4	$[[\bullet],\bullet]$		3	$f'f''(f,f)$	8	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{jl}$	1
4	$[[\bullet,\bullet]]$		1	$f''(f'f,f)$	12	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{jl}$	2
4	$[[[\bullet]]]$		1	$f'f'f'f$	24	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{kl}$	1

Table 4.1: Trees, Elementary Differentials and Coefficients.

4.2 B-Series

We are now in a position to introduce the concept of B-series, which gives a deeper insight into the qualitative and quantitative behavior of numerical methods and allows for extensions to more general classes of methods.

We study power series in $h^{|\tau|}$ containing elementary differentials $F(\tau)$ and arbitrary coefficients which are now written in the form $a(\tau)$. Such a power series will be called a $B-Series$. We start with

$$B(a,\mathbf{x}) = \mathbf{x} + a(\bullet)h\mathbf{f}(\mathbf{x}) + a([\bullet])h^2(\mathbf{f}'\mathbf{f})(\mathbf{x}) + ... = \mathbf{x} + \delta,$$

and get by Taylor expansion

$$h\mathbf{f}(B(a,\mathbf{x}) = h\mathbf{f}(\mathbf{x}+\delta) = h\mathbf{f}(\mathbf{x}) + h\mathbf{f}'(\mathbf{x})\delta + \frac{h}{2!}\mathbf{f}''(\mathbf{x})(\delta,\delta) + ...$$

Inserting the expression for δ and multiplying out, we obtain

$$h\mathbf{f}(B(a,\mathbf{x}) = h\mathbf{f}(\mathbf{x}) + a(\bullet)h^2\mathbf{f}'\mathbf{f} + a([\bullet])h^3\mathbf{f}'\mathbf{f}'\mathbf{f} + a(\bullet)^2\frac{h^3}{2!}\mathbf{f}''(\mathbf{f},\mathbf{f}) + a(\bullet)a([\bullet])h^4\mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f}) + ...$$

This formula is still not quite perfect yet for two reasons. First, there is a factor of $\frac{1}{2!}$ in the fourth term. This factor appears because of the symmetry of the tree  . We therefore introduce the symmetry coefficients (originally defined by Butcher, 1987). Secondly, there is no first term $\mathbf{x}$. Thus, we make use of the coefficient $a(\varnothing)$.

Definition 20. (Symmetry Coefficients). The symmetry coefficients $\sigma(\tau)$ are defined by $\sigma(\bullet) = 1$ and, for $\tau = [\tau_1, \tau_2, ..., \tau_m]$,

$$\sigma(\tau) = \sigma(\tau_1) \cdot ... \cdot \sigma(\tau_m) \cdot \mu_1!\mu_2! \cdot ...,$$

where the integers $\mu_1, \mu_2, ...$ count equal trees among $\tau_1, \tau_2, ..., \tau_m$.

Example 21. As an example, consider the tree of order 9 from the previous example. This tree may be represented as $\tau = [[\bullet], \bullet, [\bullet, [\bullet, \bullet]]]$ so that $\tau_1 = [\bullet]$, $\tau_2 = \bullet$ and $\tau_3 = [\bullet, [\bullet, \bullet]]$. Then $\sigma(\tau_1) = \sigma(\tau_2) = 1$. The tree τ_3 is of the form $[\tau_{1a}, \tau_{2a}]$, where $\tau_{1a} = \bullet$ and $\tau_{2a} = [\bullet, \bullet]$ so that $\sigma(\tau_{1a}) = 1$. Now there are two equal trees in τ_{2a}, therefore $\mu_{1a} = 2$ and $\sigma(\tau_{2a}) = \sigma(\bullet)\sigma(\bullet) \cdot 2! = 2$. Thus, $\sigma(\tau_3) = \sigma(\tau_{1a}) \cdot \sigma(\tau_{2a}) = 2$ since $\tau_{1a} \neq \tau_{2a}$. Putting all the information together we find that, since no trees among τ_1, τ_2, τ_3 are equal, $\sigma(\tau) = \sigma(\tau_1)\sigma(\tau_2)\sigma(\tau_3) = 2$.

Definition 22. (B-Series). For a mapping $a : T \cup \{\varnothing\} \to \mathbb{R}$ a formal series of the form

$$B(a, \mathbf{x}) = a(\varnothing)\mathbf{x} + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(\mathbf{x})$$

is called a B-Series.

B-Series were first introduced by Hairer & Wanner (1974), although, the main results of the theory of B-Series have their origin in the paper of Butcher (1972). The normalization factor of $\frac{1}{\sigma(\tau)}$ is due to Butcher and Sanz-Serna (1996). The next Lemma gives an alternative way of finding the order conditions.

Lemma 23. *Let $a : T \cup \{\varnothing\} \to \mathbb{R}$ be a mapping satisfying $a(\varnothing) = 1$. Then the corresponding B-Series inserted into $h\mathbf{f}(\cdot)$ is again a B-Series. That is,*

$$h\mathbf{f}(B(a, \mathbf{x})) = B(a', \mathbf{x}),$$

where $a'(\varnothing) = 0$, $a'(\bullet) = 1$, and

$$a'(\tau) = a(\tau_1)a(\tau_2) \cdot ... \cdot a(\tau_m) \qquad for \quad \tau = [\tau_1, ..., \tau_m].$$

Proof. Since $a(\varnothing) = 1$ we have $B(a, \mathbf{x}) = \mathbf{x} + \mathcal{O}(h)$, so that $h\mathbf{f}(B(a, \mathbf{x}))$can be expanded into a Taylor series around $\mathbf{x}$. We obtain

$$h\mathbf{f}(B(a, \mathbf{x})) = h \sum_{m \geq 0} \tfrac{1}{m!} \mathbf{f}^{(m)}(\mathbf{x})(B(a, \mathbf{x}) - \mathbf{x})^m$$

$$= h \sum_{m \geq 0} \tfrac{1}{m!} \mathbf{f}^{(m)}(\mathbf{x})(\sum_{\tau \in T} \tfrac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(\mathbf{x}))^m \qquad (Since\ a(\varnothing) = 1)$$

$$= h \sum_{m \geq 0} \tfrac{1}{m!} \mathbf{f}^{(m)}(\mathbf{x})((\sum_{\tau_1 \in T} \tfrac{h^{|\tau_1|}}{\sigma(\tau_1)} a(\tau_1) F(\tau_1)(\mathbf{x})) \cdot ... \cdot (\sum_{\tau_m \in T} \tfrac{h^{|\tau_m|}}{\sigma(\tau_m)} a(\tau_m) F(\tau_m)(\mathbf{x})))$$

$$= h \sum_{m \geq 0} \tfrac{1}{m!} \sum_{\tau_1 \in T} \cdots \sum_{\tau_m \in T} \tfrac{h^{|\tau_1| + ... + |\tau_m|}}{\sigma(\tau_1) \cdot ... \cdot \sigma(\tau_m)} \cdot a(\tau_1) \cdot ... \cdot a(\tau_m) \cdot \mathbf{f}^{(m)}(F(\tau_1)(\mathbf{x}), ..., F(\tau_m)(\mathbf{x}))$$

$$= h \sum_{m \geq 0} \tfrac{1}{m!} \sum_{\tau_1 \in T} \cdots \sum_{\tau_m \in T} \tfrac{h^{|\tau_1| + ... + |\tau_m|}}{\sigma(\tau_1) \cdot ... \cdot \sigma(\tau_m)} \cdot (\tfrac{\mu_1! \mu_2! ...}{\mu_1! \mu_2! ...}) \cdot a(\tau_1) \cdot ... \cdot a(\tau_m) \cdot \mathbf{f}^{(m)}(F(\tau_1)(\mathbf{x}), ..., F(\tau_m)(\mathbf{x}))$$

$$(Where\ the\ \mu_i's\ are\ from\ definition\ 20)$$

$$\sum_{m\geq 0}\sum_{\tau_1\in T}\cdots\sum_{\tau_m\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\cdot\frac{\mu_1!\mu_2!\ldots}{m!}\cdot a'(\tau)\cdot F(\tau)(\mathbf{x})$$

$$for\ a\ tree\ \tau=[\tau_1,\ldots,\tau_m]\ and\ a(\tau_1)\cdot\ldots\cdot a(\tau_m)=a'(\tau)$$

$$=\sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}a'(\tau)F(\tau)(\mathbf{x})=B(a',\mathbf{x}).$$

The last equality follows from the fact that there are $\begin{pmatrix} m \\ \mu_1,\mu_2,\ldots \end{pmatrix}$ possibilities for writing the tree τ in the form $\tau=[\tau_1,\ldots,\tau_m]$. For example, the trees $[\bullet,\bullet,[\bullet]]$, $[\bullet,[\bullet],\bullet]$ and $[[\bullet],\bullet,\bullet]$ appear as different terms in the upper sum, but only as one term in the lower sum.□

4.2.1 Order Conditions

Let a Runge-Kutta method, say

$$g_i=h\mathbf{f}(u_i)$$

and

$$u_i=\mathbf{x}_0+\sum_j a_{ij}g_j\qquad \mathbf{x}_1=\mathbf{x}_0+\sum_i b_i g_i,$$

be given. All quantities in the defining formulas are set up as B-Series, $g_i=B(\mathbf{g}_i,\mathbf{x}_0)$, $u_i=B(\mathbf{u}_i,\mathbf{x}_0)$, $\mathbf{x}_1=B(\phi,\mathbf{x}_0)$ with $\mathbf{u}_i(\varnothing)=1$ and $\phi(\varnothing)=1$. Using $g_i=h\mathbf{f}(u_i)$ and Lemma 23 we have that

$$\mathbf{g}_i(\tau)=\mathbf{u}_i(\tau_1)\cdot\ldots\cdot\mathbf{u}_i(\tau_m)$$

for $\tau=[\tau_1,\ldots,\tau_m]$ and $\mathbf{g}_i(\varnothing)=0$. Using $u_i=\mathbf{x}_0+\sum_j a_{ij}g_j$ we obtain

$$\mathbf{x}_0+\sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\mathbf{u}_i(\tau)F(\tau)(\mathbf{x}_0)=\mathbf{x}_0+\sum_i a_{ij}\sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\mathbf{g}_j(\tau)F(\tau)(\mathbf{x}_0)$$

and hence

$$\mathbf{u}_i(\tau)=\sum_i a_{ij}\mathbf{g}_j(\tau).$$

Finally, using $\mathbf{x}_1=\mathbf{x}_0+\sum_i b_i g_i$ we obtain

$$\mathbf{x}_0+\sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\phi(\tau)F(\tau)(\mathbf{x}_0)=\mathbf{x}_0+\sum_i b_i\sum_{\tau\in T}\frac{h^{|\tau|}}{\sigma(\tau)}\mathbf{g}_i(\tau)F(\tau)(\mathbf{x}_0)$$

and deduce that

$$\phi(\tau) = \sum_i b_i \mathbf{g}_i(\tau).$$

These formulas are identical to the formulas derived in Step 4 of Section 4.1 which justifies the starting point as B-Series.

Now assume that the exact solution of the differential equation, $\mathbf{x}' = \mathbf{f}(\mathbf{x})$, to be a B-series $B(\mathbf{e}, \mathbf{x}_0)$. The B-series is really an expansion in powers of h about the initial condition and is therefore a function of h. Thus, in taking the derivative of the exact solution we take the derivative of $B(\mathbf{e}, \mathbf{x}_0)$ with respect to h and obtain

$$\sum_{\tau \in T} |\tau| \frac{h^{|\tau|-1}}{\sigma(\tau)} \mathbf{e}(\tau) F(\tau)(\mathbf{x}_0).$$

Substituting this into the left hand side of the differential equation and $B(\mathbf{e}, \mathbf{x}_0)$ into the argument of $\mathbf{f}$ on the right hand side then multiplying both sides by h we get

$$\sum_{\tau \in T} |\tau| \frac{h^{|\tau|}}{\sigma(\tau)} \mathbf{e}(\tau) F(\tau)(\mathbf{x}_0) = h\mathbf{f}(B(\mathbf{e}, \mathbf{x}_0)).$$

An application of Lemma 23 yields

$$\mathbf{e}(\tau) = \frac{1}{|\tau|} \mathbf{e}(\tau_1) \cdot ... \cdot \mathbf{e}(\tau_m).$$

Compared to the definition of $\gamma(\tau)$ we obtain

$$\mathbf{e}(\tau) = \frac{1}{\gamma(\tau)}.$$

Thus, the B-series of the exact solution is $B(\mathbf{e}, \mathbf{x}_0)$ and repeated differentiation of this series proves

$$\mathbf{x}^{(k)}(t_0)|_{h=0} = \sum_{|\tau|=k} \frac{|\tau|!}{\sigma(\tau)\gamma(\tau)} F(\tau)(\mathbf{x}_0)$$

and the formula

$$\alpha(\tau) = \frac{|\tau|!}{\gamma(\tau)\sigma(\tau)}.$$

Now a comparison of the series for the derivatives of the exact solution and the B-series for the Runge-Kutta method will yield the order conditions of Theorem 18.

If the tools of B-series are enriched by a more general composition law then this procedure may

be applied to yet larger classes of numerical methods.

4.3 Composition Methods

We now wish to find a way of calculating the order properties of the composition of two Runge-Kutta methods. That is, find formulas for the composition of two sets of elementary weights of two Runge-Kutta methods. Suppose that, starting from an initial value $\mathbf{x}_0$, we compute a numerical solution $\mathbf{x}_1$ using a Runge-Kutta method with coefficients a_{ij}, b_i and step size h. Then, continuing from $\mathbf{x}_1$, we compute a value $\mathbf{x}_2$ using another method with coefficients a^*_{ij}, b^*_i and the same step size. This composition is now considered as a *single* method (with coefficients $\hat{a_{ij}}$, $\hat{b_i}$). How can we express the elementary weights $\hat{\phi}(\tau)$ of the composition of the two methods in terms of the elementary weights of each individual method?

If the value $\mathbf{x}_1$ from the first method is inserted into the starting value for the second method, we can see that the coefficients of the combined method are given by (here written for two stage methods)

$$\begin{array}{cccc|c} \hat{a}_{11} & \hat{a}_{12} & & & \\ \hat{a}_{21} & \hat{a}_{22} & & & \\ \hat{a}_{31} & \hat{a}_{32} & \hat{a}_{33} & \hat{a}_{34} & = \\ \hat{a}_{41} & \hat{a}_{42} & \hat{a}_{43} & \hat{a}_{44} & \\ \hline \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \hat{b}_4 & \end{array} \quad \begin{array}{cccc} a_{11} & a_{12} & & \\ a_{21} & a_{22} & & \\ b_1 & b_2 & a^*_{11} & a^*_{12} \\ b_1 & b_2 & a^*_{21} & a^*_{22} \\ \hline b_1 & b_2 & b^*_1 & b^*_1 \end{array}$$

The idea is to write the sum for $\hat{\phi}(\tau)$, say for the tree $\begin{smallmatrix} & \cdot \\ \cdot & \end{smallmatrix}$, in full detail

$$\hat{\phi}(\tau) = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{k=1}^{4}\sum_{l=1}^{4} \hat{b_i}\hat{a}_{ij}\hat{a}_{ik}\hat{a}_{kl} = ...$$

and split each sum into two different index sets. The first index set will be the sum from $i = 1$ up to 2 and the second index set will be the sum from $i = 3$ up to 4. We get $2^{|\tau|}$ different expressions:

$$\hat{\phi}(\tau) = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} ./. + \sum_{i=3}^{4}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} ./. + \sum_{i=1}^{2}\sum_{j=3}^{4}\sum_{k=1}^{2}\sum_{l=1}^{2} ./. + ...$$

We symbolize each expression by drawing the corresponding vertex of τ as a bullet for the first index set and as a circle for the second. However, due to the zero pattern in the Butcher Tableaux above, each term with "circle above bullet" can be omitted since the corresponding $\hat{a}_{ij}$'s are zero. Therefore the only combinations to be considered are those of figure 4.3. Each "circle" vertex in figure 4.3 corresponds to coefficients a^*_{ij}, b^*_i. Inserting the quantities from the right Butcher Tableaux into the expressions corresponding to the trees in figure 4.3 we obtain

$\hat{\phi}(\begin{smallmatrix} & \cdot \\ \cdot & \end{smallmatrix}) = \sum b_i a_{ij} a_{ik} a_{kl} + \sum b^*_i b_j b_k a_{kl} + \sum b^*_i a^*_{ij} b_k a_{kl} + \sum b^*_i b_j a^*_{ik} b_l + \sum b^*_i a^*_{ij} a^*_{ik} b_l + \sum b^*_i b_j a^*_{ik} a^*_{kl} +$

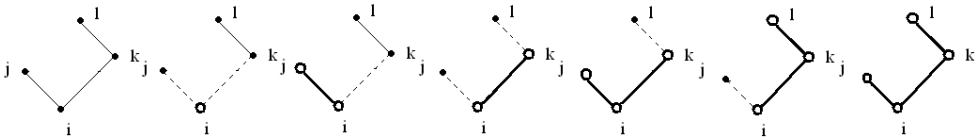


Figure 4.3: Combinations of Bullets and Circles with non-zero products.

$\sum b_i^* a_{ij}^* a_{ik}^* a_{kl}^*$

and observe that each factor of the type b_j interrupts the summation so that the terms decompose into factors of elementary weights of the individual methods as follows:

$\hat{\phi}(\ ) = \phi(\ ) + \phi^*(\bullet)\cdot\phi(\bullet)\phi(\ ) + \phi^*(\ )\cdot\phi(\ ) + \phi^*(\ )\cdot\phi(\bullet)\phi(\bullet) + \phi^*(\ )\cdot\phi(\bullet) + \phi^*(\ )\cdot\phi(\bullet) + \phi^*(\ )$.

The trees composed of the "circle" nodes of τ in figure 4.3 constitute all possible "sub-trees" θ having the same root as τ. This is the key to understanding the general result. In order to formalize the procedure of figure 4.3 we introduce the set OT of *ordered trees* recursively as follows: $\bullet \in OT$ and if $\omega_1, ..., \omega_m \in OT$, then also the ordered $m-$tuple $(\omega_1, ..., \omega_m) \in OT$. As the name suggests, in the graphical representation of an ordered tree the order of the branches leaving cannot be permuted. If we ignore the ordering, a tree can be considered as a representative of an equivalence class of ordered trees. For example, the tree of figure 4.3 has two orderings: and . Denote by $\nu(\tau)$ the number of possible orderings of a tree τ. It is defined recursively by $\nu(\bullet) = 1$ and

$$\nu(\tau) = \frac{m!}{\mu_1!\mu_2!\cdot ...}\nu(\tau_1)\cdot ...\cdot \nu(\tau_m)$$

for $\tau = [\tau_1, ..., \tau_m]$, where the integers $\mu_1, \mu_2, ...$ count equal trees among $\tau_1, ..., \tau_m$. This number is closely related to the symmetry coefficient $\sigma(\tau)$ because the product $\kappa(\tau) = \sigma(\tau)\nu(\tau)$ satisfies the recurrence relation

$$\kappa(\tau) = m!\kappa(\tau_1)\cdot ...\cdot \kappa(\tau_m).$$

We now introduce the set $OST(\omega)$ of *ordered subtrees* of an ordered tree $\omega \in OT$ by

$OST(\bullet) = \{\varnothing, \bullet\}$

$OST(\omega) = \{\varnothing\} \cup \{(\theta_1, ..., \theta_m) : \theta_i \in OST(\omega_i)\}\ for\ \omega = (\omega_1, ..., \omega_m)$.

Each ordered sub-tree $\theta \in OST(\omega)$ is naturally associated with a tree $\bar{\theta} \in T$ obtained by neglecting the ordering and the $\varnothing$-components of θ. However, the concept of ordering is only useful for finding the sub-trees corresponding to the index set expressions. Thus, for every tree $\tau \in T$ we choose, once and for all, an ordering. We denote this ordered tree by $\omega(\tau)$, and we put $OST(\tau) = OST(\omega(\tau))$. For the tree of figure 4.3, considered as an ordered tree, the ordered sub-trees correspond to the sub-trees composed of the "circle" nodes.

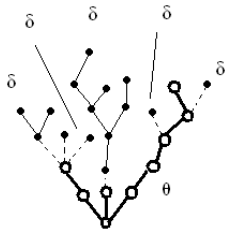


Figure 4.4: A tree with a sub-tree θ composed of "circle" nodes and sub-trees δ left over.

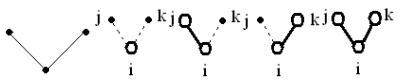


Figure 4.5: A tree with Symmetry

The general composition rule now becomes clear: for $\theta \in OST(\omega(\tau))$ we denote by $\omega(\tau) \setminus \theta$ the "forest" collecting the trees left over when θ has been removed from the ordered tree ω. Using the conventions $\phi^*(\theta) = \phi^*(\bar{\theta})$ and $\phi^*(\varnothing) = 1$ we have

$$\hat{\phi}(\tau) = \sum_{\theta \in OST(\omega(\tau))} \left(\phi^*(\theta) \cdot \prod_{\delta \in \omega(\tau) \setminus \theta} \phi(\delta) \right).$$

As an example, consider the tree in figure 4.4. Each sub-tree δ that is left over after the sub-tree θ has been removed is an element of the set $\omega(\tau) \setminus \theta$. Then the product, in the above formula, is over every sub-tree in the set $\omega(\tau) \setminus \theta$, while the sum is over every ordered sub-tree of $\omega(\tau)$.

The composition formulas for the trees up to order 3 are

$\hat{\phi}(\bullet) = \phi^*(\varnothing) \cdot \phi(\bullet) + \phi^*(\bullet)$

$\hat{\phi}(\tau_2) = \phi^*(\varnothing) \cdot \phi(\tau_2) + \phi^*(\bullet) \cdot \phi(\bullet) + \phi^*(\tau_2)$

$\hat{\phi}(\tau_3) = \phi^*(\varnothing) \cdot \phi(\tau_3) + \phi^*(\bullet) \cdot \phi(\bullet)^2 + 2\phi^*(\tau_2) \cdot \phi(\bullet) + \phi^*(\tau_3)$

$\hat{\phi}(\tau_4) = \phi^*(\varnothing) \cdot \phi(\tau_4) + \phi^*(\bullet) \cdot \phi(\tau_3) + \phi^*(\tau_2) \cdot \phi(\bullet) + \phi^*(\tau_4)$.

The factor 2 in the composition formula for the tree $\tau =$ has its origins in the symmetry of τ. The ordered sub-trees of τ are displayed in figure and we can see that the third and fourth sub-trees are topologically equivalent. Even though the node labellings of the two sub-trees are different, the convention $\phi^*(\theta) = \phi^*(\bar{\theta})$ produces the factor 2.

4.4 Composition of B-Series

Our goal is to now extend the composition law we found in the previous section to general B-series. That is, we compose a B-series with another B-series and demonstrate that the result is again a B-series. This will allow us to consider a B-series that is conjugate (by a B-series) to another B-series. Due to the similarities between the proofs of Lemma 23 and Theorem 24 (although the

proof of Theorem 24 is technically more difficult) we shall just state the Theorem for later use.

Theorem 24. (Hairer, Lubich and Wanner (2002)). *Let $a : T \cup \{\varnothing\} \to \mathbb{R}$ be a mapping satisfying $a(\varnothing) = 1$ and let $b : T \cup \{\varnothing\} \to \mathbb{R}$ be arbitrary. Then the B-series $B(a, \mathbf{x})$ inserted into $B(b, \cdot)$ is again a B-series*

$$B(b, B(a, \mathbf{x})) = B(ab, \mathbf{x}),$$

where the group operation $ab(\tau)$ is the composition law from Section 4.3, i.e.,

$$ab(\tau) = \sum_{\theta \in OST(\tau)} b(\theta) a(\tau \setminus \theta) \quad with \quad a(\tau \setminus \theta) = \prod_{\delta \in \tau \setminus \theta} a(\delta)$$

Example 25. The composition rules for the trees of order ≤ 4 are

$ab(\bullet) = b(\varnothing) \cdot a(\bullet) + b(\bullet)$

$ab(\tau_2) = b(\varnothing) \cdot a(\tau_2) + b(\bullet) \cdot a(\bullet) + b(\tau_2)$

$ab(\tau_3^{(1)}) = b(\varnothing) \cdot a(\tau_3^{(1)}) + b(\bullet) \cdot a(\bullet)^2 + 2b(\tau_2) \cdot a(\bullet) + b(\tau_3^{(1)})$

$ab(\tau_3^{(2)}) = b(\varnothing) \cdot a(\tau_3^{(2)}) + b(\bullet) \cdot a(\tau_2) + b(\tau_2) \cdot a(\bullet) + b(\tau_3^{(2)})$

$ab(\tau_4^{(1)}) = b(\varnothing) \cdot a(\tau_4^{(1)}) + b(\bullet) \cdot a(\bullet)^3 + 3b(\tau_2) \cdot a(\bullet)^2 + 3b(\tau_3^{(1)}) \cdot a(\bullet) + b(\tau_4^{(1)})$

$ab(\tau_4^{(2)}) = b(\varnothing) \cdot a(\tau_4^{(2)}) + b(\bullet) \cdot a(\tau_3^{(1)}) + b(\tau_2) \cdot a(\bullet)^2 + 2b(\tau_3^{(2)}) \cdot a(\bullet) + b(\tau_4^{(2)})$

$ab(\tau_4^{(3)}) = b(\varnothing) \cdot a(\tau_4^{(3)}) + b(\bullet) \cdot a(\bullet) a(\tau_2) + b(\tau_2) \cdot a(\tau_2) + b(\tau_2) \cdot a(\bullet)^2 + b(\tau_3^{(1)}) \cdot a(\bullet) + b(\tau_3^{(2)}) \cdot a(\bullet) + b(\tau_4^{(3)})$

$ab(\tau_4^{(4)}) = b(\varnothing) \cdot a(\tau_4^{(4)}) + b(\bullet) \cdot a(\tau_3^{(2)}) + b(\tau_2) \cdot a(\tau_2) + b(\tau_3^{(2)}) \cdot a(\bullet) + b(\tau_4^{(4)})$

Before we proceed to the next chapter it will be useful to give the following definition regarding operations on trees which is due to Butcher (1972).

Definition 26. (Butcher Product). For two trees in T, $u = [u_1, u_2, ..., u_m]$ and $v = [v_1, v_2, ..., v_l]$ we denote

$$u \circ v = [u_1, ..., u_m, v]$$

and call this the *Butcher Product*.

Example 27. Given the trees $u = [[\bullet, \bullet], \bullet] = \tau_3^{(2)}$ and $v = [\bullet, \bullet] = \tau_2$,

$u \circ v = [[\bullet, \bullet], \bullet, [\bullet, \bullet]] =$ 

$v \circ u = [\bullet, \bullet[[\bullet, \bullet], \bullet]] =$ 

Notice that the operation $u \circ v$ is simply the joining of the root of u to the root of v by an upward leaving branch.

Chapter 5

Backward Error Analysis

Backward error analysis is one of the most powerful tools in analyzing the long-time qualitative behavior of a numerical method. The construction of the modified equation (which will be explained in the following section), a formal power-series of the step size, provides a lot of insight into the numerical method. The approach taken here is the same as that of Hairer, Lubich and Wanner (2006). All Theorems, Proofs and Definitions were provided by the aforementioned authors.

5.1 The Modified Differential Equation

We start with an autonomous system of ODE's, $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, and a numerical method, $\Phi_h(\mathbf{x})$, that produces the approximations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ...$ The idea of backward error analysis is to search for a modified differential equation, $\dot{\tilde{\mathbf{x}}} = \mathbf{f}_h(\tilde{\mathbf{x}})$, of the form

$$\dot{\tilde{\mathbf{x}}} = \mathbf{f}(\tilde{\mathbf{x}}) + h\mathbf{f}_2(\tilde{\mathbf{x}}) + h^2\mathbf{f}_3(\tilde{\mathbf{x}}) + ...,$$

whose exact solution at time nh is given by the numerical method, i.e.

$$\mathbf{x}_n = \tilde{\mathbf{x}}(nh),$$

and then study the differences between the vector fields $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}_h(\tilde{\mathbf{x}})$. This gives great insight into the qualitative behavior of the numerical method and global error. It is important to note that the vector field usually diverges and one has to truncate it suitably.

For the computation of the modified vector field we put $\mathbf{x} := \tilde{\mathbf{x}}(t)$ for a fixed t, and we expand the solution of the modified equation into a Taylor series

$$\tilde{\mathbf{x}}(t+h) = \mathbf{x} + h\dot{\tilde{\mathbf{x}}} + \frac{h^2}{2!}\ddot{\tilde{\mathbf{x}}} + ...$$

$$= \mathbf{x}+h(\mathbf{f}(\tilde{\mathbf{x}})+h\mathbf{f}_2(\tilde{\mathbf{x}})+h^2\mathbf{f}_3(\tilde{\mathbf{x}})+...)+\frac{h^2}{2!}(\mathbf{f}'(\tilde{\mathbf{x}})+h\mathbf{f}_2'(\tilde{\mathbf{x}})+h^2\mathbf{f}_3'(\tilde{\mathbf{x}})+...)(\mathbf{f}(\tilde{\mathbf{x}})+h\mathbf{f}_2(\tilde{\mathbf{x}})+h^2\mathbf{f}_3(\tilde{\mathbf{x}})+...)+...$$

We also assume that the numerical method $\Phi_h(\mathbf{x})$ can be expanded as

$$\Phi_h(\mathbf{x}) = \mathbf{x} + h\mathbf{f}(\mathbf{x}) + h^2\mathbf{d}_2(\mathbf{x}) + h^3\mathbf{d}_3(\mathbf{x}) + ...$$

The functions $\mathbf{d}_j(\mathbf{x})$ are known and are expressed in terms of $\mathbf{f}$ and its derivatives. $\Phi_h(\mathbf{x})$ becomes the explicit Euler method if $\mathbf{d}_j(\mathbf{x}) = 0$ for all $j \geq 2$. Now if we require $\mathbf{x}_n = \tilde{\mathbf{x}}(nh)$ for all n then we must have that $\tilde{\mathbf{x}}(t+h) = \Phi_h(\mathbf{x})$. Comparing like powers of h in the Taylor expansion for the solution of the modified equation and the expansion of the numerical method gives

$$\mathbf{f}_2(\mathbf{x}) = \mathbf{d}_2(\mathbf{x}) - \frac{1}{2!}\mathbf{f}'\mathbf{f}(\mathbf{x})$$

$$\mathbf{f}_3(\mathbf{x}) = \mathbf{d}_3(\mathbf{x}) - \frac{1}{3!}(\mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x}) + \mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x})) - \frac{1}{2!}(\mathbf{f}'\mathbf{f}_2(\mathbf{x}) + \mathbf{f}_2'\mathbf{f}(\mathbf{x})).$$

5.2 The Modified Equation and Trees

Theorem 14 shows that the numerical solution $\mathbf{x}_1 = \Phi_h(\mathbf{x}_0)$ of a Runge-Kutta method can be written as a B-series

$$\Phi_h(\mathbf{x}) = \mathbf{x} + h\mathbf{f}(\mathbf{x}) + h^2a([\bullet])(\mathbf{f}'\mathbf{f})(\mathbf{x}) + h^3(\frac{1}{2}a([\bullet,\bullet])\mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x}) + a([[\bullet]])\mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x})) + ...$$

We can exploit the structure of $\Phi_h(\mathbf{x})$ in order to get formulas for the functions $\mathbf{f}_2(\mathbf{x}), \mathbf{f}_3(\mathbf{x}), ...$ of the modified differential equation. Comparing like powers of h in the Runge-Kutta B-series with the expansion of the solution of the modified differential equation we obtain

$$\mathbf{f}_2(\mathbf{x}) = (a([\bullet]) - \frac{1}{2})(\mathbf{f}'\mathbf{f})(\mathbf{x})$$

$$\mathbf{f}_3(\mathbf{x}) = \frac{1}{2}(a([\bullet,\bullet]) - a([\bullet]) + \frac{1}{6})\mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x}) + (a([[\bullet]]) - a([\bullet]) + \frac{1}{3})\mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x}).$$

Continuing in this fashion, it is not hard to see that the general formula is given by

$$\mathbf{f}_j(\mathbf{x}) = \sum_{|\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(\mathbf{x})$$

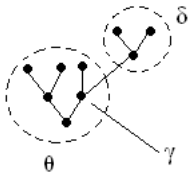


Figure 5.1: Splitting of an ordered tree ω into a sub-tree θ and $\{\delta\} = \omega \setminus \theta$

so that the modified equation becomes

$$\dot{\tilde{\mathbf{x}}} = \sum_{\tau \in T} \frac{h^{|\tau|-1}}{\sigma(\tau)} b(\tau) F(\tau)(\mathbf{x})$$

with $b(\bullet) = 1, b([\bullet]) = a([\bullet]) - \frac{1}{2}$, etc. Our goal is to find the recursion relation between the coefficients $b(\tau)$ and $a(\tau)$.

5.3 B-Series of the Modified Equation

In obtaining the recurrence relations for the coefficients $b(\tau)$ we follow the approach of Hairer (1999) and use the Lie-derivative of B-series, although, the relations were first given by Hairer (1994) and by Calvo and Sanz-Serna (1994). We shall once again make use of ordered trees and define, for a tree τ, a new set called the set of *splittings* as follows;

$$SP(\tau) = \{\theta \in OST(\tau) : \tau \setminus \theta \text{ consists of only one element}\}.$$

$OST(\tau)$ is the set of ordered sub-trees of τ as previously defined.

The following Lemma, and its accompanying proof, is given by Hairer and gives the Lie-derivative, which evaluates the change of a vector field along the flow of another, of a B-series.

Lemma 28. (Hairer (1999)). *(Lie-Derivative of B-Series). Let $b(\tau)$ (with $b(\varnothing) = 0$) and $c(\tau)$ be the coefficients of two B-Series, and let $\mathbf{x}(t)$ be a formal solution of the differential equation $h\dot{\mathbf{x}}(t) = B(b, \mathbf{x}(t))$. The Lie-derivative of the function $B(c, \mathbf{x})$with respect to the vector field $B(b, \mathbf{x})$ is again a B-series*

$$h\frac{d}{dt}B(c, \mathbf{x}(t)) = B(\partial_b c, \mathbf{x}(t)).$$

Its coefficients are given by $\partial_b c(\varnothing) = 0$ and for $|\tau| \geq 1$ by

$$\partial_b c(\tau) = \sum_{\theta \in SP(\tau)} c(\theta) b(\tau \setminus \theta).$$

Proof. It will be convenient to work with ordered trees $\omega \in OT$. Since $\nu(\tau)$ denotes the number of possible orderings of a tree $\tau \in T$, a sum $\sum_{\tau \in T} \cdot/\cdot$ becomes $\sum_{\omega \in OT} \frac{1}{\nu(\omega)} \cdot/\cdot$.

For the computation of the Lie-derivative of $B(c, \mathbf{x})$ we have to differentiate the elementary differential $F(\theta)(\mathbf{x}(t))$ with respect to t. Using the product rule, this gives $|\theta|$ terms, one for every vertex of θ. Then we need to evaluate this derivative in the direction of the B-series $B(b, \mathbf{x}(t))$, i.e. insert the series for $h\dot{\mathbf{x}}(t)$. This means that for a given tree θ in $B(c, \mathbf{x}(t))$, each tree δ appearing in $B(b, \mathbf{x}(t))$ is attached with a new branch to each vertex of θ that has been differentiated. Written out as formulas, this gives

$$h\frac{d}{dt}B(c, \mathbf{x}(t)) = \sum_{\theta \in OT \cup \{\varnothing\}} \frac{h^{|\theta|}c(\theta)}{\nu(\theta)\sigma(\theta)} \sum_{\gamma} \sum_{\delta \in OT} \frac{h^{|\delta|}b(\delta)}{\nu(\delta)\sigma(\delta)} F(\theta \circ_\gamma \delta)(\mathbf{x}(t)),$$

where $\sum_\gamma$ is the sum over all vertices of θ, and $\theta \circ_\gamma \delta$ is the ordered tree obtained by attaching the root of δ with a new branch to γ(figure 5.1). Now choose one of the $n(\gamma)+1$ possibilities of attaching δ to γ (because in terms of ordered trees, the ordering of the branches, and hence the order in which δ is attached to γ, matters), where $n(\gamma)$ denotes the number of upwards leaving branches of θ at the vertex γ. We now collect the terms with equal ordered trees $\omega = \theta \circ_\gamma \delta$, and notice that $\nu(\theta)\sigma(\theta) = \kappa(\theta)$. This gives

$$h\frac{d}{dt}B(c, \mathbf{x}(t)) = \sum_{\omega \in OT} h^{|\omega|} \left(\sum_{\theta \circ_\gamma \delta = \omega} \frac{c(\theta)b(\delta)}{(n(\gamma)+1)\kappa(\theta)\kappa(\delta)} \right) F(\omega)(\mathbf{x}(t)),$$

where $\sum_{\theta \circ_\gamma \delta = \omega}$ is over all triplets (θ, γ, δ) such that $\theta \circ_\gamma \delta = \omega$, and we have divided by the factor $n(\gamma)+1$ because we want to be able to change from sums over OT to sums over T. Because $\kappa(\omega) = \kappa(\theta)\kappa(\delta)(n(\gamma)+1)$, we get

$$h\frac{d}{dt}B(c, \mathbf{x}(t)) = \sum_{\omega \in OT} \frac{h^{|\omega|}}{\kappa(\omega)} \left(\sum_{\theta \circ_\gamma \delta = \omega} c(\theta)b(\delta) \right) F(\omega)(\mathbf{x}(t))$$

$$= \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \left(\sum_{\theta \in SP(\tau)} c(\theta)b(\tau \setminus \theta) \right) F(\tau)(\mathbf{x}(t)).$$

□

As an example of the Lie-derivative formula, consider the tree shown in figure 5.2 with five vertices, along with all possible splittings of this tree. Note that θ may be the empty tree $\varnothing$ and that always $|\delta| \geq 1$. The tree ω may be obtained in several ways: (i) differentiation of $F(\varnothing)(\mathbf{x}) = \mathbf{x}$ and adding $F(\omega)(\mathbf{x})$ as argument, (ii) differentiation of the factor corresponding to the root in $F(\theta)(\mathbf{x}) = \mathbf{f}''(\mathbf{f}, \mathbf{f})(\mathbf{x})$ and adding $F([\bullet])(\mathbf{x}) = (\mathbf{f}'\mathbf{f})(\mathbf{x})$, (iii) differentiation of all $\mathbf{f}$'s in

Figure 5.2: Splittings of an ordered tree with 5 vertices (example taken from [10])

$F(\theta)(\mathbf{x}) = \mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f})(\mathbf{x})$ and adding $F(\bullet)(\mathbf{x}) = \mathbf{f}(\mathbf{x})$ and finally, (iv) differentiation of the factor for the root in $F(\theta)(\mathbf{x}) = \mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f})(\mathbf{x})$ and adding $F(\bullet)(\mathbf{x}) = \mathbf{f}(\mathbf{x})$. This shows that

$$\partial_b c([[\bullet,\bullet],\bullet,\bullet]) = c(\varnothing)b([[\bullet,\bullet],\bullet,\bullet])) + c([\bullet,\bullet])b([\bullet]) + c([\bullet,\bullet,\bullet])b(\bullet) + 2c([[\bullet,\bullet],\bullet])b(\bullet).$$

The formulas for $\partial_b c$ are shown below for all trees up to order 3:

$$\partial_b c(\bullet) = c(\varnothing)b(\bullet)$$

$$\partial_b c([\bullet]) = c(\varnothing)b([\bullet]) + c(\bullet)b(\bullet)$$

$$\partial_b c([\bullet,\bullet]) = c(\varnothing)b([\bullet,\bullet]) + 2c([\bullet])b(\bullet)$$

Lemma 28 will allow us to determine the recursion formula for the coefficients $b(\tau)$ of the modified differential equation.

Theorem 29. (Hairer, Lubich and Wanner (2002)). *If the method $\Phi_h(\mathbf{x})$ is given as in section 5.2, the functions $\mathbf{f}_j(\mathbf{x})$ of the modified differential equation satisfy the relation given in section 5.2, where the coefficients $b(\tau)$ are recursively defined by $b(\varnothing) = 0$, $b(\bullet) = 1$ and*

$$b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|} \frac{1}{j!}\partial_b^{j-1} b(\tau),$$

where ∂_b^{j-1} is the $(j-1)$th iterate of the Lie-derivative ∂_b defined in Lemma 28.

Proof. The right-hand side of the modified differential equation is the B-series $B(b,\tilde{\mathbf{x}}(t))$ divided by h. It therefore follows from an iterative application of Lemma 28 that

$$h^j \tilde{\mathbf{x}}^{(j)}(t) = B(\partial_b^{j-1} b, \tilde{\mathbf{x}}(t)),$$

so that the Taylor series expansion $\tilde{\mathbf{x}}(t+h) = \mathbf{x} + B(\sum_{j\geq 1} \frac{1}{j!}\partial_b^{j-1} b, \mathbf{x})$, where $\mathbf{x} := \tilde{\mathbf{x}}(t)$. Since we have to determine the coefficients $b(\tau)$ in such a way that $\tilde{\mathbf{x}}(t+h) = \Phi_h(\mathbf{x}) = B(a,\mathbf{x})$, a comparison of the two B-Series gives $\sum_{j\geq 1} \frac{1}{j!}\partial_b^{j-1} b(\tau) = a(\tau)$. This proves the statement, because $\partial_b^0 b(\tau) = b(\tau)$ for $\tau \in T$ so that $b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|} \frac{1}{j!}\partial_b^{j-1} b(\tau)$, and $\partial_b^{j-1} b(\tau) = 0$ for $j > |\tau|$(as a consequence of $b(\varnothing) = 0$) so that $b(\tau)$ depends only on $b(\omega)$ for $|\omega| < |\tau|$.□

The table below shows the formula from Theorem 29 for trees up to order 3 [10].

$\tau = \bullet$	$b(\bullet) = a(\bullet)$
$\tau = [\bullet]$	$b([\bullet]) = a([\bullet]) - \frac{1}{2}b(\bullet)^2$
$\tau = [\bullet, \bullet]$	$b([\bullet, \bullet]) = a([\bullet, \bullet]) - \frac{1}{2}b([\bullet])b(\bullet) - \frac{1}{6}b(\bullet)^3$
$\tau = [[\bullet]]$	$b([[\bullet]]) = a([[\bullet]]) - \frac{1}{2}b([\bullet])b(\bullet) - \frac{1}{6}b(\bullet)^3$

5.4 Elementary Hamiltonians

Recall from chapter 1 that a Hamiltonian system is given by

$$\dot{p} = -H_q(p,q) \quad \dot{q} = H_p = (p,q),$$

where the *Hamiltonian* $H(p_1, p_2, ..., p_d, q_1, q_2, ..., q_d)$ represents the total energy; q_i are the position coordinates and p_i the momenta for $i = 1, 2, ..., d$, with d the number of degrees of freedom; H_p and H_q are the vectors of partial derivatives. With the notation $\mathbf{x} = (p, q)$ and the matrix $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$ we may write the Hamiltonian system in the form

$$\dot{\mathbf{x}} = J^{-1}\nabla H(\mathbf{x}),$$

where $\nabla H(\mathbf{x}) = H'(\mathbf{x})^T$.

We shall now consider Hamiltonian systems of the above form. According to Theorem 3.1 of Hairer, Lubich and Wanner (Pg 293, 2002), if $\dot{\mathbf{x}} = J^{-1}\nabla H(\mathbf{x})$, then the modified differential equation is again Hamiltonian (i.e. a Hamiltonian system). More precisely, there exist smooth functions $H_j : \mathbb{R}^{2d} \to \mathbb{R}$ for $j = 2, 3, ...$ such that $\mathbf{f}_j(\mathbf{x}) = J^{-1}\nabla H_j(\mathbf{x})$ and therefore the modified differential equation has the form

$$\dot{\tilde{\mathbf{x}}} = J^{-1}\nabla \tilde{H}(\mathbf{x}),$$

where

$$\tilde{H}(\mathbf{x}) = H(\mathbf{x}) + hH_2(\mathbf{x}) + h^2H_3(\mathbf{x}) + ...$$

The question is, can we find explicit formulas for $\tilde{H}(\mathbf{x})$?

As an example (taken from Hairer, Lubich and Wanner (2006)), consider the implicit midpoint rule. Written as a B-series, its coefficients are $a(\tau) = 2^{1-|\tau|}$ which is easily seen from differentiating the vector field and using the chain rule. Using this information, we are able to calculate

the coefficient functions, $\mathbf{f}_j(\mathbf{x}) = \sum_{|\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(\mathbf{x})$ using the recursive formula for $b(\tau)$, for the modified differential equation. Firstly, $\mathbf{f}_2(\mathbf{x}) = b([\bullet])\mathbf{f}'\mathbf{f}(\mathbf{x}) = 0$ since $b([\bullet]) = a([\bullet]) - \frac{1}{2}b(\bullet)^2 = \frac{1}{2} - \frac{1}{2} = 0$ and $\mathbf{f}_3(\mathbf{x}) = \frac{1}{2}b([\bullet,\bullet])\mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x}) + b([[\bullet]])\mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x}) = \frac{1}{24}\left(2\mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x}) - \mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x})\right)$ since $b([\bullet,\bullet]) = a([\bullet,\bullet]) - b([\bullet])b(\bullet) - \frac{1}{3}b(\bullet)^3 = \frac{-1}{12}$ and $b([[\bullet]]) = a([[\bullet]]) - b([\bullet])b(\bullet) - \frac{1}{6}b(\bullet)^3 = \frac{2}{24}$. Consider the function

$$H_3(\mathbf{x}) = -\frac{1}{24}H''(\mathbf{x})(J^{-1}\nabla H(\mathbf{x}), J^{-1}\nabla H(\mathbf{x})).$$

Since $\mathbf{f}(\mathbf{x}) = J^{-1}\nabla H(\mathbf{x})$, we have that $\mathbf{f}'(\mathbf{x}) = (J^{-1})^2\nabla^2 H(\mathbf{x}) = -H''(\mathbf{x})$ and $\mathbf{f}''(\mathbf{x}) = (J^{-1})^3\nabla^3 H(\mathbf{x}) = H'''(\mathbf{x})$. Therefore,

$$J^{-1}\nabla H_3(\mathbf{x}) =$$
$$-\tfrac{1}{24}(J^{-1})^3\nabla^3 H(\mathbf{x})(J^{-1}\nabla H(\mathbf{x}), J^{-1}\nabla H(\mathbf{x})) - \tfrac{2}{24}(J^{-1})^2\nabla^2 H(\mathbf{x})((J^{-1})^2\nabla^2 H(\mathbf{x}), J^{-1}\nabla H(\mathbf{x}))$$
$$= -\tfrac{1}{24}\mathbf{f}''(\mathbf{f},\mathbf{f}) + \tfrac{2}{24}\mathbf{f}'\mathbf{f}'\mathbf{f}$$
$$= \mathbf{f}_3(\mathbf{x}).$$

We have just shown that $\mathbf{f}_3(\mathbf{x}) = J^{-1}\nabla H_3(\mathbf{x})$ and have found an explicit formula for the Hamiltonian corresponding to the vector field $\mathbf{f}_3(\mathbf{x})$. The above computations lead to expressions previously introduced (although in a different context) by Sanz-Serna and Abia (1991).

Definition 30. (Elementary Hamiltonians). For a given smooth function $H : D \to \mathbb{R}$ (with open $D \subset \mathbb{R}^{2d}$) and for $\tau \in T$ we define the *elementary Hamiltonian* $H(\tau) : D \to \mathbb{R}$ by

$$H(\bullet)(\mathbf{x}) = H(\mathbf{x}), \qquad H(\tau)(\mathbf{x}) = H^{(m)}(F(\tau_1)(\mathbf{x}), ..., F(\tau_m)(\mathbf{x}))$$

for $\tau = [\tau_1, ..., \tau_m]$. Here, $F(\tau_i)(\mathbf{x})$ are elementary differentials corresponding to $\mathbf{f}(\mathbf{x}) = J^{-1}\nabla H(\mathbf{x})$.

Notice that the expression for $H_3(\mathbf{x})$ is simply the elementary Hamiltonian corresponding to the tree $[\bullet,\bullet]$.

Hairer, Lubich and Wanner (2006) give the following Theorem, for the case when a symplectic method is applied to a Hamiltonian system, regarding the coefficients $b(\tau)$ of the coefficient functions $\mathbf{f}_j(\mathbf{x})$ of the modified differential equation. We shall only quote the Theorem here and refer the reader to Hairer, Lubich and Wanner (2006), chapter VI.7 for a complete characterization of symplectic methods and the results on which the proof of this Theorem relies.

Theorem 31. (Faou, Pham and Hairer (2004)). *Suppose that for all Hamiltonians $H(\mathbf{x})$ the modified vector field (of the modified differential equation), truncated after an arbitrary power of h, is (locally) Hamiltonian. Then,*

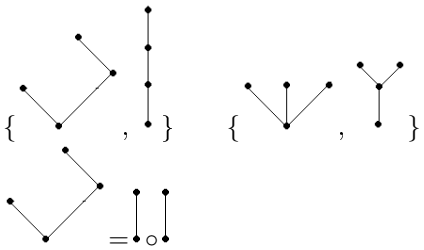


Figure 5.3: The superfluous (left) and non-superfluous (right) free trees of order 4

$$b(u \circ v) + b(v \circ u) = 0 \qquad \forall \quad u, v \in T.$$

The aim is to prove that, for symplectic methods applied to Hamiltonian systems, the coefficient functions, $\mathbf{f}_j(\mathbf{x})$, of the modified differential equation satisfy $\mathbf{f}_j(\mathbf{x}) = J^{-1}\nabla H_j(\mathbf{x})$, where $H_j(\mathbf{x})$ is a linear combination of elementary Hamiltonians.

Lemma 32. (Faou, Pham and Hairer (2004)). *Elementary Hamiltonians satisfy*

$$H(u \circ v)(\mathbf{x}) + H(v \circ u)(\mathbf{x}) = 0 \qquad \forall\, u, v \in T.$$

In particular, $H(u \circ u)(\mathbf{x}) = 0$ *for all* $u \in T$.

Proof. This follows from the fact that for $u = [u_1, ..., u_m] \in T$ and for $v \in T$ we have $H(u \circ v) = H^{(m+1)}(F(u_1), ..., F(u_m), F(v)) = F(v)^T(\nabla H)^{(m)}(F(u_1), ..., F(u_m)) = F(v)^T JF(u) = (F(u)^T J^T F(v))^T = -(F(u)^T JF(v))^T = -(H(v \circ u))^T = -H(v \circ u).\square$

Notice that the Butcher product of two trees $u, v \in T$ induces an equivalence relation on T, the smallest equivalence relation satisfying $u \circ v \sim v \circ u$ for every $u, v \in T$ ([9]). That is, $u \circ v$ and $v \circ u$ have the same graph and only differ in the root position. For two trees θ and τ, $\kappa(\theta, \tau)$ denotes the number of times the root must be shifted in order to obtain θ from τ. As an example, consider the two equivalent trees $\theta =$  , $\tau =$  . Then $\kappa(\theta, \tau) = 2$. Each equivalence class is called a *free tree* and the set of free trees of order n is denoted by FT^n. Let $\pi(\tau)$ be the free tree (equivalence class) to which $\tau \in T$ belongs. A free tree is called *superfluous* if it contains an element of the form $u \circ u$ for some $u \in T$. All other free trees are called $non - superfluous$ (Celledoni, McLachlan, Owren and Quispel (2010)). We denote the set of non-superfluous free trees by FT_{NS} Figure 5.3 displays the superfluous and non-superfluous free trees of order 4.

The following Lemma is due to Faou, Hairer and Pham (2004) and gives a description of the elementary Hamiltonian in terms of elementary differentials (as in our example of $\mathbf{f}_3(\mathbf{x})$ and $H_3(\mathbf{x})$).

Lemma 33. *For a tree τ belonging to a non-superfluous free-tree we have*

$$J^{-1}\nabla H(\tau)(\mathbf{x}) = \sigma(\tau)\sum_{\theta\sim\tau}\frac{(-1)^{\kappa(\theta,\tau)}}{\sigma(\theta)}F(\theta)(\mathbf{x}).$$

Proof. *We wish to compute $J^{-1}\nabla H(\tau)(\mathbf{x})$. The expression $H(\tau)(\mathbf{x})$ contains $|\tau|$ factors corresponding to the vertices of τ, each of which is differentiated by the product rule. Differentiation of $H^{(m)}(\mathbf{x})$ and premultiplication by the matrix J^{-1}yields $F(\tau)(\mathbf{x})$ (as in the example of $J^{-1}\nabla H_3(\mathbf{x})$ and corresponds to the differentiation of the root of τ. In order to differentiate the other factors (vertices of τ) we bring the vertex that is to be differentiated down to the root. In light of Lemma 32 this only multiplies $H(\tau)(\mathbf{x})$ by $(-1)^{\kappa(\theta,\tau)}$, and shows that differentiating the corresponding factor yields $F(\theta)(\mathbf{x})$. Since τ is a member of a non-superfluous free tree, the number of possibilities of obtaining θ from τ by exchanging roots is equal to $\frac{\sigma(\tau)}{\sigma(\theta)}$.* $\square$

We are now able to give an explicit formula for the Hamiltonian of the modified differential equation provided that the numerical method can be written as a B-series. This Theorem is due to Faou, Hairer and Pham (2004).

Theorem 34. (Faou, Pham and Hairer (2004)). *Consider a numerical method that can be written as a B-series, and that is symplectic for every Hamiltonian system $\dot{\mathbf{x}} = J^{-1}\nabla H(\mathbf{x})$. Its modified differential equation is then Hamiltonian with*

$$\tilde{H}(\mathbf{x}) = H_1(\mathbf{x}) + hH_2(\mathbf{x}) + h^2H_3(\mathbf{x}) + ...,$$

where

$$H_j(\mathbf{x}) = \sum_{|\tau|=j,\tau\ non-superfluous}\frac{b(\tau)}{\sigma(\tau)}H(\tau)(\mathbf{x}),$$

the coefficients $b(\tau)$ are those of Theorem 29 and the sum is over only one representative of each non-superfluous free tree

Proof. Apply the method $\Phi_h(\mathbf{x})$ (written as a B-series as in section 5.1) to a Hamiltonian system so that the modified differential equation is, again, (locally) Hamiltonian of the form

$$\tilde{H}(\mathbf{x}) = H_1(\mathbf{x}) + hH_2(\mathbf{x}) + h^2H_3(\mathbf{x}) + ...$$

Following the same procedure as in section 5.1 we have that

$$H_j(\mathbf{x}) = \sum_{|\tau|=j}\frac{b(\tau)}{\sigma(\tau)}F(\tau)(\mathbf{x}).$$

Since the modified differential equation is Hamiltonian, it follows from Theorem 31 that the coefficients $b(\tau)$ satisfy $b(u \circ v) + b(v \circ u) = 0$. This implies that $b(\theta) = (-1)^{\kappa(\tau,\theta)}$ whenever $\theta \sim \tau$ and that $b(\omega) = 0$ for all trees belonging to a superfluous free-tree. Putting this information into the expression for H_j

$$H_j(\mathbf{x}) = \sum_{|\theta|=j,\, \theta\ non-superfluous,\, \theta\sim\tau} \frac{b(\theta)}{\sigma(\tau)}(-1)^{\kappa(\tau,\theta)}F(\theta)(\mathbf{x})$$

$$= \sum_{|\theta|=j,\, \theta\ non-superfluous,\, \theta\sim\tau} \frac{b(\theta)}{\sigma(\theta)}(-1)^{\kappa(\tau,\theta)}\frac{\sigma(\theta)}{\sigma(\tau)}F(\theta)(\mathbf{x})$$

$$= \sum_{|\theta|=j,\, \theta\ non-superfluous} \frac{b(\theta)}{\sigma(\theta)}H(\theta)(\mathbf{x}),$$

(*by an application of Lemma* 33 *and the fact that the vector field is Hamiltonian*),

and the final sum is over only one representative of each non-superfluous free tree after applying Lemma 33 (that is, not every member of each non-superfluous free tree appears in the sum, only one member from each non-superfluous free-tree).□

Faou, Hairer and Pham (2004) also note that the elementary Hamiltonians only depend on derivatives of $H(\mathbf{x})$ and therefore the modified Hamiltonian is *globally* defined. They also obtain a simple corollary from Theorem 34 and 31.

Corollary 35. (Faou, Pham and Hairer (2004)). *The differential equation* $h\dot{\mathbf{x}} = B(b, \mathbf{x})$ *with* $b(\varnothing) = 0$ *is Hamiltonian for all vector fields* $\mathbf{f}(\mathbf{x}) = J^{-1}\nabla H(\mathbf{x})$*, if and only if*

$$b(u \circ v) + b(v \circ u) = 0 \quad \forall \quad u, v \in T.$$

5.5 First Integrals Close to the Hamiltonian

Given a Hamiltonian ODE, $\dot{\mathbf{x}} = J^{-1}\nabla H(\mathbf{x})$, and a symplectic numerical method, Φ_h, that admits a B-series, we have seen, from Theorem 34, that the modified differential equation, based on the Hamiltonian ODE, is Hamiltonian with a function of the form

$$H(c, y) = \sum_{\tau \in FT_{NS}} \frac{h^{|\tau|-1}}{\sigma(\tau)} c(\tau) H(\tau)(\mathbf{x}),$$

with $c(\tau) = b(\tau)$ and the sum being over only one representative from each non-superfluous free tree (that is, not every member of each non-superfluous free tree appears in the sum, only one member from each non-superfluous free-tree). The problem is then reset and we ask the question whether for non-symplectic methods a function of the above form, with $c(\tau)$ not necessarily equal to $b(\tau)$, can be a first integral of the modified differential equation ([9]). This question is answered in the affirmative by the following Lemma and corollary found in [9].

Lemma 36. (Faou, Pham and Hairer (2004)). *Let $\mathbf{x}(t)$ be a solution of the modified differential equation which can be written as $h\dot{\mathbf{x}} = B(b, \mathbf{x}(t))$. We then have*

$$\frac{d}{dt} H(c, \mathbf{x}(t)) = H(\delta_b c, \mathbf{x}(t))$$

where $\delta_b c(\bullet) = 0$ and, for $\tau \in FT_{NS}$, with $|\tau| > 1$,

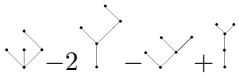
$$\delta_b c(\tau) = \sum_{\theta \sim \tau} (-1)^{\kappa(\tau,\theta)} \frac{\sigma(\tau)}{\sigma(\theta)} \sum_{\omega \in FT_{NS} \cap SP(\theta)} c(\omega) b(\theta \setminus \omega).$$

The first sum is over all trees θ that are equivalent to τ and the second sum is over all splitting of θ (as defined in section 5.3) such that the sub-tree containing the root of θ belongs to a non-superfluous free-tree.

Corollary 37. (Faou, Pham and Hairer (2004)). *The function $H(c, \mathbf{x})$ is a first integral of the differential equation for every $H(\mathbf{x})$ if and only if*

$$\delta_b c(\tau) = 0 \qquad \forall\, \tau \in FT_{NS}.$$

The proof of Lemma 36 is almost identical to the proof of Lemma 28 and a proof of corollary 37 may be found in [9]. Corollary 37 gives us a linear system to solve in order to find conditions on the coefficients $b(\tau)$ so that $H(c, y)$ is a first integral of the modified differential equation given in section 5.2. As example, let us calculate $\delta_b c$ for the tree $\tau =$. The first sum in the formula tells us to write down the sum of all trees related to τ by shifting the root and multiplying by the appropriate symmetry coefficients. We obtain



We need to find the splittings of each of these trees such that the sub-tree containing the original root of each tree belongs to a non-superfluous free-tree. Let ω denote the sub-tree containing the original root. We obtain the following splittings for each tree,

 : $(\omega =$ $, \tau \setminus \omega = \bullet)$, $(\omega =$ $, \tau \setminus \omega =$ $)$,

 : $(\omega =$ $, \tau \setminus \omega = \bullet)$, $(\omega = \bullet, \tau \setminus \omega =$ $)$

 : $(\omega =$ $, \tau \setminus \omega = \bullet)$,

 : $(\omega = \bullet, \tau \setminus \omega =$ $)$.

Putting this information into the formula we get

$\delta_b c(\tau) = c($$)b(\bullet) + c($$)b($$) - 2(c($$)b(\bullet) + c(\bullet)b($$)) - c($$)b(\bullet) + c(\bullet)b($$)$. Below are the formulas for $\delta_b c(\tau)$ for all non-superfluous trees up to order 6 (first given in [9]). Note that we

only need one representative from each free-tree since any tree members of a free-tree will give the same formula (up to sign).

$\delta_b c($$) = -2c(\bullet)b($  $)$

$\delta_b c($$) = 3c($  $)b(\bullet) - 3c(\bullet)b($  $) - 6c($  $)b(\bullet)$

$\delta_b c($$) = 4c($$)b(\bullet) - 4c(\bullet)b($  $) - 12c($$)b(\bullet)$

$\delta_b c($$) = c($$)b(\bullet) + c($  $)b($  $) - 2(c($$)b(\bullet) + c(\bullet)b($  $)) - c($$)b(\bullet) + c(\bullet)b($$)$

$\delta_b c($$) = 2c(\bullet)b($  $) - 2c($  $)b($  $)$

$\delta_b c($$) = 5c($  $)b(\bullet) - 5c(\bullet)b($  $) - 20c($  $)b(\bullet)$

$\delta_b c($$) = 3c($$)b(\bullet) + c($  $)b(\bullet) + c($$)b($  $) - 3c(\bullet)b($$) + c(\bullet)b($$)$

$\delta_b c($$) = 2c($  $)b(\bullet) + c($  $)b(\bullet) - c(\bullet)b($$) + 2c(\bullet)b($  $)$

$\delta_b c($$) = 2c($  $)b(\bullet) - c($$)b(\bullet) - c($$)b($$) - c($$)b($$) - c($$)b($$)+2c(\bullet)b($$) + c(\bullet)b($$).$

Corollary 37 gives us a linear system of equations to solve in order to find conditions on the modified differential equation (i.e. conditions on the $b(\tau)$ coefficients) for $H(c, \mathbf{x})$ to be a first integral. For completeness, Faou, Hairer and Pham (2004) solve this system for trees up to order 6. This is replicated below. Consider a consistent method, i.e. $b(\bullet) = 1$ and search for a first integral $H(c, \mathbf{x})$ close to the Hamiltonian, i.e. $c(\bullet) = 1$.

$|\tau| = 3$: The condition $\delta_b c(\tau) = 0$ for $\tau =$ gives $b($$) = 0$ and we conclude that whichever method has $H(c, \mathbf{x})$ as a first integral will have to be of order 2.

$|\tau| = 4$: The condition is satisfied by setting $b($$) = 3c($$) - 6c($$)$.

$|\tau| = 5$: The last of the order 4 conditions gives $b($$) = 0$. The first order 4 condition gives $c($$) = -b($$) - 3c($$)$ and substituting this expression into the second order 4 condition we find that we must satisfy

$b($$) + b($$) - 2b($$) = 0.$

It is important to note that this equation is satisfied by symplectic methods, i.e. methods whose coefficients satisfy $b(u \circ v) + b(v \circ u) = 0$.

$|\tau| = 6$: Performing similar substitutions for the order 5 conditions we find that the modified differential equation has $H(c, \mathbf{x})$ if and only if the following conditions is satisfied;

$5b($$) + 5b($$) + 6b($$) + 6b($$) - 12b($$) + 3b($$) - 15b($$) = 0.$

This condition is also satisfied by every symplectic method. In general, the conditions on the modified differential equation for $H(c, \mathbf{x})$ to be a first integral are satisfied by all symplectic methods.

This is encapsulated by the following Lemma of Faou, Hairer and Pham's (2004).

Lemma 38. (Faou, Pham and Hairer (2004)). *Let $c(\tau)$, τ non-superfluous, be given and assume $c(\bullet) = 1$ and $b(\varnothing) = 0$. Then, for fixed $b(\bullet)$, the linear system , $\delta_b c(\tau) = 0$ for $b(\tau)$, $\tau \in T$ has at most one solution satisfying $b(u \circ v) + b(v \circ u) = 0$ for all u, $v \in T$.*

Theorem 39. (Chartier, Faou and Murua 2005). *The only symplectic method (as a B-series) that conserves the Hamiltonian for arbitrary $H(\mathbf{x})$ is the exact flow of the differential equation.*

Proof. If the method conserves exactly the Hamiltonian, we have $H(c, \mathbf{x})$, $\delta_b c(\tau) = 0$ with $c(\bullet) = 1$ and $c(\tau) = 0$ for all other trees belonging a non-superfluous free-tree (this is because $H(c, \mathbf{x})$ is a perturbation of the Hamiltonian and we want the exact Hamiltonian to be conserved). By the uniqueness statement of Lemma 38 and the symplecticity of the method (corollary 35), we obtain $b(\tau) = 0$ for $|\tau| > 1$. Consequently, no perturbation is permitted in the modified differential equation of the method.□

Thus, we find that no B-series of the modified differential equation can be simultaneously symplectic and energy-preserving.

Chapter 6

The Algebraic Structure of B-Series

In this chapter we are interested in the algebraic structure of B-series. In particular, we wish to better understand certain properties of the elementary differentials (rooted trees) when the vector field of an ODE is Hamiltonian, and classify the elementary differentials according to these properties. Chapter 5 gave conditions for the B-series of the modified differential equation to either be Hamiltonian or energy-preserving. These conditions have implications for the elementary differentials and we shall see that certain linear combinations of elementary differentials are Hamiltonian and certain linear combinations have first integral H, where H is the Hamiltonian (although these are not the only properties elementary differentials may posses). The linear combinations of elementary differentials define linear subspaces of the vector space over rooted trees (with base field $\mathbb{R}$). These subspaces inherit the linear- and Lie-algebraic structure induced by the elementary differentials even when the original vector field $\mathbf{f}$, of the ODE, is "forgotten" and we work only with rooted trees (Celledoni, McLachlan, Owren and Quispel 2010). This chapter will explore and understand these subspace, their dimensions and their annihilators.

6.1 Energy-Preserving and Hamiltonian B-Series

Recall that T is the set of all rooted trees and T^n is the set of rooted trees of order n (i.e. trees having n vertices). Let

$$\mathcal{T}^n = \text{span}(T^n),$$

$$\mathcal{T} = \text{span}(T) = \oplus_{n=1}^{\infty} \mathcal{T}^n.$$

$\mathcal{T}$ is the real vector space generated by the set T. It consists of finite linear combinations of elements of T. We use the notation $|\cdot|$ to denote the order of a tree and the cardinality of a set, therefore $\dim(\mathcal{T}^n) = |T^n|$ (Celledoni, McLachlan, Owren and Quispel 2010).

For Hamiltonian systems, we shall consider the more general case where $\dot{\mathbf{x}} = \Omega^{-1}\nabla H$ with Ω a constant, anti-symmetric, invertible $d \times d$ matrix. The Hamiltonian vector field is defined by $\mathbf{f} = X_H = \Omega^{-1}\nabla H$. We extend the elementary differential F to $\mathcal{T}$ by linearity, e.g. $F(a\bullet + b[\bullet]) = aF(\bullet) + bF([\bullet])$. Consequently, the coefficients $b(\tau)$ are linear in τ. Also, a *forest* is an unordered, possibly empty, collection of trees where each tree can appear an arbitrary number of times. For example, $t_1 = (\bullet, \bullet, [\bullet, \bullet])$ is a forest so that $t = [t_1] = [\bullet, [\bullet, \bullet]]$. we denote the set of all forests by $\bar{T}$. Define a map $B_- : T \to \bar{T}$ by $B_-(\tau) = (t_1, ..., t_m)$ for a tree $\tau = [t_1, ..., t_m] \in T$.

The following definition of the two sub-spaces of interest comes from (Celledoni, McLachlan, Owren and Quispel 2010).

Definition 40. The energy-preserving subspace (of order n) is defined by

$$\mathcal{T}_H^n = \{t \in \mathcal{T}^n : F(t) \textit{ has first integral } H \textit{ when } f = \Omega^{-1}\nabla H\}.$$

The Hamiltonian subspace (of order n) is defined by

$$\mathcal{T}_\Omega^n = \{t \in \mathcal{T}^n : F(t) \textit{ is Hamiltonian w.r.t } \Omega \textit{ when } f = \Omega^{-1}\nabla H\},$$

with

$$\mathcal{T}_H = \oplus_{n=1}^{\infty} \mathcal{T}_H^n$$

$$\mathcal{T}_\Omega = \oplus_{n=1}^{\infty} \mathcal{T}_\Omega^n$$

In section 5.4 we calculated the coefficient functions of the modified differential equation using the implicit midpoint rule. We found that $\mathbf{f}_3(\mathbf{x}) = \frac{1}{24}\left(2\mathbf{f}'\mathbf{f}'\mathbf{f}(\mathbf{x}) - \mathbf{f}''(\mathbf{f},\mathbf{f})(\mathbf{x})\right)$ and this is a Hamiltonian combination of trees with Hamiltonian

$$H_3(\mathbf{x}) = -\frac{1}{24}H''(\mathbf{x})(J^{-1}\nabla H(\mathbf{x}), J^{-1}\nabla H(\mathbf{x})).$$

Therefore, $\mathbf{f}_3(\mathbf{x})$, written in terms of rooted trees, is an element of the subspace $\mathcal{T}_\Omega^n$. In section 5.4, Lemma 33 gives the linear combination of elementary differentials that are Hamiltonian. Written in terms of rooted trees, the Hamiltonian linear combination of trees is given by

$$X_t = \sigma(t) \sum_{\theta \sim t} \frac{(-1)^{\kappa(\theta,\tau)}}{\sigma(\theta)} \theta,$$

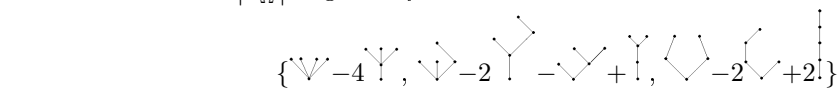
for a non-superfluous tree t, where X_t denotes the Hamiltonian combination of trees associated

with t. Since any two trees belonging to the same free tree will give the same Hamiltonian combination of trees (ignoring sign), a basis for the linear subspace of Hamiltonian trees is given X_t, for one representative t from each non-superfluous free-tree. Thus,

$$|\mathcal{T}_\Omega^n| = \begin{cases} |FT^n| & n \text{ odd} \\ |FT^n| - \left|T^{\frac{n}{2}}\right| & n \text{ even} \end{cases}$$

because for odd n there cannot exist any trees of the form $u \circ u$, $u \in T$ and for n even, an element of the form $u \circ u$ is given by taking a tree of order $\frac{n}{2}$ and taking the butcher product of this tree with itself. The number of such superfluous free-trees (with order n) is given by $\left|T^{\frac{n}{2}}\right|$.

Example 41. The non-superfluous free-trees of order 5 are  . Therefore, a basis for $\left|\mathcal{T}_\Omega^5\right|$ is given by



If the B-series of the modified differential equation is Hamiltonian then, since the numerical method is the exact solution of the modified differential equation and the flow of Hamiltonian vector fields are symplectic, we know that the numerical method is symplectic. But, as we shall see later, the Hamiltonicity of the modified differential equation does not mean the numerical method will also preserve the energy of the system.

Linear transformations from a vector space V to the base field F are very important. The next two definitions are due to Roman (2008).

Definition 42. Let V be a vector space over F. A linear transformation f, defined on V, whose value lies in the base-field F is called a *linear functional* on V. The vector space of all linear functionals on V is denoted by V^* and is called the *algebraic dual space* of V.

The functionals $f \in V^*$ are defined on vectors in V, but we may also define f on subsets M of V by letting

$$f(M) = \{f(v) : \ v \in M\}.$$

Definition 43. Let M be a non-empty subset of a vector space V. The *annihilator*, $Ann(M)$, of M is

$$Ann(M) = \{f \in V^* : \ f(M) = \{0\}\}.$$

The annihilator consists of all functionals that send every vector in M to zero. $Ann(M)$ is always a subspace of V^* even when M is not a subspace of V.

Let V be a vector space with basis $\mathcal{B} = \{v_i : i \in I\}$. For each $i \in I$, we can define a linear functional $v_i^* \in V^*$ by the orthogonality condition

$$\langle v_i^*, v_j \rangle = v_i^*(v_j) = \delta_{i,j}$$

where $\delta_{i,j}$ is the Kronecker delta function defined by

$$\delta_{i,j} = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases}$$

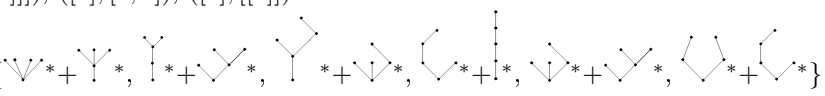
The set $\mathcal{B}^* = \{v_i^* : i \in I\}$ is clearly linearly independent (Roman (2008)).

In finite dimensional vector spaces it is customary to identify the dual of the algebraic dual space, or double dual V^{**}, with V and to think of the elements of V^{**} simply as vectors in V.

The most natural space of linear functionals defined on $\mathcal{T}$ is the space of coefficients, $b(\tau)$, where $b : \mathcal{T} \to \mathbb{R}$. We set $b(\tau) := \tau^* \in V^*$and use the pairing $\left\langle \tau^*, \frac{1}{\sigma(\theta)}\theta \right\rangle = \delta_{\tau,\theta}$, where $\delta_{\tau,\theta}$ is the Kronecker delta function defined above. From Corollary 35, section 5.4, if a linear combination of trees is Hamiltonian, then $b(u \circ v) + b(v \circ u) = 0$ and we deduce that the set of dual trees that annihilate $\mathcal{T}_\Omega^n$ is given by

$$\{(u \circ v)^* + (v \circ u)^* : \forall\ u, v \in T,\ such\ that\ |u| + |v| = n\}.$$

Example 44. For $n = 5$, the pairs (u, v) of trees satisfying $|u| + |v| = n$ are $(\bullet, [\bullet, \bullet, \bullet])$, $(\bullet, [[\bullet, \bullet]])$, $(\bullet, [\bullet, [\bullet]])$, $(\bullet, [[[\bullet]]])$, $([\bullet], [\bullet, \bullet])$, $([\bullet], [[\bullet]])$. Thus

$Ann(\mathcal{T}_\Omega^5) = \{$  $\}$

This information can be collected in the following Theorem.

Theorem 45. *A basis of $\mathcal{T}_\Omega^n$ is given by $\{X_{t_i}\}$ where one t_i is chosen from each element of FT^n. A basis of the annihilator $Ann(\mathcal{T}_\Omega^n)$ of $\mathcal{T}_\Omega^n$ is given by*

$$\{(u \circ v)^* + (v \circ u)^* : \forall\ u, v \in T,\ such\ that\ |u| + |v| = n\},$$

so that Hamiltonian B-series of the form $\sigma(t) \sum_{\theta \sim t} \frac{(-1)^{\kappa(\theta,\tau)}}{\sigma(\theta)} \theta \in T^n$, satisfy

$$b(u \circ v) + b(v \circ u) = 0\ \forall\ u, v \in T,\ such\ that\ |u| + |v| = n.$$

The dimension of $\mathcal{T}_\Omega^n$ is given by

$$\dim(\mathcal{T}_\Omega^n) = |\mathcal{T}_\Omega^n| = \begin{cases} |FT^n| & n \text{ odd} \\ |FT^n| - \left|T^{\frac{n}{2}}\right| & n \text{ even} \end{cases}.$$

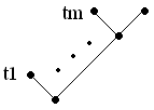
The conditions to be Energy-preserving, i.e. the conditions for the modified differential equation to preserve the Hamiltonian $H(\mathbf{x})$, are given by Lemma 36 and Corollary 37, section 5.5 with $c(\bullet) = 1$ and $c(\tau) = 0 \ \forall \ |\tau| > 1$ (we have these conditions on $c(\tau)$ if we wish to only conserve the Hamiltonian and not a perturbation of the Hamiltonian). A set of energy-preserving trees was first given by Quispel and McLaren (2008). The proof is given by Celledoni, McLachlan, Owren and Quispel (2009).

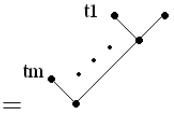
Theorem 46. (Celledoni, McLachlan, Owren, Quispel) *Let*

$$S = \{[t_1, [t_2, ..., [t_m, \bullet] \cdots] + (-1)^m [t_m, [\cdots [t_2, [t_1, \bullet] \cdots] : \ t_j \in \bar{T}\},$$

then $S \subseteq \mathcal{T}_H$.

Proof. Any element of S is of the form $t + (-1)^m \hat{t}$ with

$t =$ 

and $\hat{t} =$ 

.

We first note that t is a representation of an arbitrary tree, with the spine being the path from the root to any leaf. From the chain rule, we have that the Hamiltonian H is preserved by the flow map of a vector field $\mathbf{g}$ if and only if $H'(\mathbf{g}(\mathbf{x})) = 0$ along integral curves $\mathbf{x}$. We must prove that this is true for the vector field $F(t) + (-1)^m F(\hat{t})$. From the definition of the elementary Hamiltonians we see that $H'(F(t)) = H([t])$. Using the root shifting property $H(u \circ v) = -H(v \circ u)$ for all trees u and v, the root of $[t]$ can be moved to an adjacent vertex incurring a change of sign. We shift the root up $m+1$ places until it reaches the designated $\bullet$ in $[\hat{t}] = [[t_m, [\cdots [t_2, [t_1, \bullet] \cdots]]$. The resulting tree is $[\hat{t}]$, thus $\kappa(t, \hat{t}) = m + 1$, and we find that $H([\hat{t}]) = (-1)^{m+1} H([t])$. Using the definition of elementary Hamiltonians we conclude that $H'(F(t) + (-1)^m F(\hat{t})) = 0$, thus proving that the vector field $F(t) + (-1)^m F(\hat{t})$ preserves H.□

The table below shows the energy-preserving pairs up to order 5.

$\lvert\tau\rvert$	Energy-preserving pair
1	
2	
3	
4	
5	

Example 47.

The key to finding these energy-preserving pairs was the so called "Average Vector Field" (AVF) method. This method was first introduced by McLachlan, Quispel and Robidoux (1999) and was shown to be linear and energy-preserving by McLaren and Quispel (2008). For the ODE

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$$

the AVF method is the map $\mathbf{x} \to \mathbf{x}'$ defined by

$$\frac{\mathbf{x}' - \mathbf{x}}{h} = \int_0^1 \mathbf{f}(\xi\mathbf{x}' + (1-\xi)\mathbf{x})d\xi.$$

Celledoni, McLachlan, McLaren, Owren, Quispel and Wright (2008) give the following proof that the AVF method is a B-series method. Assume $\mathbf{x}'$ has a B-series given by

$$\mathbf{x}' = \mathbf{x} + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(\mathbf{x}).$$

This can be substituted into $\mathbf{f}(\xi\mathbf{x}' + (1-\xi)\mathbf{x})$ and an application of Lemma 23 of section 4.2 will yield the resulting coefficients. Integrating term by term we obtain

$$a(\bullet) = 1, \qquad b([t_1, ..., t_m]) = \frac{1}{m+1} a(t_1) \cdot ... \cdot a(t_m).$$

Thus, the B-series for $\mathbf{x}'$ has the form

$$\mathbf{x}' = \mathbf{x} + h\mathbf{f} + \frac{1}{2}h^2\mathbf{f}'\mathbf{f} + h^3(\frac{1}{3}\mathbf{f}''(\mathbf{f},\mathbf{f}) + \frac{1}{4}\mathbf{f}'\mathbf{f}'\mathbf{f}) + h^4(\frac{1}{4}\mathbf{f}'''(\mathbf{f},\mathbf{f},\mathbf{f}) + \frac{1}{6}\mathbf{f}'\mathbf{f}''(\mathbf{f},\mathbf{f}) + \frac{1}{6}\mathbf{f}''(\mathbf{f}'\mathbf{f},\mathbf{f}) + \frac{1}{8}\mathbf{f}'\mathbf{f}'\mathbf{f}'\mathbf{f}) + ...$$

Using the recurrence relation given in Theorem 29, section 5.3, the coefficients for the modified differential equation are therefore

$b(\bullet) = 1$

$b([\bullet]) = \frac{1}{2} - \frac{1}{2}b(\bullet)^2 = 0$

$b([\bullet,\bullet]) = \frac{1}{3} - b([\bullet]) \cdot b(\bullet) - \frac{1}{3}b(\bullet)^3 = 0$

$b([[\bullet]]) = \frac{1}{4} - b([\bullet]) \cdot b(\bullet) - \frac{1}{3}b(\bullet)^3 = \frac{-1}{12}$

$b([\bullet,\bullet,\bullet]) = \frac{1}{4} - \frac{3}{2}b([\bullet,\bullet])b(\bullet) - \frac{1}{6}(6b([\bullet])b(\bullet)^2) - \frac{1}{24}(6b(\bullet)^4) = 0$

$b([[\bullet,\bullet]]) = \frac{1}{6} - \frac{1}{2}(b([\bullet,\bullet])b(\bullet) + 2b([[\bullet]])b(\bullet)) - \frac{1}{6}(6b([\bullet])b(\bullet)^2) - \frac{1}{24}(6b(\bullet)^4) = 0$

$b([\bullet,[\bullet]]) = \frac{1}{6} - \frac{1}{2}(b([[\bullet]])b(\bullet) + b([\bullet])^2 + b([\bullet,\bullet])b(\bullet)) - \frac{1}{6}(2b([\bullet])b(\bullet) + b([\bullet])b(\bullet) + 2b([\bullet])b(\bullet)) - \frac{1}{24}(5b(\bullet)^4) = 0$

$b([[[\bullet]]]) = \frac{1}{8} - \frac{1}{2}(2b([[\bullet]])b(\bullet) + b([\bullet])^2) - \frac{1}{6}(4b([\bullet])b(\bullet)^2 + b([\bullet])b(\bullet)) - \frac{1}{24}(5b(\bullet)^4) = 0,$

etc.

giving the B-series of the modified vector field of the AVF method as

$$\bullet - \frac{1}{12}h^2[[\bullet]] + \frac{1}{720}h^4\left(9[[[[\bullet]]]] - ([[\bullet,\bullet,\bullet]] + [\bullet,\bullet,[\bullet]]) + 2([[\bullet],[\bullet]] + [[\bullet,[\bullet]]]) - 4[\bullet,[\bullet,\bullet]] + 4([[[\bullet,\bullet]]] - [\bullet,[[\bullet]]])\right).$$

The coefficients have naturally grouped the trees into energy-preserving pairs.

As was mentioned before, the conditions to be energy-preserving, i.e. for the modified differential equation to conserve the Hamiltonian $H(\mathbf{x})$ (and not a perturbation of it) is given by Lemma 36 and Corollary 37, section 5.5 with $c(\bullet) = 1$ and $c(\tau) = 0 \ \forall \ |\tau| > 1$. Therefore, in obtaining the conditions for the modified vector field to be energy-preserving we only consider splittings, of a tree τ, of the form $(\bullet, \tau \setminus \bullet)$. Since the conditions to be energy-preserving define a basis for the annihilator of the energy-preserving subspace of trees in the same way as for Hamiltonian trees, a basis for the annihilator, $Ann(\mathcal{T}_H^n)$, of the energy-preserving subspace at order n is

$$\sum_{\tau \in \phi,\ \tau = [\bar{\tau}]} (-1)^{\kappa(\tau_0,\tau)} \frac{1}{\sigma(\tau)} \bar{\tau}^*$$

where ϕ is a non-superfluous free-tree of order $n+1$, τ_0 is a designated element of ϕ and the sum is taken over all trees in ϕ having exactly one sub-tree attached to the root.

Example 48. We shall calculate the conditions to be energy-preserving for $n = 4$. The three non-superfluous trees of order 5 are

1. $\{$  $\}$
2. $\{$  $\}$
3. $\{$  $\}$

For the first non-superfluous free-tree, let $\tau_0 =$  and then the only tree with exactly one sub-tree attached to the root is $\tau =$  $= [$  $]$ with $\kappa(\tau_0, \tau) = 1$. Therefore, the first condition to be energy-preserving is

$\frac{-1}{6}$ *.

For the second non-superfluous free-tree, let $\tau_0 =$  and the only trees with exactly one sub-tree attached to the root are $\tau_1 =$  $= [$  $]$ and $\tau_2 =$  $= [$  $]$ with $\kappa(\tau_0, \tau_1) = 1$ and $\kappa(\tau_0, \tau_2) = 2$. Therefore the second energy-preserving condition is

$\frac{1}{2}$  $*-$  $*$.

Lastly, we let $\tau_0 =$  with $\tau =$  $= [$ $]$ and $\kappa(\tau_0, \tau) = 2$. Thus, the last energy-preserving condition is

 $*$

The dimension of $Ann(\mathcal{T}_H^n)$ is $\dim(\mathcal{T}_\Omega^{n+1})$. Since $\dim(\mathcal{T}^n) = \dim(\mathcal{T}_H^n) + \dim(Ann(\mathcal{T}_H^n)$, (see Roman (2008) for a full treatment of vector spaces, annihilators and their properties) it follows immediately that $\dim(\mathcal{T}_H^n) = \dim(\mathcal{T}^n) - \dim(\mathcal{T}_\Omega^{n+1})$. A basis for $\mathcal{T}_H^n$ is obtained by choosing a linearly independent set from S of the appropriate dimension. The following Theorem due to Celledoni, McLachlan. Owren and Quispel (2010) encapsulates the above information.

Theorem 49. (Celledoni, McLachlan, Owren, Quispel) *(i) A basis for the annihilator $Ann(\mathcal{T}_H^n)$ of $\mathcal{T}_H^n$ can be indexed over the non-superfluous elements of FT^{n+1} as follows*

$$\left\{ \sum_{\tau \in \phi,\ \tau = [\bar{\tau}]} (-1)^{\kappa(\tau_0, \tau)} \frac{1}{\sigma(\tau)} \bar{\tau}^* \ :\ \phi \in FT_{NS}^{n+1} \right\}$$

where τ_0 is a designated element of ϕ and the sum is taken over all trees $\tau \in \phi$ having exactly one sub-tree attached to the root.

(ii) The dimension of $\mathcal{T}_H^n$ is given by

$$\dim(\mathcal{T}_H^n) = \dim(\mathcal{T}^n) - \dim(\mathcal{T}_\Omega^{n+1}).$$

(iii) S spans $\mathcal{T}_H$.

(iv) Each $\tau \in T^n$ satisfies either (a) $\tau \in \mathcal{T}_H^n$, if $\pi([\tau])$ is non-superfluous (where $\pi([\tau])$ denotes the free-tree to which $[\tau]$ belongs); (b) $\tau^* \in Ann(\mathcal{T}_H^n)$ (that is, it appears nowhere in any energy-preserving B-series), if $\pi([\tau])$ is symmetric and non-superfluous; or (c) it appears in a basis of $\mathcal{T}_H^n$ chosen from S, if neither (a) nor (b) hold.

As a consequence of Theorem 39, section 5.5 (the only Hamiltonian B-series that preserves the energy is the exact flow of the differential equation), the energy-preserving and Hamiltonian trees only intersect in the exact flow of the differential equation:

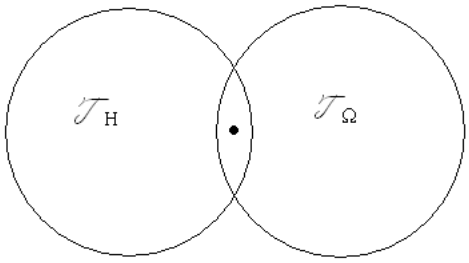


Figure 6.1: Venn diagram of the Energy-preserving and Hamiltonian subspaces

$$\mathcal{T}_H \cap \mathcal{T}_\Omega = \text{Span}(\{\bullet\}).$$

The reason for this is if the modified differential equation is Hamiltonian with respect to the same symplectic structure, Ω, as the original differential equation, then since the modified differential equation is close to the original one we have a modified Hamiltonian $\tilde{H}(\mathbf{x})$. Therefore the Hamiltonian modified differential equation preserves the same symplectic structure as the original ODE but preserves a different Hamiltonian to the original ODE. Similarly if the modified differential equation is Energy-preserving, i.e. preserves the same Hamiltonian H as the original ODE then we must have a modified symplectic structure, $\tilde{\Omega}$ in the modified differential equation.

6.2 Conjugate-to-Energy Preserving and Conjugate-to-Hamiltonian B-Series

Theorem 24 of section 4.4 has allowed us to consider B-series that are conjugate (by a B-series) to another B-series. We now wish to consider whether a B-series can be conjugate (by a B-series) to an energy-preserving of Hamiltonian B-series. Conjugate B-series do not form linear spaces, but some of their properties such as their dimension can be described using two new spaces called $\mathcal{T}_H^n$ and $\mathcal{T}_{\tilde{\Omega}}^n$ (Celledoni, McLachlan, Owren, Quispel (2010)).

A numerical integrator Φ is said to be conjugate-symplectic if there exists a map Ψ such that $\Psi\Phi\Psi^{-1}$ is symplectic. We only consider methods that admit B-series. Thus we require both Φ and Ψ have B-series. The method Φ preserves a modified symplectic form $(\Psi^{-1})^*\Omega$ not the original form Ω (Celledoni, McLachlan, Owren and Quispel (2010)). The conditions on Φ (and its modified vector field) that ensure conjugate-symplecticity have been derived by Scully (2002) and Hairer, Lubich and Wanner (2002)

At each step of the method $\Psi\Phi\Psi^{-1}$, the perturbing effect of the Ψ transformation will be corrected by the Ψ^{-1} transformation at the next step. Thus, $(\Psi\Phi\Psi^{-1})^n = \Psi\Phi^n\Psi^{-1}$. Butcher

and Sanz-Serna (1996) show that an application of the method Ψ^{-1} at the start of the iteration process adjusts the initial point to account for the error which would occur when applying the map Ψ at the last iteration. Therefore, two methods which are conjugate to each other share the same long-term behavior, irrespective of whether both methods are of a certain order, or are symplectic, or are energy-preserving, or share any other properties. Thus, even if the original B-series method Φ is not Hamiltonian (energy-preserving), it may be conjugate (by, for example, a method Ψ) to a method which is Hamiltonian (Energy-preserving) and these properties will be preserved by the using the conjugacy. Our goal then, is to describe the modified vector fields that are conjugate by a B-series to Hamiltonian or Energy-preserving B-series, i.e. eliminate the conjugacy from the description. This is a first step to determining which B-series (if any) we should conjugate to achieve Hamiltonicity or Energy preservation.

The easiest way to eliminate the conjugacy is by considering the conjugacy to be fixed up to some order and variable thereafter. This is reasonable because, given a modified vector field that is conjugate to a Hamiltonian vector field, the conjugacy is determined, essentially uniquely, order by order (Celledoni, McLachlan, Owren and Quispel (2010)).

In order to continue we will need to use Lie brackets of vector fields. The Lie bracket will be denoted by $[|f,g|]$ to distinguish them from the root-grafting operation $\bullet \circ t = [t]$. The Lie bracket of vector fields induces a Lie bracket on $\mathcal{T}$ via the elementary differentials: $[|\tau_1,\tau_2|]$ is the sum of the grafts of τ_1 onto each vertex of τ_2 minus the sum of the grafts of τ_2 onto each vertex of τ_1 (Celledoni, McLachlan, Owren, Quispel (2010)),

$$[|\tau_1,\tau_2|] = \sum_{\gamma \in V\tau_2} \tau_2 \circ_\gamma \tau_1 - \sum_{\delta \in V\tau_1} \tau_1 \circ_\delta \tau_2,$$

where $\tau_2 \circ_\gamma \tau_1$ denotes the grafting of a tree onto a vertex γ of τ_2 (similarly for the second sum), the first sum is taken over all vertices of τ_2 and the second sum is taken over all vertices of τ_1(see example 55). This definition follows directly from the Lie-derivative of B-series. Since the Hamiltonian and Energy-preserving vector fields form Lie sub-algebras, $\mathcal{T}_H$ and $\mathcal{T}_\Omega$ form graded Lie sub-algebras of $\mathcal{T}$. I shall now give the setup for the idea of conjugate Hamiltonian B-series as given in the paper "Energy-preserving Integrators and the Structure of B-Series".

Let c be the B-series whose flow conjugates the flow of $\tilde{\mathbf{f}}$:

$$e^c e^{\tilde{\mathbf{f}}} e^{-c} = e^{\hat{\mathbf{f}}}$$

where $\hat{\mathbf{f}}$ is Hamiltonian. c is called the conjugacy. Upon rearranging we get

$$e^{\tilde{\mathbf{f}}} = e^{-c} e^{\hat{\mathbf{f}}} e^{c}.$$

Therefore, the conjugate Hamiltonian B-series is

$$\left\{ log(e^{-c} e^{\hat{\mathbf{f}}} e^{c}) : \; c \in \mathcal{T}, \; \hat{\mathbf{f}} \in \mathcal{T}_\Omega \right\}.$$

This is the same as expanding the right-hand-side into the exponential series and multiplying out,

$$\left\{ \hat{\mathbf{f}} - \left[\left| c, \hat{\mathbf{f}} \right| \right] + \frac{1}{2!} \left[\left| c, \left[\left| c, \hat{\mathbf{f}} \right| \right] \right| \right] - ... : c \in \mathcal{T}, \; \hat{\mathbf{f}} \in \mathcal{T}_\Omega \right\},$$

where we have collected the appropriate terms, after multiplying out, into the Lie-bracket.

The following three restrictions have been made by Celledoni, McLachlan, Owren and Quispel (2010):

1. Only consider B-series of consistent integrators, i.e. $\hat{\mathbf{f}}(\bullet) = \tilde{\mathbf{f}}(\bullet) = 1$.

2. Only consider non-Hamiltonian conjugacies c. This is acceptable because a Hamiltonian vector field conjugated by a Hamiltonian vector field is obviously Hamiltonian and the definition of non-Hamiltonian is immaterial. We simply let $\mathcal{T}'_\Omega$ be any fixed complement of $\mathcal{T}_\Omega$ in $\mathcal{T}$ (that is, $\mathcal{T}'_\Omega$ is any vector sub-space disjoint from $\mathcal{T}_\Omega$ such that $\mathcal{T} = \mathcal{T}_\Omega \oplus \mathcal{T}'_\Omega$). Similarly, we let $\mathcal{T}'_H$ be any fixed complement of $\mathcal{T}_H$ in $\mathcal{T}$. As an example, $\mathcal{T}^3_H = \text{Span}(\{[[\bullet]]\})$, therefore we could choose $\mathcal{T}^3_H$'$= \text{Span}(\{[\bullet, \bullet]\})$ or $\mathcal{T}^3_H$'$= \text{Span}(\{[\bullet, \bullet] - [[\bullet]]\})$ or any other 1-dimensional space disjoint from $\mathcal{T}^3_H$. We choose $c \in \mathcal{T}'_\Omega$.

3. Instead of allowing the conjugacy c to range over *all* non-Hamiltonian B-series, we will take its terms of order $< n-1$ to be fixed; and instead of allowing $\hat{\mathbf{f}}$ to range over all Hamiltonian B-series, we will take its terms of order $< n$ to be fixed. This will be a useful restriction because it so happens that, given $\tilde{\mathbf{f}}$, the c that conjugates it to a Hamiltonian B-series will be determined uniquely order-by-order.

The following Lemmas are due to Celledoni, McLachlan, Owren and Quispel (2010).

Lemma 50. (Celledoni, McLachlan, Owren, Quispel). *The map* $ad_\bullet : \mathcal{T}^n \to \mathcal{T}^{n+1}$ *defined by* $ad_\bullet(\tau) = [|\tau, \bullet|]$ *is 1-1 on* $\mathcal{T}^n$ *for* $n > 1$.

Proof. (Vector Fields). Suppose the Lemma is false. Then there are distinct trees τ_1, τ_2 such that $[|\tau_1, \bullet|] = [|\tau_2, \bullet|]$. Then $[|F(\tau_1 - \tau_2), \mathbf{f}|] = 0$, i.e. $\mathbf{f}$ has the non-trivial symmetry $F(\tau_1 - \tau_2)$ for all $\mathbf{f}$. But there are $\mathbf{f}$'s with no non-trivial symmetries, a contradiction.□

The Lie bracket on $\mathcal{T}$ has the form $[|\tau_1, \tau_2|] = \sum_{\gamma \in V\tau_2} \tau_2 \circ_\gamma \tau_1 - \sum_{\delta \in V\tau_1} \tau_1 \circ_\delta \tau_2$. We can identify the two pieces of the Lie bracket with the non-associative $Left\ pre-Lie$ product $\triangleright$. We then obtain $\sum_{\gamma \in V\tau_2} \tau_2 \circ_\gamma \tau_1 = \tau_1 \triangleright \tau_2$ and $\sum_{\delta \in V\tau_1} \tau_1 \circ_\delta \tau_2 = \tau_2 \triangleright \tau_1$ so that $[|\tau_1, \tau_2|] = \tau_1 \triangleright \tau_2 - \tau_2 \triangleright \tau_1$. We may

then define $L_\bullet$ and $R_\bullet$ as the linear maps $L_\bullet : u \to \bullet \triangleright u$ and $R_\bullet : u \to u \triangleright \bullet = [u]$ for a tree $u \in \mathcal{T}$. We can now write $ad_\bullet = R_\bullet - L_\bullet$. The transpose of $ad_\bullet$ is denoted $ad_\bullet^* = R_\bullet^* - L_\bullet^*$ and is a map from the algebraic dual space, $(\mathcal{T}^{n+1})^*$ to the algebraic dual space $(\mathcal{T}^n)^*$ (bases for these spaces are obtained as described in section 6.1). We have that $ad_\bullet^*((\mathcal{T}^{n+1})^*) = (\mathcal{T}^n)^*$ due to Lemma 50. If $u \in T^{n+1}$ and $v \in T^n$, we let $r(u,v)$ count the number of leaves that, when removed from u, would yield v. Clearly we may have $r(u,v) = 0$. Then we have the following formulas for $R_\bullet^*$ and $L_\bullet^*$ applied to the dual element u^*:

$$L_\bullet^* = \sum_{v \in T^n} r(u,v) v^* \qquad\qquad R_\bullet^* = \left\{ \begin{array}{ll} \bar{u}^* \; if \; u = [\bar{u}] \\ 0 \quad otherwise \end{array} \right\}.$$

We may naturally interpret the Butcher product on dual elements as $u^* \circ v^* = (u \circ v)^*$ for any $u, v \in T$. For convenience we shall augment the basis T with the identity element $\varnothing$ of grade 0, such that $\varnothing \cdot \tau = \tau \cdot \varnothing = \tau$ for any $\tau \in \mathcal{T} \oplus \mathbf{R}\varnothing$. We then have $L_\bullet^*(\bullet^*) = \varnothing^*$ and by convention $\tau \circ \varnothing = -\varnothing \circ \tau = \tau$ for any $\tau \in T$.

Lemma 51. (Celledoni, McLachlan, Owren, Quispel). $ad_\bullet(\mathcal{T}_H\,') \cap \mathcal{T}_H = 0$.

Proof. (Vector Fields) Suppose the Lemma is false. Then there exists $\mathbf{g} = F(\tau)$, $\tau \in \mathcal{T}_H$' satisfying $[|\mathbf{g}, \mathbf{f}|]\,(H) = 0$. Then

$$0 = [|\mathbf{g}, \mathbf{f}|]\,(H) = \mathbf{g}(\mathbf{f}(H)) - \mathbf{f}(\mathbf{g}(H)) = \mathbf{f}(\mathbf{g}(H)).$$

That is, $\mathbf{f}$ has first integral $\mathbf{g}(H)$. But there exist $\mathbf{f}$ whose only independent first integral is H; in this case, $\mathbf{g}(H) = G(H)$ for some scalar function G. But $\mathbf{g}$ is an elementary differential of $\mathbf{f}$ so $G(H) = 0$. That is, $\mathbf{g} \in \mathcal{T}_H$, a contradiction.□

Lemma 52. (Celledoni, McLachlan, Owren, Quispel). $ad_\bullet(\mathcal{T}_\Omega\,') \cap \mathcal{T}_\Omega = 0$.

Proof. The vector field $\mathbf{f}$ is assumed Hamiltonian, that is, $i_{\mathbf{f}}\Omega = dH$ or $di_{\mathbf{f}}\Omega = 0$. Suppose the Lemma is false. Then there exists $\mathbf{g} = F(t)$, $t \in \mathcal{T}_\Omega$' such that $[|\mathbf{f}, \mathbf{g}|]$ is Hamiltonian, i.e.,

$$0 = di_{[|\mathbf{f},\mathbf{g}|]}\Omega$$

$$\Rightarrow 0 = d(\mathcal{L}_{\mathbf{f}} i_{\mathbf{g}}\Omega - i_{\mathbf{g}}\mathcal{L}_{\mathbf{f}}\Omega)$$

$$\Rightarrow 0 = d\mathcal{L}_{\mathbf{f}} i_{\mathbf{g}}\Omega$$

$$\Rightarrow 0 = d(di_{\mathbf{f}} i_{\mathbf{g}}\Omega + i_{\mathbf{f}} di_{\mathbf{g}}\Omega)$$

$$\Rightarrow 0 = di_{\mathbf{f}}\tilde{\Omega} \text{ where } \tilde{\Omega} = di_{\mathbf{g}}\Omega = \mathcal{L}_{\mathbf{g}}\Omega.$$

That is, the flow of the vector field $\mathbf{f}$ preserves both Ω and $\tilde{\Omega}$. (In coordinates, $\tilde{\Omega} = \Omega\mathbf{g}' + \mathbf{g}'^T\Omega$.) But there exist $\mathbf{f}$ whose flow does not preserve two independent 2-forms: for example, in $\mathbb{R}^2$ with $\Omega = dx \wedge dy$, $\tilde{\Omega} = w(x,y)dx \wedge dy$, we would need $\nabla \cdot \mathbf{f} = 1 = w(x,y)$. Regardless of t there will exist $\mathbf{f}$ for which $w(x,y) \neq 1$, a contradiction.$\square$

From the third assumption made above, we know that the conjugacy c is fixed for terms of order $< n-1$ and variable for terms of order $\geq n-1$ and the Hamiltonian vector field $\hat{\mathbf{f}}$ is fixed for terms of order $< n$ and variable for terms of order $\geq n$ in the conjugate-to-Hamiltonian B-series. Therefore, at order n, the only variable terms are $\hat{\mathbf{f}}^n$ and $\left[\left|c^{n-1}, \bullet\right|\right]$. The fixed lower order terms are all constant and can be collected into a single constant b. Since $\hat{\mathbf{f}}^n$ ranges over the space $\mathcal{T}_\Omega^n$, and c^{n-1} ranges over the space $\mathcal{T}_\Omega^{n-1\prime}$, the order n terms in the B-series for $\tilde{\mathbf{f}}$ ranges over the affine space

$$\mathcal{T}_{\tilde{\Omega}}^n + b$$

$$\mathcal{T}_{\tilde{\Omega}}^n = \mathcal{T}_\Omega^n \oplus \left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right],$$

the direct sum owing to Lemma 52. Since $ad_\bullet$ is a $1-1$ map on $\mathcal{T}$, it is a $1-1$ map on any subspace of $\mathcal{T}$ and $\dim(\left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right]) = \dim(\mathcal{T}_\Omega^{n-1\prime})$. Therefore, the dimension of $\mathcal{T}_{\tilde{\Omega}}^n$ is given by

$$\dim(\mathcal{T}_{\tilde{\Omega}}^n) = \dim(\mathcal{T}_\Omega^n) + \dim(\mathcal{T}^{n-1}) - \dim(\mathcal{T}_\Omega^{n-1}).$$

By an identical argument, we obtain a similar result for the conjugate-to-Energy-preserving subspace. That is, the order n terms in the conjugate-to-Energy-preserving B-series for $\tilde{\mathbf{f}}$ ranges over the affine space

$$\mathcal{T}_{\tilde{H}}^n + b$$

$$\mathcal{T}_{\tilde{H}}^n = \mathcal{T}_H^n \oplus \left[\left|\mathcal{T}_H^{n-1\prime}, \bullet\right|\right],$$

where

$$\dim(\mathcal{T}_{\tilde{H}}^n) = \dim(\mathcal{T}_H^n) + \dim(\mathcal{T}^{n-1}) - \dim(\mathcal{T}_H^{n-1}).$$

These results were first given by Celledoni, McLachlan, Owren and Quispel (2010) and may be collected into the following Theorem.

Theorem 53. (Celledoni, McLachlan, Owren, Quispel). *Let $n > 2$. As the conjugacy c ranges over $\mathcal{T}_\Omega$' with terms of order $< n-1$ fixed, and $\hat{\mathbf{f}}$ ranges over $\mathcal{T}_\Omega$ with $\hat{\mathbf{f}}(\bullet) = 1$ and terms of order $< n$ fixed, the order n terms in the conjugate-to-Hamiltonian B-series*

$$\hat{\mathbf{f}} - \left[\left|c, \hat{\mathbf{f}}\right|\right] + \frac{1}{2!}\left[\left|c, \left[\left|c, \hat{\mathbf{f}}\right|\right]\right|\right] - ...$$

ranges over the affine space

$$\mathcal{T}_{\tilde{\Omega}}^n + b$$

where $\mathcal{T}_{\tilde{\Omega}}^n$ is the linear space

$$\mathcal{T}_{\tilde{\Omega}}^n = \mathcal{T}_\Omega^n \oplus \left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right]$$

and $b \in \mathcal{T}^n$ is a constant depending on the lower order terms in c and $\hat{\mathbf{f}}$. The space $\mathcal{T}_{\tilde{\Omega}}^n$ is well-defined in the sense that it does not depend on the choice of complement $\mathcal{T}_\Omega^{n-1\prime}$. The dimension of $\mathcal{T}_{\tilde{\Omega}}^n$ is

$$\dim(\mathcal{T}_{\tilde{\Omega}}^n) = \dim(\mathcal{T}_\Omega^n) + \dim(\mathcal{T}^{n-1}) - \dim(\mathcal{T}_\Omega^{n-1}).$$

Let $n > 2$. As the conjugacy c ranges over $\mathcal{T}_H$' with terms of order $< n-1$ fixed, and $\hat{\mathbf{f}}$ ranges over $\mathcal{T}_H$ with $\hat{\mathbf{f}}(\bullet) = 1$ and terms of order $< n$ fixed, the order n terms in the conjugate-to-Hamiltonian B-series

$$\hat{\mathbf{f}} - \left[\left|c, \hat{\mathbf{f}}\right|\right] + \frac{1}{2!}\left[\left|c, \left[\left|c, \hat{\mathbf{f}}\right|\right]\right|\right] - ...$$

ranges over the affine space

$$\mathcal{T}_{\tilde{H}}^n + b$$

where $\mathcal{T}_{\tilde{\Omega}}^n$ is the linear space

$$\mathcal{T}_{\tilde{H}}^n = \mathcal{T}_H^n \oplus \left[\left|\mathcal{T}_H^{n-1\prime}, \bullet\right|\right]$$

and $b \in \mathcal{T}^n$ is a constant depending on the lower order terms in c and $\hat{\mathbf{f}}$. The space $\mathcal{T}_{\tilde{H}}^n$ is well-defined in the sense that it does not depend on the choice of complement $\mathcal{T}_H^{n-1\prime}$. The dimension of $\mathcal{T}_{\tilde{H}}^n$ is

$$\dim(\mathcal{T}_{\tilde{H}}^n) = \dim(\mathcal{T}_H^n) + \dim(\mathcal{T}^{n-1}) - \dim(\mathcal{T}_H^{n-1}).$$

It will be useful to define a map $X_{[\cdot]} : \mathcal{T}^n \to \mathcal{T}_\Omega^{n+1}$ by

$$X_{[\cdot]}(t) = X_{[t]} = \sigma([t]) \sum_{\theta \sim [t]} \frac{(-1)^{\kappa(\theta,[t])}}{\sigma(\theta)} \theta,$$

for all trees $t \in \mathcal{T}^n$. Then we have the following Lemma from Celledoni, McLachlan, Owren and Quispel (2010).

Lemma 54. (Celledoni, McLachlan, Owren, Quispel). *$[|t, \bullet|] - X_{[t]} \in \mathcal{T}_H$ for all trees $t \in \mathcal{T}$.*

Proof. We claim that for all $t \in \mathcal{T}$, $[|t, \bullet|] - X_{[t]} \in \mathcal{T}_H$. That is, the Hamiltonian component of $[|t, \bullet|] \in \mathcal{T}_{\tilde{H}}$ is $X_{[t]}$ and the remainder is energy-preserving. The equivalent claim for elementary differentials is established. Let $\mathbf{g} = F(t)$. The elementary differential associated with $[|t, \bullet|]$ is $[|\mathbf{g}, \mathbf{f}|]$, and the elementary differential associated with $X_{[t]}$ is $X_{H'(\mathbf{g})} = X_{\mathbf{g}(H)}$. Using $\mathbf{f} = X_H$ and $\mathbf{f}(H) = 0$ gives

$$\begin{aligned}([|\mathbf{g}, \mathbf{f}|] - X_{\mathbf{g}(H)})(H) &= \mathbf{g}(\mathbf{f}(H)) - \mathbf{f}(\mathbf{g}(H)) - X_{\mathbf{g}(H)}(H) \\ &= -X_H(\mathbf{g}(H)) - X_{\mathbf{g}(H)}(H) \\ &= -\{H, \mathbf{g}(H)\} - \{\mathbf{g}(H), H\} \\ &= 0.\end{aligned}$$

Thus, the remainder is energy-preserving and $[|t, \bullet|] - X_{[t]} \in \mathcal{T}_H$.□

Note that the above decomposition holds for any vector field $\mathbf{g}$ when $\mathbf{f}$ is Hamiltonian with respect to any symplectic structure, not just $\mathbf{g}$ an elementary differential of $\mathbf{f}$ and Ω constant, as was assumed everywhere else (Celledoni, McLachlan, Owren and Quispel 2010). For an analysis and discussion of the above decomposition see (R. I. McLachlan 2009).

We may now define another map $EP : \mathcal{T}^n \to \mathcal{T}_H^{n+1}$ by

$$EP(t) = [|t, \bullet|] - X_{[t]}$$

for all trees $t \in \mathcal{T}^n$.

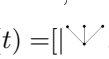
Example 55. Consider $t =$  . We have

$[|$ $, \bullet|] =$.

The Hamiltonian vector field associated with $[t]$ is

$X_{[.]}($ $) = 4$.

Therefore,

$EP(t) =[|$ $, \bullet|]-X_{[.]}($ $) = -3($  $),$

and

$[|$ $, \bullet|] = X_{[t]} + 3($ $).$

The following Lemmas give some properties of the maps $X_{[.]}$ and EP.

Lemma 56. $\mathcal{T}_H = \ker(X_{[.]})$.

Proof. Each Energy-preserving tree is a linear combination of trees of the form

1. t_1, such that $[t_1] = -[t_1]$
2. $t_2 + t_3$, such that $[t_2] = -[t_3]$
3. $t_4 - t_5$, such that $[t_4] = [t_5]$.

Now

$$X_{[t_1]} = X_{-[t_1]} = -X_{[t_1]} \Rightarrow X_{[t_1]} = 0$$

$$X_{[.]}(t_2 + t_3) = X_{[t_2+t_3]} = X_{[t_2]+[t_3]} = X_{[t_2]-[t_2]} = X_0 = 0$$

$$X_{[.]}(t_4 - t_5) = X_{[t_4-t_5]} = X_{[t_4]-[t_5]} = X_{[t_4]-[t_4]} = X_0 = 0.$$

Thus, if $t \in \mathcal{T}_H$ then $X_{[.]}(t) = 0$ and $t \in \ker(X_{[.]})$.

Suppose $t \in \ker(X_{[.]})$. Then $X_{[.]}(t) = 0$. Therefore, by Lemma 54 we have that $ad_\bullet(t) = EP(t) \in \mathcal{T}_H$. Thus, by Lemma 51 $t \in \mathcal{T}_H$.□

Lemma 57. $\mathcal{T}_\Omega = \ker(EP)$.

Proof. We claim that for all $t \in \mathcal{T}$, $[|X_t, \bullet|] = X_{[X_t]}$. We can establish the equivalent claim for elementary differentials. Let $\mathbf{g} = F(t)$, the elementary differentials associated with $[|X_t, \bullet|]$ is $[|X_\mathbf{g}, \mathbf{f}|]$ and the elementary differential associated with $X_{[X_t]}$ is $X_{H'(X_\mathbf{g})}$. Using $\mathbf{f} = X_H$ we have

$$[|X_\mathbf{g}, X_H|] = -X_{\{\mathbf{g},H\}} = X_{-\{\mathbf{g},H\}} = X_{\{H,\mathbf{g}\}} = X_{H'(X_\mathbf{g})}.$$

Thus,

$$EP(X_\mathbf{g}) = [|X_\mathbf{g}, X_H|] - X_{H'(X_\mathbf{g})} = 0$$

and $X_{\mathbf{g}} \in \ker(EP)$.

If $t \in \ker(EP)$ then $EP(t) = 0$. Therefore $ad_\bullet(t) = X_{[t]} \in \mathcal{T}_\Omega$ and by Lemma 52 $t \in \mathcal{T}_\Omega$.□

We obtain the following Theorems as a direct result of Lemmas 54, 56 and 57

Theorem 58. (Celledoni, McLachlan, Owren, Quispel). *$\mathcal{T}_{\tilde{H}}^n = \mathcal{T}_H^n \oplus \mathcal{T}_\Omega^n$ for $n > 1$*

Proof. $\mathcal{T}_{\tilde{H}}^n = \mathcal{T}_H^n \oplus \left[\left|\mathcal{T}_H^{n-1\prime}, \bullet\right|\right]$ and every element of $\left[\left|\mathcal{T}_H^{n-1\prime}, \bullet\right|\right]$ can be written as a Hamiltonian combination of trees plus an Energy-preserving combination of trees. Since $\mathcal{T}_{\tilde{H}}^n$ contains all energy-preserving trees as a subspace, we may recover the purely Hamiltonian combinations of trees from the elements of $\left[\left|\mathcal{T}_H^{n-1\prime}, \bullet\right|\right]$ as elements of $\mathcal{T}_{\tilde{H}}^n$. Since $\mathcal{T}_H^{n-1\prime}$ does not contain the kernel of the map $X_{[\cdot]}$ and $\dim(\mathcal{T}_H^{n-1\prime}) = \dim(\mathcal{T}_\Omega^n)$, every Hamiltonian combination of trees of order n appear as elements in $\mathcal{T}_{\tilde{H}}^n$. Therefore, $\mathcal{T}_{\tilde{H}}^n$ consists of all Energy-preserving trees (which form the subspace $\mathcal{T}_H^n$) and all Hamiltonian trees (which form the subspace $\mathcal{T}_\Omega^n$). Since these are the only trees in $\mathcal{T}_{\tilde{H}}^n$ we have that

$$\mathcal{T}_{\tilde{H}}^n = \mathcal{T}_H^n \oplus \mathcal{T}_\Omega^n.$$

□

Theorem 59. *$\mathcal{T}_{\tilde{\Omega}}^n = \mathcal{T}_\Omega^n \oplus EP(\mathcal{T}^{n-1})$ for $n > 1$.*

Proof. $\mathcal{T}_{\tilde{\Omega}}^n = \mathcal{T}_\Omega^n \oplus \left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right]$ and every element of $\left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right]$ can be written as a Hamiltonian combination of trees plus an Energy-preserving combination of trees. Since $\mathcal{T}_{\tilde{\Omega}}^n$ contains all Hamiltonian trees as a subspace, we may recover the purely energy-preserving trees from the elements of $\left[\left|\mathcal{T}_\Omega^{n-1\prime}, \bullet\right|\right]$ as elements of $\mathcal{T}_{\tilde{\Omega}}^n$. Therefore, we can construct $\mathcal{T}_{\tilde{\Omega}}^n$ from $\mathcal{T}_\Omega^n$ and $EP(\mathcal{T}_\Omega^{n-1\prime})$. Since $\mathcal{T}_\Omega = \ker(EP)$, $EP(\mathcal{T}_\Omega^{n-1\prime}) = EP(\mathcal{T}_\Omega^{n-1} \oplus \mathcal{T}_\Omega^{n-1\prime}) = EP(\mathcal{T}^{n-1})$. Thus

$$\mathcal{T}_{\tilde{\Omega}}^n = \mathcal{T}_\Omega^n \oplus EP(\mathcal{T}^{n-1}).$$

□

Theorem 59 tells us that $EP(\mathcal{T}^{n-1}) = \mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n \neq \varnothing$ since $\dim(EP(\mathcal{T}^{n-1})) = \dim(Ann(\mathcal{T}_\Omega^{n-1})) \neq 0$ for $n > 1$. That is, there are B-series that are Energy-preserving and conjugate-to-Hamiltonian, but are not the (reparameterized) flow of the original differential. This is good news since we saw that there cannot exist B-series that are both Energy-preserving and Hamiltonian so the next best thing is to find an Energy-preserving method that is conjugate to a Hamiltonian method. Theorem 60 tells us that such B-series may exist although we do not know how to construct them. Theorem 59 first appeared in [5], although not in the same form. It was shown that $\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n$ is the only

Order n	1	2	3	4	5	6	7	8	9	10
$\dim(\mathcal{T}^n)$	1	1	2	4	9	20	48	115	286	719
$\dim(\mathcal{T}^n_\Omega)$	1	0	1	1	3	4	11	19	47	97
$\dim(\mathcal{T}^n_H)$	1	0	1	1	5	9	29	68	189	484
$\dim(\mathcal{T}^n_{\tilde{\Omega}})$	1	0	2	2	6	10	27	56	143	336
$\dim(\mathcal{T}^n_{\tilde{H}})$	1	0	2	2	8	13	40	87	236	581
$\dim(\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}})$	1	0	1	1	3	6	16	37	96	239

Table 6.1: Dimensions of the Linear spaces spanned by the rooted trees and their 5 natural subspaces

non-empty subspace that is independent of the four naturally defined subspaces of B-series and the version of Theorem 59 that appeared described $\mathcal{T}^n_{\tilde{\Omega}}$ as $\mathcal{T}^n_\Omega \oplus (\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}})$ without knowledge of $EP(\mathcal{T}^{n-1}) = \mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}}$.

Theorem 60. *A basis for the space $\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}}$ is given by applying the EP isomorphism to the subset of $\mathcal{T}^{n-1}$ consisting of trees of the following form: (i) Each tree belonging to a superfluous free-tree is an element in the set (ii)From each non-superfluous free-tree ϕ, select any $|\phi| - 1$ trees to be elements in the set.*

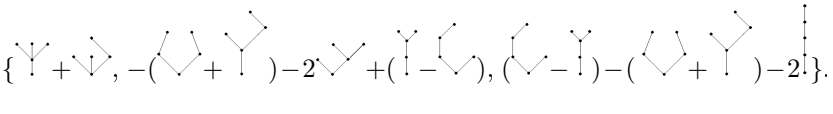
Proof. The Theorem follows from the proof of Theorem 59 and Lemma 57. □

Example 61. We shall construct $\mathcal{T}^5_H \cap \mathcal{T}^5_{\tilde{\Omega}}$. In order to construct the set of trees to be mapped by EP onto $\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}}$ we need to know about the superfluous and non-superfluous free trees of order $n-1$. There is only one non-superfluous free-tree and one superfluous free-tree. These are

$\phi = \pi($ $)$

$\psi = \pi($ $)$

The non-superfluous free-tree ϕ contains two elements, we need only one and we shall select . The superfluous free-tree ψ contains two elements and we need all two trees  and . Therefore, a basis for $\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}}$ is given by

$EP\Big($  $\Big) =$ 

The dimensions of all the natural subspaces of B-series are shown in table 6.1.

Since $EP(\mathcal{T}^{n-1}) \subseteq \mathcal{T}^n_H$ it is evident that $\mathcal{T}^n_{\tilde{\Omega}} \subseteq \mathcal{T}^n_{\tilde{H}}$. The diagram in figure 6.2 illustrates the relationship between all the subspaces.

6.3 The Annihilator of $\mathcal{T}^n_{\widetilde{H}}$

We have characterized the subspaces $\mathcal{T}^n_H$, $\mathcal{T}^n_\Omega$ and their annihilators along with the subspaces $\mathcal{T}^n_{\tilde{H}}$, $\mathcal{T}^n_{\tilde{\Omega}}$ and $\mathcal{T}^n_H \cap \mathcal{T}^n_{\tilde{\Omega}}$. Our goal is to characterize the annihilator of $Ann(\mathcal{T}^n_{\widetilde{H}})$.

Since $\mathcal{T}^n_{\tilde{H}} = \mathcal{T}^n_H \oplus \mathcal{T}^n_\Omega$, we have $Ann(\mathcal{T}^n_{\tilde{H}}) = Ann(\mathcal{T}^n_H) \cap Ann(\mathcal{T}^n_\Omega)$ (See Roman (2008) for details on Annihilators and their properties). That is, we are looking for linear combinations of dual trees

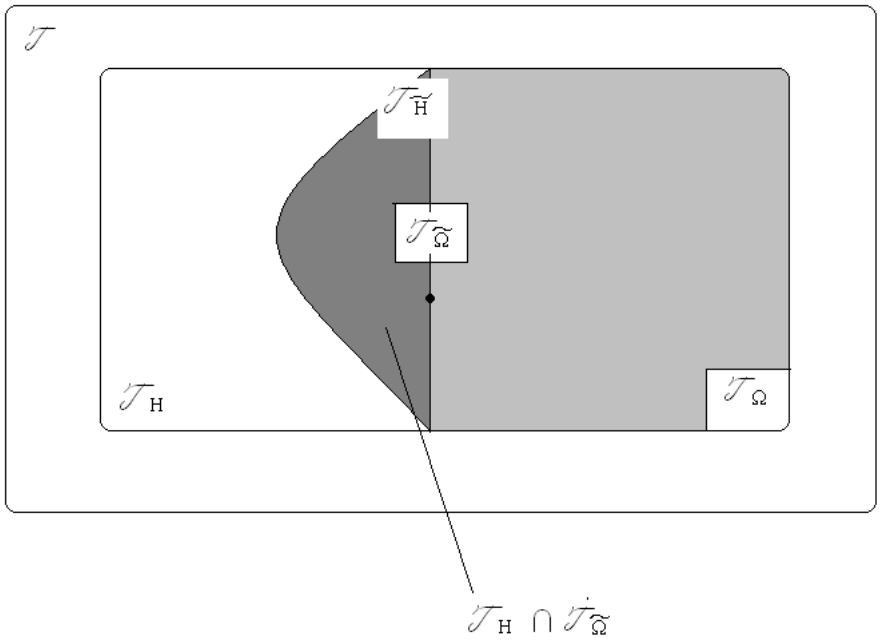
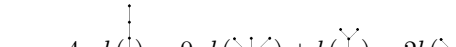
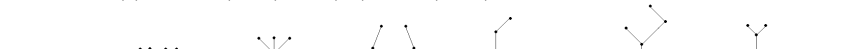


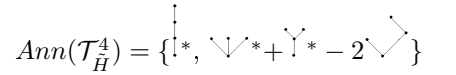
Figure 6.2: Relationship between the 5 natural subspaces of B-series. The dot represents the exact flow of the differential equation and is an element of every subspace.

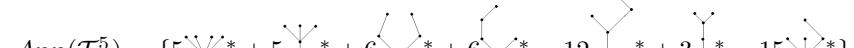
that annihilate the Energy-preserving subspace and annihilate the Hamiltonian subspace. We are able to find such trees. Recall from Corollary 37 and Lemma 38 that the B-series of the modified differential equation, $B(b, \mathbf{x})$, has first integral $H(c, \mathbf{x})$ if and only if $\delta_b c(\tau) = 0 \qquad \forall \tau \in FT_{NS}$ and that this system has a unique solution satisfying $b(u \circ v) + b(v \circ u) = 0$ for all $u, v \in T$. Comparing the expansion of the function $H(c, \mathbf{x})$ to the conjugate-to-Energy preserving B-series, we see that the coefficients $c(\tau)$ give the conjugacy. In order to find the annihilator of the conjugate-to-Energy preserving subspace we need to eliminate the conjugacy $c(\tau)$ from the system $\delta_b c(\tau) = 0$ to find conditions on $b(\tau)$. In Section 5.5 we solved this system by eliminating the conjugacy $c(\tau)$ to obtain

$n = 4 :$ 

$n = 5 :$ 

Since the coefficients $b(\tau)$ represent dual trees, these two expressions are precisely the annihilator conditions for $Ann(\mathcal{T}^n_{\widetilde{H}})$ at order 4 and 5.

$Ann(\mathcal{T}^4_{\widetilde{H}}) =$ 

$Ann(\mathcal{T}^5_{\widetilde{H}}) =$ 

The dimension of $Ann(\mathcal{T}^n_{\widetilde{H}})$ is $\dim(Ann(\mathcal{T}^n_{\widetilde{H}})) = \dim(\mathcal{T}^n) - \dim(\mathcal{T}^n_H) - \dim(\mathcal{T}^n_\Omega) = \dim(Ann(\mathcal{T}^n_H)) - |FT^n_{NS}| = \left|FT^{n+1}_{NS}\right| - |FT^n_{NS}|$. To check that we obtain the correct dimension in solving the linear system $\delta_b c(\tau) = 0$ note that we need to solve $\left|FT^{n+1}_{NS}\right|$ equations. But the only tree splittings $(\omega, \omega \setminus \theta)$ allowed are for $\pi(\omega) \in FT^n_{NS} \cup \{\bullet\}$. Since we are eliminating the conjugacy $c(\omega)$ from

each equation, we lose $|FT^n_{NS}|$ equations due to substitution. Therefore, the number of conditions obtained from the system is $\left|FT^{n+1}_{NS}\right| - |FT^n_{NS}| = \dim(Ann(\mathcal{T}^n_{\widetilde{H}}))$.

We are able to construct a basis for the subspace $Ann(\mathcal{T}^n_{\widetilde{H}})$ at any order but the price we pay for eliminating the conjugacy $c(\tau)$ is solving a linear system. We wish to characterize a basis for $Ann(\mathcal{T}^n_{\widetilde{H}})$ not involving the conjugacy $c(\tau)$. Since $Ann(\mathcal{T}^n_{\widetilde{H}}) = Ann(\mathcal{T}^n_H) \cap Ann(\mathcal{T}^n_\Omega)$ it would seem obvious to try and write the basis for $Ann(\mathcal{T}^n_\Omega)$ for certain trees in terms of the basis for $Ann(\mathcal{T}^n_H)$. However, the basis for $Ann(\mathcal{T}^n_\Omega)$ is obtained from trees of order n whilst the basis for $Ann(\mathcal{T}^n_H)$ is obtained from non-superfluous trees of order $n+1$. Thus, the annihilator bases for $Ann(\mathcal{T}^n_H)$ and $Ann(\mathcal{T}^n_\Omega)$ do not communicate in such a nice way. However, we are able to construct a basis for $Ann(\mathcal{T}^n_{\widetilde{H}})$ using the structure of the energy-preserving basis vectors for $\mathcal{T}^n_H$ and the basis vectors for $Ann(\mathcal{T}^n_\Omega)$.

Theorem 62. *A basis for $Ann(\mathcal{T}^n_{\widetilde{H}})$ may be chosen from the set*

$$\{\sigma(t_2)(t_1^*+(v\circ u)^*)+(-1)^{m+1}\sigma(t_1)(t_2^*+(w\circ x)^*) :\ t_1=(u\circ v), t_2=(x\circ w), u,v,w,x\in\mathcal{T},\ t_1+(-1)^m t_2\in\mathcal{T}^n_H\}.$$

Proof. Any element of the form $(u\circ v)^*+(v\circ u)^*$ will annihilate a Hamiltonian combination of trees. Since $Ann(\mathcal{T}^n_{\widetilde{H}}) = Ann(\mathcal{T}^n_H)\cap Ann(\mathcal{T}^n_\Omega)$, we want to find which linear combinations of elements of the form $(u\circ v)^*+(v\circ u)^*$ annihilate Energy-preserving pairs. It is enough to find the combinations of elements of the form $(u\circ v)^*+(v\circ u)^*$, $|u|+|v|=n$, that annihilate a basis for $\mathcal{T}^n_H$ as given by Quispel and McLaren (2008) and Celledoni, McLachlan, Owren and Quispel (2010).

Note that each energy-preserving pair is of the form $t_1 \pm t_2$ (depending on the parity of the spine) with $t_1 \nsim t_2$, or t_3 (where $[t_3]$ belongs to a superfluous free-tree). Thus, if t_1+t_2 is an energy-preserving pair, then it may be annihilated by the element $\sigma(t_2)(t_1^*+(v\circ u)^*)-\sigma(t_1)(t_2^*+(w\circ x)^*)$, where $t_1=u\circ v$ and $t_2=x\circ w$ for some $u,v,w,x\in\mathcal{T}$. Similarly, if t_1-t_2 is an energy-preserving pair, then it may be annihilated by the element $\sigma(t_2)(t_1^*+(v\circ u)^*)+\sigma(t_1)(t_2^*+(w\circ x)^*)$, where $t_1=u\circ v$ and $t_2=x\circ w$ for some $u,v,w,x\in\mathcal{T}$. But any trees of the form t_3 cannot be annihilated by any basis elements of $Ann(\mathcal{T}^n_\Omega)$.

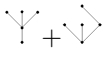
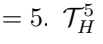
We may then construct a basis for $Ann(\mathcal{T}^n_{\widetilde{H}})$ by grouping the basis vectors of $\mathcal{T}^n_H$ into groups whose trees are related by a single root shift before/after applying the $\sigma(t_2)(t_1^*+(v\circ u)^*)\pm\sigma(t_1)(t_2^*+(w\circ x)^*)$ to a single basis vector in $\mathcal{T}^n_H$ and making use of trees that do not appear in $\mathcal{T}^n_H$. □

Example 63. $\mathcal{T}^4_H = \{$ $\}$ where $t_1 =$  and $t_2 =$  . Therefore, the element that annihilates this energy-preserving pair and $\mathcal{T}^4_\Omega$ is

( $^*+$  $^*) - 2($  $^*+$  $^*)$. For even order vector spaces the tall tree is always in $Ann(\mathcal{T}^{2k}_{\widetilde{H}})$, therefore

$Ann(\mathcal{T}^4_{\widetilde{H}}) = \{($  $^*+$  $^*) - 2($  $^*+$  $^*)$,  $\} = \{$ $^*+$ $^* - 2$ *, $\}$.

For higher orders this process becomes more complicated although quicker than calculating splittings and solving a linear system as in Faou, Hairer and Pham (2004).

Example 64. $n = 5$. $\mathcal{T}_H^5 = \{$,  ,  , , $\}$.

Recall that $\dim(Ann(\mathcal{T}_H^n) = \left|FT_{NS}^{n+1}\right| - |FT_{NS}^n| =$. Notice that there are two elements of the form t_3, with $[t_3]$ superfluous. These cannot be annihilated by any element of $Ann(\mathcal{T}_\Omega^n)$ and cannot appear in any linear combination of trees that annihilate $\mathcal{T}_H^5$ Moreover, these are. Secondly, notice that when we construct the elements that annihilate the remaining basis vectors of $\mathcal{T}_H^5$, we see that these three remaining elements are related. Start with  . The element that annihilates this pair is

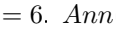
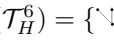
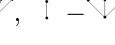
$2($  $) - 6($$) = A$. But the tree  also appears in the energy-preserving pair  and hence for A to be a basis vector of $Ann(\mathcal{T}_H^5)$ it must also annihilate this pair. Therefore we need to add $C($  $)$ to A, where C is a constant so that the new vector will annihilate both  and  . By inspection $C = 3$. Thus we have

$2($$) - 6($$) + 3($  $) = B$. But the tree  also appears in  and hence we need to add $D($  $)$ to B. By inspection $D = \frac{3}{2}$. Therefore we have

$2($$) - 6($$) + 3($  $) + \frac{3}{2}($$)$. Finally, since we cannot have the tree  appearing in any annihilator element of $Ann(\mathcal{T}_H^5)$ we can transform  using the element  $\in Ann(\mathcal{T}_\Omega^5)$ to obtain

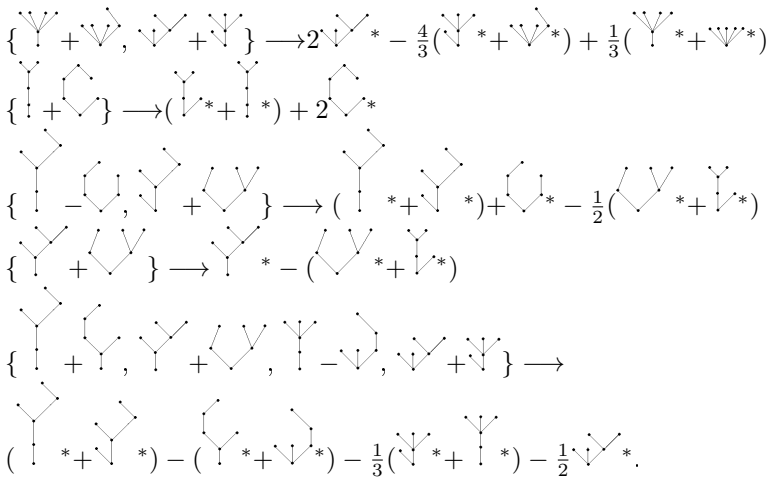
$\frac{3}{2}($  $)$. All that is left to do is update the coefficient of  as this is not related to any other trees other than  and . The updated coefficient is $\frac{5}{2}$. Therefore,

$Ann(\mathcal{T}_H^5) = \{\frac{5}{2}($$) - 6($$) + 3($$) + \frac{3}{2}($$)$.

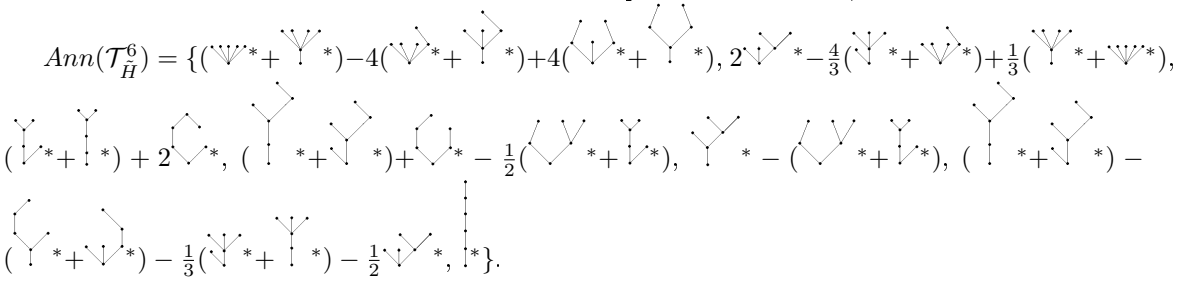
Example 65. $n = 6$. $Ann(\mathcal{T}_H^6) = \{$, , ,  ,  ,  , , , , , , $\}$.

Below are the sets of Energy-preserving pairs that are related to each other and the basis element of $Ann(\mathcal{T}_H^6)$ constructed from these sets.

$\{$, $\} \longrightarrow ($$) - 4($$) + 4($$)$



Since $n = 6$ is even, the tall tree belongs to $Ann(\mathcal{T}^6_{\tilde{H}})$ and along with the above vectors forms a basis for $Ann(\mathcal{T}^6_{\tilde{H}})$. Since the sets of related energy-preserving elements are linearly independent, we know that the constructed basis elements are independent. Therefore,



As a consequence of this construction we have a set of conditions for a B-series to be conjugate-to-energy-preserving. Recall that a basis for the annihilator of a subspace of rooted trees came from the conditions on the coefficients of a B-series to lie in that space. As an example, a B-series, $B(a, \mathbf{x})$, is Hamiltonian if and only if $a(u \circ v) + a(v \circ u) = 0$ for all $u, v \in \mathcal{T}$ and the annihilator, $Ann(\mathcal{T}^n_\Omega)$ was obtained for this space by $a(u \circ v) + a(v \circ u) = (u \circ v)^* + (v \circ u)^*$, since the coefficients are linear functionals defined $\mathcal{T}$. Therefore, in example 65 above, the conditions for a B-series, $B(b, \mathbf{x})$, to lie in the conjugate-to-Energy preserving subspace at order 6 are given by identifying the dual tree, t^*, with the coefficient $b(t) = 0$. This is in accordance with Corollary 37 and Lemma 38 of Faou, Hairer and Pham (2004).

6.4 Relationships Between the Sub-spaces.

This next section is due to an idea of Elena Celledoni. Her original idea was to study the relationships between the various subspaces which would hopefully lead to a basis for $Ann(\mathcal{T}^n_{\tilde{H}})$. Although this idea did not lead to a basis it still gives insight into the algebraic structure of B-series. Recall the adjoint transpose map $ad^*_\bullet$. By Lemma 54, $ad_\bullet = X_{[\cdot]} + EP$ so that $ad^*_\bullet = X^*_{[\cdot]} + EP^*$. The maps $X^*_{[\cdot]}$ and EP^*have the following properties (all of which follow directly from Lemmas 56, 57

and the properties of transpose maps found in Roman (2008)):

1. $im(X^*_{[\cdot]}) = Ann(\mathcal{T}^n_H)$
2. $im(EP^*) = Ann(\mathcal{T}^n_\Omega)$
3. $\ker(X^*_{[\cdot]}) = Ann(\mathcal{T}^{n+1}_\Omega)$
4. $\ker(EP^*) = Ann(\mathcal{T}^{n+1}_H \cap \mathcal{T}^{n+1}_{\tilde{\Omega}})$

We may consider $Ann(\mathcal{T}^n_{\widetilde{H}})^*$, where the elements of this space are treated as vectors in $\mathcal{T}^n$ and $Ann(\mathcal{T}^n_{\widetilde{H}})^* \cap \mathcal{T}^n_{\tilde{H}} = \varnothing$, so that $\mathcal{T}^n = \mathcal{T}^n_{\tilde{H}} \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ (Roman (2008)). $Ann(\mathcal{T}^n_{\widetilde{H}})^*$ is thought of as the canonical complement of $\mathcal{T}^n_{\widetilde{H}}$ in $\mathcal{T}^n$. Since $Ann(\mathcal{T}^n_{\widetilde{H}})^* \cap \mathcal{T}^n_{\tilde{H}} = \varnothing$, we have $Ann(\mathcal{T}^n_{\widetilde{H}})^* \cap \mathcal{T}^n_H = \varnothing$ and $Ann(\mathcal{T}^n_{\widetilde{H}})^* \cap \mathcal{T}^n_\Omega = \varnothing$.

Lemma 66. *The map $X_{[\cdot]}$ is one-to-one from $\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ onto $\mathcal{T}^{n+1}_\Omega$.*

Proof. By Lemma 56, the kernel of $X_{[\cdot]}$ on $\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ is $\{0\}$ so that $X_{[\cdot]}$ is one-to-one. Since $\dim(\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*) = \dim(\mathcal{T}^n_\Omega) + \dim(\mathcal{T}^n) - \dim(\mathcal{T}^n_H) - \dim(\mathcal{T}^n_\Omega) = \dim(\mathcal{T}^{n+1}_\Omega)$, the map is onto. □

Lemma 67. *The map EP is one-to-one from $\mathcal{T}^n_H \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ onto $\mathcal{T}^{n+1}_H \cap \mathcal{T}^{n+1}_{\tilde{\Omega}}$.*

Proof. By Lemma 57, the kernel of EP on $\mathcal{T}^n_H \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ is $\{0\}$ so that EP is one-to-one. Since the dimensions of the two spaces agree, EP is onto. Also, $\mathcal{T}^{n+1}_H \cap \mathcal{T}^{n+1}_{\tilde{\Omega}} = EP(\mathcal{T}^n) = EP(\mathcal{T}^n_H \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*)$. □

Consider the linear mapping $ann_H : \mathcal{T}^n \to Ann(\mathcal{T}^n_H)$ defined on the basis $\mathcal{T}^n$ as follows,

$$ann_H(t) = \begin{cases} \sum_{\theta \in \pi([t])\ \theta = [\bar{\theta}]} (-1)^{\kappa([t],\theta)} \frac{1}{\sigma(\theta)} \bar{\theta}^*, & \pi([t]) \in FT^{n+1}_{NS} \\ 0 & \pi([t]) \text{ Superfluous} \end{cases}$$

The map ann_H corresponds one and only one annihilator condition to each $t \in \mathcal{T}^n$ and is therefore well-defined.

Lemma 68. *Let $t_1, t_2 \in \mathcal{T}^n$. Then $ann_H(t_1) = \pm ann_H(t_2)$ if and only if t_1, t_2 is an energy-preserving pair.*

Proof. If t_1, t_2 is an Energy-preserving pair then clearly $ann_H(t_1) = \pm ann_H(t_2)$ as $[t_1] \sim [t_2]$. Now suppose $ann_H(t_1) = \pm ann_H(t_2)$. Then

$$\sum_{\theta \in \pi([t_1])\ \theta = [\bar{\theta}]} (-1)^{\kappa([t_1],\theta)} \frac{1}{\sigma(\theta)} \bar{\theta}^* = \sum_{\delta \in \pi([t_2])\ \delta = [\bar{\delta}]} (-1)^{\kappa([t_2],\delta)} \frac{1}{\sigma(\delta)} \bar{\delta}^*$$

These two sums are equal (up to sign) if and only if they are over the same non-superfluous equivalence class of trees, since free trees are disjoint equivalence classes. Thus $\pi([t_1]) = \pi([t_2])$. Therefore $[t_1] = \pm[t_2]$ and hence t_1, t_2 is an Energy-preserving.$\square$

Corollary 69. *The map ann_H is one-to-one from $\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ onto $Ann(\mathcal{T}^n_{\widetilde{H}})^*$.*

Proof. $\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$ contains no Energy-preserving trees. Thus, by Lemma 63, ann_H is one-to-one on $\mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^*$. Since the dimensions of the two spaces agree, ann_H is onto.$\square$

Theorem 70. *The map ann_H can be factorized as follows*

$$\begin{array}{ccc} Ann(\mathcal{T}^n_H) & \longleftarrow_{X^*_{[\cdot]}} & (\mathcal{T}^{n+1}_\Omega)^* \\ ann_H \uparrow & & \uparrow * \\ \mathcal{T}^n_\Omega \oplus Ann(\mathcal{T}^n_{\widetilde{H}})^* & \longrightarrow_{X_{[\cdot]}} & \mathcal{T}^{n+1}_\Omega \end{array}$$

where ann_H, $X_{[\cdot]}$ and $X^*_{[\cdot]}$ are bijective maps.

Proof. The Theorem follows directly from Corollary 69, Lemma 66 and property 1 of the transpose map $X^*_{[\cdot]}$.

Consider the linear map $ann_\Omega : \mathcal{T}^n \to Ann(\mathcal{T}^n_\Omega)$ defined on the basis $\mathcal{T}^n$ as follows

$$ann_\Omega(t) = \sum_{u,v \in \mathcal{T}\ u \circ v = t} \frac{1}{\sigma(v \circ u)}((u \circ v)^* + (v \circ u)^*).$$

To each $t \in \mathcal{T}^n$, ann_Ω corresponds one and only one annihilator condition and therefore the maps is well-defined.

Lemma 71. $\ker(ann_\Omega) = \mathcal{T}_\Omega$.

Proof. It is enough to show that the map ann_Ω sends each basis element of $\mathcal{T}^n_\Omega$ to zero. The map $X.$ maps a tree $t \in T$ to $\mathcal{T}^n_\Omega$. Then

$$ann_\Omega(X.(t)) = \sum_{\theta \sim t} (-1)^{\kappa(t,\theta)} \frac{\sigma(t)}{\sigma(\theta)} \sum_{u,v \in \mathcal{T}\ u \circ v = \theta} \frac{1}{\sigma(v \circ u)}(\theta^* + (v \circ u)^*)$$

$$= \sum_{\theta \sim t} (-1)^{\kappa(t,\theta)} \frac{1}{\sigma(\theta)} \sum_{u,v \in \mathcal{T}\ u \circ v = \theta} \frac{\sigma(t)}{\sigma(v \circ u)} \theta^* + \sum_{\theta \sim t} (-1)^{\kappa(t,\theta)} \frac{1}{\sigma(\theta)} \sum_{u,v \in \mathcal{T}\ u \circ v = \theta} \frac{\sigma(t)}{\sigma(v \circ u)} (v \circ u)^*.$$

In the expression $\sum_{u,v\in\mathcal{T}\ u\circ v=\theta}\frac{\sigma(t)}{\sigma(v\circ u)}\theta^*$, $\frac{\sigma(t)}{\sigma(v\circ u)}$ is the number of ways of obtaining the tree $v\circ u$ from t. Therefore, the sum over all u,v, so that $u\circ v=\theta$, gives $deg(\theta)$ (where $deg(\theta)$ gives the number of edges incident on the root of θ) since any root shift of t to obtain $v\circ u$ will either have to go through the root of θ or stop one root shift away from θ and $\frac{\sigma(t)}{\sigma(v\circ u)}$ then tells us how many edges are incident on the root of θ. Thus, $\sum_{u,v\in\mathcal{T}\ u\circ v=\theta}\frac{\sigma(t)}{\sigma(v\circ u)}\theta^* = deg(\theta)\theta^*$. As for the second component we shall first calculate how many times the tree θ^* appears. The tree θ will appear in the sum when a tree $\delta\sim t$ is one root shift away from θ in the Hamiltonian part of the sum. Now θ has opposite sign from its corresponding θ in the first component. Therefore the number of times the tree θ appears in the second component is given by the sum over all trees δ that are one root shift away from θ,

$$\sum_{\delta}\frac{-\sigma(t)}{\sigma(\delta)}\frac{1}{\sigma(\theta)}\theta^* = \frac{-deg(\theta)}{\sigma(\theta)}\theta^*,$$

where the equality was obtained by using the same argument as for the first component. Doing the above for all trees related to t, the second component is $-\sum_{\theta\sim t}\frac{deg(\theta)}{\sigma(\theta)}\theta^*$. Therefore

$$ann_\Omega(X_\cdot(t)) = \sum_{\theta\sim t}\frac{deg(\theta)}{\sigma(\theta)}\theta^* - \sum_{\theta\sim t}\frac{deg(\theta)}{\sigma(\theta)}\theta^* = 0$$

and $\mathcal{T}_\Omega\subseteq\ker(ann_\Omega)$.

Now suppose $t\in\mathcal{T}$ such that $ann_\Omega(t)=0$. Then

$$\sum_{u,v\in\mathcal{T}\ u\circ v=t}\frac{1}{\sigma(v\circ u)}t^* = -\sum_{u,v\in\mathcal{T}\ u\circ v=t}\frac{1}{\sigma(v\circ u)}(v\circ u)^*.$$

Now the trees on the left hand side differ from the trees on the right hand side by exactly one root shift. Therefore, these two sums can only be equal if every tree related to t appears on both sides. Multiplying both sides by $(-1)^{\kappa(t,\theta)}$ (which we need, for otherwise the trees on the left hand side will not have the same sign as the trees on the right-hand side since trees on the LHS differ from trees on the RHS by exactly one root shift), we get

$$\sum_{\theta\sim t}(-1)^{\kappa(t,\theta)}\sum_{u,v\in\mathcal{T}\ u\circ v=\theta}\frac{1}{\sigma(v\circ u)}\theta^* = \sum_{\theta\sim t}(-1)^{\kappa(t,\theta)}\sum_{u,v\in\mathcal{T}\ u\circ v=\theta}\frac{1}{\sigma(v\circ u)}(v\circ u)^*.$$

As it stands, a tree δ on the LHS side will not have the same coefficient as the δ on the RHS. For example, a tree δ in the right hand side will have a coefficient equal to

$$\sum_{\omega}\frac{1}{\sigma(\delta)}\delta^* = \frac{k}{\sigma(\delta)}\delta^*$$

where k is the number of distinct sub-trees attached to the root of δ and the sum is over all trees ω that are precisely one root shift away from δ. However, δ will have a coefficient equal to

$$\delta^* \sum_{\omega} \frac{1}{\sigma(\omega)}$$

on the RHS, where the sum is over the number of distinct sub-trees ω attached to the root of δ. Since the symmetry coefficients of two distinct trees ω and δ are, in general, not equal, δ will not have the same coefficient on both sides of the equation. However, the coefficients can only be made equal if we multiply the LHS coefficient of δ by $\frac{\sigma(t)}{\sigma(\delta)}$ and the RHS coefficient of δ by $\frac{\sigma(t)}{\sigma(\omega)}$ (under the sum) to get

$$LHS: \sum_{\omega} \frac{\sigma(t)}{\sigma(\delta)\sigma(\omega)} \delta^* = \frac{deg(\delta)}{\sigma(\delta)} \delta^*$$

$$RHS: \sum_{\omega} \frac{\sigma(t)}{\sigma(\delta)\sigma(\omega)} \delta^* = \frac{deg(\delta)}{\sigma(\delta)} \delta^*$$

Note that if we were able to make the two coefficients equal by multiplying by any other factors then these different factors would have to be in the same ratio as $\frac{\sigma(t)}{\sigma(\delta)}$ and $\frac{\sigma(t)}{\sigma(\omega)}$ thus giving the same result, only multiplied through by a constant that will appear on both sides. Putting this back into the original equation, we find that the LHS is

$$\sum_{\delta \sim t} (-1)^{\kappa(t,\theta)} \frac{\sigma(t)}{\sigma(\delta)} \sum_{u,v \in \mathcal{T}\ u \circ v = \delta} \frac{1}{\sigma(v \circ u)} \delta^*.$$

The RHS becomes, after summing over all trees $\delta \sim t$ and reversing the argument given in the first half of the proof,

$$\sum_{\delta \sim t} (-1)^{\kappa(\delta,t)} \frac{deg(\delta)}{\sigma(\delta)} \delta^* = \sum_{\delta \sim t} (-1)^{\kappa(t,\theta)} \sum_{u,v \in \mathcal{T}\ u \circ v = \delta} \frac{1}{\sigma(v \circ u)} (v \circ u)^*.$$

Now $t \in \mathcal{T}_\Omega$. Therefore $\ker(ann_\Omega) \subseteq \mathcal{T}_\Omega$. $\square$

Corollary 72. *ann_Ω is a one-to-one map from $\mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^*$ onto $Ann(\mathcal{T}_\Omega^n)$.*

Proof. By Lemma 67, the kernel of ann_Ω on $\mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^*$ is $\{0\}$ so that ann_Ω is a one-to-one map from $\mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^*$ to $Ann(\mathcal{T}_\Omega^n)$. Since the dimensions agree, the map is also onto. $\square$

Theorem 73. *The map ann_Ω factorizes over the vector spaces $\mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^*$, $\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n$ and $Ann(\mathcal{T}_\Omega^n)$* as follows

$$\begin{array}{ccc} Ann(\mathcal{T}_\Omega^n) & \longleftarrow_{EP^*} & (\mathcal{T}_H^{n+1} \cap \mathcal{T}_{\tilde{\Omega}}^{n+1})^* \\ & & \\ ann_\Omega \uparrow & & \uparrow * \\ & & \\ \mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^* & \longrightarrow_{EP} & \mathcal{T}_H^{n+1} \cap \mathcal{T}_{\tilde{\Omega}}^{n+1} \end{array}$$

where ann_Ω , EP, and EP^* are bijective maps.

Proof. The proof follows directly from Corollary 72, Lemma 67 and property 2. of the transpose maps. □

Theorem 74. *A canonical basis for $\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n$ is given by $EP(\mathcal{T}_H^n \oplus Ann(\mathcal{T}_{\tilde{H}}^n)^*)$.*

Proof. The Theorem follows directly from Lemma 67. □

Theorems 70 and 73 gives us a very interesting insight into the structure of B-series. In particular, the relationship between energy-preserving B-series, $\mathcal{T}_H$, and the Energy-preserving and conjugate-Hamiltonian B-series, $\mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$. Recall that the kernel of the Hamiltonian map, $X_{[\cdot]}$, is the space of Energy-preserving trees so that

$$EP(\mathcal{T}_H) = ad_\bullet(\mathcal{T}_H) \subset \mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}.$$

This gives a simple and explicit construction of many (but not all) Energy-preserving and conjugate-to-Hamiltonian B-series of modified vector fields and $EP(\mathcal{T})$ gives them all. Moreover, $ad_\bullet^k(\mathcal{T}_H) \subset \mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$. We are able to understand the space of Energy-preserving and conjugate-Hamiltonian trees, $\mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$, with the simplest map possible. Although the space $\mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$ is known to be non-empty (see table 4.1), an integrator whose modified B-series lies in this space is yet to be exhibited. But Theorem 73 may provide a starting point to finding this integrator. Consider the B-series of the modified vector field corresponding to the AVF method (see example 47). All trees in this B-series are either Energy-preserving or group into Energy-preserving pairs. Applying $ad_\bullet$ to this B-series we get a new B-series whose linear combinations of trees lie in $\mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$. We now have a B-series which is Energy-preserving and conjugate-to-Hamiltonian that may have a numerical method as its exact solution. We are even able to calculate the (original) B-series that constitutes the exact

solution of this B-series. If $a(t)$ are the coefficients of a B-series corresponding to a numerical method, then the coefficients, $b(t)$, of the B-series of the modified vector field corresponding to the numerical method are given by

$$b(t) = a(t) - \sum_{j=2}^{|t|} \frac{1}{j!} \partial_b^{j-1} b(t),$$

where $\partial_b^{j-1} b(t)$ is the $(j-1)$th iterate of the Lie-derivative of B-series. Therefore, if we know the coefficients of the B-series of the modified differential equation, we are able to calculate coefficients of the original B-series corresponding to the numerical method by

$$a(t) = b(t) + \sum_{j=2}^{|t|} \frac{1}{j!} \partial_b^{j-1} b(t).$$

Denote the B-series of the modified vector field corresponding to the AVF method by $B(mod)_{AVF}$ and the B-series of the AVF method by B_{AVF}. Then, diagrammatically we have

$$\begin{array}{ccccc} AVF & \rightarrow & B_{AVF} & \rightarrow_{b(t)} & B(mod)_{AVF} \\ \updownarrow ? & & & & \downarrow ad_\bullet \\ \Phi_{New} & \leftarrow_? & B_{New} & \leftarrow_{a(t)} & B(mod)_{New} \end{array}$$

where Φ_{New} is a new numerical method that would be conjugate-to-symplectic and Energy-preserving.

Although $ad_\bullet$ is the infinitesimal analogue of conjugation, it should be noted that $ad_\bullet(B(mod)_{AVF})$ is not the same as conjugating the original AVF method (B-series, B_{AVF}) with the flow of the differential equation. This is because the coefficients $a(t)$ and $b(t)$ are, in general, different and $EP = ad_\bullet$ only on the space of energy-preserving trees. But a relationship between transforming the B-series of the modified differential equation and transforming the B-series of the numerical method is sure to exist, especially since we know exactly how the AVF method, B_{AVF}, $B(mod)_{AVF}$ and $B(mod)_{New}$ are all related. This may open the possibility to some transformation from the AVF method to Φ_{New}.

Chapter 7

Conclusions

Although this thesis has given a characterization and construction of two new subspaces of rooted trees, the description of the entire vector space of rooted trees is not yet complete. Two spaces yet to be characterized are $Ann(\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n)$ and $Ann(\mathcal{T}_{\tilde{\Omega}}^n)$. A characterization of these spaces will solidify our understanding of B-series in that conditions will be obtained for a B-series to lie in either of these spaces. It will be of particular interest to find conditions on the coefficients of $B(b, \mathbf{x})$ so that this B-series lies in the conjugate-to-Hamiltonian and Energy-preserving subspace.

Secondly, the Energy-preserving and Hamiltonian spaces of rooted trees may not be the only spaces of interest, although, for the most part they are the most useful. There may be B-series that preserve other geometric properties that are worth algebraically characterizing. A search for a characterization of the spaces $Ann(\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n)$ and $Ann(\mathcal{T}_{\tilde{\Omega}}^n)$ may bring other subspaces to light.

In Chapter 6, the natural maps $ad_\bullet$, $X_{[\cdot]}$ and EP were studied and their elementary properties were given. These maps were shown to give interesting relationships between the different spaces of rooted trees. This could be a useful tool not only for classifying B-series methods but showing how different B-series methods are related to each other and maybe even obtained from each other. These relationships were initially found through attempting to describe the space $Ann(\mathcal{T}_H^n)$ and a similar attempt to describe the spaces $Ann(\mathcal{T}_H^n \cap \mathcal{T}_{\tilde{\Omega}}^n)$ and $Ann(\mathcal{T}_{\tilde{\Omega}}^n)$ may also reveal a deeper structure and link between the known spaces. Although the space $\mathcal{T}_H \cap \mathcal{T}_{\tilde{\Omega}}$ is known to be non-empty, an integrator whose modified B-series lies in this space is yet to be exhibited. The natural maps and the way they link and relate different subspaces may provide a starting point to finding this integrator. There is possibly an analogue of these natural transformations on the map level which may be worthwhile pursuing in order to understand how numerical integrators themselves may be transformed.

Bibliography

[1] J C Butcher, An algebraic theory of integration methods, *Math. Comput.* **26** (1972), 79-106.

[2] J C Butcher, *Numerical Methods for Ordinary Differential Equations*, Wiley, Chichester, England ; Hoboken, NJ, 2008.

[3] J C Butcher and J M Sanz-Serna, The number of conditions for a Runge-Kutta method to have effective order p, *Applied Numerical Mathematics* **22** (1996), 103-111.

[4] M P Calvo and J M Sanz-Serna, Canonical B-series, *Numer. Math.* **67**, *(1994), 161-175*

[5] E Celledoni, R I McLachlan, B Owren, and G R W Quispel, Energy-preserving integrators and the structure of B-series, *Found. Comput. Math*, to appear.

[6] E Celledoni, R I McLachlan, B Owren, G R W Quispel, and W M Wright, Energy-preserving Runge-Kutta methods, *Mathematical Modelling and Numerical Analysis* **43** (2009), 645-649.

[7] P Chartier, E Faou, and A Murua, An algebraic approach to invariant preserving integrators: The case of quadratic and Hamiltonian invariants, *Numer. Math* **103** (2006), 575-590.

[8] P Chartier and A Murua, Preserving first integrals and volume forms of additively split systems, *IMA Journal of Numerical Analysis* **27** (2007), 381-405.

[9] E Faou, E Hairer, and T-L Pham, Energy conservation with non-symplectic methods: examples and counter-examples, *BIT* **44** (2004) 699-709.

[10] E Hairer, C Lubich, and G Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations,* Springer, Berlin, 1st ed., 2002, 2nd ed., 2006.

[11] E Hairer, Backward analysis of numerical integrators and symplectic methods, *Annals of Numerical Mathematics* **1** (1994), 107-132.

[12] E Hairer, Backward error analysis for multistep methods, *Numer. Math.* **84** (1999), 199-232.

[13] A Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 1st ed., 1997.

[14] G R W Quispel and D I McLaren, A new class of energy-preserving numerical integration methods, *J. Phys. A* **41** (2008) 045206 (7pp).

[15] S Roman, *Advanced Linear Algebra*, Springer, New York, 3rd ed., 2008.

[16] J M Sanz-Serna and L. Abia, Order conditions for canonical Runge-Kutta schemes, *SIAM J. Numer. Anal.* **28** (1991), 1081-1096.

[17] J E Scully, A search for improved numerical integration methods using rooted trees and splitting, MSc Thesis, La Trobe University, 2002.

[18] K. Yang, A basis for the intersection of subspaces, *Mathematics Magazine* 70 (4) (1997), 297.