Non-standard finite difference methods in dynamical systems

by

Phumezile Kama

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

in the Faculty of Natural and Agricultural Sciences

in the Department of Mathematics and Applied Mathematics

University of Pretoria

Pretoria

April 2009

Supervisor: Professor Jean M-S Lubuma

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am indebted to Professor Jean M-S Lubuma, my supervisor, for introducing me to the theory of non-standard finite difference methods and for his constant support and guidance through the early period of chaos and confusion. This thesis would not have been possible without the tremendous efforts put forth by my supervisor who has devoted much of his precious time to read endless drafts, provided me with technical corrections, constructive feedback and countless suggestions for improvement of this work.

A special word of gratitude is extended to Professor R. Anguelov for his interest in the work and generating some of the figures in this thesis.

I would like to acknowledge the Numerical Analysis of Differential Equations research group at the University of Pretoria for their joint work and friendly encouragement, which gave me a better perspective on my own results.

A hearty thank you goes to Dr TA Tshifhumulo, Dr GH Maluleke and Mr PWM Chin for proofreading this thesis.

I would like to thank the National Research Foundation of South Africa for financial support.

Of course, I am grateful to my family, for their patience and *love*. Without them this work would not have been fruitful.

Dedication

To all who taught me over the years.
"I humbly thank you; well, well, well." Shakespeare, *Hamlet* (1623).

Declaration

I, the undersigned, hereby certify that the thesis submitted herewith for the degree Philosophiae Doctor to the University of Pretoria contains my own, independent work and has not been submitted for any degree at any other university.

Signature: _____

Phumezile Kama

Date: _____

Abstract

This thesis analyses numerical methods used in finding solutions of differential equations. Numerical methods are viewed as discrete dynamical systems that give useful information on continuous dynamical systems defined by systems of (ordinary) differential equations. We analyse non-standard finite difference schemes that have no spurious fixed-points compared to the dynamical system under consideration, the linear stability/instability property of the fixed-points being the same for both the discrete and continuous systems. We obtain a sharper condition for the elementary stability of the schemes. For more complex dynamical systems which are dissipative, we design schemes that replicate this property.

Furthermore, we investigate the impact of the above analysis on the numerical solution of partial differential equations. We specifically focus on reaction-diffusion equations that arise in many fields of engineering and applied sciences. Often their solutions enjoy the following essential properties: Stability/instability of the fixed points for the space independent equation, the conservation of energy for the stationary equation, and boundedness and positivity.

We design new non-standard finite difference schemes which replicate these properties. Our construction make use of three strategies: the renormalization of the denominator of the discrete derivative, nonlocal approximation of the nonlinear terms and simple functional relation between step sizes. Numerical results that support the theory are provided.

Chapter 1

Introduction

Our main interest in this thesis is the study of numerical methods for dynamical systems defined by (ordinary) differential equations. Problems as diverse as the simulation of planetary interactions, fluid flows [10] and mechanics [43], chemical reactions [16],[40], biological pattern formulation [2], [18], [33] and economic markets can all be modelled as dynamical systems [41]. For further applications of dynamical systems see [44]. In most of the systems modelled, all rates of change are assumed to be time independent, which makes the corresponding system autonomous.

Dynamical systems are concerned primarily with making qualitative study about the behaviour of systems which evolve in time given knowledge about the initial state of the system itself. It is important to know and study essential qualitative properties of the systems or more precisely their dynamics. Such properties include among others: the type of fixed points, oscillatory solutions, monotonicity of solutions, conservation of energy, dissipativity or dispersion of solution, positivity and boundedness of solutions. Our standard reference for dynamical systems is Stuart and Humphries [41] while Lambert [22] will also be used for numerical methods for ordinary differential equations. The framework of the study will include a wide range of concrete linear and non-linear models such as: logistic equation, decay equation, Hamiltonian system in ordinary differential equations as well as the Fisher equation, the reaction-diffusion equation in partial differential equations.

Existence theory is extensively developed for differential equations. However, most differential equations have no analytical solutions. As a result numerical methods are of fundamental importance in gaining understanding of dynamical systems. For contemporary numerical analysts, the understanding of differential equations from numerical methods is often limited to the study of their consistency, (zero-) stability and convergence. Unfortunately such classical numerical methods do not guarantee that the dynamics of the systems are replicated. This explains why we use the monograph [41] as our standard reference on dynamical systems, since it is one of a few classical books emphasizing the similar properties of the exact solutions that numerical schemes exhibit.

To be more explicit in this introduction, we consider the following differential equation

$$\frac{dy}{dt} = y^2(1-y) \equiv f(y). \tag{1.0.1}$$

Equation (1.0.1) is an elementary model for combustion [34]. Despite the simple nature of (1.0.1), its solution cannot be written in a closed form. The solution is expressed in the implicit form

$$\ln\left(\frac{|y|}{|y-1|}\right) + \frac{1}{y} = t + C, \tag{1.0.2}$$

where C is a constant. This equation defines a dynamical system on $(-\infty, +\infty)$ with an asymptotically stable fixed point $y = 1$ and an unstable fixed point $y = 0$. Full analysis shows that the point $y = 0$ is attracting the solutions below it and repelling those above it. (The concepts used here will be made clear in the next chapter). All these properties that represent the exact solution of (1.0.2) are visualised in Fig.1.1.

We employ for (1.0.1), the forward Euler method

$$\frac{y_{n+1} - y_n}{\Delta t} = y_n^2(1-y_n), \tag{1.0.3}$$

and the Runge-Kutta method

$$\frac{y_{n+1}-y_n}{\Delta t} = y_n + \frac{\Delta t}{2}\left[f(y_n) + f(y_n + \Delta t f(y_n))\right]. \tag{1.0.4}$$

The two classical schemes (1.0.3) and (1.0.4) are consistent, zero-stable and thus convergent. However, the discrepancy between the numerical solution by these methods and the exact solution is evident as can be seen in Figs.1.2 and 1.3. We use $\Delta t = 1.8$ in both schemes.

Our aim is to design numerical schemes that give reliable simulations, which preserve as much as possible the intrinsic properties of the dynamical systems without any limitation on the value of time step size Δt. We shall do this by considering the non-standard finite difference method which was introduced by RE Mickens [26] more than two decades ago. This approach takes advantage of specific properties of solutions of involved differential equations.

For the above mentioned combustion model (1.0.1), Fig.1.4 shows that the non-standard finite difference scheme

$$\frac{y_{n+1}-y_n}{1-e^{-\Delta t}} = y_n^2(1 + y_n - 2y_{n+1}), \tag{1.0.5}$$

proposed by Anguelov and Lubuma [8], displays better the properties of the exact solution.

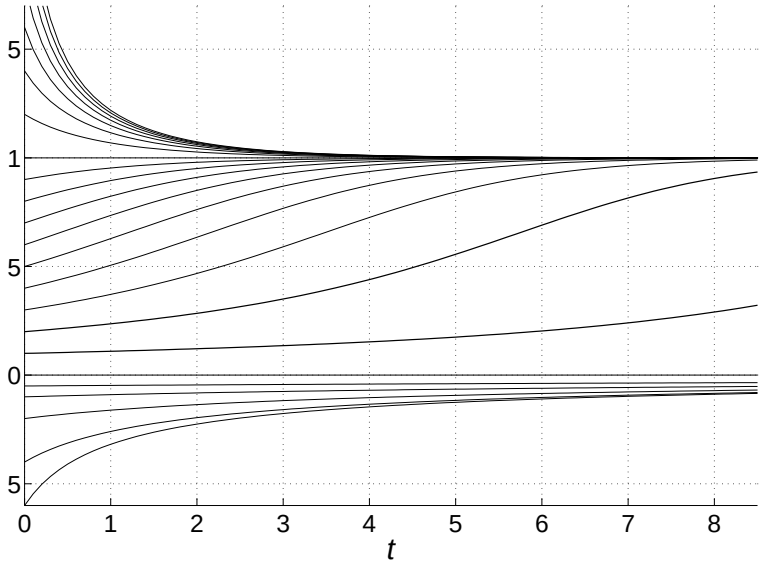

Figure 1.1: Exact solution

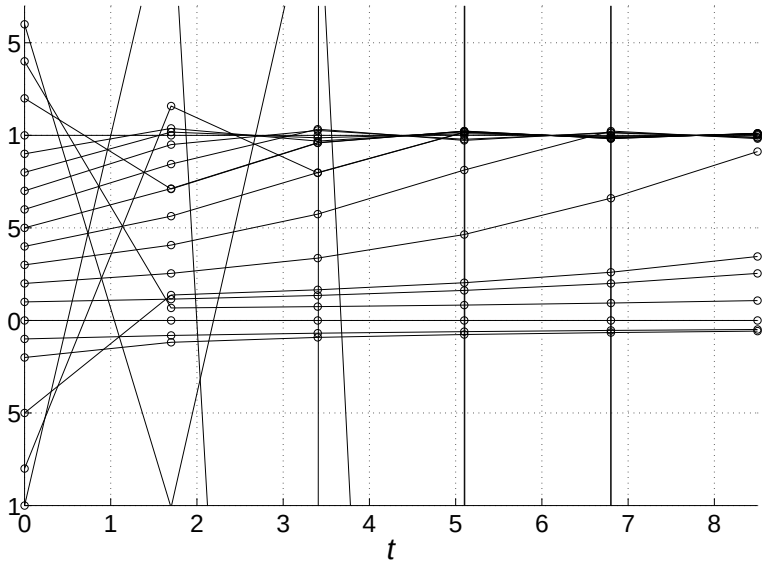

Figure 1.2: Forward Euler method

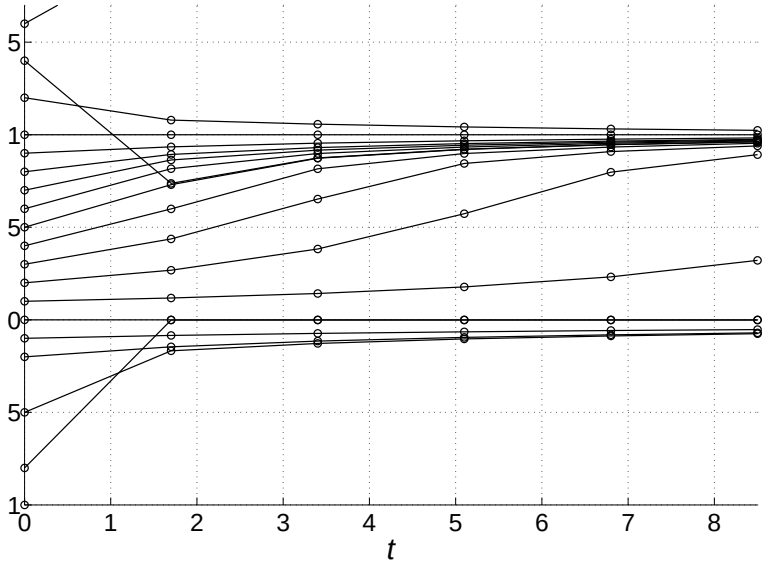

Figure 1.3: Runge-Kutta method

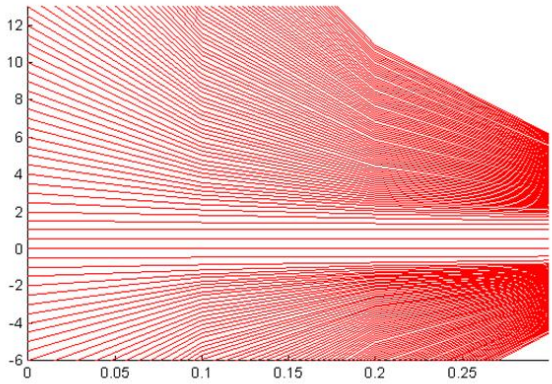

Figure 1.4: Non-standard method

Non-standard finite difference techniques developed by Mickens, have laid the foundation for designing methods that preserve the dynamics, especially the stability property of fixed points of the approximated differential system. The design of the non-standard finite difference method starts mostly with the concept of exact scheme. A major advantage of having an exact scheme for a differential equation is that questions related to the usual considerations of consistency, stability and convergence do not arise. It is to be noted that any method that is not standard could be considered non-standard. However, in this thesis when we talk about non-standard finite difference schemes, we consider those that are based on Mickens' methodology and rules as explained in the survey paper [35].

Since the publication of the monograph [26], which is the first book on this exciting topic, several authors have contributed to the study of non-standard finite difference methods. Anguelov and Lubuma [7] provided some mathematical justification for the success of empirical procedures used so far. These authors have unambiguously defined non-standard finite difference methods using two of Mickens' rules.

The edited volumes [17], [28] and [29] contain a wide range of applications of the non-standard finite difference methods, (for example, mathematical epidemiology, reaction-diffusion equations, non-smooth mechanics, singular perturbation problems, conservation law, etc). In addition to these, we mention the following works where the non-standard finite difference schemes have shown great potential: [1], [6], [7], [14] and [26].

For this thesis to be relatively self-contained, we dedicate considerable time to study classical concepts regarding dynamical systems and finite difference methods. In particular, the concept of absolute stability of linear multi-step and Runge-Kutta methods is sufficiently reviewed in view of the elementary stability which is the minimum qualitative property that non-standard finite difference methods must satisfy.

The comment made earlier about the reliable scheme (1.0.5) raises the following concerns which constitute the main focus of the thesis:

- What is a non-standard finite difference method for a dynamical system?

- How to construct a non-standard finite difference method for a dynamical system?
- How powerful are non-standard finite difference methods compared to standard finite difference methods that are used for dynamical systems?
- How can numerical methods be viewed as discrete dynamical systems of the continuous dynamical systems they approximate?
- What is the impact of the non-standard finite difference method on concrete examples of dynamical system?
- How does the study carry over to dynamical systems related to partial differential equations for dynamical systems, for example the reaction-diffusion equations?

This thesis elaborates, with extension in some cases, the author's results in the following papers: [3], [4], [5] and [6].

The thesis is organized as follows: Chapter 2 deals with the review of basic concepts, definitions and notation relating to dynamical systems which we will be using throughout this thesis. Continuous dynamical systems defined by ordinary differential equations are presented in Section 2.2 and their discrete counterparts are discussed in Section 2.3. In each case, we present qualitative properties of dynamical systems that are of interest in our work. These include, inter alia, invariant sets, fixed points, hyperbolic fixed points, linear stability and dissipativity.

In Chapter 3 we introduce finite difference schemes for ordinary differential equations. In Section 3.2, consistency, zero-stability and convergence of finite difference methods are discussed. We give a short presentation of two classical methods, namely, the linear multi-step method in Section 3.3 and the Runge-Kutta method in Section 3.4. The numerical methods are also required to behave asymptotically, like the solutions of the decay equation. This is the essence of the concept of absolute stability addressed in Section 3.5. In Section 3.6 we consider the numerical methods that define discrete dynamical systems. Finally,

the analysis in Section 3.7 is restricted to theta methods, which will be the focus for the rest of this thesis.

The first set of the author's main contribution in this thesis appear in Chapter 4. Firstly, we extend the classical theta methods. In Section 4.2, we analyse non-standard finite difference schemes that have no spurious fixed-points compared to the dynamical system under consideration, the linear stability/instability property of the fixed-points being the same for both the discrete and continuous systems. We obtain a sharper condition for the elementary stability of the schemes, a topic discussed in Section 4.3. For more complex dynamical systems which are dissipative, we design schemes that replicate this property as presented in Section 4.4. Lastly, in Section 4.5, we consider a specific class of dynamical systems which is equivalent to the simplest model of Hamiltonian systems that occur in classical mechanics. We design a non-standard finite difference scheme that replicates the underlying principle of conservation of energy. Here we use Mickens' rule about nonlocal approximation of nonlinear terms.

Chapter 5 is dedicated to a detailed analysis of the author's results given in [6]. Our point of departure is the Fisher equation, in Section 5.2, which enjoys a positivity and boundedness property. Then we move to general reaction-diffusion equations for which we construct non-standard theta methods in Section 5.3. In Section 5.4, we design non-standard finite difference schemes which are elementary stable in the limit case of space independent variable and which are stable with respect to the principle of conversation of energy in the stationary case. Furthermore, we show that our schemes replicate the positivity and boundedness properties under a more simpler functional relation between the time and space step sizes (compared to the literature).

As an alternative approach, Section 5.5 deals with approximations of the space variable by the spectral method, while the time variable is approximated via non-standard finite difference scheme. This results in what we call coupled spectral and non-standard methods which replicates the essential properties of the exact solutions.

In the last chapter, we provide concluding remarks, and a summary of our findings, a discussion on how our work fits in the literature and

possible extensions. Throughout the main chapters of the thesis, we provide numerical tests that support the theory and show superiority and reliability of our schemes compared to the classical ones.

Chapter 2

Dynamical Systems

2.1 Introduction

Dynamical systems are found in various fields of science. Usually they are given by an analytical specification or as sampled data. Dynamical systems are mainly represented by a state that evolves in time. The input as well as the current state of a dynamical system determine the evolution of the system.

An important characteristic of a dynamical system is whether it is continuous or discrete. Continuous systems (often called flows) are given by differential equations whereas discrete systems (often called maps) are specified by difference equations.

There are many possible ways to analyse such systems, for example, analysing their long term behaviour. For the analysis, it is very important to know whether a dynamical system is linear or not. Nonlinear systems typically have intricate dynamical behaviour.

The general setting of this thesis is that of continuous dynamical systems defined by a system of autonomous differential equations. We present continuous dynamical systems in the next section. After specifying general concepts and terminology, we give existence results. Thereafter, we investigate properties of dynamical systems which constitute the main qualitative properties of interest throughout this thesis. These are the stability of fixed points and their dissipative nature. Section 2.3 provides the discrete counterpart of the above study for discrete dynamical systems.

2.2 Continuous Dynamical Systems

We recall that Stuart and Humphries [41] is our standard reference for dynamical systems. Most of the classical concepts given below can be found there.

2.2.1 Generalities

Throughout this thesis, we shall be concerned with the initial-value problem for an autonomous first-order system of ordinary differential equations

$$Dy := \frac{dy}{dt} = f(y); \qquad y(0) = y_0, \tag{2.2.1}$$

where $y = y(t) = [{}^1y \cdots {}^my]^T : [0,\infty) \to \mathbb{R}^m$ is unknown, while $f = [{}^1f \cdots {}^mf]^T : \mathbb{R}^m \to \mathbb{R}^m$ and $y_0 = [{}^1y_0 \cdots {}^my_0]^T \in \mathbb{R}^m$ are given. Implicitly, we assume that f satisfies the smoothness properties that are needed. Whenever it is necessary, we will be explicit about the smoothness of f. The space $\mathbb{R}^m$ is equipped with the usual Euclidean structure through the norm $||\bullet||$ and the inner product $\langle \bullet, \bullet \rangle$.

We begin by defining a dynamical system on a subset $E \subseteq \mathbb{R}^m$.

Definition 2.2.1. *The equation (2.2.1) is said to define a* ***dynamical system*** *on a subset $E \subseteq \mathbb{R}^m$ if, for every $y_0 \in E$, there exists a unique solution of (2.2.1) which is defined for all $t \in [0,\infty)$ and $y(t) \in E$ for all $t \geq 0$.* ∎

The fact that $y(t)$ is a solution of (2.2.1) on $[0,\infty)$ implies at least the following smoothness: $y(t)$ is differentiable on $(0,\infty)$ and continuous on $[0,\infty)$.

We now introduce the concept of evolution semigroup operator for a dynamical system.

Definition 2.2.2. *For a dynamical system on E, we define its* ***evolution semigroup operator*** *or solution operator to be the map $S(t) : E \to E$ such that $y(t) = S(t)y_0$.* ∎

The terminology in Definition 2.2.2 is motivated by the following properties that can easily be checked:

i. $S(t+s) = S(t)S(s) = S(s)S(t) \qquad \forall t, s \geq 0,$

ii. $S(0) \equiv I$, the identity operator.

The evolution semigroup operator $S(t)$ is merely a convenient notation for advancing the solution through time t. In fact, for $y_0 \in E$ the set

$$\Gamma^+(y_0) := \{S(t)y_0; t \in [0, \infty)\} \subset E \tag{2.2.2}$$

is called the (positive or forward) orbit of y_0. The terminology trajectory is also used for orbit.

At this stage we need to discuss sufficient conditions for (2.2.1) to define a dynamical system. Firstly, we consider the commonly known condition stated in the following definition.

Definition 2.2.3. *A function $f : \mathbb{R}^m \to \mathbb{R}^m$ is said to be Lipschitz on $B \subset \mathbb{R}^m$ with Lipschitz constant $L \geq 0$ if*

$$\|f(x) - f(y)\| \leq L||x - y|| \qquad \forall x, y \in B.$$

*If f is Lipschitz on $\mathbb{R}^m$, then f is said to be **globally Lipschitz**. If f is Lipschitz on every bounded subset of $\mathbb{R}^m$, then f is said to be **locally Lipschitz**.* ∎

The concept of Lipschitzian functions is important in the proof of existence and uniqueness results for many problems in mathematics (see for example [48]). In our specific context, we have the following theorem.

Theorem 2.2.4. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be globally Lipschitz. Then there exists a unique solution $y(t)$ to (2.2.1) for all $t \geq 0$. Hence (2.2.1) defines a dynamical system on $\mathbb{R}^m$.*

Theorem 2.2.4 and its two corollaries below are well-known results. Given their importance in this work, we outline, for convenience, their proofs.

Proof. We employ the Banach contraction principle, see [48]. To this end, we first introduce the space $C_k(0,\infty;\mathbb{R}^m)$ consisting of continuous vector-valued functions $y:[0,\infty)\to\mathbb{R}^m$ such that

$$||y||_{C_k(0,\infty;\mathbb{R}^m)} := \sup_{t\geq 0} e^{-kt}||y(t)|| < \infty,$$

where the parameter $k>0$ will be specified shortly. It is clear that $C_k(0,\infty;\mathbb{R}^m)$ equipped with the norm $||\bullet||_{C_k(0,\infty;\mathbb{R}^m)}$ is a Banach space. Secondly, we consider the operator

$$\phi : C_k(0,\infty;\mathbb{R}^m)\to C_k(0,\infty;\mathbb{R}^m)$$

defined by

$$\phi(y)(t) = y_0 + \int_0^t f(y(s))ds.$$

It is equally clear that solving (2.2.1) is equivalent to finding fixed-points of the operator ϕ:

$$y = \phi y.$$

Using the Lipschitz condition in Definition 2.2.3 with Lipschitz constant L we have for $y, w \in C_k(0,\infty;\mathbb{R}^m)$:

$$\begin{aligned}
||\phi(y)(t)-\phi(w)(t)|| &\leq \int_0^t ||f(y(s))-f(w(s))||ds \\
&\leq L\int_0^t e^{ks}e^{-ks}||y(s)-w(s)||ds \\
&\leq L||y-w||_{C_k(0,\infty;\mathbb{R}^m)}\int_0^t e^{ks}ds \\
&= L(\frac{e^{kt}-1}{k})||y-w||_{C_k(0,\infty;\mathbb{R}^m)}.
\end{aligned}$$

Thus

$$e^{-kt}||\phi(y)(t) - \phi(w)(t)|| \leq \frac{L}{k}||y - w||_{C_k(0,\infty;\mathbb{R}^m)}$$

and

$$||\phi y - \phi w||_{C_k(0,\infty;\mathbb{R}^m)} \leq \frac{L}{k}||y - w||_{C_k(0,\infty;\mathbb{R}^m)}.$$

For the choice $k > L$, ϕ is a contraction and has therefore a unique fixed-point. □

In general, if f is only locally Lipschitz then the most we can achieve is local existence and uniqueness in the following sense:

Corollary 2.2.5. *Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is Lipschitz on the ball $\bar{B} = \bar{B}(y_0, r)$ with Lipschitz constant L_B. Consider the finite time*

$$T_B := \frac{r}{\sup_{x\in\bar{B}} ||f(x)||}.$$

Then, the initial-value problem (2.2.1) has a unique solution $y(t) \in \bar{B}$ for $t \in [0, T_B]$.

Proof. We replace $C_k(0,\infty;\mathbb{R}^m)$ by the set $C_k(0, T_B; \bar{B})$ of continuous functions $y : [0, T_B] \to \bar{B}$. Though not being a normed space, $C_k(0, T_B; \bar{B})$ is a complete metric space under the metric

$$d_k(y, w) = \sup_{0\leq t\leq T_B} e^{-kt}||y(t) - w(t)||.$$

For $y \in C_k(0, T_B; \bar{B})$, we have

$$\begin{aligned} ||\phi(y)(t) - y_0|| &\leq \int_0^t ||f(y(s))||ds \\ &\leq T_B \sup_{x\in\bar{B}} ||f(x)|| \\ &= r, \end{aligned}$$

which shows that the mapping ϕ defined earlier operates from $C_k(0, T_B; \bar{B})$ into $C_k(0, T_B; \bar{B})$. On the other hand, if $y \in C_k(0, T_B; \bar{B})$ and $w \in C_k(0, T_B; \bar{B})$, we easily obtain as in the proof of Theorem 2.2.4 that

$$d_k(\phi(y), \phi(w)) \leq \frac{L_B}{k} d_k(y, w).$$

Thus, for $k > L_B$, ϕ is a contraction. □

Whenever, some a priori bound holds for the solution, Corollary 2.2.5 permits us to obtain a global existence result in the following precise way.

Corollary 2.2.6. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be Lipschitz on an ϵ-neigbourhood E_ϵ of a bounded set $E \subseteq \mathbb{R}^m$. If for any $y_0 \in E$, the solution $y(t)$ of (2.2.1) satisfies $y(t) \in E$ for each time $t \geq 0$ where the solution exists, then (2.2.1) defines a dynamical system on E.*

Proof. For each $m = 0, 1, 2, ...$ define

$$T_m := \frac{m\epsilon}{\sup_{x \in E_\epsilon} ||f(x)||}$$

and consider the complete metric space $C_k(T_m, T_{m+1}; \bar{E}_\epsilon)$ of continuous functions $y : [T_m, T_{m+1}] \to \bar{E}_\epsilon$ equipped with the metric

$$d_k(y, w) = \sup_{t \in [T_m, T_{m+1}]} e^{-kt} ||y(t) - w(t)||.$$

Fix $y_0 \in E$. For $y \in C_k(T_m, T_{m+1}; \bar{E}_\epsilon)$, define the operator ϕ by

$$\phi(y)(t) = y_0 + \int_{T_m}^{t} f(y(s))ds.$$

As in the proof of Corollary 2.2.5, one can show that ϕ operates from $C_k(T_m, T_{m+1}; \bar{E}_\epsilon)$ into $C_k(T_m, T_{m+1}; \bar{E}_\epsilon)$ and ϕ is a contraction for the choice $k > L_{E_\epsilon}$ where L_{E_ϵ} is the Lipschitz constant of f on E_ϵ. This,

together with the assumption that the solution remains in E whenever it exists, implies that there exists a sequence $\{Y^m\}_{m\geq 0}$ of functions $Y^m : [T_m, T_{m+1}] \to E$ such that each Y^m is on $[T_m, T_{m+1}]$ the unique solution of the differential equation in (2.2.1) that satisfies the initial condition given recursively by

$$Y^0(0) = y_0$$

$$Y^m(T_m) = Y^{m-1}(T_m), \quad m = 1, 2, ...$$

Since

$$[0, \infty) = \bigcup_{m\geq 0} [T_m, T_{m+1}],$$

the function

$$y := \bigcup_{m\geq 0} Y^m : [0, \infty) \to E,$$

which is well defined in view of the above initial conditions is the unique solution of (2.2.1). Thus, (2.2.1) defines a dynamical system on E. □

The following inequality introduced by Gronwall in 1918, known as the Gronwall inequality, will be useful in the analysis of continuous dynamical systems.

Lemma 2.2.7 (Gronwall Inequality). *Let $z(t)$ be a real valued function on $[0, \infty)$ that satisfy*

$$z_t \leq az + b, \qquad z(0) = z_0,$$

for a,b constants. Then for $t \geq 0$

$$z(t) \leq z_0 e^{at} + \frac{b}{a}(e^{at} - 1), \qquad a \neq 0$$

and

$$z(t) \leq z_0 + bt, \qquad a = 0.$$

Note that an extension of the Gronwall Inequality that does not allow an exponential growth of z is known as the uniform Gronwall lemma and is provided in [44].

Theorem 2.2.8. *Suppose that (2.2.1) defines a dynamical system on $\mathbb{R}^m$ and that $f : \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz. Let $B \subset \mathbb{R}^m$ be a bounded set with the property*

$$S(t)B \subset B \quad for \quad t \in [0, T].$$

Then there exists a constant $c > 0$ depending on B and T such that

$$||S(t)y_0 - S(t)z_0|| \leq c||y_0 - z_0|| \quad \forall t \in [0, T] \quad \forall y_0, z_0 \in B. \tag{2.2.3}$$

Proof. For $y_0, z_0 \in B$, put $y(t) = S(t)y_0$ and $z(t) = S(t)z_0$ which belongs to B for $t \in [0, T]$. Using (2.2.1) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}||y(t) - z(t)||^2 &= \langle y_t - z_t, y - z\rangle \\ &= \langle f(y) - f(z), y - z\rangle \\ &\leq ||y - z||||f(y) - f(z)|| \\ &\leq L_B||y(t) - z(t)||^2 \end{aligned}$$

where $L_B \geq 0$ is the Lipschitz constant of f on B. Application of Gronwall inequality (Lemma 2.2.7) yields

$$||y(t) - z(t)||^2 \leq e^{2L_B t}||y_0 - z_0||^2 \quad \forall t \in [0, T].$$

This implies that

$$||y(t) - z(t)|| \leq e^{L_B T}||y_0 - z_0|| \quad \forall t \in [0, T].$$

□

Remark 2.2.9. When f is globally Lipschitz, the inequality (2.2.3) holds on replacing B by $\mathbb{R}^m$. The inequality (2.2.3) means that the solution of (2.2.1) depends continuously on the initial data or that the dynamical system is continuous with respect to initial data. This motivates the terminology "Lipschitz continuous dynamical system". In fact, in the rest of this thesis we deal with Lipschitz continuous dynamical systems, though the expression "Lipschitz continuous" is sometimes left out. ∎

2.2.2 Qualitative Properties

We will often be interested in the orbits or trajectories initiated at y_0 in any set $B \subseteq \mathbb{R}^m$ and the action of the evolution semigroup operator $S(t)$ on $B \subset \mathbb{R}^m$.

Definition 2.2.10. *For a dynamical system defined by (2.2.1) a set B is said to be **positively invariant** under $S(\bullet)$ if $S(t)B \subseteq B$ for all $t \geq 0$. Similarly, B is said to be **negatively invariant** if $S(t)B \supseteq B$ for all $t \geq 0$. If B is both positively and negatively invariant, so that $S(t)B = B$ for all $t \geq 0$, then B is said to be **invariant** under $S(\bullet)$.* ∎

Certain distinguished orbits play a prominent role in the qualitative theory of dynamical systems. The simplest of such orbits are fixed points which turn out to be also the simplest invariant sets.

Definition 2.2.11. *A point $\tilde{y} \in \mathbb{R}^m$ is called a **fixed point** of the dynamical system defined by (2.2.1) if $f(\tilde{y}) = 0$.* ∎

Remark 2.2.12. The terminology in Definition 2.2.11 is due to the fact that $\tilde{y} \in \mathbb{R}^m$ is a fixed point of (2.2.1) if and only if $\tilde{y}$ is a fixed point of the evolution semigroup operator, that is $S(t)\tilde{y} = \tilde{y}$. Other

terms often substituted for the term fixed point are equilibrium point, critical point, stationary point, rest point or steady state. We shall utilize the term fixed point exclusively. ∎

Given the simplicity of fixed-points as invariant sets of the dynamical system, it is natural to wonder how other trajectories compare to them. This is captured in the next definition.

Definition 2.2.13. *Let $\tilde{y} \in \mathbb{R}^m$ be a fixed point of the dynamical system (2.2.1). Then $\tilde{y}$ is said to be*

(i) ***stable*** *if, for any $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ such that if $y_0 \in B(\tilde{y}, \delta)$ then $y(t) \in B(\tilde{y}, \epsilon)$ for all $t \geq 0$;*

(ii) ***asymptotically stable*** *if (i) holds and in addition, $||y(t) - \tilde{y}|| \to 0$ as $t \to \infty$ for all $||y_0 - \tilde{y}||$ sufficiently small;*

(iii) unstable if (i) fails to hold. ∎

Remark 2.2.14. A fixed point $\tilde{y}$ is stable if all nearby solutions stay nearby. It is asymptotically stable if all nearby solutions not only stay nearby, but also tend to $\tilde{y}$ or are attracted by $\tilde{y}$. ∎

We now turn our attention to a special type of fixed point in the study of dynamical systems, called *hyperbolic fixed points*. To this end, we assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is of class C^1. Here and after, we denote the Jacobian matrix of f at the fixed point $\tilde{y}$ by

$$J \equiv Jf(\tilde{y}). \tag{2.2.4}$$

Definition 2.2.15. *If the matrix J has no eigenvalues with zero real parts, then $\tilde{y}$ is called* ***hyperbolic****. Otherwise the fixed point is called non-hyperbolic.* ∎

Suppose that f is a continuously differentiable function such that (2.2.1) generates a continuous dynamical system on $\mathbb{R}^m$. Moreover, suppose that $\tilde{y}$ is a hyperbolic fixed point of the dynamical system. If y solves (2.2.1) and setting $u = y - \tilde{y}$, we see by Taylor expansion of f about $\tilde{y}$ that

$$u' = f(u + \tilde{y}) = f(\tilde{y}) + Jf(\tilde{y})u + R(u). \tag{2.2.5}$$

That is,

$$u' = Ju + R(u) \tag{2.2.6}$$

where $R(u)/||u|| \to 0$ as $||u|| \to 0$. Because $R(u)$ is small when u is small, it is reasonable to believe that as $t \to \infty$ solutions of (2.2.6) behave similarly to solutions of

$$u' = Ju \tag{2.2.7}$$

for u near 0. Equivalently, it is reasonable to believe that solutions of (2.2.1) behave like solutions of

$$y' = J(y - \tilde{y}) \tag{2.2.8}$$

for y near $\tilde{y}$. Equation (2.2.7) or (2.2.8) is called the *linearisation* of (2.2.1) at $\tilde{y}$.

The belief expressed earlier is indeed confirmed by the following *Hartman-Grobman theorem* or linearisation theorem which is an important result about the local behaviour of dynamical systems in the neighbourhood of a hyperbolic fixed point.

Theorem 2.2.16 (Hartman-Grobman). *Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is of class C^1 and consider a hyperbolic fixed point $\tilde{y}$ of the dynamical system defined by (2.2.1). Then there exist $\delta > 0$, a neighbourhood $\mathcal{N}$ of the origin, and a homeomorphism $h : B(\tilde{y}, \delta) \to \mathcal{N}$ such that $v(t) := h(u(t))$ solves (2.2.7) if and only if $u(t)$ solves (2.2.6).*

Basically the theorem states that the behaviour as $t \to \infty$ of the solution of (2.2.1) near a fixed point is the same as the behaviour of the solution of its linearisation near the origin. Therefore when dealing

with such fixed points we can use the simpler linearisation of the system to analyse its behaviour. This observation leads us to the following result.

Theorem 2.2.17. *Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is of class C^1 and that $\tilde{y} \in \mathbb{R}^m$ is a hyperbolic fixed point of the dynamical system defined by (2.2.1). Then $\tilde{y}$ is asymptotically stable if and only if for $u(t) = e^{tJ}u_0$, solution of (2.2.7) with $||u_0|| := ||y_0 - \tilde{y}||$ small enough, we have*

$$\lim_{t\to\infty} u(t) = 0. \tag{2.2.9}$$

This is equivalent to

$$Re\lambda < 0, \qquad \forall \lambda \in \sigma(J), \tag{2.2.10}$$

where $\sigma(J)$ is the set of eigenvalues of the matrix J. The fixed-point is unstable if and only if there exists $\lambda \in \sigma(J)$ such that

$$Re\lambda > 0 \quad or \quad \lim_{t\to\infty} u(t) = \infty. \tag{2.2.11}$$

Remark 2.2.18. Note that Theorem 2.2.16 and Theorem 2.2.17 fail in the case of non-hyperbolic fixed-point $\tilde{y}$. At the same time, Theorem 2.2.16 motivates the terminology "linear stability" and "linear instability" that we will often use in place of "asymptotic stability" and instability for a hyperbolic fixed-point. ∎

Instead of considering systems for which all trajectories are asymptotic to a unique fixed point, a possible generalisation is to consider systems for which the asymptotic behaviour is confined to some bounded set, but where no restrictions are imposed on the possible dynamics within the set. Such systems are said to be dissipative and this constitutes the second type of qualitative property that we will deal with in this thesis.

Definition 2.2.19. *A dynamical system on $\mathbb{R}^m$ is **dissipative** if there exists a bounded, positively invariant set B with the property that for any bounded set $E \subseteq \mathbb{R}^m$, there exists $t^* = t^*(B, E) \geq 0$ such that $S(t)E \subseteq B$ for all $t > t^*$. The set B is called an absorbing set.* ∎

We now want to investigate when a dynamical system defined by (2.2.1) is dissipative. To this end, one needs a variety of structural assumptions on the vector field $f(\bullet)$, which arise naturally in applications. We will now consider two such structural assumptions.

First consider (2.2.1) under the assumption that there exist constants $\alpha \geq 0$ and $\beta > 0$ such that

$$\langle f(y), y\rangle \leq \alpha - \beta||y||^2 \quad \text{for all} \quad y \in \mathbb{R}^m. \tag{2.2.12}$$

Theorem 2.2.20. *Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz and satisfies (2.2.12). Then (2.2.1) defines a dynamical system on $\mathbb{R}^m$ and for any $\epsilon > 0$ and bounded set $E \subset \mathbb{R}^m$ there exists $t^* = t^*(E, \epsilon)$ such that for all $t > t^*$*

$$||y(t)||^2 < \frac{\alpha}{\beta} + \epsilon, \tag{2.2.13}$$

where $y(t)$ is the solution of (2.2.1). Hence the dynamical system (2.2.1) is dissipative with an absorbing set

$$B = B\left(0, \sqrt{\frac{\alpha}{\beta} + \epsilon}\right) \tag{2.2.14}$$

for any $\epsilon > 0$.

Proof. Given the importance of this theorem in our work, we prove it in detail following [41]. We first establish an a priori bound on the solution $y(t)$ with an initial data y_0, whenever it exists. Note that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}||y(t)||^2 &= \langle f(y(t)), y(t)\rangle \\ &\leq \alpha - \beta||y(t)||^2 \ \text{ by (2.2.12).} \end{aligned}$$

Applying the Gronwall inequality, Lemma 2.2.7, we obtain

$$||y(t)||^2 \leq \frac{\alpha}{\beta} + e^{-2\beta t}\left[||y_0||^2 - \frac{\alpha}{\beta}\right]. \tag{2.2.15}$$

Thus for $t \geq 0$,

$$||y(t)|| \leq \max\left(||y_0||, \sqrt{\frac{\alpha}{\beta}}\right). \tag{2.2.16}$$

The relation (2.2.16) shows that the solution of (2.2.1), if it exists, cannot blow up for $y_0 \in \mathbb{R}^m$. More precisely, since $\mathbb{R}^m = \bigcup_{\epsilon>0} B\left(0, \sqrt{\frac{\alpha}{\beta}+\epsilon}\right)$ so that $y_0 \in \mathbb{R}^m$ belongs to some $B_0 = B\left(0, \sqrt{\frac{\alpha}{\beta}+\epsilon_0}\right)$, (2.2.16) shows that the solution $y(t)$ must remain in this B_0. By Corollary 2.2.6, a unique solution exists indeed for all $t \geq 0$ and it remains in $B_0 \subset \mathbb{R}^m$, which means that (2.2.1) defines a dynamical system on $\mathbb{R}^m$.

From the above use of (2.2.16), we have in passing shown that each B is positively invariant: $S(t)B \subset B$.

We now show that B is absorbing. In other words for any bounded set $E \subseteq \mathbb{R}^m$, we want to show that there exists $t^* = t^*(E, \epsilon) \geq 0$ such that $t > t^*$ implies that $S(t)E \subseteq B$. That is, for $y_0 \in E$,

$$||y(t)|| < \sqrt{\frac{\alpha}{\beta}+\epsilon} \quad \text{for} \quad t > t^*. \tag{2.2.17}$$

For a bounded set E, $y_0 \in E$ and from (2.2.15), we have

$$||y(t)||^2 \leq \frac{\alpha}{\beta} + e^{-2\beta t}\left[R^2 - \frac{\alpha}{\beta}\right]$$

where

$$R = \sup_{y_0 \in E \cup B} ||y_0||.$$

Solving for t, the inequality

$$\frac{\alpha}{\beta} + e^{-2\beta t}\left[R^2 - \frac{\alpha}{\beta}\right] < \frac{\alpha}{\beta} + \epsilon,$$

it is clear that

$$t^* = \frac{1}{2\beta} \ln \frac{R^2 - \frac{\alpha}{\beta}}{\epsilon}$$

is the required time in (2.2.17). Thus the dynamical system is dissipative with B as absorbing set. □

We generalize (2.2.12) by considering a weaker condition that induces dissipativity, but for which the decay from infinity can be arbitrarily slow. Notice that if f satisfies (2.2.12) and $R \geq \sqrt{\frac{\alpha}{\beta}}$ then

$$\langle f(y), y\rangle < 0 \quad \text{for} \quad ||y|| > R. \tag{2.2.18}$$

Thus (2.2.12) implies (2.2.18) while the contrary is not true.

The theorem below shows that if f satisfies (2.2.18) then (2.2.1) defines a dissipative dynamical system.

Theorem 2.2.21. *If $f : \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz then (2.2.1), (2.2.18) defines a dissipative dynamical system and the open ball*

$$B(0, R+\epsilon)$$

is an absorbing set for any $\epsilon > 0$.

Proof. Given $\epsilon > 0$, let B denote the open ball $B(0, R+\epsilon)$. Assume that (2.2.1) has a unique solution $y(t)$ and that $y(t) \in \mathbb{R}^m \backslash B$ for $t \geq 0$. Then by (2.2.18), we have

$$\frac{d}{dt}||y(t)||^2 < 0, \tag{2.2.19}$$

because

$$\frac{1}{2}\frac{d}{dt}||y(t)||^2 \leq \langle f(y(t)), y(t)\rangle\,. \tag{2.2.20}$$

By integrating (2.2.19), we obtain

$$||y(t)||^2 \leq ||y_0||^2. \tag{2.2.21}$$

Combining this with the case when the solution $y(t)$ remains in B, we have the a priori bound

$$||y(t)|| \leq \max\{R+\epsilon, ||y_0||\} \quad \text{for} \quad t \geq 0. \tag{2.2.22}$$

Thus, in view of Corollary 2.2.6, Equations (2.2.1) and (2.2.18) define a dynamical system on $\mathbb{R}^m$.

To show that this dynamical system is dissipative, we proceed as follows. Clearly the ball B is positively invariant because of (2.2.22). For a bounded set $E \in \mathbb{R}^m$, let

$$r = \sup_{y \in E \cup B} ||y|| > R + \epsilon$$

and

$$E^* = \{y; ||y|| \leq r\}.$$

Note that E^* is positively invariant, i.e. $S(t)E^* \subset E^*$ because of (2.2.22) and of the relation (2.2.19) that holds for $y(t) \in E^* \backslash B$ (see Figure 2.1).

Furthermore, since $E^* \backslash B$ is compact, and f is continuous we deduce from (2.2.19) that there exists $\delta > 0$ such that

$$||y(t)||^2 \leq -\delta t + ||y_0||^2.$$

Now if $y_0 \in E$, we have two cases. Either $y_0 \in B$ in which case, $y(t)$ remains in B for $t \geq 0$ because B is positively invariant, or $y_0 \notin B$. In the latter case, using Definition 2.2.2, we have

$$\begin{aligned} ||S(t)y_0||^2 &= ||y(t)||^2 \\ &\leq -\delta t + ||y_0||^2 \\ &\leq -\delta t + r^2 \\ &< (R+\epsilon)^2 \end{aligned}$$

whenever

$$t > t^* := \frac{r^2 - (R+\epsilon)^2}{\delta}.$$

This shows that the dynamical system is dissipative and that $B(0, R+\epsilon)$ is an absorbing set. □

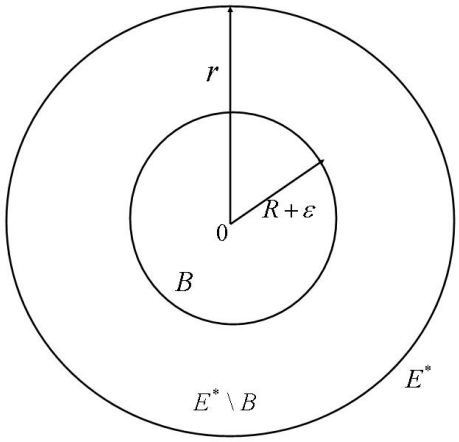

Figure 2.1: Proof of Theorem 2.2.21

2.3 Discrete Dynamical Systems

In this section we present dynamical systems generated by mappings from $\mathbb{R}^m$ to $\mathbb{R}^m$. The definitions for discrete dynamical systems are in some sense analogous to those of continuous systems on the understanding that the time variable $t \in [0, \infty)$ is now replaced by the discrete variable $n \in \mathbb{N}$. Given this analogy, we shall be concise and focus only on the main tools that we need. Once again, [41] is our main reference where most of the concepts below can be found.

2.3.1 Generalities

Let $G : \mathbb{R}^m \to \mathbb{R}^m$. Consider a sequence $\{y_n\}_{n=0}^{\infty}$ defined recursively by

$$y_{n+1} = G(y_n). \tag{2.3.1}$$

We refer to such a map or iterate as explicit mapping since y_{n+1} is given explicitly in terms of y_n. Sometimes y_{n+1} is not given by an explicit mapping of the form (2.3.1), but instead y_{n+1} is obtained from y_n through an implicit mapping of the form

$$H(y_{n+1}, y_n) = 0, \tag{2.3.2}$$

where $H : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$.

Remark 2.3.1. For (2.3.1) uniqueness of the solution sequence $\{y_n\}$ is guaranteed due to the explicit nature of the map, whereas for (2.3.2) it is necessary to establish existence and uniqueness of a solution y_{n+1} when y_n is given. ∎

Definition 2.3.2. *Equation (2.3.1) defines a* ***discrete dynamical system*** *on a subset $E \subseteq \mathbb{R}^m$ if, for every $y_0 \in E$, the sequence $\{y_n\}_{n=0}^{\infty}$ is such that y_n remains in E for all $n \geq 0$.* ∎

To deal with the problem of non-uniqueness when solving certain classes of implicit numerical methods, we consider (2.3.2) in the case when there may be multiple solutions. This motivates the following definition.

Definition 2.3.3. *Equation (2.3.2) defines a* ***generalised discrete dynamical system*** *on a subset $E \subseteq \mathbb{R}^m$ if, for every $y_0 \in E$, there exists at least one sequence $\{y_n\}$ in E that satisfies (2.3.2).* ∎

As far as the connection between discrete dynamical systems and generalised discrete dynamical system is concerned, the Implicit Function Theorem can be a powerful tool that reads as follows.

Theorem 2.3.4. *Assume that $H : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ is a function of class C^1 that satisfies the following properties:*

(i) $H(y_*, y_*) = 0$, *where* $(y_*, y_*) \in \mathbb{R}^m \times \mathbb{R}^m$ *is given;*

(ii) The determinant of the Jacobian matrix $\left(\frac{\partial H_i}{\partial z_j}(y_*, y_*)\right)_{1\leq i,j\leq m}$ *is not zero, where* $(z_1, z_2, \cdots, z_{2m})$ *denotes the variable on* $\mathbb{R}^m \times \mathbb{R}^m$. *Then there exist open neighbourhoods* $U \subset \mathbb{R}^m \times \mathbb{R}^m$ *of* (y_*, y_*) *and* $V \subset \mathbb{R}^m$. *Furthermore, there exists a* C^1 *function* $G : V \to \mathbb{R}^m$ *such that* $(y, x) \in U$ *solves* $H(y, x) = 0$ *if and only if* $y = G(x)$, $x \in V$. *Under these conditions and provided that the range of G is contained in V, $\{y_n\}$ satisfying (2.3.2) is a generalised discrete dynamical system on V if and only if $\{y_n\}$ given by (2.3.1) is a discrete dynamical system on V.*

The evolution semigroup operator of the dynamical system is an operator S^n, $n \geq 0$, that maps $\mathbb{R}^m$ into itself and enjoys the usual semigroup properties [44]. We define this more precisely.

Definition 2.3.5. *We define the* ***evolution semigroup operator*** *for the discrete dynamical system in Definition 2.3.2 to be the map $S^n : E \to E$ such that $y_n = S^n y_0$.* ∎

The evolution semigroup operator has the properties that

i. $y_{n+m} = S^n y_m = S^m y_n = S^{n+m} y_0, \qquad \forall n, m \geq 0,$

ii. $S^0 \equiv I$, the identity operator.

The discrete analogue of Gronwall inequality (Lemma 2.2.7) reads as as follows:

Lemma 2.3.6 (Gronwall Inequality). *Let a positive sequence* $\{y_n\}_{n=0}^{N}$ *satisfy*

$$y_{n+1} \leq Cy_n + D, \qquad \forall n = 0, ..., N-1$$

for some constants C *and* D *with* $C > 0$. *Then*

$$y_n \leq \frac{D}{1-C}(1-C^n) + y_o C^n, \qquad \forall n = 0, ..., N, C \neq 1$$

and

$$y_n \leq nD + y_0 \qquad \forall n = 0, ..., N, C = 1.$$

Assuming that (2.3.1) generates a discrete dynamical system on $\mathbb{R}^m$ where $G : \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz, the set $B \subset \mathbb{R}^m$ having the property $S^1 B \subset B$, it is easy to check that the discrete dynamical system is continuous with respect to initial data in the following sense: there exists a constant $c > 0$ depending on B such that

$$||S^1 y_0 - S^1 z_0|| \leq c||y_0 - z_0|| \quad \forall y_0, z_0 \in B. \tag{2.3.3}$$

In this work, we reflect (2.3.3) by using the terminology "Lipschitz continuous discrete dynamical system" though the expression "Lipschitz continuous" is often omitted. The continuity with respect to initial data stated above is to be linked to the zero-stability stated in Definition 3.2.5 below.

2.3.2 Qualitative Properties

In the results stated below, we deal once and for all with a discrete dynamical system on $\mathbb{R}^m$ defined by (2.3.1) and having evolution semigroup operator S^n.

Definition 2.3.7. *A subset* $B \subseteq \mathbb{R}^m$ *is said to be*

(a) ***positively invariant*** *if* $S^n B \subseteq B$ *for all* $n \geq 0$,

(b) ***negatively invariant*** *if* $S^n B \supseteq B$ *for all* $n \geq 0$,

(c) ***invariant*** *if* B *is both positively and negatively invariant, i.e.* $S^n B \equiv B$ *for all* $n \geq 0$. ∎

Definition 2.3.8. *A point* $\tilde{y} \in \mathbb{R}^m$ *is called a* ***fixed point*** *of the discrete dynamical system (2.3.1) if* $\tilde{y} = S^n \tilde{y}$ *for all* $n \geq 0$. ∎

Definition 2.3.9. *Let* $\tilde{y} \in \mathbb{R}^m$ *be a fixed point of the discrete dynamical system. Then* $\tilde{y}$ *is said to be*

(i) ***stable*** *if, for any* $\epsilon > 0$, *there exists* $\delta = \delta(\epsilon) > 0$ *such that if* $||y_0 - \tilde{y}|| < \delta$, *then* $||y_n - \tilde{y}|| < \epsilon$ *for* $n \geq 0$.

(ii) ***asymptotically stable*** *if (i) holds and in addition there exists* $\eta > 0$ *such that,* $||y_0 - \tilde{y}|| < \eta$ *implies* $\lim_{n\to\infty} ||y_n - \tilde{y}|| = 0$;

(iii) ***unstable*** *if (i) fails to hold.* ∎

In order to easily investigate the stability of a fixed-point $\tilde{y}$, we assume that the map G is of class C^1 and we denote by $J = JG(\tilde{y})$ the Jacobian matrix of G at $\tilde{y}$.

$$u_{n+1} = Ju_n, \quad n = 0, 1, ..., \tag{2.3.4}$$

is then a linearisation of (2.3.1) around $\tilde{y}$ where the notation $u = y - \tilde{y}$ is used as in the continuous case (see (2.2.5) - (2.2.7)).

Definition 2.3.10. *A fixed-point* $\tilde{y}$ *of the discrete dynamical system is said to be* ***hyperbolic*** *if no eigenvalues of the matrix* J *lie on the unit circle:* $|\lambda| \neq 1$, $\forall \lambda \in \sigma(J)$. *Otherwise the fixed-point is called non-hyperbolic.* ∎

Remark 2.3.11. The map G in (2.3.1) is hyperbolic if all fixed points are hyperbolic. ∎

Theorem 2.3.12 (Hartman-Grobman). *Let $G : \mathbb{R}^m \to \mathbb{R}^m$ of class C^1 have a hyperbolic fixed point $\tilde{y}$. Then there exist $\delta > 0$, a neighbourhood $\mathcal{N}$ of the origin and a homeomorphism $h : B(\tilde{y}, \delta) \to \mathcal{N}$ such that*

$$h(G(y_0)) = Jh(y_0) \quad \text{for all} \quad y_0 \in B(\tilde{y}, \delta). \tag{2.3.5}$$

Consequently, by setting

$$u_n = h(y_n) \quad \text{for all} \quad n \geq 0, \tag{2.3.6}$$

the mapping (2.3.1) in the neighbourhood $B(\tilde{y}, \delta)$ of $\tilde{y}$ is equivalent to the mapping (2.3.4) in the neighbourhood $\mathcal{N}$ of the origin.

In practice, Theorem 2.3.12 is used as follows.

Theorem 2.3.13. *Let $G : \mathbb{R}^m \to \mathbb{R}^m$ of class C^1 have a hyperbolic fixed point $\tilde{y}$. Then $\tilde{y}$ is asymptotically stable if and only if for*

$$u_n = J^n u_0, \tag{2.3.7}$$

solution of (2.3.4) with $||u_0|| := ||y_0 - \tilde{y}||$ small enough, we have

$$\lim_{n\to\infty} u_n = 0, \tag{2.3.8}$$

or equivalently,

$$|\lambda| < 1, \ \forall \lambda \in \sigma(J). \tag{2.3.9}$$

The fixed-point is unstable if and only if there exists at least one $\lambda \in \sigma(J)$ such that

$$|\lambda| > 1, \ \text{or} \lim_{n\to\infty} ||u_n|| = \infty. \tag{2.3.10}$$

To conclude this chapter, we present the definition of a discrete version of a dissipative dynamical system.

Definition 2.3.14. *A dynamical system on $\mathbb{R}^m$, is **dissipative** if there exists a bounded, positively invariant set B with the property that for any bounded set $E \subseteq \mathbb{R}^m$, there exists $n^* = n^*(B, E) \geq 0$ such that $S^n E \subseteq B$ for all $n > n^*$. The set B is called an absorbing set.* ∎

The way the stability and the dissipativity properties regarding discrete dynamical systems are crucial in our work will appear in Chapter 4 where our main contributions are presented.

Chapter 3

Finite Difference Methods

3.1 Introduction

This thesis is devoted to the study of numerical methods for dynamical systems. In this chapter, we give a short presentation of two classical methods, namely, the linear multi-step methods in Section 3.3, and the Runge-Kutta method in Section 3.4. The numerical methods we use are required to be consistent, zero-stable and thus convergent: this is discussed in Section 3.2.

The numerical methods are also required to behave asymptotically like the solutions of the decay equation: this is the essence of the concept of absolute stability addressed in Section 3.5. Finally the numerical methods are expected to define discrete dynamical systems, a topic considered in Section 3.6.

The requirements for the linear multi-step methods and the Runge-Kutta methods to be absolutely stable or to define a discrete dynamical system that is continuous with respect to initial data is subjected to a constraint on the step size Δt. However, the analytical form of this constraint can be complex for practical use. For this reason, the analysis in Section 3.7 is restricted to theta methods, which will be the focus for the rest of this thesis.

The books by Lambert [22] and Stuart and Humphries [41] are our standard references where the concepts recalled below can be found.

3.2 Basic Concepts

We consider the initial value problem for the autonomous first-order ordinary differential equations defined by (2.2.1). Numerical methods of (2.2.1) are obtained by replacing the continuous interval $[0, \infty)$ by equally-spaced grid points t_n given by

$$t_n := n\Delta t, \quad n = 0, 1, 2, ... \tag{3.2.1}$$

Δt being the stepsize. We denote by y_n an approximation to the solution $y(t_n)$ of (2.2.1) at the point t_n :

$$y_n \approx y(t_n). \tag{3.2.2}$$

The sequence $\{y_n\}_{n=0}^{\infty}$ is obtained as solution of a difference equation of the form

$$\phi(\Delta t, y_n, y_{n+1}, \cdots, y_{n+k}) = 0, \quad k \in \mathbb{N}, \tag{3.2.3}$$

coupled with appropriate initial conditions.

Thus to find the approximation y_{n+k} at the time t_{n+k} we make use of the iterates $y_{n+j}, j = 0, 1, ..., k$. If $k = 1$, the numerical method is called a *one-step method*, whereas if $k > 1$ we have a *multi-step method* or a *k-step method*, see for instance [22].

The method (3.2.3) can be explicit in which case y_{n+k} is determined recursively from the previous iterates as follows:

$$y_{n+k} = \phi(\Delta t, y_n, y_{n+1}, \cdots, y_{n+k-1}). \tag{3.2.4}$$

Otherwise the method is implicit.

For the scheme (3.2.3) to be useful, the following minimum property of fixed station convergence is required.

Definition 3.2.1. *The difference method (3.2.3) is said to be* ***convergent*** *if for each fixed* $t^* \in (0, \infty)$ *with* $t^* = t_n = n\Delta t$, *we have*

$$\lim_{\Delta t \to 0} ||y_n - y(t^*)|| = 0.$$

■

Remark 3.2.2. With the notation of Definition 3.2.1 in mind, the uniform convergence of the scheme (3.2.3) above means that

$$\sup_{t^*} ||y_n - y(t^*)|| \to 0 \text{ as } \Delta t \to 0.$$

■

Since our concern is to approximate the differential equation (2.2.1), we are mostly interested in difference equations where (3.2.3) takes the following form:

$$D_{\Delta t} y_n = F_{\Delta t}(f; y_n). \tag{3.2.5}$$

Following the notation in [7], Equation (3.2.5) is more convenient in that $D_{\Delta t} y_n$ approximates the derivative $Dy(t_n)$ of the exact solution $y(t)$ and $F_{\Delta t}(f; y_n)$ approximates $f(y(t_n))$. The notation $F_{\Delta t}(f; y_n)$ indicates that the dependence of F on y_n is through f (see, for example [22]), with

$$f_n := f(y_n). \tag{3.2.6}$$

With (3.2.5), the following further concepts of interest can be mentioned:

Definition 3.2.3. *The difference method (3.2.5) is said to be* ***consistent*** *with problem (2.2.1) if the amount by which the exact solution* $y(t)$ *fails to satisfy the discrete method is infinitely small. That is,*

$$\lim_{\Delta t \to 0} ||D_{\Delta t} y(t_n) - F_{\Delta t}(f; y(t_n))|| = 0,$$

for fixed $t^* \in [0, \infty)$ *with* $t^* = t_n = n\Delta t$. ■

The quantity

$$T_n(\Delta t) := D_{\Delta t} y(t_n) - F_{\Delta t}(f; y(t_n)), \tag{3.2.7}$$

is called the truncation error of the scheme (3.2.5).

Lemma 3.2.4. *The difference method (3.2.5) is consistent with (2.2.1) if and only if*

$$F_0(f; y) = f(y), \quad y \in \mathbb{R}, \tag{3.2.8}$$

where

$$F_0(f; y) := \lim_{\Delta t \to 0} F_{\Delta t}(f; y).$$

Definition 3.2.5. *The difference method (3.2.5) is said to be* ***zero-stable*** *if there exist $K > 0$ and $\Delta t_0 > 0$ such that for all $\Delta t \in (0, \Delta t_0]$,*

$$||z_n - \tilde{z}_n|| \leq K\epsilon \quad \textit{whenever } ||\delta_n - \tilde{\delta}_n|| \leq \epsilon$$

for a given accuracy $\epsilon > 0$ and any two perturbations δ_n and $\tilde{\delta}_n$ of the data in (2.2.1) resulting in perturbed solutions z_n and $\tilde{z}_n$. ∎

A more convenient way of proving convergence is contained in the following result:

Theorem 3.2.6. *Consistency and zero-stability are necessary and sufficient conditions for the difference method (3.2.5) to be convergent.*

3.3 Linear Multi-step Methods

To be more explicit with (3.2.5), let us consider two classical methods. We first look at the class of linear multi-step methods of order $k \geq 1$; they read as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \Delta t \sum_{j=0}^{k} \beta_j f_{n+j}, \quad n = 0, 1, 2, ..., \tag{3.3.1}$$

where $\alpha_k = 1$ and $|\alpha_0| + |\beta_0| > 0$. The parameters α_j and β_j define a particular method. If $\beta_k = 0$ then the method is explicit, otherwise it is implicit . In terms of (3.2.7), the associated truncation error of (3.3.1) is

$$T_n(\Delta t) := \sum_{j=0}^{k} \alpha_j y(t_{n+j}) - \Delta t \sum_{j=0}^{k} \beta_j f[y(t_{n+j})]. \tag{3.3.2}$$

For the multi-step method (3.3.1), consistency and zero-stability can be expressed in terms of its first and second characteristic polynomials defined by

$$\rho(z) = \sum_{j=0}^{k} \alpha_j z^j, \qquad \sigma(z) = \sum_{j=0}^{k} \beta_j z^j. \tag{3.3.3}$$

Indeed, we have the following result:

Theorem 3.3.1. *The method (3.3.1) is consistent if and only if* $\rho(1) = 0$, *and* $\sigma(1) = \rho'(1) \neq 0$. *It is also* ***zero-stable*** *if and only if all roots of* $\rho(z)$ *have modulus less than or equal to 1 and those with modulus 1 are simple.*

In the framework of this thesis numerical solutions are required to preserve the essential properties of the exact solution. The main criticism of linear multi-step methods is that they need extra initial conditions for the method to work. This could create spurious or ghost solutions, a situation which is not desirable.

For this reason we shall focus on linear one-step schemes. Specifically the two-stage θ-method that reads as follows,

$$y_{n+1} - y_n = \Delta t(\theta f_{n+1} + (1 - \theta) f_n), \tag{3.3.4}$$

where $\theta \in [0,1]$ is a given parameter. Note that for $\theta = 0$, we have the simplest method, which is referred to as the forward Euler method:

$$y_{n+1} = y_n + \Delta t f_n. \tag{3.3.5}$$

3.4 Runge-Kutta Methods

The second type of classical methods we look at are the so-called Runge-Kutta methods which have the advantage of avoiding the cost of differentiation because they do not use derivatives of f and are one-step methods.

A general k-stage Runge-Kutta method for the solution of (2.2.1) is defined by

$$y_{n+1} = y_n + \Delta t \sum_{i=1}^{k} b_i k_i \tag{3.4.1}$$

where

$$k_i = f(y_n + \Delta t \sum_{j=1}^{k} a_{ij} k_j), \quad i = 1, 2, .., k.$$

Runge-Kutta methods are often represented using the Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1k} \\ c_2 & a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ c_k & a_{k1} & a_{k2} & \dots & a_{kk} \\ \hline & b_1 & b_2 & \dots & b_k \end{array} \tag{3.4.2}$$

where we assume that the following holds:

$$c_i = \sum_{j=1}^{k} a_{ij}, \quad i = 1, 2, ..., k. \tag{3.4.3}$$

A more convenient form of (3.4.1) is

$$y_{n+1} = y_n + \Delta t \sum_{i=1}^{k} b_i f(Y_i), \tag{3.4.4}$$

where

$$Y_i = y_n + \Delta t \sum_{j=1}^{k} a_{ij} f(Y_j), \quad i = 1, 2, ..., k. \tag{3.4.5}$$

Definition 3.4.1. *The numerical method (3.4.1) - (3.4.3) is said to be explicit if*

$$a_{ij} = 0 \quad \textit{for all} \quad 1 \leq i \leq j \leq k \tag{3.4.6}$$

and implicit otherwise.

▪

It is clear that the Runge-Kutta method (3.4.1) can be written in the compact form (3.2.5) with

$$D_{\Delta t} y_n := \frac{y_{n+1} - y_n}{\Delta t}.$$

Consequently all the concepts introduced in Section 3.3 apply to Runge-Kutta methods. For convenience, we state them below.

In view of Definition 3.2.3 and Lemma 3.2.4, the Runge-Kutta method is consistent with (2.2.1) if and only if

$$\sum_{i=1}^{k} b_i = 1. \tag{3.4.7}$$

We will also use the notation

$$\mathbb{A} = \max_i \sum_{j=1}^{k} |a_{ij}| = ||A||_\infty, \tag{3.4.8}$$

and

$$||b||_1 = \mathbb{B} = \sum_{j=1}^{k} |b_i| \geq 1. \tag{3.4.9}$$

The first characteristic polynomial of the Runge-Kutta method (3.4.1) is

$$\rho(z) = z - 1 \tag{3.4.10}$$

and it always satisfies the zero-stability property or the root condition contained in Theorem 3.3.1. Consequently, Theorem 3.2.6 can be rephrased as follows:

Theorem 3.4.2. *The Runge-Kutta method is convergent if and only if it is consistent.*

A specific Runge-Kutta method that we shall deal with is the one-stage θ-method that reads as follows:

$$y_{n+1} - y_n = \Delta t f(\theta y_{n+1} + (1-\theta)y_n), \tag{3.4.11}$$

where $\theta \in [0, 1]$ is a given parameter.

Note that the two-stage θ-method (3.3.4) that is of interest to us and was presented in Section 3.3 as a linear multi-step method is also a one-stage Runge-Kutta method. In both cases the one-stage and two-stage methods with $\theta = 0$, reduce to the forward Euler method (3.3.5).

3.5 Absolute Stability

A traditional way of testing the efficiency of a numerical method is to apply it to a single model differential equation. Let us consider the model differential equation

$$y' = Jy, \quad y(0) = y_0, \tag{3.5.1}$$

where J is a constant $N \times N$ matrix with λ_s being its eigenvalues counted with their multiplicity, and satisfying the condition

$$Re\lambda_s < 0. \tag{3.5.2}$$

The reason for choosing J to be a matrix is the Hartman-Grobman Theorem (Theorem 2.2.16), which shows that the local behaviour near a fixed-point which is assumed to be $\tilde{y} = 0$ of the solution of the system (2.2.1) is given by the linearised equation (2.2.7) that has the form (3.5.1).

For simplicity, we assume that J is diagonalizable. Then there exists a transition matrix $Q = [q_1\ q_2 \cdots\ q_N]$ such that

$$Q^{-1}JQ = \Lambda := \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N). \tag{3.5.3}$$

If we make the change of dependent variable

$$y = Qz, \tag{3.5.4}$$

(3.5.1) is equivalent to

$$z' = \Lambda z, \tag{3.5.5}$$

which is an uncoupled system of N equations

$${}^s z_j' = \lambda_s\ {}^s z_j, \quad 1 \leq j, s \leq N, \tag{3.5.6}$$

having the solution

$$z(t) = {}^s z_j(0) e^{\lambda_s t}. \tag{3.5.7}$$

The behaviour of the solution

$$y(t) = e^{tJ} y_0, \tag{3.5.8}$$

of (3.5.1) as $t \to \infty$, is equivalent to that of the functions (3.5.7). In view of this behaviour of the solution we require the numerical solution to behave in a similar manner. Schemes producing such numerical solutions are roughly speaking called absolutely stable. Below we make this concept more precise for linear multi-step and Runge-Kutta methods.

3.5.1 Linear Multi-step Methods

We first discuss absolute stability in the context of linear multi-step methods. Following [22], we apply the linear multi-step method (3.3.1) to the system (3.5.1) to obtain

$$\sum_{j=0}^{k}(\alpha_j I - \Delta t\beta_j J)y_{n+j} = 0. \tag{3.5.9}$$

Using (3.5.3) - (3.5.4), (3.5.9) is equivalent to

$$\sum_{j=0}^{k}(\alpha_j I - \Delta t\beta_j \Lambda)z_{n+j} = 0. \tag{3.5.10}$$

Since both I and Λ are diagonal matrices, we may write

$$\sum_{j=0}^{k}(\alpha_j - \Delta t\beta_j \lambda_s)\ {}^s z_{n+j} = 0, \quad 1 \leq s \leq N. \tag{3.5.11}$$

The general solution for each of the difference equations in (3.5.11) takes the form

$${}^s z_n = \sum_{s=1}^{p}\left[d_{s,1} + \sum_{j=2}^{\mu_s} d_{s,j}\ n(n-1)\cdots(n-j+2)\right] r_s^n, \tag{3.5.12}$$

where $d_{s,j}$ are arbitrary complex constants and r_s are roots of the difference equation

$$\sum_{j=0}^{k}(\alpha_j - \Delta t\beta_j \lambda_s)r^j = 0 \tag{3.5.13}$$

with multiplicity μ_s, $\quad 1 \leq s \leq p$ and $\sum_{s=1}^{p}\mu_s = N$.

We define the *stability polynomial* $\pi(r, \lambda_s\Delta t)$ of the method (3.3.1) to be

$$\pi(r, \lambda_s\Delta t) = \sum_{j=0}^{k}[\alpha_j - \lambda_s\Delta t\beta_j]r^j. \tag{3.5.14}$$

This polynomial can conveniently be written in terms of the first and second characteristic polynomials ρ and σ as

$$\pi(r, \hat{h}) = \rho(r) - \hat{h}\sigma(r), \tag{3.5.15}$$

where $\hat{h} := \lambda_s \Delta t$.

The stability polynomial $\pi(r, \hat{h})$ with $\hat{h} := \lambda_s \Delta t$ permits us to better compare the solution (3.5.7) or (3.5.8) with the discrete solution (3.5.12) in the following manner.

Definition 3.5.1. *The linear multi-step method (3.3.1) is called* ***absolutely stable*** *for a given $\hat{h}$, if for that $\hat{h}$ all the roots of the stability polynomials lie within the unit circle. Otherwise the method is absolutely unstable.* ∎

3.5.2 Runge-Kutta Methods

We now discuss absolute stability of a $k-$stage Runge-Kutta method, following [41]. Applying the Runge-Kutta method (3.4.4) - (3.4.5) to (3.5.6), for each $1 \leq j \leq N$, we have

$$\begin{aligned} {}^{l}z_{j,n+1} &= {}^{l}z_{j,n} + \lambda_l \Delta t b Z_j & (3.5.16)\\ Z_j &= {}^{l}z_{j,n} e + \lambda_l \Delta t A Z_j, & (3.5.17) \end{aligned}$$

where $Z_j = [Z_{j,1}, Z_{j,2}, ..., Z_{j,k}]^T$ and $e \in \mathbb{R}^k$ with $e = [1, 1, ..., 1]^T$.

Solving the above system in Z_j gives

$${}^{l}z_{j,n+1} = {}^{l}z_{j,n}[1 + \lambda_l \Delta t b(I - \lambda_l \Delta t A)^{-1} e], \tag{3.5.18}$$

where I is the $k \times k$ unit matrix.

Note that the matrix $(I - \lambda_l \Delta t A)$ is nonsingular for Δt small enough. Unlike the linear multi-step method where we had a stability polynomial, here we have a stability function, namely

$$R(\lambda_l \Delta t) = 1 + \lambda_l \Delta t b(I - \lambda_l \Delta t A)^{-1} e. \tag{3.5.19}$$

From (3.5.18) we obtain a one-step difference equation of the form

$$ {}^{l}z_{j,n+1} = R(\lambda_l \Delta t)\ {}^{l}z_{j,n}. \tag{3.5.20} $$

Coming back to the model equation (3.5.1), we obtain from (3.5.20) and the change of variable in (3.5.3) and (3.5.4)

$$ y_{n+1} = R(\Delta t J) y_n, \tag{3.5.21} $$

with the matrix function

$$ R(\Delta t J) := Q \mathrm{diag}(R(\lambda_l \Delta t)) Q^{-1} \tag{3.5.22} $$

being the stability function. Clearly, $y_n \to 0$ as $n \to \infty$ if and only if

$$ ||R(\Delta t) J|| < 1. \tag{3.5.23} $$

The analysis above on Runge-Kutta methods motivates us to state the definition of absolute stability with $R(\lambda_l \Delta t)$ instead of $\pi(r, \lambda_s \Delta t)$ as is the case with linear multi-step method.

Definition 3.5.2. *The Runge-Kutta method (3.4.4) - (3.4.5) is said to be* ***absolutely stable*** *for a given* $\lambda \Delta t,\ Re\lambda < 0$, *if* $|R(\lambda \Delta t)| < 1$. ∎

Remark 3.5.3. Our expectation is of course to have both the linear multi-step and the Runge-Kutta methods absolutely stable for all $\lambda_l \Delta t$ where $\{\lambda_l\}_{l=1}^{N}$ are the eigenvalues of the diagonalizable matrix J in (3.5.1). ∎

We now give an alternative formula for the stability function which is more suitable for its calculation as is presented in [12].

Theorem 3.5.4. *The stability function of the Runge-Kutta method (3.4.4) - (3.4.5) is given by*

$$ R(\lambda_l \Delta t) = \frac{det(I - \lambda_l \Delta t A + \lambda_l \Delta t b^T e)}{det(I - \lambda_l \Delta t A)}. \tag{3.5.24} $$

Proof. The relation (3.5.16) - (3.5.17) can be written as the algebraic linear system:

$$\begin{bmatrix} 1-\lambda_l\Delta t a_{11} & -\lambda_l\Delta t a_{12} & \dots & -\lambda_l\Delta t a_{1k} & 0 \\ -\lambda_l\Delta t a_{21} & 1-\lambda_l\Delta t a_{22} & \dots & -\lambda_l\Delta t a_{2k} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ -\lambda_l\Delta t a_{k1} & -\lambda_l\Delta t a_{k2} & \dots & 1-\lambda_l\Delta t a_{kk} & 0 \\ -\lambda_l\Delta t b_1 & -\lambda_l\Delta t b_2 & \dots & -\lambda_l\Delta t b_k & 1 \end{bmatrix} \begin{bmatrix} Z_{j,1} \\ Z_{j,2} \\ \vdots \\ Z_{j,k} \\ {}^l z_{j,n+1} \end{bmatrix} = \begin{bmatrix} {}^l z_{j,n} \\ {}^l z_{j,n} \\ \vdots \\ {}^l z_{j,n} \\ {}^l z_{j,n} \end{bmatrix} \tag{3.5.25}$$

The denominator in (3.5.24) is given by the determinant of the matrix in (3.5.25). The numerator in (3.5.24) is the determinant of the matrix

$$\begin{bmatrix} 1-\lambda_l\Delta t a_{11} & -\lambda_l\Delta t a_{12} & \dots & -\lambda_l\Delta t a_{1k} & {}^l z_{j,n} \\ -\lambda_l\Delta t a_{21} & 1-\lambda_l\Delta t a_{22} & \dots & -\lambda_l\Delta t a_{2k} & {}^l z_{j,n} \\ \vdots & \vdots & & \vdots & \vdots \\ -\lambda_l\Delta t a_{k1} & -\lambda_l\Delta t a_{k2} & \dots & 1-\lambda_l\Delta t a_{kk} & {}^l z_{j,n} \\ -\lambda_l\Delta t b_1 & -\lambda_l\Delta t b_2 & \dots & -\lambda_l\Delta t b_k & {}^l z_{j,n} \end{bmatrix}.$$

Indeed subtraction of the last row from the first k rows leaves this determinant invariant. Cramer's rule expresses ${}^l z_{j,n+1}$ as the quotient of two determinants, so we arrive at

$${}^l z_{j,n+1} = \frac{\det(I - \lambda_l\Delta t A + \lambda_l\Delta t b^T e)}{\det(I - \lambda_l\Delta t A)}\ {}^l z_{j,n}, \tag{3.5.26}$$

which establishes (3.5.24). □

Remark 3.5.5. When A is a strictly lower triangular matrix, the matrix $I - \lambda_l\Delta t A$ is then lower triangular with all the elements of its main

diagonal being unity. It follows that $\det(I - \lambda_l \Delta t A) = 1$ and for all explicit Runge-Kutta methods the stability function is a polynomial in $\lambda_l \Delta t$. For implicit methods $\det(I - \lambda_l \Delta t A) \neq 1$ so that the stability function is a rational function of $\lambda_l \Delta t$. ∎

3.6 Numerical Methods as Dynamical Systems

We now turn our attention to discussing conditions under which numerical methods studied earlier in this chapter generate discrete dynamical systems. If numerical methods are to give useful information on Lipschitz continuous dynamical systems, it is of paramount importance that these methods are viewed as discrete dynamical systems which are continuous with respect to initial data. In this way, we are comparing dynamical systems of the same nature.

A Runge-Kutta method, applied to (2.2.1), not only defines an approximation to the solution of (2.2.1), but can also define a discrete dynamical system. We start with an explicit Runge-Kutta method (3.4.4)-(3.4.5) which, in view of Definition 3.4.1, can be written recursively as follows

$$y_{n+1} = S_{\Delta t}^{n+1} y_0,$$

where

$$S_{\Delta t}^1 y_0 = y_0 + \Delta t \sum_{j=1}^{k} b_j f(g_j(y_0)),$$

$$g_1(y_0) = y_0, \quad g_i(y_0) = y_0 + \Delta t \sum_{j=1}^{i-1} a_{ij} f(g_j(y_0)), \quad i = 2, 3, \cdots, k.$$

Assuming that $\{z_n\}$ is another sequence generated by the Runge-Kutta method from z_0, it is easy to prove that

$$||S_{\Delta t}^1 y_0 - S_{\Delta t}^1 z_0|| \leq c||y_0 - z_0||$$

for some $c > 0$ whenever f is locally Lipschitz. Consequently, we have the following result [41].

Theorem 3.6.1. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz. If the Runge-Kutta method (3.4.4), (3.4.5) is explicit, then it defines a Lipschitz continuous discrete dynamical system on $\mathbb{R}^m$.*

However, for an implicit method the Runge-Kutta method need not be uniquely solvable, and hence an implicit Runge-Kutta method need not define a discrete dynamical system. To overcome this difficulty, we impose a condition on the step size Δt as illustrated in the next result [41].

Theorem 3.6.2. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be globally Lipschitz with Lipschitz constant L. Assume that*

$$\Delta t < \frac{1}{L\mathbb{A}}, \tag{3.6.1}$$

where $\mathbb{A}$ is defined in (3.4.8). Then the Runge-Kutta method (3.4.4)-(3.4.5) is uniquely solvable for Y_i, $i = 1, .., k$. More precisely, the solution can be found as a fixed point of the iteration:

$$Y_i^{s+1} = y_n + \Delta t \sum_{j=1}^{i-1} a_{ij} f(Y_j^{s+1}) + \Delta t \sum_{j=i}^{k} a_{ij} f(Y_j^s), \tag{3.6.2}$$

for $i = 1, 2, ..., k, s = 0, 1, 2, ...$ Consequently, the Runge-Kutta method is a Lipschitz continuous discrete dynamical system on $\mathbb{R}^m$.

Proof. Consider other iterates Z_i^{s+1} of type (3.6.2) initiated at Z_i^0. It can be shown that

$$||Y^{s+1} - Z^{s+1}|| \leq \Delta t L \mathbb{A} ||Y^s - Z^s||$$

where Y^s and Z^s denote vectors in $\mathbb{R}^{mk}$ comprised of the $Y_i^s, Z_i^s \in \mathbb{R}^m$. The Contraction Mapping Theorem ([48], [49]) and (3.6.1) lead to the first part of Theorem 3.6.2.

Assuming that $\{z_n\}$ is another sequence generated by the Runge-Kutta method from z_0, (3.6.1) implies the existence of $c > 0$ such that

$$||y_{n+1} - z_{n+1}|| \leq c||y_n - z_n||,$$

which shows the Lipschitz continuity of $\{y_n\}$ with respect to the initial data. □

Regarding the linear multistep method (3.3.1), we re-write it in the form

$$y_{n+k} = \sum_{j=0}^{k-1} [\Delta t \beta_j f(y_{n+j}) - \alpha_j y_{n+j}] + \Delta t \beta_k f(y_{n+k}). \tag{3.6.3}$$

In the setting of dynamical systems, we can say that the action of the evolution semigroup operator $S^1_{\Delta t}$ on the data vector

$$Y_n := [y_n, y_{n+1}, \cdots, y_{n+k-1}]^T \in \mathbb{R}^{mk}$$

is the vector

$$Y_{n+1} := [y_{n+1}, y_{n+2}, \cdots, y_{n+k}]^T \in \mathbb{R}^{mk}.$$

That is,

$$Y_{n+1} := S^1_{\Delta t} Y_n.$$

Another sequence $\{z_n\}$ of the form (3.6.3), where z_{n+k} is generated from $(z_n, z_{n+1}, \cdots, z_{n+k-1})$ is equally considered. Proceeding as for the Runge-Kutta method, we obtain the following result [41].

Theorem 3.6.3. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz. Assume that the multi-step method (3.3.1) or (3.6.3) is explicit, i.e. $\beta_k = 0$. Then (3.3.1) or (3.6.3) defines a Lipschitz continuous discrete dynamical system $\mathbb{R}^{mk}$.*
On the other hand, we assume that f is globally Lipschitz with Lipschitz constant L and that Δt satisfies the condition

$$\Delta t < \frac{1}{|\beta_k| L}. \tag{3.6.4}$$

Then the linear multi-step method (3.6.3) defines a Lipschitz continuous discrete dynamical system on $\mathbb{R}^{mk}$ for which the solution y_{n+k} generated from the data $(y_n, y_{n+1}, \cdots, y_{n+k-1})$ can be found as a fixed point of the iteration

$$y_{n+k}^{s+1} = \sum_{j=0}^{k-1} [\Delta t \beta_j f(y_{n+j}) - \alpha_j y_{n+j}] + \Delta t \beta_k f(y_{n+k}^s), \quad s = 0, 1, 2, ... \tag{3.6.5}$$

Remark 3.6.4. For $f : \mathbb{R}^m \to \mathbb{R}^m$ locally Lipschitz, the implicit Runge-Kutta method and linear multi-step method, define Lipschitz continuous discrete dynamical systems under more restrictive conditions on Δt, (see for example, [41]). ∎

3.7 Theta Methods

The structure of the Runge-Kutta and linear multi-step methods in the previous sections has shown that a restriction must be placed on the step size Δt if the methods are to provide acceptable approximations to the solution of the Lipschitz continuous discrete dynamical systems. The expression of the restriction can be complex. For this reason, we will as from now focus the rest of the thesis to the theta methods. Within this choice, it is our aim to better understand the said restriction in order to design in the next chapters non-standard schemes which are reliable for any value of Δt.

For convenience we re-define the theta methods we mentioned in Sections 3.3 and 3.4. Consider the parameter $\theta \in [0, 1]$. The one-stage theta method for approximating (2.2.1) is defined by

$$\frac{y_{n+1} - y_n}{\Delta t} = f[\theta y_{n+1} + (1 - \theta) y_n]; \tag{3.7.1}$$

the two-stage theta method reads as follows:

$$\frac{y_{n+1} - y_n}{\Delta t} = \theta f(y_{n+1}) + (1-\theta) f(y_n). \tag{3.7.2}$$

There are two specific values of θ for which both theta methods reduce to the same scheme. More precisely, for $\theta = 0$, we have the forward explicit Euler method

$$\frac{y_{n+1} - y_n}{\Delta t} = f(y_n), \tag{3.7.3}$$

while $\theta = 1$ yields the forward implicit Euler method

$$\frac{y_{n+1} - y_n}{\Delta t} = f(y_{n+1}). \tag{3.7.4}$$

Note that the value $\theta = \frac{1}{2}$ in (3.7.1) and (3.7.2) corresponds to the so-called mid-point rule and trapezoidal rule, respectively.

When θ is different from 0 and 1, the one-stage and two-stage theta methods are still intimately related in the sense of the following theorem.

Theorem 3.7.1. *Let $\{v_n\}_{n=0}^{\infty}$ satisfy the one-stage theta method (3.7.1), then the sequence $\{y_n\}_{n=0}^{\infty}$ given by*

$$y_n = (1-\theta)v_n + \theta v_{n+1} \tag{3.7.5}$$

satisfies the two-stage theta method (3.7.2). Conversely, if $\{y_n\}_{n=0}^{\infty}$ satisfies the two-stage theta method (3.7.2). Then the sequence $\{v_n\}_{n=0}^{\infty}$ given by

$$v_n = y_n - \Delta t \theta f(y_n) \tag{3.7.6}$$

satisfies the one-stage theta method.

Proof. We follow the proof by Stuart and Humphries, [41] p 227, but in more detail. Re-writing (3.7.1) and (3.7.2), we have

$$v_{n+1} = v_n + \Delta t f[\theta v_{n+1} + (1-\theta)v_n] \tag{3.7.7}$$

$$y_{n+1} = y_n + \Delta t\theta f(y_{n+1}) + \Delta t(1-\theta)f(y_n) \tag{3.7.8}$$

respectively. Let us fix $\theta \in [0,1]$. Suppose that $\{y_n\}_{n=0}^{\infty}$ satisfies the two-stage theta method (3.7.8) and let $\{v_n\}_{n=0}^{\infty}$ be given by (3.7.6) then

$$v_{n+1} = y_{n+1} - \Delta t\theta f(y_{n+1}). \tag{3.7.9}$$

Rearranging terms in (3.7.8) and using (3.7.9) yields

$$v_{n+1} = y_n + \Delta t(1-\theta)f(y_n) = y_n + \Delta t f(y_n) - \Delta t\theta f(y_n). \tag{3.7.10}$$

Substituting the right hand side of (3.7.9) into (3.7.10), we obtain

$$v_{n+1} = v_n + \Delta t f(y_n). \tag{3.7.11}$$

Multiplying throughout by $(1-\theta)$ and using (3.7.10) produces

$$\Delta t(1-\theta)f(y_n) = v_{n+1} - y_n. \tag{3.7.12}$$

Making minor manipulation in (3.7.12) the result (3.7.5) follows.

Replacing y_n in (3.7.12) by v_n shows that $\{v_n\}_{n=0}^{\infty}$ satisfies the one-stage $\theta-$method (3.7.7) for $n \geq 0$.

Conversely, if $\{v_n\}_{n=0}^{\infty}$ satisfies the one-stage $\theta-$method (3.7.7), it can easily be shown that $\{y_n\}_{n=0}^{\infty}$ in (3.7.5) satisfies the two-stage $\theta-$method (3.7.8). Indeed from (3.7.9) we have

$$y_n = \theta v_{n+1} + (1-\theta)v_n = v_n + \theta(v_{n+1} - v_n). \tag{3.7.13}$$

From (3.7.7) we have

$$v_{n+1} - v_n = \Delta t f[\theta v_{n+1} + (1-\theta)v_n]. \tag{3.7.14}$$

Using (3.7.13) and (3.7.14), we get

$$v_n = y_n - \Delta t\theta f(y_n) \tag{3.7.15}$$

which completes the proof in view of (3.7.5). □

We saw in Section 3.3 that the two-stage theta method is a linear one-step method. We also saw in Section 3.4 that both the one-stage and the two-stage theta methods are Runge-Kutta methods with the Butcher tableau

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

and

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

respectively. Consequently, we have the following result.

Theorem 3.7.2. *The one-stage and two-stage theta methods for approximating the initial value problem (2.2.1) are convergent. In the particular case of the forward Euler method, assuming that $f : \mathbb{R}^m \to \mathbb{R}^m$ is globally Lipschitz, with Lipschitz constant L, and that the exact solution $y(t)$ of class C^2 with bounded second derivative, we have the error estimate*

$$||y(t_n) - y_n|| \leq K\Delta t(e^{Lt_n} - 1) \tag{3.7.16}$$

for some constant $K > 0$.

Proof. Although this result is known, we provide the proof here because we will not come back to convergence issues when dealing later on with non-standard schemes. The first and the second characteristic polynomials of the two-stage theta method (3.7.2), viewed as a linear multi-step method, are $\rho(z) = z-1$ and $\sigma(z) = (1-\theta)+\theta z$, respectively. In view of Theorem 3.3.1, the two-stage theta method is consistent and zero-stable and thus convergent. Furthermore, since the two-stage and one-stage theta methods are equivalent (Theorem 3.7.1) the one-stage theta method is equally convergent.

Regarding the forward Euler method

$$y_{n+1} - y_n = \Delta t f(y_n), \tag{3.7.17}$$

we proceed as follows. By (2.2.1) and Taylor expansion of $y(t_{n+1})$ about $t = t_n$, we obtain

$$y(t_{n+1}) = y(t_n) + \Delta t f(y(t_n)) + \frac{(\Delta t)^2}{2} y''(\zeta_n), \quad t_n < \zeta_n < t_{n+1}. \tag{3.7.18}$$

Letting $e_n = y(t_n) - y_n$ and subtracting (3.7.17) from (3.7.18), produce:

$$e_{n+1} = e_n + \Delta t[f(y(t_n)) - f(y_n)] + \frac{(\Delta t)^2}{2} y''(\zeta_n). \tag{3.7.19}$$

Using the Lipschitz and boundedness assumptions, we have

$$||e_{n+1}|| \leq ||e_n||(1 + L\Delta t) + c(\Delta t)^2. \tag{3.7.20}$$

By the discrete Gronwall inequality (Lemma 2.3.6), we have

$$||e_n|| \leq \frac{c(\Delta t)^2}{|1 - (1 + L\Delta t)|}[(1 + L\Delta t)^n - 1] + ||e_0||(1 + L\Delta t)^n \tag{3.7.21}$$

from which the estimate (3.7.16) follows. □

Remark 3.7.3. One could be a bit more precise about the consistency of the theta methods used in the proof of Theorem 3.7.2 in the following way. Assuming that the exact solution $y(t)$ is smooth enough and has bounded derivatives, the local truncation error

$$\tau_n = \begin{cases} \frac{y(t_{n+1}) - y(t_n)}{\Delta t} - \theta f(y(t_{n+1})) - (1 - \theta) f(y(t_n)) \\ \frac{y(t_{n+1}) - y(t_n)}{\Delta t} - f[\theta y(t_{n+1}) + (1 - \theta) y(t_n)] \end{cases}$$

of the theta methods (3.7.1) and (3.7.2) have the asymptotic behaviour

$$\tau_n = \begin{cases} O(\Delta t) & \text{if } \theta \neq \frac{1}{2} \\ O((\Delta t)^2) & \text{if } \theta = \frac{1}{2}. \end{cases} \tag{3.7.22}$$

■

These are obtained by Taylor expansion of $y(t_{n+1})$ about $t = t_n$. For example, for the two-stage method, we have the following.

$$\begin{aligned}
\tau_n &= \frac{[y(t_{n+1}) - y(t_n)}{\Delta t} - \theta f(y(t_{n+1})) - (1-\theta) f(y(t_n)) \\
&= \frac{[y(t_n) + \Delta t f(y(t_n)) + (\Delta t)^2/2 f'(y(t_n)) + (\Delta t)^3/6 f''(y(t_n)) + O(\Delta t^4)] - y(t_n)}{\Delta t} \\
&\quad -\theta f(y(t_n)) - \theta \Delta t f'(y(t_n)) - \theta (\Delta t)^2/2 f''(y(t_n)) - f(y(t_n)) + \theta f(y(t_n)) + O(\Delta t^3) \\
&= (\Delta t)/2 f'(y(t_n)) + (\Delta t)^2/6 f''(y(t_n)) - \theta \Delta t f'(y(t_n)) - \theta (\Delta t)^2/2 f''(y(t_n)) + O(\Delta t^3) \\
&= (1/2 - \theta)\Delta t f'(y(t_n)) + (1/6 - \theta/2)(\Delta t)^2/2 f''(y(t_n)) + O(\Delta t^3)
\end{aligned}$$

and from this (3.7.22) follows.

We conclude this chapter by discussing when the theta methods replicate some of the qualitative properties targeted in the previous sections for the underlying differential equation (2.2.1). Firstly, are theta methods discrete dynamical systems? A positive conditional answer is given in the next theorem [41].

Theorem 3.7.4. *Let $f : \mathbb{R}^m \to \mathbb{R}^m$ be globally Lipschitz with Lipschitz constant L. Assume that*

$$\Delta t < \frac{1}{\theta L}. \tag{3.7.23}$$

Then the one-stage and two-stage theta methods define Lipschitz continuous discrete dynamical systems on $\mathbb{R}^m$.

Proof. The theorem follows from Theorem 3.6.1 and Theorem 3.6.2. However, this can be proved directly given the simple structure of these schemes, as we show now. Let $\{y_{n+1}^s\}_{s\geq 0}$ and $\{z_{n+1}^s\}_{s\geq 0}$ be two iterates defined through the one-stage theta method by

$$y_{n+1}^{s+1} = y_n + \Delta t f[\theta y_{n+1}^s + (1-\theta) y_n] \tag{3.7.24}$$

$$z_{n+1}^{s+1} = y_n + \Delta t f[\theta z_{n+1}^s + (1-\theta) y_n]. \tag{3.7.25}$$

It is easy to check by using the Lipschitz property of f that

$$||y_{n+1}^{s+1} - z_{n+1}^{s+1}|| \leq \theta \Delta t L ||y_{n+1}^s - z_{n+1}^s||. \tag{3.7.26}$$

Under the condition (3.7.23), the Contraction Mapping Theorem shows that the one-stage theta method (3.7.1) for $\theta \neq 0$ is uniquely solvable in $\mathbb{R}^m$, with its solution $y_{n+1} \in \mathbb{R}^m$ being found as the fixed-point of the iteration (3.7.24).

Furthermore, if

$$y_{n+1} = y_n + \Delta t f[\theta y_{n+1} + (1-\theta)y_n] \tag{3.7.27}$$

$$z_{n+1} = z_n + \Delta t f[\theta z_{n+1} + (1-\theta)z_n], \tag{3.7.28}$$

we have

$$||y_{n+1} - z_{n+1}|| \leq \frac{1 + (1-\theta)L\Delta t}{1 - \theta L \Delta t} ||y_n - z_n||, \tag{3.7.29}$$

which shows the Lipschitz continuity with respect to initial data. Thus the one- stage theta method is a discrete dynamical systems on $\mathbb{R}^m$.

Given the equivalence stated in Theorem 3.7.1 between the one-stage and the two-stage theta methods, we conclude that the two-stage theta method is equally a Lipschitz continuous discrete dynamical systems on $\mathbb{R}^m$. □

Remark 3.7.5. When $\theta = 0$, there is no restriction on Δt in (3.7.23). Thus the forward Euler method is a discrete dynamical systems on $\mathbb{R}^m$. Actually, the forward Euler method is a discrete dynamical systems on $\mathbb{R}^m$ even when f is locally Lipschitz, (see Theorem 3.6.1). More generally, the theta methods can be discrete dynamical systems on $\mathbb{R}^m$ under flexible structural assumptions on f (e.g. locally Lipschitz, one-sided Lipschitz condition, etc). But we will not consider these aspects, which can be found in [41]. ▪

The next qualitative property is related to the absolute stability of the theta methods. When applied to the model linear equation (3.5.1),

both the one-stage and two-stage theta methods reduce to

$$\frac{y_{n+1}-y_n}{\Delta t} = J[\theta y_{n+1} + (1-\theta)y_n]. \tag{3.7.30}$$

Using the factorization (3.5.3) and the change of variable (3.5.4), (3.7.30) is equivalent to

$$y_{n+1} = Q\text{diag}\left[\frac{1+\Delta t(1-\theta)\lambda_1}{1-\Delta t\theta\lambda_1}, \cdots, \frac{1+\Delta t(1-\theta)\lambda_N}{1-\Delta t\theta\lambda_N}\right] Q^{-1} y_n. \tag{3.7.31}$$

The stability function (Definition 3.5.22 or 3.5.24) of the one-stage and the two-stage theta methods, viewed as Runge-Kutta methods, is then

$$R(\lambda\Delta t) = \frac{1+\Delta t(1-\theta)\lambda}{1-\Delta t\theta\lambda}, \tag{3.7.32}$$

while the stability polynomial (Definition 3.5.14) of the two-stage theta method, as a linear multi-step method, is

$$\pi(r, \lambda\Delta t) = 1 + \lambda\Delta t(1-\theta)(1-\lambda\Delta t\theta), \tag{3.7.33}$$

whose unique root is $r = R(\lambda\Delta t)$.

For a complex number λ, with $Re\lambda < 0$, we have

$$\begin{aligned} |R(\lambda\Delta t)|^2 &= \left|\frac{1+\Delta t(1-\theta)\lambda}{1-\Delta t\theta\lambda}\right|^2 \\ &= \frac{(1-\Delta t(1-\theta)|Re\lambda|)^2 + (\Delta t)^2(1-\theta)^2|Im\lambda|^2}{(1+\Delta t\theta||Re\lambda|)^2 + (\Delta t)^2\theta^2|Im\lambda|^2} \\ &= \frac{1+2\Delta t\theta|Re\lambda| - 2\Delta t|Re\lambda| + (\Delta t)^2(1-\theta)^2|\lambda|^2}{1+2\Delta t\theta||Re\lambda| + (\Delta t)^2\theta^2|\lambda|^2}. \end{aligned} \tag{3.7.34}$$

Thus we have the following result.

Theorem 3.7.6. *For $\theta \in [\frac{1}{2}, 1]$, the one-stage and the two-stage theta methods are (unconditionally) absolutely stable for any $\lambda\Delta t$ with $Re\lambda < 0$: according to the standard terminology the theta methods are $A-$stable in this case.*
For $\theta \in [0, \frac{1}{2})$, the theta methods are (conditionally) absolutely stable for $\lambda\Delta t$ with $Re\lambda < 0$, whenever

$$\Delta t < \frac{2|Re\lambda|}{(1-2\theta)|\lambda|^2}. \tag{3.7.35}$$

The last qualitative property of our interest is the dissipativity of the theta methods when the underlying dynamical system (2.2.1) is dissipative. Our point of departure is the following classical result proved in [41].

Theorem 3.7.7. *Consider (2.2.1) as a dissipative dynamical system in the setting of Theorem 2.2.21, where $R > 0$ is given and let $\theta \in [\frac{1}{2}, 1]$. Then for any $\Delta t > 0$, the one-stage and the two-stage theta methods define (generalised) dynamical systems which are dissipative in the sense of Definition 2.3.3: the closed ball $\bar{B}\,(0, R + \delta + \Delta t(1-\theta)M)$ is an absorbing set for any $\delta > 0$ and $M := \sup_{v \in \bar{B}(0,R+\delta)} ||f(v)||$.*

In the particular case when the setting is that of Theorem 2.2.20 where α and β are given and $\theta \in (\frac{1}{2}, 1]$, the above conclusion holds with any open ball $B(0, \frac{1}{2\theta-1}\sqrt{\frac{\alpha}{\beta}} + \delta)$, $\delta > 0$, being an absorbing set that does however not depend on the step size Δt.

Remark 3.7.8. It follows from Theorem 3.7.7 that within the range $\theta \in [0, \frac{1}{2})$, the dissipative property of the theta methods is not guaranteed. We will try to remedy this in the next chapter. ∎

Chapter 4

Non-standard Finite Difference Methods

4.1 Introduction

The first set of the main contributions of this thesis appear in this chapter. The chapter is based on the author's publications [3], [5], [6], as well as on the technical report [4] that is under review.

We present in Section 4.2 generalities on the non-standard finite difference method. In Section 4.3, we analyze non-standard finite difference schemes that have no spurious fixed-points compared to the dynamical system under consideration, the linear stability/instability property of the fixed-points being the same for both the discrete and continuous systems. The schemes we study are non-standard variants of the theta methods presented in the previous chapter and they are constructed by using Mickens' rule about the denominator of the discrete derivatives. We obtain a sharper condition for the elementary stability of the schemes. For more complex dynamical systems which are dissipative, we design schemes that replicate this property in Section 4.4. In a second step in Section 4.5, we consider a specific class of dynamical systems which is equivalent to the simplest model of Hamiltonian systems that occur in classical mechanics. We design a non-standard finite difference scheme that replicates the underlying principle of conversation of energy. Here we use Mickens' rules about nonlocal approximation of nonlinear terms.

4.2 Generalities

The shortcomings of the classical numerical schemes, specifically theta methods, for being reliable discrete dynamical systems were pointed out in Chapter 3. It became clear that the time step size Δt should be small enough if the schemes were to replicate qualitative properties of the exact solutions. The non-standard finite difference method introduced by Mickens [26], aims at preserving the qualitative properties at no cost with regard to the value of Δt. The following definition is due to [7]:

Definition 4.2.1. *Assume that the solution of (2.2.1) satisfies some property P. The difference scheme (3.2.5) is called* ***qualitatively stable*** *with respect to the property $\mathcal{P}$ (or $\mathcal{P}$ -stable) if for all step sizes $\Delta t > 0$, the discrete solutions for (3.2.5) satisfy the properties $\mathcal{P}$.* ∎

The term dynamic consistency with respect to $\mathcal{P}$ has been introduced recently and is sometimes used instead of that in Definition 4.2.1, (see [29] and [30]).

Significant properties of solutions of differential equations are of great importance from the practical point of view. Such properties include among others: types of fixed points, oscillatory solution, monotonicity of solutions, and conservation of energy.
The ideal situation when the discrete scheme is stable with respect to any property of the exact solution is given in the next definition.

Definition 4.2.2. *The numerical method (3.2.5) for approximating (2.2.1) is called an* ***exact scheme*** *whenever the difference equation (3.2.5) and the differential equation (2.2.1) have the same general solutions at the discrete time $t = t_n$. In particular, with $y(t)$ being the solution of the initial value problem (2.2.1), we have $y_n = y(t_n)$.* ∎

At this stage, it is essential to consider exact schemes of two model scalar equations that come often in this thesis. These are the exponential growth equation

$$y' = \lambda y, \quad y(0) = y_0, \quad \lambda \neq 0 \tag{4.2.1}$$

and the logistic equation

$$y' = \lambda y(1-y),\ y(0) = y_0, \quad \lambda > 0. \tag{4.2.2}$$

Notice that (4.2.1) was the test equation for absolute stability in Chapter 3, while (4.2.2) will appear in the Fisher equation in the next chapter. The solutions at time $t = t_{n+1}$ of (4.2.1) and (4.2.2) are

$$y(t_{n+1}) = y_0 e^{\lambda t_{n+1}} \tag{4.2.3}$$

and

$$y(t_{n+1}) = \frac{y_0}{e^{-\lambda t_{n+1}} + (1 - e^{-\lambda t_{n+1}})y_0}, \tag{4.2.4}$$

respectively. Setting $y_n := y(t_n)$ permits us to re-write (4.2.3) in an equivalent form as follows

$$\begin{aligned} y(t_{n+1}) - y(t_n) &= y_0 e^{\lambda t_{n+1}} - y(t_n) \\ &= y_0 e^{\lambda(t_n+\Delta t)} - y(t_n) \\ &= y(t_n)e^{\lambda \Delta t} - y(t_n) \\ &= y(t_n)e^{\lambda \Delta t} - y(t_n) \\ &= \lambda y(t_n)(e^{\lambda \Delta t} - 1)/\lambda \end{aligned}$$

and we have

$$\frac{y_{n+1} - y_n}{\frac{e^{\lambda \Delta t}-1}{\lambda}} = \lambda y_n. \tag{4.2.5}$$

In the similar manner from (4.2.4), we have

$$\begin{aligned} y(t_{n+1}) &= \frac{y_0 e^{\lambda t_{n+1}}}{1 + y_0 e^{\lambda t_{n+1}} - y_0} \\ &= \frac{y(t_n)e^{\lambda \Delta t}}{1 + y(t_n)e^{\lambda \Delta t} - y(t_n)}. \end{aligned}$$

Thus

$$\begin{aligned} y(t_{n+1}) &= y(t_n)e^{\lambda \Delta t} - y(t_n)y(t_{n+1})e^{\lambda \Delta t} + y(t_n)y(t_{n+1}) \\ y(t_{n+1}) - y(t_n) &= (e^{\lambda \Delta t} - 1)y(t_n)(1 - y(t_{n+1})) \\ \frac{y(t_{n+1}) - y(t_n)}{(e^{\lambda \Delta t} - 1)/\lambda} &= \lambda y(t_n)(1 - y(t_{n+1})) \end{aligned}$$

which can be written as

$$\frac{y_{n+1}-y_n}{\frac{e^{\lambda\Delta t}-1}{\lambda}} = \lambda y_n(1-y_{n+1}). \tag{4.2.6}$$

Equations (4.2.5) and (4.2.6) are exact schemes of (4.2.1) and (4.2.2), respectively. Mickens [26] established exact schemes for a substantial number of differential equations of applied sciences. For convenience, Table 4.1 of exact schemes produced in [24] is incorporated here.

Table 4.1: *Exact schemes of some ODE's and PDE's*

Differential Equations	Exact Finite Difference Schemes
$\frac{dy}{dt} = -\lambda y$	$\frac{y_{k+1}-y_k}{\left(1-e^{-\lambda\Delta t}\right)/\lambda} = -\lambda y_k$
$\frac{d^2y}{dt^2} + \omega^2 y = 0$	$\frac{y_{k+1}-2y_k+y_{k-1}}{\frac{4}{\omega^2}\sin^2\left(\frac{\Delta t\omega}{2}\right)} + \omega^2 y_k = 0$
$\frac{dy}{dt} = \lambda_1 y - \lambda_2 y^2$	$\frac{y_{k+1}-y_k}{\left(e^{\lambda_1\Delta t}-1\right)/\lambda_1} = \lambda_1 y_k - \lambda_2 y_{k+1} y_k$
$2\frac{dy}{dt} + y = \frac{1}{y}$	$\frac{2(y_{k+1}-y_k)}{(1-e^{-\Delta t})} + \frac{y_k^2}{\left(\frac{y_{k+1}+y_k}{2}\right)} = \frac{1}{\left(\frac{y_{k+1}+y_k}{2}\right)}$
$\frac{d^2y}{dt^2} = \lambda\frac{dy}{dt}$	$\frac{y_{k+1}-2y_k+y_{k-1}}{\left(\frac{e^{\lambda\Delta t}-1}{\lambda}\right)\Delta t} = \lambda\left(\frac{y_k-y_{k-1}}{\Delta t}\right)$
$u_t + u_x = u(1-u)$	$\frac{u_m^{k+1}-u_m^k}{e^{\Delta t}-1} + \frac{u_m^k-u_{m-1}^k}{e^{\Delta x}-1} = u_{m-1}^k\left(1-u_m^{k+1}\right)$ for $\Delta t = \Delta x$
$y_{tt} - y_{xx} = 0$	$y_m^{k+1} - 2y_m^k + y_m^{k-1} = y_{m+1}^k - 2y_m^k + y_{m-1}^k$
$\frac{d^2y}{dt^2} + 2\epsilon\frac{dy}{dt} + y = 0$	$\psi(\omega,\Delta t) = \frac{\epsilon e^{-\epsilon\Delta t}}{\sqrt{1-\epsilon^2}} + e^{-\epsilon\Delta t}\cos\left(\sqrt{1-\epsilon^2}\right)\Delta t, \phi(\epsilon,\Delta t)$ $= \frac{\epsilon e^{-\epsilon\Delta t}}{\sqrt{1-\epsilon^2}}\sin\left(\sqrt{1-\epsilon^2}\right)\Delta t$, $\frac{y_{k+1}-2y_k+y_{k-1}}{\phi^2} + 2\epsilon\left(\frac{y_k-\psi y_{k-1}}{\phi}\right) + \frac{2(1-\psi)y_k+(\phi^2+\psi^2-1)y_{k-1}}{\phi^2} = 0$
$\frac{\partial c}{\partial t} + P_{n-1}(t)\frac{\partial c}{\partial x} = \lambda c(1-c)$ $P_{n-1}(t) = \sum_{i=0}^{i=n-1} a_i t^i$	$\frac{C^{k+1}(x)-C^k(\tilde{x}^k)}{\left(e^{\lambda\Delta t}-1\right)/\lambda} = \lambda C^k(\tilde{x}^k)\left(1-C^{k+1}(x)\right)$, $\tilde{x}^k = x - [P_n((k+1)\Delta t) - P_n(k\Delta t)],\ P_n(t) = \int_0^t P_{n-1}(\tau)d\tau.$
$\frac{\partial c}{\partial t} + P_{n-1}(t)\frac{\partial c}{\partial x} = \lambda c$	$\frac{C^{k+1}(x)-C^k(\tilde{x}^k)}{\left(e^{\lambda\Delta t}-1\right)/\lambda} = \lambda C^k(\tilde{x}^k).$
$\frac{\partial c}{\partial t} + P_{n-1}(t)\frac{\partial c}{\partial x} = \mu + \lambda c$	$\frac{C^{k+1}(x)-C^k(\tilde{x}^k)}{\left(e^{\lambda\Delta t}-1\right)/\lambda} = \mu + \lambda C^k(\tilde{x}^k).$

The simple examples (4.2.1) and (4.2.3), (4.2.2) and (4.2.4) as well as Table 4.1 illustrate the need for the structure of the right hand side of the differential equation to be intrinsically reflected in the discrete schemes if they are required to replicate the qualitative properties of the solution of the differential equation. Equation (4.2.6) and similar equations in the table illustrate in addition the need of approximating nonlinear terms in a nonlocal way. These comments motivate the following definition due to [7].

Definition 4.2.3. *The difference method given by Equation (3.2.5) is called a* ***non-standard finite difference method*** *if at least one of the following conditions is satisfied:*

- *In the first order discrete derivative $D_{\Delta t}y_n$, the classical denominator Δt is replaced by a nonnegative function $\phi : (0,\infty) \to (0,\infty)$ satisfying*
$$\phi(\Delta t) = \Delta t + O[(\Delta t)^2]. \tag{4.2.7}$$
[e.g. $\phi(\Delta t) = 1 - e^{-\Delta t}$, $\phi(\Delta t) = (e^{\lambda \Delta t} - 1)/\lambda$].

- *In the expression $F_{\Delta t}(f, y_n)$, nonlinear terms are approximated in a nonlocal way, i.e., by suitable function of several points of mesh. e.g. $y^2(t_n) \approx y_{n+1}y_n$.*

■

In [26], Mickens set the following rules for the design of non-standard schemes:

Rule 1. The orders of the discrete derivatives should be equal to the orders of the corresponding derivatives of the differential equation.

Rule 2. Denominator functions for the discrete derivatives must, in general, be expressed in terms of more complicated functions of the step-sizes than those conventionally used.

Rule 3. Nonlinear terms should, in general, be replaced by nonlocal discrete representations.

Rule 4. Special conditions that hold for the solutions of the differential equations should also hold for the solutions of the finite difference scheme.

Rule 5. The scheme should not introduce extraneous or spurious solutions.

Remark 4.2.4. For an overview on non-standard finite difference schemes, we refer the reader to [24], [35] and the edited volumes [17], [28]. In the formal Definition 4.2.3 only two of five Mickens rules are needed because most of the other rules appear as properties of the differential equation with respect to which a discrete scheme might be qualitatively stable. ∎

Table 4.2: *Non-standard finite difference schemes*

Differential Equations	Non-standard Finite Difference Schemes
$\frac{d^2y}{dt^2} + y + \beta y^3 = 0$	$\frac{y_{k+1}-2y_k+y_{k-1}}{4\sin^2\left(\frac{\Delta t}{2}\right)} + y_k + \beta\left(\frac{\sin^2(\Delta t)}{4\sin^2\left(\frac{\Delta t}{2}\right)}\right) y_k^3 = 0$
$\frac{d^2y}{dt^2} + y + \epsilon y^2 = 0$	$\frac{y_{k+1}-2y_k+y_{k-1}}{4\sin^2\left(\frac{\Delta t}{2}\right)} + y_k + \epsilon\left(\frac{\sin^2(\Delta t)}{4\sin^2\left(\frac{\Delta t}{2}\right)}\right) y_k^2 = 0$
$\frac{d^2y}{dt^2} + y = \epsilon(1-y^2)\frac{dy}{dt}$	$\frac{y_{k+1}-2y_k+y_{k-1}}{4\sin^2\left(\frac{\Delta t}{2}\right)} + y_k = \epsilon\left(\frac{\sin(\Delta t)}{2\sin\left(\frac{\Delta t}{2}\right)}\right)(1-y_k^2)\left(\frac{y_k-\cos(\Delta t)y_{k-1}}{2\sin\left(\frac{\Delta t}{2}\right)}\right)$

Remark 4.2.5. The above-mentioned schemes were constructed by Mickens [26], who placed the emphasis on the structure of the discrete derivative. An indication on how the nonlinear terms could be approached in a nonlocal way is given in Section 4.5. In the paper [24], the schemes given in Table 4.2 should have been listed as non-standard finite difference schemes instead of exact schemes of the corresponding differential equations. Their exact schemes are not known since these do not have explicit solutions. ∎

4.3 Elementary Stable Schemes

As it was mentioned in Chapter 3, the theta methods (3.7.1) and (3.7.2) are the point of departure of our study. The popularity of the theta methods, also referred to as the weighted average method, is due in large part to their simplicity making it easy to program and efficient on large problems [9]. In this section, we introduce elementary stable non-standard theta methods and demonstrate their theoretical and practical power over the standard ones.

The terminologies we use here were clarified in Section 2.2.2 for continuous dynamical systems and in Section 2.2.3 for discrete dynamical systems. In particular, we assume, once and for all, that all fixed-points $\tilde{y}$ of the dynamical systems (2.2.1) are hyperbolic in the sense of Definition 2.3.10, each Jacobian of f at $\tilde{y}$ being denoted by J.

We would like to design for (2.2.1) numerical methods the solution of which replicate the qualitative properties of the fixed-points. We start with the following definition ([7], [26]):

Definition 4.3.1. *A difference scheme (3.2.5) for approximating (2.2.1) is called* ***elementary stable*** *if, for any value of the step size Δt, its fixed-points $\tilde{y}$ are exactly those of the differential system (2.2.1), and these fixed-points for the difference scheme have the same linear stability/instability properties as for the differential system.* ∎

In view of Definition 2.2.11 and Definition 2.3.8, it is clear that the theta methods (3.7.1) and (3.7.2) have no spurious fixed-points compared to the system (2.2.1). Indeed given $\tilde{y} \in \mathbb{R}^m$, the constant sequence $y_n = \tilde{y}$ is the solution of (3.7.1) or (3.7.2) if and only if $f(\tilde{y}) = 0$.

However, Theorem 3.7.6, shows that the classical theta methods are not elementary stable for $\theta \in [0, \frac{1}{2})$, due to the constraint (3.7.35) on the value of Δt when λ is an eigenvalue of J with $Re\lambda < 0$. On the other hand, when λ is an eigenvalue of J with $Re\lambda > 0$ and $\theta \in (\frac{1}{2}, 1]$, we have from (3.7.32)

$$\begin{aligned} |R(\lambda\Delta t)|^2 &= \frac{1+2\Delta t(1-\theta)|Re\lambda| + (\Delta t)^2(1-\theta)^2|\lambda|^2}{1-2\Delta t\theta Re\lambda + \Delta t)^2\theta^2|\lambda|^2} \\ &< 1 \end{aligned}$$

if and only if

$$\Delta t > \frac{2|Re\lambda|}{(2\theta-1)|\lambda|^2}. \tag{4.3.1}$$

Under the condition (4.3.1), the discrete solution $\{y_n\}$ of the linearised theta method (3.7.30) will tend to zero as $n \to \infty$, while the solution $y(t) = e^{tJ}y_0$ of the continuous linearization (2.2.7) diverges as $t \to \infty$. This discrepancy in the linear stability/instability properties of fixed-points for the theta methods and the differential equation means that the theta methods are equally not elementary stable for $\theta \in (\frac{1}{2}, 1]$. For $\theta = \frac{1}{2}$, the analysis above shows that the theta methods (i.e. Trapezoidal rule and mid-point rule) are elementary stable. This explains why in what follows, we implicitly assume that $\theta \neq \frac{1}{2}$.

Coming back to the general framework of the system (2.2.1), its dynamics will be captured by a fixed nonzero number

$$q \geq \max\left\{\frac{|\lambda|^2}{2|Re\lambda|}; \lambda \in E\right\}, \tag{4.3.2}$$

where

$$E = \bigcup\{\sigma(Jf(\tilde{y}));\ \tilde{y} \in \mathbb{R}^m, f(\tilde{y}) = 0\} \tag{4.3.3}$$

is the finite set of all the eigenvalues of the Jacobian matrix $Jf(\tilde{y})$ of f at all fixed-points. We also consider a non-negative function ϕ satisfying the asymptotic relation (4.2.7) as well as the property

$$0 < \phi(z) < 1, \quad \text{for} \quad z > 0. \tag{4.3.4}$$

A typical example is

$$\phi(z) = 1 - e^{-z}. \tag{4.3.5}$$

With the number q in (4.3.2) and the function ϕ in (4.3.4), we associate the function

$$\psi := \frac{\phi(q\Delta t)}{q}, \tag{4.3.6}$$

which satisfies (4.2.7). We are now in a position to introduce the following non-standard one-stage and two-stage theta methods:

$$\frac{y_{n+1}-y_n}{\psi(\Delta t)} = f[\theta y_{n+1} + (1-\theta)y_n], \tag{4.3.7}$$

and

$$\frac{y_{n+1}-y_n}{\psi(\Delta t)} = \theta f(y_{n+1}) + (1-\theta)f(y_n), \tag{4.3.8}$$

respectively. We have the following important result:

Theorem 4.3.2. *The non-standard theta method (4.3.7) and (4.3.8), where ψ is defined by (4.3.6) and (4.3.4), are elementary stable.*

Proof. As it was seen earlier for the classical schemes (3.7.1) and (3.7.2), the non-standard schemes (4.3.7) and (4.3.8) have no spurious fixed-points compared to the system (2.2.1). The linearisation of the non-standard schemes (4.3.7) and (4.3.8), about a fixed-point $\tilde{y}$ is

$$\frac{y_{n+1}-y_n}{\psi(\Delta t)} = J[\theta y_{n+1} + (1-\theta)y_n], \tag{4.3.9}$$

instead of (3.7.30). Thus the stability function in (3.7.32) becomes

$$R(\lambda\Delta t) = \frac{1+\psi(\Delta t)(1-\theta)\lambda}{1-\psi(\Delta t)\theta\lambda}. \tag{4.3.10}$$

For $\lambda = \lambda_1 + \imath\lambda_2 \in E$, we have:

$$\begin{aligned} |R(\lambda\Delta t)|^2 &\equiv \left|\frac{1+\psi(\Delta t)(1-\theta)\lambda}{1-\psi(\Delta t)\theta\lambda}\right|^2 \\ &= \frac{1+2\lambda_1\phi(q\Delta t)(1-\theta)/q + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2}{1-2\lambda_1\phi(q\Delta t)\theta/q + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2}. \end{aligned}$$

Let $\tilde{y}$ be a fixed-point of the differential equation (2.2.1). Two cases are possible. Firstly, $\tilde{y}$ can be linearly stable, which, by Theorem 2.2.16

and Remark 2.2.18, implies that $\lambda_1 < 0$ for any eigenvalue $\lambda \in \sigma(J)$. Then by (4.3.4) and (4.3.2), we have:

$$\begin{aligned} |R(\lambda\Delta t)|^2 &= \frac{1 - 2|\lambda_1|\phi(q\Delta t)(1-\theta)/q + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2}{1 + 2|\lambda_1|\phi(q\Delta t)\theta/q + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2} \\ &< 1 - 2|\lambda_1|\phi(q\Delta t)(1-\theta)/q + |\lambda|^2\phi(q\Delta t)(1-\theta)/q^2 \\ &\leq 1. \end{aligned}$$

This shows that the fixed-point $\tilde{y}$ is linearly stable for the scheme (4.3.7) and (4.3.8) in view of Theorem 2.3.13. Secondly, the fixed-point $\tilde{y}$ of (2.2.1) can be linearly unstable, i.e., there exists an eigenvalue $\lambda \in \sigma(J)$ such that $\lambda_1 > 0$. Working out the above expression of $|R(\lambda\Delta t)|^2$, we obtain

$$\frac{1 + 2\lambda_1\phi(q\Delta t)(1-\theta)/q + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2}{1 - 2\lambda_1\phi(q\Delta t)\theta/q + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2} > 1$$

if and only if

$$2\lambda_1 + |\lambda|^2\phi(q\Delta t)/q - 2|\lambda|^2\phi(q\Delta t)\theta/q > 0.$$

But

$$2\lambda_1 + |\lambda|^2\phi(q\Delta t)/q - 2|\lambda|^2\phi(q\Delta t)\theta/q \geq 2\lambda_1 - |\lambda|^2\phi(q\Delta t)/q$$

which, in view of (4.3.2) and (4.3.4), shows that

$$2\lambda_1 - |\lambda|^2\phi(q\Delta t)/q > 0.$$

Thus the fixed-point $\tilde{y}$ is linearly unstable for the scheme (4.3.7) or (4.3.8). We have thus proved that the schemes (4.3.7) and (4.3.8) are elementary stable. □

Theorem 4.3.2 is given in [13], [14] in the particular case when $\theta = 0$. By construction and the way it is involved in the proof of Theorem 4.3.2, the relation (4.3.2) is the sharpest condition compared to those in the literature for capturing the dynamics of the differential equation (see

for example [7], [26]). Thus, Theorem 4.3.2 is theoretically interesting. However, it is practically difficult to find q that meets the requirement (4.3.2) since no lower bounds are available in general for the real parts $|Re\lambda|$ of the eigenvalues of an arbitrary matrix. We want to overcome this difficulty. Following the idea in [25], it is convenient to use the identity $Re\lambda = \cos\arg\lambda$ that, in view of (4.3.2), yields the relation

$$|\cos\arg\lambda| \geq \frac{|\lambda|}{2q} \quad \text{for all} \quad \lambda \in E. \tag{4.3.11}$$

The condition (4.3.2) in its equivalent form (4.3.11), implies a restriction on the location of the eigenvalues in the complex plane in the following precise way:

Theorem 4.3.3. *The condition (4.3.2) is equivalent to saying that the eigenvalues of all the matrices J are contained in some wedge in the complex plane, i.e.*

$$E \subset W^j := \{\lambda \in \mathbb{C}; |\cos\arg\lambda| \geq \frac{j}{2}\} \tag{4.3.12}$$

for some $j \in [0, 2]$.

Proof. If q satisfies (4.3.2) and thus the inequality (4.3.11) holds, then we have the inclusion (4.3.12) with

$$j := \frac{\min\{|\lambda|; \lambda \in E\}}{q}.$$

Conversely, if (4.3.12) holds, then the number

$$q := \frac{\max\{|\lambda|; \lambda \in E\}}{j}$$

satisfies (4.3.2). □

In the following result, we present a somewhat refined version of the inclusion (4.3.12); the particular case when $j = 1$ was analysed in [6] and [25].

Theorem 4.3.4. *With a fixed real number $0 < j \leq 2$, we associate the wedges in the left and right hands complex plane defined by*

$$W_l^1 := \{\lambda \in \mathbb{C};\ Re\lambda < 0 \ and\ |\cos \arg \lambda| \geq \frac{j}{2}\} \tag{4.3.13}$$

and

$$W_r^1 := \{\lambda \in \mathbb{C};\ Re\lambda > 0 \ and\ |\cos \arg \lambda| \geq \frac{j}{2}\}. \tag{4.3.14}$$

Let the dynamics of the differential equation be captured by a number q satisfying

$$q \geq \frac{\max\{|\lambda|; \lambda \in E\}}{j}. \tag{4.3.15}$$

Then, the non-standard theta methods (4.3.7) and (4.3.8) are elementary stable whenever we have the inclusions

$$E \subset W_l^1 \cup \{\lambda \in \mathbb{C}; Re\lambda > 0\} \ for\ \theta \in [0, \frac{1}{2}) \tag{4.3.16}$$

and

$$E \subset W_r^1 \cup \{\lambda \in \mathbb{C}; Re\lambda < 0\} \ for\ \theta \in (\frac{1}{2}, 1]. \tag{4.3.17}$$

The region of elementary stability on the right hand side of (4.3.16) and (4.3.17) are shown on Fig 4.1 and Fig 4.2 for $j = 1$.

Proof. The proof works as that of Theorem 4.3.2, observing that we have to consider four cases in (4.3.16) and (4.3.17).
More precisely, in view of (4.3.10), we have for $\lambda \in \mathbb{C},\ Re\lambda < 0$,

$$|R(\lambda\Delta t)|^2 = \frac{1 - 2|\lambda|\phi(q\Delta t)(1-\theta)/q|\cos \arg \lambda| + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2}{1 + 2|\lambda|\phi(q\Delta t)\theta/q|\cos \arg \lambda| + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2} \tag{4.3.18}$$

while for $Re\lambda > 0$,

$$|R(\lambda\Delta t)|^2 = \frac{1 + 2|\lambda|\phi(q\Delta t)(1-\theta)/q|\cos \arg \lambda| + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2}{1 - 2|\lambda|\phi(q\Delta t)\theta/q|\cos \arg \lambda| + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2}. \tag{4.3.19}$$

Consider now the case when $\theta \in [0, \frac{1}{2})$ and let $\lambda \in E$ be in the right hand side of (4.3.16). This means that either $\lambda \in W_l^1$ or $Re\lambda > 0$. If $\lambda \in W_l^1$, then we have from (4.3.18)

$$\begin{aligned} |R(\lambda\Delta t)|^2 &< 1 - 2|\lambda|\phi(q\Delta t)(1-\theta)/q|\cos\arg\lambda| + |\lambda|^2(\phi(q\Delta t))^2(1-\theta)^2/q^2 \\ &< 1 - 2|\lambda|\phi(q\Delta t)(1-\theta)/q|\cos\arg\lambda| + j|\lambda|\phi(q\Delta t)(1-\theta)/q \\ &= 1 + |\lambda|\phi(q\Delta t)(1-\theta)/q(j - 2|\cos\arg\lambda|) \text{ by 4.3.15} \\ &\leq 1. \end{aligned}$$

If $Re\lambda > 0$, then it follows from (4.3.19) and the fact that $\theta \in [0, \frac{1}{2})$ that

$$|R(\lambda\Delta t)|^2 > 1.$$

Consider finally the case when $\theta \in (\frac{1}{2}, 1]$ and let $\lambda \in E$ be in the right hand side of (4.3.17), which means that either $\lambda \in W_r^1$ or $Re\lambda < 0$. When $\lambda \in W_r^1$, we use (4.3.19) and (4.3.15) to obtain

$$\begin{aligned} |R(\lambda\Delta t)|^2 &> \frac{1}{1 - 2|\lambda|\phi(q\Delta t)\theta/q|\cos\arg\lambda| + |\lambda|^2(\phi(q\Delta t))^2\theta^2/q^2} \\ &\geq \frac{1}{1 - 2|\lambda|\phi(q\Delta t)\theta/q|\cos\arg\lambda| + j|\lambda|(\phi(q\Delta t))\theta/q} \\ &= \frac{1}{1 + |\lambda|\phi(q\Delta t)\theta/q(j - 2|\cos\arg\lambda|)} \\ &\geq 1. \end{aligned}$$

For $Re\lambda < 0$, we infer directly from (4.3.18) and from $\theta \in (\frac{1}{2}, 1]$ that

$$|R(\lambda\Delta t)|^2 < 1.$$

Thus, the non-standard theta methods (4.3.7) and (4.3.8) are elementary stable . □

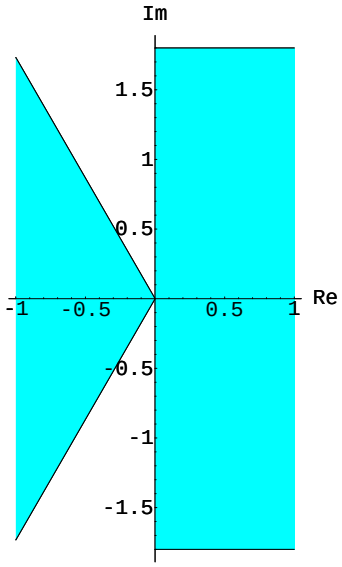

Figure 4.1: Region of elementary stability for $\theta \in [0, \frac{1}{2})$

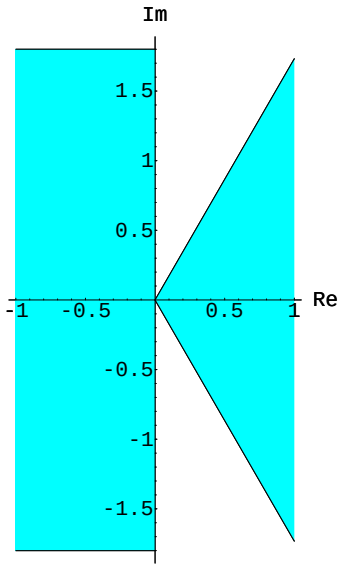

Figure 4.2: Region of elementary stability for $\theta \in [\frac{1}{2}, 1]$

Remark 4.3.5. Unlike (4.3.2), the choice of the number q in (4.3.15) is not so critical if the system is non-stiff. In practice, we may take $jq := \max\|J(g)(\widetilde{y})\|_\infty$, where $\|\cdot\|_\infty$ is the matrix norm associated with the supremum norm on $\mathbb{R}^m$. ∎

Remark 4.3.6. With the definition (4.3.13)-(4.3.14) of the wedges, the inclusions (4.3.16)-(4.3.17) for elementary stability of the scheme under consideration are in line with what is done in the classical theory of absolute stability of numerical methods for ordinary differential equations (see [22]). This observation permits us to link the extreme cases when $j = 0$ and $j = 2$ in (4.3.12) to the classical concepts of A/A_0-stable schemes. In [25], the terminology $A-$ and A_0-elementary stable schemes is used when $j = 0$ and $j = 2$, respectively.

Furthermore, from the comparative analysis in [25], it follows that the non-standard theta methods have much larger regions of absolute elementary stability than the standard ones. Some of the advantages of the non-standard theta methods over the standard ones are summarised in Table 4.3 where we recall that the case $\theta = \frac{1}{2}$ is excluded as the corresponding standard schemes preserve all the involved properties. ∎

Table 4.3: *Comparison between standard and non-standard θ-methods*

	Explicit θ-method ($\theta = 0$)		Implicit θ-method			
			$\theta \in (0, \frac{1}{2})$		$\theta \in (\frac{1}{2}, 1]$	
	Std.	Non-std.	Std.	Non-std.	Std.	Non-std.
Elementary stability	No	Yes	No	Yes	No	Yes
A-Elementary stability	No	No	No	No	Yes	Yes
A_0-Elementary stability	No	Yes	No	Yes	Yes	Yes

Remark 4.3.7. It should be noted that the non-standard theta methods (4.3.7) and (4.3.8) enjoy the consistency and convergence properties stated in Theorem 3.7.2 and Remark 3.7.3 for the classical schemes.

This is due to the property (4.2.7) that the denominator ψ satisfies. The analogy of the error estimate (3.7.16) is proved in [6]. ∎

To conclude this section, we consider an example that confirms the superiority of the non-standard approach over the standard one.

Example 4.3.8.

A typical example is the logistic equation

$$y' = 25y(1-y),\ y(0) = y_0, \tag{4.3.20}$$

whose exact solution and exact scheme are

$$y(t) = \frac{y_0}{y_0 + (1-y_0)e^{-25\Delta t}}, \tag{4.3.21}$$

and

$$\frac{y_{n+1} - y_n}{\frac{e^{25\Delta t}-1}{25}} = 25y_n(1-y_{n+1}). \tag{4.3.22}$$

The forward Euler difference scheme yields

$$\frac{y_{n+1} - y_n}{\Delta t} = 25y_n(1-y_n). \tag{4.3.23}$$

In accordance with (4.3.7) - (4.3.8) for $\theta = 0$, we introduce the non-standard scheme

$$\frac{y_{n+1} - y_n}{\frac{1-e^{-25\Delta t}}{25}} = 25y_n(1-y_n). \tag{4.3.24}$$

The exact solution for the logistic equation, the Euler forward difference scheme (4.3.23) and the non-standard finite difference schemes (4.3.24) are visualised in Fig. 4.3 for $\Delta t = 0.01$, Fig. 4.4 for $\Delta t = 0.067$, and Fig. 4.5 for $\Delta t = 0.067$, respectively using various initial conditions.

On comparing Figures 4.3, 4.4, and 4.5, it is evident that the non-standard scheme in Fig.4.5 gives a more reliable simulation of the exact solution in Fig.4.3 than the standard Euler Scheme in Fig. 4.4.

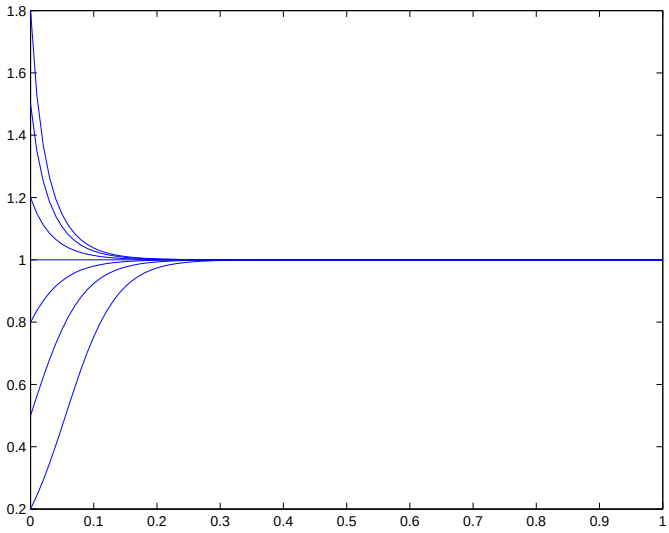

Figure 4.3: Exact solution for the logistic equation.

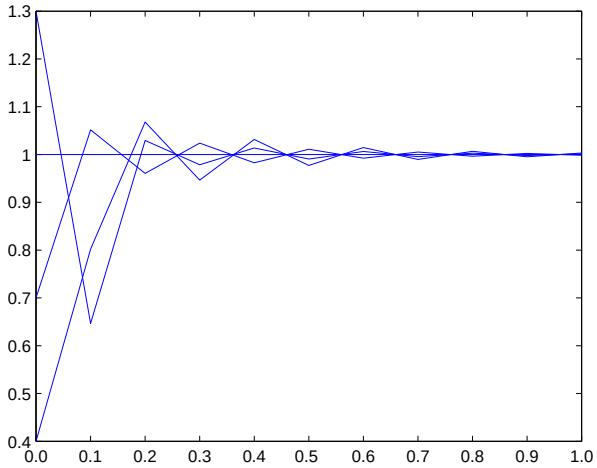

Figure 4.4: Standard Euler scheme for the logistic equation.

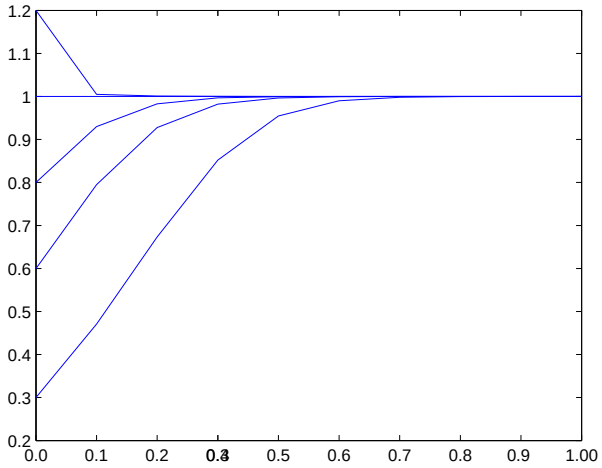

Figure 4.5: Non-standard Euler scheme for the logistic equation.

4.4 Dissipative Non-standard Theta Methods

This section is a follow up of the concern mentioned in Remark 3.7.8 regarding the dissipativity of theta methods for $\theta \in [0, \frac{1}{2})$. Firstly, when $\theta \in (\frac{1}{2}, 1]$, Theorem 3.7.7 carries over easily to the non-standard setting and we state it here for convenience.

Theorem 4.4.1. *Consider (2.2.1) as a dissipative dynamical system in the setting specified in Theorem 3.7.7. We approximate this dynamical system by the non-standard theta methods (4.3.7) or (4.3.8) where the only requirement on the denominator $\psi(\Delta t)$ is the asymptotic behaviour (4.2.7). Then, for $\theta \in (\frac{1}{2}, 1]$, these non-standard schemes are dissipative in the sense of Definition 2.3.3. The absorbing sets are those given in Theorem 3.7.7 on the understanding that Δt is replaced by $\psi(\Delta t)$ wherever this is applicable.*

Regarding the case when $\theta \in [0, \frac{1}{2})$, we managed to deal with the marginal case $\theta = 0$. More precisely, we show in what follows how the non-standard approach can help to successfully modify the simple Euler method so that it is dissipative.

To this end, we suppose that $f : \mathbb{R}^m \to \mathbb{R}^m$ satisfies the structural assumption (2.2.12) involving $\alpha > 0$ and $\beta > 0$ and we assume without loss of generality that

$$\beta < 1. \tag{4.4.1}$$

Furthermore, we assume that there exist positive constants γ and $c > 1$ such that, for every $y \in \mathbb{R}^m$:

$$||f(y)||^2 \leq \gamma + c||y||^2. \tag{4.4.2}$$

Remark 4.4.2. The condition (4.4.2) holds if the function f is Lipschitz, which is one of the widely used requirement for (2.2.1) to define a dynamical system on $\mathbb{R}^m$. ∎

We have the following important result:

Theorem 4.4.3. *For $\theta = 0$, the non-standard finite difference scheme (4.3.7) or (4.3.8) where $\psi(\Delta t)$ is given by (4.3.6) with $q := \frac{c}{\beta}$, is a dissipative dynamical system.*

Proof. From (4.3.7) or (4.3.8) with $\theta = 0$, we have

$$\frac{y_{n+1} - y_n}{\psi(\Delta t)} = f(y_n). \tag{4.4.3}$$

Multiplying (4.4.3) by y_{n+1}, we obtain

$$\frac{\langle y_{n+1} - y_n, y_{n+1}\rangle}{\psi(\Delta t)} = \langle f(y_n), y_{n+1}\rangle . \tag{4.4.4}$$

We use (4.4.3) on the left hand side of (4.4.4) and on the right hand side we apply

$$\langle u - v, u\rangle = \frac{1}{2}(||u||^2 - ||v||^2 + ||u - v||^2).$$

$$\begin{aligned}
\frac{1}{2\psi(\Delta t)}(||y_{n+1}||^2 - ||y_n||^2 + ||y_{n+1} - y_n||^2) &= \langle f(y_n), y_n\rangle + \langle f(y_n), y_{n+1} - y_n\rangle \\
\frac{1}{2\psi(\Delta t)}(||y_{n+1}||^2 - ||y_n||^2 + (\psi(\Delta t))^2||f(y_n)||^2) &= \langle f(y_n), y_n\rangle + \psi(\Delta t)\langle f(y_n), f(y_n)\rangle \\
&= \langle f(y_n), y_n\rangle + \psi(\Delta t)||f(y_n)||^2.
\end{aligned}$$

From (2.2.12), (4.3.4), (4.4.1) and (4.4.2), we obtain

$$\begin{aligned}
\frac{||y_{n+1}||^2 - ||y_n||^2}{\psi(\Delta t)} &\leq 2\alpha - 2\beta||y_n||^2 + \frac{\beta}{c}\phi(c\Delta t/\beta)(\gamma + c||y_n||^2) \\
&< 2\alpha + \frac{\beta\gamma}{c} - \beta||y_n||^2.
\end{aligned}$$

Thus

$$||y_{n+1}||^2 < \left(2\alpha + \frac{\beta\gamma}{c}\right)\psi(\Delta t) + [1 - \beta\psi(\Delta t)]||y_n||^2.$$

Applying the discrete Gronwall inequality (Lemma 2.3.6) yields

$$||y_n||^2 \leq \left(\frac{2\alpha}{\beta} + \frac{\gamma}{c}\right)[1 - (1 - \beta\psi(\Delta t))^n] + ||y_0||^2(1 - \beta\psi(\Delta t))^n.$$

Thus

$$\limsup_{n\to\infty} ||y_n||^2 \leq \frac{2\alpha}{\beta} + \frac{\gamma}{c}$$

and it follows that the discrete dynamical system under consideration is dissipative, the closed ball $\bar{B}\left(0, \sqrt{\frac{2\alpha}{\beta} + \frac{\gamma}{c}} + \epsilon\right)$ being an absorbing set for every $\epsilon > 0$. □

We have up to this point demonstrated numerically the power of the non-standard finite difference schemes over the standard ones as far as elementary stability is concerned. We now turn our attention on the dissipative property by considering two examples.

Example 4.4.4.

We consider the dynamical system defined by

$$\frac{dy_1}{dt} = 1 + 5y_2 - y_1 \tag{4.4.5}$$

$$\frac{dy_2}{dt} = 1 - 5y_1 - y_2, \tag{4.4.6}$$

whose fixed point is $(\frac{3}{13}, \frac{-2}{13})$. The right hand-side of the system is the vector function

$$f(y) = \begin{pmatrix} 1 + 5y_2 - y_1 \\ 1 - 5y_1 - y_2 \end{pmatrix}$$

which satisfies the structural assumption (2.2.12) and (4.4.1) in the following precise form:

$$\begin{aligned} \langle f(y), y\rangle &= (1 + 5y_2 - y_1)y_1 + (1 - 5y_1 - y_2)y_2 \\ &= y_1 + 5y_1y_2 - y_1^2 + y_2 - 5y_1y_2 - y_2^2 \\ &\leq \frac{1}{2}(1 + y_1^2) - y_1^2 + \frac{1}{2}(1 + y_2^2) - y_2^2 \\ &= 1 - \frac{1}{2}||y||_2^2. \end{aligned}$$

Hence $\alpha = 1$ and $\beta = \frac{1}{2}$. Furthermore, the norm of $f(y)$ can be estimated as follows:

$$\begin{aligned} ||f(y)||_2^2 &= (1 + 5y_2 - y_1)^2 + (1 - 5y_1 - y_2)^2 \\ &= 1 + 25y_2^2 + y_1^2 + 10y_2 - 2y_1 - 10y_1y_2 + 1 + 25y_1^2 + y_2^2 \\ &\quad -10y_1 - 2y_2 + 10y_1y_2 \\ &= 2 + 26(y_1^2 + y_2^2) + 10y_2 - 10y_1 \\ &\leq 2 + 26(y_1^2 + y_2^2) + 5(1 + y_2^2) + 5(1 + y_1^2) \\ &= 12 + 31||y||_2^2. \end{aligned}$$

Hence, the requirement (4.4.2) is met with $\gamma = 12$ and $c = 31$. With $\phi(\Delta t) = 1 - e^{-\Delta t}$, the non-standard scheme considered in Theorem 4.4.3 reads as

$$\frac{y_{n+1} - y_n}{\frac{1-e^{-q\Delta t}}{q}} = f(y_n), \tag{4.4.7}$$

where $q = \dfrac{c}{\beta} = 62$. Taking the step size $\Delta t = 0.1$, Fig. 4.6 and Fig. 4.7 give the phase diagrams of the numerical solutions of the system (4.4.5)-(4.4.6) by the non-standard finite difference scheme (4.4.7) using the initial conditions $y(0) = (10, 10)$ and $y(0) = (\pm 10, \pm 10)$, respectively. The dissipativity of the scheme is apparent.

For comparison, we apply to the system (4.4.5)–(4.4.6) the standard forward Euler method (3.7.17) with the same step size and initial condition $y(0) = (10, 10)$. The phase diagram of the numerical solution given in Fig.4.8 is not indicative of dissipativity.

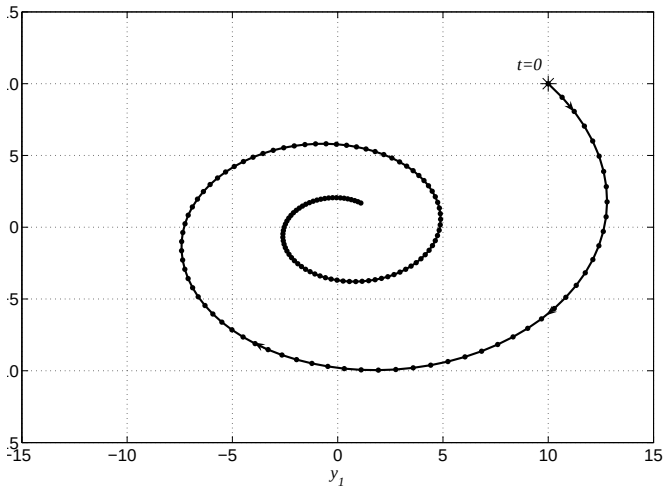

Figure 4.6: Dissipative non-standard scheme

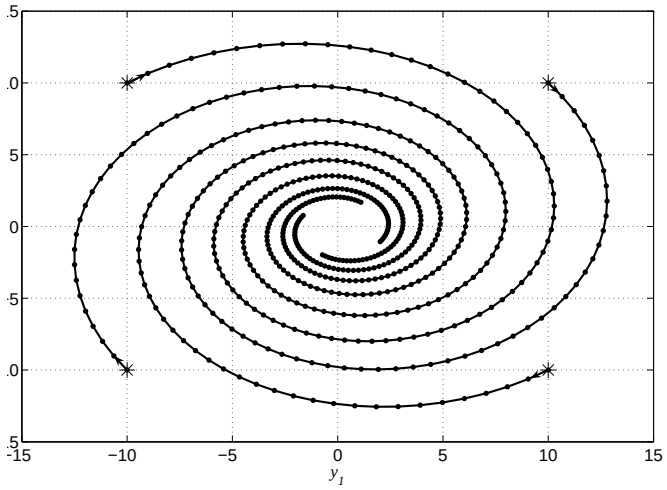

Figure 4.7: Further dissipative non-standard scheme

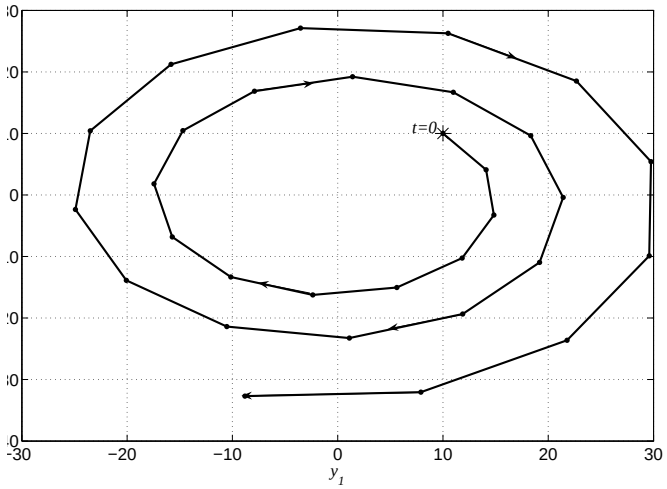

Figure 4.8: Non-dissipative standard forward Euler scheme

Example 4.4.5.

We consider the dynamical system defined by

$$y_1' = -y_1 - 5y_2 + \frac{y_1}{\sqrt{y_1^2 + y_2^2}} \tag{4.4.8}$$

$$y_2' = 5y_1 - y_2 + \frac{y_2}{\sqrt{y_1^2 + y_2^2}} \ . \tag{4.4.9}$$

Once again the conditions (2.2.12) and (4.4.1) hold. Indeed, for the right hand side

$$f(y) = \begin{pmatrix} -y_1 - 5y_2 + \dfrac{y_1}{\sqrt{y_1^2 + y_2^2}} \\ \\ 5y_1 - y_2 + \dfrac{y_2}{\sqrt{y_1^2 + y_2^2}} \end{pmatrix} ,$$

we have

$$\begin{aligned} \langle f(y), y) \rangle &= (-y_1 - 5y_2 + \frac{y_1}{\sqrt{y_1^2 + y_2^2}})y_1 \\ &\quad +(5y_1 - y_2 + \frac{y_2}{\sqrt{y_1^2 + y_2^2}})y_2 \\ &= \sqrt{y_1^2 + y_2^2} - (y_1^2 + y_2^2) \\ &= ||y|| - ||y||^2 \\ &\leq \frac{1}{2}(1 + ||y||^2) - ||y||^2 \\ &= \frac{1}{2} - \frac{1}{2}||y||^2 \ , \end{aligned}$$

i.e., $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$ in (2.2.12). Furthermore,

$$\begin{aligned}
||f(y)||_2^2 &= (-y_1 - 5y_2 + \frac{y_1}{\sqrt{y_1^2 + y_2^2}})^2 \\
&\quad +(5y_1 - y_2 + \frac{y_2}{\sqrt{y_1^2 + y_2^2}})^2 \\
&= 1 + y_1^2 + 10y_1y_2 + 25y_2^2 + 25y_1^2 - 10y_1y_2 + y_2^2 - \frac{y_1^2 + y_2^2}{\sqrt{y_1^2 + y_2^2}} \\
&= 1 + 26(y_1^2 + y_2^2) - \sqrt{y_1^2 + y_2^2} \\
&= 1 + 26||y||^2 - ||y|| \\
&\leq 1 + 26||y||^2
\end{aligned}$$

Hence (4.4.2) holds with $\gamma = 1$ and $c = 26$. Then the non-standard scheme considered in Theorem 4.4.3 is given by (4.4.7) where $q = \frac{c}{\beta} = 52$. We take $\Delta t = 0.1$ and $y(0) = (5, 5)$ or $y(0) = (0.1, 0)$. On Fig. 4.9 and Fig. 4.10 one can observe that the non-standard numerical solutions eventually belong to the absorbing set $\overline{B}(0, 1.4277... + \epsilon)$ given in Theorem 4.4.3. The ball with radius 1.55 is plotted on the figures by a dotted line. The numerical solution on Fig. 4.9 originates outside this ball and enters it after certain number of time steps, while the numerical solution on Fig. 4.10 originates inside the ball and does not leave it. We notice from Fig. 4.11 that the standard Euler method with the same step size and initial condition $y(0) = (0.1, 0)$ is not dissipative.

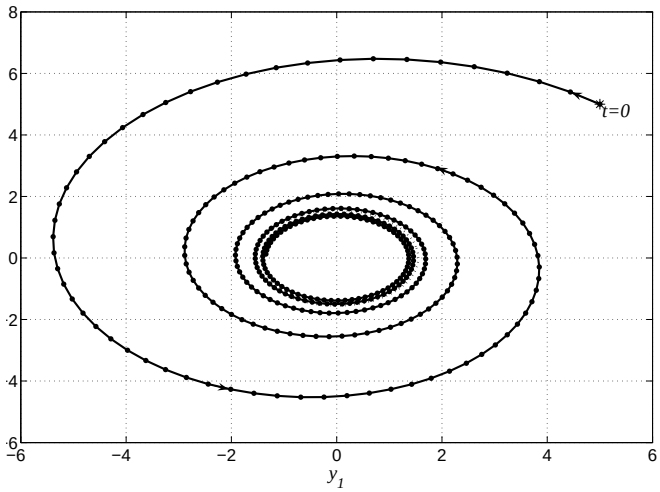

Figure 4.9: Dissipative non-standard scheme

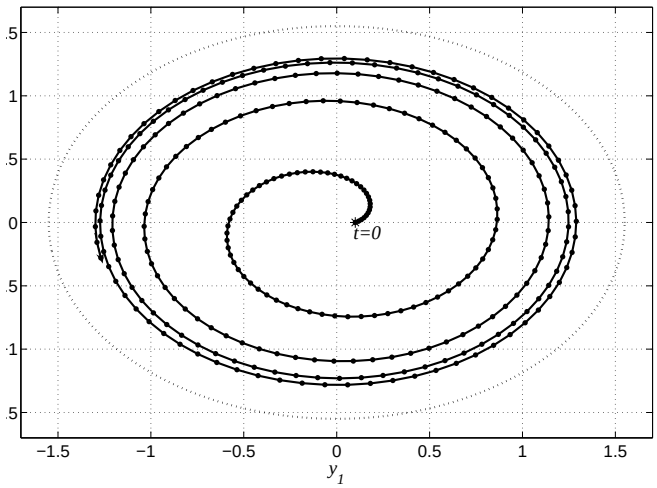

Figure 4.10: Another dissipative non-standard scheme

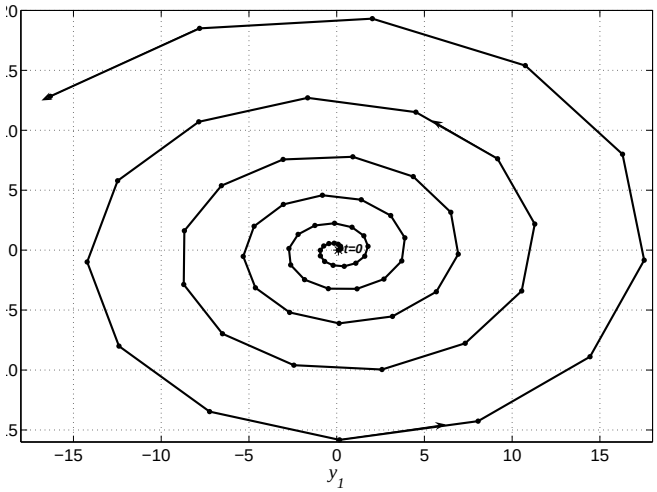

Figure 4.11: Nondissipative standard scheme

Remark 4.4.6. The absorbing sets in Examples 4.4.4 and 4.4.5 are determined by two different kinds of global attractors. In Example 4.4.4 the attractor is a hyperbolic fixed point, a case which can also be dealt with through the concept of linear stability. More precisely, equating both (4.4.5) and (4.4.6) to zero we arrive at the fixed point $(y_1, y_2) = (\frac{3}{13}, \frac{-2}{13})$. The Jacobian matrix of the system (4.4.5)-(4.4.6) is

$$J(y_1, y_2) = \begin{bmatrix} -1 & 5 \\ -5 & -1 \end{bmatrix}. \tag{4.4.10}$$

The eigenvalues of $J(\frac{3}{13}, \frac{-2}{13})$ are $\lambda = -1 \pm j5$ with $Re\lambda = -1, \quad \forall \lambda \in \sigma(J)$. Since $Re\lambda < 0$, we have a linearly stable fixed-point by Theorem 2.2.17.

However, linear stability does not yield results for Example 4.4.5, since the system does not have fixed points. In fact, it can be shown that the unit circle is a *global attractor* for this system. Notice that (see [41]) a set A is said to be an *attractor* if it is compact and invariant and attracts a neighbourhood of itself. Furthermore, a compact invariant set A is a *global attractor* for the semigroup operator $S(t)$ if it is an attractor which attracts every bounded set in $\mathbb{R}^m$. Note also that the global attractor of a dynamical system is unique if it exists. The terminology *local attractor* is sometimes used for attractors which are not global attractors. ∎

4.5 Energy Preserving Discrete Schemes

So far, our study has been concerned with systems (2.2.1) having only hyperbolic fixed-points. Non-standard schemes for such systems were designed by using mainly first part of Definition 4.2.3 on renormalization of the denominator Δt of the discrete derivative.

In this section, we consider the specific system

$$\begin{cases} \frac{dy_1}{dx} = y_2 \\ \\ \frac{dy_2}{dx} = -r(y_1), \end{cases} \tag{4.5.1}$$

where it is assumed that $(0,0)$ is the only fixed-point and that the smooth function $r : \mathbb{R} \to \mathbb{R}$ satisfies $r(0) = 0$ and $r'(0) = 1$.

The eigenvalues of the corresponding Jacobian matrix

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{4.5.2}$$

are $\lambda_{1,2} = \pm i$ and this shows that the fixed-point $\tilde{y} = (0,0)$ is non-hyperbolic. Consequently, the analysis of the previous sections does not apply. Nevertheless, by using the change of dependent variable

$$\begin{cases} y_1 = u \\ y_2 = u' \equiv \frac{du}{dx} \end{cases} \tag{4.5.3}$$

the system (4.5.1) is equivalent to the scalar equation

$$\frac{d^2u}{dx^2} + r(u) = 0, \tag{4.5.4}$$

which is the simplest model of Hamiltonian systems that occur in classical mechanics. Equation (4.5.4) is indeed equivalent to

$$H(u'(x), x) = \frac{1}{2}\left[(u')^2 + K(u)\right] = \text{constant} \tag{4.5.5}$$

where

$$K(u) = \int r(u)du. \tag{4.5.6}$$

Physically, H represents the sum of kinetic energy and potential energy of the mechanical system and (4.5.5) is the statement of conservation of

energy ([41], p 200). Consequently, (4.5.5) is one of the more important features of the system (4.5.1). Our aim is to derive finite difference methods which are stable with respect to the principle of conversation of energy. We will see that the approximation in a non-local way of nonlinear terms plays an essential role in achieving this aim.

Equation (4.5.4) is coupled with initial conditions

$$u(0) = u_0, \text{ and } u' = v_0. \tag{4.5.7}$$

Let u be a solution of (4.5.4) or (4.5.5). Fix a point x^* that can be written in the form $x^* = m\Delta x = x_m$ for different values of $m \in \mathbb{Z}$ and of the space step size Δx. Let γ be a real-valued function on $\mathbb{R}^3$ that meets the consistency condition

$$\lim_{\Delta x \to 0,\ m\Delta x = x^*} \gamma(u(x_{m-1}), u(x_m), u(x_{m+1})) = r(u(x^*)) \tag{4.5.8}$$

as well as the symmetry property

$$\gamma(u_{m-1,}u_m, u_{m+1}) = \gamma(u_{m+1}, u_m, u_{m-1}). \tag{4.5.9}$$

The notations used here are self explanatory: u_m is an approximation of the solution u at the grid point x_m.

Theorem 4.5.1. *Let ψ be a function satisfying (4.2.7). The non-standard finite difference scheme*

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{(\psi(\Delta x))^2} + \gamma(u_{m-1,}u_m, u_{m+1}) = 0, \tag{4.5.10}$$

for (4.5.4) is equivalent to the discrete principle of conservation of energy

$$\frac{1}{2}\left(\frac{u_{m+1} - u_m}{\psi(\Delta x)}\right)^2 + K_{\Delta x}(u_m) = \frac{1}{2}\left(\frac{u_m - u_{m-1}}{\psi(\Delta x)}\right)^2 + K_{\Delta x}(u_{m-1}), \tag{4.5.11}$$

where the discrete potential energy is given by

$$K_{\Delta x}(u_m) = \begin{cases} 0 & if \quad m = 0 \\ \sum_{i=1}^{m} \frac{(u_{i+1}-u_{i-1})\gamma(u_{i-1},u_i,u_{i+1})}{2} & if \quad m > 0 \\ \sum_{i=1}^{|m|} \frac{(u_{m-1+i}-u_{m+1+i})\gamma(u_{m+1+i},u_{m+i},u_{m-1+i})}{2} & if \quad m < 0. \end{cases} \tag{4.5.12}$$

Proof. A discrete principle of conversation of energy has the form

$$V_{\Delta x}(u_m) = V_{\Delta x}(u_{m-1}), \quad \forall m \geq 1 \tag{4.5.13}$$

with the discrete energy

$$V_{\Delta x}(u_m) = \frac{1}{2}\left(\frac{u_{m+1}-u_m}{\psi(\Delta x)}\right)^2 + K_{\Delta x}(u_m) \tag{4.5.14}$$

and $K_{\Delta x}(u_m)$ is given by (4.5.12). Expansion and simple manipulation of (4.5.11) yields the following equivalent relation

$$\begin{aligned} \frac{u_{m+1}^2 - u_{m-1}^2 - 2u_m(u_{m+1}-u_{m-1})}{\psi(\Delta x)} + 2(K_{\Delta x}(u_m) - K_{\Delta x}(u_{m-1})) &= 0 \\ (u_{m+1}-u_{m-1})\frac{u_{m+1}-2u_m+u_{m-1}}{\psi(\Delta x)} + 2(K_{\Delta x}(u_m) - K_{\Delta x}(u_{m-1})) &= 0 \\ \frac{u_{m+1}-2u_m+u_{m-1}}{(\psi(\Delta x))^2} + 2\frac{K_{\Delta x}(u_m) - K_{\Delta x}(u_{m-1})}{u_{m+1}-u_{m-1}} &= 0. \end{aligned} \tag{4.5.15}$$

Identification of (4.5.10) with (4.5.15) reduce to the expression

$$\frac{K_{\Delta x}(u_m) - K_{\Delta x}(u_{m-1})}{u_{m+1}-u_{m-1}} + \gamma(u_{m-1,}u_m, u_{m+1}) = 0 \tag{4.5.16}$$

which yields

$$K_{\Delta x}(u_m) = K_{\Delta x}(u_{m-1}) + (u_{m+1}-u_{m-1})\gamma(u_{m-1,}u_m, u_{m+1}) = 0. \tag{4.5.17}$$

By induction on m, with the initial-value $K_{\Delta x}(u_0) = 0$, we have

$$K_{\Delta x}(u_m) := \sum_{i=1}^{m}(u_{i+1} - u_{i-1})\gamma(u_{i-1,}u_i, u_{i+1}) = 0. \tag{4.5.18}$$

Thus, (4.5.10) is equivalent to the discrete law of conversation of energy (4.5.11). □

Remark 4.5.2. For $m > 1$ the discrete law of conservation of energy to which (4.5.10) is equivalent reads as:

$$\frac{1}{2}\left[\left(\frac{u_{m+1} - u_m}{\psi(\Delta x)}\right)^2 + \sum_{i=1}^{m}(u_{i+1} - u_{i-1})\gamma(u_{i-1,}u_i, u_{i+1})\right] =$$

$$\frac{1}{2}\left[\left(\frac{u_{m+1} - u_m}{\psi(\Delta x)}\right)^2 + \sum_{i=1}^{m}(u_{i+1} - u_{i-1})\gamma(u_{i+1,}u_i, u_{i-1})\right]. \tag{4.5.19}$$

Notice that the scheme (4.5.10) is equally equivalent to the non-standard finite difference scheme

$$\begin{cases} \frac{y_{1,m+1} - y_{1,m}}{\psi(\Delta x)} = y_{2,m+1} \\ \\ \frac{y_{2,m+1} - y_{2,m}}{\psi(\Delta x)} = -r(y_{1,m}) \end{cases} \tag{4.5.20}$$

which is closely related to (4.5.1). ∎

Remark 4.5.3. From (4.5.4) and (4.5.15) the natural choice of $\gamma(\cdot, \cdot, \cdot)$ is given in terms of (4.5.6) by the mean-value theorem:

$$\gamma(u_{m-1,}u_m, u_{m+1}) \equiv 2\frac{K_{\Delta x}(u_m) - K_{\Delta x}(u_{m-1})}{u_{m+1} - u_{m-1}} \tag{4.5.21}$$

$$= \frac{K(u_{m+1}) - K(u_{m-1})}{u_{m+1} - u_{m-1}}. \tag{4.5.22}$$

This is the approach proposed in Anguelov and Lubuma [7]. More precisely, if $r(u)$ has the form $r(u) = ug(u^2)$, these authors worked simply with $G = \int g(s)ds$ instead of K. In this case, the above leads to the scheme

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{(\psi(\Delta x))^2} + u_m \frac{G(u_m\, u_{m+1}) - G(u_m\, u_{m-1})}{u_m\, u_{m+1} - u_m\, u_{m-1}} = 0 \qquad (4.5.23)$$

equivalent to its energy preserving form

$$\frac{1}{2}\left[\left(\frac{u_{m+1} - u_m}{\psi(\Delta x)}\right)^2 + G(u_m u_{m+1})\right] = \frac{1}{2}\left[\left(\frac{u_m - u_{m-1}}{\psi(\Delta x)}\right)^2 + G(u_m u_{m-1}))\right].$$

Other non-standard finite difference schemes for conservative oscillators are investigated in [26]. ∎

Remark 4.5.4. The schemes (4.5.10) and (4.5.23) are non-standard in the sense of both Mickens rules in Definition 4.2.3. Firstly the exact scheme (see Table 4.1)

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{4\sin\frac{\Delta x}{2}} + u_m = 0 \qquad (4.5.24)$$

of the simple harmonic oscillator

$$\frac{d^2u}{dx^2} + u = 0, \qquad (4.5.25)$$

motivates the need to renormalise the denominator of the discrete derivatives in the schemes (4.5.10),(4.5.22) and (4.5.23). Secondly, the nonlinear terms that arise in $r(u)$ are approximated in a non-local way. For example, if $r(u) = u^3$, we have by (4.5.22) and (4.5.23) the respective approximations

$$r(u(x^*)) \approx \frac{(u_{m+1} + u_{m-1})(u_{m+1}^2 + u_{m-1}^2)}{4}$$

$$r(u(x^*)) \approx \frac{u_m^2(u_{m+1} + u_{m-1})}{2}.$$

∎

Remark 4.5.5. An advantage of the second choice of γ is that the three arguments $u_{m-1,}\, u_m$ and u_{m+1} appear explicitly in the analogue term of (4.5.23) contrary to (4.5.22). Furthermore, for (4.5.25), the second choice with $g(u^2) = 1$, yields the exact scheme (4.5.24). But the first choice yields the scheme

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{(\psi(\Delta x))^2} + \frac{u_{m-1} + u_{m+1}}{2} = 0. \tag{4.5.26}$$

■

Remark 4.5.6. For the implementation of (4.5.10) the initial values $u(0) = u_0$ and $u'(0) = v_0$ are usually given as indicated in (4.5.7). However, a value for u_1 is needed in order to start the scheme. In analogy with a classical procedure, we utilize the approximation

$$u'(0) = \frac{u_1 - u_{-1}}{\psi(\Delta x)}. \tag{4.5.27}$$

In (4.5.10), put $m = 0$ and replace u_{-1} with the expression obtained from (4.5.27) and this gives the missing starting value whenever the structure of γ makes (4.5.10) an explicit scheme. When the scheme (4.5.10) is not explicit, one could use the less accurate approximation

$$u'(0) = \frac{u_1 - u_0}{\psi(\Delta x)}. \tag{4.5.28}$$

■

Example 4.5.7.

As an numerical example, we consider the Duffing conservative oscillator

$$\frac{d^2u}{dx^2} + 25u(1 + 15u^2) = 0,\ u(0) = 0,\ u'(0) = 1. \tag{4.5.29}$$

With $\Delta x = 0.1$, Fig 4.12 illustrates both the stability of the non-standard scheme (4.5.23) where $\psi(\Delta x) = \frac{2}{5}\sin\frac{5\Delta x}{2}$ and the instability of the standard scheme

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{(\Delta x)^2} + 25u_m(1 + 15u_m^2) = 0 \tag{4.5.30}$$

with respect to the principle of conservation of energy. Other examples can be found in [15] where the non-standard scheme discussed in this section have been extended to more complex problems, namely, vibro-impact mechanical systems.

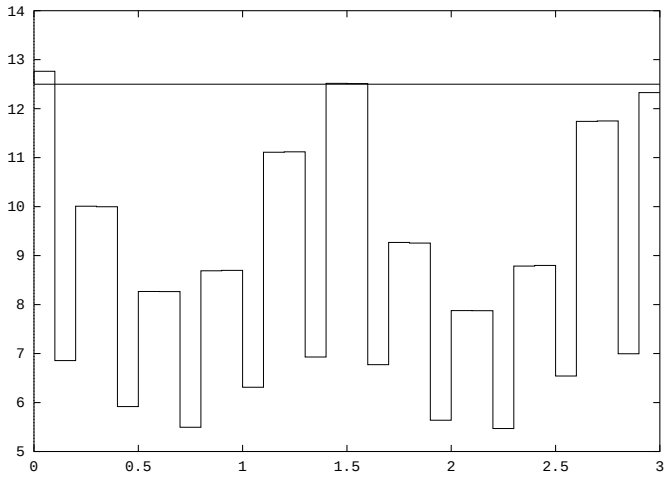

Figure 4.12: Discrete energy of the Duffing equation by standard (piecewise constant) and non-standard (constant) finite difference schemes

Chapter 5

Non-standard Finite Difference Schemes for Reaction-Diffusion Equations

5.1 Introduction

This chapter is a dedicated analysis of the author's results in [6]. We investigate the impact of the analysis of the previous chapters on the numerical solution of partial differential equations. We will specifically deal with the one dimensional reaction-diffusion equations the solutions u of which enjoy a positivity and boundedness property:

$$0 \leq u \leq 1. \tag{5.1.1}$$

A typical example is the Fisher equation for which (5.1.1) is proved in Sections 5.4. In Section 5.3, we design non-standard finite difference schemes which are elementary stable in the limit case of space independent variable and which are stable with respect to the principle of conservation of energy in the stationary case. Furthermore, we show in Section 5.4 that our schemes replicate the property (5.1.1) under a certain functional relation between the time and space step sizes.

As an alternative approach, we approximate in Section 5.5 the space variable by the spectral method, while the time variable is approximated via the non-standard finite difference scheme. This results in

what we call coupled spectral and non-standard methods. Numerical tests that show the reliability of these coupled schemes are provided.

5.2 The Fisher Equation

A classic simplest case of a non-linear reaction-diffusion equation is the Fisher equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \lambda u(1-u), \quad \lambda > 0. \tag{5.2.1}$$

The material presented in this section is based on the books [23] and [33]. Equation (5.2.1) is also referred to as the Fisher-Kolmogoroff equation. It is also the natural extension of the logistic growth model discussed in Section 4.3. It was suggested by Fisher (1937) as a deterministic version of a stochastic model for the spatial propagation of a mutant-gene in a population. Equation (5.2.1) can also be used in the analysis of travelling waves in chemical reactions.

It is indeed well known that when reaction kinetics and diffusion are coupled, travelling waves of chemical concentration exist and can effect a biochemical change, very much faster than straight diffusional processes governed by (5.2.1) without the term $\lambda u(1-u)$. Thus, we look for (5.2.1) travelling wave solutions. This means solutions of the form

$$u(x,t) = U(z), \quad z = x - ct, \tag{5.2.2}$$

where, with c being the constant speed of the wave moving in the positive $x-$direction, it is assumed that $U(z)$ is non-negative and bounded for all $z \in \mathbb{R}$.

The space independent Fisher equation (5.2.1) has fixed-points $u = 0$ and $u = 1$, which are unstable and asymptotically stable respectively. This suggests that we look for a travelling wave solution which satisfies the boundedness and positivity condition (5.1.1) as well as the conditions

$$\lim_{z\to\infty} U(z) = 0, \qquad \lim_{z\to-\infty} U(z) = 1. \tag{5.2.3}$$

Substituting this travelling wavefront (5.2.2) into (5.2.1) yields a second order ordinary differential equation for $U(z)$:

$$U'' + cU' + U(1-U) = 0, \tag{5.2.4}$$

where the range of $c \geq 0$ is to be determined. Since this equation cannot be solved in closed form, we reduce it to a pair of first order equations by defining $V = U'$ leading to the autonomous system

$$U' = V \tag{5.2.5}$$
$$V' = -cV - U(1-U). \tag{5.2.6}$$

The Jacobian matrix of the system (5.2.5) - (5.2.6) is

$$J(U,V) = \begin{bmatrix} 0 & 1 \\ 2U-1 & -c \end{bmatrix}. \tag{5.2.7}$$

The fixed-points of the system (5.2.5) - (5.2.6) are $(0,0)$ and $(1,0)$. The eigenvalues of $J(1,0)$ are

$$\lambda_{\pm} = \frac{-c \pm \sqrt{c^2+4}}{2} \tag{5.2.8}$$

and those of $J(0,0)$ are

$$\lambda_{\pm} = \frac{-c \pm \sqrt{c^2-4}}{2}, \tag{5.2.9}$$

showing that the two fixed-points are hyperbolic. Thus, Hartman-Grobman theorem (2.2.16) applies. We can therefore conclude that the fixed-point $(1,0)$ is a saddle point for any c, while $(0,0)$ is an asymptotically stable node for $c \geq 2$ and a stable spiral if $c < 2$, (see [46]) .

The case $c \geq 2$ is of interest as it follows by continuity arguments or by heuristic reasoning from the phase plane (U,V), that there exists a trajectory from the fixed-point $(1,0)$ to the fixed-point $(0,0)$ lying entirely in the quadrant $U \geq 0,\ V \leq 0$ with $0 \leq U \leq 1$ and all $c \geq 2$, (see Figs 5.1 - 5.2).

In summary, we have the following result.

Theorem 5.2.1. *For each $c \geq 2$ there exists a unique travelling wave solution $u(x,t) = U(x - ct)$ to the Fisher equation (5.2.1) with the property that in the wave variable $z = x - ct$, $U(z)$ is monotonically decreasing and satisfies (5.2.3).*

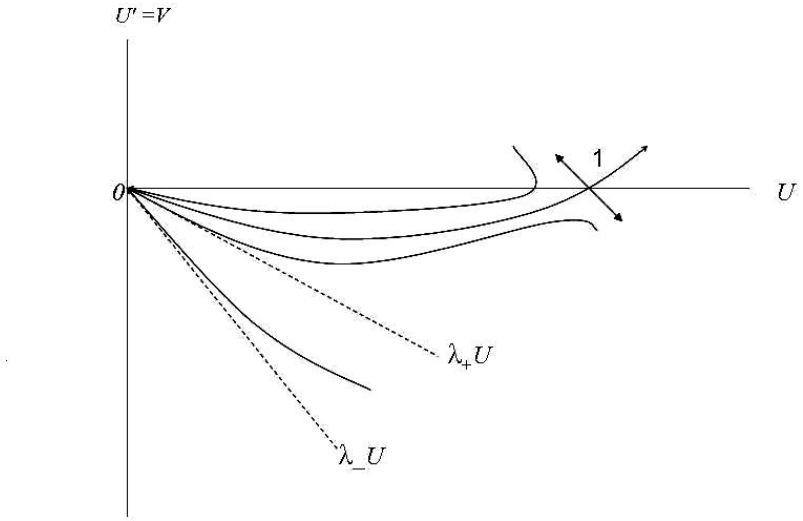

Figure 5.1: Phase plane trajectories for (5.2.5) - (5.2.6), $c \geq 2$.

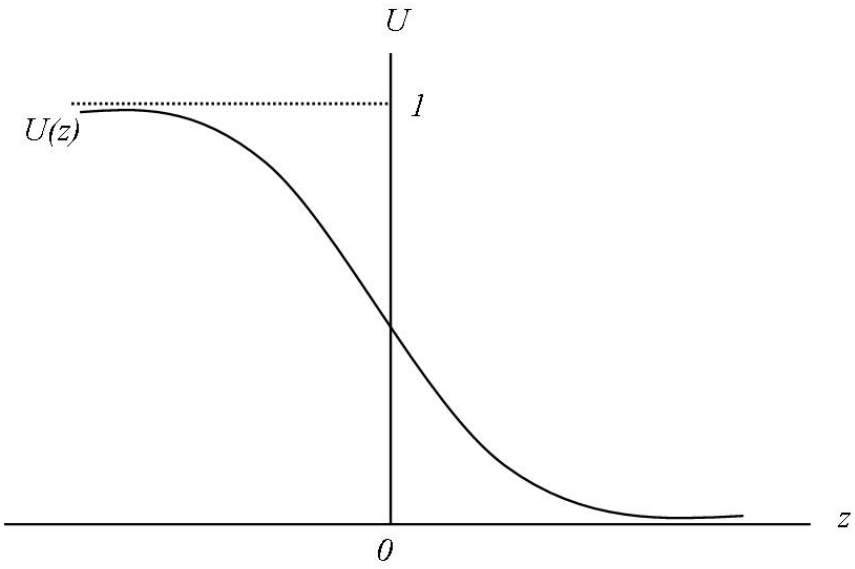

Figure 5.2: Travelling wave solution for the Fisher equation, $c \geq 2$.

5.3 Theta Methods for Reaction-Diffusion Equations

The Fisher equation considered in the previous section motivates that we now study the general one-dimensional reaction-diffusion equation

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2} + r(u), \\ u(x,0) = g(x) \end{cases} \tag{5.3.1}$$

and we assume that there exists a unique solution satisfying

$$0 \leq g \leq 1 \implies 0 \leq u \leq 1. \tag{5.3.2}$$

Equation (5.3.1) is used extensively in many areas of engineering and applied sciences to model a system on which reaction processes $r(u)$ lead to the diffusion in time of the quantity u, (see for instance [19] and [33]). We are interested in numerical schemes that produce reliable approximations u_m^k of the solution u at the time $t_k = k\Delta t$ and the space grid point $x_m = m\Delta x$. To achieve this, we use non-standard finite difference schemes following the methodology of sub-equations in [7] and [26] to address partial differential equations.

More precisely, we design non-standard schemes for the space independent equation on the one hand and for the stationary equation on the other hand. After that, we assemble them in suitable schemes for the reaction- diffusion equation.

Energy-preserving schemes for the stationary case of (5.3.1) were discussed in Section 4.5. Thus, for the equation

$$\frac{d^2 u}{dx^2} + r(u) = 0, \tag{5.3.3}$$

we have in view of (4.5.10) the non-standard scheme

$$\frac{u_{m+1} - 2u_m + u_{m-1}}{(\psi(\Delta x))^2} + \gamma(u_{m-1}, u_m, u_{m+1}) = 0. \tag{5.3.4}$$

The space independent equation of (5.3.1) is

$$\frac{du}{dt} = r(u), \quad u(0) = u_0 \tag{5.3.5}$$

which is the scalar case of (2.2.1). We approximate it using the non-standard one-stage (4.3.7) and two-stage (4.3.8) theta methods, which in this case read as follows:

$$\frac{u^{k+1} - u^k}{\frac{\phi(q\Delta t)}{q}} = r[\theta u^{k+1} + (1-\theta)u^k], \tag{5.3.6}$$

$$\frac{u^{k+1} - u^k}{\frac{\phi(q\Delta t)}{q}} = \theta r(u^{k+1}) + (1-\theta)r(u^k). \tag{5.3.7}$$

By combining (5.3.4) and (5.3.6)-(5.3.7), we arrive at the following non-standard finite difference methods for (5.3.1):

$$\begin{aligned}\frac{u_m^{k+1} - u_m^k}{\frac{\phi(q\Delta t)}{q}} &= \theta\frac{u_{m+1}^{k+1} - 2u_m^{k+1} + u_{m-1}^{k+1}}{(\psi(\Delta x))^2} + (1-\theta)\frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\psi(\Delta x))^2} + \\ &\gamma\left[\theta\ u_{m-1}^{k+1} + (1-\theta)\ u_{m-1}^k, \theta\ u_m^{k+1} + (1-\theta)\ u_m^k, \theta\ u_{m+1}^{k+1} + (1-\theta)\ u_{m+1}^k\right]\end{aligned} \tag{5.3.8}$$

$$\begin{aligned}\frac{u_m^{k+1} - u_m^k}{\frac{\phi(q\Delta t)}{q}} &= \theta\frac{u_{m+1}^{k+1} - 2u_m^{k+1} + u_{m-1}^{k+1}}{(\psi(\Delta x))^2} + (1-\theta)\frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\psi(\Delta x))^2} \\ &+\theta\gamma(u_{m-1}^{k+1}, u_m^{k+1}, u_{m+1}^{k+1}) + (1-\theta)\gamma(u_{m-1}^k, u_m^k, u_{m+1}^k).\end{aligned} \tag{5.3.9}$$

Notice that the denominator functions ϕ and the number q that captures the dynamics of the system are chosen in the manner discussed in Sections 4.3 and 4.4, namely (4.3.2) or (4.3.15) or Theorem 4.4.3 together with (4.3.4) and (4.3.6). Equally the denominator function ψ and the appropriate forms of the function γ are discussed in Section 4.5.

By construction and assuming that the function $r(u)$ satisfies the conditions of the relevant theorems in Chapter 4, we have the following result.

Theorem 5.3.1. *The non-standard finite difference schemes (5.3.8) and (5.3.9) are qualitatively stable with respect to the principle of conservation of energy in the limit case of the stationary equation. Furthermore, these non-standard schemes are elementary stable in the limit case of space independent variable. In this case, they are also qualitatively stable with respect to the dissipativity property for* $\theta = 0$ *or* $\theta \in (\frac{1}{2}, 1]$ *whenever the continuous system is in the setting of Theorem 4.4.1 and Theorem 4.4.3.*

5.4 Explicit Scheme

We would like to design schemes related in one way or another to (5.3.8)-(5.3.9) which are stable with respect to the positivity and bounded property (5.3.2). That is,

$$0 \leq u_m^0 \leq 1 \Rightarrow 0 \leq u_m^k \leq 1. \tag{5.4.1}$$

We consider the explicit case (i.e. $\theta = 0$) for which (5.3.8) and (5.3.9) reduce to

$$\frac{u_m^{k+1} - u_m^k}{\frac{\phi(q\Delta t)}{q}} = \frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\psi(\Delta x))^2} + \gamma(u_{m-1}^k, u_m^k, u_{m+1}^k). \tag{5.4.2}$$

It will be necessary to modify (5.4.2) into a new formula resulting from a somewhat convenient form. A proper choice of the function γ in (5.4.2) and Theorem 4.5.1 is essential in what follows. To this end, we assume that the function γ may be represented as

$$\gamma(u_{m-1}, u_m, u_{m+1}) = (1 - u_m)\Gamma(u_{m-1}, u_m, u_{m+1}) \tag{5.4.3}$$

for some function satisfying $\Gamma(u_{m-1}, u_m, u_{m+1}) \geq 0$ for nonnegative arguments. We also assume that the symmetry property

$$\Gamma(u_{m-1,}u_m, u_{m+1}) = \Gamma(u_{m+1,}u_m, u_{m-1})$$

holds as in (4.5.9) and that the scheme

$$\frac{u^{k+1}-u^k}{\frac{\phi(q\Delta t)}{q}} = (1-u^{k+1})\Gamma(u^k,u^k,u^k) \tag{5.4.4}$$

for (5.3.5) is elementary stable. We can in place of (5.3.4) consider

$$\frac{u_{m+1}-2u_m+u_{m-1}}{(\psi(\Delta x))^2} + (1-u_m)\Gamma(u_{m-1},u_m,u_{m+1}) = 0. \tag{5.4.5}$$

One possible combination of (5.4.4) and (5.4.5) in the spirit of (5.3.9) is

$$\frac{u_m^{k+1}-u_m^k}{\frac{\phi(q\Delta t)}{q}} = \frac{u_{m+1}^k-2u_m^k+u_{m-1}^k}{(\psi(\Delta x))^2} + (1-u_m^{k+1})\Gamma(u_{m-1}^k,u_m^k,u_{m+1}^k). \tag{5.4.6}$$

In classical finite difference methods, the quantities Δt and Δx do not vary independently [32]. It is therefore not suprising to require a certain functional relation between Δt and Δx for the scheme (5.4.6). In view of our objective to have property (5.4.1), we impose the condition

$$\frac{\phi(q\Delta t)/q}{(\psi(\Delta x))^2} = \frac{1}{2}. \tag{5.4.7}$$

Solving (5.4.6) for u_m^{k+1} yields

$$u_m^{k+1} = \frac{\frac{1}{2}\left(u_{m-1}^k+u_{m+1}^k\right)+\phi(q\Delta t)/q\Gamma(u_{m-1}^k,u_m^k,u_{m+1}^k)}{1+\phi(q\Delta t)/q\Gamma(u_{m-1}^k,u_m^k,u_{m+1}^k)}. \tag{5.4.8}$$

If $0 \leq u_m^k \leq 1$, it follows from (5.4.8) and the property of Γ that $0 \leq u_m^{k+1} \leq 1$. In summary, we have thus shown the following result.

Theorem 5.4.1. *Under condition (5.4.7) the non-standard finite difference scheme (5.4.6) is stable with respect to boundedness and positivity property (5.3.2). Furthermore, this scheme is elementary stable in the limit case of the space independent variable and it is also stable with respect to conservation of energy in the stationary case.*

Remark 5.4.2. An essential feature of the scheme (5.4.6) proposed here is that it replicates property (5.3.2) under the simple relation (5.4.7) between step sizes. Other schemes having the property (5.4.1) may be obtained but at the cost of more complicated functional relations between step sizes. For example, in the particular case of the Fisher equation (5.2.1), which satisfies property (5.3.2), an alternative scheme preserving this property is obtained in [27] but at the cost of the more complicated restriction between step sizes, namely:

$$\frac{\phi(q\Delta t)/q}{(\psi(\Delta x))^2} = \frac{1}{3}\left(1 - \frac{q(\psi(\Delta x))^2}{3}\right)^{-1}. \tag{5.4.9}$$

Furthermore, the relation (5.4.7) is a typical condition of Lax-Richtmyer stability of finite difference schemes for linear diffusion equation. Let us clarify this fact with the scheme

$$\frac{u_m^{k+1} - u_m^k}{\phi(q\Delta t)/q} = \frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\psi(\Delta x))^2} + u_m^k \tag{5.4.10}$$

applied to the linear problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u.$$

In the setting of (5.3.9), the scheme (5.4.10) corresponds to

$$\gamma(u_{m-1}^k, u_m^k, u_{m+1}^k) = u_m^k.$$

We use the Fourier series method [32]. The amplification factor for the scheme (5.4.10) is

$$\rho(\xi) = 1 - 4\nu \sin^2 \frac{\xi}{2}\Delta x + \phi(q\Delta t)/q, \forall \xi \in \mathbb{R},$$

where $\nu = \dfrac{\phi(q\Delta t)/q}{(\psi(\Delta x))^2}$. The scheme (5.4.10) is stable in the sense of Lax-Richtmyer whenever the von Neumann condition $|\rho(\xi)| \leq 1$ is satisfied. This condition is met if $\nu \leq \dfrac{1}{2} + \dfrac{\phi(q\Delta t)/q}{2}$. ∎

Remark 5.4.3. The strategy of writing the function $\gamma(\cdot,\cdot,\cdot)$ in the form (5.4.3) and of approximating it in the nonlocal way shown in (5.4.4) is being used extensively in the literature, specifically in mathematical biology, when the discrete solution is required to replicate the positivity property of the exact solution. (See for instance [7], [13], [14], [29], [30]).
■

To illustrate the analysis of the previous sections, we consider again the Fisher equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + 25u(1-u),\ u(x,0) = 0.5 + 0.5\sin 2x, \tag{5.4.11}$$

for which the solution satisfies (5.3.2). We apply various non-standard methods of the form

$$\frac{u_m^{k+1} - u_m^k}{\phi(\Delta t)} = \frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\Delta x)^2} + 25(1 - u_m^{k+1})\frac{u_{m-1}^k + u_m^k + u_{m+1}^k}{3}. \tag{5.4.12}$$

With $\phi(\Delta t) = \frac{1-e^{25\Delta t}}{25}$, the solution of the scheme (5.4.12), which corresponds to (5.4.6), is displayed in Fig.5.3, for $\Delta t = 0.061$ and $\Delta x = 0.25$.

We may choose the denominator $\phi(\Delta t) = \frac{e^{25\Delta t}-1}{25}$ which provides the exact scheme for the logistic equation (Table 4.1)

$$\frac{du}{dt} = 25u(1-u). \tag{5.4.13}$$

The solution of the resulting scheme (5.4.12) is displayed in Fig.5.4, for $\Delta t = 0.0231$ and $\Delta x = 0.25$.

The discrete scheme

$$\frac{u^{k+1} - u^k}{\Delta t} = 25u^k(1 - u^{k+1}), \tag{5.4.14}$$

for (5.4.13) is elementary stable and so we can take $\phi(\Delta t) = \Delta t$ in (5.4.12). The resulting solution is visualised in Fig.5.6 for $\Delta t = 0.031$ and $\Delta x = 0.25$.

All the results are compared with the standard scheme

$$\frac{u_m^{k+1} - u_m^k}{\Delta t} = \frac{u_{m+1}^k - 2u_m^k + u_{m-1}^k}{(\Delta x)^2} + 25u^k(1 - u_m^k), \tag{5.4.15}$$

whose solution is visualised in Fig.5.5 for $\Delta t = 0.0231$ and $\Delta x = 0.25$.

The three figures corresponding to non-standard schemes confirm elementary stability with respect to the boundedness and positivity property. On the contrary the standard scheme shown in Fig. 5.6 fails to replicate any one of these properties.

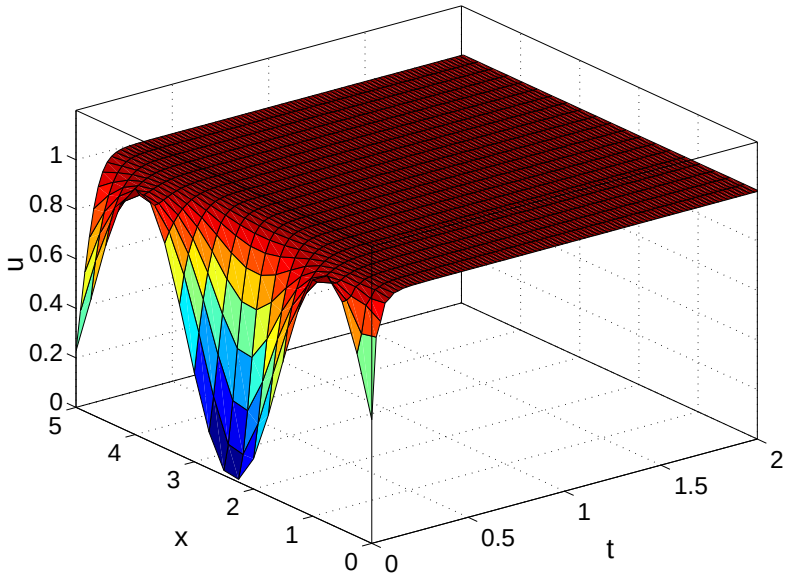

Figure 5.3: Non-standard scheme not related to exact scheme.

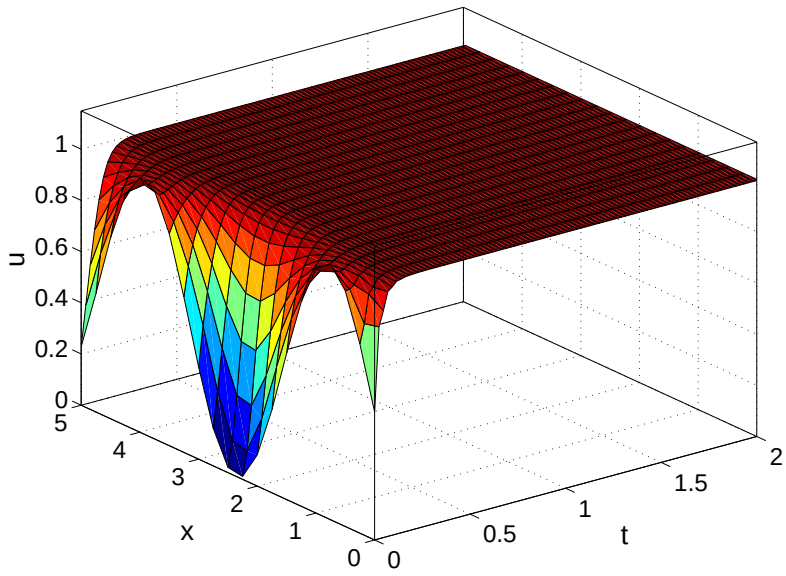

Figure 5.4: Non-standard scheme related to exact scheme.

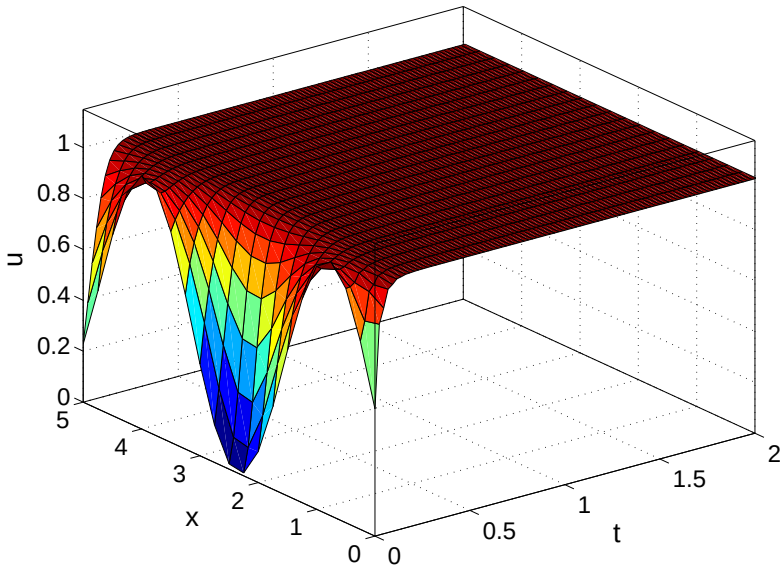

Figure 5.5: Non-standard scheme with $\phi(\Delta t) = \Delta t$.

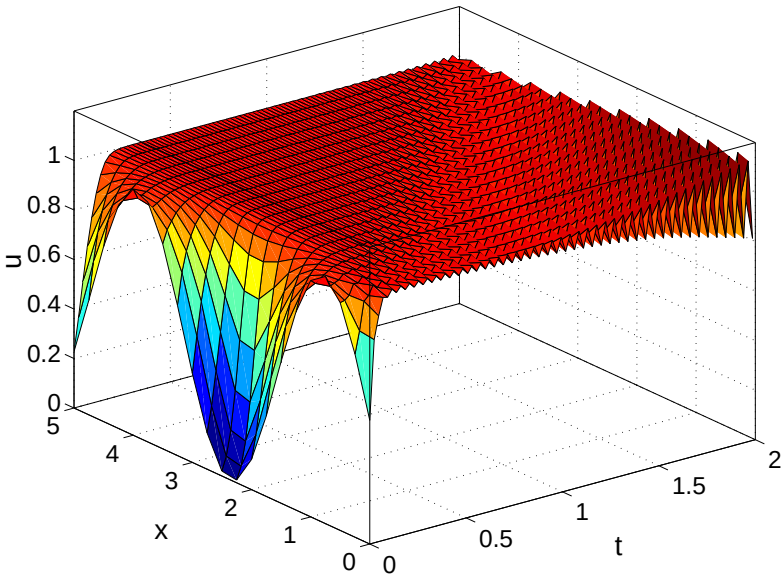

Figure 5.6: Standard scheme.

5.5 Coupled Spectral and Non-standard Methods

So far, the approximations in the space variable x were obtained by the finite difference method. In this section, we use the spectral method. We consider the reaction-diffusion problem

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + u = R(u), \quad \text{on} \quad (0, 2\pi) \times (0, T) \tag{5.5.1}$$

$$u(x, 0) = u_0(x) \quad \text{for} \quad x \in (0, 2\pi) \tag{5.5.2}$$

$$u(0, t) = u(2\pi, t) \quad \text{for } t \in (0, T), \tag{5.5.3}$$

where the function $R(\bullet)$ as well as the function u_0 in the Lebesgue space $L^2(0, 2\pi)$ with inner product $\langle \bullet, \bullet \rangle$ are given. We assume of course that problem (5.5.1)-(5.5.3) has a unique solution. In view of the numerical scheme presented below, we assume that

$$R(0) = R'(0) = 0. \tag{5.5.4}$$

Thus, (5.5.1) corresponds, in the setting of (5.3.1), to the case when $r(u) := R(u) - u$ is linearized about $u = 0$ by $-u$.

With each integer $m \in \mathbb{N}$, we associate the Fourier-Garlekin spectral approximation of the solution u, which is a semi-discrete solution given by (see [10])

$$u_m(x, t) = \sum_{k=-m}^{m} \alpha_k(t) w_k(x) \quad \text{for } (x, t) \in (0, 2\pi) \times [0, T], \tag{5.5.5}$$

where for $k \in \mathbb{Z}$,

$$w_k(x) := \frac{1}{\sqrt{2\pi}} e^{ikx} \quad \text{for} \quad x \in [0, 2\pi]. \tag{5.5.6}$$

The function $u_m(x, t)$ in (5.5.5) does not satisfy (5.5.1) and (5.5.2). But this function is an approximation of the solution u in the sense that, for $|k| \leq m$, we have

$$\begin{cases} \left\langle \dfrac{\partial u_m}{\partial t} - \dfrac{\partial^2 u_m}{\partial x^2} + u_m, w_k \right\rangle = \langle R(u_m), w_k \rangle \\ \langle u_m(0), w_k \rangle = \langle u_0, w_k \rangle . \end{cases} \tag{5.5.7}$$

Thus, with

$$\lambda_k := k^2 + 1 \tag{5.5.8}$$

the vector function $U_m = [\alpha_{-m} \;\; \alpha_{-m+1} \cdots \;\; \alpha_m]^T$ of Fourier coefficients in (5.5.5) is the unique solution of the initial-value problem for the system of $2m+1$ ordinary differential equations in $2m+1$ unknowns α_k:

$$\frac{d\alpha_k}{dt} + \lambda_k \alpha_k = \langle R(u_m), w_k \rangle \quad \text{on } (0,T), \tag{5.5.9}$$

$$\alpha_k(0) = \langle u_0, w_k \rangle \,. \tag{5.5.10}$$

Remark 5.5.1. A motivation of the spectral approximations (5.5.5) and (5.5.9)-(5.5.10) is that, in many cases, the solution u of (5.5.1)-(5.5.3) admits in $L^2(0, 2\pi)$ the Fourier series expansion

$$u(t) \equiv u(\bullet, t) = \sum_{k \in \mathbb{Z}} \alpha_k(t) w_k(\bullet). \tag{5.5.11}$$

This is in particular true for the linear diffusion equation, i.e. R in (5.5.1) is a function of the independent variables x and t only but not of the dependent variable u (See, for example [36]). ∎

To obtain a full discretisation of u, we have to approximate (5.5.9)-(5.5.10). The main source of difficulty in (5.5.9) comes from its linearised part, which is a stiff system: from (5.5.8), $1 = \lambda_0 \ll \lambda_m$ for big values of m. The condition (5.5.4) on the reaction function $R(\bullet)$ guarantees that the null vector $\widetilde{U} = \widetilde{0} \in \mathbb{R}^{2m+1}$ is a hyperbolic fixed-point of the system (5.5.9). Consequently, by Hartman-Grobman theorem (Theorem 2.2.16), this system can be qualitatively studied from its linearisation about $\widetilde{U} = 0$ which is

$$\frac{d\alpha_k}{dt} + \lambda_k \alpha_k = 0, \quad |k| \leq m. \tag{5.5.12}$$

The approach used in [26] to approximate first-order nonlinear differential equations is based on this connection between (5.5.9) and

(5.5.12). We follow this approach to avoid the above mentioned difficulty by incorporating the stiffness feature of the system (5.5.9) in the numerical scheme. More precisely, by analogy with the exact scheme

$$\frac{\alpha_{k,n+1} - \alpha_{k,n}}{(1 - e^{-\lambda_k \Delta t})/\lambda_k} + \lambda_k \alpha_{k,n} = 0 \quad n = 0, 1, 2, ..., \tag{5.5.13}$$

of the linearised part of the system (5.5.9), we consider, for the nonlinear system (5.5.9)-(5.5.10), the non-standard forward Euler method

$$\frac{\alpha_{k,n+1} - \alpha_{k,n}}{(1 - e^{-\lambda_k \Delta t})/\lambda_k} + \lambda_k \alpha_{k,n} = \langle R(u_{m,n}), w_k \rangle_0 \quad n = 0, 1, 2, ..., \tag{5.5.14}$$

where $u_{m,0}$ is obtained from (5.5.10) by taking $t = 0$ in (5.5.5), i.e.

$$u_{m,0} = \sum_{|k| \leq m} \langle u_0, w_k \rangle \, w_k(x). \tag{5.5.15}$$

This then provides

$$u_{m,n}(x) = \sum_{k=-m}^{m} \alpha_{k,n} w_k(x), \tag{5.5.16}$$

as the spectral non-standard finite difference approximation of the solution u at the point (x, t^*) where $t^* = t_n = n\Delta t$ is fixed.

Issues pertaining to the consistency, the stability and the convergence, with rates of convergence, of this coupled spectral-non-standard finite difference methods can be analysed along the lines of [10] and [36]. We do not do this analysis here. Our interest is rather in testing this approach numerically. To this end, we consider (5.5.1)-(5.5.3) with $R(u) = u^2$ and $u_0(x) = x(2\pi - x)/\pi^2$. The result of the non-standard scheme (5.5.14)-(5.5.16) is visualized in Fig.5.7, for $m = 20$ and $\Delta t = 0.1$. This is to be compared with Fig.5.8, relative to the standard scheme (5.5.16) where the traditional denominator Δt is used in the discrete derivative in (5.5.14) for the specific values $m = 5$ and $\Delta t = 0.075$, which satisfy the stability condition $\lambda_m \Delta t \leq 2$. One observes, for instance, that the non-standard scheme is elementary stable and stable with respect to the monotonicity of solution in the limit case of space independent equation contrary to the standard scheme.

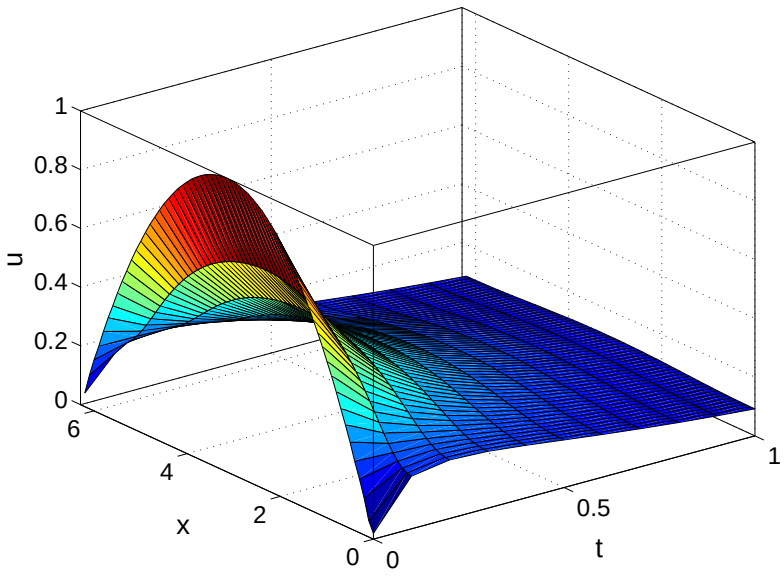

Figure 5.7: Spectral non-standard scheme based on the exact scheme.

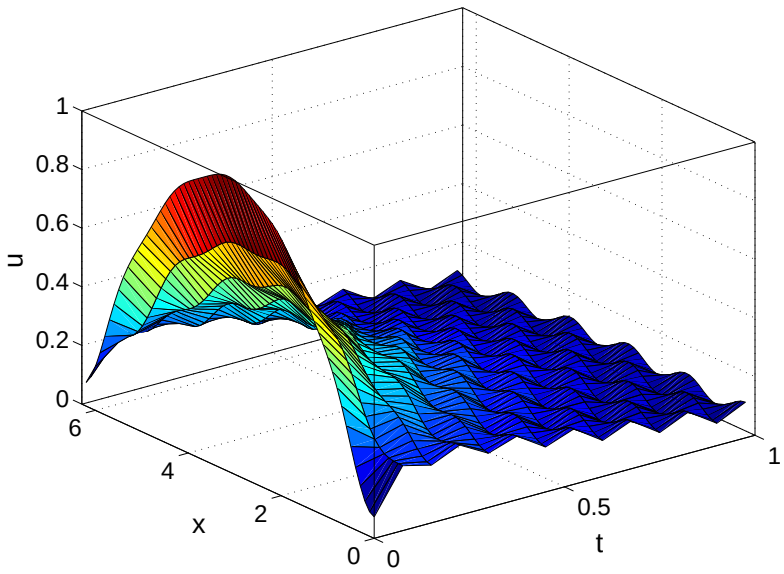

Figure 5.8: Spectral standard scheme.

Chapter 6

Conclusion

The non-standard finite difference approach was initiated more than two decades ago by Mickens. The monograph [26] constitutes a self-contained and comprehensive treatment of the non-standard finite difference method. Since the publication of this book, the non-standard approach has extensively been applied to differential models originating from problems in engineering, physics, biology, chemistry, etc. With the great potential that the non-standard finite difference schemes have been showing in replicating the essential properties of the exact solutions of the involved differential equations, we felt strongly about focusing on dynamical systems in this thesis. Indeed, dynamical systems have a wide range of important intrinsic properties, such as fixed-points and their stability, attracting sets, limit cycles, which ideally should be preserved by numerical schemes if they are to yield reliable simulations that provide qualitative information and useful insights on the exact solutions. In particular, the following facts constitute some of the specific motivations of this thesis, which show where it fits in the literature:

1. A sharper condition given in [14] for the elementary stability of the non-standard forward Euler method and a claim made therein that the condition avoids the location of the eigenvalues of the involved Jacobian matrices in some regions of the complex plane.

2. A follow up to the chapter [24] in order to investigate other types of dissipative properties of differential models, than the dissipativity of singular perturbed problems, which has a specific meaning in terms of the decay/variation of their solutions in layer regions;

3. A result on classical theta methods that restricts their dissipativity as discrete dynamical systems to the range $\theta \in (\frac{1}{2}, 1]$.(see, e.g., [41]).

4. The design in [27] of a non-standard finite difference scheme for the Fisher equation, which is stable with respect to the boundedness and positivity property of the solution under a certain functional relation between the time and space step sizes.

For this thesis to be relatively self-contained, we dedicated considerable time to overview classical concepts on finite difference schemes, continuous dynamical systems and discrete dynamical systems. We also studied the mathematical foundations of the non-standard finite difference method summarized in [7] by the triple question below. What is a non-standard finite difference method? In which way are non-standard schemes powerful compared to the standard ones? How to construct systematically non-standard finite difference methods?

However, the main contributions of this thesis are as follows. To address the issues 1-3, we constructed non-standard one-stage and two-stage theta methods for stiff and non-stiff systems of ordinary differential equations. The schemes were obtained by using Mickens'rule about the denominator of the discrete derivatives. On the one hand, we showed that the condition in [14] is equally sufficient for the elementary stability in this general setting of non-standard theta methods. On the other hand, we proved that the stated condition is equivalent to having the eigenvalues of the Jacobian matrices located in some wedges of the complex plane and we explained how the condition can be used in practice.

For a particular class of dynamical systems, which have non-hyperbolic fixed-points and which is equivalent to some specific Hamiltonian systems, we derived energy-preserving non-standard finite difference schemes. The schemes were constructed by using Mickens' rule about the nonlocal approximation of nonlinear terms [26].

Unlike the work in [24], the term dissipative is used in this thesis to express the fact that the gross asymptotics of a dynamical system

are independent of initial conditions with everything ending up inside some absorbing set. We showed that, for θ taking the smallest value 0 in the forbidden interval $[0, \frac{1}{2})$, our explicit scheme i.e., the non-standard forward Euler scheme, replicates the dissipative property of the continuous dynamical system.

As for the issue no. 4, we used a much simpler functional relation between step sizes and we proposed a systematic procedure of designing new qualitatively stable schemes for the general reaction-diffusion equations that involve arbitrary reaction terms. The positivity and the boundedness of the non-standard discrete solutions was established in this general setting. Furthermore, we designed for this general case an alternative method. It consists of a spectral method (in the space variable) and a non-standard finite difference method (in the time variable) in which the stiffness feature of the linearised system of Fourier coefficients is exactly incorporated.

Throughout the thesis, we presented numerical tests that support the theory. The accomplishment of this thesis has raised some concerns for future research. Among them, we can mention the following:

1. The design of non-standard finite difference schemes for dynamical systems with non-hyperbolic fixed-points. This is actually an open problem.

2. The design of non-standard finite difference schemes that preserve global attractors of continuous dynamical systems.

3. The investigation of the dissipativity of the non-standard theta methods for any value of the parameter θ.

4. The design of dissipative schemes for evolution partial differential differential equations.

5. The design of schemes, which display the boundedness and positivity property of solutions for the convective/advection-reaction-diffusion equation considered in [21] and for the Burger equation, as pointed out in [27].

Bibliography

[1] H. Al-Kahby, F. Dannan and S. Elaydi, Non-standard Discretization Methods for Some Biological Models, In: R.E. Mickens (Ed), *Applications of Non-standard Finite Difference Schemes*, World Scientific, Singapore, 2000, 155-180.

[2] R.M. Anderson and R.M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, 1991.

[3] R. Anguelov, J.K. Djoko, P. Kama and J.M.-S. Lubuma, On Elementary Stable and Dissipative Non-standard Finite Difference Schemes for Dynamical Systems, *Proceedings of the International Conference of Computational Methods in Science and Engineering* (Crete, Greece, 27 October-1 November 2006), Lecture Series on Computer and Computational Sciences, Vol 7A, VSP International Science Publishers,Utrecht, 2006, 24-27.

[4] R. Anguelov R, J.K. Djoko, P. Kama,J.M-S Lubuma, On Finite Difference Schemes Having the Correct Linear Stability and Dissipative Properties of Dynamical Systems.*University of Pretoria Technical Report No. UPWT*, 2007/02, submitted.

[5] R. Anguelov, P. Kama and J.M-S. Lubuma, Non-standard Theta Methods and Related Discrete Schemes for the Reaction-Diffusion Equations, In T.E. Simos (Ed),*Proceedings of the International Conference of Computational Methods in Science and Engineering* (Kastoria, Greece, 12-16 September 2003), World Scientific, Singapore, 2003, 24-27.

[6] R. Anguelov, P. Kama and J.M-S. Lubuma, On Non-standard Finite Difference Models of Reaction-Diffusion Equations, *Journal Computational and Applied Mathematics*, **175** (2005), 11-29.

[7] R. Anguelov and J.M-S. Lubuma, Contributions to the Mathematics of the Non-standard Finite Difference Methods and Applications, *Numer. Methods Partial Differential Equations*, **17** (2001), 518-543.

[8] R. Anguelov, J.M-S. Lubuma, Nonstandard Finite Difference Method by Nonlocal Approximation, *Mathematics and Computers in Simulation* **61**(3-6) (2003), 465-475.

[9] G.J. Barclay, D.F. Griffiths and D.J. Higham. Theta Method Dynamics, *London Mathematical Society J. comput. Math*, **3**(2000), 27-43.

[10] C. Canuto, M.Y. Hussani, A. Quarteroni and T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.

[11] R. Dautray and J-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology,*Vol 6, Springer,New York, 1988.

[12] K. Dekker and J.G.Verwer *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*. Elsevier Science Pub. Co, Amsterdam, 1984.

[13] D. T. Dimitrov and H.V. Kojouharov, Positive and Elementary Stable Non-standard Numerical Methods with Applications to Predator-Prey Models, *Journal of Computational and Applied Mathematics*, **189** (2006), 98-108.

[14] D. T. Dimitrov, H.V. Kojouharov and B. M. Chen-Charpentier, Reliable Finite Difference Schemes with Applications in Mathematical Biology, In: R.E. Mickens (Ed.), *Advances in the Applications of Nonstandard Finite Difference Schemes*, World Scientific, Singapore, 2005, 249-285.

[15] Y. Dumont amd J.M-S. Lubuma, Non-standard Finite Difference Schemes for Vibro-Impact Problems, *Proceedings of the Royal Society A*, **461** (2005), 1927-1950.

[16] L.C.Evans, *Partial Differential Equations*, Vol 19, American Mathematical Society, Providence, Rhode Island, 1998.

[17] A.B. Gumel (Ed.), *Journal of Difference Equations and Application*, Volume 9, 2003, Special Issue no 11-12 dedicated to Prof R.E. Mickens on the occasion of his 60th birthday.

[18] H.W. Hethcote, *The Mathematics of Infectious Diseases*, SIAM Review **42**(2000), 599-653.

[19] D.S. Jones and B.D. Sleeman, *Differential Equations and Mathematical Biology* Chapman and Hall/CRC, New York, 2003.

[20] A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, New York, 1995.

[21] H.V. Kojouharov and B. M. Chen, Non-standard Methods for Advection-Diffusion-Reaction Equations, In: R.E. Mickens (Ed.), *Applications of Nonstandard Finite Difference Schemes*, World Scientific Publishing Company, London, 2000, 55-108.

[22] J. D. Lambert. *Numerical Methods for Ordinary Differential Systems*, John Wiley and Sons, New York, 1991.

[23] J.D. Logan, *Nonlinear Differential Equations*, Wiley-Interscience, New York, 1994.

[24] J.M.-S. Lubuma and K.C. Patidar, Contributions to the theory of non-standard finite difference methods and applications to singular perturbation problems, In: R.E.Mickens (Ed.), *Advances in the applications of nonstandard finite difference schemes*, World Scientific, Singapore, 2005, 513-560.

[25] J.M-S Lubuma and A. Roux, An Improved Theta Method for the Systems of Ordinary Differential equations, *J Difference Equations and Applications*, **9**(11) (2003), 1023-1035.

[26] R.E. Mickens, *Nonstandard Finite Difference Models of Differential Equations,* World Scientific, Singapore, 1994.

[27] R.E. Mickens, Relation Between the Time and Space Step-Sizes in Nonstandard Finite Difference Schemes for the Fisher Equation.*Numerical Methods for Partial Differential Equations*, **13**(1)(1997), 51-55.

[28] R.E. Mickens (Ed.),*Applications of Nonstandard Finite Difference Schemes*, World Scientific, Singapore, 2000.

[29] R.E. Mickens, Nonstandard Finite Difference Methods, In: R.E. Mickens (Ed), *Advances in the Applications of Nonstandard Finite Difference Schemes*, World Scientific, Singapore, 2005, 1-9.

[30] R.E. Mickens, Discrete Models of Differential Equations: the Roles of Dynamic Consistency and Positivity, In: L.J.S. Allen, B. Aulbach, S. Elaydi and R. Sacker (Eds), *Difference Equations and Discrete Dynamical Systems* (Proceedings of the 9th International Conference), Los Angeles, USA, 2-7 August 2004, World Scientific, Singapore, 2005, 51-70.

[31] A. R. Mitchell and D. F. Griffiths, *Finite Difference Methods in Partial Differential Equations*; Wiley, New York, 1980.

[32] K.W. Morton and D.F. Mayers. *Numerical Solution of Partial Differential Equations*, Cambridge University Press, New York, 1994.

[33] J.D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1989.

[34] R.E. O'Malley Jr., *Thinking About Ordinary Differential Equations,* Cambridge University Press, New York, 1997.

[35] K.C. Patidar, On the Use of Non-standard Finite Difference Methods, *Journal of Difference Equations and Applications*, **11**(8) (2005, 735-758.

[36] P.A. Raviart and J.M. Thomas, *Introduction a L'analyse Numerique des Equations Auxderivees Partielles*, Masson, 1983.

[37] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*; Wiley-Interscience, New York,1967.

[38] A. Roux, Fourier Series and Spectral-Finite Difference Methods for the General Linear Diffusion Equation, *University of Pretoria Internal Report* **UPWI 2002/03.**

[39] G.D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, Oxford University Press, New York, 1985.

[40] J.Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

[41] A.M. Stuart and A.R. Humphries. *Dynamical Systems and Numerical Analysis*, Cambridge University Press, New York, 1998.

[42] A.M. Stuart and A.T. Peplow. The Dynamics of the Theta Methods, *SIAM J. Sci. Stat. Comput.*, **12**(1991), 1351-1372.

[43] R. Temam, *Navier-Stokes Equations Theory and Numerical Analysis*, North-Holland, Amsterdam, 1984.

[44] R. Temam. *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Berlin, 1988.

[45] J.W. Thomas. *Numerical Partial Differential Equations*, Springer, New York, 1988.

[46] W. Walter, *Ordinary Differential Equations*, Springer-Verlag, New York, 1998.

[47] S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

[48] E. Zeidler. *Applied Functional Analysis: Applications to Physics*, Springer-Verlag, New York, 1995.

[49] E. Zeidler. *Applied Functional Analysis: Main Principles and their Applications*, Springer-Verlag, New York, 1995.

Summary

Non-standard finite difference methods in dynamical systems

Student: Phumezile Kama

Supervisor: Professor Jean M-S Lubuma

Department: Mathematics and Applied Mathematics

Degree: Philosophiae Doctor

Date submitted: April 2009

This thesis is devoted to the study of numerical methods for dynamical systems. The numerical methods are expected to define discrete dynamical systems that are required to preserve the essential properties of the exact solution. The shortcomings of the classical numerical methods, specifically the theta methods, for being reliable discrete dynamical systems is that the step size is subjected to a constraint. The time step size should be small enough if the schemes were to replicate qualitative properties of the exact solutions.

The schemes we study are non-standard variants of the theta methods. The non-standard finite difference method aims at preserving the qualitative properties at no cost with regard to the value of time step size. We analyse non-standard finite difference schemes that have no spurious fixed-points compared to the dynamical system under consideration, the linear stability/instability property of the fixed-points

being the same for both the discrete and continuous systems. We obtain a sharper condition for the elementary stability of the schemes. For more complex dynamical systems which are dissipative, we design schemes that replicate this property.

We consider a specific class of dynamical systems which is equivalent to the simplest model of Hamiltonian systems that occur in classical mechanics. We design a non-standard finite difference scheme that replicates the underlying principle of conversation of energy.

We analyse the Fisher equation which enjoys a positivity and boundedness property. For the reaction-diffusion equation we obtain non-standard finite difference schemes that are elementary stable in the limit case of space independent variable and which are stable with respect to the principle of conservation of energy in the stationary case. As an alternative approach, we approximate the space variable by the spectral method, while the time variable is approximated via the non-standard finite difference scheme.

Throughout the thesis, we provide numerical experiments that support the theory.