A Lyapunov exponent based stability theory for ordinary differential equation initial value problem solvers

By

Andrew J Steyer

Submitted to the Department of Mathematics and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Erik Van Vleck, Chairperson

Weizhang Huang

Weishi Liu

Hongguo Xu

David Mechem

Date defended: August 2nd, 2016

The Dissertation Committee for Andrew J Steyer certifies
that this is the approved version of the following dissertation :

A Lyapunov exponent based stability theory for ordinary differential equation initial value problem solvers

____

Erik Van Vleck, Chairperson

Date approved: August 2nd, 2016

Abstract

In this dissertation we consider the stability of numerical methods approximating the solution of bounded, stable, and time-dependent solutions of ordinary differential equation initial value problems. We use Lyapunov exponent theory to determine conditions on the maximum allowable step-size that guarantees that a one-step method produces a decaying numerical solution to an asymptotically contracting, time-dependent, linear problem. This result is used to justify using a one-dimensional asymptotically contracting real-valued nonautonomous linear test problem to characterize the stability of a one-step method. The linear stability result is applied to prove a stability result for the numerical solution of a class of stable nonlinear problems. We use invariant manifold theory to show that we can obtain similar stability results for strictly stable linear multistep methods approximating asymptotically contracting, time-dependent, linear problems by relating their stability to the stability of an underlying one-step method. The stability theory for one-step methods is used to devise a procedure for stabilizing a solver that fails to produce a decaying solution to a linear problem when selecting step-size using standard error control techniques. Additionally, we develop an algorithm that selects step-size for the numerical solution of a decaying nonautonomous scalar test problem based on accuracy and the stability theory we developed.

Acknowledgements

Throughout my time in graduate school I have been the beneficiary of direct and indirect support from many individuals. I am especially thankful to my adviser, Erik Van Vleck, for his guidance and direction. He has taught me so much about research, mathematics, and life. Any theory I discover, any product I design, or project I complete will always be at least indirectly influenced by his mentorship. There is no way can express the gratitude that I have for all the time, funding, and patience he has given me throughout my time in graduate school.

In addition to my adviser, I would like to thank the members of my thesis committee, David Mechem, Hongguo Xu, Weishi Liu, and Weizhang Huang for their comments and constructive criticism.

My sincere thanks also goes to the office support staff: Gloria Prothe, Kerrie Breicheisen, and Lori Springs for their help in filing paperwork, meeting deadlines, and generally keeping the department running smoothly.

I would also like to thank my fiancé Alexandria Tremble for her love and support during the latter half of my studies and while I wrote this dissertation. Her help for the last several months especially has been crucial and I am ecstatic that we will soon be married. I am quite lucky to be the recipient of her ongoing support in my research career.

I am also grateful to my cat, Bucky, for always being there for me even when I did not want him to be. I would also like to thank the other graduate students in the Department of Mathematics for being a source of friendship, community, and support. Finally, I would like to thank Trey Anastasio for some dank guitar jams.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The topic of this dissertation is the stability of numerical methods for the approximation of time-dependent (nonautonomous) initial value problems (IVPs) for ordinary differential equations (ODEs). Differential equations are ubiquitous in the modeling of real-world physical processes and dynamical systems. For physical systems that evolve in time they provide mathematical tools for forecasting the future state of the system from some initially prescribed state (or estimate therof), usually referred to as the initial condition. The exact solution to most differential equations that model physical systems is impossible to know exactly. Hence, numerical methods for the approximation of solutions to differential equations are essential tools for researchers seeking to model and predict the evolution of a dynamical system.

An enhanced theoretical understanding of time-dependent stability for the numerical solution of ODE IVPs has far reaching ramifications. In Steyer, A.J. & Van Vleck, E.S (2015), Lyapunov exponent theory is used to develop a step-size selection algorithm for nonautonomous scalar test problems based on Lyapunov exponent stability as well as accuracy. This motivated a better theoretical understanding of the time-dependent stability of ODE IVP solvers. This led to the development of a Lyapunov exponent based stability theory for one-step ODE IVP solvers and general linear methods is developed in Steyer, A.J. & Van Vleck, E.S (2016b) and Steyer, A.J. & Van Vleck, E.S (2016a) respectively. Applications of time-dependent stability theory for ODE IVP solvers includes work on nonautonomous bifurcations and tipping phenomena in Hoyer-Leitzel, A. et al. (2016), global error analysis Chung Y.-M. et al. (2016), the computation of inertial manifolds Chung, Y.-M. et al. (2016), and data assimilation Dubinkin, S. et al. (2016). Understanding the time-dependent stability of ODE IVP solvers can lead to greater computational performance, more robust algorithms, and better resolution of dynamics.

Ordinary differential equations are important and useful in directly modeling real-life dynamical systems and for the numerical solution of partial differential equations (PDEs) by the method of lines. There are always errors in the approximation of the solution of an ODE IVP and if these errors accumulate, then a numerical method can introduce instabilities that are unrelated to the dynamics of the ODE. If the the numerical solution fails to accurately resolve the stability of a trajectory, then there can be no confidence that the output of a solver will remain accurate over a long time interval. Thus, understanding and preserving the stability of ODE IVP solvers is crucial for dynamical systems arising in areas such as climate and earth system modeling where simulations forecast the state of the system for long periods of time.

Consider the ordinary differential equation (ODE) initial value problem (IVP)

$$\begin{cases} \dot{x}(t) = f(x(t),t), \quad t > t_0 \\ x(t_0) = x_0 \end{cases} \tag{1.1}$$

where $t_0 \in \mathbb{R}$, $d \geq 1$, $f : \mathbb{R}^d \times (t_0,\infty) \to \mathbb{R}^d$ has derivatives of all orders, and $f(x,\cdot)$ is bounded on (t_0,∞) for each fixed $x \in \mathbb{R}^d$. We that the solution $x(t;x_0)$ of (1.1) is bounded and Lyapunov stable in the sense that the solution $x(t;y_0)$ of the ODE $\dot{x}(t) = f(x(t),t)$ with initial condition y_0 remains near to $x(t;x_0)$ for all $t > t_0$ whenever y_0 is sufficiently close to x_0.

We fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$ and use the same symbol $\|\cdot\|$ to denote the induced matrix norm on $\mathbb{R}^{d\times d}$ and we also fix an arbitrary orthogonal basis $\{e_1,\ldots,e_d\}$ of $\mathbb{R}^d$. The standard Lyapunov stability analysis of the solution of (1.1) begins with linearization in space about the solution $x(t;x_0)$ to obtain the associated linear variational equation

$$\dot{u}(t) = A(t)u(t), \quad t > t_0 \tag{1.2}$$

where the coefficient matrix is defined as $A(t) := Df(x(t;x_0),t)$ where $D := \partial/\partial x$. The matrix-valued function $A(t)$ is bounded and continuous since $x(t;x_0)$ is bounded and Df is continuous. Under mild conditions on $A(t)$ and the nonlinear terms $N(x,t) := f(x,t) - A(t)x$ we can infer the stability of the solution of (1.1) from the stability of the zero solution $u(t) \equiv 0$ of (1.2). For general time-dependent linear systems of the form (1.2) the stability of the zero solution is not generally dependent on the eigenvalues of $A(t)$ and, in fact, examples in Coppel (1978); Kreiss (1978) show that eigenvalues may give counter-indicative stability information. This has led to the development of several alternative spectral stability theories for characterizing the stability of the zero solution of (1.2). The spectral stability theory we consider in this paper is the theory of Lyapunov exponents.

Our contribution in this dissertation is to apply the approximation theory for Lyapunov exponents by QR methods to develop a time-dependent stability theory for numerical methods approximating a class of stable initial value problems. This theory allows us to weaken the hypotheses made in AN-stability and B-stability theory requiring that the differential equation is uniformly contracting at the expense of acquiring an inherent step-size restriction. First we use local error estimates to give conditions on the maximum allowable step-size so that a one-step method produces an asymptotically decaying solution when solving a test problem of the form

$$\dot{x}(t) = \lambda(t)x(t), \quad \lambda(t) \in \mathbb{R} \tag{1.3}$$

where $\lambda(t)$ is asymptotically contracting in the sense that its time average is negative on all sufficiently large time intervals. For such test problems the coefficient function $\lambda(t)$ is allowed to take on positive values for infinitely many t and because of this we can show that, in contrast to A-stability theory, there are no Runge-Kutta methods that produce a decaying numerical solution to all such test problems without a step-size restriction. We employ a time-dependent orthogonal change of variables to transform to a corresponding linear system with an upper triangular coefficient matrix and use this system to justify characterizing the numerical stability of a one-step method approximating a time-dependent linear problem (1.2) using test problems of the form (1.3). This is contrasted to the time-independent case where a similarity transformation (i.e. a time-independent change of variables) to the Jordan canonical form is used to justify using time-independent scalar test problems to characterize the stability of methods approximating time-independent linear problems. In general, numerical stability is not automatically preserved under a time-dependent orthogonal change of variables since the step-size must be small enough so that this change of variables resolves the geometry of the change of variables is preserved at the discrete level. We show that under generic hypotheses on (1.2) we can determine an additional step-size restriction so that this change of variables accurately preserves the underlying geometry. The linear theory we develop is then used to prove a stability result for Runge-Kutta methods solving a nonlinear system of the form (1.1) whose linear part satisfies the hypotheses of the linear theory we develop. The linear theory for one-step methods is extended to strictly stable linear multistep methods by reducing their analysis when applied to linear problems to that of one-step methods. We then use the theory as the basis for step-size selection algorithms we develop based on controlling the Lyapunov exponent stability.

The stability analysis of numerical methods that approximate the solution of (1.1) for time-dependent and time-independent problems is a well-developed field. The earliest work on the stability of time-stepping methods solving initial value problems is due in Dahlquist (1963) where A-stability and the theory of linear stability domains for linear multistep methods is developed and it is shown how to characterize the numerical stability of linear multistep methods approximating the solution of a time-independent linear problem $\dot{x}(t) = Ax(t)$ where $A \in \mathbb{R}^{d\times d}$ by using test problems of the form $\dot{z}(t) = \lambda z(t)$ where $\lambda \in \mathbb{C}$. The notion of A-stability and linear stability domains was extended to Runge-Kutta methods independently in Ehle, B.L (1968); Ehle, B.L. (1973) and Axelsson (1969). Following this, stability theories for Runge-Kutta methods such as AN-stability, B-stability, and algebraic stability and deeper results on A-stability were developed in Butcher, J.C.

& Burrage, K. (1979); Butcher, J.C. (1975, 1987); Crouzeix (1979); Nevanlinna (1977); Nevanlinna, O. & Liniger, W. (1978, 1979); Nevanlinna, O. & Jeltsch, R. (1982); Nevanlinna & Sipilä, A.H. (1974); Scherer (1979); Wanner (1976) and in many other works. For an extensive survey of these classical theories see Hairer, E. et al. (1987); Hairer, E. & Wanner, G. (1991). The theories of algebraic stability, B-stability, and AN-stability all deal with the numerical stability of nonautonomous linear and dissipative nonlinear problems that are uniformly contracting in some sense. Runge-Kutta methods that are, e.g., B-stable and used to solve a uniformly contracting problem will generally produce a contracting numerical solution at each step with no step-size restriction. While this is quite a desirable property for a method to have it is also quite restrictive. For instance no explicit method is B-stable. There are relatively recent results such as González, C.and Palencia, C. (1999); González, C. & Palencia, C. (2000) and Boutelje, B.R. & Hill, A.T. (2010) that allow for a somewhat larger class of methods, but still require that the problem is uniformly contracting.

In recent years the approximation theory of Lyapunov exponents by QR methods has been developed extensively (Dieci, L. & Van Vleck, E.S. (2002, 2003); Dieci & Van Vleck, E.S. (2005); Dieci, L. & Van Vleck, E.S. (2006, 2007); Dieci & Van Vleck, E.S. (2009); Badawy, M. & Van Vleck, E.S. (2012)). For any fundamental matrix solution $X(t)$ of a nonautonomous linear system of the form (1.2) there exists unique QR factorization $X(t) = Q(t)R(t)$ where $Q(t)$ is orthogonal and $R(t)$ upper triangular with positive diagonal entries. The linear system $\dot{y}(t) = B(t)y(t)$ that results from the change of variables $x(t) = Q(t)y(t)$ has an upper triangular coefficient matrix and generically the Lyapunov exponents can be expressed in terms of the diagonal elements of $B(t)$ (see Dieci, L. & Van Vleck, E.S. (2003)). Continuous QR methods approximate $Q(t)$ by solving an additional system of differential equations that depends on $A(t)$ and then approximate the Lyapunov exponents using the resulting approximations to the diagonal entries of $B(t)$.

While relying heavily on the methods and techniques used in the analysis of QR methods in Dieci & Van Vleck, E.S. (2005) and Van Vleck, E.S. (2010), this paper still constitutes a substantial body of original research. Our focus is to apply the existing theory to determine step-size restrictions for the numerical preservation of asymptotic decay as opposed to finding conditions on the local error so that a numerical method approximates the exact Lyapunov exponents of a continuous time system. We use the theory of QR methods to determine step-size restrictions, prove rigorous decay estimates, and justify characterizing the stability of a method using a scalar test problem. This is analogous to the time-independent stability theory for Runge-Kutta methods which relies on eigenvalues, linear algebra, and similarity transformations to obtain estimates and justify using complex scalar test problems to characterize the stability of a method. Additionally, our results are used to provide a practical method to stabilize a solver that unstably solves an asymptotically contracting linear problem and to develop a method for selecting step-size based on accuracy and Lyapunov exponent stability.

Stability theories for time-stepping methods solving (1.2) typically assume that the differential equation is uniformly contracting. However, there are many stable and decaying problems that are only decaying in an asymptotic limit. Understanding the numerical stability of such problems is important since there exist non-uniformly decaying differential equations for which a one-step method with adaptive step-size error control can still fail to produce a decaying numerical solution. For instance, consider the following linear problem,

$$\dot{x}(t) = A(t)x(t) \equiv [Q(t)B(t)Q(t)^T + \dot{Q}(t)Q(t)^T]x(t) \tag{1.4}$$

where

$$B(t) = \begin{bmatrix} a_1 + b_1\cos(\omega_1 t) & \beta \\ 0 & a_2 + b_2\cos(\omega_2 t) \end{bmatrix}, \quad Q(t) = \begin{bmatrix} \cos(\omega_3 t) & \sin(\omega_3 t) \\ -\sin(\omega_3 t) & \cos(\omega_3 t) \end{bmatrix}. \tag{1.5}$$

Each solution $x(t)$ of (1.4) satisfies $\|x(t)\| \leq K_1 e^{a_1 t} + K_2 e^{a_2 t}$ for positive constants K_1 and K_2. In Figure 1.1 we show the results of a Matlab experiment of the numerical solution of (1.4) where $a_1, a_2 < 0$ so that every exact solution of (1.4) decays exponentially fast. The solver used in the experiment was the Matlab ode15s solver using BDF integration formulas using a maximum order

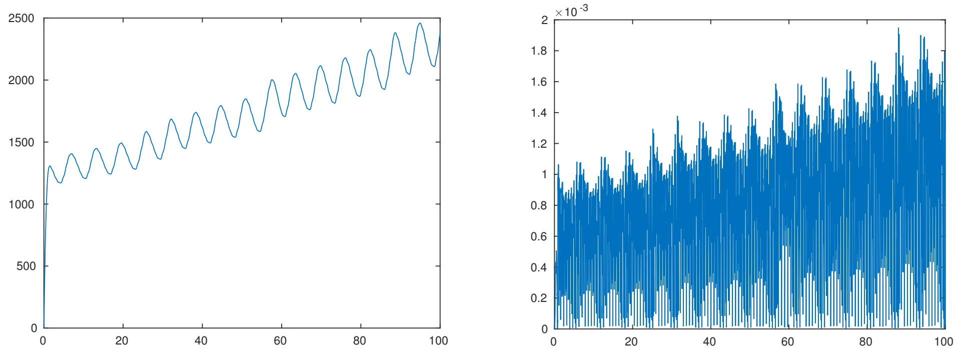


Figure 1.1: Left: Plot of the norm of the numerical solution versus time. Right: Plot of the approximate local truncation error versus time. Numerical solutions were computed with the Matlab solver ode15s using BDF's with a maximum order of 1 (implicit Euler method). Absolute and relative tolerances used in the solver was 10^{-6}, 190368 time-steps were used, and the parameter values $a_1 = -0.2$, $a_2 = -0.3$, $b_1 = 0.21$, $b_2 = 0.31$, $\omega_1 = \omega_2 = 1$, $\omega_3 = 2$, and $\beta = 3 \cdot 10^3$ were used. The initial condition used was $(-1,1)^T/\sqrt{2}$.

of 1 which is the implicit Euler method. The plots of Figure 1.1 show that the AN-stable and B-stable implicit Euler method produces an unstable numerical solution even while using adaptive step-size selection so that the local truncation error is bounded by $2 \cdot 10^{-3}$. This dissertation seeks to provide a theoretical understanding for this anomalous instability phenomenon and provide an efficient method for stabilizing the solver.

The remainder of this work is organized as follows. In Section 1.2 we review the background on the standard stability theory for ODE IVP solvers. In Chapter 2 we review the necessary background on Lyapunov exponents of continuous and discrete time systems and introduce the notions of integral separation from zero and asymptotic contraction. In Chapter 3 we prove a stability result for one-step methods solving an asymptotically contracting, nonautonomous, scalar linear test equation and justify characterizing the numerical stability of a one-step method solving an asymptotically contracting, linear equation of the form (1.2) by d such test equations. In Section 4.1 we use the discrete variation of constants formula combined with the linear stability results of Chapter 3 to prove a stability result for Runge-Kutta methods solving a class of stable nonlinear problems whose linear part satisfies the hypotheses of the linear theory. In Section 4.2 we apply invariant manifold theory so that the analysis of strictly stable linear multistep methods approximating linear problems becomes a corollary of the analysis for one-step methods. In Chapter 5 we develop and test algorithms for selecting step-size for the control of numerical stability using the theory developed in Chapters 3-4. We present the results of some experiments that show how our theory and algorithms can be used to explain and correct the lack of numerical stability of (1.4) and also explore how we can use our theory to characterize the stiffness of a nonlinear problem (1.1) on a time interval. We conclude this work in Chapter 6 with brief summary and some remarks on future work related to the topic of this dissertation.

1.2 Background on the stability theory for time-stepping initial value problem solvers

Time-stepping methods are numerical methods for the numerical approximation of the solution of (1.1) that advance the numerical solution step-by-step in time. They broadly fall into two classes: one-step and multistep methods. One-step methods advance the approximate solution using only the approximate value of the solution from the previous step. A k-step multistep method advances the approximate solution using the approximate values of the solution from k previous steps. The two most important and widely used time-stepping methods are the s-stage Runge-Kutta methods which are a one-step method that take the form

$$\begin{cases} x_{n+1} = x_n + h_n \sum_{j=0}^{s} \tilde{b}_j f(g_n^j, t_n + c_j h_n) \\ g_n^i = x_n + h_n \sum_{j=0}^{s} \tilde{a}_{i,j} f(g_n^j, t_n + c_j h_n), \quad i = 1, \ldots, s \end{cases} \tag{1.6}$$

where the step-sizes h_n are chosen adaptively based on local error tolerances and, the k-step linear multistep methods which take the form

$$\sum_{i=0}^{k} \alpha_i x_{n+i} = h \sum_{i=0}^{k} \beta_i f(x_{n+i}, t_{n+i}) \tag{1.7}$$

where for simplicity the step-size $h > 0$ is fixed. The coefficients $\{\tilde{a}_{i,j}\}_{i,j=0}^{s}$, $\{\tilde{b}_j\}_{j=0}^{s}$, $\{c_j\}_{j=0}^{s}$ in the case (1.6) and $\{\alpha_i\}_{i=0}^{k}$, $\{\beta_i\}_{i=0}^{k}$ in the case of (1.7) are chosen so that the method matches the Taylor series of the exact solution to a certain order or satisfies some other desirable qualities. Often a Runge-Kutta method (1.6) is expressed using its so-called Butcher tableaux

$$\begin{array}{c|c} c & \tilde{A} \\ \hline & \tilde{b}^T \end{array}$$

where $\tilde{A} = (\tilde{a}_{i,j})$, $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_s)^T$, and $c = (c_1, \ldots, c_s)^T$. Both Runge-Kutta and linear multistep methods are types of general linear methods. A k-step and s-stage general linear method with fixed step-size $h > 0$ takes the form

$$\begin{cases} x_i^{(n+1)} = \sum_{j=1}^{k} \alpha_{i,j} x_j^{(n)} + h \sum_{j=1}^{s} \beta_{i,j} f(g_j^{(n)}, t_n + c_j h), \quad i = 1, \ldots, k \\ g_i^{(n)} = \sum_{j=1}^{k} \tilde{a}_{i,j} x_j^{(n)} + h \sum_{j=1}^{s} \tilde{b}_{i,j} f(x_j^{(n)}, t_n + c_j h), \quad i = 1, \ldots, s \end{cases} \tag{1.8}$$

with coefficient matrices denoted by $\mathscr{A} = (\alpha_{i,j})$, $\mathscr{B} = (\beta_{i,j})$, $\tilde{A} = (\tilde{a}_{i,j})$, and $\tilde{B} = (\tilde{b}_{i,j})$. General linear methods provide a common framework for unifying and generalizing the standard theories to Runge-Kutta and linear multistep methods. We do not pursue their analysis in this work, but use them as a way of simplifying the presentation of this section.

The stability theory for numerical methods approximating the solution of (1.1) is motivated by a simple observation common to other fields of numerical analysis which is that over time small errors can become magnified and then subsequently corrupt an approximation. For ODE IVP solvers there are certain problems and methods for which it is possible to construct an approximate solution of (1.1) that initially is locally accurate, that is, it satisfies a specified local error tolerance at each step, but over time these errors accumulate and, for instance, the numerical approximation of a problem that has a bounded and decaying exact solution may become unbounded.

This review of the stability theory for general linear methods closely follows that found in Hairer, E. & Wanner, G. (1991). The stability of a time-stepping method has typically been characterized by determining what, if any, step-size restriction is necessary so that the method produces an asymptotically decaying numerical solution when it is applied to certain type of test problem. The oldest and most well known such test problem is the complex, linear, scalar, test problem

$$\dot{z}(t) = \lambda z(t), \quad \lambda \in \mathbb{C}. \tag{1.9}$$

The test problem (1.9) is meant to serve as a caricature of the numerical stability of a time-stepping method solving a linear problem $\dot{x}(t) = Ax(t)$ where $A \in \mathbb{R}^{d\times d}$. From the change of variables $x = Pz$ where P is such that $J = P^{-1}AP$ is the Jordan form of A, it follows that the stability of the exact solution of $\dot{x} = Ax$ is governed by the stability of d linear complex scalar problems of the form $\dot{z}_i(t) = \lambda_i z(t)$ where $\lambda_i \in \mathbb{C}$ is an eigenvalue of A. Since the stability of solutions of a general linear method (1.8) is preserved by a linear time-independent change of variables $x_n = Pz_n$, it follows that the numerical stability of Runge-Kutta and linear multistep methods solving a linear problem $\dot{x} = Ax$ is characterized by stability of the method applied to d complex linear scalar test problems. Classically, this observation led to the development of linear stability domains and A-stability theory.

Definition 1. *The linear stability domain of a general linear method* (1.8) *is the set of all* $z = h\lambda \in \mathbb{C}$ *such that if the method is applied to solve* (1.9) *using the step-size* $h > 0$*, then the numerical solution* $\{z_n\}_{n=0}^{\infty}$ *satisfies that* $z_n \to 0$ *as* $n \to \infty$*.*

Definition 2. *A general linear method is A-stable if its linear stability domain contains the left half complex plane* $\mathbb{C}^- := \{z \in \mathbb{C} : Re(z) < 0\}$.

Linear stability domains characterize the steps-size restriction due to stability of a general linear method solving either autonomous linear problems or autonomous nonlinear problems with an initial condition nearby a fixed point. For the numerical stability of nonlinear and nonautonomous problems there have been several classes of test problems that have been proposed. One such test problem is a d dimensional nonlinear ODE $\dot{x} = f(x,t)$ where $f(x,t)$ satisfies a one-sided Lipschitz condition

$$\langle f(x,t) - f(y,t), x - y\rangle \leq 0. \tag{1.10}$$

where $\langle\cdot\rangle$ is some inner product that induces a norm $\|\cdot\|$ on $\mathbb{R}^d$. If f satisfies the estimate (1.10), then $\dot{x} = f(x,t)$ is a dissipative ODE and given any two initial conditions x_0 and y_0 and $s \leq t$, then the solutions $x(\cdot;x_0)$ and $x(\cdot,y_0)$ through x_0 and y_0 will satisfy the estimate

$$\|x(t;x_0) - x(t;y_0)\| \leq \|x(s;x_0) - x(s;y_0)\| \tag{1.11}$$

For a symmetric, positive definite matrix $G = (g_{i,j})$ in $\mathbb{R}^{d\times d}$ we let the G-norm $\|\cdot\|_G$ on $\mathbb{R}^d$ be defined by

$$\|u\|_G^2 := \sum_{i=1}^{d}\sum_{j=1}^{d} g_{i,j}\langle u_i, u_j\rangle. \tag{1.12}$$

A numerical method is called G-stable if whenever f satisfies (1.10), then there exists a real symmetric positive definite matrix G so that the numerical solutions x_n and y_n of $\dot{x}(t) = f(x(t),t)$ using the initial conditions x_0 and y_0 respectively satisfy that for any fixed step-size $h > 0$ we have

$$\|x_{n+1} - y_{n+1}\|_G \leq \|x_n - y_n\|_G, \quad n \geq 0.$$

A sufficient condition in terms of the coefficients of (1.8) for a method to be G-stable is that it is algebraically stable, which means that there is a real symmetric positive definite matrix G and a real, non-negative definite matrix D so that the matrix M defined as

$$M = \begin{bmatrix} G - \mathscr{A}^T G \mathscr{A} & \tilde{A}^T D - \mathscr{A}^T G \mathscr{B} \\ D\tilde{A} - \mathscr{B}^T G \mathscr{A} & D\tilde{B} + \tilde{B}^T D - \mathscr{B}^T G \mathscr{B} \end{bmatrix}$$

is non-negative definite. The stability of general linear methods solving problems that satisfy (1.10) can often be characterized by a linear, nonautonomous, scalar, complex test problem

$$\dot{z}(t) = \lambda(t) z(t), \quad \lambda(t) \in \mathbb{C}, \quad \mathrm{Re}(\lambda(t) \leq 0. \tag{1.13}$$

Solving (1.13) with the general linear method (1.8) produces a numerical solution $\{z_n\}_{n=0}^{\infty}$ that satisfies a linear difference equation

$$z_{n+1} = S(Z)z_n, \quad Z = h(\lambda(t_n + c_1 h), \ldots, \lambda(t_n + c_s h))^T \tag{1.14}$$

A general linear method (1.8) is said to be AN-stable if there exists a real, symmetric, positive definite G so that $\|S(Z)u\|_G \le \|u\|_G$ for every $Z = (z_1, \ldots, z_s)^T \in \mathbb{C}^s$ with $\mathrm{Re}(z_i) \le 0$ for $i = 1, \ldots, s$ and $z_i = z_j$ whenever $c_i = c_j$. A method (1.8) that is AN-stable will produce a bounded and/or decaying numerical solution to the test problem $\dot{z}(t) = \lambda(t)z(t)$ where $\lambda(t) \in \mathbb{C}$ with $\mathrm{Re}(\lambda(t) \le 0$ for any step-size $h > 0$. To justify using the test problem (1.13) and AN-stability theory to characterize the stability of the general linear method (1.8) applied to solve the problem (1.1) where f satisfies the condition (1.10) we introduce the following terminology.

Definition 3. *General linear methods for which there exists i and j such that $c_i = c_j$ are referred to as non-confluent. General linear methods for which there exists $\xi \in \mathbb{R}^k$ such that $\mathscr{A}\xi = \xi$ and $\tilde{A}\xi = (1, \ldots, 1)^T$ are referred to as preconsistent.*

Theorem 1. *The following implications hold for all general linear methods*

$$\textit{algebraic stability} \Rightarrow \textit{G-stability} \Rightarrow \textit{AN-stability} \Rightarrow \textit{A-stability}$$

Furthermore, if the method (1.8) *is preconsistent and non-confluent, then AN-stability, G-stability, and algebraic stability are all equivalent.*

□

The test problems (1.13) and $\dot{x} = f(x,t)$ where $f(x,t)$ satisfies (1.10) correspond to the case where the exact solution is uniformly decaying. There are many classical dissipative problems, such as the Lorenz 1963 system which first appeared in Lorenz (1963) and the Van der Pol oscillator which appears in Van der Pol, B. (1927), that do not satisfy one-sided Lipschitz conditions. For such problems, the Lyapunov stability of the exact solution is governed by a general nonautonomous linear equation $\dot{x}(t) = A(t)x(t)$ and solutions may not be uniformly decaying. This work seeks to find conditions so that the numerical stability of one-step methods and strictly stable linear multistep methods solving such a linear problem can be characterized by an asymptotically decaying scalar test problem of the form $\dot{x}(t) = \lambda(t)x(t)$ and then find the step-size restriction under which the method preserves asymptotic decay.

Chapter 2

Lyapunov exponents

This chapter is a review of the background on Lyapunov exponents of continuous time and discrete time systems necessary for the stability theory for one-step methods we develop in Chapter 3. Additionally, we recall the concepts of integral separation from zero and asymptotic contraction that are useful in estimating Lyapunov exponents and in quantifying asymptotic decay.

2.1 Continuous time systems

In this section we review the necessary background on Lyapunov exponents for continuous time systems. For a detailed account of the general theory of Lyapunov exponents, see Adrianova (1995) and for references on the continuity and numerical approximation of Lyapunov exponents see Dieci, L. & Van Vleck, E.S. (2002, 2003, 2006, 2007). Consider a linear nonautonomous ODE

$$\dot{x}(t) = A(t)x(t), \quad t > t_0 \tag{2.1}$$

where $A : (t_0, \infty) \to \mathbb{R}^{d\times d}$ is bounded and continuous. We discuss how to compute the Lyapunov exponents of (2.1) without constructing fundamental matrix solutions. For systems $\dot{y}(t) = B(t)y(t)$ where $B(t)$ is upper triangular, the Lyapunov exponents generically are given in terms of the diagonal elements of the coefficient matrix $B(t)$.

Theorem 2. *(Theorem 5.1 in Dieci, L. & Van Vleck, E.S. (2007)) Consider* $\dot{y}(t) = B(t)y(t)$ *where* $B : (t_0, \infty) \to \mathbb{R}^{d \times d}$ *is bounded, continuous, and upper triangular. Suppose that for every* $i < j$ *one of the two following conditions hold:*

1. $B_{i,i}$ *and* $B_{j,j}$ *are integrally separated, that is, there exists* $a_{i,j} > 0$ *and* $b_{i,,j} \in \mathbb{R}$ *so that if* $t \geq s > t_0$*, then*
$$\int_s^t B_{i,i}(\tau)d\tau - B_{j,j}(\tau)d\tau \geq a_{i,j}(t-s) - b_{i,j}. \tag{2.2}$$

2. *For every* $\varepsilon > 0$ *there exists* $M_{i,j}(\varepsilon) > 0$ *so that if* $t \geq s > t_0$*, then*
$$\left| \int_s^t B_{i,i}(\tau) - B_{j,j}(\tau)d\tau \right| \leq M_{i,j} + \varepsilon(t-s). \tag{2.3}$$

Then the Lyapunov exponents $\mu_1, \ldots, \mu_d$ *of* $\dot{y}(t) = B(t)y(t)$ *are continuous and given by the formulas*
$$\mu_i = \limsup_{t \to \infty} \frac{1}{t - t_0} \int_{t_0}^t B_{i,i}(\tau)d\tau, \quad i = 1, \ldots, d. \tag{2.4}$$

A system $\dot{y}(t) = B(t)y(t)$ satisfying the hypotheses of Theorem 2 is referred to as a system that has an integral separation structure and the coefficient matrix $B(t)$ is said to have an integral separation structure. The system $\dot{y}(t) = B(t)y(t)$ is integrally separated if each pair of diagonal elements of $B(t)$ are integrally separated. Integral separation is a generic property for systems of the form (2.1) in the same way that generically $d \times d$ real-valued matrices $M \in \mathbb{R}^{d \times d}$ of autonomous systems $\dot{x} = Mx$ have distinct eigenvalues, see page 21 of Palmer (1979). Integrally separated systems have distinct Lyapunov exponents that are continuous with respect to perturbations in the entries of the coefficient matrix $A(t)$.

For general systems (2.1) where $A(t)$ is not upper triangular, we can construct a time-dependent change of variables that transforms the original problem to one with an upper triangular coefficient matrix. Consider a fundamental matrix solution $X(t)$ of (2.1) and let $X(t) = Q(t)R(t)$ be the unique continuous QR factorization of $X(t)$ where $Q(t)$ is orthogonal and $R(t)$ is upper triangular with positive diagonal entries. Then $x(t) = Q(t)y(t)$ is a Lyapunov transformation and $\dot{y}(t) = B(t)y(t)$

where $B(t) = Q(t)^T A(t)Q(t) - Q(t)^T \dot{Q}(t)$ is upper triangular. Furthermore, it can be shown that $Q(t)$ satisfies the differential equation

$$\dot{Q}(t) = Q(t)S(Q(t),A(t)), \quad S(Q,A)_{ij} = \begin{cases} (Q^T AQ)_{i,j}, & i > j \\ 0, & i = j \\ -(Q^T AQ)_{i,j}, & i < j \end{cases} \tag{2.5}$$

If $B(t)$ is such that $B(t) = Q^T(t)A(t)Q(t) - Q^T(t)\dot{Q}(t)$ where $Q(t)$ is orthogonal for all t and satisfies (2.5) for some initial condition $Q(t_0)$, then we refer to the system $\dot{y}(t) = B(t)y(t)$ as a corresponding upper triangular system to (2.1). Since $x(t) = Q(t)y(t)$ is a Lyapunov transformation, every upper triangular system corresponding to (2.1) has the same Lyapunov exponents as (2.1).

Generically, a corresponding upper triangular system to (2.1) has an integral separation structure and thus it is a natural assumption to make. Theorem 2 is useful since allows us to consider problems that have continuous and possibly indistinct Lyapunov exponents and the Lyapunov exponents are given as formulas in terms of the coefficient matrix $B(t)$. We use Theorem 2 and the hypothesis that (2.1) has a corresponding upper triangular system with an integral separation structure as the basis for the main structural assumption we place on the linear problem (2.1) that we use in our numerical stability analysis in Section 3.

2.2 Discrete time systems

In this section we review necessary background on Lyapunov exponents of discrete time systems. Consider a nonautonomous linear difference equation of the form

$$x_{n+1} = \Phi^A(t_n)x_n, \tag{2.6}$$

where $x_n \in \mathbb{R}^d$, $\Phi^A(t_n) \equiv \Phi^A(n) \in \mathbb{R}^{d\times d}$ is a bounded sequence of invertible matrices, and $\{t_n\}_{n=0}^{\infty}$ is a sequence such that there exists $0 < h_{\min} \leq h_{\max} < \infty$ so that $h_{\min} \leq t_{n+1} - t_n \leq h_{\max}$ for all

$n \geq 0$. We refer to such a sequence $\{t_n\}_{n=0}^{\infty}$ as a time sequence and remark that the system (2.6) depends on time-sequence that is used. We have the following definition of integral separation structure in discrete time analogous to that found in Badawy, M. & Van Vleck, E.S. (2012).

Definition 1. *Consider $y_{n+1} = \Phi^B(n)y_n$ where $\Phi^B(n) \in \mathbb{R}^{d\times d}$ is bounded and upper triangular and suppose that the diagonal entries $\Phi^B_{i,i}(n)$ are all positive and have uniformly bounded inverses. Suppose that for every $i < j$ one of the two following conditions hold:*

1. *$\Phi^B_{i,i}(n)$ and $\Phi^B_{j,j}(n)$ are discretely integrally separated, that is, there exists $b_{i,j} \in (0,1]$ and $a_{i,j} > 0$ so that if $n \geq m$, then*

$$\prod_{j=m}^{n} \Phi^B_{i,i}(j)(\Phi^B_{i+1,i+1}(j))^{-1} \geq b_{i,j}e^{a_{i,j}(t_n-t_m)} \tag{2.7}$$

2. *$\Phi^B_{i,i}(n)$ and $\Phi^B_{j,j}(n)$ satisfy that for every $\varepsilon > 0$, there exists $M_{i,j} > 0$ and $h^* > 0$ so that if $n \geq m$ and $h_{max} \leq h^*$, then*

$$|\prod_{k=m}^{n} \Phi^B_{i,i}(k)(\Phi^B_{j,j}(k))^{-1}| \leq e^{M_{i,j}+\varepsilon(t_n-t_m)}. \tag{2.8}$$

We refer to such a system as a system with an approximate discrete integral separation structure and say that $\Phi^B(n)$ has an approximate discrete integral separation structure.

The following theorem follows from the results proved in Badawy, M. & Van Vleck, E.S. (2012); Dieci, L. & Van Vleck, E.S. (2007); Dieci & Van Vleck, E.S. (2005).

Theorem 3. *If the system $y_{n+1} = \Phi^B(n)y_n$ is a system with an approximate discrete integral separation structure, then for every $\varepsilon >$ there exists $h^* > 0$ so that if $h_{max} \leq h^*$, then the discrete Lyapunov exponents satisfy that*

$$|\mu_i - \limsup_{n\to\infty} \frac{1}{t_n - t_0} \sum_{j=0}^{n} \ln(\Phi^B_{i,i}(j))| < \varepsilon, \quad 1 = 1,\ldots,d.$$

If the diagonal elements of $y_{n+1} = \Phi^B(n)y_n$ are all discretely integrally separated, then there is no restriction on h^* and we can take $\varepsilon = 0$.

We now discuss how an approximate discrete integral separation structure is preserved under perturbations of the coefficient matrix. Consider the perturbed system

$$z_{n+1} = (\Phi^A(n) + F_n)z_n. \tag{2.9}$$

Additionally assume that both $\Phi^A(n)$ and $\Phi^A(n) + F_n$ are bounded and invertible for all $n \geq 0$. Fix $Q_0 = \overline{Q}_0$ orthogonal and inductively construct QR factorizations $\Phi^A(n)Q_n = Q_{n+1}R^A(n)$ and $(\Phi^A(n) + F_n)\overline{Q}_n = \overline{Q}_{n+1}\overline{R}^A(n)$ where Q_n and $\overline{Q}_n$ are orthogonal and $R^A(n)$ and $\overline{R}^A(n)$ are upper triangular with positive diagonal entries. We shall refer to $v_{n+1} = R^A(n)v_n$ as a corresponding upper triangular system to (2.6).

Define $E_n := \overline{Q}^T_{n+1}F_n\overline{Q}_n$ and suppose that $\|F_n\| = \|E_n\|$ is small for all $n \geq 0$. Then we would expect that $R^A(n) \approx \overline{R}^A(n)$ and $Q_n \approx \overline{Q}_n$ for n sufficiently small. The following theorem, which follows from the estimates in the proof Theorem 7.7 in Badawy, M. & Van Vleck, E.S. (2012) and Theorem 4.1 in Van Vleck, E.S. (2010), says that for systems (2.6) where the corresponding upper triangular factor $R^A(n)$ has an approximate discrete integral separation structure, that this is indeed the case, and in fact, there are global uniform bounds on the differences $Q_n - \overline{Q}_n$ and $R^A(n) - \overline{R}^A(n)$

Theorem 4. *Suppose that the discrete QR process for both of the systems* (2.6) *and* (2.9) *is well-defined and suppose that $\tilde{R}^A(n)$ has an approximate discrete integral separation structure. If $F := sup_{n\geq 0}\|F_n\|$ is sufficiently small, then there exists an $h^* > 0$ so that if $h_{max} \leq h^*$, then there exists an orthogonal sequence of matrices $\{\tilde{Q}_n\}_{n=0}^{\infty}$ and $K > 0$ such that*

$$\tilde{Q}_{n+1}R^A(n) = [\overline{R}^A(n) + E_n]\tilde{Q}_n$$

and $\|\tilde{Q}_n - I\| \leq K\|E_n\| = K\|F_n\|$. If the diagonal elements of $\tilde{R}^A(n)$ are all discretely integrally *separated, then there is no restriction on $h^* > 0$.*

Using the estimate in Badawy, M. & Van Vleck, E.S. (2012) we can actually approximate how small $F > 0$ must be taken for the conclusion of Theorem 4 to hold. In Section 3.2 we apply Theorem 4 to relate the numerical stability of the numerical solution of the system (2.1) and a corresponding upper triangular system.

2.3 Integral separation from zero and asymptotic contraction

In this section we define the notions of asymptotic contraction and integral separation from zero.

Definition 2. *We say that the system $\dot{x}(t) = \lambda(t)x(t)$ is integrally separated from zero if $\lambda : (t_0, \infty) \to \mathbb{R}$ is bounded and continuous and there exists $L_1, L_2 \in \mathbb{R}$ and $D_1, D_2 \in \mathbb{R}$ with $D_1 \leq D_2$ and $L_1 \leq L_2$ so that if $t \geq s > t_0$, then*

$$D_1 + L_1(t-s) \leq \int_s^t \lambda(\tau)d\tau \leq D_2 + L_2(t-s). \tag{2.10}$$

We say that the system $\dot{x}(t) = \lambda(t)x(t)$ is asymptotically contracting if $L_2 < 0$.

If $\lambda(t)$ is integrally separated from zero and satisfies the estimate (2.10), then the Lyapunov exponent of $\dot{x}(t) = \lambda(t)x(t)$ lies in the interval $[L_1, L_2]$ and if $\lambda(t)$ is asymptotically contracting, then the Lypaunov exponent is negative. We generalize Definition 2 to systems of the form (2.1) as follows.

Definition 3. *We say that* (2.1) *is integrally separated from zero if there exists a corresponding upper triangular system $\dot{y}(t) = B(t)y(t)$ has an integral separation structure and each of the d diagonal systems $\dot{y}_i(t) = B_{i,i}(t)y_i(t)$ are integrally separated from zero. We say that* (2.1) *is asymptotically contracting if in addition the systems $\dot{y}_i(t) = B_{i,i}(t)y_i(t)$ are all asymptotically contracting.*

Suppose that (2.1) is integrally separated from zero and the diagonal elements $B_{i,i}(t)$ of a corresponding upper triangular system $\dot{y}(t) = B(t)y(t)$ satisfy that for $i = 1, \ldots, d$ there exists $L_{i,1} \leq L_{i,2}$

and $D_{i,1} \leq D_{i,2}$ so that if $t \geq s > t_0$, then

$$D_{i,1} + L_{i,1}(t-s) \leq \int_s^t B_{i,i}(\tau)d\tau \leq D_{i,2} + L_{i,2}(t-s).$$

It follows that the Lyapunov exponents $\mu_1, \ldots, \mu_d$ of (2.1) satisfy that $\mu_i \in [L_{i,1}, L_{i,2}]$ for $i = 1, \ldots, d$ and that the Lyapunov exponents are all negative if (2.1) is asymptotically contracting and $L_{i,2} < 0$ for $i = 1, \ldots, d$. Asymptotic contraction gives us a way of establishing uniform estimates on the growth and decay rates of systems with an integral separation structure.

We close this section by remarking that we can analogously define asymptotic contraction and integral separation from zero for discrete time systems (2.6) using the same types of estimates as in Section 2.2.

Chapter 3

Stability of one-step methods

In this chapter we analyze the stability of one-step methods solving asymptotically contracting linear problems of the form (2.1). One-step methods solving nonautonomous linear differential equations of the form (2.1) take the form of a nonautonomous linear difference equation $x_{n+1} = \Phi^A(t_n)x_n$ where $\Phi^A(t_n) \in \mathbb{R}^{d\times d}$. The matrix sequence $\Phi^A(t_n) \equiv \Phi^A(t_n, h_n) \equiv \Phi^A(n)$ depends on the current time t_n and $A(t_n)$. Throughout, we let $h_{\max} := \sup_{n\geq 0} h_n$ and $h_{\min} := \inf_{n\geq 0} h_n$ where $h_n := t_{n+1} - t_n$ and assume that $h_{\min} > 0$ and $h_{\max} < \infty$. This chapter is organized as follows. First, in Section 3.1, we find conditions on the maximum allowable step-size so that the numerical solution of an asymptotically contracting scalar test problem is discretely asymptotically contracting. Subsequently, in Section 3.2, we determine the maximum allowable step-size so that the numerical solution an asymptotically contracting system (2.1) using a one-step method is discretely asymptotically contracting and then justify using d asymptotically contracting scalar test problems to characterize the maximum allowable step-size.

3.1 Stability analysis for an asymptotically contracting scalar test problem

In this section we consider the numerical stability of a scalar test problem

$$\dot{x}(t) = \lambda(t)x(t) \tag{3.1}$$

solved with a one-step method $\mathscr{M}$. We assume that $\lambda : (t_0, \infty) \to \mathbb{R}$ is asymptotically contracting and satisfies the estimates

$$D_1 + L_1(t-s) \int_s^t \lambda(\tau)d\tau \leq D_2 + L_2(t-s), \quad t \geq s > t_0 \tag{3.2}$$

where $L_1 \leq L_2 < 0$ and $D_1 \leq D_2$. The numerical solution of (3.1) with $\mathscr{M}$ using a sequence of step-sizes $\{h_n\}_{n=0}^{\infty}$ takes the form

$$x_{n+1} = \Phi^{\lambda}(n)x_n.$$

The test problem should be thought of as one of the problems $\dot{y}_i(t) = B_{i,i}(t)y_i(t)$ where $B_{i,i}(t)$ is a diagonal element of an upper triangular coefficient matrix $B(t)$ of an upper triangular system corresponding to (2.1). In Section 3.2 we rigorously justify this intuition when $B(t)$ has an integral separation structure.

We remark that test problems of the form (3.1) already appear in the literature of AN-stability theory. Our analysis differs from that found in the literature on AN-stability in two main ways. The first way, as we shall show in 3.2, we have a method for computing the test problem for a given system (2.1) and our method justifies considering only the case where $\lambda(t)$ is real-valued as opposed to complex valued in AN-stability theory. The second way of analysis differs from AN-stability theory is that we assume only that $\lambda(t)$ is asymptotically contracting rather than nonpositive for all $t > t_0$ which allows for $\lambda(t)$ such as $\lambda(t) = \cos(t) - 1/2$ that take on positive values for infinitely many t even though the solution is asymptotically contracting. This apparently minor difference turns out to have an substantial impact on the analysis since, as we show in the proof of the following theorem, there are no Runge-Kutta methods that produce a bounded or decaying solution to every problem of the form (3.1) that satisfies an estimate (3.2) without the introduction of a step-size restriction.

Theorem 5. *Given any convergent Runge-Kutta method $\mathscr{M}$, any step-size $h > 0$, and $L_2 < 0$ we can find D_1, D_2, and L_1 so that $\lambda(t)$ satisfies* (3.2) *and the numerical solution of* (3.1) *using $\mathscr{M}$ with fixed step-size $h > 0$ and initial condition $x(t_0) = x_0 \neq 0$ becomes unbounded.*

Proof. Let $R(\cdot)$ be the classical stability function of $\mathscr{M}$ and let $h > 0$. $R(\cdot)$ is a Padé approximation to the exponential and therefore there exists $\delta > 0$ so that $R(x) > 1$ for all $x \in \mathbb{R}$ with $0 < x < \delta$. Let $\lambda(t) = D\cos(2\pi t/h) + L_2$ where $D > -L_2$ and $h(D + L_2) < \delta$. Then $\lambda(t)$ satisfies (3.2) with $L_1 = L_2, D_1 = -hD/\pi$ and $D_2 = hD/\pi$. The numerical solution of (3.1) with the method $\mathscr{M}$ using the fixed step-size h is $x_{n+1} = R(h(D+L_2))x_n$. Since $0 < h(D+L_2) < \delta$ implies $R(h(D+L_2)) > 1$ and $x_0 \neq 0$ it follows that $|x_n| \to \infty$ as $n \to \infty$. □

The $\lambda(t)$ constructed in Theorem 5 shows that time-dependent oscillations in the coefficient function $\lambda(t)$ may trigger instabilities in the numerical solution. Such oscillations produce an inherent step-size restriction in any Runga-Kutta method for solving initial value problems and may occur in the presence of "small" exponential growth and decay rates; these oscillations may not be damped out by normal stiff integrators. Theorem 5 is the main reason that we use error estimates for stability control since for Runge-Kutta methods there does not seem to be a straightforward way of controlling the stability of an asymptotically contracting, nonautonomous, linear, scalar test problem without some type of error control.

Although Theorem 5 paints a pessimistic picture for numerically preserving the asymptotic decay of time-dependent problems, the following theorem says that the next best thing we would hope for is true, that for all sufficiently small step-sizes we can guarantee that a one-step method with local truncation error of order $p \geq 1$ is discretely asymptotically contracting when applied to solve the problem (3.1).

Theorem 6. *Suppose that $\mathscr{M}$ has local truncation error of order $p \geq 1$. Then there exists $h^* > 0$ so that if $h_{max} \leq h^*$, then the numerical solution $x_{n+1} = \Phi^\lambda(n)x_n$ is discretely asymptotically contracting.*

Proof. We can write $\Phi^\lambda(n) = \exp\left(\int_{t_n}^{t_{n+1}} \lambda(\tau)d\tau\right) + E_n \equiv I_n + E_n$ so that

$$\prod_{j=n}^{m} \Phi^\lambda(j) = \prod_{j=n}^{m}(1 + E_j I_j^{-1}) \prod_{j=n}^{m} I_j, \quad n \geq m > 0.$$

If there exists $K > 0$ so that $|E_n I_n^{-1}| < Kh_n < 1/2$ for all $n \geq 0$, then the estimate (3.2) implies that if $n \geq m \geq 0$, then

$$e^{D_1 + (L_1 - 2K)(t_n - t_m)} \leq \prod_{j=n}^{m} \Phi^\lambda(j) \leq e^{D_2 + (L_2 + K)(t_n - t_m)}.$$

To show that $\Phi^\lambda(n)$ is discretely asymptotically contracting it suffices to show that we can find $h^* > 0$ so that if $h_{\max} \leq h^*$, then there exists $K > 0$ so that $K + L_2 < 0$ and $|E_n I_n^{-1}| < Kh_n < 1/2$ for all $n \geq 0$. Since the method $\mathscr{M}$ is of order $p \geq 1$, there exists $\tilde{h} > 0$ so that if $h_{\max} \leq \tilde{h}$, then for all $n \geq 0$ we have $E_n = T_n h_n^{p+1}$ and $|T_n| \leq C$ for some $C > 0$. If $h_{\max} \leq \tilde{h}$,then it follows from (3.2) that $|E_n I_n^{-1}| \leq Ch_n^{p+1} e^{-D_1 - L_1 h_n}$. We can then choose $h^* > 0$ with $h^* \leq \tilde{h}$ so that if $h_{\max} \leq h^*$, then $Ch_n^p e^{-D_1 - L_1 h_n} < \min\{-L_2, 1/2\}$.

□

The term $E_n I_n^{-1}$ that appears in the proof of Theorem 6 is the product of a stability term I_n, and an accuracy term E_n. The term I_n^{-1} provides a measure of stiffness for the solution of (3.1) in the interval $[t_n, t_{n+1}]$. If $\lambda(t)$ is negative and has a large magnitude on $[t_n, t_{n+1}]$, then I_n^{-1} will be very large and the step-size must be taken much smaller to compensate for this. We explore this intuition more in the experiments in Section 5.3.

We close this section by discussing an alternative to restricting the step-size to guarantee that the numerical solution (3.1) is discretely asymptotically contracting. The alternative approach is to allow the coefficients of the one-step method to vary at each time step and then selecting the values for these coefficients in a judicious way. Such variable coefficient methods appear in the literature (see e.g. Lambert (1974); Wambecq (1978); Hairer (1980); Calvo, M. & Quemada, M. Mar (1982); Verwer, J.G. & Dekeker, K. (1983)) under the names of rational Runge-Kutta formulas and arise in the contexts of monotone and conservative methods and also in preserving the orthogonality of $Q(t)$ in the numerical integration of (2.5), Dieci, L. & Van Vleck, E.S. (1995). We do note pursue the analysis of such methods in this work, but we remark that they may provide a viable alternative to using error control to guarantee asymptotic decay of numerical solutions of (3.1).

3.2 Justification for the test problem

Fix some one-step method $\mathscr{M}$ with local truncation error of order $p \geq 1$. In this section we justify using d asymptotically contracting, nonautonomous, linear, scalar test problems of the form $\dot{x}_i(t) = \lambda_i(t)x_i(t)$ to characterize the numerical stability of $\mathscr{M}$ applied to solve (2.1). In addition we show how to compute the coefficients $\lambda_1(t), \ldots, \lambda_d(t)$ of the test problems for a given (2.1). For the remainder of this section make the following assumption on (2.1).

Assumption 1. *Assume that the coefficient matrix $A(t)$ in* (2.1) *is bounded and $p+1$ times differentiable. Suppose that there is a fundamental matrix solution $X(t)$ with QR factorization $X(t) = Q(t)R(t)$ so that under the change of variables $x(t) = Q(t)y(t)$ the corresponding upper triangular problem*

$$\dot{y}(t) = B(t)y(t) \tag{3.3}$$

is asymptotically contracting with the diagonal elements satisfying the estimates

$$D_{1,i} + L_{1,i}(t-s) \leq \int_s^t B_{i,i}(\tau) \leq D_{2,i} + L_{2,i}(t-s) \tag{3.4}$$

with $D_{1,i} \leq D_{2,i}$ and $L_{1,i} \leq L_{2,i} < 0$ for $i = 1, \ldots, d$ and an integral separation structure defined by the following estimates. For $i < j$ we either have $B_{i,i}$ and $B_{j,j}$ are integrally separated with

$$\int_s^t B_{i,i}(\tau)d\tau - B_{j,j}(\tau)d\tau \geq a_{i,j}(t-s) - b_{i,j} \tag{3.5}$$

for all $t \geq s > t_0$ where $a_{i,j} > 0$ and $b_{i,j} \in \mathbb{R}$ or if $B_{i,i}$ and $B_{j,j}$ are not integrally separated, then for every $\varepsilon > 0$, there exists $M_{i,j}(\varepsilon) > 0$ so that if $t \geq s > t_0$, then

$$\left| \int_s^t B_{i,i}(\tau) - B_{j,j}(\tau) d\tau \right| \leq M_{i,j}(\varepsilon) + \varepsilon(t-s). \tag{3.6}$$

Assumption 1 implies that the Lyapunov exponents of (2.1) are all negative and can be computed from formula (2.4). We remark that in all this section we assume that $A(t)$ and $B(t)$ are known exactly. In practice, $A(t)$ and $Q(t)$ are approximated simultaneously from an approximate nonlinear trajectory and then used to form an approximate $B(t)$. The additional issues that arise from this are studied in more detail in Dieci & Van Vleck, E.S. (2005).

Let $x_{n+1} = \Phi^A(n)x_n$ denote the numerical solution of (2.1) using the method $\mathscr{M}$ with the time sequence $\{t_n\}_{n=0}^{\infty}$ and let $y_{n+1} = \Phi^B(n)y_n$ denote numerical solution of (3.3) using the method $\mathscr{M}$ with the same time sequence and $x_0 = Q(t_0)y_0$. We shall assume that each $h_{\max} > 0$ is always so small that $\Phi^A(n)$ and $\Phi^B(n)$ are both bounded and invertible for all $n \geq 0$. The matrices $\Phi^B(n)$ are upper triangular since $B(t)$ is upper triangular and each diagonal entry $\Phi^B_{j,j}(n)$ is such that $y^j_{n+1} = \Phi^B_{j,j}(n)y^j_n$ is the numerical solution of the scalar problem $\dot{y}_j(t) = B_{j,j}(t)y_j(t)$ using $\mathscr{M}$ with the same time-sequence.

Since $\Phi^A(n)$ is invertible we can inductively construct unique QR factorizations of $\Phi^A(n)Q_n$ as $\Phi^A(n)Q_n = Q_{n+1}R^A(n)$ where each Q_n is orthogonal, $Q_0 = Q(t_0)$, and $R^A(n)$ is upper triangular with positive diagonal entries. The stability of the zero solution of $x_{n+1} = \Phi^A(n)x_n$ is equivalent to the the stability of the zero solution of the upper triangular system $z_{n+1} = R^A(n)z_n$ since $x_n = Q_n z_n$ and Q_n is orthogonal and therefore defines a discrete Lyapunov transformation. The essence of our theory is to determine conditions on the maximum allowable step-size so that $R^A(n)$ is discretely asymptotically contracting by estimating the difference between the diagonal entries of $R^A(n)$ and $\Phi^B(n)$.

We factor the fundamental matrix solutions $X(t)$ of (2.1) and $R(t)$ of (3.3) from Assumption 1 on the time sequence $\{t_n\}_{n=0}^{\infty}$ to establish a relation between these factorizations and the local approximation properties of $\Phi^A(n)$ and $\Phi^B(n)$. Consider the sequence of matrix IVPs:

$$\begin{cases} \dot{\Psi}(t) = A(t)\Psi(t), \quad t > t_n \\ \quad \Psi(t_n) = I_{d\times d} \end{cases} \tag{3.7}$$

and let $X(t,t_n)$ be the unique solution of (3.7) for all $n \geq 0$. Then

$$X(t_n) = X(t_n,t_{n-1})\cdot\ldots\cdot X(t_1,t_0)X(t_0).$$

Similarly for corresponding upper triangular system $\dot{y}(t) = B(t)y(t)$ consider the sequence of matrix IVPs:

$$\begin{cases} \dot{\Phi}(t) = B(t)\Phi(t), \quad t > t_n \\ \quad \Phi(t_n) = I_d \end{cases} \tag{3.8}$$

with the unique exact solution $R(t,t_n)$. We can express $R(t_n)$ as

$$R(t_n) = R(t_n,t_{n-1})\cdot\ldots\cdot R(t_1,t_0)R(t_0).$$

Notice that we have $X(t,t_n) = Q(t)R(t,t_n)Q(t_n)^T$ for $n \geq 0$. Consider the local error expressions

$$\Phi^A(n) = X(t_{n+1},t_n) + E_n^A, \quad \Phi^B(n) = R(t_{n+1},t_n) + E_n^B. \tag{3.9}$$

Combining (**??**) with the relation $X(t,t_n) = Q(t_{n+1})R(t,t_n)Q(t_n)^T$ implies that

$$\Phi^B(n) = Q(t_{n+1})^T(\Phi_n^A + F_n)Q(t_n) \tag{3.10}$$

where $F_n = -E_n^A + Q(t_{n+1})E_n^B Q(t_n)^T$. So, if the diagonal entries of $\Phi^B(n)$ are all positive, then it is the unique upper triangular factor of the discrete QR process applied to the unperturbed system

$$y_{n+1} = \tilde{\Phi}^A(n)y_n \tag{3.11}$$

where $\tilde{\Phi}^A(n) := \Phi^A(n) + F_n$ with the orthogonal factor given by $Q(t_n)$ and the corresponding perturbed system

$$x_{n+1} = (\tilde{\Phi}^A(n) + \tilde{F}_n)x_n \tag{3.12}$$

where $\tilde{F}_n := \Phi^A(n) - \tilde{\Phi}^A(n) = -F_n$. Before we can apply Theorem 4 to estimate the difference $R^A(n) - \Phi^B(n)$ we show that we can always choose $h_{\max} > 0$ so that Φ^B_n has an approximate discrete integral separation structure.

Lemma 1. *The system $y_{n+1} = \Phi^B(n)y_n$ has an approximate discrete integral separation structure.*

Proof. Express $\Phi^B_{i,i}(n)$ in the form

$$\Phi^B_{i,i}(n) = \exp\left(\int_{t_n}^{t_{n+1}} B_{i,i}(\tau)d\tau\right) + E^i_n \equiv I^i_n + E^i_n, \quad i = 1, \ldots, d$$

If $n \geq m$, then for $i < j$ we have

$$\begin{aligned}\prod_{k=m}^{n} \Phi^B_{i,i}(k)(\Phi^B_{j,j}(k))^{-1} &= \left[\prod_{k=m}^{n} I^i_k(I^j_k)^{-1}\right]\left[\prod_{k=m}^{n} \frac{1 + E^i_k(I^i_k)^{-1}}{1 + E^j_k(I^j_k)^{-1}}\right] \\ &= e^{\int_{t_m}^{t_n} B_{i,i}(\tau) - B_{j,j}(\tau)d\tau}\left[\prod_{k=m}^{n} \frac{1 + E^i_k(I^i_k)^{-1}}{1 + E^j_k(I^j_k)^{-1}}\right].\end{aligned}$$

For $l = 1, \ldots, d$ suppose that there exists $K_l > 0$ so that $|E^l_n(I^l_n)^{-1}| < K_l h_n < 1/2$. Then for $i < j$ we have

$$e^{-(2K_i + K_j)h_n} \leq \frac{1 + E^i_n(I^i_n)^{-1}}{1 + E^j_n(I^j_n)^{-1}} \leq e^{(K_i + 2K_j)h_n}.$$

Suppose that $B_{i,i}(t)$ and $B_{j,j}(t)$ are integrally separated and satisfy the estimate (3.5). Then

$$\prod_{k=m}^{n} \Phi^B_{i,i}(k)(\Phi^B_{j,j}(k))^{-1} \geq e^{(a_{i,j} - 2K_i - K_j)(t_n - t_m) - b_{i,j}}.$$

So $\Phi^B_{i,i}(n)$ and $\Phi^B_{j,j}(n)$ will be discretely integrally separated if $a_{i,j} - 2K_i - K_j > 0$. On the other hand if $B_{i,i}(t)$ and $B_{j,j}(t)$ are not integrally separated, but instead satisfy the estimate (3.6). Then, given $\varepsilon > 0$ we can choose $M_{i,j}(\varepsilon)$ so that

$$\prod_{k=m}^{n} |\Phi^B_{i,i}(k)(\Phi^B_{j,j}(k))^{-1}| \leq e^{M_{i,j}(\varepsilon)+\frac{\varepsilon}{2}(t_n-t_m)+\sum_{k=m}^{n}(K_i+2K_j)h_k}.$$

So, in particular, if $K_i + 2K_j \leq \frac{\varepsilon}{2}$, then $\sum_{k=m}^{n}(K_i+2K_j)h_k \leq \frac{\varepsilon}{2}(t_n - t_m)$ so that $\Phi^B_{i,i}(n)$ and $\Phi^B_{j,j}(n)$ will satisfy an estimate of the form (2.8). The final condition for $\Phi^B(n)$ to have an approximate discrete integral separation structure is that $\Phi^B_{i,i}(n) = I^i_n + E^i_n > 0$ for $i = 1,\ldots,d$ and $n \geq 0$ which follows from the relation that $|E^i_n(I^i_n)^{-1}| < 1/2$.

The diagonal elements $\Phi^B_{i,i}(n)$ are the numerical solutions of the scalar problems $\dot{y}_i(t) = B_{i,i}(t)y_i(t)$ using $\mathscr{M}$ with the same sequence of step-sizes $\{h_n\}_{n=0}^{\infty}$. Therefore, since the method $\mathscr{M}$ has local truncation error of order $p \geq 1$ and $A(t)$ (and therefore $Q(t)$ and $B(t)$) is $p+1$ times differentiable, we can choose $\tilde{h} > 0$ so that if $h_{\max} \leq \tilde{h}$ and $i = 1,\ldots,d$ we have $E^i_n = T^i_n h^{p+1}_n$ where $|T^i_n| \leq C_i$ for some $C_i > 0$. Because $B(t)$ is bounded we can choose $M^i_B > 0$ so that $|B_{i,i}(t)| \leq M^i_B$ for all $t > t_0$. So, if $h_{\max} \leq \tilde{h}$, then for $i = 1,\ldots,d$ we have $|E^i_n(I^i_n)^{-1}| \leq C_i h^{p+1}_n e^{h_n M^i_B}$.

For $l = 1,\ldots,d$ let $K_l = K_l(n) = C_l h^p_n e^{M^l_B h_n} > 0$. For $i < j$ and each $\varepsilon > 0$, let $h_{i,j} \in (0,1)$ be so small that if $h_{\max} \leq h_{i,j}$, then $C_l h^{p+1}_n e^{M^l_B h_n} < 1/2$ for $l = 1,\ldots,d$ and so that if $B_{i,i}$ and $B_{j,j}$ are integrally separated, then

$$2K_i + K_j = 2C_i h^p_n e^{M^i_B h_n} + C_j h^p_n e^{M^j_B h_n} < a_{i,j}$$

and if $B_{i,i}$ and $B_{j,j}$ are not integrally separated

$$K_i + 2K_j = C_i h^p_n e^{M^i_B h_n} + 2C_j h^p_n e^{M^j_B h_n} < \varepsilon/2.$$

If $h^* > 0$ is such that $h^* = \min\{\{h_{i,j} : i < j\}, \tilde{h}\}$ then it follows that the diagonal entries of $\Phi^B(n)$ are positive, have uniformly bounded inverses, and satisfy either (2.7) or (2.8). It follows that $\Phi^B(n)$ has an approximate discrete integral separation structure. □

The size that $h^* > 0$ must be taken in Lemma 1 depends on the strength of the integral separation. Stronger integral separation between diagonal elements of $B(t)$ implies a milder step-size restriction to preserve integral separation of $\Phi^B(n)$ at the discrete level. As discussed in Section 3.1, an alternative to restricting the step-size to guarantee that our method has an approximate discrete integral separation structure would be to allow the coefficients of the method to vary between time-steps.

There is an interesting connection between Lemma 1 and the conditioning of boundary value problems (BVPs) of ODEs. A classic result in Ascher, U. et al. (1988) states that BVPs for linear nonautonomous ODEs are well-conditioned if and only if there is a dichotomy. The integral separation structure of $B(t)$ gives us a way of quantifying an exponential dichotomy of the zero solution of (2.1) and the matrix $Q(t)$ may be used to define the related projections, see Dieci, L. et al. (2010). Our result (1) can be interpreted as saying that if the system has a stronger dichotomy, then there is a weaker step-size restriction due to stability.

Under additional constraints on $h_{\max} > 0$, we can use Theorem 4 to obtain bounds on the difference of the diagonal elements of $R^A(n)$ and $\Phi^B(n)$.

Lemma 2. *There exists $h^* > 0$ so that $h_{max} \leq h^*$, then $G_n := R^A(n) - \Phi^B_n$ satisfies that $\|G_n\| = C_G h_n^{p+1}$ for some $C_G > 0$.*

Proof. By Lemma 1 and the definition of local truncation error, we have that $\Phi^B(n)$ has an approximate discrete integral separation structure and for $i = 1, \ldots, d$ we have $\Phi^B_{i,i}(n) = \exp\left(\int_{t_n}^{t_{n+1}} B_{i,i}(\tau)d\tau\right) + E_n^i$ where $E_n^i = T_n^i h_n^{p+1}$ where $|T_n^i| \leq C_i$ for some $C_i > 0$.

Consider the unperturbed system $y_{n+1} = \tilde{\Phi}^A(n) y_n$ and the perturbed system $x_{n+1} = (\tilde{\Phi}^A(n) + \tilde{F}_n)x_n$ where $\tilde{\Phi}^A(n)$ and $\tilde{F}_n$ are as defined in (3.11) and (3.12). Using the definition of local truncation error and the fact that $A(t)$ and $B(t)$ are bounded and $p+1$ times differentiable, there exists $\overline{h} > 0$ so that if $h_{\max} \leq \overline{h}$, then E_n^A and E_n^B from (3.9) satisfy that $E_n^A = C_n^A h_n^{p+1}$ and $E_n^B = C_n^B h_n^{p+1}$ where $\|C_n^A\| \leq C^A$ and $\|C_n^B\| \leq C^B$ for constants $C^A, C^B > 0$. Therefore we can choose $0 < h^* < \overline{h}$ so that if $h_{\max} \leq h^*$, then we can bound the sequence $\tilde{F}_n = E_n^A - Q(t_{n+1})E_n^B Q(t_n)^T$ as $\tilde{F}_n = C_n^F h_n^{p+1}$ where $\|C_n^F\| \leq (C^A + C^B)h_{\max}^{p+1}$ and the conclusion of Theorem 4 holds: there exists a sequence $\{\tilde{Q}_n\}_{n=0}^{\infty}$ with each $\tilde{Q}_n$ a real orthogonal $d \times d$ matrix such that $\tilde{Q}_{n+1}\Phi^B(n) = (R^A(n) + E_n)\tilde{Q}_n$

where $E_n = -Q(t_{n+1})\tilde{F}_n Q(t_n)^T$ and $\|\tilde{Q}_n - I\| \le K\|\tilde{F}_n\|$ for some $K > 0$. From this it follows that there exists $C > 0$ so that $\|G_n\| = \|R^A(n) - \Phi^B(n)\| \le C\|\tilde{F}_n\|$ whenever $h_{\max} \le h^*$ and it follows that if $h_{\max} \le h^*$, then $\|G_n\| \le C_G h_n^{p+1}$ for some $C_G > 0$. □

We are now ready to prove our two main theorems showing that we can always select an $h_{\max} > 0$ so that $R^A(n)$ has an approximate discrete integral separation structure and is asymptotically contracting.

Theorem 7. *$R^A(n)$ has an approximate discrete integral separation structure.*

Proof. By Lemma 2 and Lemma 1, there exists $\tilde{h} > 0$ so small that if $h_{\max} \le \tilde{h}$, then for $i = 1, \dots, d$ we have

$$R^A_{i,i}(n) = \Phi^B_{i,i}(n) + (G_n)_{i,i}$$

where G_n is such that $|G_n| \le C_G h_n^{p+1}$ for some $C_G > 0$ and

$$\Phi^B_{i,i}(n) = \exp\left(\int_{t_n}^{t_{n+1}} B_{i,i}(\tau)d\tau\right) + T_n^i h_n^{p+1}$$

where T_n^i is such that $\|T_n^i\| \le T_i$ for some $T_i > 0$. If $h_{\max} \le \tilde{h}$ we have for $i = 1, \dots, d$ that

$$R^A_{i,i}(n) = \exp\left(\int_{t_n}^{t_{n+1}} B_{i,i}(\tau)d\tau\right) + (G_n)_{i,i} + T_n^i h_n^{p+1}$$

where $\|(G_n)_{i,i} + E_n^i h_n^{p+1}\| \le (C_G + C_i)h_n^{p+1}$. By repeating the argument in Lemma 1 we can show that there exists $h^* > 0$ with $h^* \le \tilde{h}$ so that if $h_{\max} \le h^*$, then the diagonal entries of $R^A(n)$ are positive, have bounded inverses, and satisfy an estimate of the form (2.7) or (2.8). It follows that $R^A(n)$ has an approximate discrete integral separation structure. □

The following corollary follows from Lemma 1, Theorem 3, and Theorem 7 and their proofs.

Corollary 1. *Let $\mu_1^A, \dots, \mu_d^A$ denote the Lyapunov exponents of $x_{n+1} = \Phi^A(n)x_n$ and $\mu_1^B, \dots, \mu_d^A$ denote the Lyapunov exponents of $y_{n+1} = \Phi^B(n)y_n$ and $\mu_1, \dots, \mu_d$ denote the Lyapunov exponents* *of* (2.1). *There exists $h^* > 0$ so that if $h_{max} \leq h^*$ and $i = 1,\ldots,d$, then*

$$\mu_i^A = \limsup_{n\to\infty} \frac{1}{t_n - t_0} \sum_{j=0}^{n} \ln(R_{i,i}^A(j)) + \mathscr{O}(h_{max}^p), \quad \mu_i^B = \limsup_{n\to\infty} \frac{1}{t_n - t_0} \sum_{j=0}^{n} \ln(\Phi_{i,i}^B(j)) + \mathscr{O}(h_{max}^p) \tag{3.13}$$

and $\mu_i^A = \mu_i^B + \mathscr{O}(h_{max}^p) = \mu_j + \mathscr{O}(h_{max}^p)$. If the diagonal elements of $B(t)$ are all integrally separated, then we can omit the $\mathscr{O}(h_{max}^p)$ in (3.13).

□

The next theorem states that for sufficiently small step-sizes the numerical solution $x_{n+1} = \Phi^A(n)x_n$ inherits the asymptotic contraction of the diagonal of $B(t)$.

Theorem 8. *There exists $h^* > 0$ so that if $h_{max} \leq h^*$, then $x_{n+1} = \Phi_n^A x_n$ is discretely asymptotically contracting.*

Proof. Use the estimates of Lemma 2 and those in the proof Theorem 7 together with (3.4). □

Under Assumption 1 on the problem (2.1) we may characterize the numerical stability of $\mathscr{M}$ applied to solve (2.1) as follows. Theorem 7 implies that for all sufficiently small $h_{\max} > 0$ the diagonal entries of $R^A(n)$ differ from the diagonal entries of $\Phi^B(n)$ by a term of the same order as the local truncation error of the method. Therefore, if the step-sizes are sufficiently small, each diagonal entry $R_{i,i}^A(n)$ of $R^A(n)$ corresponds to a single step in the numerical solution of the real-valued nonautonomous, linear, scalar test problem $\dot{y}_j(t) = B_{j,j}(t)y_j(t)$ by a one-step method with local error of the same order as the the method $\mathscr{M}$. Theorem 8 then implies that for all sufficiently small $h_{\max} > 0$ the system $v_{n+1} = R^A(n)v_n$ is discretely asymptotically contracting whenever the problems $\dot{y}_j(t) = B_{j,j}(t)y_j(t)$ are each asymptotically contracting. It follows that $x_{n+1} = \Phi^A(n)x_n$ is discretely asymptotically contracting. Whenever the local error of a method is sufficiently small the numerical stability of the one-step method $\mathscr{M}$ applied to solve a problem (2.1) that satisfies Assumption 1 is characterized by the numerical stability of a one-step method of the same order solving the d real-valued, asymptotically contracting, nonautonomous, scalar test problems $\dot{y}_i(t) = B_{i,i}(t)y_i(t)$. The coefficients $B_{i,i}(t)$ can be approximated by computing $Q(t)$ as the solution of (2.5) and then forming $B(t) := Q^T(t)A(t)Q(t) - Q^T(t)\dot{Q}(t)$ or by running a discrete QR iteration directly on the numerical solution $x_{n+1} = \Phi^A(n)x_n$

Remark 1. *Generically, systems of the form* (2.1) *are integrally separated if the coefficient matrix $A(t)$ is bounded and continuous. Therefore, if $A(t)$ is smooth and bounded, then generically any corresponding upper triangular system has an integral separation structure. The proofs of Lemma 1, Theorem 1, and Theorem 7 only use the assumption that $B(t)$ has an integral separation structure and therefore their conclusions hold under the generic assumption that* (2.1) *has a corresponding upper triangular system that is integrally separated. Then as shown above, its numerical stability is characterized by the numerical stability of d real-valued scalar test problems. However, without the additional hypothesis of asymptotic contraction, the stability of these test problems becomes more difficult to characterize. In subsequent work we hope to determine whether or not the asymptotic contraction is a generic hypothesis for systems* (2.1) *that have all negative Lyapunov exponents.*

Chapter 4

Stability of nonlinear problems and linear multistep methods

In this chapter we analyze the stability of Runge-Kutta methods solving a class of nonlinear problems and linear multistep methods solving a linear problem (2.1) that has an integral separation structure.

4.1 Nonlinear Problems

In this section we use the linear numerical stability theory to prove a numerical stability result for Runge-Kutta methods solving a class of nonlinear problems. Similar nonlinear stability results for other linear one-step methods can be shown using similar arguments as long as the structure of the method is known.

Consider the nonlinear initial value problem (1.1) and without loss of generality we can assume that $x_0 = 0$. We let $A(t) := Df(x(t;0),t)$ and rewrite $f(x,t)$ as

$$f(x,t) = A(t)x + N(x,t)$$

where $N(x,t) = f(x,t) - A(t)x$. In light of the hypotheses placed on $f(x,t)$ in Section 1.1 and the assumption that $x(t;x_0)$ is Lypaunov stable, we make the following assumption for the remainder of the section:

Assumption 2. *The function $N(x,t)$ is of the form $N(x,t) = n_1(x,t) + n_2(t)$ where $\|n_1(x,t)\| \leq K\|x\|^2$ and $\|n_2(t)\| \leq K$ for some positive constant $K > 0$.*

Consider the following ODE

$$\dot{x}(t) = f(x(t),t) \equiv A(t)x(t) + N(x(t),t), \quad t > t_0 \tag{4.1}$$

and let $x(t;y_0)$ be the solution of (4.1) with the initial condition $x(t_0) = y_0$. Consider a Runge-Kutta method with Butcher tableaux

$$\begin{array}{c|c} c & \tilde{A} \\ \hline & b^T. \end{array} \tag{4.2}$$

The numerical solution $\{y_n\}_{n=0}^{\infty}$ of (4.1) using the method (4.2) with any initial condition y_0 with the sequence of step-sizes $\{h_n\}_{n=0}^{\infty}$ takes the form

$$\begin{cases} y_{n+1} = y_n + h_n \sum_{j=1}^{s} b_j (A_{n,j} g_{n,j} + N(g_{n,j}, t_n + c_j h_n)), & n \geq 0 \\ g_{n,i} = y_n + h_n \sum_{j=1}^{s} \tilde{a}_{i,j} (A_{n,j} g_{n,j} + N(g_{n,j}, t_n + c_j h_n)), & i = 1, \ldots s \end{cases} \tag{4.3}$$

The numerical solution of $\dot{u}(t) = A(t)u(t)$ using the method (4.2) with the same initial condition y_0 and the same sequence of step-sizes $\{h_n\}_{n=0}^{\infty}$ is of the form $u_{n+1} = \Phi^A(n)u_n$. Assumption 2 together with the implicit function theorem imply that there exists $h^* > 0$ so that if $h_{\max} \leq h^*$, then the numerical solution y_n satisfies the difference equation

$$y_{n+1} = \Phi^A(n)y_n + h_n \tilde{N}(y_n, t_n) \tag{4.4}$$

where $\tilde{N}(x_n,t_n)$ is of the form $\tilde{N} = \tilde{n}_1 + \tilde{n}_2$ where $\|n_1(y_n,t_n)\| \leq \tilde{K}\|y_n\|^2$ and $\|n_2(y_n,t_n)\| \leq \tilde{K}$ for some $\tilde{K} > 0$. We now have the following theorem that shows that the numerical solution of (4.1) with the initial condition x_0 is Lyapunov stable.

Theorem 9. *Let* $\{x_n\}_{n=0}^{\infty} \equiv 0$ *and* $\{y_n\}_{n=0}^{\infty}$ *denote the numerical solutions obtained from solving* (4.1) *with the method* (4.2) *using the sequence of step-sizes* $\{h_n\}_{n=0}^{\infty}$ *with initial conditions* $x_0 = 0$ *and* y_0 *respectively. Then, given* $\varepsilon > 0$*, there exists* $\delta > 0$ *and* $h^* > 0$ *so that if* y_0 *is such that* $|y_0| < \delta$ *and* $h_{max} \leq h^*$*, then the numerical solutions satisfy that* $\|y_n\| = \|x_n - y_n\| < \varepsilon$ *for all* $n \geq 0$*. In other words, the numerical solution* $\{x_n\}_{n=0}^{\infty}$ *is Lyapunov stable.*

Proof. Let $\{z_n\}_{n=0}^{\infty}$ denote the numerical solution of (4.1) using the method (4.3). By the above work and Theorem 8, there exists $\tilde{h}_{z_0} > 0$ so that if $h_{\max} \leq \tilde{h}_{z_0}$, then

$$z_{n+1} = \Phi^A(n) z_n + h_n \tilde{N}(z_n, t_n)$$

and additionally the linear system $u_{n+1} = \Phi^A(n) u_n$ is discretely asymptotically contracting. Thus if $h_{\max} \leq \tilde{h}_{z_0}$, then there exists $C_A > 0$ and $L > 0$ so that $\| \prod_{j=n}^{m} \Phi_j^A \| \leq C_A e^{-L(t_n - t_m)}$ for all $n \geq m$.

By the discrete variation of constants formula, for $n > 0$ we have:

$$z_n = \left[\prod_{j=n-1}^{0} \Phi^A(j) \right] z_0 + \sum_{i=0}^{n-1} \left[\prod_{j=n-1}^{i+1} \Phi^A(j) \right] h_i N(z_i, t_i).$$

where the product $\prod_{j=n-1}^{i+1} \Phi^A(j)$ is taken be 1 when $i \geq n-1$. It follows that

$$\|z_n\| \leq C_A e^{-L(t_{n-1} - t_0)} \|z_0\| + C_A \sum_{i=0}^{n-1} e^{-L(t_{n-1} - t_{i+1})} h_i \tilde{K}(\|z_i\|^2 + 1). \tag{4.5}$$

where sum $\sum_{i=0}^{n-1} e^{-L(t_{n-1} - t_{i+1})}$ is convergent and satisfies that $\sum_{i=0}^{n-1} e^{-L(t_{n-1} - t_{i+1})} \leq \tilde{C}_A$ where $\tilde{C}_A > 0$.

Let $\varepsilon > 0$ be given. Let $\delta_{z_0} > 0$ be so small that $\delta_{z_0} < \min\{\varepsilon/4C_A, \varepsilon/2\}$ and let $0 < h_z^* < \min\{\tilde{h}_{z_0}, \varepsilon C_A \tilde{C}_A \tilde{K}/8\}$. Suppose that $\|z_0\| < \delta_{z_0}$ and consider the set $N_z = \{n : \|z_n\| \geq \varepsilon\}$. Suppose for contradiction that N_z is nonempty. Then there is a minimal element N_z^* of N_z and $N_z^* > 0$ since $\|z_0\| < \delta_{z_0}$. By (4.5) and the definition of N_z^* we have

$$\varepsilon/2 \leq \|y_{N_z^*}\| \leq C_A \|z_0\| + h_n C_A \tilde{C}_A \tilde{K}(\sigma^2 + 1) \leq C_A \delta_{z_0} + 2h_z^* C_A \tilde{C}_A \tilde{K} < \varepsilon/2$$

which is a contradiction. It follows that $\|z_n\| < \varepsilon/2$ for all $n \geq 0$. From this work it follows that if y_0 is so small that $\|y_0\| < \delta$ where $\delta := \delta_{y_0} > 0$ and $h_{\max} \leq \min\{h_0^*, h_{y_0}^*\}$, then $\|y_n - x_n\| < \varepsilon$ for all $n \geq 0$. □

Determining the step-size $h^* > 0$ so that the conclusion of Theorem 9 holds depends on knowing $A(t)$ exactly or equivalently knowing the exact solution. The point of this result is to abstractly show that for small enough step-sizes the global error in the approximation of a Lyapunov stable trajectory whose linear variational equation satisfies Assumption 1 is bounded for a large class of nonlinearities. Typically what is done in practice is to linearize around the numerical solution at each time step and form $C_n := Df(x_n, t_n)$ so that in the numerical solution of (4.1) instead of forming $\Phi^A(n)$ we are forming $\Phi^C(n)$ where $C(t)$ is some matrix with $C(t_n) = C_n$. Using shadowing-type arguments and an assumption of ergodicity (see e.g. Dieci & Van Vleck, E.S. (2005)) it can be shown that the two systems $x_{n+1} = \Phi^A(n)x_n$ and $v_{n+1} = \Phi^C(n)v_n$ have Lyapunov exponents whose differences are bounded in terms of the local error.

4.2 Linear multistep methods

In this section we study the numerical stability of a linear multistep method of the form (1.7) solving the nonautonomous linear problem (2.1). Without loss of generality we assume that $\alpha_k = 1$. The linear multistep method (1.7) applied to solve (2.1) is given as

$$\sum_{i=0}^{k} (\alpha_i I_d - h\beta_i A_{n+i})x_{n+i} = 0 \tag{4.6}$$

where $A_n := A(t_n)$ for all $n \geq 0$. The numerical stability analysis for linear multistep methods turns out to be more challenging than for one-step methods since, as written in the form (4.6), linear multistep methods do not define discrete time dynamical systems. There are two main strategies that have been used to get past this hurdle. The first strategy is to use the structure of the equation (4.6) to express the multistep method as nonautonomous, linear, difference equation in a higher dimensional space. The second strategy is to use invariant manifold theory for maps to associate to each strictly stable method (1.7) a one-step method that governs the long-term dynamics. A strictly stable linear multistep method (1.7) is a method for which the zeros of the polynomial $\rho(z) := \sum_{i=0}^{k} \alpha_i z^i$ all have modulus less than or equal to 1 and the only zero of modulus 1 is $z = 1$ and it is a simple root. We pursue the second strategy since then the numerical stability theory for stictly stable linear multistep methods solving time-dependent problems follows a corollary to the one-step theory we developed in Section 3.

The invariant manifold theory for linear multistep methods was pioneered in Kirchgraber (1986) Eirola , T. & Nevanlinna, O. (1988). This theory allows us to associate to each strictly stable linear multistep method a one-step method with local truncation error of the same order, called the underlying one-step method, that governs the stability of the linear multistep method. Using a one-step method to characterize the stability of a strictly stable method (1.7) allows us to apply the theory developed in Section 3, although an additional restriction on the step-size $h > 0$ may be incurred to guarantee the existence of the underlying one-step method. In the remainder of this section we prove the existence of an underlying one-step method for a linear multistep method (1.7) approximating (2.1) with fixed step-size $h > 0$ and show how to determine this additional step-size restriction.

We first rewrite the nonautonomous linear equation (2.1) of dimension d as an equivalent system autonomous system of dimension $d+1$ using the standard trick of using $\dot{t}(\tau) = 1$ as the differential equation for time:

$$\begin{cases} \dot{x}(\tau) = A(\tau)x(\tau) \\ \dot{t}(\tau) = 1 \end{cases} \tag{4.7}$$

The method (1.7) applied to the system (4.7) produces a numerical solution of the form $(x_n^T, t_n)^T$. If (1.7) is a consistent multistep method, then $\dot{t}(\tau) = 1$ is integrated exactly and therefore $t_n = t(\tau_n) = t_0 + nh$.

We assume that the method (1.7) is strictly stable and also assume that the method is of order

p, that is, for every $(p+1)$-times differentiable function $v(t)$ and $h>0$ sufficiently small we have

$$\sum_{i=0}^{k} \alpha_i v(t_n+ih) - h\sum_{i=0}^{k} \beta_i v'(t_n+ih) = \mathscr{O}(h^{p+1})$$

Consider an autonomous vector field $\dot{x}(t) = f(x)$ and let $\{x_n\}_{n=0}^{\infty}$ be its solution by the method (1.7) using starting values $x_0,\ldots,x_{k-1}$ and fixed step-size $h>0$. Let $X_n := (x_n^T,\ldots,x_{n+k-1}^T)^T$ so that we can express the method (1.7) as

$$X_{n+1} = (L\otimes I)X_n + h\tilde{R}(X_n, X_{n+1}, t_n) \tag{4.8}$$

where $\otimes$ denotes the Kronecker matrix product and

$$L = \left[\begin{array}{cccc|c} 0 & 1 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \\ \hline -\alpha_0 & \ldots & \ldots & -\alpha_{k-2} & -\alpha_{k-1} \end{array}\right], \quad \tilde{R}(X_n, X_{n+1}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sum_{i=0}^{k} \beta_i f(x_{n+i}) \end{bmatrix}.$$

Suppose that we apply (1.7) to the autonomous problem (4.7) using the fixed step-size $h>0$ and let $A(t_n) = A_n$. Then, assuming that $h>0$ is so small that $(I - h\beta_k A_{n+k})$ is always invertible, we may then solve (4.8) for X_{n+1} as

$$X_{n+1} = (L\otimes I)X_n + R(X_n) \tag{4.9}$$

where in this setting $X_n := (x_n^T, t_n, \dots, x_{n+k-1}^T, t_{n+k-1})^T$ and

$$R(X_n) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sum_{i=0}^{k-1} \begin{bmatrix} \alpha_i(I-(I-h\beta_k A_{n+k})^{-1}) + h\beta_i A_{n+i} & 0 \\ 0 & h\beta_i \end{bmatrix} \begin{bmatrix} x_{n+i}^T \\ 1 \end{bmatrix} \end{bmatrix} \equiv \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sum_{i=0}^{k-1} r_{n,i} \begin{bmatrix} x_{n+i}^T \\ 1 \end{bmatrix} \end{bmatrix}$$

Notice that

$$R(X_n) - R(Y_n) = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & & 0 \\ r_{n,0} & \dots & r_{n,k-1} \end{bmatrix} (X_n - Y_n), \quad r_{n,i} = \begin{bmatrix} h(\beta_i A_{n+i} - \alpha_i(I-h\beta_k A_{n+k})^{-1}\beta_k A_{n+k}) & 0 \\ 0 & 0 \end{bmatrix}.$$

Since $A(t)$ is bounded, it follows that R is Lipschitz with constant hP where $P > 0$ depends on A, the coefficients of (1.7), and the norm $\|\cdot\|$.

To apply invariant manifold theory to prove the existence of an underlying one-step method we construct a change of variables that puts the matrix L into a special linearly decoupled form. Our approach closely follows that of Chapter 4 in Humphries, A.R. & Stuart, A.M. (1998). Because (1.7) is strictly stable, $z = 1$ is a zero of the polynomial $\rho(z) = \sum_{i=0}^{k} \alpha_i z^i$ and so we can write $\rho(z) = (z-1)p(z)$ where the zeros of $p(z) = \sum_{i=0}^{k-1} a_i z^i$ all have modulus strictly less than 1. Define the matrices

$$D = \left[\begin{array}{cccc|c} -1 & 1 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & -1 & 1 & 0 \\ & & & -1 & 1 \\ \hline a_0 & \dots & \dots & a_{k-2} & a_{k-1} \end{array}\right], \quad C = \left[\begin{array}{cccc|c} 0 & 1 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & 0 & 1 & \vdots \\ a_0 & \dots & \dots & a_{k-2} & 0 \\ \hline 0 & \dots & \dots & 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} C_1 & \\ \hline & 1 \end{array}\right]$$

We have the following Lemma that can be proved by direct computation with D, L, and C and by using the fact that the method (1.7) is strictly stable.

Lemma 3. *The matrix D is invertible, $DL = CD$, and the spectrum of C_1 lies in* $(0,1)$.

□

Under the change of variables $U_n = (D \otimes I)X_n$, we obtain

$$U_{n+1} = (C \otimes I)U_n + (D \otimes I)R\left((D^{-1} \otimes I)U_n\right) \tag{4.10}$$

If we let $U_n = (V_n^T, W_n^T)^T$ where $V_n \in \mathbb{R}^{(d+1)(k-1)}$ and $W_n \in \mathbb{R}^{d+1}$ this leads to the following linearly decoupled system:

$$\begin{cases} V_{n+1} = (C_1 \otimes I_{k-1})V_n + N_1(V_n, W_n) \\ W_{n+1} = W_n + N_2(V_n, W_n) \end{cases} \tag{4.11}$$

where $N = (N_1^T, N_2^T)^T$ is defined as $N(U_n) := R((D^{-1} \otimes I)U_n)$. It follows that N has Lipschitz constant bounded above by $hP\|D^{-1} \otimes I\| := hK$. The stability properties of the systems (4.11) and the original system (4.8) are identical since $U_n = (D \otimes I)X_n$ and therefore we focus on analyzing the stability of (4.11). Hence, it suffices to show that there exists φ so that $V_n = \varphi(W_n)$ for all n and that the one-step method defined by $W_{n+1} = W_n + N_2(\varphi(W_n), W_n)$ has local truncation error of order $\mathscr{O}(h^{p+1})$.

We now state a general theorem on invariant manifolds for maps that appears in Stoffer (1993) that we use to show the existence function φ such that $V_n = \varphi(W_n)$ that is invariant under the map defined by (4.11). For a proof see Stoffer, D. & Nipp, K. (1992).

Theorem 10. *Consider the map $F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n$ defined by $(\tilde{x}, \tilde{y}) = F(x, y)$ where*

$$\begin{cases} \tilde{x} = Ax + \tilde{f}(x, y) \\ \tilde{y} = g(x, y) \end{cases} \tag{4.12}$$

Let $\|\cdot\|$ be a norm on $\mathbb{R}^{n+m}$ and assume that A, $\tilde{f}$, and g satisfy the following:

1. *The matrix A is invertible and* $\|A^{-1}\| \leq \alpha$.

2. *The functions* $\tilde{f}$ *and g are k time differentiable and their derivatives are bounded.*

3. *The functions* $\tilde{f}$ *and g satisfy the following Lipschitz conditions*

$$\|\tilde{f}(x,y) - \tilde{f}(u,v)\| \leq L_{11}|x-u| + L_{12}|y-v|$$
$$\|g(x,y) - g(u,v)\| \leq L_{21}|x-u| + L_{22}|y-v|.$$

4. *The constants* α *and* L_{ij} *where* $i, j = 1,2$ *satisfy the following estimate*

$$L_{11} + L_{22} + 2\sqrt{L_{12}L_{21}} < \alpha^{-1}$$

and

$$L_{22} + \frac{2L_{12}L_{21}}{\alpha^{-1} - L_{11} - L_{22}} < min\left\{1, \left(\alpha^{-1} - L_{11} - \frac{2L_{12}L_{21}}{\alpha^{-1} - L_{11} - L_{22}}\right)\right\}.$$

Then there exists a function $\varphi : \mathbb{R}^m \to \mathbb{R}^n$ *for which the following holds:*

1. *The manifold defined by the graph of* φ *is invariant under the map F.*

2. *The function* φ *is k times differentiable and the derivatives are bounded.*

3. *The graph of* φ *is exponentially attractive with constant* ξ *given by*

$$\xi = L_{22} + \frac{2L_{12}L_{21}}{\alpha^{-1} - L_{11} - L_{22}} < 1$$

4. *All points in the phase space are attracted to the manifold at an exponential rate: There exists* $c > 0$ *so that for any* $(x_0, y_0) \in \mathbb{R}^m \times \mathbb{R}^n$ *there exists* $(\tilde{x}_0, \tilde{y}_0) \in \mathbb{R}^{m \times n}$ *with* $\tilde{y}_0 = s(\tilde{x}_0)$ *so that*

$$\|x_k - \tilde{x}_k\| \leq c\xi^k \|y_0 - s(x_0)\|$$
$$\|y_k - \tilde{y}_k\| \leq \xi^k \|y_0 - s(x_0)\|$$

where (x_k, y_k) is the k^{th} iterate of the map F applied to (x_0, y_0) and $(\tilde{x}_n, \tilde{y}_n)$ is the n^{th} iterate of the map F applied to $(\tilde{x}_0, \tilde{y}_0)$

□

Corollary 1. *Suppose that (1.7) is strictly stable and has local truncation error of order $p \geq 1$. Then there exists $h^* > 0$ so small that if $0 < h < h^*$, then there exists a function $\varphi : \mathbb{R}^{d+1} \to \mathbb{R}^{(d+1)(k-1)}$ so that*

1. *The graph of φ is invariant under the map (4.11)*
2. *The function φ is a smooth map*
3. *All points $(V_n, W_n) \in \mathbb{R}^k \times \mathbb{R}^{(d-1)k}$ are attracted to the graph of φ at an exponential rate.*
4. *The one-step method defined by $W_{n+1} = W_n + N_2(\varphi(W_n), W_n)$ has local truncation error of order p*

Proof. The map defined by (4.11) is of the form (4.12) where $W_n = x$, $V_n = y$, $A = I_{d+1}$, $\alpha = 1$ and where $L_{11} = L_{12} = L_{21} = hK$ and $L_{22} = \rho + hK$ where $\rho := \|(C_1 \otimes I_{k-1}\|$. By Lemma 3, the spectrum of C_1 lies in $(0,1)$ and it follows that there exists $\rho \in (0,1)$ such that $\|C_1 \otimes I_{k-1}\| \leq \rho$. Suppose that h_0 is so small that $2h_0K < 1$ and so that

$$h_0 < \frac{(1-\rho)}{4K}, \quad hK\left(1 + \frac{2hK}{1-2hK}\right) < \frac{1-\rho}{2}.$$

Then, if $0 < h \leq h_0$, Theorem 10 implies the existence of a function $\varphi : \mathbb{R}^k \to \mathbb{R}^{(d-1)k}$ satisfying 1-3. The conclusion 4. follows by substituting a continuous function $v(t)$ on the graph of φ into the map (4.11) and noting the the method (1.7) is of order $p \geq 1$. □

Notice that by definition of $D \times I$, since the method is consistent, we can write W_n in the form $W_n = (\tilde{W}_n^T, t_0 + nh)^T$. Then, corollary (1) implies that if h is so small that $hK + \sqrt{hK(1+hK)} < 1/2$, then the stability of the system (4.11), and equivalently, the stability of (1.7) applied to solve

(2.1), is characterized by the one-step method defined by

$$\tilde{W}_{n+1} = \tilde{W}_n + N_2(\varphi(\tilde{W}_n, t_n), \tilde{W}_n, t_n) \equiv \theta(\tilde{W}_n, t_n) \tag{4.13}$$

which has local truncation error of order $\mathscr{O}(h^{p+1})$ and $t_n = t_0 + nh$. From the definition of truncation error, it follows that we can write (4.13) in the form

$$\tilde{W}_n = \Phi^A(n)\tilde{W}_n \tag{4.14}$$

Once in this form we can apply the one-step theory developed in Chapter 3 to find additional restrictions on the step-size h so that (4.14) is discretely asymptotically contracting.

A significant drawback to using underlying one-step methods to analyze the stability of linear multistep methods solving (2.1) is that it is unclear how to directly extend the linear theory to the nonlinear case was done in Section 4.1 for Runge-Kutta methods. To apply Theorem 10 to the case where the method (1.7) is approximating a nonlinear problem we must assume that the derivatives of the nonlinearity are bounded, which it was not necessary to do in Theorem 9. Additionally, Theorem 4.1 only applies to Runge-Kutta methods and since the underlying one-step method is not necessarily a Runge-Kutta method (indeed, it can be 'quite exotic' Eirola , T. & Nevanlinna, O. (1988)), it is unclear how to extend such a theorem to an underlying one-step method. We hope to address this issue in subsequent work.

Chapter 5

Numerical methods and experiments

In this chapter we use the theory from Chapter 3 to develop two novel stability based methods for step-size selection. We present the results of several numerical experiments that highlight the utility these methods. The chapter is concluded with an exploratory section on how our theory might be useful as a way of characterizing stiffness in the numerical solution of time-dependent nonlinear problems.

5.1 Two-dimensional linear example

Standard initial value problem solvers select step-size based on the local accuracy of the numerical solution. For the numerical solution of (2.1), this means that the solver exerts no direct control over the local accuracy of $Q(t)$ and the diagonal of $B(t)$ and hence there is no direct control over in the error in the Lyapunov exponents. In this section we describe an efficient step-size selection procedure that gives a solver control over its discrete Lyapunov exponents and demonstrate the efficacy of this procedure by showing that it is able to produce a decaying numerical solution in a situation where standard step-size selection fails to do so.

When using QR methods for the computation of Lyapunov exponents of continuous or discrete time dynamical systems of dimension d, typically the first $p \leq d$ diagonal entries of the upper triangular matrices in the QR decomposition correspond to the p largest Lyapunov exponents, see e.g. Dieci, L. & Van Vleck, E.S. (1995). In the notation of Section 3.1, this implies that typically the largest Lyapunov exponent $\dot{y}(t) = B(t)y(t)$ is the Lyapunov exponent of the scalar equation $\dot{y}_1(t) = B_{1,1}(t)y_1(t)$ which is approximately the discrete Lyapunov exponent of $y^1_{n+1} = \Phi^B_{1,1}(n)y^1_n$. Similarly, the largest discrete Lyapunov exponent of the numerical solution of (2.1) is typically the discrete Lyapunov exponent of $v_{n+1} = R^A_{1,1}(n)v_n$. Therefore the size of $|R^A_{1,1}(n) - \Phi^B_{1,1}(n)|$ provides a way to measure how accurately a one-step method solving (2.1) is resolving the stability of the differential equation.

Both $R^A_{1,1}(n)$ and $\Phi^B_{1,1}(n)$ can be computed efficiently by using only the first column in the orthogonal factor of a discrete and continuous QR process respectively. To form $\Phi^B_{1,1}(n)$ we must compute the numerical solution of $\dot{y}_1(t) = B_{1,1}(t)y_1(t)$. This requires that we have an approximation to $B_{1,1}(t)$ which satisfies the equation $B_{1,1}(t) = Q_1(t)^T A(t)Q_1(t) - Q_1(t)^T \dot{Q}_1(t)$ where $Q_1(t)$ is the first column of $Q(t)$. It can be shown (see Dieci & Van Vleck, E.S. (1999)) that $Q_1(t)$ satisfies the differential equation $\dot{Q}_1(t) = A(t)Q_1(t) - Q_1(t)(Q_1(t)^T A(t)Q_1(t) - S(Q_1(t), A(t)) := S_1(Q_1, A)$, where $S(Q,A)$ is defined as in (2.5), we can approximate $\Phi^B_{1,1}(n)$ by solving $\dot{y}_1 = B_{1,1}(t)y_1(t)$ and $\dot{Q}_1(t) = S_1(Q(t), A(t))$ simultaneously. We can form $R^A_{1,1}(n)$ by letting $Q^1_n = Q_1(t_0) \in \mathbb{R}^{d\times 1}$ and then inductively forming partial QR factorization $\Phi^A(n)Q^1_n = Q^1_{n+1}R^A_{1,1}(n)$ where $Q^1_{n+1} \in \mathbb{R}^{d\times 1}$ is orthogonal and $R^A_{1,1}(n)$ is a scalar.

Consider the extended system

$$\begin{cases} \dot{x}(t) = A(t)x(t) \\ \dot{Q}(t) = A(t)Q(t) - Q(t)(Q(t)^T A(t)Q(t) - S(Q(t), A(t)) \\ \dot{y}_1(t) = (Q(t)^T A(t)Q(t) - Q^T(t)\dot{Q}(t))y_1(t) \end{cases} . \tag{5.1}$$

Our procedure for selecting step-size based only on the accuracy of the numerical solution of (2.1) is as follows. Solve the extended system (5.1) with the initial conditions $x(t_0) = x_0$, $Q_1(t_0) = Q_1(t_0)$ and $y_1(t_0) = Q_1(t_0)^T x(t_0)$ where $Q_1(t_0)$ is a random orthogonal vector. Then at each candidate time-step we form approximations to $R^A_{1,1}(n)$ and $\Phi^B_{1,1}(n)$ as described in the previous paragraph and let $\varepsilon_n := |R^A_{1,1}(n) - \Phi^B_{1,1}n)|$. If ε_n is below a specified tolerance TOLQR, then we continue. If not, then the step-size is reduced and a new smaller candidate step-size is selected. We refer to this procedure as QR error control.

Although solving the equation (5.1) to find the solution of (1.4) requires solving a $2d+1$ dimensional system as opposed to a d dimensional linear system, it may be advantageous to do so, especially since a result in Dieci & Van Vleck, E.S. (2009) proves that the global error in the approximation of the Q-equation $\dot{Q}(t) = A(t)Q(t) - Q(t)(Q(t)^T A(t)Q(t) - S(Q(t),A(t))$ is bounded in terms of the local error. To ensure the orthogonality of our approximation to $Q(t)$ we modify ode15s to be a step-and-project type method where after the integrator forms an approximation of $Q(t_n)$ we project this value by forming its QR factorization and taking this orthogonal factor as the orthogonal approximation to $Q(t_n)$. The results in Dieci, L. & Van Vleck, E.S. (2002) show that this step and project type procedure does not affect the order of the local error and hence the standard error control algorithms will still work in the same way .

Consider the two-dimensional nonautonomous linear problem (1.4) with $B(t)$ and $Q(t)$ as defined in (1.5) and with the parameter values as listed in Figure 1.1. Under the change of variables $x(t) = Q(t)y(t)$, the system (1.4) is transformed to the corresponding upper triangular system

$$\dot{y}(t) = B(t)y(t). \tag{5.2}$$

From this it follows that the system is is integrally separated and asymptotically contracting and satisfies the hypotheses of Assumption 1. For our experiments we fix the initial conditions $Q_1(0) = (1,0)^T$ and $x(0) = (1,0)^T$ and $y(0) = (1,0)^T$. Let $R(t)$ be an upper triangular fundamental matrix solution of (5.2) and factor $R(t)$ as

$$R(t_n) = R(t_n, t_{n-1}) \cdot \ldots \cdot R(t_1, t_0)R(0)$$

where $R(t, t_{n-1})$ is the solution of (5.2) with initial condition $R(t_{n-1}, t_{n-1}) = I$. If $Q_1(0) = (1,0)^T$

and the first column R_1 of R satisfies $R_1(0) = (1,0)$, then we can express $R(t_n,t_{n-1})$ exactly as

$$R(t_n,t_{n-1}) = \exp\left(-.20h_n + .21(\sin(t_n) - \sin(t_{n-1}))\right), \quad h_n = t_n - t_{n-1}.$$

We use the quantity $R(t_n,t_{n-1})$ as a way of measuring the accuracy of $R^A_{1,1}(n)$.

In Table 5.1 we present the results of a Matlab experiment. The approximate discrete Lyapunov exponent was found by taking the maximum value of the quantity $\mu_n := \ln(R^A_{1,1}(n))/T(n)$ for values of n such that $T(n) > 50$. In all the tested cases, $\mu_n > 0$ corresponded to a numerical solution of (1.4) with norm growing at an exponential rate and, conversely, $\mu_n < 0$ corresponded to a numerical solution of (1.4) with norm decaying at an exponential rate. Therefore, the values of the approximate discrete Lyapunov exponents are indicative of the stability or instability of the numerical solution of (1.4).

The results indicate that QR error control is an efficient method for preserving the numerical stability of the numerical solution of (1.4) using ode15s with a maximum BDF order of 1. The standard, unmodified ode15s solver fails to produce a decaying numerical solution for tested values of TOL less that 10^{-7}. When ode15s is modified to use QR error control it produces a decaying numerical solution for all tested values of TOL at the expense of using many more time-steps at lower tolerances. This extra expense is justified since by using QR error control the modified ode15s solver is able to produce a numerical solution that correctly preserves asymptotic decay using fewer time-steps than the unmodified ode15s solver.

We can explain the superior performance of the modified ode15s solver as follows. In Table 5.1, one can see that while the local error, measured by LTE(max) and LTE(mean), of the numerical solution of (1.4) is approximately the same for both the modified and unmodified solvers. For values of TOL less than 10^{-7} the modified solver produces a much more accurate approximation to $R^A_{1,1}(n)$, with values of LTEB(max) and LTEB(mean) an order of magnitude or more smaller than those of the unmodified solver. This indicates that the local accuracy of the numerical solution itself is not the only quantity one should be monitoring for the preservation of stability. It runs counter to the heuristic that a loss of stability in a numerical method will manifest itself as a spike in the value of the local truncation error. Numerical instabilities can accumulate in slow and subtle ways for time-dependent problems.

Problem	TOL	LTE(max)	LTE(mean)	LTEB(max)	LTEB(mean)	Appr. DLE	Nsteps
Solution of (1.4)	$1E-3$	$3.02E-1$	$1.65E-2$	$2.79E-3$	$1.37E-3$	$7.61E-3$	$1.52E4$
	$1E-4$	$1.52E-1$	$3.60E-2$	$2.70E-3$	$1.22E-3$	$1.15E-2$	$1.74E4$
	$1E-5$	$8.02E-3$	$3.24E-3$	$8.00E-4$	$3.63E-4$	$2.61E-3$	$5.60E4$
	$1E-6$	$1.95E-3$	$7.67E-4$	$2.30E-4$	$1.09E-4$	$6.85E-3$	$1.90E5$
	$1E-7$	$1.63E-4$	$3.10E-5$	$6.98E-5$	$2.64E-5$	$-3.52E-2$	$6.23E5$
Solution of (5.1)	$1E-3$	$3.02E-1$	$1.49E-4$	$9.41E-3$	$8.72E-5$	$-3.68E-2$	$1.89E5$
	$1E-4$	$5.07E-2$	$8.39E-5$	$1.76E-3$	$5.8632E-5$	$-4.25E-2$	$2.70E5$
	$1E-5$	$6.63E-3$	$8.76E-5$	$5.98E-4$	$5.8639E-5$	$-3.89E-2$	$2.75E5$
	$1E-6$	$1.04E-3$	$7.27E-5$	$2.05E-4$	$5.07E-5$	$-2.75E-2$	$3.42E5$
	$1E-7$	$1.58E-4$	$2.15E-5$	$4.50E-5$	$2.3E-5$	$-4.44E-2$	$6.84E5$

Table 5.1: Table of values for various error tolerances (TOL) of the maximum (LTE(max)) and mean (LTE(mean)) local truncation error of the solution of (1.4), the maximum (LTEB(max)) and mean (LTEB(mean)) of the error of $\varepsilon_n := |R^A_{1,1}(n) - R_{1,1}(t_{n+1}, t_n)|$, the approximate value of the largest discrete Lyapunov exponent of the numerical method $x_{n+1} = \Phi^A(n)x_n$, and the number of time steps taken (Nsteps). TOL is the absolute and relative error tolerance of the integrator and and the solution interval was $[0, 100]$. The integrator used to solve (1.4) was the Matlab solver ode15s using BDFs with a maximum order of 1. The integrator used to solve (2.5) was a modified version Matlab's ode15s using BDFs with a maximum order of 1 where the modifications were to project the candidate $Q_1(t_n)$ at each time-step to ensure its orthogonality and to control the step-size so that ε_n satisfies a tolerance of $\text{TOLQR} = 3E-5$.

5.2 Stability based step-size control for asymptotically contracting scalar test problems

There have been many methods proposed for selecting the step-size for initial value problem solvers. Most step-size selection strategies for solving the ODE initial value problem rely on some guess for the step-size followed by refinement based upon accuracy or stability consideration, see Gustafsson et al. (1988); Hall (1985, 1986). For nonautonomous problems, selecting step-size based upon stability considerations is difficult as there does not seem to be a good time-dependent characterization of the stability region. In this section we briefly review a classical algorithm for selecting step-size based upon local accuracy and then devise a new algorithm that selects step-size by monitoring Lyapunov exponent of the numerical method.

One of the most well known algorithms for selecting the step-size for a numerical initial value problem solver is known as Milne's method which briefly review. Suppose that we simultaneously use two different solvers, one with local error of order p and the other with local error of order $\tilde{p} > p$. The solution generated using the higher order method is treated as a proxy for the exact solution. If x_n is the solution obtained at time-step n using the lower order solver and y_n is the solution obtained at time-step n using the higher order solver, then the difference $\varepsilon_n = x_n - y_n$ is used as an estimate of the local error of x_n at time-step n. The step-size can then be adjusted based upon whether or not ε_n satisfies a given tolerance. An implementation of Milne's method is given in Algorithm 1.

Algorithm 1 Milne's Device

```
Input: $x_0$, $h_0$, TOL, $t0$, $T$, $h_{\min}$, $h_{\max}$
  Set $n = 1, x_0 = y_0$
  while $t < T$ do
    if $h_n < h_{\min}$ then
      $h_n = h_{\min}$
    end if
    if $h_n > h_{\max}$ then
      $h_n = h_{\max}$
    end if
    Compute approximate solution $x_n$ with local error of order $p$ and approximate solution $y_n$ with local error of order at least $p+1$
    Set $\kappa$ to be measure of the error using $x_n$ and $y_n$
    if $\kappa >$ TOL and $h_n \geq 2h_{\min}$ then
      $h_n = h_n/2$
    else if $\kappa <$ TOL$/10$ and $h_n \leq h_{\max}/10$ then
      $h_n = 10h_n$
    else
      $h_n = 0.9h_{n-1}(TOL/\kappa)^{1/(p+1)}$
      $n = n+1$
      $t_n = t_{n-1} + h_{n-1}$
    end if
  end while
```

Consider the scalar test problem

$$\dot{x}(t) = \lambda(t)x(t) \tag{5.3}$$

where $\lambda(t)$ is asymptotically contracting. The importance of preserving the sign of a scalar test problem is highlighted in Section 5.1 where we devised a procedure for stabilizing an unstable numerical solution by comparing the first entry of the upper triangular factor of its QR iteration to the numerical solution of a scalar test problem. We now develop an algorithm based off of Milne's method that selects step-size in a way that numerically preserves the stability of (5.3)

Recall Theorem 6 from that estimates the discrete Lyapunov exponent μ of a one-step method solving an asymptotically contracting scalar test problem of the form (5.3). For all sufficiently small step-sizes we can estimate μ as

$$\mu \leq \limsup_{n\to\infty} \frac{1}{n+1} \sum_{j=0}^{n} \left(\int_{t_j}^{t_{j+1}} \lambda(\tau)d\tau + C_j h_j^{p+1} \right).$$

The stability of the test problem is determined by the sign of μ which is determined by what sign the expression

$$S_n = \sum_{j=0}^{n} \left(\int_{t_j}^{t_{j+1}} \lambda(\tau)d\tau + C_j h_j^{p+1} \right)$$

assumes for large values of n. If $p \geq 1$, we assume that the term $C_j h_j^{p+1}$ is negligible and estimate

$$S_n \approx \sum_{j=0}^{n} h_j \lambda(t_j) =: s_n.$$

Fix some tolerances $s_{\text{tol}} > 0$ and $\lambda_{\text{tol}} > 0$. If $s_n < -s_{\text{tol}}$ and $|\lambda(t_n)| \geq \lambda_{\text{tol}}$, then we can solve the above equation to determine an approximate $h_{n+1} > 0$ so that $s_{n+1} < 0$. If $s_n \geq -s_{\text{tol}}$ or $|\lambda(t_n)| < \lambda_{\text{tol}}$, then we can select step-size based using Milne's device. This leads us to the following algorithm for selecting step-size based upon accuracy and Lyapunov stability.

Algorithm 2 functions as follows. If the approximate Lyapunov exponent is not less than $-s_{\text{TOL}}$, then we select step-size based upon accuracy in the same way as Algorithm 1. If the approximate Lyapunov exponent is less than $-s_{\text{TOL}}$, then we select step-size based upon the stability

Algorithm 2 Algorithm for selecting step-size based upon Lyapunov exponent stability

```
Input: $x_0$, $h_0$, TOL, $t_0$, $T$, $h_{\min}$, $h_{\max}$, $s_{\text{tol}}$, $\lambda_{\text{tol}}$
  Set $s_0 = 0$ and $n = 1$
  while $t < T$ do
    Compute $x_n$ and $y_n$
    if $s_{n-1} \geq -s_{\text{tol}}$ and $|\lambda(t_{n-1})| < \lambda_{\text{tol}}$ then
      if $h_n < h_{\min}$ then
        $h_n = h_{\min}$
      end if
      if $h_n > h_{\max}$ then
        $h_n = h_{\max}$
      end if
      Set $\kappa$ to be measure of the error using $x_n$ and $y_n$
      if $\kappa >$ TOL then
        $h_n = h_n/2$
      else if $\kappa <$ TOL$/10$ then
        $h_n = 10h_n$
      else
        $h_n = 0.9h_{n-1}(TOL/\kappa)^{1/(p+1)}$
        $t_n = t_{n-1} + h_{n-1}$
        $n = n+1$
      end if
    else
      Set $h_n = -s_{n-1}/|\lambda(t_{n-1})|$
      $s_n = s_{n-1} + h_{n-1}\lambda(t_{n-1})$
      $t_n = t_{n-1} + h_{n-1}$
      $n = n+1$
    end if
  end while
```

considerations outlined above.

We present and discuss the numerical results of several experiments using Algorithm 2. We use the method of Bogacki and Shampine derived in Bogacki, P. & Shampine, L. (1989) that computes the numerical approximation to solution using a Runge-Kutta method of order 2 and estimates the local error by comparing this to the output of Runge-Kutta method of order 3. We use an implementation of Algorithm 1 to compare with an implementation of Algorithm 2. Matlab's ODE solver ode23 which is a more advanced implementation of the method of Bogacki and Shampine than the implementation we used for Algorithm 1 and Algorithm 2. At the end of this section we compare the number of steps taken by Algorithms 1 and 2 with the number of steps taken by ode23 to give some insight into how Algorithm 2 performs against a commercial ODE solver.

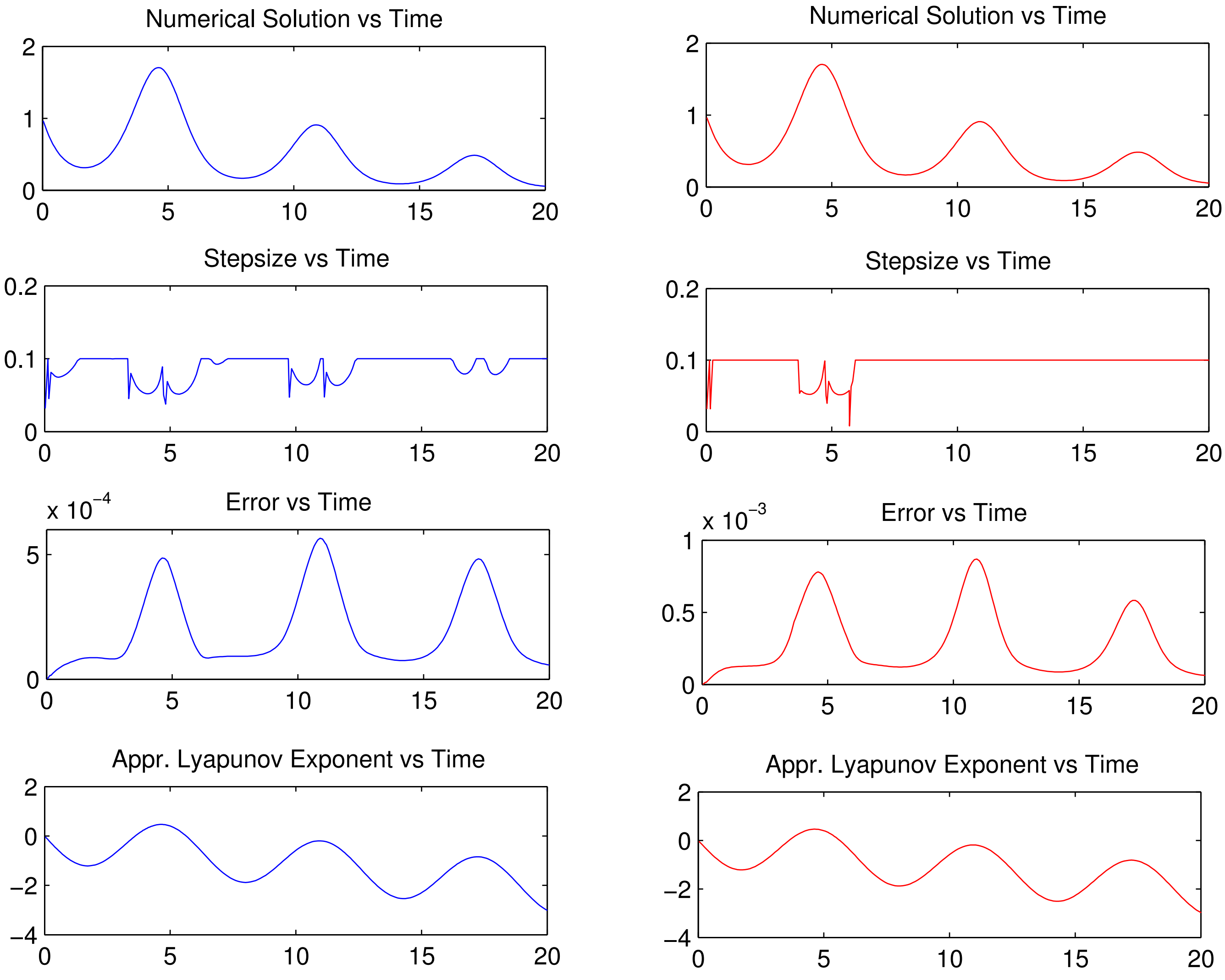


Figure 5.1: Experiment 1: $\dot{x}(t) = (-\cos(t) - 0.1)x(t)$. The left column has the results for Algorithm 1 and the right for Algorithm 2. The error κ was measured by the absolute error. The algorithm inputs we used were $x_0 = 1$, TOL$= 1E-5$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$. Algorithm 1 took 246 steps while Algorithm 2 took 222 steps.

We now make a few remarks on the results in Figures 1-6. First of all, it is clear that Algorithm 1 will generally produce a more accurate solution and only when the numerical solutions decay to be very close to 0 can the accuracy of the numerical solution produced by Algorithm 2 recover the same order of accuracy. In Figures 5.1 and 5.2 where $\lambda(t)$ has lower amplitude and lower frequency oscillations, Algorithm 2 produces a less accurate solution than Algorithm 1 with only a

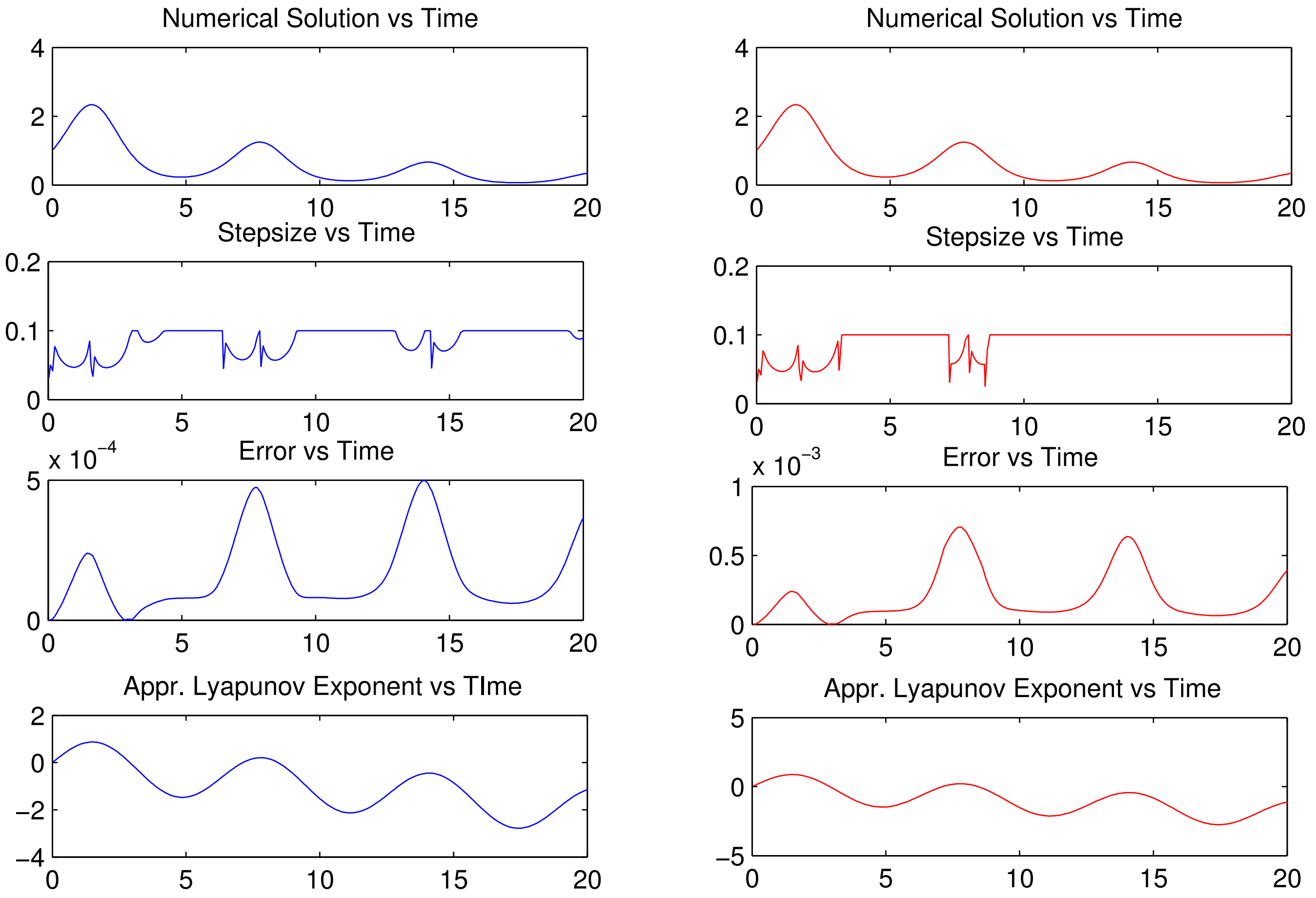


Figure 5.2: Experiment 2: $\dot{x}(t) = (\cos(t) - 0.1)x(t)$. The left column has the results for Algorithm 1 and the right for Algorithm 2. The error κ was measured by the absolute error. The algorithm inputs we used were $x_0 = 1$, TOL$= 1E-5$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$. The Milne method took 260 steps while Algorithm 2 took 244 steps.

slight decrease in the number of steps required. In Figure 5.3, when $\lambda(t)$ has has faster oscillations and in Figure 5.4 where $\lambda(t)$ has larger amplitude and higher frequency oscillations, Algorithm 2 has a large gain in efficiency, getting away with a much larger stepsize, although the accuracy still declines by a factor of 10. Thus Algorithm 2 is more advisable for use in the presence of a solution with high frequency or large amplitude oscillations where preserving stability using accuracy, which is a local property, will be much harder than preserving stability using Lyapunov exponents

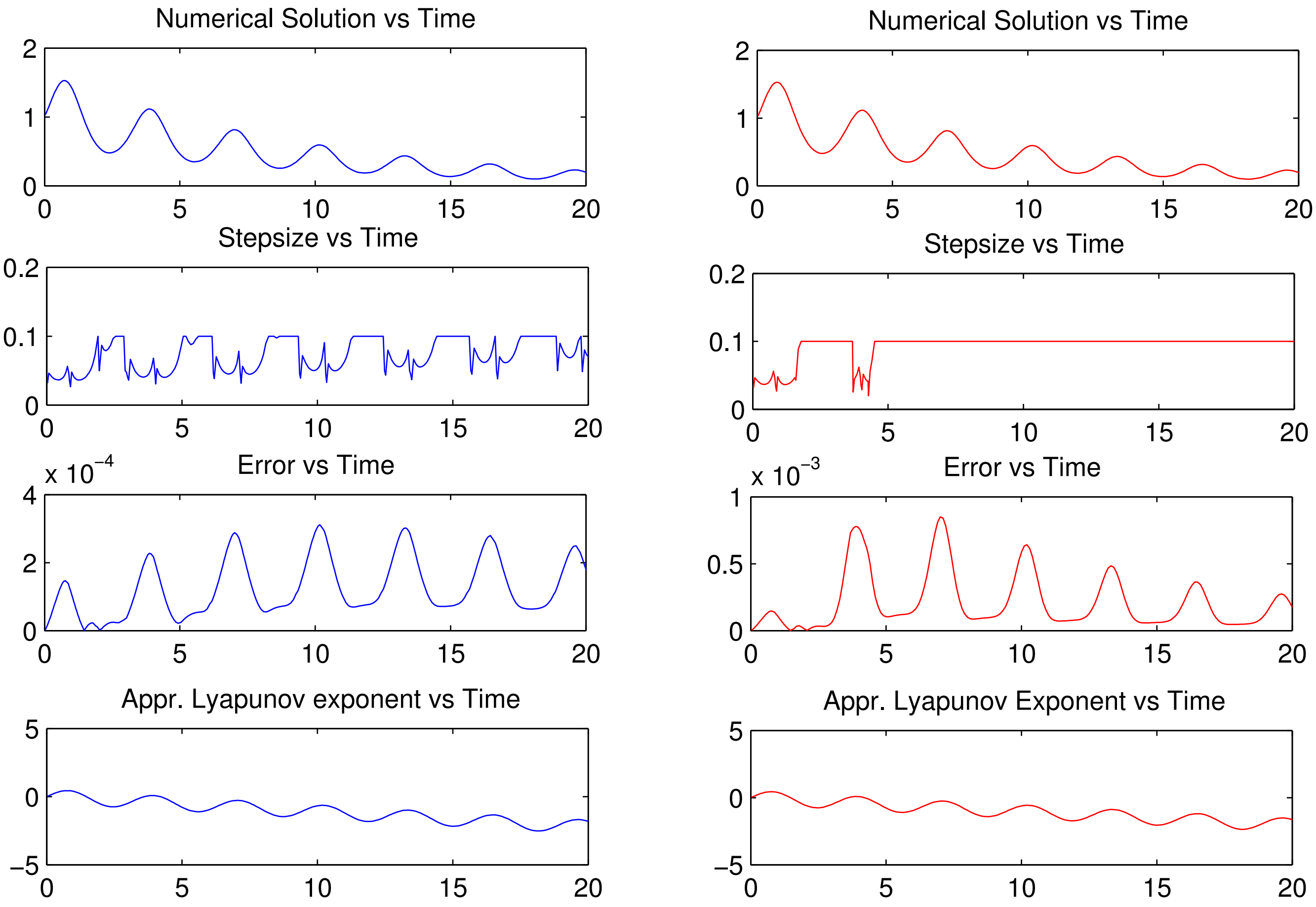


Figure 5.3: Experiment 3: $\dot{x}(t) = (\cos(2t) - 0.1)x(t)$. The left column has the results for Algorithm 1 and the right for Algorithm 2. The error κ was measured by the absolute error. The algorithm inputs we used were $x_0 = 1$, TOL$= 1E-5$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$. Algorithm 1 took 338 steps while Algorithm 2 took 240 steps.

which are global quantities.

Figures 5.5 and 5.6 are meant to demonstrate performance of the algorithms under various error tolerances. Matlab's ode23 is used for comparison, as ode23 implements the same embedded Runge-Kutta method of Bogacki-Shampine we used in Algorithms 1 and 2. For lower error tolerances Algorithm 2 requires about the same number of steps as Algorithm 1 and Matlab's ode23

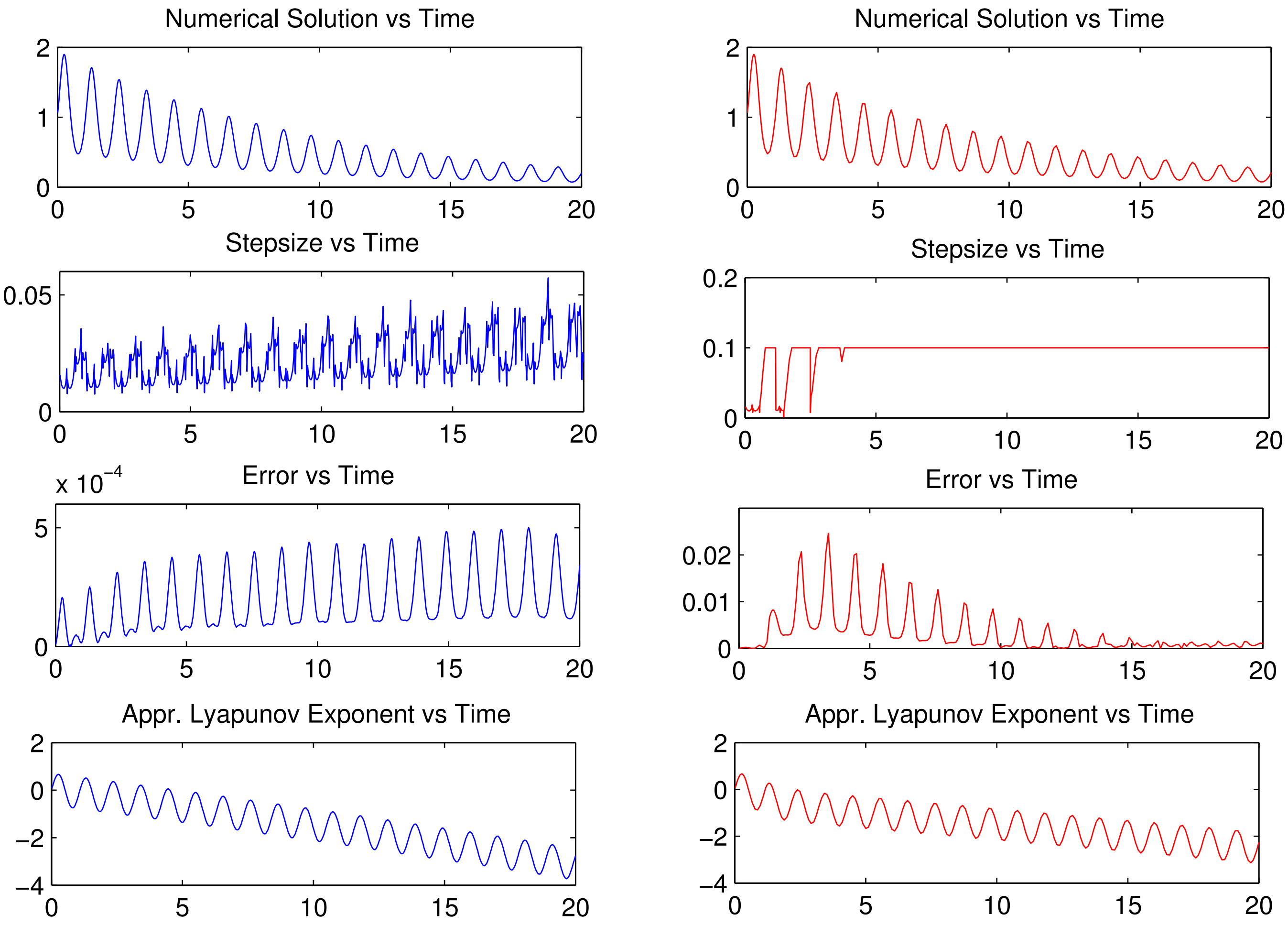


Figure 5.4: Experiment 4: $\dot{x}(t) = (4\cos(6t) - 0.1)x(t)$. The left column has the results for Algorithm 1 and the right for Algorithm 2. The error κ was measured by the absolute error. The algorithm inputs we used were $x_0 = 1$, TOL$= 1E-5$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$. Algorithm 1 took 1242 steps while Algorithm 2 took 287 steps.

requires the fewest steps. However, when high tolerances are used, Algorithm 2 outperforms the other two algorithms. So, if preserving the stability and asymptotic properties of a numerical solution are more important than accuracy, then using Algorithm 2 may be a more efficient choice than using the standard algorithms with a high error tolerance.

Table 1			
Method\Tol	1E-4	1E-5	1E-6
Algorithm 2	204	222	266
Algorithm 1	202	308	633
ode23	110	228	486

Table 2			
Method\Tol	1E-4	1E-5	1E-6
Algorithm 2	204	244	326
Algorithm 1	202	310	645
ode23	111	233	492

Figure 5.5: Table 1 corresponds to the equation $\dot{x}(t) = (-\cos(t) - 0.1)x(t)$ and Table 2 corresponds to $\dot{x}(t) = (\cos(t) - 0.1)x(t)$. The tables record the number of steps taken by Algorithms 2 and 1 and Matlab's ode23 for comparison for various values of TOL where TOL is the relative and absolute error tolerance. The algorithm inputs we used were $x_0 = 1$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$.

Table 3			
Method\Tol	1E-4	1E-5	1E-6
Algorithm 2	209	240	291
Algorithm 1	216	445	869
ode23	153	312	652

Table 4			
Method\Tol	1E-4	1E-5	1E-6
Algorithm 2	255	285	361
Algorithm 1	838	1581	2997
ode23	513	1065	2254

Figure 5.6: Table 1 corresponds to the equation $\dot{x}(t) = (\cos(2t) - 0.1)x(t)$ and Table 2 corresponds to $\dot{x}(t) = (4\cos(6t) - 0.1)x(t)$. The tables record the number of steps taken by Algorithms 2 and 1 and Matlab's ode23 for comparison for various values of TOL where TOL is the relative and absolute error tolerance. The algorithm inputs we used were $x_0 = 1$, $h_{\min} = 1E-4$, $h_{\max} = 1E-1$, $h_0 = \sqrt{h_{\max} h_{\min}}$, $t_0 = 0$, $T = 20$.

5.3 Forced Van der Pol Equation

Consider the forced Van der Pol oscillator from Van der Pol, B. (1927):

$$\ddot{x}(t) - \mu(1 - x(t)^2)\dot{x}(t) + x(t) - F\sin(\omega t) = 0 \tag{5.4}$$

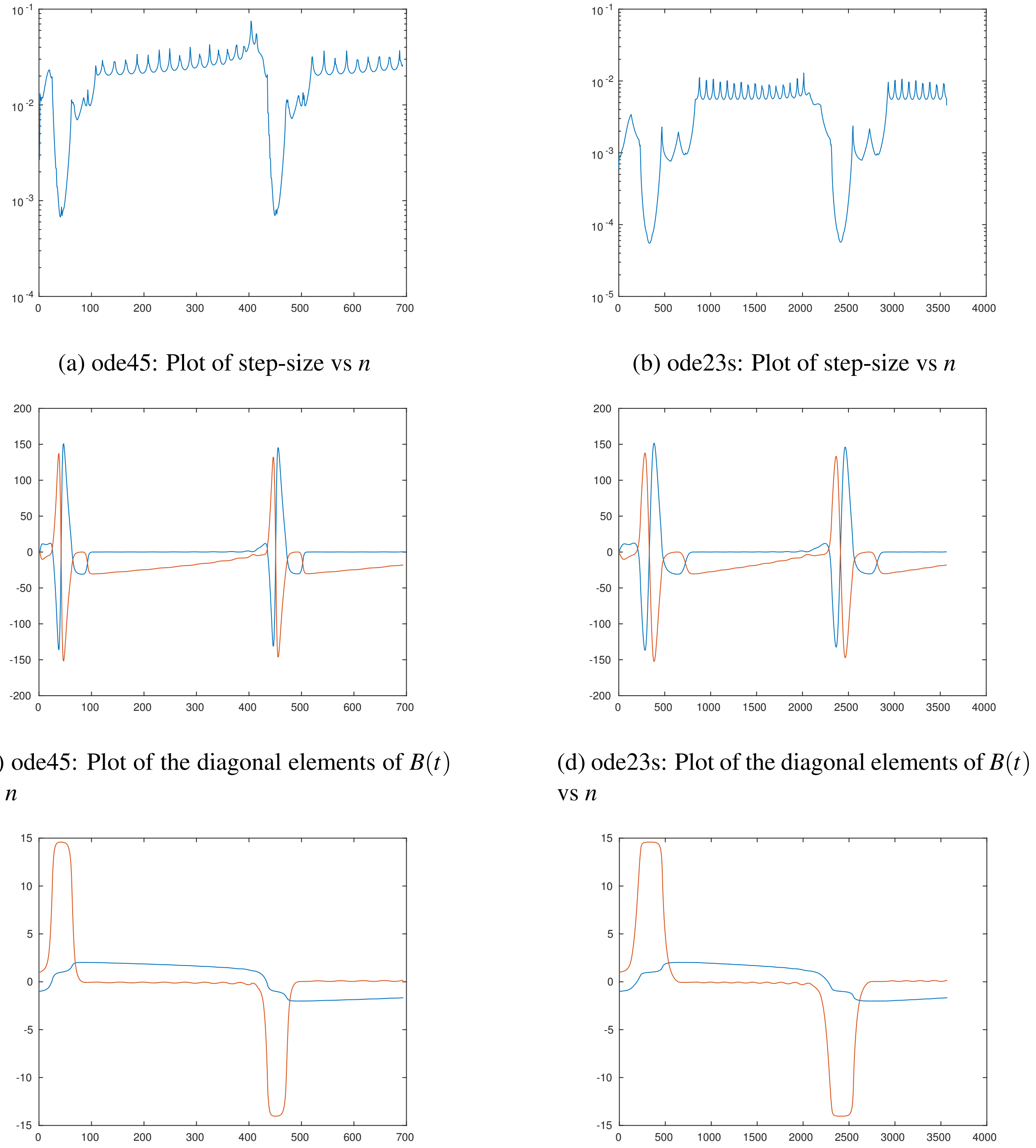
where μ, F, and ω are real constants. By introducing the relation $y(t) = \dot{x}(t)$ the equation (5.4) can be expressed as the equivalent two-dimensional system

$$\begin{cases} \dot{x}(t) = y(t) \\ \dot{y}(t) = \mu(1 - x(t)^2)y(t) - x(t) + F\sin(\omega t). \end{cases} \tag{5.5}$$

For large values of μ and $F = 0$, the system (5.5) is a classically stiff nonlinear equation; the ratio of real parts of the smallest and largest eigenvalues of the system linearized at the equilibrium

$(0,0)^T$ is large when $F = 0$ and μ is large. When μ is small the system is not stiff in the classical sense, although the step-sizes may need to be taken quite small along certain time intervals..

In Figure 5.5 we present the results of a Matlab experiment where, using both the nonstiff solver ode45 and the stiff solver ode23s, we solve (5.5) coupled the the Q-equation (2.5) so that we are able to form approximations to $B(t)$. The results show that periodically both the stiff and the nonstiff solver must reduce the step-size from approximately 10^{-2} to either 10^{-4} or 10^{-3} respectively. Contrary to what might be expected, this step-size restriction is most severe where the solution is flat relatively flat and occurs between, but not during, intervals over which rapid growth or decay happen. A better indicator of when the step-size restriction occurs is the magnitude of the diagonal elements of $B(t)$. This is consistent with the theoretical results of Section 3.1, where the local error in addition to the inverse of the quantities $I_n^i := \int_{t_n}^{t_{n+1}} B(\tau)d\tau$ must be controlled to control the numerical stability. This suggests the efficiency of integrators can be improved by designing methods that can handle spikes in the values of I_n^i.



(a) ode45: Plot of step-size vs n

(b) ode23s: Plot of step-size vs n

(c) ode45: Plot of the diagonal elements of $B(t)$ vs n

(d) ode23s: Plot of the diagonal elements of $B(t)$ vs n

(e) ode45: Plot of the components of the numerical solution vs n

(f) ode23s: Plot of the components of the numerical solution vs n

Figure 5.7: Results of a Matlab experiment of the solution of (5.5) coupled with (2.5) using ode45 and ode23s with absolute and relative error tolerances of 10^{-6} and the parameter values of $\mu = 10$, $F = 1$, and $\omega = 2\pi$. Figures vs n where n denotes the n^{th} time-step of the numerical solution.

Chapter 6

Conclusion

In this work we have developed a stability theory for one-step and linear multistep methods approximating the solution of time-dependent ODE IVPs using Lyapunov exponents theory. The majority of our effort was spent on the stability analysis of one-step methods approximating nonautonomous linear problems. Analogous stability theories for linear multistep methods and for numerical methods solving nonlinear problems were also developed. Our theory explains how a solver with a standard step-size selection strategy can fail to produce a decaying numerical solution to an asymptotically contracting, time-dependent, linear problem that falls outside the previously existing framework for numerical stability. We are able to apply the theory we developed to devise an efficient method for selecting step-size that stabilizes a solver in this context.

There are still many open avenues of investigation in the stability theory for numerical methods solving time-dependent IVPs. The natural continuation of this work would be to analyze the preservation of the exponential dichotomy or Sacker-Sell spectrum by one-step and linear multistep methods. As mentioned in Section 3.2 there are interesting similarities between the conditioning of BVP solvers and the step-size restriction due to strength of the integral separation in IVP solvers and it would be an interesting research project to investigate this connection. Another natural continuation of this work would be to investigate the stability of numerical methods for the solution of PDE IVPs. Such a theory would undoubtedly make heavy use of the QR perturbation theory on infinite dimensional Hilbert spaces developed in Badawy, M. & Van Vleck, E.S. (2012) and would involve making restrictions on the spatial discretization as well as the size of the time-steps. We would also like to investigate the theory for asymptotically contracting linear systems more deeply and determine whether or not asymptotic contraction holds generically for linear systems whose Lyapunov exponents are all negative.

The analysis of linear multistep methods in this work is essentially treated as a corollary of the one-step theory. It follows from the application of invariant manifold theory and relies on an $\mathscr{O}(h)$ Lipschitz estimate even when the method has local truncation error of order $p > 1$. It would be desirable to develop a time-dependent theory for linear multistep methods that does not resort to treating them as one-step methods or use $\mathscr{O}(h)$ estimates. This would relax the additional step-size restriction and also provide a way of analyzing the stability of linear mulitistep methods approximating nonlinear problems. A way forward along these lines may be as follows. In classical time-independent stability theory for linear multistep methods the Kreiss Matrix Theorem and its corollaries are used to bound products of companion matrices that are formed from linear multistep methods applied to time-independent scalar test problems. It should be possible to apply techniques from QR perturbation theory to develop a time-dependent theory for bounding the products of companion matrices that result from solving a time-dependent scalar test problem. This would facilitate the analysis of the stability of linear multistep methods solving time-dependent problems without making use of invariant manifold theory and underlying one-step methods.

References

Adrianova, L. (1995). *Introduction to Linear Systems of Differential Equations*, volume 146. AMS, Providence, R.I.

Ascher, U., Mattheij, R., & Russel, R. (1988). *Numerical solution of boundary value problems of ordinary differential equations*. Prentice hall, Englewood Cliffs, New Jersey.

Axelsson, O. (1969). A class of A-stable methods. *BIT*, 9, pp. 185–199.

Badawy, M. & Van Vleck, E.S. (2012). Perturbation theory for the approximation of stability spectra by QR methods for sequences of linear operators on a Hilbert space. *Linear Algebra and its Applications*, 437, pp. 37–59.

Bogacki, P. & Shampine, L. (1989). A 3(2) pair of Runge-Kutta formulas. *Applied Mathematical Letters*, 2, pp. 321–325.

Boutelje, B.R. & Hill, A.T. (2010). Nonautonomous stability of linear multistep methods. *IMA J. Numer. Anal.*, 30(2), pp. 525–542.

Butcher, J.C. (1975). A stability property of implicit Runge-Kutta methods. *BIT*, (pp. vol. 27, pp. 358–361.).

Butcher, J.C. (1987). The equivalence of algebraic stability and AN-stability. *BIT*, (pp. vol. 27, pp. 510–533).

Butcher, J.C. & Burrage, K. (1979). Stability criteria for implicit Runge-Kutta methods. *SIAM J. Numer. Anal.*, 16, vol. 27, pp. 46–57.

Calvo, M. & Quemada, M. Mar (1982). On the stability of rational Runge-Kutta methods. *J. Comp. Appl. Math.*, 8, pp. 289–293.

Chung Y.-M., Steyer, A.J., Tubbs, M., Van Vleck, E.S., & Vedantam, M. (2016). Global error analysis and inertial manifold reduction. *J. Comp. Appl. Math.*, 307.

Chung, Y.-M., Steyer, A.J., & Van Vleck, E.S. (2016). Nonautonomous inertial manifold reduction. *Preprint*.

Coppel, W. (1978). *Lecture Notes in Mathematics # 629: Dichotomies in Stability Theory*, volume 629. Springer-Verlag, Berlin.

Crouzeix, M. (1979). Sur la B-stabilité des mé methodes de Runge-Kutta. *Numer. Math.*, 32, pp. 75–82.

Dahlquist, G. (1963). A special stability problem for linear multistep methods. *BIT*, 3, pp. 27–43.

Dieci, L. & Van Vleck, E.S. (1999). Computation of orthonormal factors for fundamental matrix solutions. *Numer. Math.*, 83, pp 599–620.

Dieci, L. & Van Vleck, E.S. (2005). On the error in computing Lyapunov exponents by QR methods. *Numer. Math.*, 101, pp 619–642.

Dieci, L. & Van Vleck, E.S. (2009). On the error in QR integration. *SIAM J. Numer. Anal.*, 46, no 3., pp 1166–1189.

Dieci, L. & Van Vleck, E.S. (1995). Computation of a few Lyapunov exponents for continuous and discrete dynamical systems. *Appld. Numer. Math.*, 17, pp. 275–291.

Dieci, L. & Van Vleck, E.S. (2002). Lyapunov and other spectra: A survey. *Collected Lectures on the Preservation of Stability under Discretization, A Volume Published by SIAM*, (pp. pp. 197–218.).

Dieci, L. & Van Vleck, E.S. (2003). Lyapunov spectral intervals: Theory and computation. *SIAM J. Numer. Anal.*, 40, pp. 516–542.

Dieci, L. & Van Vleck, E.S. (2006). Perturbation theory for approximation of Lyapunov exponents by QR methods. *J. Dynam. Diff. Eq.*, 18, pp 815–840.

Dieci, L. & Van Vleck, E.S. (2007). Lyapunov and Sacker-Sell spectral intervals. *J. Dyn. Diff. Eqn.*, 19, pp. 263–295.

Dieci, L., Van Vleck E.S., & Cinzia, E. (2010). Exponential dichotomy on the real line: Svd and qr methods. *J. Differential Equations*, 248, pp. 287–308.

Dubinkin, S., Frank, J., Leeuw B., Steyer, A.J., Tu, X., & Van Vleck, E.S. (2016). Hybrid shadowing based data assimilation with assimilation in the non-strongly stable subspace. *preprint*.

Ehle, B.L (1968). High order A-stable methods for the numerical solution of systems of DEs. *BIT*, 8, pp. 276–278.

Ehle, B.L. (1973). A-stable methods and Padé approximations to the exponential. *SIAM J. Math. Anal.*, 4, pp. 671–680.

Eirola , T. & Nevanlinna, O. (1988). What do multistep methods approximate? *Numer. Math.*, 53, pp. 559–569.

González, C. & Palencia, C. (2000). Stability of Runge-Kutta methods for quasilinear parabolic problems. *Math. Comp.*, 69, pp. 609–628.

González, C.and Palencia, C. (1999). Stability of Runge-Kutta methods for abstract time-dependent parabolic problems: The Hölder case. *Math. Comp.*, 68, pp. 73–89.

Gustafsson, K., Lundh, M., & Söderlind, G. (1988). A PI stepsize control for the numerical solution of ordinary differential equations. *BIT*, 28, pp. 270–287.

Hairer, E. (1980). Unconditionally stable explicit methods for parabolic equations. *Numer. Math.*, 35, pp. 57–68.

Hairer, E. & Wanner, G. (1991). *Solving Ordinary Differential Equations II: Stiff and Differential-algebraic problems*. Springer-Verlag, Berlin.

Hairer, E., Nørsett, S.P., & Wanner, G. (1987). *Solving Ordinary Differential Equations I: Nonstiff problems*. Springer-Verlag, Berlin.

Hall, G. (1985). Equilibrium states of runge-kutta schemes. *ACM Trans. Math. Software*, 11, pp. 289–301.

Hall, G. (1986). Equilibrium states of runge-kutta schemes, part ii. *ACM Trans. Math. Software*, 12, pp. 183–192.

Hoyer-Leitzel, A., Nadeau, A., Roberts, A., & Steyer, A.J. (2016). Connections between rate-induced tipping and nonautonomous bifurcation. *preprint*.

Humphries, A.R. & Stuart, A.M. (1998). Dynamical systems and numerical analysis.

Kirchgraber, U. (1986). Multi-step methods are essentially one-step methods. *Numer. Math.*, 48, pp. 85–90.

Kreiss, H.-O. (1978). Difference methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.*, 15(1), pp. 21–58.

Lambert, J. (1974). Two unconventional classes of methods for stiff systems. In R. Willoughby (Ed.), *Stiff differential systems*. New York: Plenum Press.

Lorenz, E. (1963). Deterministic nonperiodic flow. *J. Atm. Sci.*, 20, pp. 130–141.

Nevanlinna, O. (1977). On the behaviour of global errors at infinity in the numerical integration of stable initial value problems. *Numer. Math.*, 28, pp. 445–454.

Nevanlinna, O. & Sipilä, A.H. (1974). A nonexistence theorem for explicit A-stable methods. *Math. Comp.*, 28, pp. 1053–1055.

Nevanlinna, O. & Jeltsch, R. (1982). Stability and accuracy of time discretizations for initial value problems. *Numer. Math.*, 40, pp. 245–296.

Nevanlinna, O. & Liniger, W. (1978). Contractive methods for stiff differential equations I. *BIT*, 18, pp. 457–474.

Nevanlinna, O. & Liniger, W. (1979). Contractive methods for stiff differential equations II. *BIT*, 19, pp. 53–72.

Palmer, K. (1979). The structurally stable systems on the half-line are those with exponential dichotomy. *J. Diff. Eqn.*, 33, pp. 16–25.

Scherer, R. (1979). A necessary condition for B-stability. *BIT*, 19, pp. 111–115.

Steyer, A.J. & Van Vleck, E.S (2015). A step-size selection strategy for explicit Runge-Rutta methods based on Lyapunov exponent theory. *J. Comp. Appl. Math.*, 292.

Steyer, A.J. & Van Vleck, E.S (2016a). A Lyapunov exponent based stability theory for general linear methods. *preprint*.

Steyer, A.J. & Van Vleck, E.S (2016b). A Lyapunov exponent based stability theory for one-step ordinary differential equation initial value problem solvers. *preprint*.

Stoffer, D. (1993). General linear methods: connection to one-step methods and invariant curves. *Numer. Math.*, 64, pp. 395–407.

Stoffer, D. & Nipp, K. (1992). Attractive invariant manifolds for maps: existence, smoothness and continuous dependence on the map. research report. *Applied Mathematics, ETH-Zurich*, (pp. pp. 92–111).

Van der Pol, B. (1927). Forced oscillations in a circuit with non-linear resistance (receptions with reactive triode). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science Ser. 7*, 3, pp. 65–80.

Van Vleck, E.S. (2010). On the error in the product QR decomposition. *SIAM J. Matrix Anal. Appl.*, 31(4), pp. 1775–1791.

Verwer, J.G. & Dekeker, K. (1983). Two unconventional classes of methods for stiff systems. In K. Strehmel (Ed.), *Proceedings zweiter Seminar Numerische Behandlung von Differentialgleichungen*. Martin Luther Unversität Halle-Wittenberg.

Wambecq, A. (1978). Rational Runge-Kutta methods for solving sytems of ordinary differential equations. *Computing*, 20, pp. 333–342.

Wanner, G. (1976). A short proof on nonlinear A-stability. *BIT*, 16, pp. 226–227.