

**On the Global Error
of Discretization Methods
for Ordinary Differential Equations**

Jitse Niesen
Trinity Hall
University of Cambridge

Submitted March 2004
Revised June 2004

A dissertation presented in the fulfilment of the requirements for the degree of
Doctor of Philosophy at the University of Cambridge

Abstract

Discretization methods for ordinary differential equations are usually not exact; they commit an error at every step of the algorithm. All these errors combine to form the global error, which is the error in the final result. The global error is the subject of this thesis.

In the first half of the thesis, accurate *a priori* estimates of the global error are derived. Three different approaches are followed: to combine the effects of the errors committed at every step, to expand the global error in an asymptotic series in the step size, and to use the theory of modified equations. The last approach, which is often the most useful one, yields an estimate which is correct up to a term of order h^{2p} , where h denotes the step size and p the order of the numerical method. This result is then applied to estimate the global error for the Airy equation (and related oscillators that obey the Liouville–Green approximation) and the Emden–Fowler equation. The latter example has the interesting feature that it is not sufficient to consider only the leading global error term, because subsequent terms of higher order in the step size may grow faster in time.

The second half of the thesis concentrates on minimizing the global error by varying the step size. It is argued that the correct objective function is the norm of the global error over the entire integration interval. Specifically, the L_2 norm and the L_∞ norm are studied. In the former case, Pontryagin’s Minimum Principle converts the problem to a boundary value problem, which may be solved analytically or numerically. When the L_∞ norm is used, a boundary value problem with a complementarity condition results. Alternatively, the Exterior Penalty Method may be employed to get a boundary value problem without complementarity condition, which can be solved by standard numerical software. The theory is illustrated by calculating the optimal step size for solving the Dahlquist test equation and the Kepler problem.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Jitse Niesen

Preface

It is a great pleasure to be able to write this preface. It not only signals that the thesis has come to completion, but it also gives me the opportunity to thank all the people without whose help I would not have come to this point.

My PhD supervisor is Arie Iserles. The research described in this thesis built on his ideas, and he has been a source of assistance and inspiration throughout. Per Christian Moan has put me on the right track in the first few months of my PhD, while Arie was abroad, and he has been very helpful ever since. Arie and Per Christian, thank you. I am also grateful to the many colleagues who discussed my work with me, including Sergio Blanes, Chris Budd, Stig Faltinsen, Tony Humphries, Tobias Jahnke, Erjen Lefeber, Christian Lubich, Marcel Oliver, Brynjulf Owren, Matthew Piggot, Divakar Viswanath, and Will Wright. The Numerical Analysis group of the University of Cambridge is a great place to work and I thank its present and former members for providing a pleasant atmosphere; they are Arie, Per Christian, Sergio, and Stig whom I have already mentioned, and Brad Baxter, Aurelian Bejancu, Nicoleta Bîlă, Coralia Cartis, Anita Faul, Simon Foucart, Hans-Martin Gutmann, Raphael Hauser, Mike Powell, Malcolm Sabin, and Alexei Shadrin. I am especially grateful for the company of Cora, with whom I shared an office for four years. She went through her PhD at the same time and was a great support for me.

I thank the VSB Funds and Nuffic in the Netherlands, the EPSRC in Britain, and various other Dutch sponsors for their financial support. I also thank Simon Malham for letting me continue my work on the thesis while at Heriot-Watt. Last but not least, I wish to thank my friends and family. They may not have contributed directly to the thesis, but they always were there to give support. I would not have managed to endure this ordeal without their help.

Thanks to all of you.

Edinburgh, March 2004
Jitse Niesen

Contents

Abstract	i
Preface	iii
1 Introduction	1
Part I Estimating the global error	6
2 The numerical solution of ordinary differential equations	7
2.1 Basic theory	7
2.2 Runge–Kutta methods	10
2.3 Butcher trees and B-series	11
2.4 Backward error analysis	15
2.5 Variable step-size methods	19
3 Three methods to estimate the global error	22
3.1 Lady Windermere’s fan	23
3.2 Asymptotic expansion	31
3.3 Modified equations and the global error	33
4 Applications of global error estimates	38
4.1 Error growth in periodic orbits	38
4.2 The Airy equation and related oscillators	39
4.3 The Emden–Fowler equation	48

Part II	Minimizing the global error	60
5	Formulation of the optimization problem	61
5.1	Minimizing the final error	62
5.2	Minimizing the error at all intermediate points	64
5.3	Formulation as an optimal control problem	66
6	Minimizing the error in the L_2 norm	69
6.1	Optimal control problems	70
6.2	Analytic treatment	74
6.3	Numerical treatment	77
7	Minimizing the maximal error	88
7.1	State-constrained optimal control problems	89
7.2	Analytic treatment	93
7.3	Numerical treatment	98
8	Conclusion and pointers for further research	107
8.1	Estimating the global error	107
8.2	Minimizing the global error	109
	Bibliography	111

Chapter 1

Introduction

Many phenomena, like the weather, the circulation of blood through one's body, or the movement of the planets in the solar system, can be modelled by differential equations. Typically, these equations are so complicated that we cannot write down the exact solution. However, we can use a numerical method to compute an approximate solution on the computer. The computed solution will usually deviate from the exact solution.

We will assume that we know the situation at a certain instant, and that we wish to compute how the situation changes subsequently (this is called an *initial value problem*). This is commonly achieved by *time-stepping*: starting with the known situation, we apply some formula to calculate the state a little while later, then we apply the formula again to find the situation still a bit later, and so on until we have covered the whole period of interest. Of course, the formula that we apply at every step is not exact, otherwise we would be able to find the exact solution. The error that is committed at a particular step is called the *local truncation error*. We will neglect round-off errors, which are caused by the fact that computers can store numbers with only a finite precision, because these are typically small compared to the truncation errors. We also neglect all other sources of errors, such as discrepancies between the mathematical model and the reality.

So, an error is committed in the first step. In the second step, another error is committed. However, we began the second step with a value which was slightly wrong, because of the error committed in the first step. So, the result of the second step is contaminated by both the error from the first step and the error from the second step. In general, the result of some step is contaminated by the errors from all the previous steps. The combined effect of all the local errors is called the *global error*. The global error is the subject of this thesis.

The global error is the crucial quantity to study if one wants to assess the quality of some numerical method. However, this is not so easy, as the following quote by Lambert [58, p. 57] indicates.

The LTE [local truncation error] and the starting errors accumulate to produce the GTE [global truncation error], but this accumulation process is very complicated, and we cannot hope to obtain any usable general expression for the GTE.

Therefore, we will settle for an *estimate* for the global error, instead of seeking an exact formula.

Generally, we can distinguish two classes of error estimates. Some estimates use the information obtained during the numerical solution of the differential equation, while other estimates use the analytic solution or at least some knowledge about it. These are called *a posteriori* estimates and *a priori* estimates, respectively. Both types have their respective strengths. If one has actually computed some numerical solution and wants to know how far it deviates from the exact solution, one probably should use an *a posteriori* estimate. However, for the purpose of comparing different methods, or that of devising methods which are particularly suited for a certain class of problems, there is often no choice but to use *a priori* estimates.

Various approaches for obtaining *a posteriori* estimates are discussed in the reviews by Skeel [80], Enright, Higham, Owren and Sharp [26, §4], and Calvo, Higham, Montijano and Randez [17]. However, in this thesis, we will only consider *a priori* estimates. The classical texts on this class of estimates include the work of Henrici [48, 49], Gragg [34], Stetter [81], Dahlquist [24], and Hairer and Lubich [40]. A couple of years ago, Hairer and Lubich [41, 42] made the connection to the new theory of modified equations. Recent developments are due to Viswanath [84], Moon, Szepessy, Tempone and Zouraris [69], and Iserles [54]. Indeed, the research of Iserles motivated the work described in this thesis.

Our aim in the first half of the thesis is to derive accurate *a priori* estimates for the global error. The accuracy of the estimates is probed by applying them to certain specific equations. The solutions of these equations are highly oscillatory. We can expect that the local errors oscillate as well, so they may cancel when combining to form the global error. This makes estimating the global error challenging.

The second half of the thesis illustrates another use of estimates for the global error. Recall that a local error is committed with every step. Naturally, the step size influences the local error, and indirectly the global error. So, we may try to keep the error in check by choosing the step size wisely.

Commonly, *a posteriori* estimates for the local error are used to choose the step size. The idea is that an efficient method commits errors of roughly equal size in every step; this is called *equidistribution*. But actually, we want to control the global error, not the local error. So the question becomes: how to choose the step size such that the global error is as small as possible?

This question has been considered before. Morrison [70], Greenspan, Hafner and Ribarič [35], Fujii [30], and Gear [31] show how to vary the step size in order to minimize the global error at some given instant. Butcher [14] extends their work to the situation where one can not only vary the step size, but also switch from one method to another, possibly with different order. The latest addition is the interesting paper by Moon, Szepessy, Tempone and Zouraris [68], who give a rigorous analysis of the complexity of their algorithm.

However, the size of the global error at a single instant is not always a good indicator for the quality of the solution, as we will show. Instead, one should look at the global error over the whole time interval. This does make the problem rather more complicated, and only few people have studied it from this angle. Eriksson, Estep, Hansbo and Johnson [28] study optimal step size strategies for discontinuous Galerkin methods; however, these methods are rarely used to solve initial value problems for ordinary differential equations. Lindberg [61] searches for the strategy which minimizes the maximum of the global error. He manages to characterize the optimal strategy for some equations using techniques from the calculus of variations. Dahlquist [24] attempts to use this characterization to construct a practical method for step size selection. Takaki, Utumi and Kawai [83] also look at the global error over the whole time interval, but they use a rather unnatural expression for evaluating the step size strategies. As they mention in the same paper, the root mean square value of the global error is of more interest.

In the second half of this thesis, we study the problem of determining the optimal step size strategy. We concentrate on two expressions for measuring the performance of different strategies: the root mean square value of the global error, as suggested by Takaki, Utumi and Kawai [83], and the maximal value of the global error, as used by Lindberg [61]. These two objectives correspond to the L_2 norm and the L_∞ norm of the global error, respectively. The resulting problem can be viewed as an optimal control problem (this point of view is also taken by Utumi, Takaki and Kawai).

Dikusar [25] and Gustafsson, Lundh and Söderlind [37, 38] also use Control Theory to select the step size, but their goals differ from ours. Dikusar tries to control the *local* error, while Gustafsson *et al.* want to eliminate violent oscillations in the step size sequence by adding a closed-loop control.

Above, we considered the problem of finding the optimal step size for solving a given differential equation. One could also ask for an optimal method for selecting the step size that does not use any *a priori* knowledge of the differential equation. This question can be posed in the framework of information-based complexity (see, for instance, the book of Werschulz [85] for details). However, the results obtained using this approach are not in agreement with the current practice. The work of Moon, Szepessy, Tempone and Zouraris [68], referred to above, may be considered as an effort to bridge this gap.

The plan of the thesis is as follows. There are two parts, each divided in three chapters. Part I is devoted to the estimation of the global error. We start with the standard theory on the numerical solution of ordinary differential equations in Chapter 2. In Chapter 3, we derive *a priori* estimates for the global error. These estimates are applied in Chapter 4. Part II starts with Chapter 5, in which we discuss how to measure the performance of different step size strategies. This chapter serves as motivation for the problems studied in the next two chapters. Chapter 6 investigates the step size strategy which minimizes the L_2 norm of the global error, both from an analytic and a numerical point of view. Chapter 7 does the same for the L_∞ norm. Finally, Chapter 8 brings the thesis to a conclusion by summarizing the main results and providing pointers for further research.

To end this introduction, we list the main contributions of this thesis to the field of Numerical Analysis. To the best of the author's knowledge, the following results are new:

- the global error estimate of Theorem 3.4 for fixed step-size methods;
- the explicit expression for the global error in terms of modified equations, as stated in Theorem 3.10 for general method and in Corollary 3.12 for Runge–Kutta methods;
- the *proof* of Theorem 4.1, describing the global error when tracking a periodic orbit (a different proof has been given before by Cano and Sanz-Serna [20]);
- Theorem 4.6, estimating the global error *up to order* h^{2p} , where h is the step size and p the order of the method, for the Airy equations and generalizations;
- Theorem 4.8, estimating the global error for the Emden–Fowler equation and yielding a concrete, nontrivial example where the leading global error term does not dominate the next term;

- the characterization of the step size which minimizes the L_2 norm of the global error in Theorem 6.5;
- the numerical computation of this step size, as described in Section 6.3;
- the corresponding results for the L_∞ norm in Theorem 7.5 and Section 7.3.

Parts of the material in Sections 3.3 and 4.3 were published in [72].

Part I

Estimating the global error

Chapter 2

The numerical solution of ordinary differential equations

The subject of this thesis is the numerical solution of initial value problems for ordinary differential equations, i.e., problems of the form

$$y' = f(t, y), \quad y(t_0) = y_0 \in \mathbf{R}^d. \quad (2.1)$$

In this chapter, we briefly describe the standard theory of the subject in order to set the stage for the investigation carried out in the subsequent chapters. For more background information, the reader is referred to the text books by Iserles [53] or Lambert [58], the book by Butcher [15], or the extensive two-volume monograph by Hairer, Nørsett and Wanner [44, 46]. Recent developments, especially the theory of modified equations, are treated in Hairer, Lubich and Wanner [43].

2.1 Basic theory

Given an initial time $t_0 \in \mathbf{R}$, a final time $t_f \in \mathbf{R}$, and a set $U \subset \mathbf{R}^d$, the function f is said to be *Lipschitz continuous* in $[t_0, t_f] \times U$ if there is a number L such that

$$\|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\|, \quad \text{for all } t \in [t_0, t_f], y_1, y_2 \in U, \quad (2.2)$$

where $\|\cdot\|$ denotes any vector norm on \mathbf{R}^d . If f is both continuous and Lipschitz continuous in $[t_0, t_f] \times U$, then the differential equation (2.1) has a unique solution (which may only be defined on a subinterval $[t_0, t)$ with $t < t_f$). We will always assume that this condition is indeed met. In fact, we will even assume that f is sufficiently smooth to justify all manipulations that we wish to perform. This is certainly the case if f is analytic, but in most situations less stringent requirements will suffice.

The *flow map* $\Phi_{t_0}^{t_f}$ is the function which associates to the initial value $y_0 \in \mathbf{R}^d$ the corresponding solution value at time t_f . It is defined by the relations

$$\Phi_{t_0}^{t_0}(y_0) = y_0 \quad \text{and} \quad \frac{d}{dt}\Phi_{t_0}^t(y_0) = f(t, \Phi_{t_0}^t(y_0)). \quad (2.3)$$

Since the solution of equation (2.1) is unique, the flow map enjoys the following composition property

$$\Phi_{t_2}^{t_3} \circ \Phi_{t_1}^{t_2} = \Phi_{t_1}^{t_3} \quad \text{for all } t_1, t_2, \text{ and } t_3. \quad (2.4)$$

If the original differential equation (2.1) is linear, meaning that the right-hand side $f(t, y)$ depends linearly on y , then the flow map is also linear. In this case, the corresponding matrix is commonly called the *fundamental matrix* or the *resolvent*.

The derivative of the flow map is called the *variational flow*. We denote it by $D\Phi_{t_0}^{t_f}$. By differentiating (2.3), we find that it satisfies

$$D\Phi_{t_0}^{t_0}(y_0) = I \quad \text{and} \quad \frac{d}{dt}D\Phi_{t_0}^t(y_0) = \frac{\partial f}{\partial y}(t, \Phi_{t_0}^t(y_0))D\Phi_{t_0}^t(y_0). \quad (2.5)$$

The variational flow determines the effect of perturbations of the differential equation, as specified in the Alekseev–Gröbner lemma below. The lemma provides a generalization of the variations-of-constants formula for linear differential equations. A proof can be found in e.g. [44, §I.14].

Lemma 2.1 (Alekseev–Gröbner). *Let the differential equation (2.1) and a perturbed equation*

$$\tilde{y}' = f(t, \tilde{y}) + \hat{f}(t, \tilde{y}), \quad \tilde{y}(t_0) = y_0 \quad (2.6)$$

be given. The solutions of (2.1) and the perturbed equation (2.6) are connected by

$$\tilde{y}(t_f) = y(t_f) + \int_{t_0}^{t_f} D\Phi_{t_0}^{t_f}(\tilde{y}(t)) \hat{f}(t, \tilde{y}(t)) dt. \quad (2.7)$$

Now suppose that we want to solve the differential equation (2.1) numerically. An easy way is the (*forward*) *Euler method*. We partition the interval $[t_0, t_f]$ in N subintervals of equal length. Denote the intermediate points by $t_k = t_0 + kh$, $k = 1, 2, \dots, N$, where $h = (t_f - t_0)/N$ denotes the length of each of the subintervals. We now approximate $y(t_1)$, the solution at time t_1 , by

$$y_1 = y_0 + hf(t_0, y_0). \quad (2.8)$$

The difference between this approximation and the real solution is called the *local error*. It is defined by

$$L_h(t_0, y_0) = y_1 - y(t_0 + h).$$

To calculate the local error for the Euler method, we first find y'' by differentiating the given equation (2.1). The result can be substituted in the Taylor series of y ,

$$y(t_0+h) = y(t_0) + hf(t_0, y_0) + \frac{1}{2}h^2 \left(\frac{\partial f}{\partial t}(t_0, y_0) + \frac{\partial f}{\partial y}(t_0, y_0) f(t_0, y_0) \right) + \mathcal{O}(h^3). \quad (2.9)$$

Together with the definition of $L_h(t_0, y_0)$, we find that the local error satisfies

$$L_h(t_0, y_0) = -\frac{1}{2}h^2 \left(\frac{\partial f}{\partial t}(t_0, y_0) + \frac{\partial f}{\partial y}(t_0, y_0) f(t_0, y_0) \right) + \mathcal{O}(h^3). \quad (2.10)$$

The $\mathcal{O}(h^3)$ term can be found by further expanding the Taylor series (2.9).

We apply the Euler step (2.8) iteratively to obtain approximations to the solution at t_2, t_3, \dots, t_N by the recursive formula

$$y_{k+1} = y_k + hf(t_k, y_k). \quad (2.11)$$

The local errors committed in all steps accumulate to the total difference between the numerical approximation y_N and the exact solution $y(t_f)$ at $t_f = t_N$. This difference is called the *global error*, and its definition is

$$G_h(t_f) = y_N - y(t_f) \quad \text{where} \quad t_f = t_N. \quad (2.12)$$

Note that the dependence of $G_h(t_f)$ on the initial data (t_0, y_0) is not explicit in this notation. For the Euler method, one can prove that $G_h(t_f) = \mathcal{O}(h)$.

At this point we can state the goal of the first part of this thesis: *We want to obtain precise estimates for the global error.* However, we first need to generalize the above description to include other numerical methods.

A general *one-step method* has the same form as the Euler method, except that (2.11) is replaced by

$$y_{k+1} = \Psi_h(t_k, y_k), \quad (2.13)$$

where Ψ_h is some function $\mathbf{R} \times \mathbf{R}^d \rightarrow \mathbf{R}^d$. As for the function f , we assume throughout the thesis that Ψ_h is sufficiently smooth.

The local error is again the difference between the numerical approximation and the exact solution, so

$$L_h(t, y) = \Psi_h(t, y) - \Phi_t^{t+h}(y). \quad (2.14)$$

The global error is defined by (2.12) as before. The local and global errors are related by the following fundamental theorem. For a proof, the reader is referred to [44, Theorem II.3.4].

Theorem 2.2. *If $L_h(t, y) = \mathcal{O}(h^{p+1})$, then $G_h(t_f) = \mathcal{O}(h^p)$.*

As indicated before, we always assume that f is sufficiently smooth (in the above theorem, it is enough to require that f be Lipschitz continuous in a neighbourhood of the solution). This theorem leads to the following definition: a method is said to have *order* p if $L_h(t, y) = \mathcal{O}(h^{p+1})$, and hence $G_h(t_f) = \mathcal{O}(h^p)$, for all sufficiently smooth f .

A further generalization is to consider *multistep methods*, which have the form $y_{k+\nu} = \Psi_h(t_k, y_k, \dots, y_{k+\nu-1})$. However, we will not concern ourselves with them in this work.

2.2 Runge–Kutta methods

An important family of one-step methods is formed by the so-called *Runge–Kutta methods* (abbreviated RK-methods). Given an integer ν , denoting the number of stages, a $(\nu \times \nu)$ -matrix $[a_{ij}]$ and two vectors $[b_i]$ and $[c_i]$ of length ν , the corresponding Runge–Kutta method is given by

$$\begin{aligned} \xi_i &= f\left(t_k + c_i h, y_k + h \sum_{j=1}^{\nu} a_{ij} \xi_j\right), \quad i = 1, 2, \dots, \nu, \\ y_{k+1} &= y_k + h \sum_{i=1}^{\nu} b_i \xi_i. \end{aligned} \tag{2.15}$$

It is customary to collect the coefficients in an arrangement called the *RK-tableau*. The tableau corresponding to general RK-method (2.15) is

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1\nu} \\ \vdots & \vdots & \ddots & \vdots \\ c_\nu & a_{\nu 1} & \cdots & a_{\nu\nu} \\ \hline & b_1 & \cdots & b_\nu \end{array}$$

For example, the Euler method (2.11) can be considered as a Runge–Kutta method with $\nu = 1$ stage, $a_{11} = c_1 = 0$, and $b_1 = 1$, so its RK-tableau is

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

We list below some other Runge–Kutta methods, which will be mentioned in this thesis.

- *Runge’s method* (also known as the explicit midpoint rule), a second-order method with two stages due to Carl Runge, is given by

$$\begin{aligned} \xi_1 &= f(t_k, y_k) & 0 & \left| \begin{array}{cc} 0 & 0 \end{array} \right. \\ \xi_2 &= f\left(t_k + \frac{1}{2}h, y_k + \frac{1}{2}h\xi_1\right) & 1/2 & \left| \begin{array}{cc} 1/2 & 0 \end{array} \right. \\ y_{k+1} &= y_k + h\xi_2 & & \left| \begin{array}{cc} 0 & 1 \end{array} \right. \end{aligned} \tag{2.16}$$

- *Heun's method*, a third-order method with three stages due to Karl Heun, is given by

$$\begin{array}{l}
 \xi_1 = f(t_k, y_k) \\
 \xi_2 = f(t_k + \frac{1}{3}h, y_k + \frac{1}{3}h\xi_1) \\
 \xi_3 = f(t_k + \frac{2}{3}h, y_k + \frac{2}{3}h\xi_2) \\
 y_{k+1} = y_k + \frac{1}{4}h(\xi_1 + 3\xi_3)
 \end{array}
 \quad
 \begin{array}{c|ccc}
 0 & 0 & 0 & 0 \\
 1/3 & 1/3 & 0 & 0 \\
 2/3 & 0 & 2/3 & 0 \\
 \hline
 & 1/4 & 0 & 3/4
 \end{array}
 \quad (2.17)$$

- Martin Kutta designed a fourth-order method with four stages which is so popular that it is commonly called *the standard Runge–Kutta method*. It is given by

$$\begin{array}{l}
 \xi_1 = f(t_k, y_k) \\
 \xi_2 = f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}h\xi_1) \\
 \xi_3 = f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}h\xi_2) \\
 \xi_4 = f(t_k + h, y_k + h\xi_3) \\
 y_{k+1} = y_k + \frac{1}{6}h(\xi_1 + 2\xi_2 + 2\xi_3 + \xi_4)
 \end{array}
 \quad
 \begin{array}{c|cccc}
 0 & 0 & 0 & 0 & 0 \\
 1/2 & 1/2 & 0 & 0 & 0 \\
 1/2 & 0 & 1/2 & 0 & 0 \\
 1 & 0 & 0 & 1 & 0 \\
 \hline
 & 1/6 & 1/3 & 1/3 & 1/6
 \end{array}
 \quad (2.18)$$

In principle, the local error of any Runge–Kutta method can be obtained in the same way as we did above for the Euler method. However, the calculation soon becomes highly cumbersome, so we need a device to keep the complexity in check. The theory of Butcher trees, explained in the next section, is one way to achieve this. An interesting alternative is provided by the framework of Albrecht [2], which is also explained in Lambert [58].

2.3 Butcher trees and B-series

As a first step, we transform the nonautonomous equation (2.1) to an autonomous equation by appending t to the dependent variables as follows,

$$\begin{bmatrix} y \\ t \end{bmatrix}' = \begin{bmatrix} f(t, y) \\ 1 \end{bmatrix}.$$

So it suffices to restrict ourselves to *autonomous* equations $y' = f(y)$.

The second device to simplify the computation is provided by the so-called *Butcher trees*. To introduce them, we look again at the Taylor series of the exact solution, cf. (2.9). In the autonomous case, the first terms of the series are

$$\begin{aligned}
 y(t+h) &= y(t) + hf + \frac{1}{2}h^2 f'(f) + \frac{1}{6}h^3 \left(f''(f, f) + f'(f'(f)) \right) \\
 &\quad + \frac{1}{24}h^4 \left(f'''(f, f, f) + 3f''(f, f'(f)) + f'(f''(f, f)) + f'(f'(f'(f))) \right) \\
 &\quad + \mathcal{O}(h^5).
 \end{aligned}
 \quad (2.19)$$

Here the arguments $(y(t))$ of the function f and its derivatives have been suppressed, and the derivatives of f are considered as multilinear operators. For instance, the term $f''(f, f'(f))$ stands for a vector whose j th component is given by

$$[f''(f, f'(f))]_j = \sum_{k=1}^d \sum_{\ell=1}^d \sum_{m=1}^d \frac{\partial^2 f_j}{\partial y_k \partial y_\ell}(y(t)) f_k(y(t)) \frac{\partial f_\ell}{\partial y_m}(y(t)) f_m(y(t)), \quad (2.20)$$

where the subscripts denote the various components of the vectors. Terms like the above, which appear in the Taylor expansion of the solution $y(t)$, are called *elementary differentials*.

The idea is now to represent every elementary differential with a (rooted) tree, a graph without cycles with one designated vertex called the *root*. For instance, the tree corresponding to (2.20) is



Here, and in the rest of this thesis, the root is the vertex depicted at the bottom.

With this convention, the Taylor series (2.19) can be written as:

$$\begin{aligned} y(t+h) = & y(t) + hF(\bullet)(y) + \frac{1}{2}h^2F(\bullet\bullet)(y) + \frac{1}{6}h^3\left(F(\bullet\swarrow\searrow)(y) + F(\bullet\ddot{\bullet})(y)\right) \\ & + \frac{1}{24}h^4\left(F(\bullet\swarrow\searrow\swarrow)(y) + 3F(\bullet\swarrow\searrow\ddot{\bullet})(y) + F(\bullet\ddot{\bullet}\ddot{\bullet})(y) + F(\bullet\swarrow\searrow\swarrow\swarrow)(y)\right) \\ & + \mathcal{O}(h^5). \end{aligned} \quad (2.22)$$

It remains to formalize the above.

We define the set of trees, denoted by \mathbf{T} , to be the smallest set with the following properties.

- The tree with one vertex, called the *unit tree* and denoted \bullet , is a member of \mathbf{T} ;
- If the trees τ_1, \dots, τ_n are in \mathbf{T} , then so is the tree formed by connecting all the roots of τ_1, \dots, τ_n to a new vertex, which becomes the root of the newly formed tree. This tree is denoted $[\tau_1, \dots, \tau_n]$. Note that some of the τ_1, \dots, τ_n may be equal, and that the result does not depend on the ordering, so $[\tau_1, \tau_2] = [\tau_2, \tau_1]$.

For example, the tree in (2.21) is in \mathbf{T} and it is denoted $[\bullet, [\bullet]]$.

We now define the following functions acting on the set of trees. They are also defined recursively.

- The number of vertices of a tree τ is called its *order*, and denoted $\rho(\tau)$. Its definition is

$$\rho(\bullet) = 1 \quad \text{and} \quad \rho([\tau_1, \dots, \tau_n]) = 1 + \rho(\tau_1) + \dots + \rho(\tau_n).$$

- The *symmetry coefficient* $\sigma(\tau)$ is defined by

$$\sigma(\bullet) = 1 \quad \text{and} \quad \sigma([\tau_1, \dots, \tau_n]) = \sigma(\tau_1) \dots \sigma(\tau_n) \mu_1! \mu_2! \dots,$$

where the integers μ_1, μ_2, \dots count equal trees among τ_1, \dots, τ_n .

- The *order product* $\gamma(\tau)$ is defined by

$$\gamma(\bullet) = 1 \quad \text{and} \quad \gamma([\tau_1, \dots, \tau_n]) = \rho(\tau) \gamma(\tau_1) \dots \gamma(\tau_n).$$

- Given the function $f : \mathbf{R}^d \rightarrow \mathbf{R}^d$, the *elementary differential* corresponding to a tree τ is the function $F(\tau) : \mathbf{R}^d \rightarrow \mathbf{R}^d$ with

$$F(\bullet)(y) = f(y), \quad F([\tau_1, \dots, \tau_n])(y) = f^{(n)}(y)(F(\tau_1)(y), \dots, F(\tau_n)(y)).$$

- Given a Runge–Kutta method (2.15), the *elementary weight* corresponding to a tree τ is $\varphi(\tau) = \sum_{i=1}^{\nu} b_i \tilde{\varphi}_i(\tau)$, where

$$\tilde{\varphi}_i(\bullet) = 1 \quad \text{and} \quad \tilde{\varphi}_i([\tau_1, \dots, \tau_n]) = \sum_{j_1, \dots, j_n=1}^{\nu} a_{ij_1} \dots a_{ij_n} \tilde{\varphi}_{j_1}(\tau_1) \dots \tilde{\varphi}_{j_n}(\tau_n).$$

Note that for the Euler method (2.11), we have $\varphi(\tau) = 0$ if $\tau \neq \bullet$. More generally, for an explicit Runge–Kutta method with ν stages, the elementary weight $\varphi(\tau)$ vanishes whenever $\rho(\tau) > \nu$.

In Table 2.1 we list all the trees of order up to 5 with the values of the functions defined above. It is often convenient to extend the domain of definition of the above functions with the empty tree, denoted \emptyset . The results of evaluating the functions on \emptyset can also be found in Table 2.1.

We can now write the Taylor expansion of the exact solution as

$$y(t+h) = y(t) + \sum_{\tau \in \mathbf{T}} \frac{h^{\rho(\tau)}}{\sigma(\tau) \gamma(\tau)} F(\tau)(y(t)). \quad (2.23)$$

Of course, this requires y to be analytic, which is the case if f is analytic. But also the numerical solution can be expanded in a Taylor series. For the Runge–Kutta method (2.15), we find

$$y_{k+1} = y_k + \sum_{\tau \in \mathbf{T}} \frac{\varphi(\tau)}{\sigma(\tau)} h^{\rho(\tau)} F(\tau)(y_k). \quad (2.24)$$

$\rho(\tau)$	τ	$\sigma(\tau)$	$\gamma(\tau)$	$F(\tau)$	$\varphi(\tau)$
0	\emptyset	1	1	id	1
1	\bullet	1	1	f	$\sum_i b_i$
2	$[\bullet] = \begin{array}{c} \bullet \\ \\ \bullet \end{array}$	1	2	$f'f$	$\sum_{ij} b_i a_{ij}$
3	$[\bullet, \bullet] = \begin{array}{c} \bullet \quad \bullet \\ \diagdown \quad \diagup \\ \bullet \end{array}$	2	3	$f''(f, f)$	$\sum_{ijk} b_i a_{ij} a_{ik}$
3	$[[\bullet]] = \begin{array}{c} \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	1	6	$f'f'f$	$\sum_{ijk} b_i a_{ij} a_{jk}$
4	$[\bullet, \bullet, \bullet] = \begin{array}{c} \bullet \quad \bullet \quad \bullet \\ \diagdown \quad \diagup \quad \diagup \\ \bullet \end{array}$	6	4	$f'''(f, f, f)$	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{il}$
4	$[[\bullet], \bullet] = \begin{array}{c} \bullet \quad \bullet \\ \diagdown \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	1	8	$f''(f'f, f)$	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{jl}$
4	$[[\bullet, \bullet]] = \begin{array}{c} \bullet \quad \bullet \\ \diagdown \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	2	12	$f'f''(f, f)$	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{jl}$
4	$[[[\bullet]]] = \begin{array}{c} \bullet \\ \\ \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	1	24	$f'f'f'f$	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{kl}$
5	$[\bullet, \bullet, \bullet, \bullet] = \begin{array}{c} \bullet \quad \bullet \quad \bullet \quad \bullet \\ \diagdown \quad \diagup \quad \diagup \quad \diagup \\ \bullet \end{array}$	24	5	$f''''(f, f, f, f)$	$\sum b_i a_{ij} a_{ik} a_{il} a_{im}$
5	$[[\bullet], \bullet, \bullet] = \begin{array}{c} \bullet \quad \bullet \quad \bullet \\ \diagdown \quad \diagup \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	2	10	$f'''(f'f, f, f)$	$\sum b_i a_{ij} a_{jk} a_{il} a_{im}$
5	$[[\bullet, \bullet], \bullet] = \begin{array}{c} \bullet \quad \bullet \quad \bullet \\ \diagdown \quad \diagup \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	2	15	$f''(f''(f, f), f)$	$\sum b_i a_{ij} a_{ik} a_{kl} a_{km}$
5	$[[[\bullet]], \bullet] = \begin{array}{c} \bullet \\ \\ \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	1	30	$f''(f'f'f, f)$	$\sum b_i a_{ij} a_{jk} a_{kl} a_{im}$
5	$[[\bullet], [\bullet]] = \begin{array}{c} \bullet \quad \bullet \\ \diagdown \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	2	20	$f''(f'f, f'f)$	$\sum b_i a_{ij} a_{jk} a_{il} a_{lm}$
5	$[[\bullet, \bullet, \bullet]] = \begin{array}{c} \bullet \quad \bullet \quad \bullet \\ \diagdown \quad \diagup \quad \diagup \\ \bullet \\ \\ \bullet \end{array}$	6	20	$f'f'''(f, f, f)$	$\sum b_i a_{ij} a_{jk} a_{jl} a_{jm}$
5	$[[[\bullet], \bullet]] = \begin{array}{c} \bullet \\ \\ \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	1	40	$f'f''(f'f, f)$	$\sum b_i a_{ij} a_{jk} a_{jl} a_{lm}$
5	$[[[\bullet, \bullet]]] = \begin{array}{c} \bullet \quad \bullet \\ \diagdown \quad \diagup \\ \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	2	60	$f'f'f''(f, f)$	$\sum b_i a_{ij} a_{jk} a_{kl} a_{km}$
5	$[[[[\bullet]]]] = \begin{array}{c} \bullet \\ \\ \bullet \\ \\ \bullet \\ \\ \bullet \\ \\ \bullet \end{array}$	1	120	$f'f'f'f'f$	$\sum b_i a_{ij} a_{jk} a_{kl} a_{lm}$

Table 2.1: Trees, and the values of various functions defined on them.

The local error of a Runge–Kutta method can now be found by simply subtracting (2.23) from (2.24). A succinct formulation of the order conditions follows: a Runge–Kutta method has order p if and only if its elementary weights satisfy $\varphi(\tau) = 1/\gamma(\tau)$ for all trees τ of order $\rho(\tau) \leq p$.

If we compare the Taylor series (2.23) and (2.24), we see that they have a similar form. This motivates the following definition. A *B-series* is the formal series

$$B(a, y) = a(\emptyset)y + \sum_{\tau \in \mathbf{T}} \frac{h^{\rho(\tau)}}{\sigma(\tau)} a(\tau) F(\tau)(y), \tag{2.25}$$

where a is a function $\mathbf{T} \cup \{\emptyset\} \rightarrow \mathbf{R}$. Different normalizations are in use in the literature. Here we follow the convention used in Butcher and Sanz-Serna [16] and Hairer, Lubich and Wanner [43].

With the definition (2.25), we can reformulate the series (2.23) and (2.24) as $y(t+h) = B(\frac{1}{\gamma}, y(t))$ and $y_{k+1} = B(\varphi, y_k)$, respectively; here, $\frac{1}{\gamma}$ stands for the

function mapping the tree τ to $\frac{1}{\gamma(\tau)}$.

Many operations on trees and B-series have been considered in the literature, see e.g. Butcher [15], but we will only need the *Lie derivative* of a B-series as defined in Hairer, Lubich and Wanner [43]. Recall that given a function a on \mathbf{R}^d , the Lie derivative of a with respect to f , denoted $\partial_f a$, is defined to be the derivative of a along solutions of $y' = f(y)$, so

$$\partial_f a(y_0) = \left. \frac{d}{dt} a(y(t)) \right|_{t=0}, \text{ where } y'(t) = f(y(t)), y(0) = y_0.$$

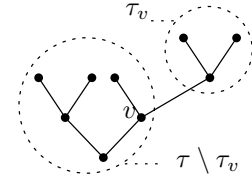
Mirroring this definition, we define the Lie derivative of the function $B(a, y)$ with respect to $B(b, y)$ as the B-series

$$B(\partial_b a, y_0) = h \left. \frac{d}{dt} B(a, y(t)) \right|_{t=0}, \text{ where } hy'(t) = B(b, y(t)), y(0) = y_0. \quad (2.26)$$

It turns out that there indeed exists a function $\partial_b a : \mathbf{T} \cup \{\emptyset\} \rightarrow \mathbf{R}$ such that (2.26) is satisfied if $b(\emptyset) = 0$. Moreover, we can compute this function explicitly as

$$\partial_b a(\tau) = \sum_{v \in V(\tau)} a(\tau \setminus \tau_v) b(\tau_v). \quad (2.27)$$

Here $V(\tau)$ is the set containing the $\rho(\tau)$ vertices of the tree τ . Furthermore, τ_v denotes the subtree of τ having v as its root, and $\tau \setminus \tau_v$ is whatever is left from τ after the subtree τ_v is removed. The picture to the right illustrates these definitions. For the trees up to order 3, the formulae for $\partial_b a$ are



$$\begin{aligned} \partial_b a(\bullet) &= a(\emptyset)b(\bullet) \\ \partial_b a(\updownarrow) &= a(\emptyset)b(\updownarrow) + a(\bullet)b(\bullet) \\ \partial_b a(\swarrow \searrow) &= a(\emptyset)b(\swarrow \searrow) + 2a(\updownarrow)b(\bullet) \\ \partial_b a(\updownarrow \updownarrow) &= a(\emptyset)b(\updownarrow \updownarrow) + a(\bullet)b(\updownarrow) + a(\updownarrow)b(\bullet). \end{aligned}$$

The Lie derivative is used to compute the modified equation, which is the subject of the next section.

2.4 Backward error analysis

There is a theory in numerical ODEs akin to the backward error analysis in numerical linear algebra initiated by Wilkinson. While a forward error analysis is concerned with the error in the solution space, backward error analysis consists of

a study of the error in the problem space. In the context of numerical ODEs, the development of the theory of backward error analysis is fairly recent and therefore not contained in older books. The reader is referred to Hairer, Lubich and Wanner [43] for a more extensive treatment than is possible here.

More precisely, given an autonomous differential equation $y' = f(y)$ and a numerical method $y_{k+1} = \Psi_h(y_k)$, the idea is to search for a *modified equation* $\tilde{y}' = \tilde{f}_h(\tilde{y})$ whose exact solution equals the numerical solution $\{y_k\}$ of the original equation at the grid points $\{t_k\}$. It turns out that usually the modified equation only exists in a formal sense, so instead we are looking for an equation whose exact solution is $\mathcal{O}(h^r)$ -close to the numerical solution where r is an arbitrary positive integer. If both the right-hand side f and the numerical method Ψ_h are analytic in some neighbourhood, then there exist functions $\tilde{f}_0, \tilde{f}_1, \dots, \tilde{f}_{r-1}$, such that the difference between the exact solution of

$$\tilde{y}' = \tilde{f}_0(\tilde{y}) + h\tilde{f}_1(\tilde{y}) + h^2\tilde{f}_2(\tilde{y}) + \dots + h^{r-1}\tilde{f}_{r-1}(\tilde{y})$$

and the numerical solution of the original equation is $\tilde{y}(t_n) - y_n = \mathcal{O}(h^r)$.

Note that the solution of the original equation is already $\mathcal{O}(h^p)$ -close to the numerical solution, where p is the order of the method. Hence, the modified equation in fact takes the form

$$\tilde{y}' = f(\tilde{y}) + h^p\tilde{f}_p(\tilde{y}) + h^{p+1}\tilde{f}_{p+1}(\tilde{y}) + \dots \quad (2.28)$$

However, the series on the right-hand side does not converge in general.

Example 2.3. The results in this chapter and the next are illustrated by a running example, which is also considered by Calvo, Murua and Sanz-Serna [19], Griffiths and Sanz-Serna [36], and Hairer, Lubich and Wanner [42, 43]. We are seeking a numerical solution of the scalar differential equation

$$y'(t) = (y(t))^2, \quad y(0) = 1. \quad (2.29)$$

The exact solution of this initial value problem is $y(t) = 1/(1-t)$, which has a singularity at $t = 1$.

Suppose that Runge's second order method (2.16) is used to solve (2.29). Then the modified equation is given by

$$\tilde{y}' = \tilde{y}^2 - \frac{3}{4}h^2\tilde{y}^4 + \frac{5}{4}h^3\tilde{y}^5 - \frac{7}{8}h^4\tilde{y}^6 + \dots \quad (2.30)$$

In Example 2.4, we will explain how to find this equation. Another method can be found in [43, §IX.1].

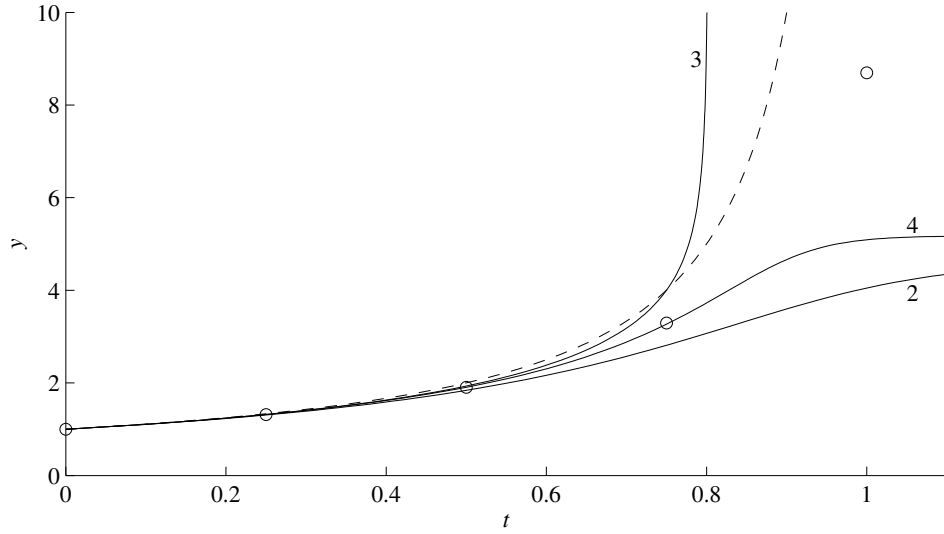


Figure 2.2: The exact solution of $y' = y^2$ (dashed line), the result of Runge's method with step size $h = 0.25$ (circles), and the exact solution of the modified equation (2.30) truncated after two, three, and four terms (solid lines marked 2, 3 and 4, respectively).

In Figure 2.2, we compare the numerical results with the exact solutions of the original equation $y' = y^2$ and the modified equation (2.30). It is clear that the latter approximates the numerical results better. However, this approximation breaks down at $t = 1$, reflecting the singularity in the exact solution (of course, the numerical solution has no finite-time singularity). \diamond

Let us now suppose that the numerical method can be expressed as a B-series (2.25), say $y_{k+1} = \Psi_h(y_k) = B(a, y_k)$, as is the case for Runge–Kutta methods. Then the modified equation (2.28) can also be written in terms of a B-series, namely as

$$\tilde{y}' = \frac{1}{h} B(b, \tilde{y}) = \sum_{\tau \in \mathbf{T}} \frac{h^{\rho(\tau)-1}}{\sigma(\tau)} b(\tau) F(\tau)(\tilde{y}), \quad (2.31)$$

but with different coefficients $b(\tau)$. In fact, we have $b(\emptyset) = 0$ and

$$b(\tau) = a(\tau) - \sum_{j=2}^{\rho(\tau)} \frac{1}{j!} \partial_b^{j-1} b(\tau), \quad (2.32)$$

where ∂_b^{j-1} denotes the $(j-1)$ -th iterate of the Lie derivative ∂_b defined in (2.26).

The coefficients for trees up to order 3 are

$$\begin{aligned}
 b(\bullet) &= a(\bullet) \\
 b(\mathfrak{1}) &= a(\mathfrak{1}) - \frac{1}{2}b(\bullet)^2 &= a(\mathfrak{1}) - \frac{1}{2}a(\bullet)^2 \\
 b(\mathfrak{2}) &= a(\mathfrak{2}) - b(\mathfrak{1})b(\bullet) - \frac{1}{3}b(\bullet)^3 = a(\mathfrak{2}) - a(\mathfrak{1})a(\bullet) + \frac{1}{6}a(\bullet)^3 \\
 b(\mathfrak{3}) &= a(\mathfrak{3}) - b(\mathfrak{1})b(\bullet) - \frac{1}{6}b(\bullet)^3 &= a(\mathfrak{3}) - a(\mathfrak{1})a(\bullet) + \frac{1}{3}a(\bullet)^3.
 \end{aligned} \tag{2.33}$$

Example 2.4. We return to the equation $y' = y^2$. The elementary differentials can easily be computed with the definition given on page 13.

$$F(\bullet)(y) = y^2, \quad F(\mathfrak{1})(y) = 2y^3, \quad F(\mathfrak{2})(y) = 2y^4, \quad F(\mathfrak{3})(y) = 4y^4.$$

The corresponding elementary weights for Runge's second order method are

$$\varphi(\bullet) = 1, \quad \varphi(\mathfrak{1}) = \frac{1}{2}, \quad \varphi(\mathfrak{2}) = \frac{1}{4}, \quad \varphi(\mathfrak{3}) = 0.$$

We can now use (2.33) to find the B-series coefficients of the modified equation.

$$b(\bullet) = 1, \quad b(\mathfrak{1}) = 0, \quad b(\mathfrak{2}) = -\frac{1}{12}, \quad b(\mathfrak{3}) = -\frac{1}{6}.$$

Hence, the modified equation is, cf. (2.31),

$$\tilde{y}' = \tilde{y}^2 - \frac{3}{4}h^2\tilde{y}^4 + \mathcal{O}(h^3).$$

To find the next term, we repeat this calculation for the trees of order four. The results are

$$\begin{aligned}
 F(\mathfrak{4}) &= 0, & \varphi(\mathfrak{4}) &= \frac{1}{8}, & b(\mathfrak{4}) &= 0; \\
 F(\mathfrak{5}) &= 4y^5, & \varphi(\mathfrak{5}) &= 0, & b(\mathfrak{5}) &= 0; \\
 F(\mathfrak{6}) &= 4y^5, & \varphi(\mathfrak{6}) &= 0, & b(\mathfrak{6}) &= \frac{1}{8}; \\
 F(\mathfrak{7}) &= 8y^5, & \varphi(\mathfrak{7}) &= 0, & b(\mathfrak{7}) &= \frac{1}{8}.
 \end{aligned}$$

So, the next term is $\frac{5}{4}h^3\tilde{y}^4$, and the modified equation reads

$$\tilde{y}' = \tilde{y}^2 - \frac{3}{4}h^2\tilde{y}^4 + \frac{5}{4}h^3\tilde{y}^4 + \mathcal{O}(h^4).$$

This agrees with (2.30). ◇

2.5 Variable step-size methods

In the previous sections, we divided the time interval $[t_0, t_f]$ in a number of subintervals of equal length. However, almost all programs used in practice do not space the intermediate points t_k uniformly, because it is often more efficient to concentrate them where the problem is harder. This means that, instead of having a constant step size h throughout the computation, we have a different step size h_k at every step. The time stepping formula changes from $y_{k+1} = \Psi_h(t_k, y_k)$ to

$$t_{k+1} = t_k + h_k \quad \text{and} \quad y_{k+1} = \Psi_{h_k}(t_k, y_k).$$

In this context, the global error depends not on a single variable h , but on all the step sizes h_k . We still denote the global error by $G_h(t_f)$, but now h represents the vector $(h_0, h_1, \dots, h_{k-1})$. Theorem 2.2, which states that a local error of order h^{p+1} implies a global error of order h^p , is still valid in the form

$$L_{h_k}(t_k, y_k) = \mathcal{O}(h_k^{p+1}) \quad \text{implies} \quad G_h(t_f) = \mathcal{O}(h_{\max}^p), \quad (2.34)$$

where $h_{\max} = \max_k h_k$ denotes the maximal step size.

For a further analysis of the variable step-size method, some knowledge on how the step size h_k is determined is required. The idea of most programs is to try and keep the local error, possibly normalized by dividing it by the step size, below a certain value specified by the user. The details, which may be rather intricate, often introduce a dependency of h_k on the size of the previous step, h_{k-1} . However, Stoffer and Nipp [82] prove that under some assumptions h_k is asymptotically independent of h_{k-1} . This suggests that we consider methods of the form

$$\begin{aligned} h_k &= \varepsilon_h h(t_k, y_k), \\ t_{k+1} &= t_k + h_k, \\ y_{k+1} &= \Psi_{h_k}(t_k, y_k), \end{aligned} \quad (2.35)$$

where ε_h is a reference step size, reflecting the user-specified tolerance. Furthermore, we will assume that the function h is bounded from below, i.e., that there is a $C > 0$ such that $h(t, y) \geq C$ for all t and y . This model is commonly used when analysing variable step-size methods; see for instance [44, §II.8] and [43, §VIII.2].

The variable step-size method can generally not be expanded in a B-series, even if Ψ_h is a Runge–Kutta method. That is, there is no coefficient function $a : \mathbf{T} \rightarrow \mathbf{R}$ such that

$$y_{k+1} = y_k + \sum_{\tau \in \mathbf{T}} \frac{a(\tau)}{\sigma(\tau)} \varepsilon_h^{\rho(\tau)} F(\tau)(y_k).$$

Hence the theory of Section 2.3 is not applicable to variable step-size methods, which complicates their analysis considerably.

The theory of Section 2.4 on backward error analysis remains valid for variable step-size methods, in that there exists a modified equation (2.28) whose (formal) solution coincides with the result of the numerical method, as proved by Hairer and Stoffer [45]. Of course, formula (2.31) expressing the modified equation in terms of the B-series of the numerical method cannot be applied.

Example 2.5. We continue with the equation $y' = y^2$. Again, Runge's second order method is employed, but now with variable step size. We suppose that the step size is given by $h_k = \varepsilon_h (y_k)^{-2}$, so that the method is indeed of the form (2.35). This causes the step size to decrease as the singularity is approached, as illustrated in Figure 2.3.

The modified equation can be found by using the same technique as explained for constant step-size methods in [43, §IX.1]. We find that it is given by

$$\tilde{y}' = \tilde{y}^2 - \frac{3}{4}\varepsilon_h^2 - \frac{1}{4}\varepsilon_h^3\tilde{y}^{-1} + \frac{1}{8}\varepsilon_h^4\tilde{y}^{-2} + \dots \quad (2.36)$$

Its solution, and the solution of the original equation, is compared to the numerical results in Figure 2.3. We see that the solution of the modified equation follows the numerical results closely.

One might naively think that the modified equation (2.36) for the variable step-size method can be deduced from the modified equation (2.30) for the constant step-size equation by the substitution $h = \varepsilon_h \tilde{y}^{-2}$. A comparison of (2.36) and (2.30) shows that this is not the case. However, and this holds in general, the term of order ε_h^p (remember that p stands for the order of the numerical method) can indeed be derived in this way. \diamond

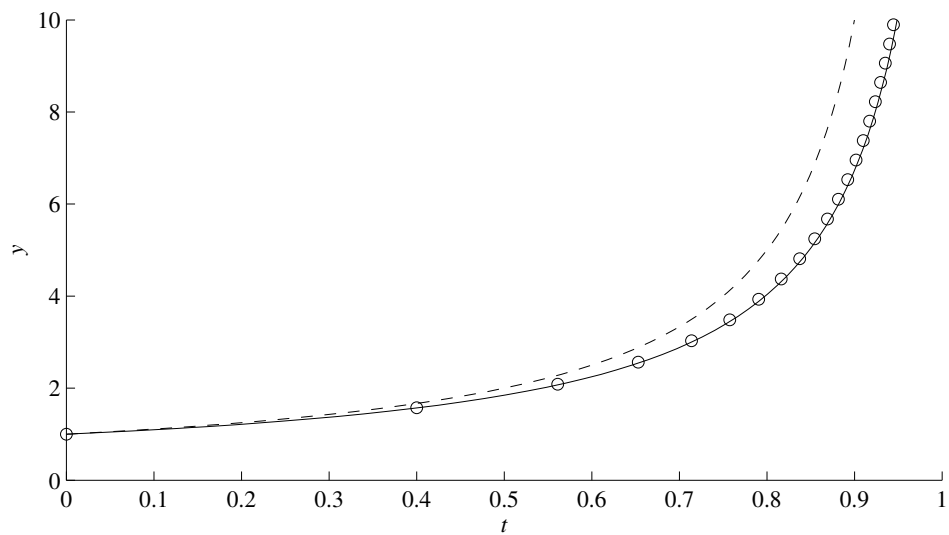


Figure 2.3: The exact solution of $y' = y^2$ (dashed line), the result of Runge's method with step size $h_k = \varepsilon_h (y_k)^{-2}$ where $\varepsilon_h = 0.4$ (circles), and the exact solution of the modified equation (2.36) truncated after the fourth term (solid line).

Chapter 3

Three methods to estimate the global error

This chapter forms the heart of the first part of the thesis. Its purpose is to find estimates for the global error $G_h(t)$, the difference between the result of the numerical method and the exact solution. The basic ingredient of the estimates is the local error of a method. Parts of the discussion are valid for both constant step-size methods of the form (2.13) and variable step-size methods given by (2.35), but the greater part assumes that the step size is kept constant.

Recall from Chapter 2 that the global error of a constant step-size method satisfies $G_h(t_f) = \mathcal{O}(h^p)$ if the method has order p . We are seeking an estimate $\tilde{G}_h(t_f)$ that at least satisfies $\tilde{G}_h(t_f) - G_h(t_f) = \mathcal{O}(h^{p+1})$. However, considerable efforts will be put in the quest for more precise estimators. Similarly, for variable step-size methods, we are seeking an estimate $\tilde{G}_{\varepsilon_h}(t_f)$ with $\tilde{G}_{\varepsilon_h}(t_f) - G_{\varepsilon_h}(t_f) = \mathcal{O}(\varepsilon_h^{p+1})$, or better.

In each of the three sections in this chapter, a different line of attack will be pursued. First, the most straightforward approach will be taken: for every step, we compute the contribution of the local error at that step to the global error at time t_f , and then we sum these contributions to find the global error. In Section 3.2, we expand the global error in powers of h , and we seek to obtain the terms in this power series. The third approach uses the theory of backward error analysis, which was explained in Section 2.4, to get yet another estimate for the global error.

3.1 Lady Windermere's fan

The approach described here is rather straightforward: we add the contributions of the local errors committed at every step to find the global error. Many of the classical text books (see, for example, Isaacson and Keller [52]) use this approach. The usual result is a bound of the form

$$\|G_h(t_f)\| \leq Ch_{\max}^p e^{L(t_f-t_0)}, \quad (3.1)$$

where L is a Lipschitz constant as defined in (2.2), and C is some constant depending on the size of the derivatives of f . In fact, this is the standard way to prove Theorem 2.2.

It is well known that the bound (3.1) often yields a gross overestimate of the error. The problem is that the Lipschitz constant L often does not describe the function f well. Hence, the estimate can be improved if one takes a more sophisticated approach, based for instance on a one-sided Lipschitz condition. Details can be found in e.g. Hairer, Lubich and Wanner [44, §I.10], who follow Dahlquist [23], or Iserles and Söderlind [55].

None of these approaches take into account that the errors committed at various steps may (partially) cancel each other. This is why their results may still be too pessimistic. In other words, these approaches give *bounds* on the global error, while we are looking for *estimates*.

Below, we will derive an estimate of the global error. The idea is illustrated in Figure 3.1. At every time t_k , a local error is committed. This error is transported along the flow of the differential equation to the end of the integration interval, and then all the transported local errors are summed to get the global error $G_h(t_f)$. As the local errors are quite small, the linearized flow can be used to transport them. This gives the following error estimate.

Lemma 3.1. *If a one-step method of order p is employed, then the global error satisfies $G_h(t_f) = \tilde{G}_h(t_f) + \mathcal{O}(h_{\max}^{2p})$, where \tilde{G}_h is given by*

$$\tilde{G}_h(t_f) = \sum_{k=0}^{N-1} D\Phi_{t_{k+1}}^{t_f}(y(t_{k+1}))L_{h_k}(t_k, y(t_k)), \text{ with } t_N = t_f. \quad (3.2)$$

Proof. We decompose the global error as suggested by Figure 3.1,

$$G_h(t_f) = y_N - \Phi_{t_0}^{t_f}(y_0) = \sum_{k=0}^{N-1} R_k, \text{ where } R_k = \Phi_{t_{k+1}}^{t_f}(y_{k+1}) - \Phi_{t_k}^{t_f}(y_k).$$

Note that R_k represents the local error committed at t_{k+1} , transported to t_f . Using the definition of the local error (2.14) and the flow composition property (2.4), the

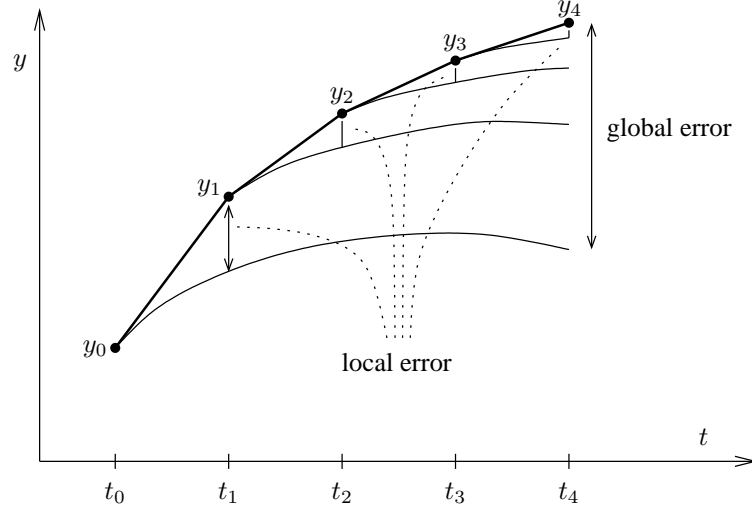


Figure 3.1: The local errors committed at every step are combined to form the global error. This picture is called *Lady Windermere's fan*, after a play by Oscar Wilde, in [44].

expression for R_k can be rewritten as

$$R_k = \Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k) + L_{h_k}(t_k, y_k)) - \Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k)).$$

The main theorem of calculus allows us to write this as an integral,

$$R_k = \int_0^1 D\Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k) + \xi L_{h_k}(t_k, y_k)) L_{h_k}(t_k, y_k) d\xi. \quad (3.3)$$

The local error is $L_{h_k}(t_k, y_k) = \mathcal{O}(h_k^{p+1})$ because the numerical method has order p . Furthermore, the smoothness of f implies that the flow map Φ is also smooth, so

$$D\Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k) + \xi L_{h_k}(t_k, y_k)) = D\Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k)) + \mathcal{O}(h_k^{p+1}).$$

Hence we can approximate the integral in (3.3) as follows.

$$R_k = D\Phi_{t_{k+1}}^{t_f} (\Phi_{t_k}^{t_{k+1}}(y_k)) L_{h_k}(t_k, y_k) + \mathcal{O}(h_k^{2p+2}). \quad (3.4)$$

We also have $y_k = \Phi_{t_0}^{t_k}(y_0) + G_h(t_k) = \Phi_{t_0}^{t_k}(y_0) + \mathcal{O}(h_{\max}^p)$ because of Theorem 2.2. Thus, the composition property (2.4) of the flow map implies that $\Phi_{t_k}^{t_{k+1}}(y_k) = \Phi_{t_0}^{t_{k+1}}(y_0) + \mathcal{O}(h_{\max}^p)$. Hence, we can rewrite (3.4) as

$$R_k = D\Phi_{t_{k+1}}^{t_f} (\Phi_{t_0}^{t_{k+1}}(y_0)) L_{h_k}(t_k, y_k) + \mathcal{O}(h_{\max}^{2p+1}).$$

Furthermore, using the smoothness of the numerical method, and hence of the local error, we can write

$$R_k = D\Phi_{t_{k+1}}^{t_f}(\Phi_{t_0}^{t_{k+1}}(y_0))L_{h_k}(t_k, \Phi_{t_0}^{t_k}(y_0)) + \mathcal{O}(h_{\max}^{2p+1})$$

Finally, the global error is retrieved via $G_h(t_f) = \sum_k R_k$, where the summation goes over $N \leq (t_f - t_0)/h_{\max}$ terms. This gives us (3.2). \square

The same argument, up to (3.4), can also be used to prove Theorem 2.2.

It should be borne in mind that the constant hidden in the \mathcal{O} symbol in the equation $G_h(t_f) = \tilde{G}_h(t_f) + \mathcal{O}(h_{\max}^{2p})$ depends on the final time t_f . This constant often grows exponentially as t_f increases. In this case, the estimate $\tilde{G}_h(t_f)$ becomes meaningless from a certain point (see for instance Example 4.7).

Example 3.2. We return to the same example as in Chapter 2: solving the equation $y' = y^2$, $y(0) = 1$, with Runge's second order method. For the moment, we restrict ourselves to the case where the step size is constant. We want to use Lemma 3.1 to estimate the global error.

We set $t_0 = 0$, so $t_f = Nh$. The estimate for the global error becomes $G_h(t_f) = \sum_{k=0}^{N-1} A_k + \mathcal{O}(h^4)$ where

$$A_k = D\Phi_{(k+1)h}^{Nh}(y((k+1)h))L_h(kh, y(kh)).$$

Recall that the exact solution is $y(t) = 1/(1-t)$. The flow and the variational flow of the equation $y' = y^2$ are

$$\Phi_s^t(y) = \frac{y}{1-(t-s)y} \quad \text{and} \quad D\Phi_s^t(y) = \frac{1}{(1-(t-s)y)^2}. \quad (3.5)$$

We also need to know the local error of Runge's method. This can be determined by comparing the B-series of the exact and the numerical solution, or by a straightforward Taylor expansion. Both methods yield

$$L_h(t, y) = -\frac{3}{4}h^3y^4 - h^4y^5 + \mathcal{O}(h^6). \quad (3.6)$$

Combining all these expressions, we find that

$$\begin{aligned} A_k &= -\left(\frac{1-(k+1)h}{1-Nh}\right)^2 \left(\frac{3}{4}h^3 \left(\frac{1}{1-kh}\right)^4 + h^4 \left(\frac{1}{1-kh}\right)^5\right) \\ &= -\frac{1}{(1-t_f)^2} \left(\frac{3}{4}h^3 \left(\frac{1}{1-kh}\right)^2 - \frac{1}{2}h^4 \left(\frac{1}{1-kh}\right)^3\right) + \mathcal{O}(h^5). \end{aligned} \quad (3.7)$$

We now need to sum the A_k . However, here we encounter a problem: the sum $\sum_k (1-kh)^{-n}$ cannot be expressed in elementary functions. One possible way

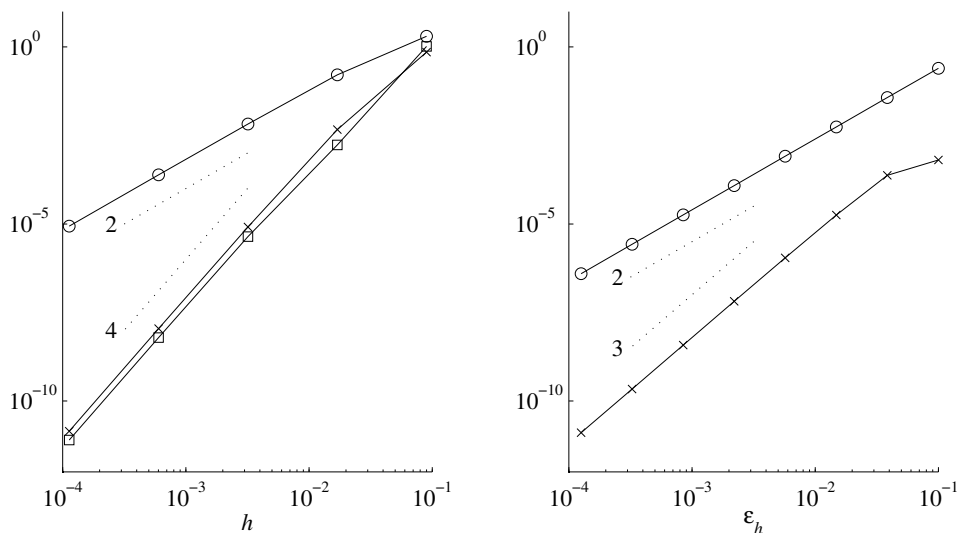


Figure 3.2: The plot on the left shows the global error at $t_f = 0.9$ committed by Runge's method with constant step size when applied to $y' = y^2$ (circles), the difference between the estimate (3.8) and the true global error (crosses), and the difference between (3.11) and the true global error (squares). The dotted reference lines have slopes 2 and 4. On the right, the global error with step size $h_k = \varepsilon_h/y_k^2$ (circles) and the difference between the estimate (3.15) and the global error (crosses) are shown. The dotted reference lines have slopes 2 and 3.

around this obstacle is to use the Euler–MacLaurin theorem, as will be explained in Example 3.5. We can also use the digamma function ψ , which is defined by $\psi(x) = \Gamma'(x)/\Gamma(x)$, where $\Gamma(x)$ denotes Euler's gamma function (see for instance [1, Ch. 6]). In terms of the digamma function,

$$\begin{aligned}
 G_h(t_f) &= \sum_{k=0}^{N-1} A_k + \mathcal{O}(h^4) \\
 &= -\frac{h}{(1-t_f)^2} \left(\frac{3}{4} \psi' \left(\frac{1-t_f}{h} + 1 \right) - \frac{3}{4} \psi' \left(\frac{1}{h} + 1 \right) \right. \\
 &\quad \left. + \frac{1}{4} \psi'' \left(\frac{1-t_f}{h} + 1 \right) - \frac{1}{4} \psi'' \left(\frac{1}{h} + 1 \right) \right) + \mathcal{O}(h^4). \tag{3.8}
 \end{aligned}$$

This estimate is illustrated on the left-hand side of Figure 3.2, which shows the difference between the estimate (3.8) and the exact global error at $t_f = 0.9$ for various values of the step size h . Comparison with the reference line shows that the remainder term is indeed of order h^4 . \diamond

As illustrated in the above example, it is often rather difficult to evaluate the summation in (3.2). This problem can be mitigated by converting the sum to an integral. The Euler–MacLaurin theorem achieves this conversion. A proof of this theorem can be found in [57, §7.7].

Theorem 3.3 (Euler–MacLaurin). *Let ψ be a function on $[0, 1]$. If ψ is C^{2n} , then for some $\xi \in (0, 1)$ we have*

$$\int_0^1 \psi(t) dt = \frac{1}{2}(\psi(0) + \psi(1)) - \sum_{k=1}^{n-1} \frac{b_{2k}}{(2k)!} (\psi^{(2k-1)}(1) - \psi^{(2k-1)}(0)) - \frac{b_{2n}}{(2n)!} \psi^{(2n)}(\xi), \quad (3.9)$$

where b_k denote the Bernoulli numbers.

Two equivalent definitions for the Bernoulli numbers are

$$\sum_{k=0}^{\infty} \frac{b_k}{k!} x^k = \frac{x}{e^x - 1} \quad \text{and} \quad \sum_{k=0}^n \binom{n+1}{k} b_k = 0 \quad (\text{for } n \geq 1).$$

The first values are $b_0 = 1$, $b_1 = -\frac{1}{2}$, $b_2 = \frac{1}{6}$, $b_3 = 0$, $b_4 = -\frac{1}{30}$, $b_5 = 0$, $b_6 = \frac{1}{42}$. In general, b_k vanishes for odd $k \geq 3$.

The Euler–MacLaurin theorem can readily be used if the numerical method employs a constant step size.

Theorem 3.4. *If we are using a constant step-size method of order p to solve the equation $y' = f(t, y)$, then the global error satisfies $G_h(t_f) = \tilde{G}_h(t_f) + \mathcal{O}(h^{2p})$ with*

$$\begin{aligned} \tilde{G}_h(t_f) &= \frac{1}{h} \int_{t_0}^{t_f-h} \rho_h(t) dt + \frac{1}{2}(\rho_h(t_f - h) + \rho_h(t_0)) \\ &\quad + \sum_{k=1}^{p-1} \frac{b_{2k}}{(2k)!} h^{2k} (\rho_h^{(2k-1)}(t_f - h) - \rho_h^{(2k-1)}(t_0)), \end{aligned}$$

where the function $\rho_h : [t_0, t_f - h] \rightarrow \mathbf{R}^d$ is defined by

$$\rho_h(t) = D\Phi_{t+h}^{t_f}(y(t+h)) L_h(t, y(t)).$$

Proof. Summing the Euler–MacLaurin formula (3.9) yields the following summation formula (see e.g. [57, §7.4])

$$\begin{aligned} \int_0^N \psi(x) dx &= \frac{1}{2}\psi(0) + \sum_{k=1}^{N-1} \psi(k) + \frac{1}{2}\psi(N) \\ &\quad - \sum_{k=1}^{n-1} \frac{b_{2k}}{(2k)!} (\psi^{(2k-1)}(N) - \psi^{(2k-1)}(0)) - \frac{b_{2n}}{(2n)!} N\psi^{(2n)}(\xi). \quad (3.10) \end{aligned}$$

For constant step-size methods, we have $t_k = t_0 + kh$. Hence, (3.2) takes the form $\tilde{G}_h(t_f) = \sum_{k=0}^{N-1} \rho_h(t_0 + kh)$. The theorem follows by applying (3.10), rescaled to the time interval $[t_0, t_f - h]$. \square

Example 3.5. We continue Example 3.2. From (3.5) and (3.6), it follows that the function ρ_h in Theorem 3.4 is given by, cf. (3.7),

$$\rho_h(t) = -\frac{1}{(1-t_f)^2} \left(\frac{3h^3}{4(1-t)^2} - \frac{h^4}{2(1-t)^3} \right) + \mathcal{O}(h^5).$$

Integrating this expression is a tedious but straightforward computation, which results in

$$\int_0^{t_f-h} \rho_h(t) dt = -\frac{3h^3 t_f}{4(1-t_f)^3} + \frac{h^4}{4(1-t_f)^2} \left(\frac{1}{(1-t_f)^2} - 4 \right) + \mathcal{O}(h^5).$$

Hence, we find that the global error satisfies

$$\begin{aligned} G_h(t_f) &= \frac{1}{h} \int_{t_0}^{t_f-h} \rho_h(t) dt + \frac{1}{2} (\rho_h(t_f - h) + \rho_h(t_0)) + \mathcal{O}(h^4) \\ &= -\frac{3t_f}{4(1-t_f)^3} h^2 + \frac{5t_f(2-t_f)}{8(1-t_f)^4} h^3 + \mathcal{O}(h^4). \end{aligned} \quad (3.11)$$

Again, we plot the difference between the above estimate and the true global error in the left-hand plot of Figure 3.2. This figure supports the error estimate (3.11). Note that the estimates (3.8) and (3.11) are not identical; the difference is subsumed in the $\mathcal{O}(h^4)$ remainder terms. \diamond

The situation is more complicated for variable step-size methods. The following result is a straightforward application of the Euler–MacLaurin theorem, but it discards all information of order h^{p+1} , and retains only the leading error term, which has order h^p .

Theorem 3.6. *Suppose that we are employing a variable step-size method of the form (2.35). If the local error is $L_h(t, y) = h^{p+1} \ell(t, y) + \mathcal{O}(h^{p+2})$ and the step size function h is bounded below, then the global error satisfies the estimate $G_{\varepsilon_h}(t_f) = \varepsilon_h^p g(t_f) + \mathcal{O}(\varepsilon_h^{p+1})$ with*

$$g(t_f) = \int_{t_0}^{t_f} h(t, y(t))^p D\Phi_t^{t_f}(y(t)) \ell(t, y(t)) dt. \quad (3.12)$$

Proof. According to Lemma 3.1, the global error satisfies

$$G_{\varepsilon_h}(t_f) = \sum_{k=0}^{N-1} g_k + \mathcal{O}(h_{\max}^{2p}) \quad \text{with} \quad g_k = D\Phi_{t_{k+1}}^{t_f}(y(t_{k+1})) L_{h_k}(t_k, y(t_k)).$$

We have $D\Phi_{t_k}^{t_{k+1}} = I + \mathcal{O}(h_k)$ because of the differential equation (2.5) determining the variational flow. So we can write

$$\begin{aligned} g_k &= D\Phi_{t_k}^{t_f}(y(t_k)) L_{h_k}(t_k, y(t_k)) + \mathcal{O}(h_k^{p+2}) \\ &= h_k^{p+1} D\Phi_{t_k}^{t_f}(y(t_k)) \ell(t_k, y(t_k)) + \mathcal{O}(h_k^{p+2}). \end{aligned}$$

Now consider the sequence $\{t_k\}$. It is generated by the recurrence

$$t_{k+1} = t_k + h_k, \quad y_{k+1} = \Psi_{h_k}(t_k, y_k) \quad \text{and} \quad h_k = \varepsilon_h h(t_k, y_k).$$

Since the method is convergent, we have $y_k = y(t_k) + \mathcal{O}(\varepsilon_h)$. If the sequence $\{\hat{t}_k\}$ is defined by

$$\hat{t}_0 = t_0 \quad \text{and} \quad \hat{t}_{k+1} = \hat{t}_k + \varepsilon_h h(\hat{t}_k, y(\hat{t}_k)), \quad (3.13)$$

then $\hat{t}_k = t_k + \mathcal{O}(\varepsilon_h)$. But (3.13) is the Euler method (2.11) with constant step size ε_h applied to the differential equation $\frac{dt}{d\kappa} = h(t, y(t))$, where t is the dependent variable and κ is the independent variable. Denoting the solution of this equation by $t(\kappa)$, we have $\hat{t}_k = t(k\varepsilon_h) + \mathcal{O}(\varepsilon_h)$ because the Euler method is a convergent method. Hence $g_k = \psi(k\varepsilon_h) + \mathcal{O}(\varepsilon_h^{p+2})$ with

$$\psi(\kappa) = h(t(\kappa), y(t(\kappa)))^{p+1} D\Phi_{t(\kappa)}^{t_f}(y(t(\kappa))) \ell(t(\kappa), y(t(\kappa))).$$

Going back to the global error, we have

$$G_{\varepsilon_h}(t_f) = \sum_{k=0}^{N-1} \psi(k\varepsilon_h) + \mathcal{O}(\varepsilon_h^{p+1}), \quad (3.14)$$

since $N = \mathcal{O}(\varepsilon_h^{-1})$. It follows from the Euler-MacLaurin summation formula (3.10) with $n = 1$ that

$$\int_0^N \psi(k\varepsilon_h) dk = \frac{1}{2}\psi(0) + \sum_{k=1}^{N-1} \psi(k\varepsilon_h) + \frac{1}{2}\psi(N\varepsilon_h) + \frac{1}{12}Nh^2\psi''(\xi),$$

for some $\xi \in [0, N]$. This is the (composite) trapezoid rule. We now use this to rewrite (3.14), remembering that $\psi(x) = \mathcal{O}(\varepsilon_h^{p+1})$. This gives

$$G_{\varepsilon_h}(t_f) = \int_0^N \psi(k\varepsilon_h) dk + \mathcal{O}(\varepsilon_h^{p+1}).$$

Finally, changing the variable of integration from k to t gives the estimate in the theorem. \square

Example 3.7. We return to Runge's second-order method applied to the differential equation $y' = y^2$, but this time we vary the step size according to $h_k = \varepsilon_h/y_k^2$ (as in Example 2.5), so we have $h(t, y) = y^{-2}$. The variational flow is given in (3.5), and from (3.6) we have $\ell(t, y) = -\frac{3}{4}y^4$. Hence the integral (3.12) evaluates to

$$g(t_f) = -\frac{3}{4} \int_0^{t_f} \left(\frac{1-t}{1-t_f} \right)^2 dt = -\frac{1}{4} \left(\frac{1}{(1-t_f)^2} - (1-t_f) \right).$$

It follows from Theorem 3.6 that the global error satisfies

$$G(t_f) = -\frac{1}{4} \varepsilon_h^2 \left(\frac{1}{(1-t_f)^2} - (1-t_f) \right) + \mathcal{O}(\varepsilon_h^3). \quad (3.15)$$

The accuracy of this estimate is shown in the right-hand plot of Figure 3.2. \diamond

Theorem 3.6 also holds if the step size happens to be constant. In that case, the result is

$$G_h(t_f) = h^p \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) \ell(t, y(t)) dt + \mathcal{O}(h^{p+1}). \quad (3.16)$$

This is also an immediate corollary of Theorem 3.4.

The estimate (3.16) for the global error is not new; Iserles [54] provides a proof of it. In fact, the research described in this thesis was inspired by this paper.

In an earlier paper, Viswanath [84] proves the following bound on the global error. Given any t_f and $\varepsilon > 0$, and assuming that the magnitude of the local error is bounded above by Kh^{p+1} , the error satisfies

$$G_h(t) < (\mathcal{E}(t) + \varepsilon)Kh^p \quad (3.17)$$

for any $t \in [t_0, t_f]$ and sufficiently small h . In the framework adopted in this thesis, Viswanath's definition for $\mathcal{E}(t)$ is equivalent to

$$\mathcal{E}(t_f) = \int_{t_0}^{t_f} \|D\Phi_t^{t_f}(y(t))\| dt.$$

With this definition, the bound (3.17) on the global error follows from (3.16). Viswanath [84] uses (3.17) to prove that the global error as a function of t bounded above by a linear function or by a constant in various situations, e.g. stable, hyperbolic cycles.

Differentiating (3.12) gives the following equation for the leading term of the global error

$$g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t, y(t))^p \ell(t, y(t)), \quad g(t_0) = 0. \quad (3.18)$$

This differential equation is well known in the literature, see e.g. [44, §II.8]. In the case of constant step size, where we have $h \equiv 1$, it goes back to Henrici [48, Thm. 3.4]. The equation (3.18) will play an essential role in Part II of this thesis.

3.2 Asymptotic expansion

In this section, we assume that the step size h does not vary.

Another approach is to expand the global error $G_h(t)$ in a power series

$$G_h(t) \sim g_p(t) h^p + g_{p+1}(t) h^{p+1} + g_{p+2}(t) h^{p+2} + \dots \quad (3.19)$$

As denoted by the symbol \sim , this is an *asymptotic expansion* in the sense of Poincaré, meaning that

$$G_h(t) = \sum_{k=p}^N g_k(t) h^k + \mathcal{O}(h^{N+1}) \quad \text{for every } N,$$

but the infinite series $\sum_{k=p}^{\infty} g_k(t) h^k$ diverges in general.

Henrici [48] and Gragg [34] are among the first who rigorously analysed this approach. We will follow the treatment of Hairer, Nørsett and Wanner [44, §II.8].

Theorem 3.8 (Gragg [34]). *If we apply a convergent method with constant step size, then the global error admits an asymptotic expansion of the form (3.19), where the function g_k solves a differential equation of the form $g'_k = \frac{\partial f}{\partial y}(t, y) g_k + d_k$, $g_k(t_0) = 0$, for suitably chosen functions d_k .*

Proof (after [44]). Suppose that the first r terms of the expansion have already be found, so we know functions g_k such that $\hat{g}(t) = \sum_{k=1}^r g_k(t) h^k$ satisfies $G_h(t) = \hat{g}(t) + \mathcal{O}(h^{r+1})$. Note that we can always start with $r = 0$ and $\hat{g}(t) = 0$, because $G_h(t) = \mathcal{O}(h)$ as the method is convergent. We will show how to find the next term in the expansion (3.19), thus proving the theorem by induction.

Recall that $\Psi_h(t, y)$ denotes the numerical method. Now consider another method, defined by

$$\hat{\Psi}_h(t, y) = \Psi_h(t, y + \hat{g}_h(t)) - \hat{g}_h(t + h). \quad (3.20)$$

Application of this method yields the sequence of values $\{\hat{y}_n\}$ with $\hat{y}_0 = y_0$ and $\hat{y}_{n+1} = \hat{\Psi}_h(t_n, \hat{y}_n)$. These values satisfy $\hat{y}_n = y_n - \hat{g}_h(t_n)$, as can easily be shown by induction. Therefore, its global error is

$$\hat{y}_n - y(t_n) = (\hat{y}_n - y_n) + (y_n - y(t_n)) = -\hat{g}_h(t_n) + G_h(t_n) = \mathcal{O}(h^{r+1}). \quad (3.21)$$

We conclude that the numerical method (3.20) has order $r + 1$, and that its leading global error term is $g_{r+1}(t) h^{r+1}$, the term which we are looking for. However, the function g_{r+1} satisfies the differential equation, cf. (3.18),

$$g'_{r+1}(t) = \frac{\partial f}{\partial y}(t, y(t)) g_{r+1}(t) + h^{p+1} \hat{\ell}(t, y(t)), \quad g_{r+1}(t_0) = 0,$$

where $\hat{\ell}(t, y)$ is defined by requiring that the local error of the method (3.20) be $h^{p+2}\hat{\ell}(t, y) + \mathcal{O}(h^{p+3})$. Hence, we can find $g_{r+1}(t)$ by solving this equation. This completes the induction step. \square

Note that the differential equation $g'_k = \frac{\partial f}{\partial y}(t, y) g_k + d_k$, $g_k(t_0) = 0$ can easily be solved. The variation of constants formula gives

$$g_k(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) d_k(t) dt. \quad (3.22)$$

The theorem is still valid if we use a variable step-size method, provided that the step size is chosen according to $h_k = \varepsilon_h h(t_k)$; in other words, it should be independent of the current position y_k (see [44]).

Example 3.9. We return to the equation $y' = y^2$, solved with Runge's second-order method (2.16) with constant step size.

From (3.6) it follows that $\ell(t, y) = -\frac{3}{4}y^4$ and hence the differential equation (3.18) reads

$$g'(t) = 2y(t)g(t) - \frac{3}{4}y(t)^4 = \frac{2}{1-t}g(t) - \frac{3}{4(1-t)^4}, \quad g(0) = 0.$$

The solution of this equation is $g(t) = -\frac{3}{4}t(1-t)^{-3}$, so the global error is $-\frac{3}{4}h^2t(1-t)^{-3} + \mathcal{O}(h^3)$. To find the next term, we construct the method (3.20) with $\hat{g}(t) = -\frac{3}{4}h^2t(1-t)^{-3}$. We find

$$\begin{aligned} \hat{\Psi}_h(t, y) = y - \frac{3t}{4(1-t)^3} + h \left(y - \frac{3t}{4(1-t)^3} + \frac{1}{2}h \left(y - \frac{3t}{4(1-t)^3} \right)^2 \right)^2 \\ + \frac{3(t+h)}{4(1-t-h)^3}. \end{aligned}$$

After some tedious algebra, we find that

$$\begin{aligned} \hat{\Psi}_h(t, y(t)) = \frac{1}{1-t} + \frac{1}{(1-t)^2}h + \frac{1}{(1-t)^3}h^2 \\ + \frac{1}{(1-t)^4}h^3 + \frac{9}{4(1-t)^5}h^4 + \mathcal{O}(h^5). \end{aligned}$$

We now subtract the Taylor series for $y(t+h)$, which reads

$$y(t+h) = \frac{1}{1-t} + \frac{1}{(1-t)^2}h + \frac{1}{(1-t)^3}h^2 + \frac{1}{(1-t)^4}h^3 + \frac{1}{(1-t)^5}h^4 + \mathcal{O}(h^5).$$

This gives the local error for the method $\hat{\Psi}$,

$$\hat{\Psi}_h(t, y(t)) - y(t+h) = \frac{5}{4(1-t)^5}h^4 + \mathcal{O}(h^5).$$

So, the method $\hat{\Psi}$ indeed has order 3. We can again find the leading global error term by solving the equation (3.18). For $\hat{\Psi}$, this equation reads

$$g_3'(t) = \frac{2}{1-t}g_3(t) + \frac{5}{4(1-t)^5}, \quad g_3(0) = 0.$$

The solution is $g_3(t) = \frac{5}{8}t(2-t)(1-t)^{-4}$.

We can proceed in this way to find more and more terms. However, instead of boring the reader with the calculations, we just state the result of the computation.

$$G_h(t) = -\frac{3t}{4(1-t)^3}h^2 + \frac{5t(2-t)}{8(1-t)^4}h^3 - \frac{t(23t^2 - 96t + 42)}{48(1-t)^5}h^4 - \frac{3t(2t^3 - 18t^2 + 27t + 12)}{32(1-t)^6}h^5 + \mathcal{O}(h^6). \quad (3.23)$$

This estimate is illustrated in the left-hand plot of Figure 3.3. The dotted line shows the global error of Runge's method. The topmost solid line shows the first term of the estimate (3.23), the second line shows the sum of the h^2 and h^3 terms, and so on. We can conclude from this picture that (3.23) is correct. \diamond

The Euler–MacLaurin formula (cf. Theorem 3.3) can be proved with a similar computation that considers the trapezoidal rule

$$y_{n+1} = y_n + \frac{1}{2}h(f(t_n, y_n) + f(t_{n+1}, y_{n+1}))$$

applied to the equation $y' = f(t)$.

Comparing Theorem 3.4 with Theorem 3.8 above, we see that the latter theorem has the clear advantage that it gives the complete asymptotic expansion, while Theorem 3.4 does not allow one to go beyond the term of order h^{2p} . On the other hand, we only need to compute one integral when evaluating the estimate in Theorem 3.4, the rest is straightforward but tedious algebra (provided that the exact solution is known). In contradistinction, the method of Theorem 3.8 requires us to solve a differential equation, or, equivalently, the integral (3.22), for every term in the asymptotic expansion.

3.3 Modified equations and the global error

In this section, we describe the third approach to estimating the global error. This approach uses the theory of backward error analysis, which was explained in Section 2.4. A similar strategy is used by Calvo and Hairer [18], and Hairer and Lubich [42].

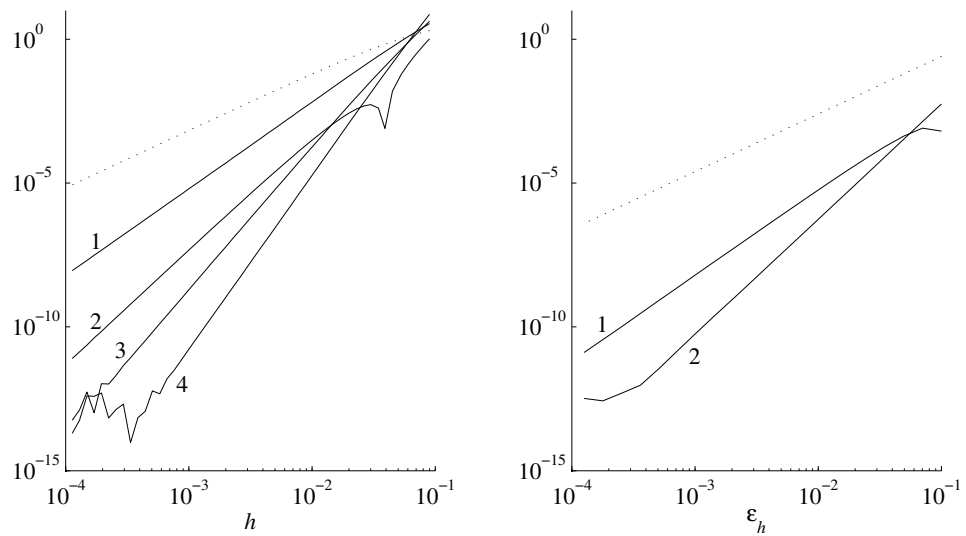


Figure 3.3: On the left, the dotted line shows the global error at $t_f = 0.9$ committed by Runge's method with constant step size when applied to $y' = y^2$. The solid curve marked 1 shows the difference between the global error and the first term of the estimate (3.23); in other words, it shows $G_h(t) + \frac{3}{4}t(1-t)^{-3}h^2$. Similarly, the curves marked 2, 3 and 4 show the difference between the global error and the estimate (3.23) truncated after the second, third, and fourth term, respectively. On the right, a similar picture compares the global error of Runge's method with step size $h_k = 1/y_k^2$ with the estimate (3.25).

The idea is rather simple. The Alekseev–Gröbner lemma describes the effect of perturbing a differential equation. The modified equation allows us to view the numerical method as a perturbation of the original differential equation. This leads to the following theorem. For simplicity, we assume that the differential equation is autonomous.

Theorem 3.10. *Suppose that we solve the differential equation $y' = f(y)$ with a variable step-size method of the form (2.35), which produces the values $\{y_k\}$. If the method has order p , and the solution $\tilde{y}(t)$ of the (truncated) modified equation $\tilde{y}' = \tilde{f}_{\varepsilon_h}(\tilde{y})$ satisfies $\tilde{y}(t_k) = y_k + \mathcal{O}(\varepsilon_h^{2p})$ for all k , then*

$$G_{\varepsilon_h}(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) \delta_{\varepsilon_h}(y(t)) dt + \mathcal{O}(\varepsilon_h^{2p}), \quad (3.24)$$

where $\delta_{\varepsilon_h}(y) = \tilde{f}_{\varepsilon_h}(y) - f(y)$.

Of course, this theorem is also valid for constant step-size methods.

Proof. The Alekseev–Gröbner lemma on page 8 shows that

$$\tilde{y}(t_f) - y(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f}(\tilde{y}(t)) \delta_{\varepsilon_h}(\tilde{y}(t)) dt.$$

The expression on the left-hand side is $G_{\varepsilon_h}(t_f) + \mathcal{O}(\varepsilon_h^{2p})$ because the solution of the modified equation is $\mathcal{O}(\varepsilon_h^{2p})$ -close to the numerical solution. Furthermore, we have $y(t) - \tilde{y}(t) = \mathcal{O}(\varepsilon_h^p)$, since the method is of order p . Finally, it follows from the theory of modified equation in Section 2.4 that $\delta_{\varepsilon_h}(y) = \mathcal{O}(\varepsilon_h^p)$. Together, this proves (3.24). \square

Example 3.11. We again consider Runge's second-order method applied to the differential equation $y' = y^2$. First, we will assume that the step size is held constant. In Example 2.3, we found that the modified equation is given by (2.30). Hence we have

$$\delta_h(y) = -\frac{3}{4}h^2y^4 + \frac{5}{4}h^3y^5 - \frac{7}{8}h^4y^6 + \mathcal{O}(h^5).$$

The integral in (3.24) evaluates to

$$-\frac{3t_f}{4(1-t_f)^3}h^2 + \frac{5t_f(2-t_f)}{8(1-t_f)^4}h^3 - \frac{7t_f(t_f^2-3t_f+3)}{24(1-t_f)^5}h^4 + \mathcal{O}(h^5).$$

If we compare this with the estimate (3.23), which we obtained in the last section, we see that the first two terms are correct, but the third one is not. So (3.24) provides an $\mathcal{O}(h^4)$ -estimate of the global error, just as the theorem states.

Now suppose we choose the step size according to $h_k = 1/y_k^2$ as in Example 2.5. The modified equation is given in (2.36), and a similar computation as in the constant step-size case yields

$$G_{\varepsilon_h}(t_f) = -\frac{t_f(t_f^2 - 3t_f + 3)}{4(1 - t_f)^2} \varepsilon_h^2 + \frac{t_f(t_f^3 - 4t_f^2 + 6t_f - 4)}{16(1 - t_f)^2} \varepsilon_h^3 + \mathcal{O}(\varepsilon_h^4). \quad (3.25)$$

A numerical experiment was performed to check this estimate. The result, reported on the right-hand side of Figure 3.3, indicates that the estimate is indeed correct. \diamond

We know from Section 2.4, that if the numerical method can be expanded in a B-series (as is the case for all Runge–Kutta methods), the modified equation can also be written in terms of a B-series. Combining this with the above theorem yields the following result.

Corollary 3.12. *Suppose that a consistent numerical method with constant step size is used, and that this method can be expanded in a B-series with coefficient function $a : \mathbf{T} \cup \{\emptyset\} \rightarrow \mathbf{R}$. Define the function $b : \mathbf{T} \cup \{\emptyset\} \rightarrow \mathbf{R}$ by $b(\emptyset) = 0$ and*

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

If the numerical method has order p , then the global error satisfies

$$G_h(t_f) = \sum_{k=p}^{2p-1} h^k \sum_{\tau \in \mathbf{T}_{k+1}} \frac{b(\tau)}{\sigma(\tau)} \mathcal{I}(\tau)(t_f) + \mathcal{O}(h^{2p}), \quad (3.26)$$

$$\text{where } \mathcal{I}(\tau)(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) F(\tau)(y(t)) dt.$$

Here, \mathbf{T}_k denotes the set of all trees with order k .

Proof. The modified equation is $\tilde{y}' = \frac{1}{h} B(b, \tilde{y})$ with b as defined in the theorem, as the discussion around (2.31) shows. The first term of the B-series $B(b, \tilde{y})$ vanishes, since $b(\emptyset) = 0$. For the second term, we have $b(\bullet) = a(\bullet) = 1$ (because the method is consistent) and $F(\bullet)(y) = f(y)$. Hence, we have

$$\delta_h(y) = \tilde{f}_h(y) - f(y) = \sum_{k=2}^{\infty} \sum_{\tau \in \mathbf{T}_k} h^{k-1} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(y).$$

However, we know that $b(\tau)$ vanishes if $\rho(\tau) \leq p$, because the first term in the modified equation has order h^p . We now substitute this expression in the estimate in Theorem 3.10, and move the scalar factors out of the integral (remember that the variational flow $D\Phi_t^{t_f}$ is linear). This yields the estimate (3.24). \square

The nice thing about the error estimate (3.24) is that it cleanly separates the numerical method from the particular problem that we want to solve. The method only enters the estimate via the coefficients $b(\tau)$. On the other hand, the value of the integrals $\mathcal{I}(\tau)$ is completely determined by the particular differential equation (i.e., the function f) under consideration. We will call $\mathcal{I}(\tau)$ the *elementary integral* associated with τ , because its role in the global error estimate (3.24) is similar to the role of the elementary differential $F(\tau)(y)$ in the local error.

Example 3.13. We consider again the differential equation $y' = y^2$ with initial condition $y(0) = 1$. An easy calculation shows that the elementary differentials are given by $F(\tau)(y) = C(\tau) y^{\rho(\tau)+1}$, where $C(\tau)$ is a (possibly vanishing) constant which depends on the tree τ . Hence, the elementary integrals are

$$\mathcal{I}(\tau)(t_f) = \frac{C(\tau)}{\rho} \left(\frac{1}{(1-t_f)^{\rho(\tau)+2}} - \frac{1}{(1-t_f)^2} \right).$$

It now follows from Corollary 3.12 that the global error of every constant step-size Runge–Kutta method of order p is

$$G_h(t_f) = \sum_{k=p}^{2p-1} C_k h^k \left(\frac{1}{(1-t_f)^{k+3}} - \frac{1}{(1-t_f)^2} \right) + \mathcal{O}(h^{2p}),$$

where the constants C_k depend on the coefficients of the method. This agrees with the result for Runge’s method with constant step size which we found in Example 3.11. \diamond

The above example shows the merit of Corollary 3.12: it allows us to find the global error of *any* Runge–Kutta method in one calculation. The disadvantage when compared to the method of the previous section, is that the global error estimate is only exact up to a term of order h^{2p} , while the theory of Section 3.2 allows us to find an estimate which approximates the global error with arbitrary order.

Chapter 4

Applications of global error estimates

This chapter contains various applications of the estimates for the global error which were derived in the previous chapter. The first application is of a more theoretical nature: we prove a key lemma which can be used to study the growth of the global error when tracking a periodic orbit. The other two applications pertain to specific classes of equations with highly oscillatory solutions. In Section 4.2, we study the Airy equation $y'' + ty = 0$, and related equations which are amenable to Liouville–Green analysis. In Section 4.3, we look at the nonlinear Emden–Fowler equation $y'' + t^\nu y^n = 0$ in the oscillatory regime. The goal in both cases is to obtain estimates for the global error.

4.1 Error growth in periodic orbits

Cano and Sanz-Serna [20] study the growth of the global error when the numerical integrator is tracking a periodic orbit. Their main technical result is reproduced as Theorem 4.1 below. They use the asymptotic expansion of the global error (cf. Section 3.2) to prove this result. Specifically, they prove that the coefficients $g_k(t)$ in the expansion (3.19) are periodic under the assumptions of the theorem. Here, we will give an alternative proof via modified equations, based on Theorem 3.10.

Theorem 4.1 (Cano and Sanz-Serna). *Suppose we are solving an autonomous differential equation of the form $y' = f(y)$, and that the exact solution is T -periodic, meaning that $y(t+T) = y(t)$ for all t . Assume for simplicity that $t_0 = 0$. If we use a variable step-size method of the form (2.35) that has order p , then the*

global error satisfies

$$G_{\varepsilon_h}(NT) = \left(\sum_{k=0}^{N-1} M^k \right) G_{\varepsilon_h}(T) + \mathcal{O}(\varepsilon_h^{2p}).$$

where M is the monodromy matrix, defined by $M = D\Phi_0^T(y_0)$.

Proof. Let $\tilde{y}' = \tilde{f}_{\varepsilon_h}(\tilde{y})$ be the modified equation, truncated after the term of order ε_h^{2p} . Define $\delta_{\varepsilon_h}(y) = \tilde{f}_{\varepsilon_h}(y) - f(y)$. It follows from Theorem 3.10 that

$$\begin{aligned} G_{\varepsilon_h}(NT) &= \int_0^{NT} D\Phi_t^{NT}(y(t)) \delta_{\varepsilon_h}(y(t)) dt + \mathcal{O}(\varepsilon_h^{2p}) \\ &= D\Phi_{(N-1)T}^{NT} \int_0^{(N-1)T} D\Phi_t^{(N-1)T}(y(t)) \delta_{\varepsilon_h}(y(t)) dt \\ &\quad + \int_{(N-1)T}^{NT} D\Phi_t^{NT}(y(t)) \delta_{\varepsilon_h}(y(t)) dt + \mathcal{O}(\varepsilon_h^{2p}) \\ &= M G_{\varepsilon_h}((N-1)T) + G_{\varepsilon_h}(T) + \mathcal{O}(\varepsilon_h^{2p}). \end{aligned}$$

Here, we used the composition property (2.4) and the T -periodicity of the flow. The theorem now follows by induction on N . \square

Having established Theorem 4.1, Cano and Sanz-Serna [20] apply this formula to Hamiltonian systems. They conclude that for many Hamiltonian systems general integrators possess quadratic error growth, while energy-conserving methods and symplectic methods only lead to linear error growth when following a periodic orbits. A similar result holds for reversible systems.

It should be stressed that techniques from the theory of Dynamical Systems, in particular the KAM-theory, have been successfully applied to explain the behaviour of numerical integrators for Hamiltonian and reversible systems. Two excellent recent references are the works of Hairer, Lubich and Wanner [43], and Moan [67].

4.2 The Airy equation and related oscillators

In this section, we study the global error of Runge–Kutta methods with constant step size when applied to a certain class of equations of the form $y'' + \eta(t)y = 0$, a class which includes the (time-reversed) Airy equation $y'' + ty = 0$. All these equations have oscillatory solutions. However, we do not know this solution; we only have the asymptotic solution as $t \rightarrow \infty$. Nevertheless, we will be able to derive accurate estimates for the global error.

This section builds on the work of Iserles [54], who uses the estimate (3.16) to study the global error of Runge–Kutta, Magnus, and modified Magnus methods.

However, this estimate gives only the leading error term of order h^p . Here, we will apply Corollary 3.12, which enables us to find a more accurate estimate of the global error.

We start by stating the assumptions needed to carry through the computation. Informally, we require $\eta(t)$ to be large for large t , while its derivatives are small. The precise requirements are as follows.

Assumption 4.2. *The function η and its derivatives satisfy the following growth conditions as $t \rightarrow \infty$: $\eta \rightarrow \infty$, $\eta' = o(\eta)$ and $\eta^{(\ell)} = o(\eta')$ for $\ell = 2, 3, 4, \dots$. In addition,*

$$\int_{t_0}^{\infty} \left| \frac{1}{\eta^{1/4}} \frac{d^2}{dt^2} \left(\frac{1}{\eta^{1/4}} \right) \right| dt \text{ converges.} \quad (4.1)$$

Note that this assumption is satisfied by $\eta(t) = t^\alpha$ with $\alpha > 0$ (the choice $\alpha = 1$ retrieves the Airy equation) and $\eta(t) = \log t$.

We now rewrite the second-order equation $y'' + \eta(t)y = 0$ as a system of first-order equations,

$$y' = \begin{bmatrix} 0 & 1 \\ -\eta(t) & 0 \end{bmatrix} y. \quad (4.2)$$

The asymptotic solution of this equation is given by the following result.

Theorem 4.3 (Liouville–Green approximation). *If η satisfies Assumption 4.2, then the asymptotic solution of the equation (4.2) is¹*

$$y(t) \sim \Lambda(t) R(\theta(t)) s_0 \quad \text{as } t \rightarrow \infty, \quad (4.3)$$

where $s_0 \in \mathbf{R}^2$ is a vector whose value depends on the initial conditions of (4.2), $\theta(t) = \int_{t_0}^t \sqrt{\eta(s)} ds$, and

$$\Lambda(t) = \begin{bmatrix} (\eta(t))^{-1/4} & 0 \\ 0 & (\eta(t))^{1/4} \end{bmatrix} \quad \text{and} \quad R(\theta) = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}.$$

The Liouville–Green approximation is also known as the WKB– or WKBJ–approximation (the letters stand for Wentzel, Kramers, Brillouin, and Jeffrey), especially among theoretical physicists. The central idea in deriving this estimate is that, if we set $v(t) = \Lambda(t) y(t)$, the differential equation (4.2) transforms to

$$\frac{dv}{d\theta} = \begin{bmatrix} \frac{1}{4}\eta^{-3/2} & 1 \\ -1 & -\frac{1}{4}\eta^{-3/2} \end{bmatrix} v.$$

¹The notation $f(t) \sim g(t)$ means that $\lim_{t \rightarrow \infty} \frac{f(t)}{g(t)} = 1$. Do not confuse this with the use of the \sim symbol to indicate an asymptotic expansion in Section 3.2.

If we neglect the $\pm \frac{1}{4}\eta^{-3/2}$ entries, this is the harmonic oscillator with solution $v(\theta) = R(\theta)s_0$, which corresponds to (4.3). A rigorous proof of Theorem 4.3 can be found in Olver [76, §6.3]. This proof makes clear why we need to impose the condition (4.1); the other growth conditions in Assumption 4.2 are in fact not necessary for Theorem 4.3 but are used to prove Theorem 4.6 later. The book by Hinch [51], which may be more accessible, takes another approach to derive the Liouville–Green approximation.

Example 4.4. If we take $\eta(t) = t$ in (4.2), we get the Airy equation. We plot the solution of (4.2) with initial condition $y(0) = [1 \ 0]^\top$ in Figure 4.1. Note that the amplitude of the oscillations decreases, while their frequency increases. The solution can be expressed in terms of the standard Airy functions (see e.g. [1, §10.4]),

$$y_1(t) = \frac{1}{2}3^{1/6}\Gamma\left(\frac{2}{3}\right) (\sqrt{3} \operatorname{Ai}(-t) + \operatorname{Bi}(-t)). \quad (4.4)$$

For the Airy equation, the Liouville–Green approximation (4.3) reads

$$y_1(t) \sim t^{-1/4} (s_{0,1} \cos(\frac{2}{3}t^{3/2}) + s_{0,2} \sin(\frac{2}{3}t^{3/2})). \quad (4.5)$$

This estimate is also shown in Figure 4.1 with

$$s_0 = \frac{1}{4} \sqrt{\frac{2}{\pi}} 3^{1/6} \Gamma\left(\frac{2}{3}\right) \begin{bmatrix} \sqrt{3} + 1 \\ \sqrt{3} - 1 \end{bmatrix}. \quad (4.6)$$

Note that the Liouville–Green approximation, being an asymptotic approximation, gives no information about which s_0 corresponds to a particular initial condition; another method is required to find the correct s_0 (in this case, the initial values and asymptotic expansions of the Airy functions, listed in [1, §10.4], are used).

We see that (4.5) approximates the real solution very well for $t \gtrsim 5$. \diamond

We want to use the Liouville–Green approximation to evaluate the global error estimate in Corollary 3.12. This estimate is reproduced below for the reader’s convenience.

$$G_h(t_f) = \sum_{k=p}^{2p-1} h^k \sum_{\tau \in \mathbf{T}_{k+1}} \frac{b(\tau)}{\sigma(\tau)} \mathcal{I}(\tau)(t_f) + \mathcal{O}(h^{2p}), \quad (3.26)$$

$$\text{where } \mathcal{I}(\tau)(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) F(\tau)(y(t)) dt.$$

We first need to transform the equation (4.2) in an autonomous system by adding a dummy variable (cf. the first paragraph of Section 2.3),

$$\begin{cases} y'_1 = y_2, \\ y'_2 = \eta(y_3) y_1, \\ y'_3 = 1. \end{cases} \quad (4.7)$$

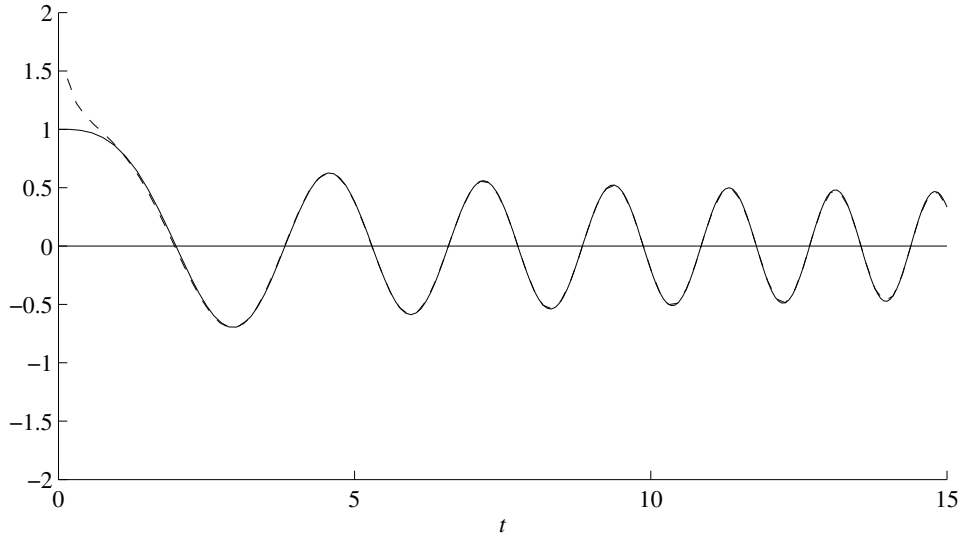


Figure 4.1: The solid line shows the first component of the solution of (4.2) with $\eta(t) = t$, while the dashed line displays the corresponding Liouville–Green approximation (4.5).

The next step is to evaluate the elementary integrals $\mathcal{I}(\tau)$. First, we want to find out which tree from \mathbf{T}_k gives the dominating contribution to the error estimate in order to save us some work.

We define the *height* of a vertex to be the distance to the root, and the height of a tree to be the maximum of the heights of its vertices. The lemma below shows that we can bound the growth of the elementary differentials in terms of the height of the corresponding tree.

Lemma 4.5. *Suppose that Assumption 4.2 is satisfied. If the tree τ has height m , with $m \geq 2$, then*

$$F(\tau)(y(t)) = \begin{bmatrix} \mathcal{O}(\eta(t)^{m/2-1/4}) \\ \mathcal{O}(\eta(t)^{m/2+1/4}) \\ 0 \end{bmatrix}. \tag{4.8}$$

For $m = 1$, the same result holds except that $F_3(\bullet)(y(t)) = 1$.

Proof. We start by evaluating the simplest elementary differential,

$$F(\bullet)(y) = f(y) = \begin{bmatrix} y_2 \\ -\eta y_1 \\ 1 \end{bmatrix} \tag{4.9}$$

where f denotes the right-hand side of (4.7) and the argument t is left out. Note that from the Liouville–Green formula (4.3) we have $y_1 = \mathcal{O}(\eta^{-1/4})$ and $y_2 = \mathcal{O}(\eta^{1/4})$, so we have already proved the statement of the lemma for $m = 1$.

To tackle the case $m \geq 2$, we use induction. Assume that the lemma has been proved for all trees of height up to $m - 1$. If the root of τ has degree 1, then τ is of the form $[\tau_1]$, and the associated elementary differential is

$$F\left(\begin{array}{c} \tau_1 \\ \bullet \end{array}\right)(y) = \left[\sum_j \frac{\partial f_j}{\partial y_j} F_j(\tau_1)(y) \right]_i = \begin{bmatrix} F_2(\tau_1)(y) \\ -\eta F_1(\tau_1)(y) - \eta' y_1 F_3(\tau_1)(y) \\ 0 \end{bmatrix}. \quad (4.10)$$

But the height of τ_1 is one less than the height of $\tau = [\tau_1]$. So we can apply the induction hypothesis, and establish (4.8).

Now suppose that the root has degree k , with $k \geq 2$. In this case,

$$\left[F\left(\begin{array}{c} \tau_1 \cdots \tau_k \\ \bullet \end{array}\right)(y) \right]_i = \sum_{j_1=1}^3 \cdots \sum_{j_k=1}^3 \frac{\partial^k f_i}{\partial y_{j_1} \cdots \partial y_{j_k}} F_{j_1}(\tau_1)(y) \cdots F_{j_k}(\tau_k)(y).$$

Note that the only nonvanishing partial derivatives of f of order k are

$$\frac{\partial^k f_2}{\partial y_3^k} = -\eta^{(k)} y_1 \quad \text{and} \quad \frac{\partial^k f_2}{\partial y_1 \partial y_3^{k-1}} = -\eta^{(k-1)}.$$

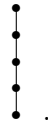
However, $F_3(\tilde{\tau})(y) = 0$ unless $\tilde{\tau}$ is the unit tree (the unique tree of order 1). So, the elementary differential $F(\tau)(y)$ vanishes except if τ is of the form $[\tau_1, \bullet, \dots, \bullet]$, or $\tau = [\bullet, \bullet, \dots, \bullet]$. The associated elementary differentials are

$$F\left(\begin{array}{c} \bullet \cdots \bullet \tau_k \\ \bullet \end{array}\right)(y) = \begin{bmatrix} 0 \\ -\eta^{(k-1)} F_1(\tau_k)(y) \\ 0 \end{bmatrix}$$

$$F\left(\begin{array}{c} \bullet \cdots \bullet \bullet \\ \bullet \end{array}\right)(y) = \begin{bmatrix} 0 \\ -k\eta^{(k-1)} y_2 - \eta^{(k)} y_1 \\ 0 \end{bmatrix}.$$

In both cases, the order estimate (4.8) holds. □

Because of the above lemma, we can expect that among the trees of order k , the highest tree dominates the error estimate. Obviously, the height of a tree of order k is bounded above by $k - 1$. There is only one such tree; it consists of only a trunk without any branches. We denote this tree by τ_k^\perp . For example, τ_5^\perp is the tree



The next step is to calculate the elementary differentials associated to τ_k^\perp . For $k = 1$, this is done in (4.9). For $k > 1$, we can use the recurrence relation (4.10) to find

$$F(\tau_{2m+1}^\perp)(y) = \begin{bmatrix} (-\eta)^m y_2 - \eta' (-\eta)^{m-1} y_1 \\ (-\eta)^{m+1} y_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(\eta^{m+1/4}) \\ \mathcal{O}(\eta^{m+3/4}) \\ 0 \end{bmatrix} \quad (4.11)$$

$$F(\tau_{2m+2}^\perp)(y) = \begin{bmatrix} (-\eta)^{m+1} y_1 \\ (-\eta)^{m+1} y_2 - \eta' (-\eta)^m y_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(\eta^{m+3/4}) \\ \mathcal{O}(\eta^{m+5/4}) \\ 0 \end{bmatrix}. \quad (4.12)$$

Comparing with Lemma 4.5, we conclude that, among the elementary differential associated with trees of a given order, the one associated to the branchless tree dominates.

As noted by Orel [77], this knowledge can be used to design better methods. In particular, if a method of order p has $a(\tau_{p+1}^\perp) = 0$, then the leading error term is knocked out. Therefore, methods with this property will perform better on the differential equation (4.2) than general methods.

Note that the third component of $F(\tau)(y)$ is always zero (except if τ is the unit tree). We shall henceforth drop this component; in other words, we return to the nonautonomous formulation (4.2).

The next step is to calculate the elementary integrals $\mathcal{I}(\tau_k^\perp)(y)$. First, assume that k is even. From (4.12), we deduce that $F(\tau_{2m}^\perp) \sim (-\eta)^m y$, where we dropped the term with η' which is $o(\eta^{m-1/4})$. Furthermore, from (4.3) we can calculate the variational flow

$$D\Phi_s^t \sim \Lambda(t) R(\theta(t) - \theta(s)) \Lambda^{-1}(s). \quad (4.13)$$

Putting both facts together, we can evaluate the elementary integral,

$$\begin{aligned} \mathcal{I}(\tau_{2m}^\perp)(t_f) &= \int_{t_0}^{t_f} D\Phi_t^{t_f}(y(t)) F(\tau_{2m}^\perp)(y(t)) dt. \\ &\approx \int_{t_0}^{t_f} \Lambda(t_f) R(\theta(t_f) - \theta(t)) \Lambda^{-1}(t) \cdot (-\eta(t))^m \Lambda(t) R(\theta(t)) s_0 dt. \\ &= (-1)^m \int_{t_0}^{t_f} (\eta(t))^m dt y(t_f). \end{aligned} \quad (4.14)$$

Now, assume that k is odd. From (4.11), we deduce that

$$F(\tau_{2m+1}^\perp)(y(t)) = \begin{bmatrix} 0 & (-\eta(t))^m \\ (-\eta(t))^{m+1} & 0 \end{bmatrix} y(t).$$

A similar computation as above yields

$$\mathcal{I}(\tau_{2m+1}^\perp)(t_f) \approx (-1)^m \int_{t_0}^{t_f} (\eta(t))^{m+1/2} dt y_R(t_f), \quad (4.15)$$

where $y_R(t)$ denotes the solution with opposite phase, that is, $y_R(t)$ is the solution of (4.2) which satisfies

$$y_R(t) \sim \Lambda(t) R(\theta(t) + \frac{1}{2}\pi) s_0. \quad (4.16)$$

Combining (4.14) and (4.15) into the error estimate in Corollary 3.12, we get the following estimate for the global error.

Theorem 4.6. *Suppose we are solving the differential equation (4.2) where η satisfies Assumption 4.2. If a constant step-size Runge–Kutta method of order p is employed, then the global error is*

$$\begin{aligned} G_h(t_f) &\approx \sum_{m=\lceil p/2 \rceil}^{p-1} (-1)^m b(\tau_{2m+1}^\dagger) h^{2m} \int_{t_0}^{t_f} (\eta(t))^{m+1/2} dt y_R(t_f) \\ &+ \sum_{m=\lceil p/2 \rceil}^{p-1} (-1)^{m+1} b(\tau_{2m+2}^\dagger) h^{2m+1} \int_{t_0}^{t_f} (\eta(t))^{m+1} dt y(t_f) + \mathcal{O}(h^{2p}), \end{aligned} \quad (4.17)$$

where the B-series coefficients $b(\tau)$ are defined in Corollary 3.12, and $y_R(t_f)$ is as defined in (4.16).

It should be noted that the remainder term $\mathcal{O}(h^{2p})$ in (4.17) is not uniform in t_f . This is illustrated in the following example.

Example 4.7. As in Example 4.4, we consider the Airy equation, that is, we set $\eta(t) = t$. The error estimate (4.17) becomes (assuming that t_0 is small)

$$\begin{aligned} G_h(t_f) &\approx \sum_{m=\lceil p/2 \rceil}^{p-1} \frac{(-1)^m b(\tau_{2m+1}^\dagger)}{m + \frac{3}{2}} h^{2m} t_f^{m+3/2} y_R(t_f) \\ &+ \sum_{m=\lceil p/2 \rceil}^{p-1} \frac{(-1)^{m+1} b(\tau_{2m+2}^\dagger)}{m + 2} h^{2m+1} t_f^{m+2} y(t_f) + \mathcal{O}(h^{2p}). \end{aligned}$$

For example, for Runge's second order method, we have $p = 2$, $b(\tau_3^\dagger) = -\frac{1}{6}$ and $b(\tau_4^\dagger) = \frac{1}{8}$ (cf. Example 2.4), so

$$G_h(t_f) \approx \frac{1}{15} h^2 t_f^{5/2} y_R(t_f) + \frac{1}{24} h^3 t_f^3 y(t_f) + \mathcal{O}(h^4). \quad (4.18)$$

With enough perseverance, the same result can also be derived using the method of Section 3.2. However, the calculation is more complicated than the method described above, and it seems impossible to derive the general estimate (4.17) with the method of Section 3.2.

We see from (4.18) that the global error oscillates with ever-increasing frequency, like the true solution. The amplitude of the oscillations of the leading term grows like $t^{9/4}$ (remember that the amplitude of the true solution decays like $t^{-1/4}$). The h^3 term is negligible for small t_f , but it grows faster than the h^2 term. Hence, the h^3 term will overtake the h^2 term at some point, namely at $t_f \approx \frac{64}{25}h^{-2}$ (but we will see later that this is irrelevant).

To check this estimate, we solve the Airy equation with the initial condition $y(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ over the time interval $[0, 2000]$ with various step sizes. The numerical solution is compared with the exact solution (4.4) to compute the global error. The first component of the error is shown in Figure 4.2. The left column shows the time interval $[0, 50]$, and on the right the larger interval $[0, 2000]$ is displayed.

We can see in the left-hand column that the global error oscillates, just as discussed under (4.18). The envelope of these oscillations, as predicted by the estimate (4.18), is shown by the thick curve in the left-hand column in Figure 4.2. We conclude that the estimate (4.18) describes the actual error accurately.

The right-hand column of Figure 4.2 shows a much larger time interval. Here the oscillations of the error are compressed so heavily that the error appears as a grey blob. The black line shows the estimate (4.18). We see that this estimate breaks down around $t_f = 800$ for $h = 1/1000$ and $t_f = 1400$ for $h = 1/2000$.

If we look again at (4.18), we see that the amplitude of the leading error term reaches the amplitude of the solution when $t_f \approx 15^{2/5}h^{-4/5}$. For $h = 1/1000$ and $h = 1/2000$, this evaluates to $t_f \approx 742$ and $t_f \approx 1292$, respectively. We conclude that the estimate (4.18) ceases to be valid around the time at which the numerical solution has become meaningless because the error is as big as the solution itself. The reason for this break-down of the error estimate lies in the $\mathcal{O}(h^4)$ remainder term in (4.18), which grows faster than the h^2 and h^3 terms.

Therefore, the fact that the h^3 term overtakes the h^2 term at $t_f \approx \frac{64}{25}h^{-2}$ is irrelevant, as the estimate (4.18) has already become meaningless by that time. So, it turns out that it was not necessary to compute the h^3 term, and that we could have restricted ourselves to the leading error term. Of course, we did not know this in advance. In the next section, we show an example where the leading term does not dominate, and other terms in the expansion have to be calculated in order to get an accurate approximation of the global error. \diamond

Iserles [54] gives error estimates and numerical results for the standard fourth-order method (2.18), the fourth-order Gauss–Legendre method, the Magnus method, and the modified Magnus method, when applied to either the Airy equation or the equation (4.2) with $\eta(t) = \log(t + 1)$.

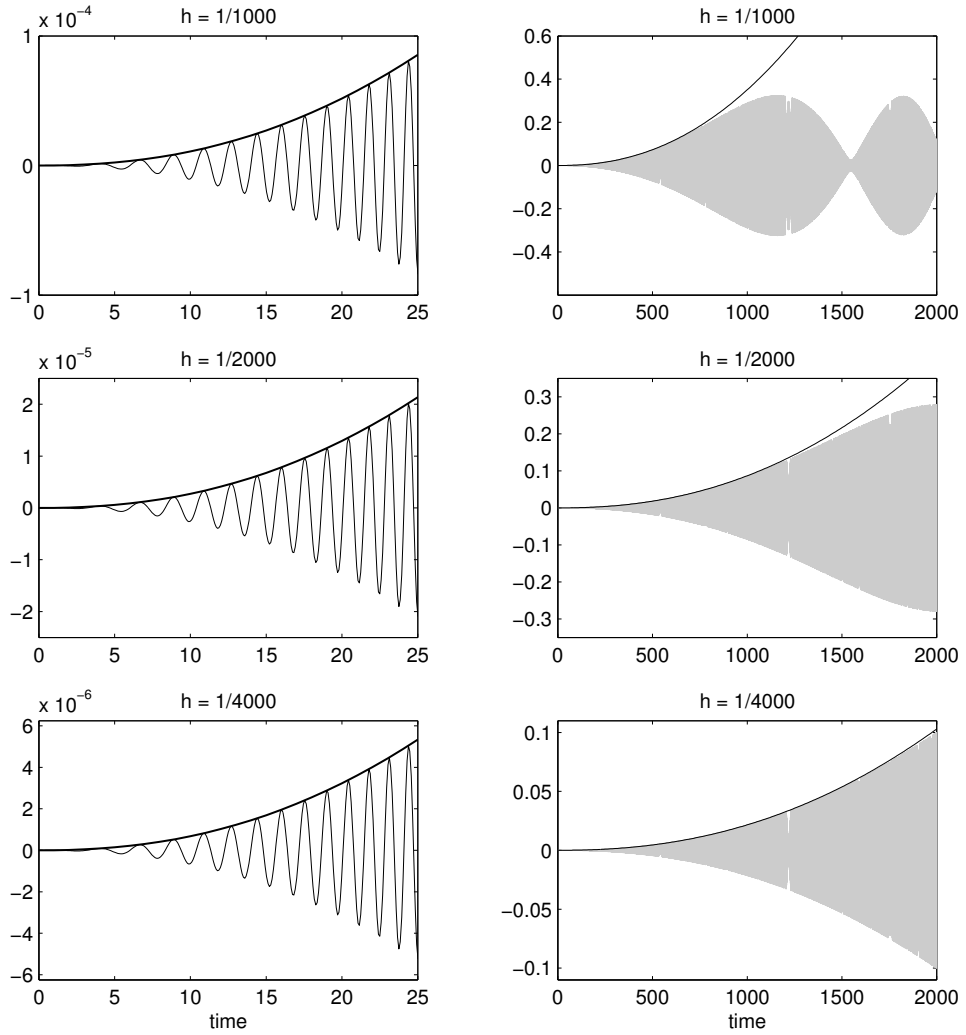


Figure 4.2: On the left, the oscillating curve shows the first component of the global error committed by Runge’s second-order method for various step sizes, when applied to the Airy equation. The thick curve shows the envelope of the oscillations as predicted by (4.18). On the right, a different time scale is used. The true error is shown in grey and the solid curve shows the error estimate (4.18).

4.3 The Emden–Fowler equation

In this section, we study the Emden–Fowler equation. This equation can be viewed as a nonlinear, and hence more challenging, variant of the Airy-like equations studied in the previous section. Again, the goal is to find an accurate estimate for the global error when solving this equation. In contrast to the Airy equation, we find that for the Emden–Fowler equation, the leading error term is sometimes dominated by the second term. This shows that we cannot always be satisfied with computing only the leading error term.

The history of the Emden–Fowler equation starts with a model of the Sun derived by Lane [59]. Suppose that the Sun is a radially symmetric body of gas with radius R . The pressure P at a point at distance r_0 of the centre is given by the weight of the column of gas above it, so $P = \int_{r_0}^R g \rho \, dr$ where g is the gravitational acceleration and ρ is the density (both g and ρ depend on r). The acceleration is given by $g = -\frac{d\varphi}{dr}$, where φ is the gravitational potential, and hence $\frac{dP}{d\varphi} = \rho$. We now assume that the gas satisfies a polytropic equation of state, meaning that its pressure and density are related by $P = K\rho^\gamma$, where K and γ are empirical constants. Substituting this in $\frac{dP}{d\varphi} = \rho$ and solving the resulting equation using the fact that $P = \rho = 0$ at the surface of the Sun, yields

$$\rho = \left(\frac{\gamma - 1}{\gamma K} \varphi \right)^n, \quad (4.19)$$

where $n = 1/(\gamma - 1)$. Finally, we use that the gravitational potential φ is given by $\nabla^2 \varphi = -4\pi G\rho$, where G is the gravitational constant. This reduces by spherical symmetry to $\frac{d^2\varphi}{dr^2} + 2r^{-1}\frac{d\varphi}{dr} = -4\pi G\rho$. Substituting (4.19) in this equation yields

$$\frac{d^2\varphi}{dr^2} + \frac{2}{r} \frac{d\varphi}{dr} = -4\pi G \left(\frac{\gamma - 1}{\gamma K} \varphi \right)^n.$$

Setting $r = (4\pi G)^{-\frac{1}{2}} \left(\frac{\gamma - 1}{\gamma K} \right)^{-\frac{n}{2}} t$ simplifies the equation to $\varphi'' + 2t^{-1}\varphi' + \varphi^n = 0$, where the primes denote derivatives with respect to t . This equation is commonly called the Lane–Emden equation. The substitution $\varphi = t^{-1}y$ reduces the Lane–Emden equation to $y'' + t^{1-n}y^n = 0$. An obvious generalization of this equation is $y'' + t^\nu y^n = 0$, or, in first-order form,

$$y_1' = y_2 \quad \text{and} \quad y_2' = -t^\nu y_1^n. \quad (4.20)$$

This is the Emden–Fowler equation, named after Robert Emden and Ralph Howard Fowler, who contributed significantly to its analysis in the beginning of the twentieth century.

The Emden–Fowler equation has recently appeared in the study of spherical gas clouds that cool slowly by radiation (see Meerson, Megged and Tajima [66]), phase transition in critical adsorption in the mean-field framework (see Gnutzmann and Ritschel [32]), and spherically symmetric space-time manifolds with constant scalar curvature (see Goenner and Havas [33]). More applications of the Emden–Fowler equation (including nuclear physics and the study of chemically reacting systems) can be found in the review by Wong [86] and references therein. Note that the choice $\nu = n = 1$ reduces the Emden–Fowler equation to the Airy equation, which was studied in the previous section.

From now on, we will assume that n is an odd integer, that $n \geq 3$, and that $\nu > -\frac{1}{2}(n+3)$. These conditions assure that oscillatory solutions exist (see Wong [86] for details). We remark incidentally that this remains true when the requirement that n be an integer is dropped, provided we replace y_1^n by $|y_1|^n \operatorname{sgn}(y_1)$ in (4.20), where $\operatorname{sgn}(y_1)$ denotes the sign of y_1 . However, this destroys the analyticity of the equation at $y_1 = 0$.

Inspired by the Liouville–Green approximation discussed on page 40, we seek a transformation which allows us to find the asymptotic solution of the Emden–Fowler equation (see also the work of Aripov and Èshmatov [3], who do a similar analysis in the nonoscillatory regime). From now on, we set

$$\beta = \frac{\nu}{n+3} \quad \text{and} \quad \lambda = 1 + 2\beta.$$

The conditions on n and ν imply that $\beta > -\frac{1}{2}$ and $\lambda > 0$. Now consider the transformation given by

$$\begin{aligned} y_1(t) &= \lambda^{2/(n-1)} t^{-\beta} v_1(t^\lambda), \\ y_2(t) &= \lambda^{(n+1)/(n-1)} t^\beta v_2(t^\lambda). \end{aligned}$$

This transforms the differential equation (4.20) into

$$\begin{aligned} v_1' &= v_2 + \beta \lambda^{-1} t^{-\lambda} v_1, \\ v_2' &= -v_1^n - \beta \lambda^{-1} t^{-\lambda} v_2. \end{aligned}$$

If we neglect the last term in both equations (remember that $\lambda > 0$), we are left with the equations $v_1' = v_2$ and $v_2' = -v_1^n$. The expression $I = 2v_1^{n+1} + (n+1)v_2^2$ is an invariant of this equation, so the solutions trace the level curves of I . However, I is only approximately constant on solutions of the original equation.

Note that the system $v_1' = v_2$, $v_2' = -v_1^n$ can be written as the single second-order equation $v'' + v^n = 0$. We will denote the solution of this equation that satisfies the initial conditions $v(0) = 0$ and $v'(0) = 1$ by $w_n(t)$, and note for

further reference that it is an odd, periodic function. The general solution of the system $v'_1 = v_2$, $v'_2 = -v_1^n$ is then given by

$$\begin{aligned} v_1(t) &= c_1^{2/(n-1)} w_n(c_1 t + c_2), \\ v_2(t) &= c_1^{(n+1)/(n-1)} w'_n(c_1 t + c_2). \end{aligned}$$

Note that c_1 determines the amplitude of the oscillations, while c_2 determines the phase. In other words, (c_1, c_2) are the action-angle coordinates of the Hamiltonian system $v'_1 = v_2$, $v'_2 = -v_1^n$.

It follows that the solution of the Emden–Fowler equation (4.20) is asymptotically (as $t \rightarrow \infty$) given by

$$\begin{aligned} y_1(t) &\approx (\lambda c_1)^{2/(n-1)} t^{-\beta} w_n(c_1 t^\lambda + c_2), \\ y_2(t) &\approx (\lambda c_1)^{(n+1)/(n-1)} t^\beta w'_n(c_1 t^\lambda + c_2). \end{aligned} \tag{4.21}$$

In the remainder of this section, we assume that the above asymptotic solution is in fact exact. The numerical experiments described at the end of this section, will show that this is a valid approximation.

The next steps are to calculate the elementary differentials and integrals. We can then apply Corollary 3.12 to find an estimate for the global error. Unfortunately, the whole computation is rather tedious. While studying the details, the reader may want to refer to Table 4.3 on page 54, where some elementary differentials and integrals for the specific case $\nu = 1$ and $n = 3$ are listed.

To compute the elementary differentials, we need to convert (4.20) to a system of autonomous equations by introducing a third variable representing time,

$$\begin{cases} y'_1 = y_2, \\ y'_2 = y_3^\nu y_1^n, \\ y'_3 = 1. \end{cases} \tag{4.22}$$

The first components of the elementary differentials satisfy the following recurrence relations (where the argument y is deleted)

$$F_1(\bullet) = y_2, \quad F_1(\overset{\tau}{\downarrow}) = F_2(\tau), \quad F_1(\overset{\tau_1 \dots \tau_k}{\downarrow}) = 0 \quad (\text{for } k \geq 2). \tag{4.23}$$

For the third component, the situation is even simpler, as we have

$$F_3(\bullet) = 1 \quad \text{and} \quad F_3(\tau) = 0 \quad \text{for all } \tau \text{ with } \rho(\tau) \geq 2. \tag{4.24}$$

Finally, for the second component, we have $F_2(\bullet) = -y_3^\nu y_1^n$ and

$$\begin{aligned} F_2(\overset{\tau_1 \dots \tau_k}{\downarrow}) &= - \sum_{S \subset \{1, \dots, k\}} \left((\nu - |S| + 1)_{|S|} (n - |S^c| + 1)_{|S^c|} \right. \\ &\quad \left. \cdot y_3^{\nu - |S|} y_1^{n - |S^c|} \prod_{i \in S} F_3(\tau_i) \prod_{i \in S^c} F_1(\tau_i) \right), \end{aligned} \tag{4.25}$$

where $|\cdot|$ denotes the cardinality, $S^c = \{1, \dots, k\} \setminus S$ is the complement of S , and $(x)_n$ denotes the Pochhammer symbol

$$(x)_n = x(x+1)(x+2) \dots (x+n-2)(x+n-1).$$

We now compare the various terms in the summation in (4.25). Let S be a nonempty subset of $\{1, \dots, k\}$ and pick an arbitrary $i \in S$. If τ_i is not the unit tree, then $F_3(\tau_i)$ vanishes, so the term in the sum corresponding to S is zero. On the other hand, if τ_i is the unit tree, then shifting i from S to S^c results in replacing the factor $y_3^{-1}F_3(\bullet) = t^{-1}$ by $y_1^{-1}F_1(\bullet) = \mathcal{O}(t^{2\beta})$, disregarding the constant. In both cases, we find that the term corresponding to S is dominated by the term corresponding to $S \setminus \{i\}$. Hence, the sum in (4.25) is dominated by the term corresponding to $S = \emptyset$, and we have

$$F_2\left(\begin{array}{c} \tau_1 \dots \tau_k \\ \bullet \end{array}\right) = -(n-k+1)_k y_3^\nu y_1^{n-k} \prod_{i=1}^k F_1(\tau_i) \cdot (1 + \mathcal{O}(t^{-1-2\beta})). \quad (4.26)$$

After solving the recurrence relations (4.23), (4.24), and (4.26), and substituting the approximate solution (4.21), we find that the elementary differentials are given by

$$F(\tau)(y(t)) = \begin{bmatrix} C_{1,\tau} t^{\beta(2\rho-1)} w_n^{(n+1)d-\rho+1}(\tilde{t}) w_n^{\prime \rho-2d}(\tilde{t}) + \mathcal{O}(t^{\beta(2\rho-3)-1}) \\ C_{2,\tau} t^{\beta(2\rho+1)} w_n^{n\rho-(n+1)d}(\tilde{t}) w_n^{\prime 2d-\rho+1}(\tilde{t}) + \mathcal{O}(t^{\beta(2\rho-1)-1}) \end{bmatrix}. \quad (4.27)$$

Here ρ denotes the order of the tree τ , and d is the number of vertices with odd height. Furthermore, $\tilde{t} = c_1 t^\lambda + c_2$. We dropped the third component, because it does not contribute to the global error. It should be noted that the constants $C_{1,\tau}$ and $C_{2,\tau}$ may vanish; in fact, the only trees for which both constants are nonzero, are the branchless trees τ_ρ^\perp .

The growth rate of the elementary differential (4.27) is determined by the exponent of t . Note that the variable d does not enter in this exponent. The surprising conclusion is that all trees of the same order contribute a term with the same growth rate, independent of their shape. This is in stark contrast to the linear case treated in the previous section, where the differential corresponding to the branchless tree τ_ρ^\perp dominates, as discussed after Lemma 4.5.

The next step is to calculate the elementary integral $\mathcal{I}(\tau)$. For this, we need to multiply the above differential with the variational flow matrix and integrate the resulting expression, cf. (3.26). To compute the variational flow, we introduce the map $X_t : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ defined by

$$X_t(c_1, c_2) = \begin{bmatrix} (\lambda c_1)^{2/(n-1)} t^{-\beta} w_n(c_1 t^\lambda + c_2) \\ (\lambda c_1)^{(n+1)/(n-1)} t^\beta w_n'(c_1 t^\lambda + c_2) \end{bmatrix}. \quad (4.28)$$

So X_t maps the parameter space to the solution space at time t , cf. (4.21). It follows that the flow map satisfies $\Phi_s^t = X_t \circ X_s^{-1}$. Hence we can write the elementary integral as

$$\mathcal{I}(\tau)(t_f) = DX_{t_f}(y(t_f)) \int_{t_0}^{t_f} DX_t^{-1}(y(t)) F(\tau)(y(t)) dt. \quad (4.29)$$

To find the integrand in the above expression, we multiply the inverse of the Jacobian matrix of (4.28) with the elementary differential (4.27). The result is

$$DX_t^{-1} F(\tau)(y) = \left[\begin{array}{l} t^{2\beta\rho} \left(C_{3,\tau} w_n^{(n+1)(d+1)-\rho} (w'_n)^{\rho-2d} + \mathcal{O}(t^{-\lambda}) \right) \\ t^{2\beta\rho+\lambda} \left(C_{4,\tau} w_n^{(n+1)(d+1)-\rho} (w'_n)^{\rho-2d} + \mathcal{O}(t^{-\lambda}) \right) \end{array} \right] \quad (4.30)$$

where the functions w_n and w'_n are evaluated at $\tilde{t} = c_1 t^\lambda + c_2$. In the calculation, we used $w_n'' = -w_n^n$ and the fact that $2w_n^{n+1} + (n+1)(w'_n)^2$ is a first integral.

The next step is to integrate (4.30). But consider the exponents of w_n and w'_n . If ρ is even, then these exponents are also even and hence the integrand is nonnegative. However, if ρ is odd, then the exponents of w_n and w'_n are also odd. Since w_n is odd and periodic, this implies that the integrand oscillates around zero. Thus we can expect cancellations if ρ is odd, but not if ρ is even. We stress that this phenomenon does not occur in the linear case, analysed in the previous section.

More precisely, we have

$$\int_0^{t_f} w_n^\ell(t) w_n^m(t) dt = \begin{cases} \tilde{C}_{\ell mn}(t_f), & \text{if either } \ell \text{ or } m \text{ is odd,} \\ C_{\ell mn} t_f + \tilde{C}_{\ell mn}(t_f), & \text{if both } \ell \text{ and } m \text{ are even,} \end{cases}$$

where $\tilde{C}_{\ell mn}(t)$ denotes an oscillatory function with the same period as $w_n(t)$, and $C_{\ell mn}$ is a constant. After the substitution $\tilde{t} = c_1 t^\lambda + c_2$ and integration by parts, we find that

$$\int_0^{t_f} t^k w_n^\ell(\tilde{t}) w_n^m(\tilde{t}) dt = \begin{cases} \tilde{C}_{k\ell mn}(\tilde{t}_f) t_f^{k-2\beta} + \mathcal{O}(t_f^{k-2\beta-\lambda}), & \text{if } \ell \text{ or } m \text{ odd,} \\ C_{k\ell mn} t_f^{k+1} + \mathcal{O}(t_f^{k+1-\lambda}), & \text{if } \ell \text{ and } m \text{ even,} \end{cases}$$

where again $\tilde{C}_{k\ell mn}$ and $C_{k\ell mn}$ denote a periodic function and a constant, respectively, and $\tilde{t}_f = c_1 t_f^\lambda + c_2$.

We can use this result to integrate (4.30), which yields (under the assumption that $t_0 \ll t_f$)

$$\int_{t_0}^{t_f} DX_t^{-1} F(\tau)(y) dt = \begin{cases} \left[\begin{array}{l} \tilde{C}_{5,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1-\lambda} + \mathcal{O}(t_f^{2\beta\rho+1-2\lambda}) \\ \tilde{C}_{6,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1} + \mathcal{O}(t_f^{2\beta\rho+1-\lambda}) \end{array} \right], & \text{if } \rho \text{ odd,} \\ \left[\begin{array}{l} C_{5,\tau} t_f^{2\beta\rho+1} + \mathcal{O}(t_f^{2\beta\rho+1-\lambda}) \\ C_{6,\tau} t_f^{2\beta\rho+1+\lambda} + \mathcal{O}(t_f^{2\beta\rho+1}) \end{array} \right], & \text{if } \rho \text{ even.} \end{cases} \quad (4.31)$$

To compute the elementary integral $\mathcal{I}(\tau)$, we need to premultiply the above integral with DX_τ , cf. (4.29). But the expression (4.31) has an interpretation by itself. Recall that X_t maps the parameter space to the solution space. So the integral (4.31) represents the error in parameter space. As the first parameter represents the amplitude, or energy, we conclude that the energy error associated with the tree τ grows as $t^{2\beta\rho+1}$ if ρ is even, and as $t^{2\beta\rho+1-\lambda}$ if ρ is odd. The second component of (4.31) gives the phase error.

Multiplying the Jacobian matrix of the map X_t with the integral (4.31) gives us the elementary integrals,

$$\mathcal{I}(\tau)(t_f) = \begin{cases} \begin{bmatrix} \tilde{C}_{7,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1-\beta} + \mathcal{O}(t_f^{2\beta\rho+1-\beta-\lambda}) \\ \tilde{C}_{8,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1+\beta} + \mathcal{O}(t_f^{2\beta\rho+1+\beta-\lambda}) \end{bmatrix}, & \text{if } \rho \text{ odd,} \\ \begin{bmatrix} \tilde{C}_{7,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1-\beta+\lambda} + \mathcal{O}(t_f^{2\beta\rho+1-\beta}) \\ \tilde{C}_{8,\tau}(\tilde{t}_f) t_f^{2\beta\rho+1+\beta+\lambda} + \mathcal{O}(t_f^{2\beta\rho+1+\beta}) \end{bmatrix}, & \text{if } \rho \text{ even.} \end{cases} \quad (4.32)$$

Finally, we can find an estimate for the global error by adding the contributions of all trees, according to Corollary 3.12.

Theorem 4.8. *Suppose we are solving the differential equation (4.20) where $n \geq 3$ is an odd integer and $\nu > -\frac{1}{2}(n+3)$. If a constant step-size Runge–Kutta method of order p is employed, then the global error is*

$$G_h(t_f) \approx \sum_{m=\lceil p/2 \rceil}^{p-1} h^{2m} \begin{bmatrix} \tilde{C}_{2m}^1(\tilde{t}_f) t_f^{4\beta m + \beta + 1} \\ \tilde{C}_{2m}^2(\tilde{t}_f) t_f^{4\beta m + 3\beta + 1} \end{bmatrix} + \sum_{m=\lfloor p/2 \rfloor}^{p-1} h^{2m+1} \begin{bmatrix} \tilde{C}_{2m+1}^1(\tilde{t}_f) t_f^{4\beta m + 5\beta + 2} \\ \tilde{C}_{2m+1}^2(\tilde{t}_f) t_f^{4\beta m + 7\beta + 2} \end{bmatrix} + \mathcal{O}(h^{2p}). \quad (4.33)$$

Here \tilde{C}_k^i denotes a periodic function, $\tilde{t}_f = c_1 t_f^\lambda + c_2$, and $\beta = \nu/(n+3)$.

Like in Theorem 4.6, the remainder term $\mathcal{O}(h^{2p})$ in (4.33) is not uniform in t_f . Furthermore, we see that the error coefficients of odd powers of h grow faster than the coefficients of even powers. Both remarks are illustrated in the following example.

Example 4.9. We choose the parameters $n = 3$ and $\nu = 1$, so we are solving the equation $y_1' = y_2$, $y_2' = -ty_1^3$. In this case, the function w_n , which solves $w_n'' + w_n^n = 0$ with initial conditions $w_n(0) = 0$ and $w_n'(0) = 1$, can be expressed in terms of Jacobi elliptic functions (see e.g. Neville [71]). In fact, we have $w_3(t) = \text{sd}(t | \frac{1}{2})$. The parameter $\frac{1}{2}$ will be dropped from now on. As a consequence, we can calculate the elementary integral associated with any given tree explicitly. For the first couple of trees, this yields the results listed in Table 4.3.

Tree τ	Elementary differential $F(\tau)(y)$	Elementary integral $\mathcal{I}(\tau)(t_f) = \int_{t_0}^{t_f} D\Phi_t^{t_f} F(\tau)(y) dt$
	$\begin{bmatrix} -y_1^3 y_3 \\ -3y_1^2 y_2 y_3 - y_1^3 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{1/2}) \\ \mathcal{O}(t^{5/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{128}{675} \sqrt{2} c_1^4 \chi t_f^{17/6} \text{sd}'(\tilde{t}_f) \\ \frac{256}{2025} \sqrt{2} c_1^5 \chi t_f^{19/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} 0 \\ -6y_1 y_2^2 y_3 - 6y_1^2 y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{7/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{32}{135} \sqrt{2} c_1^4 \chi t_f^{11/6} \text{sd}'(\tilde{t}_f) \\ \frac{64}{405} \sqrt{2} c_1^5 \chi t_f^{13/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} -3y_1^2 y_2 y_3 - y_1^3 \\ 3y_1^5 y_3^2 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{5/6}) \\ \mathcal{O}(t^{7/6}) \end{bmatrix}$	$\begin{bmatrix} -\frac{208}{135} \sqrt{2} c_1^4 \chi t_f^{11/6} \text{sd}'(\tilde{t}_f) \\ \frac{416}{405} \sqrt{2} c_1^5 \chi t_f^{13/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} 0 \\ -6y_2^3 y_3 - 18y_1 y_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} -\frac{4096}{14553} \sqrt{2} c_1^6 t_f^{7/2} \text{sd}'(\tilde{t}_f) \\ \frac{8192}{43659} \sqrt{2} c_1^7 t_f^{23/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} 0 \\ 6y_1^4 y_2 y_3^2 + 3y_1^5 y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} \frac{4096}{43659} \sqrt{2} c_1^6 t_f^{7/2} \text{sd}'(\tilde{t}_f) \\ -\frac{8192}{130977} \sqrt{2} c_1^7 t_f^{23/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} -6y_1 y_2^2 y_3 - 6y_1^2 y_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{7/6}) \\ 0 \end{bmatrix}$	$\begin{bmatrix} -\frac{4096}{43659} \sqrt{2} c_1^6 t_f^{7/2} \text{sd}'(\tilde{t}_f) \\ \frac{8192}{130977} \sqrt{2} c_1^7 t_f^{23/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$
	$\begin{bmatrix} 3y_1^5 y_3^2 \\ 9y_1^4 y_2 y_3^2 + 3y_1^5 y_3 \end{bmatrix} = \begin{bmatrix} \mathcal{O}(t^{7/6}) \\ \mathcal{O}(t^{3/2}) \end{bmatrix}$	$\begin{bmatrix} \frac{16384}{43659} \sqrt{2} c_1^6 t_f^{7/2} \text{sd}'(\tilde{t}_f) \\ -\frac{32768}{130977} \sqrt{2} c_1^7 t_f^{23/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}$

Table 4.3: Trees of order ≤ 4 , with their elementary differentials and integrals for the Emden–Fowler equation (4.20) with $n = 3$ and $\nu = 1$. In the last column, $\chi = \frac{1}{4K} \int_0^{4K} \text{sd}^2(t) dt$ where $4K$ is the period of the function sd , $\tilde{t}_f = c_1 t_f^{4/3} + c_2$ where c_1 and c_2 depend on the initial condition, and only the term of leading order is displayed.

We start with Runge’s second-order method, defined in (2.16). Substituting the B-series coefficients of the modified equation (cf. Example 2.4) and the elementary integrals from Table 4.3 in the estimate (3.26) of Corollary 3.12, we obtain the following global error estimate

$$G_h(t_f) \approx h^2 \begin{bmatrix} \frac{4}{15} \sqrt{2} c_1^4 \chi t_f^{11/6} \text{sd}'(\tilde{t}_f) \\ -\frac{8}{45} \sqrt{2} c_1^5 \chi t_f^{13/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix} + h^3 \begin{bmatrix} \frac{256}{6237} \sqrt{2} c_1^6 t_f^{7/2} \text{sd}'(\tilde{t}_f) \\ -\frac{512}{18711} \sqrt{2} c_1^7 t_f^{23/6} \text{sd}^3(\tilde{t}_f) \end{bmatrix}. \quad (4.34)$$

Here $\chi = \frac{1}{4K} \int_0^{4K} \text{sd}^2(t) dt$ where $4K$ is the period of the function sd , and furthermore, $\tilde{t}_f = c_1 t_f^{4/3} + c_2$.

As in the previous section, we perform a numerical experiment to check this estimate. We solve the equation (4.20) with $\nu = 1$ and $n = 3$ with Runge’s method (2.16). The initial condition is $y(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, which leads to a solution with $c_1 \approx 0.7$. The numerical solution is compared to the result of the standard fourth-order Runge–Kutta method (2.18) with $h = 1/10000$. According to Corollary 3.12, this would give an error of about 10^{-9} , so we can consider this to be the exact solution. The global error G_h is computed by subtracting the result of Runge’s method from the “exact” solution. The first component of the global error is depicted in Figure 4.4. Again, the left column shows the time interval $[0, 50]$, and on the right the larger interval $[0, 2000]$ is displayed.

As we can see in the left-hand column, the estimate (4.34) describes the actual error accurately over the interval $[0, 50]$. In the right-hand column, the oscillations of the global error are again compressed to a grey blob. The dashed curve shows the first, leading term of the estimate (4.34), and the solid curve shows the sum of both terms. We conclude that the leading h^2 term of the estimate does not describe the actual error correctly, but that the error is predicted accurately if the h^3 term is included. For $h = 1/1000$ the latter estimate breaks down around $t_f = 1200$. At this point, the h^3 term of the error estimate (4.34) and the actual solution are of equal magnitude, so the numerical solution is meaningless beyond this point. In this respect we are in the same situation as in the example in the previous section, but there is a marked difference: because of the different growth rates of the h^2 and h^3 terms, the h^3 term overtakes the h^2 term when t_f is of the order $h^{-3/4}$, which happens well before the numerical solution becomes meaningless. We conclude that for the Emden–Fowler equation, we cannot restrict our attention to the leading term of the global error, which can easily be computed using (3.16), because the second term in the error expansion dominates (at least for Runge’s method). This shows the utility of Corollary 3.12.

Another way to put it is to say that Runge’s method is essentially behaving

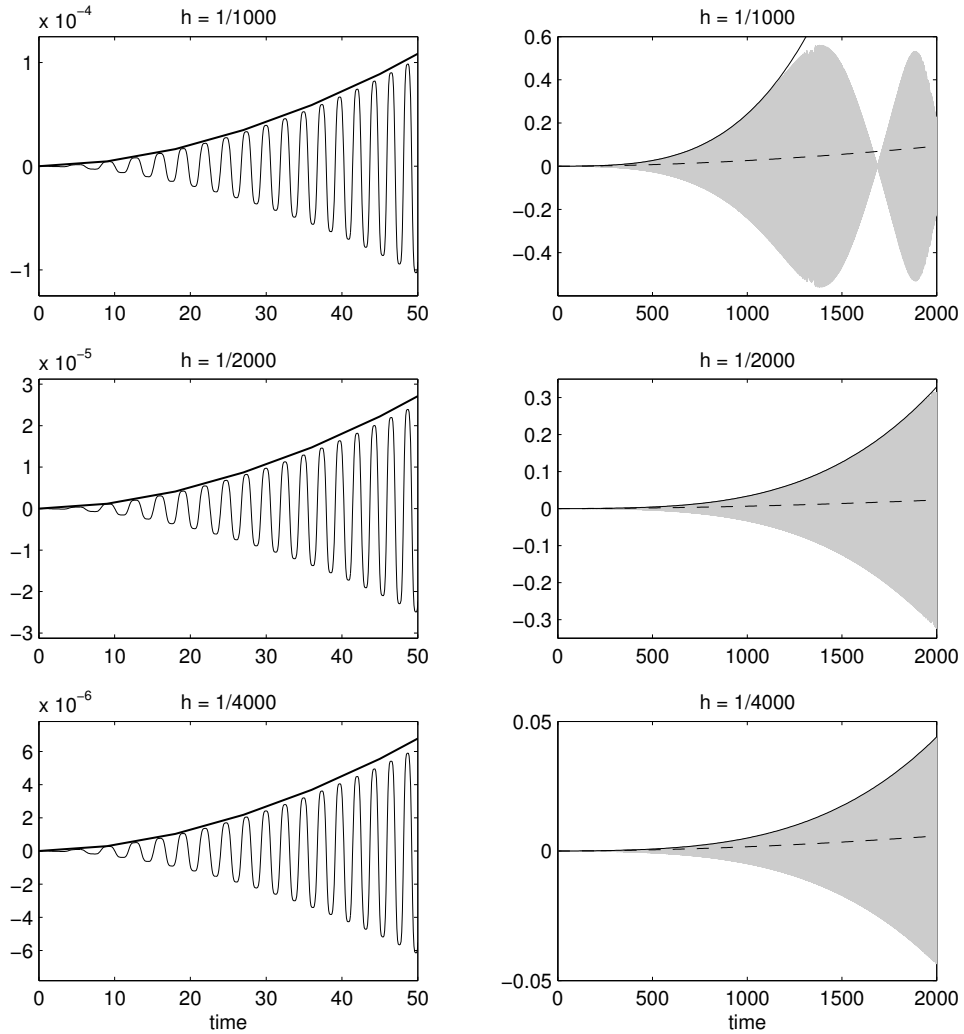


Figure 4.4: On the left, the oscillating curve shows the first component of the global error committed by Runge’s second-order method for various step sizes when applied to (4.20) with $\nu = 1$ and $n = 3$. The thick curve shows the envelope of the oscillations as predicted by (4.34). On the right, a different time scale is used. The true error is shown in grey, the dashed curve shows the first, leading term of the error estimate (4.34), and the solid curve shows the sum of both terms.

as a third-order method for large enough t_f . To check this, we compare Runge’s method with Heun’s third-order method, given in (2.17). A similar calculation as for Runge’s method, based on Theorem 4.8, yields the following estimate for the global error committed by Heun’s method,

$$E_h(t_f) \approx h^3 \left[-\frac{512}{35721} \sqrt{2} c_1^6 t^{7/2} \text{sd}'(\tilde{t}_f) \right] + h^4 \left[\frac{50208}{229635} \sqrt{2} c_1^6 t^{5/2} \text{sd}'(\tilde{t}_f) \right] \\ + h^5 \left[\frac{557056}{34543665} \sqrt{2} c_1^8 \chi t^{25/6} \text{sd}'(\tilde{t}_f) \right] \\ - \left[\frac{1024}{107163} \sqrt{2} c_1^7 t^{23/6} \text{sd}^3(\tilde{t}_f) \right] - \left[-\frac{60416}{688905} \sqrt{2} c_1^7 t^{17/6} \text{sd}^3(\tilde{t}_f) \right] - \left[-\frac{1114112}{103630995} \sqrt{2} c_1^9 \chi t^{9/2} \text{sd}^3(\tilde{t}_f) \right]. \quad (4.35)$$

The actual error and the above estimate, for step size $h = 1/2000$, are displayed in the second row of Figure 4.5. We see that the error estimate (4.35) again provides an excellent description of the actual error. When the second row in Figure 4.5 is compared with the first row, which corresponds to Runge’s second order method discussed before, the difference in order clearly shows for small values of t_f (see the left-hand column). For larger values of t_f however (cf. the right-hand column), Runge’s method behaves essentially as a third-order method and we see indeed that the difference between the two methods is much smaller.

The bottom row of Figure 4.5 shows a specially tuned method. Remember that for the linear oscillator (4.2), which was studied in Section 4.2, the contribution of the branchless tree τ_k^\perp to the global error dominates the contribution of the other trees of the same order. This is not the case for the Emden–Fowler oscillator (4.20). However, if we study the elementary integrals $\mathcal{I}(\tau)$ in Table 4.3 carefully, we see that they are scalar multiples of each other. Indeed, we have

$$4\mathcal{I}(\bullet \searrow \bullet \nearrow \bullet)(t_f) = -12\mathcal{I}(\bullet \searrow \bullet \nearrow \bullet)(t_f) = 12\mathcal{I}(\bullet \searrow \bullet \nearrow \bullet)(t_f) = -3\mathcal{I}(\bullet \vdots \bullet)(t_f).$$

Therefore, bearing in mind the factor $\sigma(\tau)$ in Corollary 3.12, the h^3 term in the global error estimate (4.33) will be killed for a third-order method with

$$b(\bullet \searrow \bullet \nearrow \bullet) - 2b(\bullet \searrow \bullet \nearrow \bullet) + b(\bullet \searrow \bullet \nearrow \bullet) - 8b(\bullet \vdots \bullet) = 0. \quad (4.36)$$

According to the theory of backward error analysis of ODEs, discussed in Section 2.4, a method has order three if

$$b(\bullet) = 1 \quad \text{and} \quad b(\bullet \vdots \bullet) = b(\bullet \searrow \bullet \nearrow \bullet) = b(\bullet \vdots \bullet) = 0.$$

Together, these are five conditions. An explicit 3-stage Runge–Kutta method has six free parameters, so there is some hope that we can find such a method satisfying

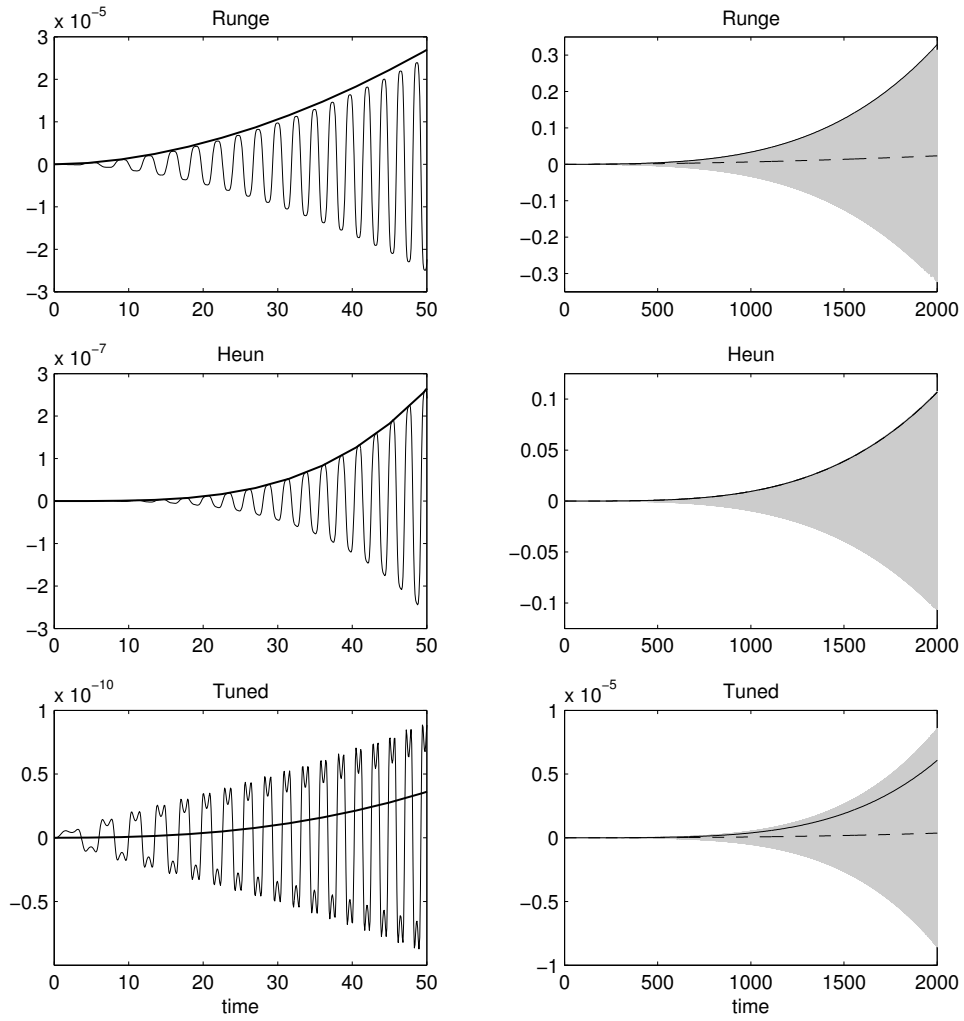


Figure 4.5: The first component of the global error committed by Runge's method (2.16), Heun's method (2.17), and the specially tuned third-order method (4.37), all with step size $h = 1/2000$, together with their respective error estimates (4.34), (4.35), and (4.38). The lines in the plots have the same meaning as in Figure 4.4.

these five conditions. Indeed, it turns out that there exists a one-parameter family of these methods. A particular instance is given by

$$\begin{array}{l|lll}
 \xi_1 = f(t_k, y_k) & 0 & 0 & 0 \\
 \xi_2 = f(t_k + h, y_k + h\xi_1) & 1 & 1 & 0 \\
 \xi_3 = f(t_k + \frac{3}{2}h, y_k + \frac{9}{4}h\xi_1 - \frac{3}{4}h\xi_2) & 3/2 & 9/4 & -3/4 \\
 y_{k+1} = y_k + \frac{7}{18}h\xi_1 + \frac{5}{6}h\xi_2 - \frac{2}{9}h\xi_3 & \hline & 7/18 & 5/6 & -2/9
 \end{array} \tag{4.37}$$

Note that the last stage of this method evaluates f at $t = t_k + \frac{3}{2}h$, which lies outside the interval $[t_k, t_{k+1}]$. In fact, all 3-stage, third order Runge–Kutta methods satisfying (4.36) do this. This is generally a bad idea because it may cause instability, but the numerical results indicate that it does not matter in this case.

If we use Theorem 4.8 to compute the estimate for the global error of this method, we find

$$G_h(t_f) \approx h^4 \left[\begin{array}{l} -\frac{5008}{25515} \sqrt{2} c_1^6 t^{5/2} \text{sd}'(\tilde{t}_f) \\ \frac{10016}{76545} \sqrt{2} c_1^7 t^{17/6} \text{sd}^3(\tilde{t}_f) \end{array} \right] + h^5 \left[\begin{array}{l} -\frac{78848}{1279395} \sqrt{2} c_1^8 \chi t^{25/6} \text{sd}'(\tilde{t}_f) \\ \frac{157696}{3838185} \sqrt{2} c_1^9 \chi t^{9/2} \text{sd}^3(\tilde{t}_f) \end{array} \right]. \tag{4.38}$$

The h^3 term has disappeared, which is indeed how we designed the method. This estimate, together with the actual error committed by (4.37), is displayed in the bottom row of Figure 4.5. We see that the global error of this method is much smaller than that of Heun’s method, though both methods are of third order. The other thing to note is that the error estimate (4.38) is not as accurate as the corresponding estimates for the other methods, even though it does predict the right order of magnitude.

The reader should keep in mind that the *local* error of the method (4.37) is still $\mathcal{O}(h^4)$, as for all third-order methods. In fact, the B-series coefficient function a of the method (4.37) satisfies

$$a(\text{tree with 3 nodes}) = \frac{1}{3}, \quad a(\text{tree with 4 nodes}) = 2, \quad a(\text{tree with 5 nodes}) = 2, \quad a(\text{tree with 6 nodes}) = 0.$$

So the local error of this method is (cf. Section 2.3)

$$B(a - \frac{1}{\gamma}, y) = h^4 \left[\begin{array}{l} -\frac{1}{8}y_1^5 y_3^2 - \frac{1}{4}y_1 y_2^2 y_3 - \frac{1}{4}y_1^2 y_2 \\ \frac{3}{8}y_1^4 y_2 y_3^2 + \frac{1}{4}y_1^5 y_3 + \frac{1}{6}y_2^3 y_3 + \frac{1}{2}y_1 y_2^2 \end{array} \right] + \mathcal{O}(h^5).$$

It is only when the individual local errors are combined to form the global error, that something special happens with the method (4.37). While for a general third-order method, the leading global error term is of the order $h^3 [t_f^{7/2} t_f^{23/6}]^\top$, this term disappears for the method (4.37). There is still an h^3 contribution to the global error, but it is only of the order $h^3 [t_f^{13/6} t_f^{3/2}]^\top$; this is the remainder term in (4.29). ◇

Part II

Minimizing the global error

Chapter 5

Formulation of the optimization problem

In the second part of this thesis, as in the first part, we study the numerical solution of the initial value problem

$$y' = f(t, y), \quad y(t_0) = y_0 \in \mathbf{R}^d. \quad (2.1)$$

We are considering one-step methods with variable step size of the form described in Section 2.5, namely

$$\begin{aligned} h_k &= \varepsilon_h h(t_k, y_k), \\ t_{k+1} &= t_k + h_k, \\ y_{k+1} &= \Psi_{h_k}(t_k, y_k). \end{aligned} \quad (2.35)$$

In the first part, we derived estimates for the global error committed by such methods. The global error is defined by

$$G_h(t_f) = y_N - y(t_f) \quad \text{where} \quad t_f = t_N. \quad (2.12)$$

Recall that ε_h is the reference step size, reflecting the user-specified tolerance. We assume that ε_h is small, so all results are only valid in the limit $\varepsilon_h \rightarrow 0$.

Now, we switch from observing to controlling. Specifically, we try to vary the step size in such a way, that the global error is as small as possible. This yields an optimization problem: how to choose the step size sequence in order to minimize the global error?

The aim of this chapter is to formulate this optimization problem precisely; the next two chapters are about solving the problem. In particular, we need to describe the objective function: exactly which quantity do we want to minimize? First, we take the most straightforward choice, namely the error at the end of integration

interval. However, as we will see in Section 5.1, this yields unsatisfactory results. The next attempt takes the error at every instant t_k into account. Unfortunately, this approach also has its shortcomings. Finally, in Section 5.3, we consider the global error as a function defined on the whole time interval, instead of only at the intermediate points t_k . We then minimize the norm of this function, yielding an optimal control problem. We conclude that this formulation is the most appropriate, and we will return to it in the next two chapters.

5.1 Minimizing the final error

The most obvious choice for the objective is to minimize $\|G_h(t_f)\|$, the Euclidean norm of the global error at the end of the integration interval, subject to the requirement that a certain number of steps, say N , be used. This requirement needs to be stipulated, as otherwise the error can be made arbitrarily small by taking a very large number of tiny steps. The optimization problem can be formulated as follows

$$\underset{h}{\text{minimize}} \|G_h(t_f)\| \quad \text{subject to} \quad t_N = t_f. \quad (5.1)$$

Recall from Section 2.5 that $h \in \mathbf{R}^N$ is the vector containing the step sizes h_0, h_1, \dots, h_{N-1} . So (5.1) is an N -dimensional optimization problem.

Morrison [70] appears to be the first one who considered this problem. Greenspan, Hafner and Ribarič [35] did some further investigations, and Gear [31] considered the case $d > 1$. Fujii [30] extended the analysis to multistep methods, while Butcher [14] considered methods that vary both the step size and the order. Also relevant is the work of Utumi, Takaki and Kawai [83], to which we will return in Section 5.3.

In the simplest case, when we are solving Dahlquist's test equation $y' = \beta y$, this works fine. We can prove that the optimal strategy is to use constant step size.

Theorem 5.1. *Suppose that we are solving the equation $y' = \beta y$ with $\beta \in \mathbf{R}$. If we are using the Euler method, then the solution of the optimization problem (5.1) is given by $h_k = t_f/N$.*

Proof. For this equation, the Euler method is $y_{k+1} = (1 + \beta h_k)y_k$. Assume that the initial condition is given by $y(0) = y_0$. We have, by the algebraic-geometric mean inequality,

$$y_N = \left(\prod_{k=0}^{N-1} (1 + \beta h_k) \right) y_0 \leq \left(1 + \frac{\beta t_f}{N} \right)^N y_0. \quad (5.2)$$

The inequality $1 + x < e^x$ shows that the right-hand side of (5.1) is strictly smaller than $y(t_f) = y_0 e^{\beta t_f}$. Therefore, the difference between y_N and $y(t_f)$ is minimized if we have equality in (5.1), which happens when the h_k are all equal. \square

The same result holds if we minimize the leading term of the global error instead of minimizing the global error. This was already proved by Morrison [70]. The work of Greenspan, Hafner and Ribarič [35] implies that this can be generalized to any Runge–Kutta method.

However, problems arise when we study slightly more complicated equations. Generally, there are many step-size vectors h such that $G_h(t_f) = 0$, suggesting that $\|G_h(t_f)\|$ is not a good choice for the objective function. The following example elaborates on this.

Example 5.2. The Lotka–Volterra equation, a simple model for the growth of animal species, reads¹

$$u' = u(v - 2) \quad \text{and} \quad v' = v(1 - u). \quad (5.3)$$

The solutions of this equation are periodic for strictly positive u and v (except at the equilibrium point). If we take the initial values $u(0) = 3$ and $v(0) = 2$, then the period is $t_* \approx 4.956$ (see Figure 5.1).

Suppose we are using the Euler method (2.11) to solve this differential equation. Consider the carefully chosen step size sequence

$$h_0 = 3.83287\dots, \quad h_1 = 0.50531\dots, \quad h_2 = 0.61785\dots \quad (5.4)$$

If we take three steps of lengths h_0 , h_1 and h_2 respectively, then we return to the point (u_0, v_0) , as shown in Figure 5.1. Furthermore, $t_3 = h_0 + h_1 + h_2 = t_*$. We conclude that $u_3 = u(t_3)$ and $v_3 = v(t_3)$, so the global error at t_3 is zero. Hence, (5.4) is a solution for the optimization problem (5.1) with $N = 3$ and $t_f = t_*$. Even worse, for any $N \geq 3$, a solution for (5.1) is given by taking three steps according to (5.4) and then taking $N - 3$ steps of length zero. Of course, the numerical solution is ridiculously far from the exact solution, even though the objective function is zero. \diamond

We conclude that, in general, it is not sufficient to look only at the global error committed at the end point. We need to replace (5.1) by another formulation.

¹We have chosen the constants 2 and 1 arbitrarily.

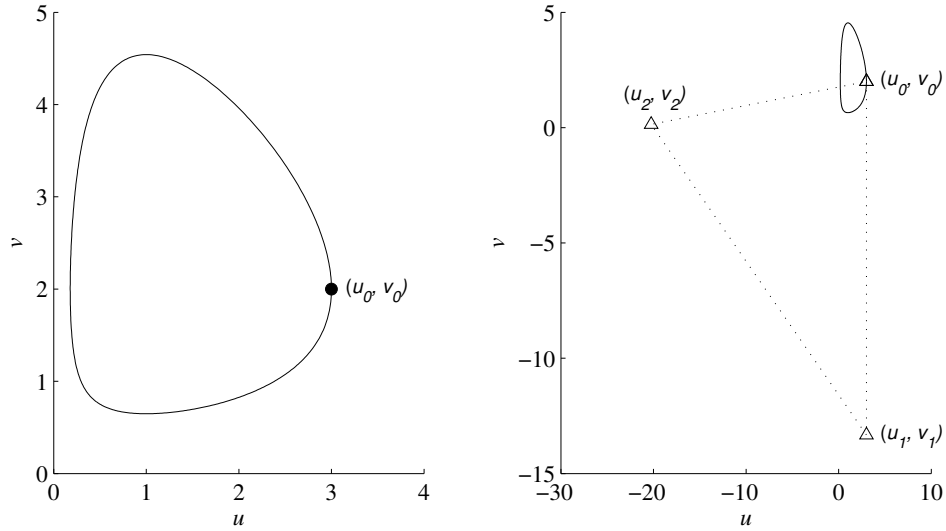


Figure 5.1: The exact solution of the Lotka–Volterra equation (5.3) with initial values $u(0) = 3$ and $v(0) = 2$ (left) and the numerical solution produced by the Euler method with three steps, whose lengths are given by (5.4).

5.2 Minimizing the error at all intermediate points

The example in the previous section shows that we cannot consider only the error at the end point. The next step is to take the global error at all the points t_k with $k = 1, \dots, N$ into account. We can formulate this as the following optimization problem,

$$\text{minimize}_h \sum_{k=1}^N \|G_h(t_k)\| \quad \text{subject to} \quad t_N = t_f. \tag{5.5}$$

Of course, different formulations are also possible. For instance, we may replace the sum in (5.5) by the sum of squared errors instead, or by a weighted sum.

Preliminary numerical experiments indicate that the optimization problem (5.5) is rather hard to solve when N is about a hundred or more. However, there is a more fundamental problem with the formulation (5.5), as will become apparent in the following example.

Example 5.3. Consider a unit mass particle moving in the double-well potential $V(q) = \frac{1}{24}q^2(3q^2 + 2q - 9)$. The motion of this particle is governed by the equations

$$q' = p \quad \text{and} \quad p' = -V'(q) = \frac{1}{2}q\left(q + \frac{3}{2}\right)(q - 1), \tag{5.6}$$

where q stands for the position and p denotes the momentum. The initial conditions are $q(0) = 1.290908997$ and $p(0) = -0.9206021281$.

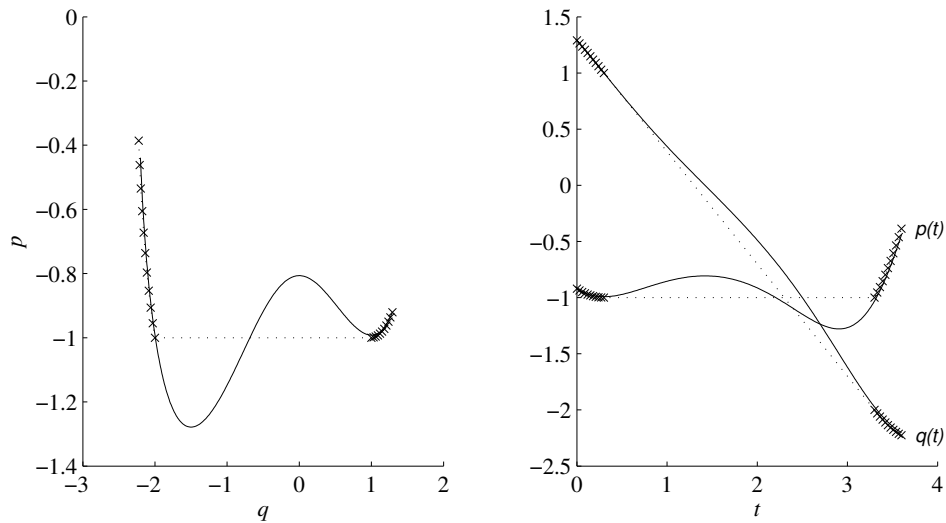


Figure 5.2: The solid curves show the exact solution of (5.6), while the crosses connected by the dotted lines show the numerical results of the Euler method with step sizes given by (5.7).

Suppose we solve the equation (5.6) with the Euler method, and that we take twenty-one steps: ten short steps, followed by one very long step, and finally ten more short steps. Specifically, we take

$$h_0 = \cdots = h_9 = 0.03, \quad h_{10} = 3, \quad \text{and} \quad h_{11} = \cdots = h_{20} = 0.03. \quad (5.7)$$

The numerical results and the exact solution are shown in Figure 5.2. The picture shows that the numerical results (the crosses) are fairly close to the exact solution (the solid line). Hence, the global error G_h is fairly small at the intermediate points t_k , and the same goes for the objective function in (5.5). However, the numerical solution is far worse than this suggests, because it does not approximate the exact solution well in the middle of the integration interval. \diamond

The conclusion from the above example must be that the optimization problem (5.5) is not a good model either. We need to consider not only the error at the end point t_f , as in (5.1), or at the intermediate points t_k , as in (5.5), but throughout the whole interval $[t_0, t_f]$.

5.3 Formulation as an optimal control problem

In this section, we minimize the global error over the whole integration interval. Our objective function is thus $\|G_h\|$, where $\|\cdot\|$ denotes some norm of the function G_h . In particular, we will consider in Chapter 6 the L_2 norm

$$\|G_h\|_2 = \left(\int_{t_0}^{t_f} \|G_h(t)\|^2 dt \right)^{1/2},$$

and in Chapter 7 the maximum norm

$$\|G_h\|_\infty = \max_{t \in [t_0, t_f]} \|G_h(t)\|.$$

This implies that we are now considering the global error as a function defined on the whole interval $[t_0, t_f]$. We will explain below how to interpret $G_h(t)$ when t is not one of the grid points t_k .

We also embed the step size h_k in a continuous function $h(t)$. So, we suppose that the numerical method has the form (2.35). However, we will assume that the step size is given by $h_k = \varepsilon_h h(t_k)$ instead of $h_k = \varepsilon_h h(t_k, y_k)$. The latter expression may appear to be more general, but in fact, all possible step size strategies for a given initial value problem can be written in the form $h_k = \varepsilon_h h(t_k)$.

Two things need to be specified before we can give a complete formulation of the optimization problem, namely, how to formulate the constraint that the number of steps be fixed, and how exactly to interpret $G_h(t)$ at every $t \in [t_0, t_f]$.

Define $t(\kappa)$ to be the solution of $\frac{dt}{d\kappa} = h(t)$ with initial condition $t(0) = t_0$. It was shown in the proof of Theorem 3.6 that $t_k = t(k\varepsilon_h) + \mathcal{O}(\varepsilon_h)$. Neglecting the remainder term, we conclude that the integration interval $[t_0, t_f]$ is traversed in N steps if $t(N\varepsilon_h) = t_f$. However, the differential equation for $t(\kappa)$ can easily be solved by separation of variables. The result is that the condition $t(N\varepsilon_h) = t_f$ is equivalent to $\int_{t_0}^{t_f} \frac{1}{h(t)} dt = N\varepsilon_h$.

Having conceded an error of order ε_h in the definition of the constraint, we can allow for a similar error in the objective function. Thus, we can restrict ourselves to the leading term of the global error. As before, we will denote this term by $\varepsilon_h^p g(t)$, where p is the order of the numerical method. At the end of Section 3.1, we found that $g(t)$ satisfies the differential equation $g' = \frac{\partial f}{\partial y} g + h^p \ell$, cf. (3.18), if the local error of the numerical method is $L_h(t, y) = h^{p+1} \ell(t, y) + \mathcal{O}(h^{p+2})$.

Summarizing, the minimization problem that we want to solve, is

$$\begin{aligned} & \underset{h}{\text{minimize}} \|\varepsilon_h^p g\| \quad \text{subject to} \quad \int_{t_0}^{t_f} \frac{1}{h(t)} dt = N\varepsilon_h \\ & \text{where} \quad g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned}$$

We can remove the constant factor ε_h^p in the objective function, because minimizing $\|\varepsilon_h^p g\|$ is equivalent to minimizing $\|g\|$. Furthermore, we can assume that $N\varepsilon_h = t_f - t_0$ (with this normalization, ε_h denotes the average step size). Indeed, if $h^*(t)$ is the solution of the above minimization problem when $N\varepsilon_h = t_f - t_0$, then the solution for another value of ε_h is given by $\frac{t_f - t_0}{N\varepsilon_h} h^*(t)$. This reduces the problem to

$$\begin{aligned} & \underset{h}{\text{minimize}} \|g\| \quad \text{subject to} \quad \int_{t_0}^{t_f} \frac{1}{h(t)} dt = t_f - t_0 \\ & \text{where} \quad g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned} \tag{5.8}$$

Optimal Control Theory studies problems of this form, where one is minimizing a functional of some function (here g), which is determined via a differential equation by another function (here h), which one is allowed to vary. Sections 6.1 and 7.1 give a short introduction to this field. The remainder of Chapters 6 and 7 are about the solution of (5.8), if the norm in the objective function is either the L_2 or the maximum norm.

The structure of the discussion in this part of the thesis is outlined in Figure 5.3, showing how we leap back and forth between the continuous and the discrete points of view. We start with some differential equation, which has a continuous solution. This solution is approximated by a discrete process, the numerical method. The numerical method commits some error, $G_h(t_k)$, which is then embedded in the continuous function $g(t)$. This function is used to formulate the optimal control problem (5.8). In Chapters 6 and 7, we will see that this problem can be converted to a two-point boundary value problem. Finally, we use collocation to discretize the problem, so that we can solve it on a computer. This gives us an approximation to the optimal step size strategy.

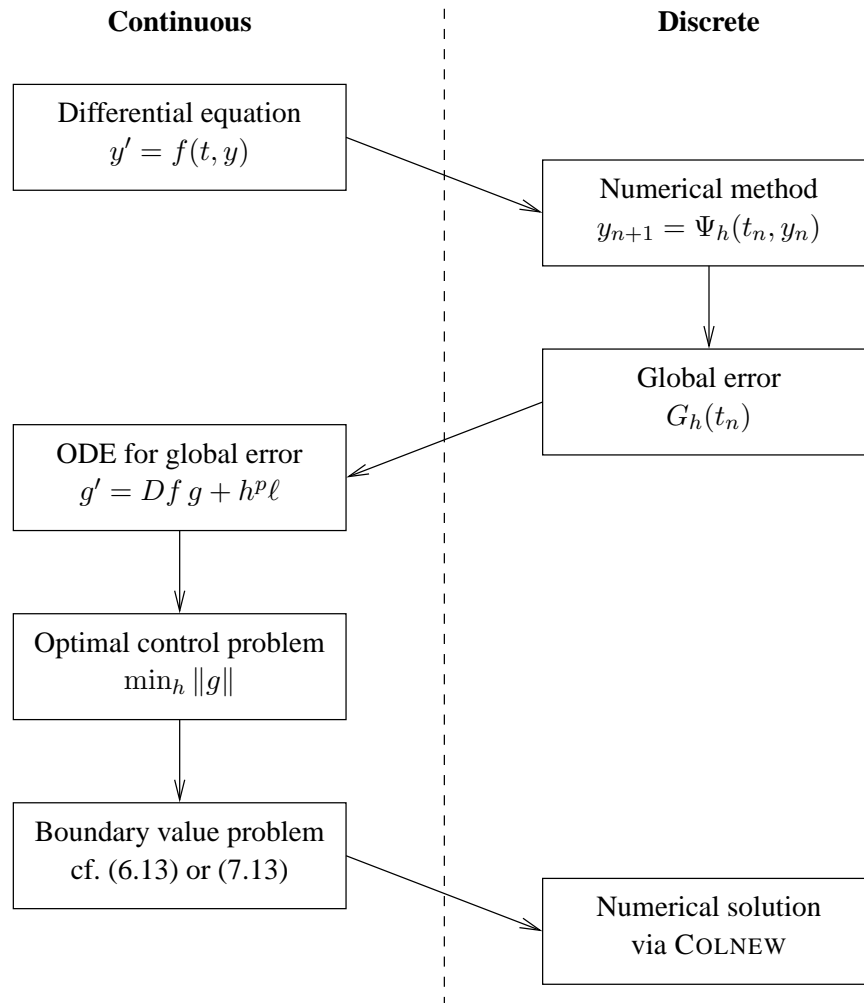


Figure 5.3: Overview of the structure of Part II of this thesis. The last two boxes will be treated in Chapters 6 and 7.

Chapter 6

Minimizing the error in the L_2 norm

In this chapter, we want to minimize the global error measured in the L_2 norm. Specifically, we want to solve the optimization problem (5.8), where the norm in the objective function is the L_2 norm. So, the problem under consideration is

$$\begin{aligned} & \underset{h}{\text{minimize}} \int_{t_0}^{t_f} \|g(t)\|^2 dt \quad \text{subject to} \quad \int_{t_0}^{t_f} \frac{1}{h(t)} dt = t_f - t_0 \\ & \text{where } g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned} \quad (6.1)$$

The values of $\frac{\partial f}{\partial y}(t, y(t))$ and $\ell(t, y(t))$ are given; they depend on the differential equation being solved and the employed numerical method, respectively.

Problems of this sort are studied in Optimal Control Theory, so we start in Section 6.1 with an introduction to this theory, concentrating on the parts which we need in the remainder of the chapter. Then, in Section 6.2, we apply this theory to the optimal control problem (6.1). This leads to the main result of the chapter, Theorem 6.5, which states a boundary value problem equivalent to (6.1). Unfortunately, the boundary value problem can rarely be solved analytically, so in the final section we discuss how to treat it numerically. We illustrate the procedure by applying it to two differential equations, namely the trivial equation $y' = ky$ and the Kepler problem.

6.1 Optimal control problems

This section gives a short introduction to Optimal Control Theory. All results in this section are well known and treated in many books, including the very readable text book by Macki and Strauss [62], and the monographs by Cesari [22] and Lee and Markus [60], which cover the subject in more depth.

In Control Theory, one studies dynamical systems of the form¹

$$x'(t) = \hat{f}(t, x(t), u(t)), \quad x(t_0) = x_0.$$

Here, $x(t)$ denotes the *state* of the system at a certain time t . The problem is to choose the function u , so that a certain goal is reached. For instance, one can ask to steer the state to a given target. We call u the *control* variable.

For example, consider a ship which has to be brought to some port. The state $x(t)$ models the current velocity and the direction that the ship is facing, while the control $u(t)$ models the direction of the rudder and the thrust of the engine. We want to find a control u , such that the ship is at rest at a certain position in the harbour at a given time. Navigational problems of this sort were originally studied by Zermelo [87].

In general, there are many controls which reach the required target. This freedom can be used to achieve a secondary goal. In the above example with the ship, we can ask for the fuel consumption to be as small as possible. In general, we are considering problems of the form

$$\underset{u \in \mathcal{U}}{\text{minimize}} \int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) dt \quad \text{subject to} \quad x(t_f) \in \mathcal{T} \quad (6.2)$$

$$\text{where} \quad x'(t) = \hat{f}(t, x(t), u(t)), \quad x(t_0) = x_0.$$

The function \hat{f}_0 models the objective, and the set \mathcal{T} denotes the target that we want to reach. The set \mathcal{U} of admissible controls will be specified later. We will denote the integral $\int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) dt$ by J .

Note that the problem (6.1), which is being studied in this chapter, is almost of the form (6.2), if we take the global error $g(t)$ as the state of the system, and the step size $h(t)$ as the control variable. Only the constraint $\int_{t_0}^{t_f} \frac{1}{h(t)} dt = 1$ does not fit. Fortunately, there is a standard trick which converts (6.1) to the form (6.2): an extra variable $k(t)$ is introduced to keep track of the integral $\int \frac{1}{h(t)} dt$. So, instead

¹We denote the right-hand side with \hat{f} instead of the customary f to avoid confusion with the differential equation (2.1) being solved by the numerical method.

of (6.1), we consider the problem

$$\begin{aligned} & \underset{h}{\text{minimize}} \int_{t_0}^{t_f} \|g(t)\|^2 dt \quad \text{subject to} \quad k(t_f) = t_f - t_0 \\ & \text{where} \quad g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0, \quad (6.3) \\ & \quad \quad \quad k'(t) = \frac{1}{h(t)}, \quad k(t_0) = 0. \end{aligned}$$

The problem (6.3) is indeed in the standard form (6.2), with the state being the pair (g, k) . The target set \mathcal{T} consists of all states with $k = t_f - t_0$.

Optimal Control Theory studies problems of the form (6.2) and their generalizations, like the problem (7.3) treated in the next chapter. The basic questions are the same as in other disciplines of Optimization Theory. On the theoretical side, we want to have existence theorems, telling us that under certain conditions the problem (6.2) has a solution. However, these existence theorems are typically nonconstructive, so we cannot use them to find a solution. Therefore, we are also seeking necessary and sufficient conditions for a pair (x, u) to be a solution of the optimal control problem (6.2).

Before we turn to these questions, we need to discuss some technical details. We assume that the function \hat{f} is continuously differentiable, that the function \hat{f}_0 is continuous, and that the target set \mathcal{T} is a smooth manifold. There are various possibilities to define the set \mathcal{U} of admissible control functions, with more or less generality. For our purposes, it is best to take \mathcal{U} to be the set of measurable functions (the example in Section 7.2 shows that we cannot require u to be smooth). If u is a measurable function, and the state x is an absolutely continuous function² satisfying the equation $x' = \hat{f}(t, x, u)$ for almost all t ,³ then the pair (x, u) is called an *admissible pair*. The optimal control problem (6.2) asks for an admissible pair which minimizes the objective $J = \int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) dt$ among all the admissible pairs. Such a pair is called a *optimal pair*.

The control u determines the state x via the equation $x' = \hat{f}(t, x, u)$, and hence, indirectly, the objective J . If the mapping $u \mapsto J$ is a lower semi-continuous function with a compact domain, then the optimal control problem (6.2) has a solution. Many existence theorems in Optimal Control Theory can be thought of as a consequence of this basic fact. A typical representative is the following.

²A function x is *absolutely continuous* on an interval I if for all $\epsilon > 0$ there exists a $\delta > 0$ such that $\sum_i |x(\beta_i) - x(\alpha_i)| \leq \epsilon$ whenever $(\alpha_1, \beta_1), \dots, (\alpha_N, \beta_N)$ are disjoint subintervals of I whose total length is at most δ . An absolutely continuous function has a derivative almost everywhere.

³A condition, depending on a variable t , is said to hold for *almost all* t , if the set where the condition does *not* hold has measure zero. We will often use the abbreviation *a.a.* for *almost all*.

Theorem 6.1. *If there is at least one admissible pair, if all admissible pairs (x, u) satisfy a priori bounds $\|x(t)\| \leq x_{max}$, $\|u(t)\| \leq u_{max}$, for all t , and if the set*

$$\{(y_0, y) \mid y_0 \geq \hat{f}_0(t, x, u) \text{ and } y = \hat{f}(t, x, u) \text{ for some admissible } u\}$$

is convex for all (t, x) , then there exists an optimal pair for the optimal control problem (6.2).

This theorem is stated and proved by Berkovitz [10, Ch. III, Thm. 5.1].

Theorem 6.1 guarantees the existence of an optimal pair, but does not tell us how to find it. For this, we need Pontryagin's Minimum Principle, which gives a necessary condition for an admissible pair to be optimal.

Theorem 6.2 (Pontryagin's Minimum Principle). *Consider the problem (6.2). Suppose that the set \mathcal{U} of admissible controls has the form*

$$\mathcal{U} = \{u \mid u \text{ is measurable and } u(t) \in \hat{\mathcal{U}} \text{ for a.a. } t\}, \quad (6.4)$$

for some set $\hat{\mathcal{U}}$. If (x, u) is an optimal pair for (6.2), then there exist a $\lambda_0 \geq 0$ and an absolutely continuous function λ , not both zero, such that

$$\frac{d}{dt} \lambda(t) = -\frac{\partial H}{\partial x} \text{ for a.a. } t, \quad \text{and} \quad (6.5)$$

$$\lambda(t_f) \text{ is orthogonal to } \mathcal{T} \text{ at } x(t_f), \quad \text{and} \quad (6.6)$$

$$H(t, x(t), u(t), \lambda_0, \lambda(t)) = \min_{v \in \hat{\mathcal{U}}} H(t, x(t), v, \lambda_0, \lambda(t)) \text{ for a.a. } t, \text{ and} \quad (6.7)$$

$$H(t, x(t), u(t), \lambda_0, \lambda(t)) = -\int_{t_0}^t \lambda(s)^\top \frac{\partial \hat{f}}{\partial t}(s, x(s), u(s)) ds, \quad (6.8)$$

where the Hamiltonian H is defined by $H = \lambda_0 \hat{f}_0 + \lambda^\top \hat{f}$.

This result is called the Minimum Principle because of equation (6.7), which says that the optimal control at almost any time is the control which minimizes the Hamiltonian. A proof is given by Lee and Markus [60, §5.1].

The Hamiltonian H is said to be *regular* if the minimizer $v \in \hat{\mathcal{U}}$ in (6.7) is unique. In this case, the conclusion in Theorem 6.2 can be simplified considerably.

Theorem 6.3. *Suppose that all assumptions of Theorem 6.2 are satisfied. If furthermore the Hamiltonian is regular, then the optimal control u is continuous, and both the state x and the costate λ are C^1 . Hence, the conditions (6.5) and (6.7) are satisfied for all t .*

Proof. Jacobson, Lele, and Speyer [56] prove (in a more general setting) that the optimal control is continuous (see also Theorem 7.3). The other conclusions of the theorem follow by Remark 2 in [22, §4.2C]. \square

The above theorems give only a *necessary* condition for optimality. However, under some extra assumptions, it is also a *sufficient* condition. The following theorem, due to Lee and Markus [60, §5.2], gives the details.

Theorem 6.4. *Suppose that the optimal control problem has the form*

$$\underset{u \in \mathcal{U}}{\text{minimize}} \int_{t_0}^{t_f} \hat{f}_0(t, x(t)) + \hat{h}_0(t, u(t)) \, dt \quad \text{subject to} \quad x(t_f) \in \mathcal{T} \quad (6.9)$$

$$\text{where} \quad x'(t) = A(t)x(t) + \hat{h}(u(t), t), \quad x(t_0) = x_0.$$

Suppose furthermore that \mathcal{U} has the form (6.4), the target set \mathcal{T} is convex, and the function \hat{f}_0 is convex in x for all t . If (x, u) is an admissible pair and there exist a $\lambda_0 \geq 0$ and an absolutely continuous function λ , not both zero, which satisfy the conditions (6.5), (6.6), (6.7), and (6.8) of Pontryagin's Minimum Principle, then (x, u) is an optimal pair.

Optimal Control Theory is closely connected to the Calculus of Variations. Most variational problems can be considered as optimal control problems, and vice versa, as described by Hestenes [50].

An alternative approach is to use techniques from Dynamic Programming. This field was originally conceived by Bellman [9] as an effective computational method for dealing with optimal decision making in discrete time processes. The theory was later extended to continuous time via a limiting process. The basic idea is the *Principle of Optimality*, which says that from any point on an optimal trajectory, the remaining trajectory is optimal for the corresponding problem initiated at that point. This leads to the definition of the *value function*, which associates to a time t and a state x , the cost incurred by an optimal pair for the problem initiated at that point. This value function satisfies a partial differential equation, called the *Bellman equation*. By solving this equation, one can find the value function, and subsequently the optimal control. A popular reference for Dynamic Programming is the two-volume work by Bertsekas [12].

The Dynamic Programming approach might be considered quite natural for the problem at hand, as the limiting process for going from discrete time to continuous time reflects the limit where the step size of the numerical method goes to zero. Nevertheless, it seems that using Pontryagin's Minimum Principle is the more fruitful approach, because the Bellman equation is a complicated partial differential equation. For this reason, we abandon the Dynamic Programming point of view.

6.2 Analytic treatment

In this section, we apply the theory of the preceding section to the optimal control problem (6.1), which asks for the step size $h(t)$ that minimizes the L_2 norm of the global error. This follows a suggestion in the appendix of a paper by Utumi, Takaki and Kazai [83].

We found in the previous section that (6.1) is equivalent to (6.3), which is in the standard form (6.2). The first question under consideration is whether the existence of an optimal pair can be guaranteed. If we could establish an *a priori* bound on the step size function h , say $h(t) \leq h_{\max}$ for all t , then this would imply a bound on the global error $g(t)$. Indeed, the solution of the differential equation for g is (cf. Theorem 3.6)

$$g(t) = \int_{t_0}^t h(s, y(s))^p D\Phi_s^t(y(s)) \ell(s, y(s)) ds,$$

and both the variational flow matrix $D\Phi$ and the local error $\ell(t, y)$ are bounded. In this case, the existence of an optimal pair would follow from Theorem 6.1. However, there seems to be no reason to assume an *a priori* bound on the step size. In fact, as we will see after the forthcoming Theorem 6.5, the optimal step size is unbounded as t approaches the final time t_f . For this reason, we are unable to make any claims on the existence of a solution to the optimal control problem (6.3).

We now turn to the second theorem mentioned in the previous section, namely Pontryagin's Minimum Principle (Theorem 6.2). Let $\gamma \in \mathbf{R}^d$ and $\kappa \in \mathbf{R}$ denote the adjoint variables of g and k , respectively. Then the Hamiltonian is

$$H(t, g, k, h, \lambda_0, \gamma, \kappa) = \lambda_0 \|g\|^2 + \gamma^\top \left(\frac{\partial f}{\partial y}(t, y(t)) g + h^p \ell(t, y(t)) \right) + \frac{\kappa}{h}. \quad (6.10)$$

Hence the adjoint variables evolve according to

$$\gamma'(t) = -\frac{\partial H}{\partial g} = -\left(\frac{\partial f}{\partial y}(t, y(t)) \right)^\top \gamma(t) - 2\lambda_0 g(t) \quad \text{and} \quad \kappa'(t) = -\frac{\partial H}{\partial k} = 0; \quad (6.11)$$

this is equation (6.5) from Theorem 6.2. Furthermore, equation (6.6) from the same theorem reads $\gamma(t_f) = 0$. Finally, the actual minimum principle, i.e. (6.7), states that the step size h minimizes the Hamiltonian (6.10). If both κ and the inner product $\gamma^\top \ell$ are positive, then the minimum for the Hamiltonian is attained when

$$h = \left(\frac{\kappa}{p\gamma^\top \ell} \right)^{1/(p+1)}.$$

Furthermore, this minimizer is unique, so the Hamiltonian is regular and Theorem 6.3 applies. This implies that the optimal control h is continuous, and that

the differential equations (6.11) are valid for all t . It follows that κ is a constant function. If we now set $\kappa_0 = (\kappa/p)^{1/(p+1)}$, then the above formula for h reads $h = \kappa_0(\gamma^\top \ell)^{-1/(p+1)}$. On the other hand, if $\gamma^\top \ell \leq 0$ then H is a decreasing function of h , so the Hamiltonian has no minimum and condition (6.7) can never be satisfied.

Summarizing, Pontryagin's Maximum Principle implies that if (g, k, h) is a solution to the optimal control problem (6.3), then there exist a continuously differentiable function $\gamma : \mathbf{R} \rightarrow \mathbf{R}^d$ and constants λ_0 and κ_0 , such that

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), & g(t_0) &= 0, \\ k'(t) &= \frac{1}{h(t)}, & k(t_0) &= 0, \quad k(t_f) = t_f - t_0, \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - 2\lambda_0 g(t), & \gamma(t_f) &= 0, \\ h(t) &= \kappa_0(\gamma(t)^\top \ell(t, y(t)))^{-1/(p+1)} & \text{and } \gamma(t)^\top \ell(t, y(t)) &> 0. \end{aligned} \tag{6.12}$$

We now simplify the system (6.12) by exploiting its scaling symmetries. Note that if $(g, k, h, \gamma, \lambda_0, \kappa_0)$ is a solution, then so is $(g, k, h, \alpha\gamma, \alpha\lambda_0, \alpha^{1/(p+1)}\kappa_0)$. Hence, we can assume that $\lambda_0 = 1$. If we neglect the constraint $k(t_f) = t_f - t_0$ for the moment, then we can find another scaling symmetry, namely

$$(g, k, h, \gamma, \lambda_0, \kappa_0) \mapsto (\alpha^p g, \alpha^{-1} k, \alpha h, \alpha^p \gamma, \lambda_0, \alpha^{-p/(p+1)} \kappa_0).$$

So, as long as we neglect the constraint $k(t_f) = t_f - t_0$, we can assume without loss of generality that $\kappa_0 = 1$. This also decouples the equation for $k(t)$ from the other differential equations. After we found a solution (g, γ) , we reintroduce the equation for $k(t_f)$. We can now use the freedom in κ_0 to scale the step size and thus ensure that the constraint $k(t_f) = t_f - t_0$ is satisfied.

Finally, note that the system (6.3) satisfies the conditions in Theorem 6.4. Hence, the necessary conditions from Pontryagin's Minimum Principle are in fact also sufficient conditions. We summarize our results in the following theorem.

Theorem 6.5. *Consider the optimal control problem (6.1). If (g^*, h^*) is an optimal pair for (6.1), then there exist continuously differentiable functions $g, \gamma : \mathbf{R} \rightarrow \mathbf{R}^d$ that solve the following two-point boundary value problem,*

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t)) g(t) + (\gamma(t)^\top \ell(t, y(t)))^{-p/(p+1)} \ell(t, y(t)), & g(t_0) &= 0, \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - 2g(t), & \gamma(t_f) &= 0, \end{aligned} \tag{6.13}$$

and for some value of κ_0 , the pairs (g^*, h^*) and (g, γ) are connected by

$$g^*(t) = \kappa_0^p g(t) \quad \text{and} \quad h^*(t) = \kappa_0(\gamma(t)^\top \ell(t, y(t)))^{-1/(p+1)}. \tag{6.14}$$

Furthermore, the following condition is satisfied, ensuring that the fractional power in (6.13) is well-defined,

$$\gamma(t)^\top \ell(t, y(t)) > 0 \text{ for all } t < t_f. \quad (6.15)$$

Conversely, if a pair (g, γ) satisfies (6.13) and (6.15), then there is a constant κ_0 such that (g^*, h^*) as defined in (6.14) is an optimal pair for (6.1).

The differential equation (6.13) implies that $\gamma(t) \rightarrow 0$ as $t \rightarrow t_f$. Hence, by (6.14), the step size $h(t)$ becomes unbounded as t approaches the final time t_f . This may be surprising at first sight, but it should be borne in mind that notwithstanding the unboundedness of $h(t)$, the integration interval $[t_0, t_f]$ is traversed in a finite number of steps because $h(t)$ satisfies the constraint $\int_{t_0}^{t_f} \frac{1}{h(t)} dt = t_f - t_0$.

Furthermore, (6.14) implies that the expression

$$(h^*(t))^{p+1} \gamma(t)^\top \ell(t, y(t)) \quad (6.16)$$

is constant. Recall that the local error is $h^{p+1}(t) \ell(t, y(t))$, so we can interpret the above expression as a weighted local error. The condition that this weighted local error be constant is reminiscent of *equidistribution*. This term refers to the idea that, when solving a differential equation numerically, an efficient method commits an equal error in every subinterval (or subdomain, for partial differential equations). With other words, the error is distributed equally over the subintervals. Eriksson, Estep, Hansbo and Johnson [27, 28] describe this idea in great detail.

A similar computation can be carried out for the more general case in which we are minimizing the L_s norm of the global error, with $s \in (1, \infty)$, meaning that the optimal control problem (6.1) is replaced by

$$\text{minimize}_h \int_{t_0}^{t_f} \|g(t)\|^s dt \quad \text{subject to} \quad \int_{t_0}^{t_f} \frac{1}{h(t)} dt = t_f - t_0 \quad (6.17)$$

$$\text{where } g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0.$$

The case $s = 1$ needs to be excluded, as the objective function fails to be continuously differentiable when $s = 1$. Theorem 6.5 still holds for (6.17) with $s \in (1, \infty)$ if the boundary value problem (6.13) is replaced by

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t)) g(t) + (\gamma(t)^\top \ell(t, y(t)))^{-p/(p+1)} \ell(t, y(t)), & g(t_0) &= 0, \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - s \|g(t)\|^{s-2} g(t), & \gamma(t_f) &= 0. \end{aligned} \quad (6.18)$$

For the rest of the chapter however, we consider only the special case $s = 2$.

Unfortunately, the boundary value problem (6.13) is quite hard to solve analytically. Even determining the optimal step size for the simple equation $y' = ky$ is a problem, as the following example shows.

Example 6.6. Suppose that we are solving the differential equation $y' = ky$ with $k \in \mathbf{R} \setminus \{0\}$. We take $y(0) = 1$ as initial condition, so the exact solution is $y(t) = e^{kt}$.

The leading local error term of any Runge–Kutta method applied to this equation is proportional to $h^{p+1}y$, where p is the order of the method. We can assume without loss of generality that the constant of proportionality is one. In that case, $\ell(t, y) = y$ and the boundary value problem (6.13) reads

$$\begin{aligned} g'(t) &= kg(t) + (e^{kt}\gamma(t))^{-p/(p+1)}, & g(0) &= 0, \\ \gamma'(t) &= -k\gamma(t) - 2g(t), & \gamma(t_f) &= 0. \end{aligned} \quad (6.19)$$

These equations can be simplified in various ways. For instance, the substitution

$$\begin{aligned} g(t) &= k|k|^{-(2p+2)/(2p+1)} e^{-s} \tilde{g}(s), \\ \gamma(t) &= |k|^{-(2p+2)/(2p+1)} e^{-s} \tilde{\gamma}(s)^{p+1}, \\ t &= \frac{2p+1}{kp} s, \end{aligned}$$

transforms the problem (6.19) in

$$\begin{aligned} p\tilde{g}'(s) &= (3p+1)\tilde{g}(s) + (2p+1)\tilde{\gamma}(s)^{-p}, & \tilde{g}(0) &= 0, \\ p\tilde{\gamma}'(s) &= -\tilde{\gamma}(s) - \frac{4p+2}{p+1}\tilde{\gamma}(s)^{-p}\tilde{g}(s), & \tilde{\gamma}\left(\frac{kp t_f}{2p+1}\right) &= 0. \end{aligned}$$

Compared to (6.19), we got rid of the fractional power, and we removed the dependency of the differential equation on the parameter k . Furthermore, the transformed system is autonomous, so we can write it as a single equation if we consider \tilde{g} to be a function of $\tilde{\gamma}$,

$$\frac{d\tilde{g}}{d\tilde{\gamma}} = -(p+1) \frac{(3p+1)\tilde{\gamma}^p\tilde{g} + 2p+1}{(4p+2)\tilde{g} + \tilde{\gamma}^{p+1}}. \quad (6.20)$$

Unfortunately, neither (6.20) nor the original system (6.19) can be solved analytically. So, in the next section, we investigate the numerical solution of the boundary value problem (6.13). We then return to the system (6.19) in Example 6.7. \diamond

6.3 Numerical treatment

In this section, we describe how the boundary value problem (6.13), and hence the optimal control problem (6.1), can be solved numerically. Below, we give a very short introduction to the numerical solution of boundary value problems. More information can be found in the extensive literature that covers this field, which includes the book by Ascher and Petzold [6].

It was mentioned in Section 2.1 that initial value problems for ordinary differential equations have a unique solution if the right-hand side is differentiable. This is no longer true for a boundary value problem like (6.13), which may have zero or multiple solutions. This reflects the lack of an existence result for the optimal control problem (6.1).

With this warning in mind, we proceed with the numerical solution of (6.13). The basic numerical methods for boundary value problems are shooting and finite difference methods. A *shooting method* for (6.13) starts from a guess for $\gamma(t_0)$, and then uses any numerical method for initial value problems to solve the differential equation. In general, the solution will not satisfy the end-point condition $\gamma(t_f) = 0$, so we correct our guess for $\gamma(t_0)$ and try again, until we find a solution with $\gamma(t_f) = 0$. The shooting method has the disadvantage that it is not very robust, because it might take a long time (or even forever) before it hits the correct guess for $\gamma(t_0)$.

A *finite difference method* divides the time interval $[t_0, t_f]$ in a number of subintervals. At each of the intermediate points t_1, \dots, t_{N-1} , the derivatives occurring in the boundary value problem (6.13) are replaced by a finite difference formula. Together with the boundary conditions, this yields a large system of equations, which is subsequently solved to get a numerical solution to the boundary value problem. Solving the system of equations is feasible because of its band structure, but this method is more difficult to implement than a shooting method.

Both the shooting and the finite difference method can be extended. Here, we choose to use a *collocation method*, an extension of the finite difference method, because of the availability of an excellent implementation which performs well in practice. The idea behind collocation methods is to approximate the solution in every subinterval $[t_k, t_{k+1}]$ by a polynomial of some fixed degree d , and to require that the polynomial satisfy the differential equation at d predetermined points in the subinterval and that all the polynomials in the different subintervals fit together to form a continuous function satisfying the boundary conditions.

We use the COLNEW code by Bader and Ascher [7], which is a newer version of the COLSYS code by Ascher, Christiansen, and Russell [4, 5]. This program performs collocation at the five Gauss–Legendre points. The resulting collocation equations are solved with the damped Newton method. The program then estimates the error of the numerical solution, and compares it with the tolerance level requested by the user. If the estimated error exceeds the tolerance level, then the program determines a new subdivision of the interval $[t_0, t_f]$, either by halving the subintervals in the previous iteration or by redistributing the intermediate points t_k so that they concentrate in the regions where the error estimate is high, and an-

other iteration is started. Otherwise, the solution is accepted and returned to the user. The COLNEW code differs from the COLSYS code by using another basis to represent the solution, which speeds up the solution of the linearized system.

There is one remaining detail that needs to be taken care of, namely, the fractional power in (6.13). If the inner product $\gamma(t)^\top \ell(t, y(t))$ is negative for some t , then raising it to a fractional power results in a complex number for the step size, which does not make any sense. Theorem 6.5 guarantees us that, if there is a solution to the optimal control problem (6.1), the boundary value problem (6.13) has a solution for which the inner product is never negative, cf. (6.15). Nevertheless, it is quite possible that one of the iterates produced by the COLNEW program violates the condition (6.15). This would lead to complex numbers and terminate the iteration process, unless we make some special arrangement to circumvent the problem.

Here, we employ the following trick to overcome this difficulty. Denoting the inner product $\gamma(t)^\top \ell(t, y(t))$ by x , we replace the inner product in (6.13) by $\psi_\sigma(x)$, where ψ_σ denotes a family of real functions that satisfies

$$\psi_\sigma(x) > 0 \text{ for all } x \in \mathbf{R} \quad \text{and} \quad \psi_\sigma(x) = x \text{ for all } x \geq \sigma > 0. \quad (6.21)$$

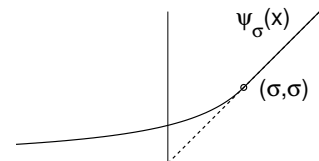
The first condition ensures that $\psi_\sigma(x)$ can be raised to a fractional power without problems, while the second condition implies that the introduction of the function ψ_σ in (6.13) has no effect as long as the inner product x is larger than the cut-off parameter σ . However, we know that x is strictly positive on (t_0, t_f) for the solution that we are looking for. Hence, we may expect that for small but strictly positive values of σ , the solution of the new boundary value problem

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t)) g(t) + \left(\psi_\sigma(\gamma(t)^\top \ell(t, y(t))) \right)^{-p/(p+1)} \ell(t, y(t)), & g(t_0) &= 0, \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t)) \right)^\top \gamma(t) - 2g(t), & \gamma(t_f) &= 0, \end{aligned} \quad (6.22)$$

is close to the solution of the original boundary value problem (6.13). On the other hand, if σ is chosen too small, then the bad scaling may cause numerical problems. For the two examples in this section, the precise choice of σ is not very important. However, the choice of the parameters is a more delicate issue for the examples in Section 7.3.

Specifically, we define ψ_σ by

$$\psi_\sigma(x) = \begin{cases} \frac{\sigma^2}{2\sigma - x}, & \text{if } x < \sigma, \\ x, & \text{if } x \geq \sigma. \end{cases}$$



This function, which is depicted on the right, is continuously differentiable and possesses the required property (6.21).

Example 6.7. In Example 6.6 in the previous section, we sought the optimal step size for solving the equation $y' = ky$ with $k \in \mathbf{R} \setminus \{0\}$ and initial condition $y(0) = 1$. The corresponding boundary value problem was found to be (6.19), but we could not solve this problem analytically. Here, we discuss its numerical solution.

We consider the Euler method (2.11) for solving the equation $y' = ky$. The Euler method has order $p = 1$ and local error $\ell(t, y) = -\frac{1}{2}k^2y$, cf. (2.10). Hence, the boundary value problem (6.22) reads

$$\begin{aligned} g'(t) &= kg(t) - \frac{1}{2}k^2e^{kt} \left(\psi_\sigma \left(-\frac{1}{2}k^2e^{kt}\gamma(t) \right) \right)^{-1/2}, & g(0) &= 0, \\ \gamma'(t) &= -k\gamma(t) - 2g(t), & \gamma(t_f) &= 0, \end{aligned} \quad (6.23)$$

We use COLNEW to solve this boundary value problem, picking $t_f = 1$, $k = -2$, and $\sigma = 10^{-3}$. COLNEW also needs an initial guess for the solution to start the iteration. Lacking much inspiration, we provide as initial guess $g(t) \equiv 0$ and $\gamma(t) = \ell(t, y(t))$, which ensures that the condition $\ell^\top \gamma > 0$ in Theorem 6.5 is satisfied.

This results in the numerical solution shown in the left-hand plot in Figure 6.1. The corresponding step size function can be retrieved using (6.14), which for this example reads

$$h(t) = \kappa_0 \left(-\frac{1}{2}k^2e^{kt}\gamma(t) \right)^{-1/2}.$$

The constant κ_0 is determined by the requirement $\int_0^{t_f} \frac{1}{h(t)} dt = t_f$. In our case, we find $\kappa_0 \approx 0.5082$, and the resulting step size function is also plotted in Figure 6.1. For comparison, the optimal step size when $k = 2$ is also displayed in this figure.

The picture shows that the step size is an increasing function, and that it becomes very large as t approaches the end of the integration interval; in fact, $h(t)$ is unbounded as $t \rightarrow t_f$ for the original problem (6.13). Nevertheless, the global error $g(t)$ remains bounded.

The following informal argument explains why the step size increases as t grows. Any error committed at some instant t_* contaminates the numerical solution over the interval $[t_*, t_f]$. Thus, it is important to avoid early errors, since they cause contamination over a long time interval. On the other hand, an error committed near the end of the integration interval increases the objective $\int_{t_0}^{t_f} |g(t)| dt$ only slightly, so one should not worry too much about taking large steps near the final time t_f . It pays therefore to concentrate one's efforts at the beginning of the integration interval.

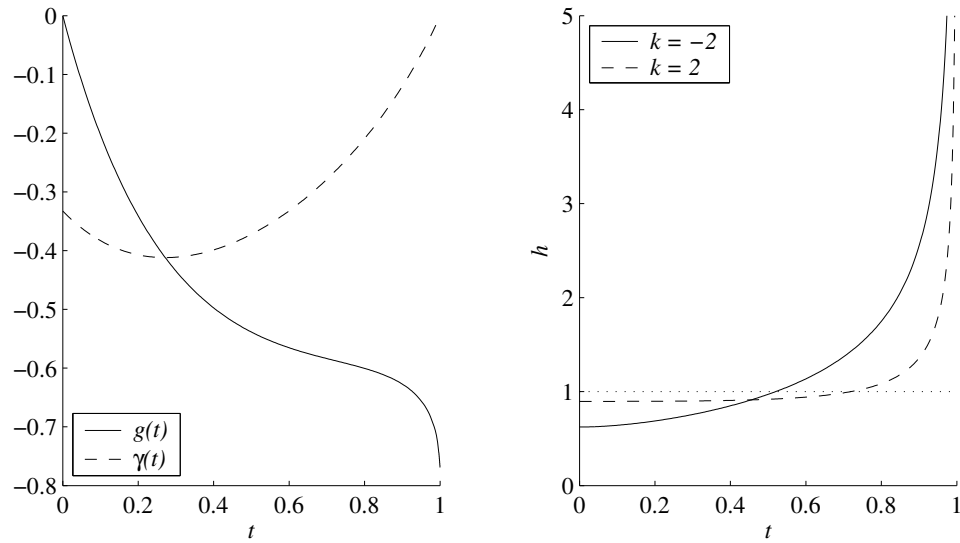


Figure 6.1: On the left, the solution of (6.23) with $t_f = 1$, $k = -2$, and $\sigma = 10^{-3}$. On the right, the corresponding step size function, and the optimal step size function when $k = 2$.

When interpreting the above results, it should be borne in mind that the step size sequence is given by $h_k = \varepsilon_h h(t_k)$, where h is the function depicted in Figure 6.1. The differential equation (3.18) describing the dynamics of the global error is only valid in the limit $\varepsilon_h \rightarrow 0$. In this limit, the global error at time t is $\varepsilon_h^p g(t) + \mathcal{O}(\varepsilon_h^{p+1})$. In particular, one cannot take the step size sequence given by $h_k = h(t_k)$ and expect the global error to be $g(t)$; this amounts to taking $\varepsilon_h = 1$ and violates the condition that ε_h be sufficiently small.

To determine the effect of the cut-off parameter σ , we repeat the numerical simulation with $\sigma = 10^{-6}$ instead of $\sigma = 10^{-3}$. The results are nearly identical, except when t is very close to the end of the interval. Indeed, the function ψ_σ acts as the identity unless its argument is less than σ . For $\sigma = 10^{-3}$, this happens only in the tiny interval $(0.998, 1)$. In conclusion, we can be confident that the solution of (6.22), which includes the cut-off function ψ_σ , with σ around 10^{-6} or 10^{-3} does not differ markedly from the solution of (6.13).

In the next chapter, specifically in Example 7.7 and Figure 7.3, the above results for the L_2 -optimal step size are compared to other step size strategies, including the L_∞ -optimal step size. \diamond

The above example concerns a very simple equation. Next, we consider a more complicated example, namely the Kepler two-body problem.

Example 6.8. The dynamics of a particle moving in the gravitational field of another particle of unit mass, fixed at the origin, is given by

$$r''(t) = -\frac{1}{\|r(t)\|^3} r(t). \quad (6.24)$$

Here, $r(t) \in \mathbf{R}^3$ is the position of the moving particle, $\|\cdot\|$ denotes the Euclidean norm, and we assume that the units are chosen such that the universal gravitational constant G equals one.

In fact, the particle will stay in a plane containing the origin, so we may assume without loss of generality that $r(t) = (y_1(t), y_2(t), 0)$. If we set $y_3 = y_1'$ and $y_4 = y_2'$, then the equation (6.24) is equivalent to the following system of first-order equations,

$$\begin{aligned} y_1'(t) &= y_3(t), \\ y_2'(t) &= y_4(t), \\ y_3'(t) &= -\frac{y_1(t)}{(y_1(t)^2 + y_2(t)^2)^{3/2}}, \\ y_4'(t) &= -\frac{y_2(t)}{(y_1(t)^2 + y_2(t)^2)^{3/2}}. \end{aligned} \quad (6.25)$$

To complete the formulation of the problem, we pick the following initial conditions,

$$y_1(0) = 2, \quad y_2(0) = 0, \quad y_3(0) = 0, \quad y_4(0) = \frac{1}{2}. \quad (6.26)$$

With these initial conditions, the particle describes an ellipse around the origin with eccentricity $\frac{1}{2}$, as depicted in Figure 6.2. The time to complete one revolution is $T \approx 9.674$.

We want to solve the problem (6.25) with initial conditions 6.26 numerically. We first assume that the Euler method (2.11) is employed. The local error committed by the Euler method is, cf. (2.10),

$$\ell(t, y) = -\frac{1}{2} Df(y) f(y) = \frac{1}{2} \begin{bmatrix} y_1 r^{-3} \\ y_2 r^{-3} \\ y_3 r^{-3} - 3y_1(y_1 y_3 + y_2 y_4) r^{-5} \\ y_4 r^{-3} - 3y_2(y_1 y_3 + y_2 y_4) r^{-5} \end{bmatrix}, \quad (6.27)$$

where $r = \sqrt{y_1^2 + y_2^2}$ denotes the distance to the origin.

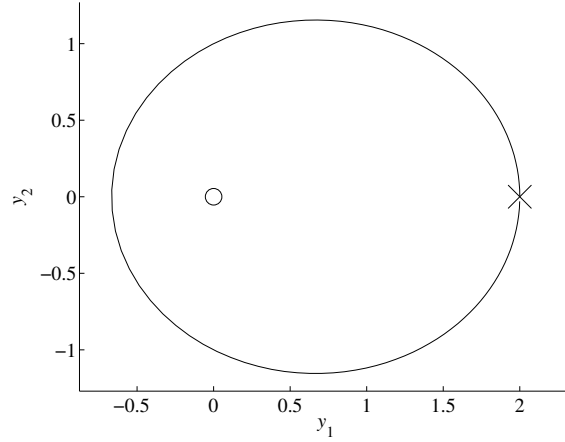


Figure 6.2: The orbit of a satellite in a gravitational field, as described by (6.25). The cross indicates the initial condition (6.26), and the circle at the origin represents the massive object generating the gravitational field.

Hence, the boundary value problem (6.22) reads

$$\begin{aligned}
g'_1 &= g_3 + \frac{1}{2}h y_1 r^{-3}, & g_1(0) &= 0, \\
g'_2 &= g_4 + \frac{1}{2}h y_2 r^{-3}, & g_2(0) &= 0, \\
g'_3 &= (3y_1^2 r^{-5} - r^{-3})g_1 + 3y_1 y_2 r^{-5} g_2 \\
&\quad + \frac{1}{2}h(y_3 r^{-3} - 3y_1(y_1 y_3 + y_2 y_4)r^{-5}), & g_1(t_f) &= 0, \\
g'_4 &= 3y_1 y_2 r^{-5} g_1 + (3y_2^2 r^{-5} - r^{-3})g_2 \\
&\quad + \frac{1}{2}h(y_4 r^{-3} - 3y_2(y_1 y_3 + y_2 y_4)r^{-5}), & g_2(t_f) &= 0, \quad (6.28) \\
\gamma'_1 &= -2g_1 - (3y_1^2 r^{-5} - r^{-3})\gamma_3 - 3y_1 y_2 r^{-5} \gamma_4, & \gamma_1(t_f) &= 0, \\
\gamma'_2 &= -2g_2 - 3y_1 y_2 r^{-5} \gamma_3 - (3y_2^2 r^{-5} - r^{-3})\gamma_4, & \gamma_2(t_f) &= 0, \\
\gamma'_3 &= -2g_3 - \gamma_1, & \gamma_3(t_f) &= 0, \\
\gamma'_4 &= -2g_4 - \gamma_2, & \gamma_4(t_f) &= 0, \\
\text{where } h &= (\psi_\sigma(y_1 r^{-6} \gamma_1 + y_2 r^{-6} \gamma_2 - \frac{1}{2} y_3 \gamma_3 - \frac{1}{2} y_4 \gamma_4))^{-1/2}
\end{aligned}$$

The current example differs from the previous one in that we do not have an analytic expression for the solution of the differential equation (6.25). Hence, we first need to solve (6.25) numerically. We use the DOP853 code for this task. This routine, described by Hairer, Nørsett and Wanner [44], implements an explicit eighth-order Runge–Kutta method due to Dormand and Prince [79]. We run this code with the stringent tolerance requirement of 10^{-10} , and we use its dense output routine to sample the solution with a frequency of 1000. These samples are stored in a table.

We can now use COLNEW to solve the boundary value problem (6.28). Whenever the solution of the Kepler problem (6.25) at a certain time is required, we look it up in the table of samples constructed before, using cubic interpolation if necessary. We track the satellite over three revolutions, so we take $t_f = 3T$. As in the previous example, we set $\sigma = 10^{-3}$, and we use as initial guess $g(t) \equiv 0$ and $\gamma(t) = \ell(t, y(t))$. Finally, the tolerance level for COLNEW is set to a modest 10^{-3} , which is quite enough for graphical purposes. The resulting solution is plotted in Figure 6.3.

We see that the step size in one period is different from the step size at the corresponding time in another period. In fact, 54% of the steps are taken while traversing the first revolution around the origin. The corresponding percentages for the second and third revolution are 37% and 9% respectively. The reason is the same as in Example 6.7: errors committed at the beginning of the integration interval contaminate the numerical solution over a longer time interval.

Figure 6.3 shows that, apart from the increasing trend, the step size decreases around $t = \frac{1}{2}T$, $t = 1\frac{1}{2}T$, and $t = 2\frac{1}{2}T$. These times correspond to the left-most point of the ellipse in Figure 6.2, where the satellite is closest to the origin. Equation (6.27) shows that the local error is large when r is small. It is thus not surprising that the optimal strategy is to take smaller steps when the satellite is close to the origin.

We next consider a variant of the Euler method (2.11), given by

$$y_{k+1} = y_k + h_k f(t_k, (y_k)_1, (y_k)_2, (y_{k+1})_3, (y_{k+1})_4), \quad (6.29)$$

where $(y_k)_i$ denotes the i th component of the four-dimensional vector y_k , and f refers to the right-hand side of the Kepler equation (6.25). This method is called the *symplectic Euler* method. The reason for this name is that (6.29) is very similar to the standard Euler method, but it possesses a property which the Euler method misses, namely symplecticity. A method is said to be *symplectic* if the time-stepping map Ψ_h satisfies $(D\Psi_h)^{-1} J D\Psi_h = J$ with $J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$, where I denotes the identity matrix of dimension $\frac{1}{2}d$, whenever the method is applied to a Hamiltonian differential equation, i.e., an equation of the form $y' = J^{-1}\nabla H(y)$ with $H : \mathbf{R}^d \rightarrow \mathbf{R}$. The Kepler equation (6.25) is Hamiltonian with

$$H(y) = -\frac{1}{\sqrt{y_1^2 + y_2^2}} + \frac{1}{2}y_3^2 + \frac{1}{2}y_4^2.$$

The flow Φ of a Hamiltonian equation satisfies $(D\Phi)^{-1} J D\Phi = J$, so it may be expected that symplectic methods perform well when applied to Hamiltonian equations. Indeed, as was mentioned in Section 4.1, the global error of symplectic

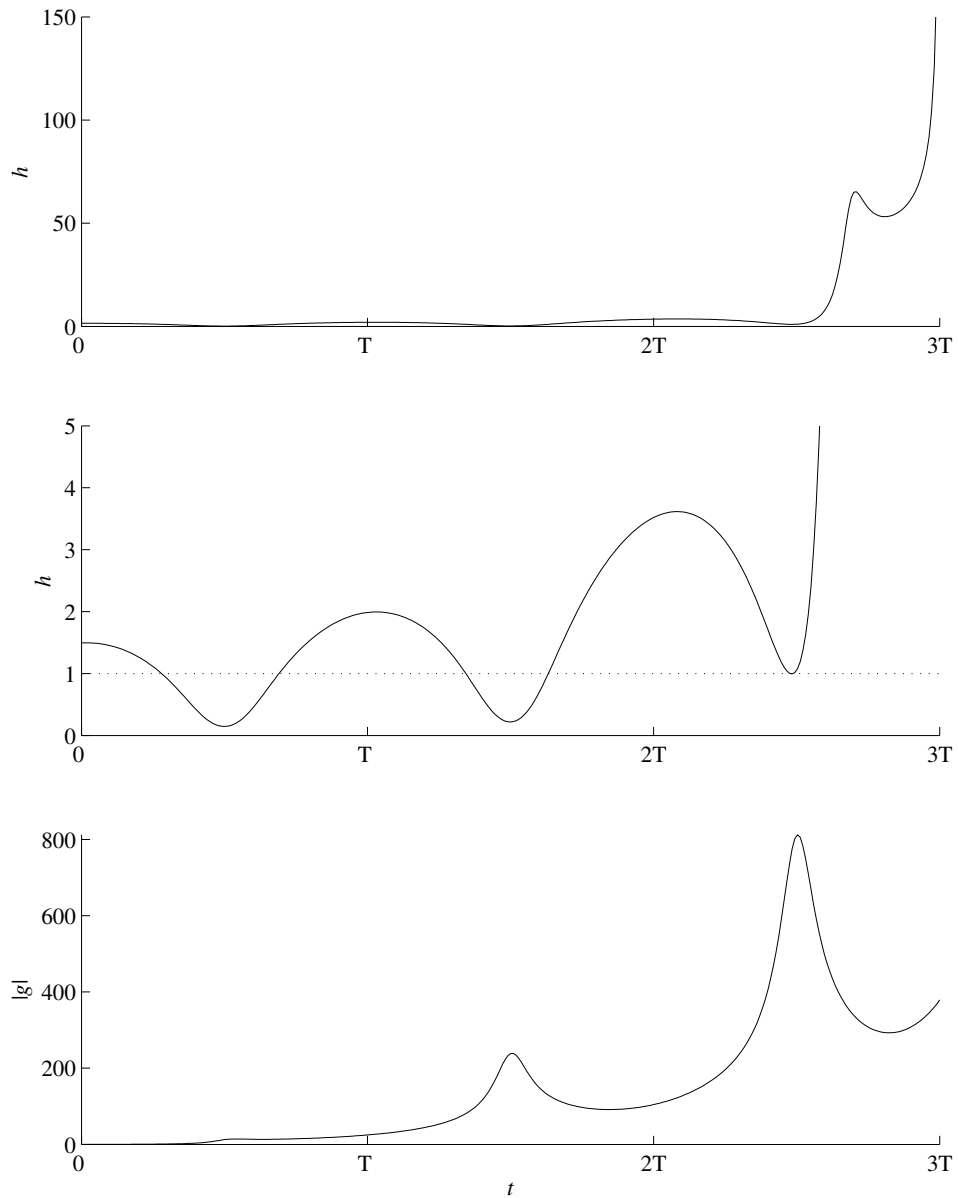


Figure 6.3: The top picture shows the optimal step size (that is, the step size which minimizes the L_2 norm of the global error) when solving the Kepler problem (6.25) with the standard Euler method (2.11). The middle picture is identical, except that a different vertical scaling is used. The bottom picture shows the norm of the global error when the optimal step size is used.

methods accumulates more slowly than the error of other methods. More information can be found in [20, 43, 67] to which the text of §4.1 refers.

The local error of the symplectic Euler method, when applied to the Kepler problem, is given by

$$\ell(t, y) = \frac{1}{2} \begin{bmatrix} -y_1 r^{-3} \\ -y_2 r^{-3} \\ y_3 r^{-3} - 3y_1(y_1 y_3 + y_2 y_4) r^{-5} \\ y_4 r^{-3} - 3y_2(y_1 y_3 + y_2 y_4) r^{-5} \end{bmatrix}. \quad (6.30)$$

If we compare this with the local error of the standard Euler method, cf. (6.27), we find that it differs very little: only the sign of the first two components of $\ell(t, y)$ is reversed.

The optimal step size for the symplectic Euler method can be determined in the same way as we did before for the standard Euler method. The results are shown in Figure 6.4. The difference with Figure 6.3 is great, even though the local errors of the standard and the symplectic Euler method are almost the same, showing the importance of using a symplectic method. The global error committed by the symplectic Euler method is far smaller, and the optimal step size function does not vary as wildly. Other features do not change: the step size still shows an increasing trend, and it drops when the satellite approaches the origin.

We return to the Kepler problem in Example 7.8 in the next Chapter. There we compare the step size strategy which we discussed here, with other step size strategies (see Figures 7.4 and 7.6). \diamond

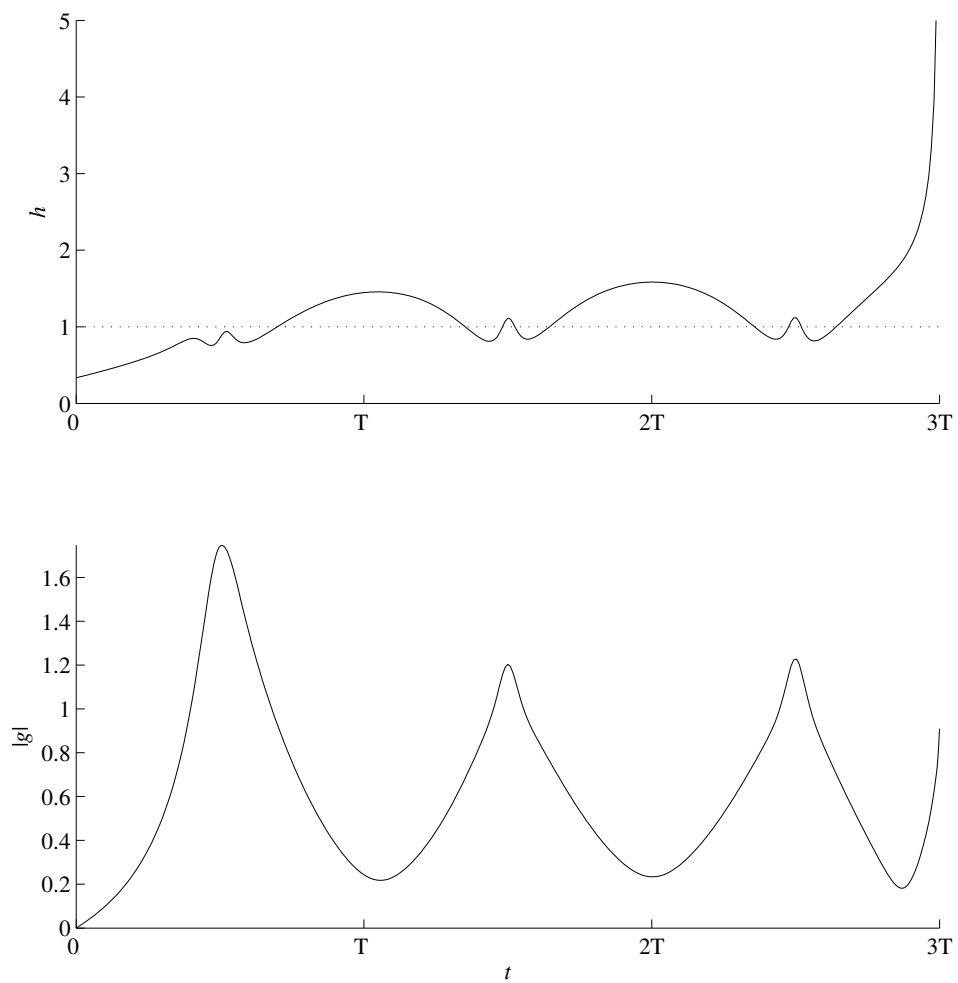


Figure 6.4: The optimal step size and the corresponding error when solving the Kepler problem with the *symplectic* Euler method (6.29).

Chapter 7

Minimizing the maximal error

In the previous chapter, we minimized the global error as measured in the L_2 norm. In this chapter, we consider the L_∞ norm. In other words, we want to find the step size function h for which the maximum of the norm of the global error over the whole time interval is minimized. This leads to the following optimal control problem,

$$\begin{aligned} & \underset{h}{\text{minimize}} \max_{t \in [t_0, t_f]} \|g(t)\|^2 \quad \text{subject to} \quad \int_{t_0}^{t_f} \frac{1}{h(t)} dt = t_f - t_0 \\ & \text{where} \quad g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned} \quad (7.1)$$

Note that we use the Euclidean norm for $g(t) \in \mathbf{R}^d$ and the L_∞ norm for the function g .

The theory of Section 6.1 is not immediately applicable to this problem because of the form of the objective function. This problem may be resolved by considering the L_∞ norm as the limit of the L_s norm as $s \rightarrow \infty$. So, we could consider (7.1) as the limit of the problem (6.17) as $s \rightarrow \infty$. The results of Baron and Ishii [8] imply that the minimal value of the objective function $\|g\|_s$ of (6.17) converges to the minimal value of the objective function $\|g\|_\infty$ of (7.1). However, there is no guarantee that the optimal pair (g, h) of (6.17) converges to an optimal pair of (7.1). Another problem is that it is not clear how to interpret the boundary value problem (6.18) associated with (6.17) in the limit $s \rightarrow \infty$.

Here, we take a different approach to avoid these problems. Following Lindberg [61], we replace (7.1) by the problem

$$\begin{aligned} & \underset{h}{\text{minimize}} \int_{t_0}^{t_f} \frac{1}{h(t)} dt \quad \text{subject to} \quad \|g(t)\| \leq 1 \text{ for all } t \in [t_0, t_f] \\ & \text{where} \quad g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned} \quad (7.2)$$

In words, instead of minimizing the maximal error using a fixed number of steps, we minimize the number of steps subject to a bound on the global error. A moment of reflection shows that the optimal pairs (g, h) of (7.1) and (7.2) coincide up to a rescaling (recall that the differential equation $g' = \frac{\partial f}{\partial y}g + h^p \ell$ is invariant under the scaling $(g, h) \mapsto (\alpha^p g, \alpha h)$).

The theory of Section 6.1 cannot be applied to the optimal control problem (7.2) either, even though the objective function has the right form, because of the constraint $\|g(t)\| \leq 1$. Fortunately, the theory can be extended to take this constraint into account. This generalization is described in Section 7.1. Again, we find that we can convert the optimal control problem to an equivalent boundary value problem. In parallel with the previous chapter, Section 7.2 describes the analytic solution of this boundary value problem, and Section 7.3 its numerical solution. The examples in this chapter are also the same as those considered in the previous chapter, namely the equation $y' = ky$ and the Kepler problem.

7.1 State-constrained optimal control problems

In Section 6.1, we studied optimal control problems of the form

$$\begin{aligned} & \underset{u \in \mathcal{U}}{\text{minimize}} \int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) \, dt \quad \text{subject to} \quad x(t_f) \in \mathcal{T} \\ & \text{where} \quad x'(t) = \hat{f}(t, x(t), u(t)), \quad x(t_0) = x_0, \end{aligned} \quad (6.2)$$

where the set \mathcal{U} of admissible controls is of the form

$$\mathcal{U} = \{u \mid u \text{ is measurable and } u(t) \in \hat{\mathcal{U}} \text{ for a.a. } t\}, \quad (6.4)$$

for some set $\hat{\mathcal{U}}$. The set $\hat{\mathcal{U}}$ allows us to place constraints on the control variable. However, the problem (7.2) being studied in this chapter includes a constraint on the global error g , which plays the role of a state variable. So we need to consider a generalization of (6.2) that includes state constraints. Specifically, we study optimal control problems of the form

$$\begin{aligned} & \underset{u}{\text{minimize}} \int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) \, dt \\ & \text{subject to} \quad b(t, x(t)) \geq 0 \quad \text{for all } t \in [t_0, t_f] \\ & \text{where} \quad x'(t) = \hat{f}(t, x(t), u(t)), \quad x(t_0) = x_0. \end{aligned} \quad (7.3)$$

As in the previous chapter, we assume that the functions \hat{f}_0 and \hat{f} are continuously differentiable with respect to all their arguments. Furthermore, the function b defining the constraint is assumed to be C^2 . A pair (x, u) is called *admissible* if x is

an absolutely continuous function, u is a measurable function, and the inequality constraint and differential equation in (7.3) are satisfied. An *optimal pair* is a pair which minimizes the objective function over all admissible pairs.

The problem (7.2) is indeed of this form, with g acting as the state variable x and h as the control variable u , while $b(t, x) = 1 - \|x\|^2$ (the square is needed to ensure smoothness at $x = 0$).

Obviously, the problem (7.3) can be generalized further. We will not discuss this here, as it is not needed to solve the problem (7.2). Instead, we refer to the excellent survey by Hartl, Sethi, and Vickson [47], which mentions the forthcoming Theorems 7.1–7.4 in a more general setting, and also provides the interested reader with additional background and references in state-constrained optimal control theory.

For any time $t \in [t_0, t_f]$, the inequality $b(t, x(t)) \geq 0$ describes a closed subset of state space. A state $x(t)$ satisfying the inequality can lie either in the interior or on the boundary of this subset. This leads to the following definitions. Fix a trajectory $x : [t_0, t_f] \rightarrow \mathbf{R}^d$. We call a subinterval $(t_1, t_2) \subset [t_0, t_f]$ an *interior interval* if $b(t, x(t)) > 0$ for all $t \in (t_1, t_2)$. Similarly, a subinterval $[t_1, t_2] \subset [t_0, t_f]$ is a *boundary interval* if $b(t, x(t)) = 0$ for all $t \in [t_1, t_2]$. An instant t_* is called an *entry time* if there is an interior interval ending at $t = t_*$ and a boundary interval starting at $t = t_*$. The term *exit time* describes the opposite situation, where the state exits the boundary at $t = t_*$. If $x(t)$ is in the interior for all t in a neighbourhood of a certain instant t_* , while $x(t_*)$ lies on the boundary, then $t = t_*$ is called a *contact time* (in this case, the trajectory only touches the boundary). The term *junction time* is used to refer to entry times, exit times, and contact times taken together.

Given an optimal control problem without state constraints, Theorem 6.1 guarantees the existence of a solution to this problem under certain conditions. A similar existence result holds for problems with state constraints.

Theorem 7.1 (Filippov–Cesari Theorem). *Consider the optimal control problem (7.3). If there exists at least one admissible pair, if for certain constants x_{max} and u_{max} all admissible pairs (x, u) satisfy the bounds $\|x(t)\| \leq x_{max}$ and $\|u(t)\| \leq u_{max}$ for all t , and if the set*

$$\{(y_0, y) \mid y_0 \geq \hat{f}_0(t, x, u) \text{ and } y = \hat{f}(t, x, u) \text{ for some admissible } u\}$$

is convex for all (t, x) , then there exists an optimal pair (x, u) for (7.3).

A proof of this theorem is given by Cesari [22].

Let us now turn to the Pontryagin Minimum Principle (Theorem 6.2), which gives a necessary condition for optimality. An analogous result for the state-constrained optimal control problem (7.3) is as follows.

Theorem 7.2. *Suppose that (x, u) is an optimal pair for (7.3). Assume that u is right-continuous with left-hand limits and that x has only finitely many junction times. Then there exist a constant $\lambda_0 \geq 0$, a piecewise absolutely continuous function $\lambda : [t_0, t_f] \rightarrow \mathbf{R}^d$, a piecewise continuous function $\mu : [t_0, t_f] \rightarrow \mathbf{R}$, a number η_f , and a number $\eta(\tau_i)$ for every point τ_i of discontinuity of λ , such that $(\lambda_0, \lambda(t), \mu(t), \eta_f, \eta(\tau_1), \eta(\tau_2), \dots) \neq 0$ for all t and*

$$\frac{d\lambda}{dt} = -\frac{\partial H}{\partial x} - \mu \frac{\partial b}{\partial x} \text{ for a.a. } t, \quad \text{and} \quad (7.4)$$

$$\mu(t) \geq 0 \text{ and } \mu(t) b(t, x(t)) = 0 \text{ for a.a. } t, \quad \text{and} \quad (7.5)$$

$$H(t, x(t), u(t), \lambda_0, \lambda(t)) = \min_v H(t, x(t), v, \lambda_0, \lambda(t)) \text{ for a.a. } t, \quad (7.6)$$

where the Hamiltonian H is defined by $H = \lambda_0 \hat{f}_0 + \lambda^\top \hat{f}$. Furthermore, at any junction time τ , the dual variable λ might be discontinuous, in which case the following jump condition is satisfied

$$\lim_{t \downarrow \tau} \lambda(t) - \lim_{t \uparrow \tau} \lambda(t) = -\eta(\tau) \frac{\partial b}{\partial x} \quad \text{with} \quad \eta(\tau) \geq 0. \quad (7.7)$$

Finally, at the terminal time t_f , we have

$$\lambda(t_f) = \eta_f \frac{\partial b}{\partial x} \quad \text{with} \quad \eta_f \geq 0 \quad \text{and} \quad \eta_f b(t_f, x(t_f)) = 0. \quad (7.8)$$

According to [47], the above theorem has been proved (in a more general setting) by Maurer [63].

We see that the minimum principle (6.7) from the last chapter, where we studied optimal control problem *without* state constraints, carries over unchanged when a state constraint is added. However, the state constraint does lead to an additional multiplier, namely μ , which influences the dynamics of the costate λ via (7.4). Condition (7.5) is a complementarity condition, which says that either the state constraint is satisfied strictly and the multiplier μ vanishes, or the state is on the boundary of the allowed region, in which case μ may take any positive value.

A further complication is that the costate may exhibit jumps, cf. (7.7). The next theorem rules out this possibility in certain cases. As in Section 6.1, we call the Hamiltonian H *regular* if the minimizer $v \in \hat{\mathcal{U}}$ in (7.6) is unique. The state constraint $b(t, x(t)) \geq 0$ is said to be of *first order* if

$$\frac{\partial b}{\partial x}(t, x) \frac{\partial \hat{f}}{\partial u}(t, x, u) \neq 0. \quad (7.9)$$

This condition is equivalent to requiring that the time derivative of $b(t, x(t))$ depends on the control u .

Theorem 7.3. *Suppose that (x, u) is an optimal pair for (7.3). If the Hamiltonian H is regular, then x is C^1 and u is continuous. If, furthermore, the state constraint is of first order at a junction time τ , then the costate λ as defined in Theorem 7.2 is continuous at τ and $\eta(\tau) = 0$.*

This result is due to Jacobson, Lele and Speyer [56, §5.2 and §6].

Contrary to the case where no state constraints are present (cf. Theorem 6.3), the costate λ need not be differentiable, even if the Hamiltonian is regular, because the multiplier μ is not continuous in general. In fact, the next section contains an example in which μ is not continuous and λ is not differentiable (see (7.21) and Figure 7.1).

In Section 7.3, where we are seeking a numerical solution, we follow a different approach. We introduce a penalty parameter $\nu > 0$ and replace the problem (7.3) by the unconstrained optimal control problem

$$\begin{aligned} & \underset{u}{\text{minimize}} \int_{t_0}^{t_f} \hat{f}_0(t, x(t), u(t)) + \nu(b(t, x(t)))_+ dt \\ & \text{where } x'(t) = \hat{f}(t, x(t), u(t)), \quad x(t_0) = x_0, \end{aligned} \quad (7.10)$$

where the notation $(\cdot)_+$ is defined by $(a)_+ = 0$ if $a \leq 0$ and $(a)_+ = a$ if $a \geq 0$. This approach is called the *exterior penalty* approach: those parts of the trajectory that lie outside the region $\{x \in \mathbf{R}^d : b(x, t) \geq 0\}$ are penalized by the extra term in the objective function.

The parameter ν determines the weight of this penalty term. As ν grows larger, violations of the state constraint are penalized more heavily. Intuitively speaking, any violation is penalized infinitely heavily in the limit $\nu \rightarrow \infty$, so in this limit, the optimal pair (x, u) will not violate the constraint $b(t, x) \geq 0$ and thus also be a solution to the state-constrained problem (7.3). This idea is substantiated by the following theorem, which is due to Okamura [75].

Theorem 7.4. *Let $\{\nu_k\}$ be an unbounded increasing sequence of real numbers. Suppose that for every k , the unconstrained problem (7.10) with $\nu = \nu_k$ has a piecewise continuous optimal control u_k . If the sequence $\{u_k\}$ converges in the L_1 norm, then the limit is an optimal control for the constrained problem (7.3).*

We conclude this section by repeating that more information on state-constrained optimal control theory can be found in the survey article by Hartl, Sethi, and Vickson [47] and references therein.

7.2 Analytic treatment

We return to the problem of choosing the step size function of a numerical integrator in such a way that the maximal global error is as small as possible. In the introduction to this chapter, we formulated this as the minimax problem (7.1), which is equivalent to the state-constrained optimal control problem (7.2). In the previous section, we briefly described the theory on problems of this kind. Now, we apply this theory on the problem (7.2). This gives a characterization of the optimal step size function (see Theorem 7.5). We then apply this result to the differential equation $y' = ky$. In contrast to the L_2 case (cf. Example 6.6), we can solve the boundary value problem. The result is that the optimal step size is given by (7.21).

We recall that the problem (7.2) is of the form (7.3), if we identify the error g as the state variable x and the step size h as the control u . The state constraint is $1 - \|g\|^2 \geq 0$.

As in the L_2 case, treated in the Chapter 6, we cannot guarantee the existence of an optimal step size function. Theorem 7.1 does not apply, as we do not have *a priori* bounds on the optimal pair. The usual generalizations of this existence result do not apply either. However, we will see in Examples 7.6 and 7.8 that the optimal step size function for the Dahlquist test equation and the Kepler problem are bounded, in contrast to the L_2 case where the optimal step size is unbounded as $t \rightarrow t_f$ (as remarked immediately after Theorem 6.5). So, it might still be possible that some proof of an *a priori* bound on the optimal step size function will be found, thus establishing the existence of a solution of (7.2).

Next, we turn to the minimum principle, stated in Theorem 7.2. We again denote the adjoint variable of g by γ . The Hamiltonian is given by

$$H(t, g, h, \lambda_0, \gamma) = \frac{\lambda_0}{h} + \gamma^\top \left(\frac{\partial f}{\partial y}(t, y(t)) g + h^p \ell(t, y(t)) \right). \quad (7.11)$$

Condition (7.6) says that the optimal control minimizes the above Hamiltonian. If $\gamma^\top \ell$ is either negative or zero, then H does not have a minimum, and condition (7.6) cannot be satisfied. Thus, $\gamma^\top \ell$ has to be positive, in which case the minimum is attained at

$$h = \left(\frac{\lambda_0}{p \gamma^\top \ell} \right)^{1/(p+1)}.$$

We write this as $h = \kappa_0 (\gamma^\top \ell)^{-1/(p+1)}$ with $\kappa_0 = (\lambda_0/p)^{1/(p+1)}$. This minimum is unique, so the Hamiltonian (7.11) is regular and thus Theorem 7.3 implies that the optimal control h is continuous. Furthermore, the costate λ is continuous at a junction time τ if the state constraint is first order which is the case if $g(\tau)^\top \ell(\tau, y(\tau)) \neq 0$, cf. (7.9).

The dynamics of the costate is given by (7.4), which for the problem under consideration reads

$$\gamma'(t) = -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) + 2\mu(t)g(t).$$

The multiplier $\mu(t)$, introduced in the above equation, is nonnegative and it vanishes whenever the strict inequality $|g(t)| < 1$ holds, cf. (7.5).

The terminal condition (7.8) reads

$$\gamma(t_f) = -2\eta_f g(t_f) \quad \text{with} \quad \eta_f \geq 0 \quad \text{and} \quad \eta_f(1 - |g(t_f)|^2) = 0.$$

The complementarity condition means that either $\eta_f = 0$ or $1 - |g(t_f)|^2 = 0$. However, we can prove that the latter equality holds. Indeed, suppose that (g, h) is an optimal pair. If $|g(t_f)| < 1$, then, by continuity, there exists an $\varepsilon > 0$ such that $|g(t)| < 1$ for all $t \in [t_f - \varepsilon, t_f]$. Let $\delta > 0$ be sufficiently small. We define a new step size function \hat{h} by

$$\hat{h}(t) = \begin{cases} h(t), & \text{if } t \leq t_f - \varepsilon, \\ h(t) + \delta, & \text{if } t > t_f - \varepsilon, \end{cases}$$

and denote the corresponding global error by \hat{g} . We have $\hat{g} = g$ for $t < t_f - \varepsilon$. Furthermore, the constraint $|\hat{g}(t)| \leq 1$ will still be satisfied inside $[t_f - \varepsilon, t_f]$ if δ is small enough. Thus, the constraint $|\hat{g}(t)| \leq 1$ is satisfied for all $t \in [t_0, t_f]$, so (\hat{g}, \hat{h}) is an admissible pair. However, since δ is positive, we have

$$\int_{t_0}^{t_f} \frac{1}{\hat{h}(t)} dt < \int_{t_0}^{t_f} \frac{1}{h(t)} dt.$$

This is in contradiction with the optimality of (g, h) . We conclude that an optimal pair necessarily satisfies $|g(t_f)| = 1$. This condition automatically implies that $\eta_f(1 - |g(t_f)|^2) = 0$.

We summarize the above discussion in the following theorem. The reader may want to compare it to the analogous result for the L_2 case, Theorem 6.5.

Theorem 7.5. *Consider the optimal control problem (7.1), and let (g^*, h^*) be an optimal pair, such that $g^*(\tau)^\top \ell(\tau, y(\tau)) \neq 0$ at all junction times τ . Then there exist piecewise absolutely continuous functions $g, \gamma : [t_0, t_f] \rightarrow \mathbf{R}$, which for some value of κ_0 are related to (g^*, h^*) by*

$$g^*(t) = \kappa_0^p g(t) \quad \text{and} \quad h^*(t) = \kappa_0 (\gamma(t)^\top \ell(t, y(t)))^{-1/(p+1)}. \quad (7.12)$$

The functions g and γ satisfy the following differential equation

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t))g(t) + (\gamma(t)^\top \ell(t, y(t)))^{-p/(p+1)} \ell(t, y(t)), \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - 2\mu(t)g(t), \end{aligned} \quad (7.13)$$

with boundary conditions

$$g(t_0) = 0, \quad |g(t_f)| = 1, \quad \text{and} \quad \gamma(t_f) = -2\eta_f g(t_f). \quad (7.14)$$

where the multiplier η_f is nonnegative, and $\mu : [t_0, t_f] \rightarrow \mathbf{R}$ is a piecewise continuous function satisfying the complementarity condition

$$\text{either} \quad \left\{ \begin{array}{l} \mu(t) = 0 \\ |g(t)| \leq 1 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} \mu(t) \geq 0 \\ |g(t)| = 1 \end{array} \right\}. \quad (7.15)$$

Furthermore, the following condition is satisfied, ensuring that the fractional power in (7.13) is well-defined,

$$\gamma(t)^\top \ell(t, y(t)) > 0 \text{ for all } t < t_f. \quad (7.16)$$

Lindberg [61] obtains very similar results in the specific case that the local error is proportional to $y^{(p+1)}(t)$ and the underlying differential equation is either linear and autonomous (i.e., $f(t, y) = Ay$ for some matrix A) or scalar (i.e., $d = 1$).

As in the previous chapter, we can interpret (7.13) as requiring the equidistribution of the weighted local error

$$(h^*(t))^{p+1} \gamma(t)^\top \ell(t, y(t)).$$

Example 7.6. In Examples 6.6 and 6.7, we considered the differential equation $y' = ky$ with $k \in \mathbf{R} \setminus \{0\}$ and initial condition $y(0) = 1$. We solved the boundary value problem of Theorem 6.5, which determines the step size function that minimizes the L_2 norm of the global error. In this example, we apply Theorem 7.5 and determine the step size function that minimizes the maximal global error.

The exact solution is given by $y(t) = e^{kt}$. If we use a Runge–Kutta method to solve the equation, then the leading local error term is proportional to $h^{p+1}y$. The boundary value problem (7.13), (7.14) has the scaling symmetry

$$(\ell, g, \gamma, \mu, \eta_f) \mapsto (\alpha^p \ell, g, \alpha \gamma, \alpha \mu, \alpha \eta_f). \quad (7.17)$$

So, we can assume without loss of generality that the constant of proportionality is one. This implies that $\ell(t, y(t)) = y(t)$, which is positive, and thus, by (7.16), we have $\gamma(t) > 0$ for all $t < t_f$. Hence, the boundary value problem (7.13), (7.14) reduces to

$$\begin{aligned} g'(t) &= kg(t) + e^{kt/(p+1)} (\gamma(t))^{-p/(p+1)}, & g(0) &= 0, \quad |g(t_f)| = 1, \\ \gamma'(t) &= -k\gamma(t) - 2\mu(t)g(t), & \gamma(t_f) &= 0. \end{aligned} \quad (7.18)$$

Solving the differential equation for g yields

$$g(t) = \int_0^t \exp\left(kt - \frac{kps}{p+1}\right) (\gamma(s))^{-p/(p+1)} ds. \quad (7.19)$$

The integrand is positive, so $g(t) > 0$ for all $t > t_0$. Hence, $g(t)\ell(t, y(t))$ is positive, so the state constraint is of first order and the global error g is C^1 by Theorem 7.3. Furthermore, the terminal condition $|g(t_f)| = 1$ is equivalent to $g(t_f) = 1$.

Let t_* denote the first instant t at which $g(t) = 1$. Then $(0, t_*)$ is an interior interval, so μ vanishes on this interval by the complementarity condition (7.15). Hence, the costate γ satisfies $\gamma' = -k\gamma$ on $(0, t_*)$, so $\gamma(t) = \gamma_0 e^{-kt}$ for $t \in [0, t_*]$ where $\gamma_0 = \gamma(0)$. Substituting this in (7.19) reveals that

$$g(t) = \gamma_0^{-p/(p+1)} t e^{kt} \quad \text{for } t \in [0, t_*]. \quad (7.20)$$

This function is increasing if $kt > -1$ and decreasing if $kt < -1$. Furthermore, the condition $g(t_*) = 1$ implies that $\gamma_0 = (t_* e^{kt_*})^{(p+1)/p}$.

Suppose that $t_* = t_f$. Then $g(t) < 1$ for $t < t_f$ and $g(t_f) = 1$, so $g'(t_f) \geq 0$, which implies that $kt_f \geq -1$. If, on the other hand, t_* is strictly smaller than t_f , then $g'(t_*) = 0$ since g is a continuously differentiable function having a maximum at t_* , and hence $kt_* = -1$. We conclude that $t_* = t_f$ if k is positive or $t_f \leq -1/k$ and that $t_* = -1/k$ otherwise.

In the case $t_* = t_f$, equation (7.20) is valid on the whole interval $[0, t_f]$ and we have constructed a solution to the boundary value problem of Theorem 7.5.

Now assume that $t_* = -1/k < t_f$. Suppose that the interval $[t_*, t_f]$ contains an interior interval, say $[t_1, t_2]$. We can now reason as above. Inside this interval, the multiplier μ vanishes. Hence γ satisfies $\gamma' = -k\gamma$, therefore we have $\gamma(t) = C_1 e^{-kt}$ for some C_1 , and finally $g(t) = C_2 t e^{kt}$ for some positive constant C_2 . However, we already found out that this is a decreasing function on $[t_*, t_f]$, so $g(t_2) < g(t_1) = 1$. If $t_2 < t_f$, then the interior interval $[t_1, t_2]$ is followed by a boundary interval, but this would require that $g(t_2) = 1$. On the other hand, if $t_2 = t_f$, then $g(t_2) < 1$ is in contradiction with the terminal condition $g(t_f) = 1$. We conclude that $[t_*, t_f]$ contains no interior interval, so $g(t) = 1$ for all $t \in [t_*, t_f]$. It now follows from (7.18) that $\gamma(t) = (-k)^{-(p+1)/p} e^{kt/p}$ and $\mu(t) = \frac{p+1}{p} \left(-\frac{1}{k} e^{kt}\right)^{1/p}$ for $t \in (t_*, t_f]$. Note that k is negative in this case, so the fractional power is well-defined and the multiplier μ is indeed positive.

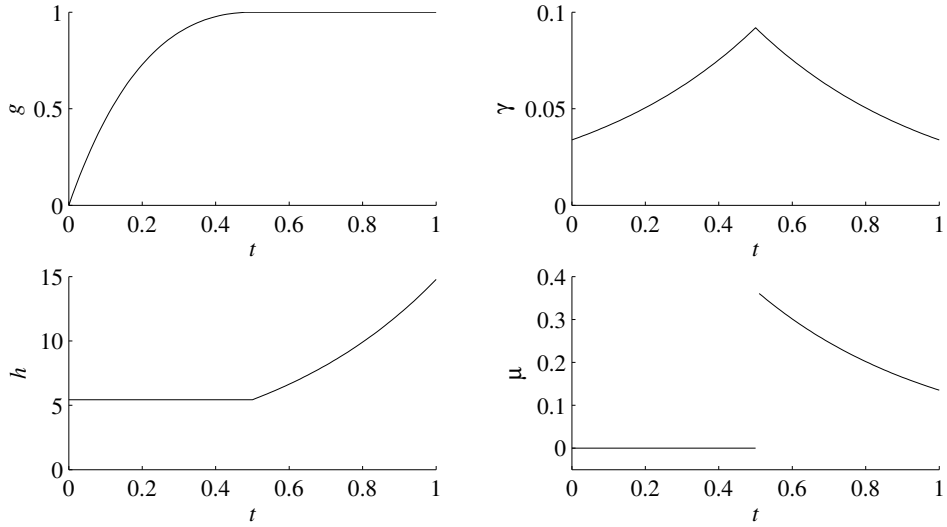


Figure 7.1: The functions (7.21) for $k = -2$, $t_f = 1$, and $p = 1$.

Summarizing, the unique solution of (7.13)–(7.16) is

$$\begin{aligned}
 g(t) &= \begin{cases} \frac{t}{t_*} e^{k(t-t_*)}, & \text{if } t \leq t_*, \\ 1, & \text{if } t > t_*, \end{cases} \\
 \gamma(t) &= \begin{cases} (t_* e^{kt_*})^{(p+1)/p} e^{-kt}, & \text{if } t \leq t_*, \\ t_*^{(p+1)/p} e^{kt/p}, & \text{if } t > t_*, \end{cases} \\
 h(t) &= \begin{cases} (t_* e^{kt_*})^{-1/p}, & \text{if } t \leq t_*, \\ (t_* e^{kt})^{-1/p}, & \text{if } t > t_*, \end{cases} \\
 \mu(t) &= \begin{cases} 0, & \text{if } t \leq t_*, \\ \frac{p+1}{p} (t_* e^{kt})^{1/p}, & \text{if } t > t_*, \end{cases} \\
 \text{where } t_* &= \begin{cases} t_f, & \text{if } k > 0 \text{ or } k < -1/t_f, \\ -1/k, & \text{otherwise.} \end{cases}
 \end{aligned} \tag{7.21}$$

The same solution can be derived from the results of Lindberg [61].

We show the above solution with $k = -2$, $t_f = 1$, and $p = 1$ in Figure 7.1. Note that the global error g is continuously differentiable, while both γ and h are only continuous and the multiplier μ makes a jump at $t = \frac{1}{2}$. This is in agreement with Theorem 7.3.

In Example 7.7 in the next section, we compare the above strategy with the strategy found in the previous chapter. \diamond

Of course, not all equations are as simple as $y' = ky$, and the boundary value problem of Theorem 7.5 cannot always be solved analytically. The alternative is to solve it numerically. This is the subject of the next section.

7.3 Numerical treatment

In this section, we want to find a numerical solution to the problem (7.1), or the equivalent problem (7.2), which asks for the step size function that minimizes the maximal error. This complements Section 7.2, where we solve the problem analytically, but also Section 6.3, where the objective is to minimize the L_2 norm of the global error over the interval $[t_0, t_f]$.

Oberle and Grimm [73, 74], Pesch [78], and Maurer, Büskens and Feichtinger [64] solve state-constrained optimal control problems by converting them to a boundary value problem with Theorem 7.2. In our case, the boundary value problem is given by (7.13)–(7.16). This problem is then solved by multiple shooting. The obvious difficulty in this approach is the implementation of the complementarity condition (7.15). The authors cited at the beginning of this paragraph all assume that the number and type of junction times in the interval $[t_0, t_f]$ is known and that only the precise instant at which the solution shifts from one alternative in (7.15) to the other, needs to be determined. Unfortunately, this knowledge is not available for the problem under consideration. Indeed, Example 7.8 shows that the number of interior intervals depends subtly on the numerical method being used to solve the differential equation. Therefore, we have to find an alternative algorithm for solving the boundary value problem.

Consider the exterior penalty approach introduced at the end of Section 7.1. This approach replaces the problem (7.2) by

$$\begin{aligned} & \underset{h}{\text{minimize}} \int_{t_0}^{t_f} \frac{1}{h(t)} + \nu(\|g(t)\|^2 - 1)_+ \, dt \\ & \text{where } g'(t) = \frac{\partial f}{\partial y}(t, y(t)) g(t) + h(t)^p \ell(t, y(t)), \quad g(t_0) = 0. \end{aligned} \tag{7.22}$$

Recall that the notation $(\cdot)_+$ is defined by $(x)_+ = 0$ if $x \leq 0$ and $(x)_+ = x$ if $x \geq 0$. The optimal control problem (7.22) does not contain any state constraints, just like the problem (6.1) being considered in the previous chapter, so we can use the same method as in the previous chapter to solve (7.22). We apply Theorem 6.2 to convert (7.22) to a boundary value problem. The Hamiltonian is

$$\lambda_0 \left(\frac{1}{h} + \nu(\|g\|^2 - 1)_+ \right) + \gamma^\top \left(\frac{\partial f}{\partial y}(t, y(t)) g + h^p \ell(t, y(t)) \right),$$

where we introduced the costate $\gamma \in \mathbf{R}^d$. The evolution of the costate is given by

$$\gamma'(t) = -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - 2\lambda_0\nu H(|g(t)|^2 - 1) g(t),$$

where $H(\cdot)$ denotes the Heaviside function,¹ defined by

$$H(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0. \end{cases}$$

Assuming that the inner product $\gamma^\top \ell(t, y(t))$ is positive, the Hamiltonian is minimized when

$$h = \left(\frac{\lambda_0}{p\gamma^\top \ell(t, y(t))}\right)^{1/(p+1)},$$

Furthermore, the Hamiltonian is regular, so any optimal control will be continuous by Theorem 6.3. Hence, Theorem 7.4 states that if the problem (7.22) has a sequence of optimal step size functions which converges as $\nu \rightarrow \infty$, then the limit of this sequence solves the original problem (7.2).

Finally, we can use a scaling symmetry to set $\lambda_0 = p$. This results in the following boundary value problem,

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial y}(t, y(t)) g(t) + \left(\psi_\sigma(\gamma(t)^\top \ell(t, y(t)))\right)^{-p/(p+1)} \ell(t, y(t)), & g(t_0) &= 0, \\ \gamma'(t) &= -\left(\frac{\partial f}{\partial y}(t, y(t))\right)^\top \gamma(t) - 2\nu p H(|g(t)|^2 - 1) g(t), & \gamma(t_f) &= 0. \end{aligned} \tag{7.23}$$

As in Section 6.3, we introduce the function ψ_σ , which is defined on page 79, to circumvent any problems with the fractional power when $\gamma^\top \ell$ becomes negative. Note the similarity with the boundary value problem (6.22), which yields the optimal step size function in the L_2 norm.

As in the previous chapter, we use the COLNEW routine to solve the problem (7.23). Ideally, we would like to use a high value for the penalty parameter ν in (7.23), since we want to find the solution as $\nu \rightarrow \infty$. However, COLNEW is unable to solve (7.23) for high values of ν , unless the initial guess for the solution is accurate. This suggests the use of *continuation*: start with a modest value of ν , say $\nu = 1$, use the solution as an initial guess for a slightly larger value of ν , and repeat until the desired value of ν is reached.

Further details are described in the two examples which make up the rest of this chapter. As in Section 6.3, we first study the Dahlquist test equation $y' = ky$, before turning to the Kepler system.

¹Hopefully, the reader will not confuse $H(\cdot)$ denoting the Heaviside function with H denoting the Hamiltonian.

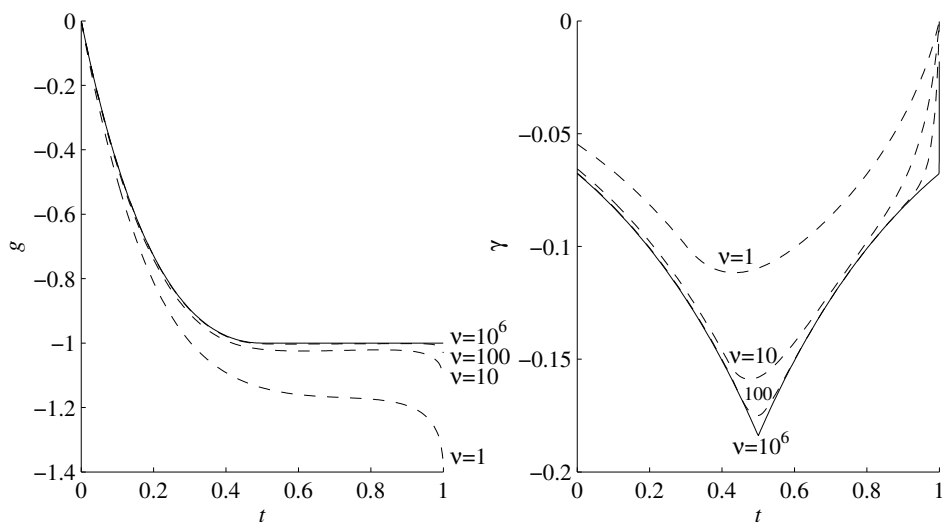


Figure 7.2: The numerical solution of the boundary value problem (7.23) for $\nu = 1$, $\nu = 10$, $\nu = 100$ (dashed lines), and $\nu = 10^6$ (solid line). The values of the other parameters are $k = -2$ and $t_f = 1$.

Example 7.7. In Example 6.6, we tried to find the step size function h which minimizes the L_2 norm of the global error when solving the equation $y' = ky$, $y(0) = 1$, with $k \in \mathbf{R} \setminus \{0\}$ using the Euler method. That attempt was unsuccessful, as we were not able to solve the boundary value problem (6.19) analytically, but we did find a numerical solution in Example 6.7. However, if we replace the L_2 norm by the L_∞ norm, the problem can be solved analytically, as described in Example 7.6. We now complete this series of examples by finding the L_∞ -optimal step size numerically.

We have $y(t) = e^{kt}$, $\ell(t, y) = -\frac{1}{2}k^2y$, $\frac{\partial f}{\partial y} = k$, and $p = 1$, so the boundary value problem (7.23) reads

$$\begin{aligned} g'(t) &= kg(t) - \frac{1}{2}k^2e^{kt} \left(\psi_\sigma \left(-\frac{1}{2}k^2e^{kt}\gamma(t) \right) \right)^{-1/2}, & g(t_0) &= 0, \\ \gamma'(t) &= -k\gamma(t) - 2\nu H(|g(t)|^2 - 1)g(t), & \gamma(t_f) &= 0. \end{aligned} \quad (7.24)$$

We use the same parameters as in the previous examples, namely $k = -2$, $t_f = 1$, and $\sigma = 10^{-3}$. Furthermore, we provide the following initial guess for the solution to COLNEW: $g(t) \equiv 0$ and $\gamma(t) = \ell(t, y(t))$. It turns out that the program is not able to find a solution when $\nu = 10^6$, so we start with $\nu = 1$ and use continuation, multiplying ν by a factor 10 at every step. The result is shown in Figure 7.2.

The plot suggests that the solution converges (in the L_1 norm) as $\nu \rightarrow \infty$. When comparing the limit with the analytic solution (7.21), depicted in Figure 7.1,

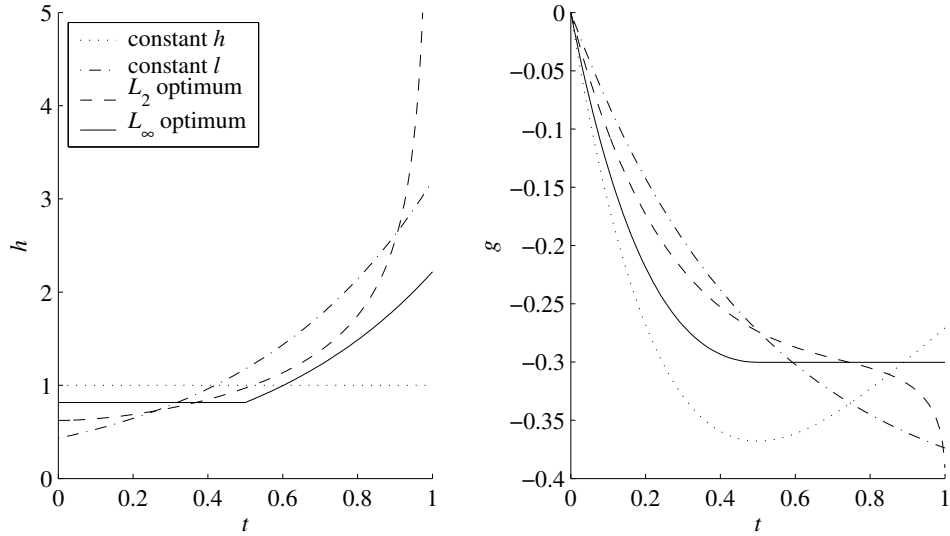


Figure 7.3: Comparison of different step size strategies for solving the equation $y' = -2y$ with the Euler method. The strategies are: constant step size (dotted line), constant local error (dash-dotted line), L_2 -optimal step size (dashed line), and L_∞ -optimal step size (solid line). The left-hand plot shows how the step size varies over time, and the right-hand plot depicts the global error.

it should be kept in mind that the analytic solution was derived under the assumption $\ell(t, y) = y$. However, for the Euler method, we have $\ell(t, y) = -\frac{1}{2}k^2y$. So, we have to apply the scaling symmetry (7.17) when comparing the solutions. Apart from this scaling, the analytic and numerical solution agree. \diamond

To place these results in perspective, we contrast four different step size strategies to solve the equation $y' = ky$ with the Euler method. All strategies use the same number of steps in order to produce a fair comparison. Specifically, we enforce the normalization condition $\int_0^1 \frac{1}{h(t)} dt = 1$. Figure 7.3 shows plots of the step size and the global error against time for all four strategies.

1. The simplest method, which barely qualifies to be called a strategy, is to use the same step size throughout the integration interval. The normalization implies that $h \equiv 1$. Furthermore, the global error satisfies $g' = \frac{\partial f}{\partial y}g + h^p\ell$, cf. (3.18). For the Euler method applied to $y' = ky$, this differential equation reduces to $g' = kg - \frac{1}{2}k^2e^{kt}h$. Solving this equation yields that the global error is $g(t) = -\frac{1}{2}k^2te^{kt}$.
2. Another strategy is to choose the step size so that $L_h(t, y)/h$, the local error per unit step, is kept constant. A lot of numerical software in practical use

is based on this idea. If we neglect higher order terms and use the approximation $L_h(t, y) \approx h^{p+1}\ell(t, y)$, then we find that the step size is proportional to $(\ell(t, y))^{-p}$. In the present situation, we find that $h(t) = Ce^{-kt}$. The normalization condition implies that the proportionality constant C equals $-k/(e^{-kt} - 1)$. Finally, the global error committed by this strategy is given by $g(t) = \frac{1}{2}Ck(1 - e^{kt})$.

3. In the previous chapter, we considered choosing the step size so that $\|g\|_2$, the global error in the L_2 norm, is minimized. The step size that achieves this was calculated in Example 6.7.
4. In this chapter, the maximum global error was minimized. Example 7.6 gives an explicit formula for the resulting step size and global error, namely (7.21).

The step size chosen by these four strategies and the resulting global error are depicted in Figure 7.3. In this figure, the dotted, dash-dotted, dashed, and solid lines correspond to strategies 1, 2, 3, and 4, respectively. The right-hand graph shows that the L_2 -optimal strategy has indeed the smallest error in the L_2 norm, while the L_∞ -optimal strategy has the smallest value of $\max_t |g(t)|$, thus corroborating our computations.

Example 7.8. In the final example in this thesis, we return to the Kepler problem (6.24). This equation was also studied in Example 6.8, where the objective was to minimize $\|g\|_2$ for the standard Euler method (2.11) and the symplectic Euler method (6.29). The results can be found in Figures 6.3 and 6.4.

Here, we seek to find the step size that minimizes the maximal global error by solving the boundary value problem (7.23). We substitute the local error of the standard Euler method for the Kepler problem, which is given by (6.27). As in Example 6.8 in the previous chapter, we need to solve the Kepler problem numerically. We then use COLNEW to solve the resulting boundary value problem.

The parameters for the Kepler problem are the same as in Example 6.8: the initial condition is $y(0) = (2, 0, 0, \frac{1}{2})$ and the integration interval is $[0, 3T]$, where $T \approx 9.674$ denotes the time for one revolution.

The parameters ν and σ have to be chosen carefully, as the boundary value problem (7.23) is not an easy problem to solve numerically. We start with $\nu = \frac{1}{10}$ and $\sigma = 1$, and use continuation. At every iteration, we double ν , we halve σ , and we call COLNEW with the previous solution as first guess. We stop after the thirtieth call of COLNEW, when $\nu \approx 5 \cdot 10^7$ and $\sigma \approx 2 \cdot 10^{-9}$. The result of this computation is depicted in Figure 7.4. This figure shows also the other three strategies discussed above.

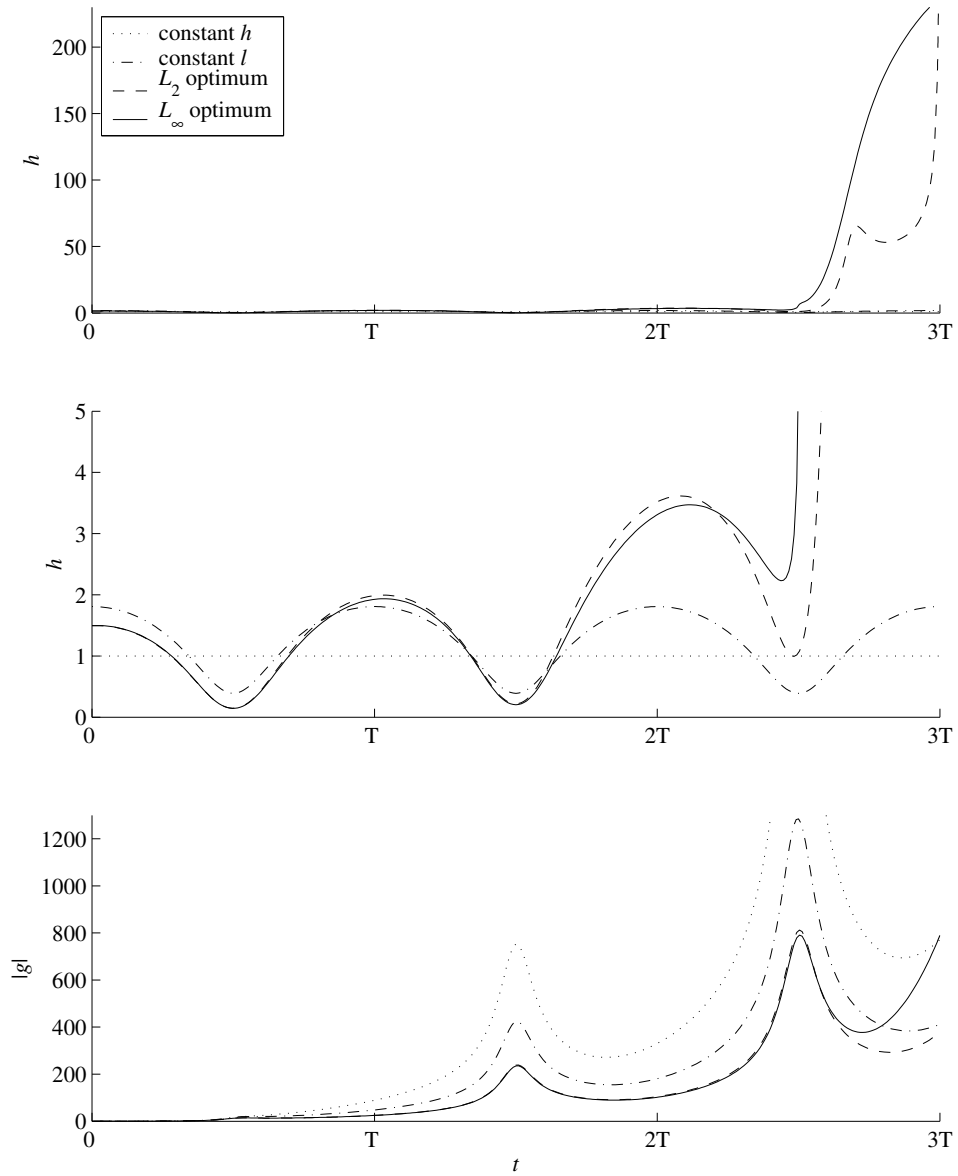


Figure 7.4: The top two plots show (on a different scale) four different step size strategies for the Kepler problem with the (standard) Euler method: constant step size (dotted line), constant local error (dash-dotted line), L_2 -optimal step size (dashed line), and L_∞ -optimal step size (solid line). The bottom plot shows the corresponding global error.

Figure 7.4 clearly shows that keeping the step size constant results in the largest error. The second strategy (constant local error) entails taking small steps around $t = \frac{1}{2}T$, $t = 1\frac{1}{2}T$, and $t = 2\frac{1}{2}T$, when the satellite is closest to the origin, and larger steps in between (see also Figure 7.5). This results in a smaller global error (with the same number of steps). The L_2 -optimal strategy also decreases the step size when the satellite approaches the origin, but it adds a tendency to increase the step size throughout the integration interval. This leads to a further decrease in the global error. Finally, consider the L_∞ -optimal strategy. It is very similar to the L_2 -optimal strategy up to $t = 2\frac{1}{2}T$, where the global error achieves its maximum. After $t = 2\frac{1}{2}T$, the L_∞ -optimal strategy takes even bigger steps than the L_2 -optimal strategy, because it only needs to keep the global error under the maximum achieved around $t = 2\frac{1}{2}T$. This freedom is used to take slightly smaller steps before $t = 2\frac{1}{2}T$. In fact, the L_∞ -optimal strategy takes 55%, 39%, and 6% of the steps in the first, second and third period respectively. The corresponding percentages for the L_2 -optimal strategy are 54%, 37%, and 9%, whereas the other two strategies take the same number of steps in each of the three periods.

Budd, Leimkuhler and Piggott [13] note that the Kepler equation (6.25) is invariant under the rescaling

$$(t, y_1, y_2, y_3, y_4) \mapsto (\alpha t, \alpha^{2/3}y_1, \alpha^{-1/3}y_2, \alpha^{2/3}y_3, \alpha^{-1/3}y_4).$$

They argue that the choice of step size should reflect this scaling invariance, which leads them to take the step size proportional to $r^{3/2}$. This strategy is almost the same as keeping the local error constant (see Figure 7.5), and the resulting global error differs by only a few percent.

We repeat the calculation for the symplectic Euler method (6.29), which was introduced at the end of the previous chapter. The boundary value problem is the same as for the standard Euler method, except that the local error (6.27) is replaced by (6.30). Now, the computation is even more sensitive to the choice of the parameters, especially the penalty parameter ν . In fact, COLNEW is unable to solve the boundary value problem at the twenty-fourth call, when $\nu \approx 8 \cdot 10^5$: more and more collocation points are added by the program, until it runs out of memory. Hence, we take the result of the twenty-third call as the L_∞ optimum. Surprisingly, it is counterproductive to multiply ν by $\frac{3}{2}$ (instead of 2) at every iteration; in that case, COLNEW gives up at $\nu \approx 6 \cdot 10^3$. Other multiplicative factors are also not effective.

The solid curves in Figure 7.6 show the L_∞ -optimal step size and the resulting global error. The bottom plot shows that there are three boundary intervals around $t = \frac{1}{2}T$, $t = 1\frac{1}{2}T$, and $t = 2\frac{1}{2}T$, where the norm of the global error reaches its

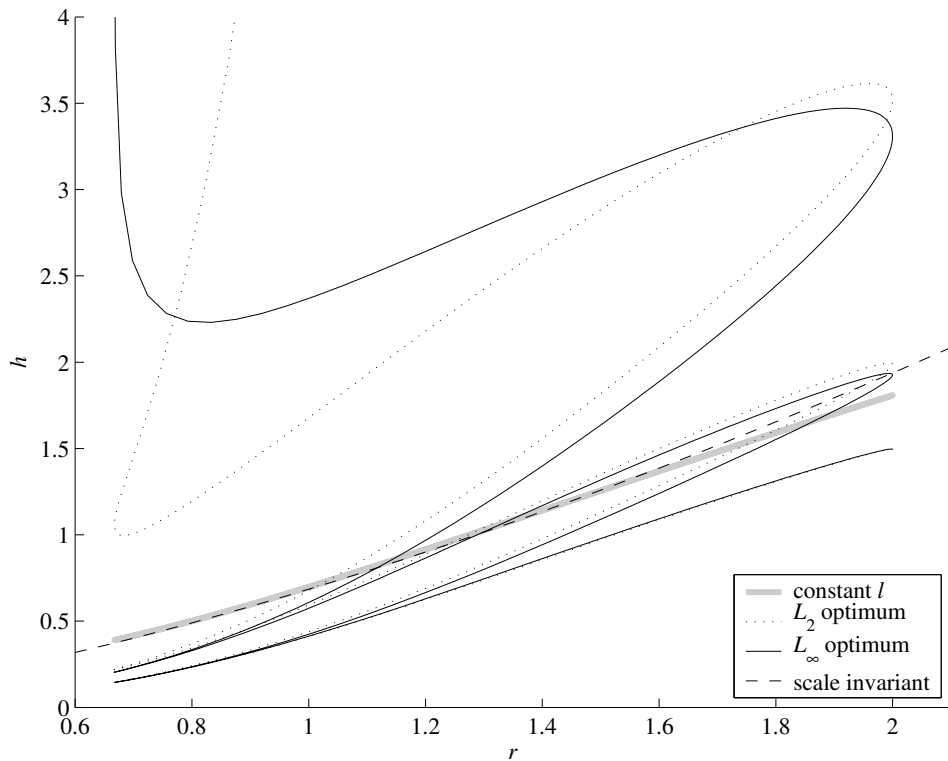


Figure 7.5: This plot shows the step size h as a function of $r = \sqrt{y_1^2 + y_2^2}$ for four different strategies: constant local error (thick grey line), L_2 -optimal step size (dotted line), L_∞ -optimal step size (solid black line), and the scale-invariant choice $h = r^{3/2}$ of Budd, Leimkuhler and Piggott [13] (dashed line).

maximum. This maximum is also reached at t_f , the end point of the integration interval. In contrast, the optimal solution for the (standard) Euler method has no boundary intervals; the norm of the global error only touches the maximum around $t = 2\frac{1}{2}T$ (see Figure 7.4).

Figure 7.6 also shows the three other strategies. Again, it is obvious that the symplectic Euler method behaves differently from the standard Euler method. We see that the simple strategy of keeping the step size constant is doing remarkably well, whereas varying the step size to keep the local error per step constant is disastrous: the norm of the global error peaks above 100 around $t = 2\frac{1}{2}T$. This should not come as a surprise, as it is well known that symplectic methods do not perform well when combined with standard automatic step size selection strategies (see for instance [43, §VIII.1]). Finally, the L_2 -optimal step size strategy computed in Example 6.8 behaves similarly to the L_∞ -optimal strategy. \diamond

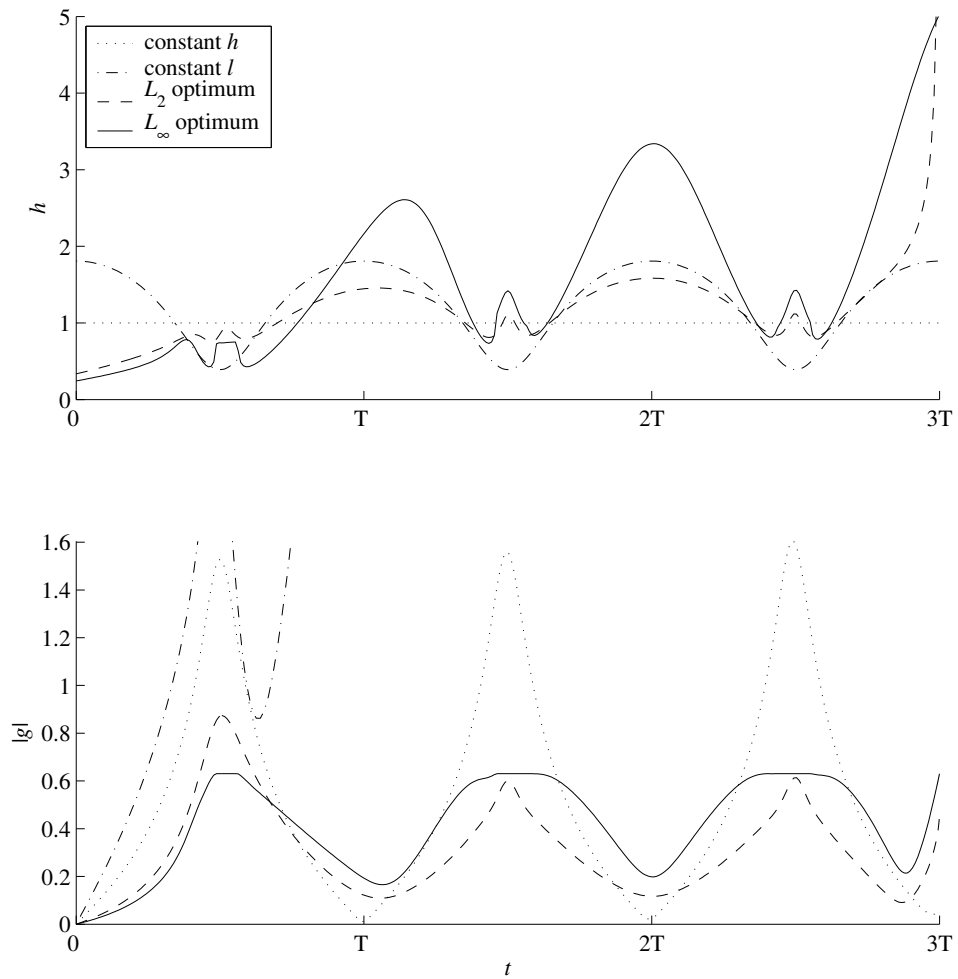


Figure 7.6: Four different step size strategies and the corresponding global error for the Kepler problem with the *symplectic* Euler method. The strategies are the same as in Figure 7.4.

Chapter 8

Conclusions and pointers for further research

In this chapter, the main results of the thesis are summarized. Furthermore, we specify how these results can be used in practice, and we pose some questions that might form a basis for future research.

The chapter is divided in two sections. The first section refers to Part I of the thesis, where we derived estimates for the global error of a numerical method for solving ordinary differential equations. The second section contains the conclusions of Part II on the minimization of the global error.

8.1 Estimating the global error

In Chapter 3, we described three methods for deriving *a priori* estimates for the global error committed by a discretization method for solving ordinary differential equations.

Using Lady Windermere's fan (cf. Figure 3.1), we found an expression for the global error with a remainder term of order h^{2p} , if a constant step-size method is used (cf. Theorem 3.4). The corresponding expression for variable step-size methods is given in Theorem 3.6. Unfortunately, this expression has a remainder term of order ε_h^{p+1} while the global error is of order ε_h^p , so Theorem 3.6 gives only the leading term of the global error.

Another possibility is provided by Theorem 3.8 due to Gragg, which states that the global error has an asymptotic expansion. Every term in this asymptotic expansion can be found by solving a differential equation, in which the previous term in the expansion appears as a source term. This approach works for all methods with step size rules of the form $h_k = \varepsilon_h h(t_k)$.

The third approach uses the theory of modified equations. Theorem 3.10 gives an estimate for the global error with a remainder term of order ε_h^{2p} , if the modified equation is known. If we are using a Runge–Kutta method with fixed step size, we can write this estimate as a linear combination of so-called elementary integrals (cf. Corollary 3.12). The coefficients in this linear combination depend on the coefficients of the Runge–Kutta method.

The second approach has the clear advantage that we can find an estimate for the global error with a remainder term of arbitrary order, at least in theory. However, in practice it may prove hard to solve the differential equations yielding the terms in the asymptotic expansion. In contrast, the estimate obtained using modified equations is only valid for Runge–Kutta methods and up to order h^{2p} , but it is easier to evaluate as the examples show. The first approach does not seem very useful for constant step-size methods, but it still yields an estimate for general variable step-size methods, unlike the other two approaches.

There are two obvious deficiencies in the theory developed in Chapter 3. Firstly, we did not find an accurate estimate for variable-step size methods. Secondly, we only gave the order of the remainder term in the estimates, but we did not give any bounds. A resolution of either of these issues would be very welcome.

It would also be useful to extend the global error estimates to other methods. Cano and Sanz-Serna [21] explain how to use Gragg’s asymptotic expansion for multistep methods. Similarly, the modified equations approach could be generalized using the theory of modified equation for multistep method developed by Hairer [39]. One could also look for estimates for Lie group methods (Berland and Owren [11] describe the corresponding modified equations) or methods for partial differential equations (see De Frutos and Sanz-Serna [29] for a generalization of Theorem 3.6 in this context).

We can use the estimates derived in Chapter 3 to analyse existing methods and to develop methods that perform especially well when applied to a certain class of problems. Chapter 4 describes applications along these lines. We first proved a theorem describing the accumulation of the global error when tracking a periodic orbit. Next, we considered two families of equations with oscillatory behaviour, namely Airy-like equations of the form $y'' + \eta(t)y = 0$ and the Emden–Fowler equation $y'' + t^\nu y^n = 0$. There is no closed-form expression for the exact solution of these equations. Nevertheless, we can still find an estimate for the global error via the modified-equations approach by using the asymptotic solution as $t \rightarrow \infty$. Both the remainder term of order h^{2p} in the estimate and the difference between the asymptotic and the exact solution of the differential equation cause the estimate for the global error to deviate from the actual global error. We did not find a

bound on this deviation. Consequently, we cannot predict with confidence in what time range the estimates are valid, but the experiments show that the estimates are very accurate over a long interval. The estimates show that the error of most Runge–Kutta methods grows at the same rate, but that some methods (like (4.37) for the Emden–Fowler equation) accumulate error at a slower rate. Further research is required to answer questions like whether this only happens for these specific equation, or whether a similar phenomenon happens for a more general class of equations.

8.2 Minimizing the global error

The aim in the second part of the thesis is to make the global error as small as possible by varying the step size of the numerical method in a particular way. But precisely what do we mean with a “small error”? We argued in Chapter 5 that the solution given by a numerical method may be of low quality even though it is close to the exact solution at the final point of the integration interval, or even at all the grid points. This suggests considering the global error as a continuous function of time, and minimizing the norm of this function over the entire integration interval. But which norm should be used? We did not answer this question, because the correct answer probably depends on the specific application. However, in the remainder of the thesis, we concentrated on the two most natural norms, namely the L_2 and L_∞ norms. These are also the norms suggested in the literature.

This choice for the objective turns the problem of determining the optimal step size into an optimal control problem (with state constraints, if the L_∞ norm is used). We used Pontryagin’s Minimum Principle to characterize the optimal step size as the solution of a boundary value problem in Theorems 6.5 and 7.5 for the L_2 and L_∞ norm respectively. The result for the L_2 norm has the counterintuitive implication that the optimal step size becomes unbounded as we approach the end of the integration interval. On the other hand, we had to assume a technical condition (namely that the state constraint is of first order) to derive the result for the L_∞ norm. No interpretation of this condition was given; we hope that future work will enable us to do so. Another deficiency is that we were unable to prove the existence of an optimal step size, but we did hint at a possible strategy: it suffices to obtain an *a priori* bound for the optimal step size (for the L_2 norm, this requires a rescaling to eliminate the unboundedness at the end of the integration interval). Note that in fact routines for solving ordinary differential equations often include an upper bound for the step size. We also found an interesting parallel between the solution of the optimal control problem and the idea of equidistribu-

tion, as described by Eriksson, Estep, Hansbo and Johnson [27, 28]. It is probably worthwhile to explore this connection further.

Generally, the boundary value problem cannot be solved analytically, so we have to find recourse in a numerical method. The boundary value problem for the L_2 norm is of standard form, and the currently available software can solve it. However, the problem for the L_∞ norm includes a complementarity condition, rendering it in nonstandard form, so we used a different approach. We went back to the optimal control problem, and solved it with an exterior penalty method. The resulting method does manage to solve the examples that we considered, but it is not very robust. It would probably be better to solve the boundary value problem with the complementary condition. This can probably be done with a method based on collocation. If such an algorithm is developed, it would not only be useful to compute the optimal step size, but also for solving other state-constrained optimal control problem.

The step size selection algorithm is an important part of practical codes for solving ordinary differential equations. It is natural to ask oneself how the results obtained in this thesis can be used to improve the current mechanisms for choosing the step size. The numerical algorithms for determining the optimal step size, which are developed in Sections 6.3 and 7.3, cannot be implemented straight away in a numerical integrator because they are based on *a priori* estimates of the global error. Specifically, the estimates require the exact solution to be known. Of course, they may be replaced by *a posteriori* estimates, for instance by using the numerical solution instead of the exact solution. But the resulting algorithm will be slow, since it entails the solution of a boundary value problem (albeit at a low precision), which is far more costly than the solution of the original initial value problem. Nevertheless, the algorithms from Sections 6.3 and 7.3 might provide a basis on which a practical method may be built.

However, the main application envisaged for the results of the second part of the thesis is in the analysis of the error control mechanisms that are being used in practice. We can compare the step size that is chosen by these practical methods with the optimal step size. This will allow us to assess the quality of the current algorithms for selecting the step size. A first step in this direction would be to calculate the optimal step size for a wide variety of test problems, for instance those in the Bari test set [65], and to compare the results with the step size used by various numerical integrators. One may well be able to distinguish broad classes of problems for which the current methods are fairly bad or near-optimal, and use this knowledge to improve the current methods.

Bibliography

- [1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, 1964.
- [2] P. Albrecht. The Runge–Kutta theory in a nutshell. *SIAM J. Numer. Anal.*, 33(5):1712–1735, 1996.
- [3] M. Aripov and D. Éshmatov. On WKB-solutions of a generalized equation of Emden–Fowler type. *Dokl. Akad. Nauk UzSSR*, (9):4–6, 1988. In Russian.
- [4] U. M. Ascher, J. Christiansen, and R. D. Russell. A collocation solver for mixed order systems of boundary value problems. *Math. Comp.*, 33(146):659–679, 1979.
- [5] U. M. Ascher, J. Christiansen, and R. D. Russell. Collocation software for boundary-value ODEs. *ACM Trans. Math. Software*, 7(2):209–222, 1981.
- [6] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.
- [7] G. Bader and U. M. Ascher. A new basis implementation for a mixed order boundary value ODE solver. *SIAM J. Sci. Stat. Comput.*, 8(4):483–500, 1987.
- [8] E. N. Barron and J. Ishii. The Bellman equation for minimizing the maximum cost. *Nonlinear Anal.*, 13(9):1067–1090, 1989.
- [9] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- [10] L. D. Berkovitz. *Optimal Control Theory*. Springer-Verlag, New York, 1974.
- [11] H. Berland and B. Owren. Algebraic structures on ordered rooted trees and their significance to Lie group integrators. Numerics preprint 3/2003, Norwegian University of Science and Technology, Trondheim, Norway, 2003.

-
- [12] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, Mass, 1995. Two volumes.
- [13] C. Budd, B. Leimkuhler, and M. Piggott. Scaling invariance and adaptivity. *Appl. Numer. Math.*, 39(3-4):261–288, 2001.
- [14] J. C. Butcher. Optimal order and stepsize sequences. *IMA J. Numer. Anal.*, 6(4):433–438, 1986.
- [15] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Chichester, 2003.
- [16] J. C. Butcher and J. M. Sanz-Serna. The number of conditions for a Runge–Kutta method to have effective order p . *Appl. Numer. Math.*, 22(1–3):103–111, 1996.
- [17] M. Calvo, D. J. Higham, J. I. Montijano, and L. Randez. Global error estimation with adaptive explicit Runge–Kutta methods. *IMA J. Numer. Anal.*, 16(1):47–63, 1996.
- [18] M. P. Calvo and E. Hairer. Accurate long-term integration of dynamical systems. *Appl. Numer. Math.*, 18:95–105, 1995.
- [19] M. P. Calvo, A. Murua, and J. M. Sanz-Serna. Modified equations for ODEs. In *Chaotic Numerics*, volume 172 of *Contemp. Math.*, pages 53–74, Geelong, 1993, 1994. Amer. Math. Soc., Providence, RI.
- [20] B. Cano and J. M. Sanz-Serna. Error growth in the numerical integration of periodic orbits, with application to Hamiltonian and reversible systems. *SIAM J. Numer. Anal.*, 34(4):1391–1417, 1997.
- [21] B. Cano and J. M. Sanz-Serna. Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems. *IMA J. Numer. Anal.*, 18(1):57–75, 1998.
- [22] L. Cesari. *Optimization—Theory and Applications: Problems with Ordinary Differential Equations*, volume 17 of *Applications of Mathematics*. Springer, New York, 1983.
- [23] G. Dahlquist. Stability and error bounds in the numerical integration of ordinary differential equations. Kungl. Tekn. Högsk. Handl. Stockholm, No. 130, 1959.

- [24] G. Dahlquist. On the control of the global error in stiff initial value problems. In G. Watson, editor, *Numerical Analysis*, volume 912 of *Lecture Notes in Mathematics*, pages 38–49, Dundee, 1982. Springer-Verlag, Berlin.
- [25] V. V. Dikusar. Methods of control theory for the numerical integration of ordinary differential equations. *Differentsial'nye Uravneniia*, 30(12):2116–2121, 1994. In Russian. English translation in *Differential Equations*, 30(12):1944–1949.
- [26] W. H. Enright, D. J. Higham, B. Owren, and P. W. Sharp. A survey of the explicit Runge–Kutta method. Technical Report 291/94, Department of Computer Science, University of Toronto, 1994.
- [27] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to adaptive methods for differential equations. *Acta Numerica*, pages 105–158, 1995.
- [28] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Cambridge University Press, 1996.
- [29] J. de Frutos and J. M. Sanz-Serna. Accuracy and conservatino properties in numerical integration: The case of the Korteweg-de Vries equation. *Numer. Math.*, 75(4):421–445, 1997.
- [30] M. Fujii. Optimal choice of mesh points for Adams-type-methods with variable step size. *Bull. Fukuoka Univ. Ed. III*, 20:31–46, 1970.
- [31] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [32] S. Gnutzmann and U. Ritschel. Analytic solution of Emden–Fowler equation and critical adsorption in spherical geometry. *Z. Phys. B*, 96(3):391–393, 1995.
- [33] H. Goenner and P. Havas. Spherically symmetric space-times with constant curvature scalar. *J. Math. Phys.*, 42(2):1837–1859, 2001.
- [34] W. B. Gragg. On extrapolation algorithms for ordinary initial value problems. *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, 2(3):384–404, 1965.
- [35] H. Greenspan, W. Hafner, and M. Ribarič. On varying stepsize in numerical integration of first order differential equations. *Numer. Math.*, 7:286–291, 1965.

- [36] D. F. Griffiths and J. M. Sanz-Serna. On the scope of the method of modified equations. *SIAM J. Sci. Stat. Comput.*, 7(3):994–1008, 1986.
- [37] K. Gustafsson. Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods. *ACM Trans. Math. Software*, 17(4):533–554, 1991.
- [38] K. Gustafsson, M. Lundh, and G. Söderlind. A PI stepsize control for the numerical solution of ordinary differential equations. *BIT*, 28(2):270–287, 1988.
- [39] E. Hairer. Backward error analysis for multistep methods. *Numer. Math.*, 84(2):199–232, 1999.
- [40] E. Hairer and C. Lubich. Asymptotic expansions of the global error of fixed-stepsize methods. *Numer. Math.*, 45:345–360, 1984.
- [41] E. Hairer and C. Lubich. The life-span of backward error analysis for numerical integrators. *Numer. Math.*, 76:441–462, 1997.
- [42] E. Hairer and C. Lubich. Asymptotic expansions and backward analysis for numerical integrators. In R. de la Llave, L. Petzold, and J. Lorenz, editors, *Dynamics of algorithms (Minneapolis, MN, 1997)*, volume 118 of *IMA Vol. Math. Appl.*, pages 91–106. Springer, New York, 2000.
- [43] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2002.
- [44] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer-Verlag, Berlin, second edition, 1993.
- [45] E. Hairer and D. Stoffer. Reversible long-term integration with variable step-sizes. *SIAM J. Sci. Comput.*, 18(1):257–269, 1997.
- [46] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, second edition, 1996.
- [47] R. F. Hartl, S. P. Sethi, and R. G. Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995.

- [48] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons, New York, 1962.
- [49] P. Henrici. *Error Propagation for Difference Methods*. The SIAM series in applied mathematics. John Wiley & Sons, New York, 1963.
- [50] M. R. Hestenes. *Calculus of Variations and Optimal Control Theory*. Wiley, New York, 1966.
- [51] E. Hinch. *Perturbation Methods*. Cambridge texts in applied mathematics. Cambridge University Press, Cambridge, UK, 1991.
- [52] E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. John Wiley & Sons, New York, 1966.
- [53] A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.
- [54] A. Iserles. On the global error of discretization methods for highly-oscillatory ordinary differential equations. *BIT*, 42(3):561–599, 2002.
- [55] A. Iserles and G. Söderlind. Global bounds on numerical error for ordinary differential equations. *J. Complexity*, 9:97–112, 1993.
- [56] D. H. Jacobson, M. M. Lele, and J. L. Speyer. New necessary conditions of optimality for control problems with state-variable inequality constraints. *J. Math. Anal. and Appl.*, 35(2):255–284, 1971.
- [57] D. Kincaid and W. Cheney. *Numerical Analysis*. Brooks/Cole Series in Advanced Mathematics. Brooks/Cole, Pacific Grove, CA, third edition, 2002.
- [58] J. D. Lambert. *Numerical Methods for Ordinary Differential Equations: The Initial Value Problem*. John Wiley & Sons, Chichester, 1991.
- [59] J. H. Lane. On the theoretical temperature of the Sun under the hypothesis of a gaseous mass maintaining its volume by its internal heat and depending on the laws of gases known to terrestrial experiment. *Amer. J. Sci. and Arts*, 2nd series, 50:57–74, 1870.
- [60] E. B. Lee and L. Markus. *Foundations of Optimal Control Theory*. Wiley, New York, 1967.
- [61] B. Lindberg. Characterization of optimal stepsize sequences for methods for stiff differential equations. *SIAM J. Numer. Anal.*, 14(5):859–887, 1977.

- [62] J. Macki and A. Strauss. *Introduction to Optimal Control Theory*. Undergraduate Texts in Mathematics. Springer, New York, 1982.
- [63] H. Maurer. On the minimum principle for optimal control problems with state constraints. Schriftenreihe des Rechenzentrums der Univ. Münster, Nr. 41, Münster, Germany, 1979.
- [64] H. Maurer, C. Büskens, and G. Feichtinger. Solution techniques for periodic control problem: A case study in production planning. *Optimal Control Appl. Methods*, 19(3):185–203, 1998.
- [65] F. Mazzia and F. Iavernaro. Test set for initial value problem solvers. Technical Report 40/2003, Department of Mathematics, University of Bari, Italy, 2003. Web site at <http://pitagora.dm.uniba.it/~testset/>.
- [66] B. Meerson, E. Megged, and T. Tajima. On the quasi-hydrostatic flows of radiatively cooling self-gravitating gas clouds. *Astrophys. J.*, 456:321–331, 1996.
- [67] P. C. Moan. *On Backward Error Analysis and Nekhoroshev Stability in Numerical Analysis of Conservative ODEs*. PhD thesis, University of Cambridge, UK, 2002.
- [68] K.-S. Moon, A. Szepessy, R. Tempone, and G. E. Zouraris. Convergence rates for adaptive approximation of ordinary differential equations. *Numer. Math.*, 96(1):99–129, 2003.
- [69] K.-S. Moon, A. Szepessy, R. Tempone, and G. E. Zouraris. A variational principle for adaptive approximation of ordinary differential equations. *Numer. Math.*, 96(1):131–152, 2003.
- [70] D. Morrison. Optimal mesh size in the numerical integration of an ordinary differential equation. *J. Assoc. Comput. Mach.*, 9(1):98–103, 1962.
- [71] E. H. Neville. *Jacobian Elliptic Functions*. Clarendon Press, Oxford, 1944.
- [72] J. Niesen. A priori estimates for the global error committed by Runge–Kutta methods for a nonlinear oscillator. *LMS J. Comput. Math.*, 6:18–28, 2003.
- [73] H. J. Oberle. Numerical solution of minimax optimal control problems by multiple shooting technique. *J. Optim. Theory Appl.*, 50(2):331–357, 1986.

- [74] H. J. Oberle and W. Grimm. BNDSO – A program for the numerical solution of optimal control problems. Internal report 515-89/22, Institute for Flight Systems Dynamics, DLR, Oberpfaffenhofen, Germany, 1989.
- [75] K. Okamura. Some mathematical theory of the penalty method for solving optimum control problems. *J. Soc. Indust. Appl. Math. Ser. A Control*, 2(3):317–331, 1965.
- [76] F. W. Olver. *Asymptotics and Special Functions*. Academic Press, New York, 1974.
- [77] B. Orel. Runge–Kutta and Magnus methods for oscillatory ODEs. Talk delivered at the International Conference on Scientific Computation and Differential Equations (SciCADE), Vancouver, 2001.
- [78] H. J. Pesch. Real-time computation of feedback controls for constrained optimal control problems. *Optimal Control Appl. Methods*, 10(2):129–171, 1989.
- [79] P. J. Prince and J. R. Dormand. High order embedded Runge–Kutta formulae. *J. Comput. Appl. Math*, 7(1):67–75, 1981.
- [80] R. D. Skeel. Thirteen ways to estimate global error. *Numer. Math.*, 48(1):1–20, 1986.
- [81] H. J. Stetter. *Analysis of Discretization Methods for Ordinary Differential Equations*, volume 23 of *Springer tracts in natural philosophy*. Springer, Berlin, 1973.
- [82] D. Stoffer and K. Nipp. Invariant curves for variable step size integrators. *BIT*, 31:169–180, 1991. Erratum in vol. 32, pp. 367–369.
- [83] M. Utumi, R. Takaki, and T. Kawai. Optimal time step control for the numerical solution of ordinary differential equations. *SIAM J. Numer. Anal.*, 33(4):1644–1653, 1996.
- [84] D. Viswanath. Global errors of numerical ODE solvers and Lyapunov’s theory of stability. *IMA J. Numer. Anal.*, 21(1):387–406, 2001.
- [85] A. G. Werschulz. *The Computational Complexity of Differential and Integral Equations. An Information-Based Approach*. Oxford University Press, 1991.
- [86] J. S. W. Wong. On the generalized Emden–Fowler equation. *SIAM Review*, 17(2):339–360, 1975.

-
- [87] E. Zermelo. Über das Navigationsproblem bei ruhender oder veränderlicher Windverteilung. *Z. Angew. Math. Mech.*, 22(2), 1931.