

**Springer Series in
Computational
Mathematics**

33

Editorial Board

R. Bank, La Jolla (CA)
R.L. Graham, La Jolla (CA)
J. Stoer, Würzburg
R. Varga, Kent (Ohio)
H. Yserentant, Tübingen

Springer-Verlag Berlin Heidelberg GmbH

Willem Hundsdorfer
Jan Verwer

Numerical Solution of Time-Dependent Advection-Diffusion- Reaction Equations



Springer

Willem Hunds dorfer
Jan Verwer
Center for Mathematics
and Computer Science (CWI)
Kruislaan 413
1098 SJ Amsterdam
The Netherlands
e-mail: willem.hunds dorfer@cwi.nl
jan.verwer@cwi.nl

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Mathematics Subject Classification (2000): 65L05, 65L06, 65L20, 65M06, 65M12,
65M20, 65M60

ISSN 0179-3632

ISBN 978-3-642-05707-6 ISBN 978-3-662-09017-6 (eBook)

DOI 10.1007/978-3-662-09017-6

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2003

Originally published by Springer-Verlag Berlin Heidelberg New York in 2003.

Softcover reprint of the hardcover 1st edition 2003

The use of general descriptive names, registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: *design&production*, Heidelberg

Typeset by the authors

Printed on acid-free paper 46/3142LK-543210

Preface

This book deals with numerical methods for solving partial differential equations (PDEs) coupling advection, diffusion and reaction terms, with a focus on time-dependency. A combined treatment is presented of methods for hyperbolic problems, thereby emphasizing the one-way wave equation, methods for parabolic problems and methods for stiff and non-stiff ordinary differential equations (ODEs). With regard to time-dependency we have attempted to present the algorithms and the discussion of their properties for the three different types of differential equations in a unified way by using semi-discretizations, i.e., the method of lines, whereby the PDE is transformed into an ODE by a suitable spatial discretization. In addition, for hyperbolic problems we also discuss discretizations that use information based on characteristics. Due to this combination of methods, this book differs substantially from more specialized textbooks that deal exclusively with numerical methods for either PDEs or ODEs. We treat integration methods suitable for both classes of problems.

This combined treatment offers a clear advantage. On the one hand, in the field of numerical ODEs highly valuable methods and results exist which are of practical use for solving time-dependent PDEs, something which is often not fully exploited by numerical PDE researchers. Although many problems can be solved by Euler's method or the Crank-Nicolson method, better alternatives are often available which can significantly reduce the computational effort needed to solve practical problems. On the other hand, many numerical ODE researchers are unaware of the vast amount of highly interesting results on discretization methods for PDEs. Moreover, when solving PDEs, discretizations in space and time have to be matched, and different spatial discretizations may require different temporal discretizations. It is our hope that this book bridges these gaps, if not fully, then at least partially.

With regard to applications our aim has been to present material specifically directed at solving so-called transport-chemistry problems, i.e., problems where the transport part is based on advection and diffusion processes and the chemistry part on chemical reaction processes modelled by ordinary differential equations. Such transport-chemistry problems are frequently used in environmental modelling, notably in connection with pollution of atmospheric air, surface water and groundwater. Similar problem types are also found in mathematical biology, for instance with chemo-taxis problems that are used to study bacterial growth, tumour growth and related biochemical phenomena. Hence throughout the book our dependent variables mostly rep-

resent concentrations of chemical species, and we therefore give considerable attention to monotonicity and positivity properties of numerical schemes, i.e., to the question how to prevent spurious, negative numerical concentrations associated with spurious temporal and spatial oscillations, the plague of many discretization methods for differential equations with dominating hyperbolic terms.

The focus on advection-diffusion-reaction problems enabled us to keep the size of the text within reasonable limits. As a consequence, however, for certain important classes of PDEs, like the Maxwell and Navier-Stokes equations, only some aspects are touched. Moreover, the main focus in this book is on time-dependency. Spatial discretizations are obtained by finite differences or finite volumes, and, to a lesser degree, by finite elements. Spectral methods and their variants are not treated, nor are elliptic problems and the associated, specialized numerical linear algebra solvers.

The book has five chapters of which the first gives an introduction to the wide field of numerical solution of evolutionary PDEs and to the more specialized material presented in later chapters. This first chapter is written primarily for readers and students of applied and numerical mathematics and the exact sciences with little numerical training; this chapter contains exercises in footnotes. Once the first chapter is digested, the remaining four should be accessible. More experienced readers and students could skip most of Chapter I and concentrate on the more specialized subjects of Chapters II – V, dealing with time integration methods, advection-diffusion discretizations, splitting methods and stabilized explicit Runge-Kutta methods. These subjects have been chosen because of their practical relevance for solving advection-diffusion-reaction problems. But they obviously also reflect our personal taste and research history.

Each chapter is divided into numbered sections and subsections. Numbering of items like formulas, theorems and figures is done section-wise per chapter. Cross references to numbered items in other chapters are given explicitly with the chapter number in front.

We wish to thank all our colleagues who have helped us with the preparation of the book by reading parts of the manuscript and by providing us with many corrections, helpful remarks and suggestions for improvement. In alphabetical order we mention: Assyr Abdulle, Joke Blom, Jason Frank, Alf Gerisch, Piet Hemker, Barry Koren, Johannes Krottje and, last but not least, Ben Sommeijer. Moreover, Alf Gerisch provided great assistance with the numerical examples presented in Section IV.6, and Assyr Abdulle and Ben Sommeijer have given us valuable support when we used their codes for the experiments in Chapter V. We are also grateful to Paul Zegeling for carrying out the numerical experiment with moving grids for Section III.7. Lastly we acknowledge the pleasant co-operation with Springer-Verlag.

Table of Contents

I Basic Concepts and Discretizations	1
1 Advection-Diffusion-Reaction Equations	1
1.1 Nonlinear Reaction Problems from Chemistry	3
1.2 Model Equations for Advection-Diffusion	9
1.3 Multi-dimensional Problems	14
1.4 Examples of Applications	18
2 Basic Discretizations for ODEs	23
2.1 Initial Value Problems and Euler's Method	23
2.2 Norms and Matrices	27
2.3 Perturbations on ODE Systems	30
2.4 The θ -Method and Stiff Problems	35
2.5 Stability of the θ -Method	37
2.6 Consistency and Convergence of the θ -Method	42
2.7 Nonlinear Results for the θ -Method	44
2.8 Concluding Remarks	46
3 Basic Spatial Discretizations	48
3.1 Discrete Fourier Decompositions	49
3.2 The Advection Equation	52
3.3 The Diffusion Equation	62
3.4 The Advection-Diffusion Equation	66
4 Convergence of Spatial Discretizations	71
4.1 Stability, Consistency and Convergence	71
4.2 Advection-Diffusion with Constant Coefficients	74
4.3 Advection with Variable Coefficients	77
4.4 Diffusion with Variable Coefficients	81
4.5 Variable Coefficients and Higher-Order Schemes	83
5 Boundary Conditions and Spatial Accuracy	84
5.1 Refined Global Error Estimates	85
5.2 Outflow with Central Advection Discretization	86
5.3 Boundary Conditions with the Heat Equation	88
5.4 Boundary Conditions and Higher-Order Schemes	92
6 Time Stepping for PDEs	94
6.1 The Method of Lines and Direct Discretizations	94
6.2 Stability, Consistency and Convergence	99

VIII Table of Contents

6.3	Stability for MOL – Stability Regions	103
6.4	Von Neumann Stability Analysis	111
7	Monotonicity Properties	116
7.1	Positivity and Maximum Principle	116
7.2	Positive Semi-discrete Systems	118
7.3	Positive Time Stepping Methods	121
7.4	Numerical Illustrations	124
8	Numerical Test Examples	127
8.1	The Nonlinear Schrödinger Equation	128
8.2	The Angiogenesis Model	134
II	Time Integration Methods	139
1	Runge-Kutta Methods	139
1.1	The Order Conditions	140
1.2	Examples	142
1.3	The Stability Function	144
1.4	Step Size Restrictions for Advection-Diffusion	149
1.5	Rosenbrock Methods	151
2	Convergence of Runge-Kutta Methods	155
2.1	Order Reduction	155
2.2	Local Error Analysis	158
2.3	Global Error Analysis	161
2.4	Concluding Notes	166
3	Linear Multistep Methods	170
3.1	The Order Conditions	171
3.2	Examples	173
3.3	Stability Analysis	174
3.4	Step Size Restrictions for Advection-Diffusion	181
3.5	Convergence Analysis	182
4	Monotone ODE Methods	185
4.1	Linear Positivity for One-Step Methods	185
4.2	Nonlinear Positivity for One-Step Methods	189
4.3	Positivity for Multistep Methods	192
4.4	Related Monotonicity Results	196
5	Variable Step Size Control	197
5.1	Step Size Selection	197
5.2	An Explicit Runge-Kutta Example	200
5.3	An Implicit Multistep Example	203
5.4	General Purpose ODE Codes	205
6	Numerical Examples	206
6.1	A Model for Antibodies in Tumorous Tissue	206
6.2	The Nonlinear Schrödinger Equation	209

III Advection-Diffusion Discretizations	215
1 Non-oscillatory MOL Advection Discretizations	215
1.1 Spatial Discretization for Linear Advection	215
1.2 Numerical Examples	222
1.3 Positivity and the TVD Property	226
1.4 Nonlinear Scalar Conservation Laws	233
2 Direct Space-Time Advection Discretizations	239
2.1 Optimal-Order DST Schemes	239
2.2 A Non-oscillatory Third-Order DST Scheme	243
2.3 Explicit Schemes with Unconditional Stability	248
3 Implicit Spatial Discretizations	250
3.1 Order Conditions	251
3.2 Examples	253
3.3 Stability and Convergence	258
3.4 Monotonicity	261
3.5 Time Integration Aspects	263
4 Non-uniform Grids – Finite Volumes (1D)	264
4.1 Vertex Centered Schemes	265
4.2 Cell Centered Schemes	272
4.3 Numerical Illustrations	278
4.4 Higher-Order Methods and Limiting	281
5 Non-uniform Grids – Finite Elements (1D)	283
5.1 The Basic Galerkin Method	283
5.2 Standard Galerkin Error Estimates	288
5.3 Upwinding	291
6 Multi-dimensional Aspects	292
6.1 Cartesian Grid Discretizations	293
6.2 Diffusion on Cartesian Grids	295
6.3 Advection on Cartesian Grids	303
6.4 Transformed Cartesian Grids	308
6.5 Unstructured Grids	311
7 Notes on Moving Grids and Grid Refinement	316
7.1 Dynamic Regridding	316
7.2 Static Regridding	321
IV Splitting Methods	325
1 Operator Splitting	325
1.1 First-Order Splitting	325
1.2 Second-Order Symmetrical Splitting	329
1.3 Higher-Order Splittings	330
1.4 Abstract Initial Value Problems	331
1.5 Advection-Diffusion-Reaction Splittings	335
1.6 Dimension Splitting	337
1.7 Boundary Values and Stiff Terms	344
2 LOD Methods	348

2.1	The LOD-Backward Euler Method	348
2.2	LOD Crank-Nicolson Methods	351
2.3	The Trapezoidal Splitting Method	359
2.4	Boundary Correction Techniques	365
2.5	Numerical Comparisons	367
3	ADI Methods	369
3.1	The Peaceman-Rachford Method	369
3.2	The Douglas Method	373
4	IMEX Methods	383
4.1	The IMEX- θ Method	383
4.2	IMEX Multistep Methods	386
4.3	Notes on IMEX Runge-Kutta Methods	391
4.4	Concluding Remarks and Tests	393
5	Rosenbrock AMF Methods	398
5.1	One-Stage Methods of Order One and Two	398
5.2	Two-Stage Methods of Order Two and Three	400
5.3	A Three-Stage Method of Order Two	403
5.4	Concluding Remarks and Tests	405
6	Numerical Examples	409
6.1	Two Chemo-taxis Problems from Biology	409
6.2	The Numerical Methods	411
6.3	Numerical Experiments	412
V	Stabilized Explicit Runge-Kutta Methods	419
1	The RKC Family	420
1.1	Stability Polynomials	420
1.2	Integration Formulas	426
1.3	Internal Stability and Full Convergence Properties	430
2	The ROCK Family	433
2.1	Stability Polynomials	433
2.2	Integration Formulas	435
2.3	Internal Stability and Convergence	436
3	Numerical Examples	438
3.1	A Combustion Model	439
3.2	A Radiation-Diffusion Model	441
	Bibliography	447
	Index	465

I Basic Concepts and Discretizations

This chapter gives a first introduction to the numerical solution of time-dependent advection-diffusion-reaction problems. Our goal in this chapter is to discuss important basic concepts and discretizations for ordinary differential equations and for advection and diffusion equations in one spatial dimension. More advanced material will be treated in later chapters.

1 Advection-Diffusion-Reaction Equations

In this first section we will consider some properties of solutions of linear advection and diffusion equations and nonlinear chemical reaction equations and briefly mention some application fields.

The standard advection-diffusion-reaction model deals with the time evolution of chemical or biological species in a flowing medium such as water or air. The mathematical equations describing this evolution are partial differential equations (PDEs) that can be derived from mass balances. Consider a concentration $u(x, t)$ of a certain species, with space variable $x \in \mathbb{R}$ and time $t \geq 0$. Let $h > 0$ be a small number, and consider the average concentration $\bar{u}(x, t)$ in a cell $[x - \frac{1}{2}h, x + \frac{1}{2}h]$,

$$\bar{u}(x, t) = \frac{1}{h} \int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} u(s, t) ds = u(x, t) + \frac{1}{24} h^2 \frac{\partial^2}{\partial x^2} u(x, t) + \dots$$

If the species is carried along by a flowing medium with velocity $a(x, t)$, then the mass conservation law implies that the change of $\bar{u}(x, t)$ per unit of time is the net balance of inflow and outflow over the cell boundaries,

$$\frac{\partial}{\partial t} \bar{u}(x, t) = \frac{1}{h} \left[a(x - \frac{1}{2}h, t) u(x - \frac{1}{2}h, t) - a(x + \frac{1}{2}h, t) u(x + \frac{1}{2}h, t) \right],$$

where $a(x \pm \frac{1}{2}h, t)u(x \pm \frac{1}{2}h, t)$ are the mass fluxes over the left and right cell boundaries. Now, if we let $h \rightarrow 0$, it follows that the concentration satisfies

$$\frac{\partial}{\partial t} u(x, t) + \frac{\partial}{\partial x} \left(a(x, t) u(x, t) \right) = 0.$$

This is called an advection equation (or convection equation).¹⁾ In a similar way we can consider the effect of diffusion. Then the change of $\bar{u}(x, t)$ is caused by gradients in the solution and the fluxes across the cell boundaries are $-d(x \pm \frac{1}{2}h, t)u_x(x \pm \frac{1}{2}h, t)$ with $d(x, t)$ the diffusion coefficient. The corresponding diffusion equation is

$$\frac{\partial}{\partial t}u(x, t) = \frac{\partial}{\partial x}\left(d(x, t)\frac{\partial}{\partial x}u(x, t)\right).$$

There may also be a local change in $u(x, t)$ due to sources, sinks and chemical reactions, which is described by

$$\frac{\partial}{\partial t}u(x, t) = f(x, t, u(x, t)).$$

The overall change in concentration is described by combining these three effects, leading to the advection-diffusion-reaction equation

$$\begin{aligned} \frac{\partial}{\partial t}u(x, t) + \frac{\partial}{\partial x}\left(a(x, t)u(x, t)\right) \\ = \frac{\partial}{\partial x}\left(d(x, t)\frac{\partial}{\partial x}u(x, t)\right) + f(x, t, u(x, t)). \end{aligned} \quad (1.1)$$

We will consider (1.1) in a spatial interval $\Omega \subset \mathbb{R}$ with time $t > 0$. An initial profile $u(x, 0)$ will be given and we also assume that suitable boundary conditions are provided. In Section 1.3 the extension to more spatial dimensions is discussed.

In the notation we will usually omit the explicit dependence of x and t , and partial derivatives will be denoted by sub-indices. Thus (1.1) will be written as

$$u_t + (au)_x = (du_x)_x + f(u). \quad (1.2)$$

Occasionally we will also use the notation $\partial_x = \frac{\partial}{\partial x}$, $\partial_{xx} = \frac{\partial^2}{\partial x^2}$ for the spatial differential operators.

In these equations it is assumed that the advection and diffusion coefficients, $a(x, t)$ and $d(x, t)$, are given and independent of the concentration $u(x, t)$. Hence the advection and diffusion terms are linear. In many applications this assumption is valid. On the other hand, there are also many nonlinear problems where the coefficients will depend on the concentrations, but we will at first only regard linear advection and diffusion.

Some simple examples of actual models are presented at the end of this section. First the individual effect of reaction, advection and diffusion will be discussed.

¹⁾ The terms ‘advection’ and ‘convection’ are used indiscriminately in the numerical analysis literature. In meteorology, advection is the passive transport by horizontal wind whereas convection refers to vertical transport, which is usually caused by localized vertical heat gradients.

1.1 Nonlinear Reaction Problems from Chemistry

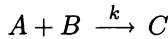
Instead of a single concentration we will consider in general a vector of concentrations $u = (u_1, u_2, \dots, u_s)^T$. In the description of chemical reactions it is usually assumed that the concentrations are homogeneous in x , due to stirring for instance. In that case there will be no spatial dependence and so (1.1) reduces to a system of ordinary differential equations (ODEs)

$$u'(t) = f(t, u(t)), \quad t > 0, \quad (1.3)$$

with given initial value $u(0)$ in \mathbb{R}^s , and with $u'(t)$ standing for the time derivative. This ODE system will often be written as $u' = f(t, u)$ and the equations will be considered on time intervals $(0, T]$ or $(0, \infty)$. We will describe here some characteristic features of nonlinear ODE systems for chemical reactions, such as linear conservation properties and positivity.

A Single Chemical Reaction

First we consider, by way of example, the single chemical reaction



describing the reaction of two species A and B into a species C with reaction rate constant k . Let a, b, c denote the concentrations of the three species. According to the mass action law of chemical kinetics – see Aris (1965) and Gavalas (1968), for example – the speed of the reaction is proportional to ab , leading to the following ODE system

$$\begin{aligned} a'(t) &= -k a(t) b(t), \\ b'(t) &= -k a(t) b(t), \\ c'(t) &= k a(t) b(t). \end{aligned}$$

Note that we have $a'(t) + c'(t) = 0$ and $b'(t) + c'(t) = 0$, or equivalently,

$$a(t) + c(t) = a(0) + c(0), \quad b(t) + c(t) = b(0) + c(0).$$

These two relations are linear invariants for any exact solution. They are a consequence of the fact that in a closed system of chemical reactions the total number of atoms must remain constant. Such relations are therefore also called *mass conservation laws*, expressing this property of conservation of molecular mass.

Another important property of chemical systems is *positivity*, by which we actually mean preservation of non-negativity for all components. This is of course a natural property for chemical concentrations, and thus it should also hold for the mathematical description. More precisely, we should have $a(t), b(t), c(t) \geq 0$ for all $t \geq 0$ whenever $a(0), b(0), c(0) \geq 0$. For the above

system this can be demonstrated from the exact solution. Denote $d(t) = a(t) - b(t)$. We already know that $d(t) = d(0)$ for all $t \geq 0$. Eliminating b from $a' = -k ab$ gives the scalar equation

$$a'(t) = -k a(t)^2 + k d(0)a(t).$$

The exact solution of this equation is given by

$$a(t) = \frac{a(0)}{1 + b(0)q(t)}, \quad q(t) = \frac{1 - e^{-kt d(0)}}{d(0)},$$

where $q(t)$ is to be replaced by kt if $d(0) = 0$. For component a the positivity property now follows from the observation that in all circumstances $q(t) \geq 0$, and hence $a(t) \geq 0$ for all $t \geq 0$ provided that $a(0), b(0) \geq 0$. This short computation can be redone for components b and c , showing the positivity property for the whole system.

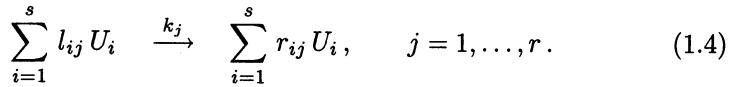
This example can also be used to illustrate the important fact that negative initial values are not allowed as they can lead to an unstable solution and even to blow-up in finite time. For component a we have blow-up in finite time if the denominator $1 + b(0)q(t)$ vanishes for a finite positive value of t . This holds if the equation

$$e^{-kt d(0)} = \frac{a(0)}{b(0)}$$

is satisfied for a finite positive value of t , which is easily shown to be true for any choice of $a(0), b(0) < 0$. Later on we will see that the requirement of positivity imposes severe restrictions on numerical methods. Even a small numerical error leading to a small negative value can cause instability and blow-up if the numerical method mimics the above unstable behaviour.

General Reaction Systems

Next consider r chemical reactions between s species U_i , $i = 1, \dots, s$, with concentrations u_i . Assume that the s species simultaneously interact in the r reactions



Here l_{ij}, r_{ij} are non-negative integers (the so-called stoichiometric coefficients) describing loss and gain of the number of molecules U_i in the j th reaction, and $k_j > 0$ is the rate constant of the reaction which may depend on the time t due to external influences, such as temperature or sunlight. According to the mass action law of chemical kinetics the speed of the j th reaction is given by the rate function

$$g_j(t, u) = k_j(t) \prod_{n=1}^s (u_n)^{l_{nj}},$$

which is proportional to the product of all concentrations on the left-hand side of the reaction. By considering the net result of all reactions on U_i , this yields the set of ODEs

$$u'_i(t) = \sum_{j=1}^r (r_{ij} - l_{ij}) g_j(t, u(t)), \quad i = 1, \dots, s.$$

We can write this in system form, with $u = (u_1, \dots, u_s)^T$,

$$u'(t) = S g(t, u(t)), \quad u(0) \text{ given}, \quad (1.5)$$

where $S = (r_{ij} - l_{ij})$ is an $s \times r$ matrix and $g(t, u) = (g_j(t, u)) \in \mathbb{R}^r$. The matrix S is called the stoichiometric matrix. Because in most chemistry models only second-order (bimolecular) reactions are taken into account for which $\sum_j l_{ij} \leq 2$, one mostly encounters functions that are quadratic in u . With only first-order (monomolecular) reactions, for which $\sum_j l_{ij} \leq 1$, the system is linear in u . If, in addition, the reaction constants do not depend on t , the system is of constant coefficient linear type.

System (1.5) is often presented in the so-called production-loss form

$$u'(t) = p(t, u(t)) - L(t, u(t)) u(t), \quad u(0) \text{ given}, \quad (1.6)$$

with production vector $p(t, u) = (p_i(t, u)) \in \mathbb{R}^s$ and diagonal loss matrix $L(t, u) = \text{diag}(L_i(t, u)) \in \mathbb{R}^{s \times s}$ given by

$$p_i(t, u) = \sum_{j=1}^r r_{ij} g_j(t, u), \quad L_i(t, u) = \sum_{j=1}^r l_{ij} g_j(t, u)/u_i.$$

The functions $p_i(t, u)$ and $L_i(t, u)$ are polynomials in the arguments u_j with non-negative coefficients,²⁾ and hence we have $p(t, u) \geq 0$, $L(t, u) \geq 0$ whenever $u \geq 0$.³⁾ If we have at most second-order reactions the loss matrix L is linear in u .

Because (1.6) models chemical reactions, one might be tempted to believe that all problems of type (1.6) possess a solution for all $t > 0$. We already have seen that with negative initial values this cannot be guaranteed. In general there is no guarantee either that it is true for positive initial values. Consider, for example, the simple scalar equation

$$u'(t) = \kappa u(t)^2, \quad u(0) \text{ given},$$

which would result from $2U + X \rightarrow 3U + Y$ if we assume that the concentration of X is constant (abundantly available). The solution of the equation reads

$$u(t) = \frac{u(0)}{1 - \kappa u(0) t}.$$

²⁾ To see that L_i is polynomial, rewrite it as $L_i(t, u) = \sum_j k_j l_{ij} (u_i)^{l_{ij}-1} \prod_{n \neq i} (u_n)^{l_{nj}}$.

³⁾ If we write an inequality property like ≥ 0 for a vector or a matrix, this is meant to apply to all vector components or matrix entries.

Hence if $\kappa > 0$ and $u(0) > 0$, the solution only exists for $0 \leq t < 1/(\kappa u(0))$ and it blows up when t approaches $1/(\kappa u(0))$. This divergence is the result of inaccurate modelling: in this simple reaction the molecules X will disappear at a rate proportional to u^2 , and therefore the assumption that the concentration of X is constant should only be used if u is sufficiently small.

We now proceed with the properties of *positivity* and *mass conservation* for the general chemical kinetic system. For a vector u we write $u \geq 0$ if all its components are non-negative. The positivity property

$$u(0) \geq 0 \implies u(t) \geq 0 \text{ for all } t > 0$$

can be demonstrated from the production-loss form (1.6) by observing that if $u_i \downarrow 0$ then the loss term $L_i(t, u)u_i$ will disappear and thus only a production term $p_i(t, u) \geq 0$ will remain, preventing u_i to become negative; a formal proof will be given in Section 7. For discussing mass conservation the form (1.5) is more appropriate. Consider a weight vector $v = (v_1, \dots, v_s)^T$ with constant entries. We want to establish when for this vector the linear sum

$$v^T u(t) = \sum_{i=1}^s v_i u_i(t) = \text{constant}, \quad (1.7)$$

indicating a linear invariant. The answer lies in the null-space of the transpose S^T of the stoichiometric matrix S . Suppose v belongs to this null-space, i.e., $S^T v = 0$, and thus $v^T S = 0$. Multiplying (1.5) by v^T then gives $v^T u'(t) = 0$ for all t and hence (1.7) is valid. Consequently, the possible linear invariants are given by the set of linearly independent vectors v that span the null-space.

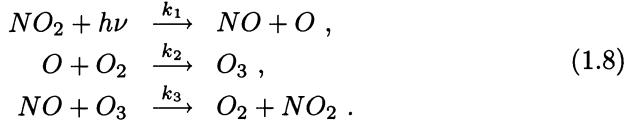
Later on it will be shown that standard numerical integration methods for ODE systems preserve the linear invariants and hence are mass conservative. There also exist integration methods for which this is not entirely true, for instance if steady-state approximations are used for fast reactive terms, but such methods will not be considered. With respect to numerical integration, the positivity property is in general of greater concern because of the danger of instability which can completely ruin a numerical calculation. With only few exceptions, standard numerical integration methods for ODE systems do *not* guarantee positivity.

An Example from Atmospheric Chemistry

In atmospheric chemistry research one studies chemical reactions between trace gases, such as ozone, nitrogen oxides, methane, hydrocarbons, etc. An important concern is air pollution due to anthropogenic (man-made) emissions of primary and secondary polluting species. Ozone levels in the lower troposphere are of particular concern, as ozone is dangerous for humans and animals during short term smog episodes and can damage crops when over longer seasonal periods levels are too high. Ozone itself is not emitted but formed in many different chemical reactions. Ozone is also a greenhouse gas,

similar as methane, carbon dioxide and other species. Air pollution models are therefore also used in connection with climate studies. A nice introduction to the field of atmospheric chemistry and important environmental issues related to air pollution can be found in Graedel & Crutzen (1995).

Example 1.1 We illustrate the mass action law by the following three reactions between oxygen O_2 , atomic oxygen O , nitrogen oxide NO , and nitrogen dioxide NO_2 :



These reactions are basic to any tropospheric air pollution model. The first reaction is photochemical and says that NO and O are formed from NO_2 by photo-dissociation caused by solar radiation, indicated by $h\nu$. This depends on the time of the day and therefore $k_1 = k_1(t)$. We consider the concentrations $u_1 = [O]$, $u_2 = [NO]$, $u_3 = [NO_2]$, $u_4 = [O_3]$. The oxygen concentration $[O_2]$ is treated as constant, and we assume there is a constant source term σ_2 simulating emission of nitrogen oxide. The rate functions and stoichiometric matrix are

$$g(t, u) = \begin{pmatrix} k_1(t)u_3 \\ k_2u_1 \\ k_3u_2u_4 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}.$$

The corresponding ODE system reads

$$\begin{aligned} u'_1(t) &= k_1(t)u_3(t) - k_2u_1(t), \\ u'_2(t) &= k_1(t)u_3(t) - k_3u_2(t)u_4(t) + \sigma_2, \\ u'_3(t) &= k_3u_2(t)u_4(t) - k_1(t)u_3(t), \\ u'_4(t) &= k_2u_1(t) - k_3u_2(t)u_4(t). \end{aligned}$$

We see immediately that

$$u'_1(t) + u'_3(t) + u'_4(t) = 0, \quad u'_2(t) + u'_3(t) = \sigma_2,$$

hence $[O] + [NO_2] + [O_3]$ is a conserved quantity while $[NO] + [NO_2]$ growths with $\sigma_2 t$. By considering the null-space of S^T it is seen that these are the two mass laws for this chemical model.

The photochemical nature of the first reaction is special in the sense that it creates a diurnal cycle with rapid changes in concentration values at each sunset and sunrise. Figure 1.1 illustrates this cycle, which is typical for atmospheric chemical kinetic systems. The physical units are seconds for time and number of molecules per cm^3 for the concentrations. The figure shows

the time evolution of the concentrations over approximately 6 days for the set of initial and source values

$$u(0) = (0, 1.3 \cdot 10^8, 5.0 \cdot 10^{11}, 8.0 \cdot 10^{11})^T, \quad \sigma_2 = 10^6,$$

and reaction coefficients

$$k_3 = 10^{-16}, \quad k_2 = 10^5, \quad k_1(t) = \begin{cases} 10^{-5} e^{7.0 \sec(t)} & \text{in daytime,} \\ 10^{-40} & \text{during the night,} \end{cases}$$

where

$$\sec(t) = (\sin(\frac{1}{16}\pi(\bar{t}_h - 4)))^{0.2}, \quad \bar{t}_h = t_h - 24\lfloor t_h/24 \rfloor, \quad t_h = t/3600,$$

with daytime between 4 a.m. and 8 p.m. and with $\lfloor t_h/24 \rfloor$ denoting the largest integer $\leq t_h/24$. Hence $\sec(t)$ is periodic with a period of 24 hours, but defined only during daytime. The maximum for k_1 is equal to ≈ 0.01 and occurs at noon. The concentrations are plotted over approximately six days, from sunrise at day 1 (initial value) until sunset at day 6 (time interval $14400 \leq t \leq 504000$ sec.).

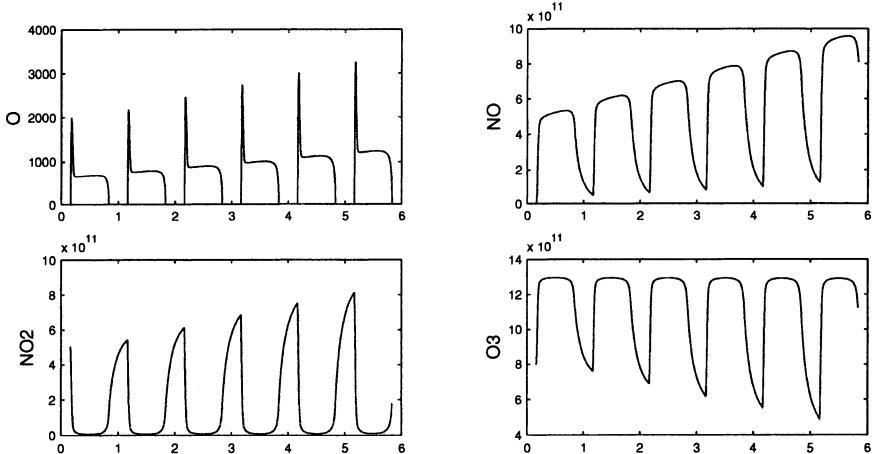


Fig. 1.1. Time evolution, from 4 a.m. at day 1 till 8 p.m. at day 6, for the concentrations of the atmospheric chemistry problem (1.8).

Although this chemistry system is too limited from the atmospheric chemical point of view, the variation of the concentrations due to the diurnal cycle are more or less realistically simulated and the concentration values and reaction coefficients approximate their counterparts used in real-life models. Because the oxygen concentration $[O_2]$ is considered constant, which is very realistic, k_2 contains the total number of O_2 molecules per cm^3 and is therefore much larger than k_1 and k_3 . For our system this means that the product

O_3 is formed much faster than the products of the first and second reaction. In all real-life models such large differences between rate coefficients occur causing a large spread in time-scales. This is related to the notion of stiffness, which is, as we will see in the next section, an important property to consider for numerical ODE methods. \diamond

1.2 Model Equations for Advection-Diffusion

We proceed with some basic properties of advection-diffusion equations, starting with simple constant coefficient cases in one space dimension (1D).

The scalar, 1D advection-diffusion equation

$$u_t + a u_x = d u_{xx} \quad (1.9)$$

with constant $a \in \mathbb{R}$ and $d \geq 0$ is an important test model for numerical schemes. We will consider this equation for $t > 0$ and $x \in \mathbb{R}$ with given initial function $u(x, 0) = u_0(x)$ and with the periodicity condition

$$u(x \pm 1, t) = u(x, t). \quad (1.10)$$

The reason for considering spatial periodicity is mainly the ease of presentation. Boundary conditions cause additional theoretical and numerical problems, as we will see gradually in later sections. Note that with (1.10) it is sufficient to consider $u(x, t)$ on the spatial interval $[0, 1]$. This interval can be best viewed as a ring with the points $x = 0$ and $x = 1$ glued together, so that it is clear that there are no genuine boundary conditions here, even though this periodicity condition is often referred to as a *periodic boundary condition*.

Because the advection-diffusion equation underlies the law of mass conservation, it is also called a *mass balance* equation. In particular, if $u(x, t)$ is a concentration then the integral

$$M(t) = \int_0^1 u(x, t) dx$$

represents the mass in $[0, 1]$ at time t . Since we assume periodicity, it is easily seen that M is a conserved quantity:

$$\begin{aligned} \frac{d}{dt} M(t) &= \int_0^1 u_t(x, t) dx = \int_0^1 (-au_x(x, t) + du_{xx}(x, t)) dx \\ &= -a(u(1, t) - u(0, t)) + d(u_x(1, t) - u_x(0, t)) = 0. \end{aligned}$$

Positivity of the advection-diffusion equation will be discussed in Section 7, but we mention here already that solutions satisfy the *maximum principle*

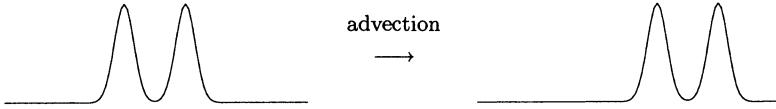
$$\min_{0 \leq \xi \leq 1} u(\xi, 0) \leq u(x, t) \leq \max_{0 \leq \xi \leq 1} u(\xi, 0), \quad (1.11)$$

and so we have in particular $u(x, t) \geq 0$ if $u(x, 0) \geq 0$ for all x .

The solution of the scalar advection (test-) equation

$$u_t + a u_x = 0 \quad (1.12)$$

is particularly simple for constant a . It is easily seen that $u(x, t) = u(x - at, 0)$ satisfies the equation, showing that the prescribed initial profile $u(x, 0)$ is merely shifted in time with velocity a without any change of shape. The lines $x - at = \text{const.}$ in the (x, t) -plane are called the *characteristics* of this advection equation. Along these characteristics the solution $u(x, t)$ is constant.



Next consider the scalar diffusion (test-) equation

$$u_t = d u_{xx} \quad (1.13)$$

with a constant $d > 0$. This equation is also known as the *heat equation* since it is a model for the evolution of the temperature distribution in a thin (one-dimensional) rod, with d denoting the heat conductivity. We will mainly interpret this equation as the result of molecular diffusion caused by Brownian motion of particles.

Fourier Decompositions

Insight in the solution of the diffusion equation (1.13) can be obtained by Fourier decompositions. Here we briefly recall some basic properties that are relevant to our one-dimensional advection-diffusion model problem. For this we consider the function space $L_2[0, 1]$ consisting of all square integrable complex functions on $[0, 1]$ with inner product and norm given by

$$(v, w) = \int_0^1 \overline{v(x)} w(x) dx, \quad \|v\| = (v, v)^{1/2}.$$

Further consider the functions

$$\varphi_k(x) = e^{2\pi i k x} \quad \text{for } k \in \mathbb{Z}. \quad (1.14)$$

These functions will be called the *Fourier modes* or *harmonics*. It is easily seen that the Fourier modes form an orthonormal set, $(\varphi_j, \varphi_k) = 0$ if $j \neq k$ and $\|\varphi_j\| = 1$. These modes are actually a basis for $L_2[0, 1]$, that is, any function $v \in L_2[0, 1]$ can be written as

$$v(x) = \sum_{k \in \mathbb{Z}} \alpha_k \varphi_k(x),$$

where the right-hand side is a convergent series, which we now call the Fourier series.⁴⁾ The Fourier coefficients are given by $\alpha_k = (\varphi_k, v)$ and we have

$$\|v\|^2 = \sum_{k \in \mathbb{Z}} |\alpha_k|^2.$$

This relation is known as Parseval's identity. Proofs for these statements can be found in standard analysis textbooks, for example Dym & McKean (1972), Pinkus & Zafrany (1997).

Consider the diffusion equation (1.13) with the periodicity condition (1.10) and with initial function $u(x, 0) = \varphi_k(x)$. To find the solution we make the ansatz (a motivated guess that will turn out right) of separation of variables,

$$u(x, t) = \gamma(t)\varphi_k(x), \quad \gamma(0) = 1. \quad (1.15)$$

Inserting this into (1.13) we see that $\gamma(t)$ must satisfy

$$\gamma'(t) = -4\pi^2 k^2 d \gamma(t), \quad \gamma(0) = 1,$$

which has the solution

$$\gamma(t) = e^{-4\pi^2 k^2 dt}.$$

It follows that with a single Fourier mode $\varphi_k(x)$ as initial condition, the periodic diffusion problem (1.13), (1.10) has the solution

$$u(x, t) = e^{-4\pi^2 k^2 dt} \varphi_k(x). \quad (1.16)$$

It is important to note that all Fourier modes with $k \neq 0$ are damped. The larger the wave number k , the stronger the damping and for $t \rightarrow \infty$ all solutions (1.16) with $k \neq 0$ vanish.

Now consider a given initial function $u(x, 0) = u_0(x)$ in $L_2[0, 1]$ with Fourier series

$$u_0(x) = \sum_{k \in \mathbb{Z}} \alpha_k \varphi_k(x).$$

We already have found a solution for each of its terms. Combining these through the *superposition* principle then gives as solution for the periodic diffusion problem the Fourier series

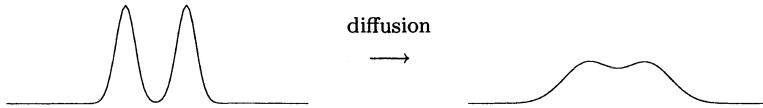
$$u(x, t) = \sum_{k \in \mathbb{Z}} \alpha_k e^{-4\pi^2 k^2 dt} \varphi_k(x).$$

Application of Parseval's identity shows that the L_2 -norm of $u(\cdot, t)$ is non-increasing in time,

$$\|u(\cdot, t)\|^2 = \sum_{k \in \mathbb{Z}} |\alpha_k e^{-4\pi^2 k^2 dt}|^2 \leq \sum_{k \in \mathbb{Z}} |\alpha_k|^2 = \|u(\cdot, 0)\|^2. \quad (1.17)$$

⁴⁾ Fourier's result was originally stated less precisely and it was initially not fully appreciated by his contemporaries, see Kline (1972, Ch. 28). Convergence of the series is to be understood in the L_2 sense. On the space $L_2[0, 1]$, two functions are identified if they differ only in isolated points (set of measure zero).

Note that the initial function is only required to be in $L_2[0, 1]$, so it does not have to be differentiable or even continuous. Therefore we have here in fact a generalized solution of the diffusion equation. The series expansion is very instructive as it shows that the solution $u(x, t)$ will become smoother and smoother for evolving time because high frequency modes, with high wave numbers k , are damped faster than modes with low wave numbers. This smoothing property of the diffusion equation is consistent with the physical interpretation of (1.13) as a heat flow or molecular diffusion describing Brownian motion of particles.



Next consider the advection equation (1.12) for $t > 0$ with the periodicity condition (1.10) and initial profile $u(x, 0) = \varphi_k(x)$ for some wave number k . Repeating the computation based on the separation of variables (1.15) we then find the solution

$$u(x, t) = e^{-2\pi i k a t} \varphi_k(x) = \varphi_k(x - at). \quad (1.18)$$

As we already saw for the general solution of (1.12), Fourier modes are now only shifted, they are not damped as with diffusion. Furthermore, all modes are shifted with the same velocity a . If $a > 0$ the modes travel to the right, if $a < 0$ to the left. Because wave-type functions $e^{2\pi i k(x-at)}$ with evolution in one direction are exact solutions of the equation, the advection equation (1.12) is also called the *one-way wave equation*.⁵⁾

For the advection-diffusion model problem (1.9), (1.10) with a Fourier mode as initial profile, $u(x, 0) = \varphi_k(x)$, we find as superposition of the previous cases

$$u(x, t) = e^{(-2\pi i k a - 4\pi^2 k^2 d)t} \varphi_k(x) = \underbrace{e^{-4\pi^2 k^2 d t}}_{\text{damping}} \underbrace{\varphi_k(x - at)}_{\text{shift}}.$$

Hence, for the advection-diffusion problem all Fourier modes are shifted with the same velocity and damped according to their frequency.

With an arbitrary initial profile $u_0(x) = \sum_k \alpha_k \varphi_k(x)$ in $L_2[0, 1]$, the solution of (1.9), (1.10) is given by

$$u(x, t) = \sum_{k \in \mathbb{Z}} \alpha_k e^{-4\pi^2 k^2 d t} \varphi_k(x - at), \quad (1.19)$$

⁵⁾ The (standard) wave equation is $u_{tt} = b^2 u_{xx}$ with wave-type solutions $e^{2\pi i k(x \pm bt)}$, so here waves traveling to the left and to the right occur simultaneously.

and so again the L_2 -norm of $u(\cdot, t)$ will be non-increasing in time for any $d \geq 0$, see (1.17). Note that the sign of the advection coefficient a merely decides the direction of the shift, but having $d \geq 0$ is essential. If d were negative, then all Fourier modes with $k \neq 0$ would be amplified for increasing time and for general initial values this leads to immediate blow-up. A solution could still be defined if the initial profile $u_0(x)$ is such that $\alpha_k = 0$ for $|k| \geq K$, but with arbitrary small perturbations on $u_0(x)$ the Fourier series can again have an infinite number of terms. A diffusion problem with a negative diffusion coefficient is *not well-posed*.

Fourier analysis is one of the corner stones of applied mathematics. Fourier modes are eigenfunctions of derivative operators and (as we will see later) of the corresponding finite difference operators, and therefore Fourier analysis has become extremely useful for examining solutions of partial differential equations and their approximating difference schemes.

Fourier decompositions can also be regarded for functions in $L_2(\mathbb{R})$ without the periodicity condition. Then the above Fourier series are to be replaced by Fourier integrals, see for instance Pinkus & Zafrany (1997) or Strikwerda (1989, Ch. 2). Finally we note that consideration of the complex function space L_2 and complex Fourier modes has been made only for the ease of presentation. Usually we will deal with real valued solutions of PDEs and instead of complex exponentials it is then also possible to use real expansions with sine and cosine functions.

Remark 1.2 If the function v is j times differentiable with $v^{(j)} \in L_2[0, 1]$ and $v(x) = \sum_k \alpha_k \varphi_k(x)$, then⁶⁾

$$|\alpha_k| \leq \frac{1}{|2\pi k|^j} \max_{0 \leq x \leq 1} |v^{(j)}(x)|.$$

In fact, if $v^{(j)}$ is piecewise differentiable one can show that $|\alpha_k| \sim |k|^{-(j+1)}$. Thus for smooth functions v the coefficients α_k are very small for large $|k|$.

Consider for any given function v on $[0, 1]$ the coefficients $\alpha_k = (\varphi_k, v)$ and the truncated Fourier series

$$v_K(x) = \sum_{k=-K}^K \alpha_k \varphi_k(x).$$

In the above discussion we have considered the function class $L_2[0, 1]$ and so writing $v(x) = \sum_{k \in \mathbb{Z}} \alpha_k \varphi_k(x)$ means that in the L_2 -norm we have

$$\|v - v_K\|^2 = \int_0^1 |v(x) - v_K(x)|^2 dx \rightarrow 0 \quad \text{as } K \rightarrow \infty.$$

By restricting the function class, one can also prove convergence of the Fourier series in a pointwise sense. For example, if v is periodic and differentiable with

⁶⁾ Exercise: Derive this inequality by considering the inner product of $v^{(j)}$ with φ_k . If $v^{(j)}$ is piecewise differentiable, refine the inequality by applying integration by parts.

a piecewise continuous first derivative, then the convergence will be uniform pointwise, see Pinkus & Zafraany (1997).

For smooth functions the truncated Fourier series gives good approximations already with relatively small values of K , but near discontinuities or sharp gradients we get an oscillatory approximation. This is called the *Gibbs phenomenon*. An illustration is given in Figure 1.2 for a block function and a hat function. The truncated series are plotted for $K = 2, 4, 8, 16$. The block function and hat function itself are represented by grey lines. With the hat function the truncated series with $K = 8, 16$ are already hard to distinguish from the actual function, except near the corner points.

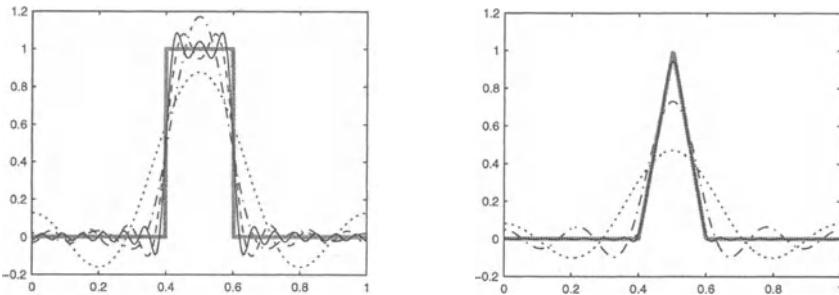


Fig. 1.2. Truncated Fourier series with block function and hat function for $K = 2$ (dotted), $K = 4$ (dash-dots), $K = 8$ (dashed) and $K = 16$ (solid).

By direct calculation of the Fourier coefficients (left as exercise) it follows here that $|\alpha_k| \leq |\pi k|^{-1}$ for the block function and $|\alpha_k| \leq |\pi k|^{-2}$ for the hat function. \diamond

Remark 1.3 Linear PDEs in 1D with constant coefficients, such as our advection and diffusion equation, admit wave-type functions

$$u(x, t) = e^{2\pi i(kx - ct)}$$

as solution. Here k is commonly called the *wave number* and the number c , multiplying the time variable, the *frequency*. The sort of equation determines the dependence $c = c(k)$ of the frequency on the wave number. This is called the *dispersion relation* for the equation. For the diffusion equation $u_t = du_{xx}$ we have found the imaginary values $c = -2\pi idk^2$. Dispersion relations are mostly regarded for advection equations and for $u_t + au_x = 0$ we have $c = ak$, which is just a restatement of the fact that all Fourier modes are traveling with the same velocity $c/k = a$. \diamond

1.3 Multi-dimensional Problems

In applications one is mainly interested in higher space dimensional problems with velocities and diffusion coefficients which vary in space and time.

Suppressing the dependence on the time and space variables, the variable-coefficient 3D counterpart of (1.9) reads

$$u_t + (a_1 u)_x + (a_2 u)_y + (a_3 u)_z = (d_1 u_x)_x + (d_2 u_y)_y + (d_3 u_z)_z. \quad (1.20)$$

This equation can be derived as in the 1D case illustrated in the beginning of this section by considering mass balances over small cubes.

For brevity we will often use the classical vector analysis notation, as can be found for instance in Apostol (1964). Let $\nabla = (\partial_x, \partial_y, \partial_z)^T$ be the gradient operator and denote by $\underline{a} \cdot \underline{b}$ the standard Euclidean inner product of vectors $\underline{a}, \underline{b} \in \mathbb{R}^3$. Then the *gradient (vector)* of a differentiable scalar function u on \mathbb{R}^3 is defined by

$$\text{grad } u \equiv \nabla u = (u_x, u_y, u_z)^T,$$

and the *divergence* of a differentiable vector function $\underline{a} = (a_1, a_2, a_3)^T$ is given by

$$\text{div } \underline{a} \equiv \nabla \cdot \underline{a} = \frac{\partial a_1}{\partial x} + \frac{\partial a_2}{\partial y} + \frac{\partial a_3}{\partial z}.$$

Further $\Delta = \nabla \cdot \nabla$ stands for the *Laplace operator*,

$$\Delta u = u_{xx} + u_{yy} + u_{zz}.$$

With these definitions the multi-dimensional advection-diffusion equation (1.20) can be rewritten in the shorter form

$$u_t + \nabla \cdot (\underline{a} u) = \nabla \cdot (D \nabla u), \quad (1.21)$$

where D is the diagonal matrix $D = \text{diag}(d_1, d_2, d_3)$.

The diffusion equation

$$u_t = \nabla \cdot (D \nabla u) \quad (1.22)$$

is called the multi-dimensional heat equation. Sometimes D is a full matrix to cater for so-called cross-diffusion terms. More often, D is a scalar, in which case the amount of diffusion is the same in all coordinate directions; if D is constant the equation will then be written as $u_t = D \Delta u$. With constant coefficients and periodicity conditions Fourier decompositions can be derived similar as in one space dimension.

Since it underlies the mass conservation law, the advection equation

$$u_t + \nabla \cdot (\underline{a} u) = 0 \quad (1.23)$$

is said to be in *conservative form* (also called *conservation* or *flux* form). From the identity

$$\nabla \cdot (\underline{a} u) = u \nabla \cdot \underline{a} + \underline{a} \cdot \nabla u,$$

it follows that we can bring (1.23) in the so-called *advective form*

$$u_t + \underline{a} \cdot \nabla u = 0 \quad (1.24)$$

if the vector field \underline{a} is *divergence-free*, that is if

$$\nabla \cdot \underline{a} = 0. \quad (1.25)$$

In many applications this divergence-free condition holds.

An advantage of the advective form (1.24) is that, similar as for the 1D problem (1.12), it admits the characteristic solution approach. If we define the characteristics $(\underline{\xi}(t), t)$ in the $(\underline{x}, t) = (x, y, z, t)$ space to be solutions of the ordinary differential equation

$$\frac{d}{dt} \underline{\xi}(t) = \underline{a}(\underline{\xi}(t), t),$$

then it follows that

$$\frac{d}{dt} u(\underline{\xi}(t), t) = 0,$$

and hence the solution $u(\underline{x}, t)$ is constant along the characteristics.

Remark 1.4 Also with a velocity field that is not divergence-free, both the conservative and the advective form have a physical meaning. Consider a fluid with density $\rho > 0$ and a dissolved chemical species with concentration u . Then we can define the *mixing ratio* $v = u/\rho$, which is a dimensionless number (like a percentage). By mass conservation it follows that

$$\rho_t + \nabla \cdot (\underline{a}\rho) = 0, \quad u_t + \nabla \cdot (\underline{a}u) = 0.$$

Writing the last equation in terms of $u = v\rho$ we obtain

$$(v_t + \underline{a} \cdot \nabla v)\rho + v(\rho_t + \nabla \cdot (\underline{a}\rho)) = 0,$$

and thus we see that the mixing ratio v satisfies the advective form

$$v_t + \underline{a} \cdot \nabla v = 0.$$

If the fluid is incompressible, that is, ρ is constant, then we have $\nabla \cdot \underline{a} = 0$ and u/v is constant.

With a velocity field $\underline{a}(\underline{x}, t)$ that is not divergence-free, the advective form is no longer mass conservative. As we will see later on, numerical methods based on the advective form are in general not mass conservative even if the velocity field is divergence-free. Since chemical reactions are mostly defined in terms of concentrations we will usually deal with the conservative form of the advection equation. \diamond

Boundary Conditions

Periodicity conditions, such as (1.10), are mainly considered for theoretical and test purposes. In general we will deal with an open, bounded spatial region Ω and on the boundary $\Gamma = \partial\Omega$ it will be assumed that u satisfies

appropriate boundary conditions. A partial differential equation with given initial and boundary conditions is called an *initial-boundary value problem*.

Let \underline{n} be the outward normal vector on boundary points $\underline{x} \in \Gamma$. The set of boundary points for which $\underline{n} \cdot \underline{a} < 0$ is called the *inflow boundary*. Consider a partitioning of Γ into Γ_D, Γ_N where Γ_D contains the inflow boundary, and let γ_D, γ_N be given functions on the corresponding boundary parts. Typical boundary conditions are the *Dirichlet condition*

$$u = \gamma_D \quad \text{on } \Gamma_D, \quad (1.26)$$

and the *Neumann condition*

$$\underline{n} \cdot \nabla u = \gamma_N \quad \text{on } \Gamma_N. \quad (1.27)$$

Occasionally we can also have on a part $\Gamma_M \subset \Gamma$ a *mixed condition*, also called *Robin condition*, for instance

$$\underline{n} \cdot (\underline{a}u - D\nabla u) = \gamma_M \quad \text{on } \Gamma_M. \quad (1.28)$$

Concrete examples and general remarks will follow later on. Here we mention already that for the pure advection equation (1.23) specification of the values on the inflow boundary is sufficient. With non-zero diffusion coefficients conditions on the other boundary parts must also be specified.

Advection-Diffusion-Reaction Systems

If we put the advection-diffusion equation (1.21) and the general reaction system (1.5) together, we get the general advection-diffusion-reaction system

$$\frac{\partial}{\partial t} u_j + \nabla \cdot (\underline{a}_j u_j) = \nabla \cdot (D_j \nabla u_j) + f_j(u), \quad j = 1, 2, \dots, s. \quad (1.29)$$

Here $\underline{a}_j = (a_{1j}, a_{2j}, a_{3j})^T$ is the velocity field for the species concentration u_j and likewise D_j stands for the corresponding diffusion matrix. The possible explicit dependence of \underline{a}_j, D_j and f_j on \underline{x} and t is suppressed in the notation. Recall that $u = (u_1, \dots, u_s)^T$ now represents a vector of species concentrations and the s equations are coupled through the nonlinear chemistry part. This chemistry part is usually extended with source and sink terms, and the system of course also needs initial and boundary conditions. If the velocity field and diffusion coefficients are the same for all species we will write (1.29) also as

$$u_t + \nabla \cdot (\underline{a}u) = \nabla \cdot (D \nabla u) + f(u), \quad (1.30)$$

where the spatial operators should still be interpreted component-wise. Without chemistry the system becomes uncoupled and u may be read as a scalar quantity. It is also possible in (1.29) that \underline{a}_j and D_j depend on the species vector u . Then the coupling extends to the advection-diffusion part.

The numerical solution of these coupled systems is the main topic studied in this book. The number of unknowns in a numerical simulation can become very large. In particular for 3D equations the design of highly efficient algorithms is a prerequisite to reduce CPU times to feasible levels, even for high-performance massively parallel computers.

1.4 Examples of Applications

Applications of the general advection-diffusion-reaction system (1.29) are numerous. Here we briefly consider three examples from environmental studies, biology and chemistry.

Pollutant Transport-Chemistry Models

Many mathematical environmental studies use PDE systems of type (1.30) to model pollutant transport in the atmosphere, groundwater and surface water. The space-time dependent velocities \underline{a} of the transport medium (water or air) are either given in a data archive or computed alongside by solving flow equations, such as the atmospheric flow equations used in weather forecast and climate models or the equations describing flows in porous media. These flow equations are nonlinear but again advection and diffusion are of primary importance. The diffusion coefficient matrices D are mostly the result of parameterizations of sub-grid scale phenomena that cannot be resolved on practical grids. These coefficients are usually constructed by the modellers and may include for instance parameterizations of turbulence.

In real-life atmospheric air pollution models the number of species concentrations u_j may be quite large. As many as about 100 species (trace gases) are nowadays used in the study of air pollution caused by anthropogenic emissions. This large number makes air pollution modelling highly expensive in computer usage, both with respect to memory and CPU. The monograph of Zlatev (1995) gives an account of the state of the art of numerical atmospheric air pollution modelling. A recent survey on numerical methods for atmospheric air pollution models focusing on advection, diffusion and chemistry computations and high performance computing is Verwer, Blom & Hundsdorfer (2002).

Chemo-taxis Problems from Mathematical Biology

The analysis and computation of solutions of PDEs from mathematical biology is of increasing importance for the understanding of biological processes, for the verification of hypotheses about the underlying biology and also for the application of such models to patient specific data in medicine. The complexity of the models nearly always necessitates the application of efficient

numerical methods. Interesting problems, where similar as in (1.29) advection, diffusion and bio-chemistry occurs are so-called (chemo-)taxis problems. These problems take the form of advection-diffusion-reaction systems

$$\begin{aligned} \rho_t + \nabla \cdot \left(\rho \sum_{i=1}^l f_i(c) \nabla c_i \right) &= \varepsilon \Delta \rho + f_0(\rho, c), \\ c_t &= D \Delta c + g(\rho, c), \end{aligned} \quad (1.31)$$

where ρ denotes the density of a cell population and c is a vector of l concentrations or densities of certain bio-chemicals.

A characteristic property of (1.31) is that the evolution of ρ depends on gradients ∇c_i of the chemical concentrations – a process known as chemo-taxis. Consequently, these equations are nonlinearly coupled not only in the reaction part but also in the advective chemo-taxis part in the population equation. When considering c as being given, we have linear advection in the population equation,

$$\rho_t + \nabla \cdot (\underline{a} \rho) = 0, \quad \underline{a} = \sum_{i=1}^l f_i(c) \nabla c_i,$$

if $\varepsilon = 0$, $f_0 = 0$. The functions $f_i(c)$, $i = 1, 2, \dots, l$, describe the strength and the sign of the tactic influence of each chemical c_i on the population density ρ .

- The reaction term $f_0(\rho, c)$ in (1.31) accounts for creation or loss of entities in the population. The chemical concentrations in c can also change by diffusion (D is a diagonal matrix with non-negative entries D_i), or be spatially immobile (if $D_i = 0$). Finally, reactions between the concentrations and the population density leading to a change in c are modelled through the vector-valued function $g(\rho, c)$.

Specific applications of this chemo-taxis model concern tumour invasion, tumour angiogenesis and bacterial pattern formation, see for instance Anderson et al. (2000), Chaplain & Stuart (1993) and Tyson et al. (1999). A recent numerical study is Gerisch & Verwer (2002), see also references therein. An interesting numerical case arises when the diffusion in the population equation, governed by the coefficient ε , is much smaller than the speed of migration induced by the taxis term or when there is no diffusion in the population at all. This situation may lead to steep moving fronts for the population which in general are numerically difficult.

As an illustration we consider the 1D model for tumour angiogenesis from Chaplain & Stuart (1993). Angiogenesis is the process of blood vessel development. The model describes the case that this process is induced by a tumour which aims to establish a connection to the blood network – and hence nutrient supply – in order to be able to grow further. The model has two components. The first, ρ , is the concentration of endothelial cells which line

the blood vessels, and hence ρ is a measure of the density of the developing network. The second, c , is the concentration of a tumour angiogenesis factor (TAF) which is secreted by the tumour and stimulates blood vessel growth. With the scaling $x \in [0, 1]$ the two PDEs are given by

$$\begin{aligned}\rho_t &= \varepsilon \rho_{xx} - (\kappa c_x \rho)_x + \mu \rho(1 - \rho) \max(0, c - c^*) - \beta \rho, \\ c_t &= \delta c_{xx} - \lambda c - \frac{\alpha \rho c}{\gamma + c},\end{aligned}\tag{1.32}$$

describing chemo-taxis (advection) for ρ and diffusion, losses and biochemical reactions for ρ and c .

Following Chaplain & Stuart (1993) we simulate the following set-up. At the initial time $t = 0$ a tumour is located at $x = 0$, given by the initial condition for the TAF concentration $c(x, 0) = \cos(\frac{1}{2}\pi x)$. At $t = 0$ a blood vessel is located at $x = 1$, as given by the initial concentration

$$\rho(x, 0) = \begin{cases} 0 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x = 1. \end{cases}$$

Dirichlet boundary conditions are used which match these initial conditions,

$$\rho(0, t) = 0, \rho(1, t) = 1 \quad \text{and} \quad c(0, t) = 1, c(1, t) = 0,$$

and the parameter values are

$$\varepsilon = 10^{-3}, \delta = 1, \alpha = 10, \beta = 4, \gamma = 1, \kappa = 0.75, \lambda = 1, \mu = 100, c^* = 0.2.$$

These non-dimensionalized initial and boundary functions and parameter values are the same as used in Chaplain & Stuart (1993). With the above choices we always have $c_x < 0$, and thus the boundary conditions for ρ at $x = 1$ and $x = 0$ act, respectively, as inflow and outflow conditions in the taxis part. The simulation time for the described set-up is $T = 0.7$. For much longer times the assumptions underlying the model do not hold anymore

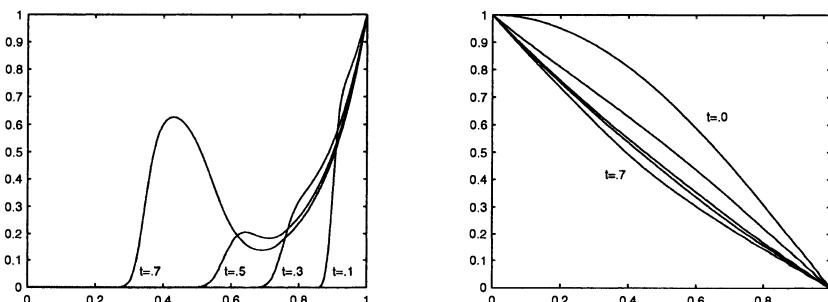


Fig. 1.3. Solutions ρ (left), c (right) of the tumour angiogenesis problem for $\delta = 1$.

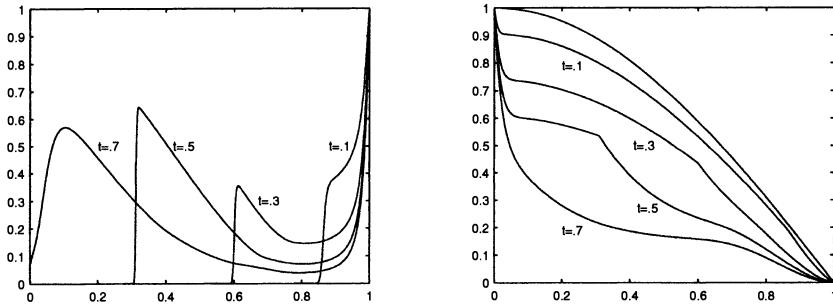


Fig. 1.4. Solutions ρ (left), c (right) of the tumour angiogenesis problem for $\delta = 10^{-3}$.

because the blood vessel has then reached the tumour and other processes take over.

For diffusion coefficient $\delta = 1$, Figure 1.3 shows $\rho(x, t)$ and $c(x, t)$ as functions of x at time $t = 0, 0.1, 0.3, 0.5, 0.7$. We see that the TAF concentration c is taken up while more or less retaining its initial smoothness. On the other hand the blood vessel density ρ develops a left traveling front moving up the gradient of the TAF concentration. In the course of time the front smears out, collapses and then grows. The initial collapse and subsequent growth is governed by the term containing threshold c^* in the equation for ρ (no cell proliferation takes place in the beginning because the TAF concentration at the cells is below the threshold). For $t = 0.7$ we still have zero values near $x = 0$, showing that the front region has not yet reached the left boundary.

Using a smaller diffusion coefficient $\delta = 0.001$ leads locally to a stronger uptake of TAF and a less smooth concentration profile, which in turn results in a significantly steeper profile for ρ , see Figure 1.4. The solution behaviour is roughly the same as with $\delta = 1$, but now the blood vessel front first smears out and collapses, then grows and steepens up, and smears out again. The intermediate steepening is caused by the sudden change in the slope of c (as visible near $(x, t) = (0.3, 0.5)$ in the right plot). Also the speed of the front has become larger; it has now reached the left boundary at time $t = 0.7$.

Pattern Formation with Reaction-Diffusion Equations

Reaction-diffusion equations can exhibit a large variety of solutions that are of interest to applied mathematical analysts and to modellers from physics, chemistry and biology. A simple model with interesting dynamics is the *Gray-Scott model* described in Pearson (1993). The model involves two chemical species, with concentrations u and v , and we consider it here in two spatial dimensions. The equations read

$$\begin{aligned} u_t &= D_1 \Delta u - uv^2 + \gamma(1-u), \\ v_t &= D_2 \Delta v + uv^2 - (\gamma + \kappa)v. \end{aligned} \tag{1.33}$$

We take the spatial region $0 \leq x, y \leq 2.5$ and $t \geq 0$. At the boundaries a homogeneous Neumann condition is imposed for both u and v . The initial value is

$$u(x, y, 0) = 1 - 2v(x, y, 0),$$

$$v(x, y, 0) = \begin{cases} \frac{1}{4} \sin^2(4\pi x) \sin^2(4\pi y) & \text{if } 1 \leq x, y \leq 1.5, \\ 0 & \text{elsewhere.} \end{cases}$$

The v -component thus consists initially of four spots near the center of the domain. The parameters are taken as $D_1 = 8 \cdot 10^{-5}$, $D_2 = 4 \cdot 10^{-5}$, $\gamma = 0.024$ and $\kappa = 0.06$. With these parameters and initial values the solution gives repeated replication of the initial spots. Figure 1.5 shows the time evolution of the v -component by contour lines in the (x, y) -plane at various times; for clarity the solutions are only displayed for $0.5 \leq x, y \leq 2$.

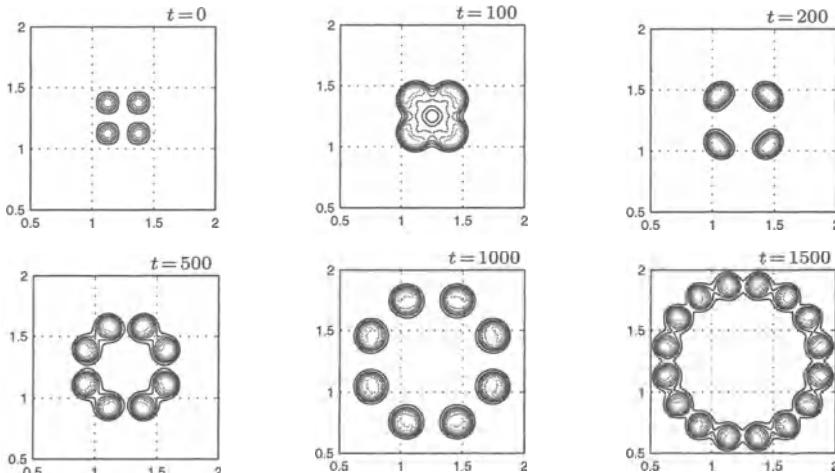


Fig. 1.5. Time evolution (v -component) in the Gray-Scott model.

Other patterns, with moving stripes for example, are obtained with different choices of parameters and initial values. Related reaction-diffusion models will be used in later chapters for numerical illustrations. Many more examples from biology and biochemistry can be found in Murray (1989).

The common aspect of these pattern formation problems is the subtle interplay between diffusion and reactions. To obtain qualitatively correct solutions a fine spatial and temporal resolution may be needed. In particular for 3-dimensional problems this poses a challenging computational task.

2 Basic Discretizations for ODEs

In the numerical solution of advection-diffusion-reaction problems it is often the reaction part that is the most expensive in terms of CPU time. As explained in the previous section, this reaction part is formed by systems of ordinary differential equations (ODEs). In this section we will briefly outline some simple methods and main concepts in the numerical solution of initial value problems for ODEs. More advanced methods will be presented in Chapter II.

2.1 Initial Value Problems and Euler's Method

Initial Value Problems for ODEs

The general formulation of an initial value problem for a system of ODEs is

$$w'(t) = F(t, w(t)), \quad t > 0, \quad w(0) = w_0, \quad (2.1)$$

with given $F : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ and $w_0 \in \mathbb{R}^m$. Such systems not only arise from reaction equations, but, as we will see in the next sections, the full advection-diffusion-reaction equations also lead to such ODE systems if the spatial derivatives are approximated by difference quotients. So in general the dimension of this ODE system can be very much larger than the number of species in equation (1.3).

First let us recall some basic facts on the existence and uniqueness of solutions. Let $\|\cdot\|$ be a vector norm on \mathbb{R}^m and consider a time interval $[0, T]$. For $K_0 > 0$, let the cylinder \mathcal{C}_0 be given by

$$\mathcal{C}_0 = \{(t, v) \in \mathbb{R} \times \mathbb{R}^m : 0 \leq t \leq T, \|v - w_0\| \leq K_0\},$$

and consider the condition

$$\|F(t, \tilde{v}) - F(t, v)\| \leq L \|\tilde{v} - v\| \quad \text{for all } (t, \tilde{v}), (t, v) \in \mathcal{C}_0. \quad (2.2)$$

A condition of this type is called a *Lipschitz condition* with L the Lipschitz constant. If F is continuous on \mathcal{C}_0 then (2.1) has a solution on some interval $[0, T^*]$ with $T^* > 0$. If $\|F(t, v)\| \leq M_0$ on \mathcal{C}_0 we can take $T^* = \min(T, K_0/M_0)$. If F satisfies a Lipschitz condition on \mathcal{C}_0 then the solution is unique. Moreover, if the function F is q times differentiable, the solution w will be $q + 1$ times differentiable on $[0, T^*]$. Proofs of these statements can be found in many textbooks on ODEs, for instance in Coddington & Levinson (1955), Coppel (1965), and in the first chapter of the numerical ODE textbook of Hairer, Nørsett & Wanner (1993).

Example 2.1 To show the relevance of the Lipschitz condition we consider the scalar problem

$$w'(t) = -3|w(t)|^{2/3}, \quad w(0) = 1. \quad (2.3)$$

A solution is given by $w(t) = (1-t)^3$ and this is the unique solution up to $t = 1$. However we have $w(1) = 0$ and the function $F(v) = -3|v|^{2/3}$ does not satisfy a Lipschitz condition around $v = 0$ due to the fact that there $F'(v)$ becomes unbounded. Indeed, after $t = 1$ the solution is no longer unique; for any $s \geq 1$ the continuously differentiable function

$$w(t) = \begin{cases} (1-t)^3 & \text{for } 0 \leq t \leq 1, \\ 0 & \text{for } 1 \leq t \leq s, \\ (s-t)^3 & \text{for } t \geq s \end{cases}$$

is also a solution of the initial value problem. \diamond

Euler's Method

In the following, we consider numerical approximations w_n to the exact solution values $w(t_n)$ at the points $t_n = n\tau$, $n = 0, 1, 2, \dots$ with $\tau > 0$ being the step size. For simplicity this step size τ is taken constant. Convergence properties of the numerical schemes will only be considered on bounded time intervals $[0, T]$. If the exact solution $w(t)$ is q times continuously differentiable on $[0, T]$ we will write $w \in C^q[0, T]$.

The most simple numerical method for solving the initial value problem is Euler's method

$$w_{n+1} = w_n + \tau F(t_n, w_n), \quad n = 0, 1, 2, \dots . \quad (2.4)$$

This method can be seen to result from truncating the Taylor series

$$w(t_{n+1}) = w(t_n) + \tau w'(t_n) + \mathcal{O}(\tau^2)$$

after the first derivative term. We will show that Euler's method does convergence to the exact solution on any bounded time interval $[0, T]$ if F satisfies a Lipschitz condition.

To obtain a first indication of the error with Euler's method, we insert the exact solution values in the Euler scheme to get

$$w(t_{n+1}) = w(t_n) + \tau F(t_n, w(t_n)) + \tau \rho_n \quad (2.5)$$

with a residual $\tau \rho_n$. The quantity ρ_n will be called the (*local*) *truncation error*. Using the fact that $F(t_n, w(t_n)) = w'(t_n)$ it follows from the Taylor series expansion that

$$\rho_n = \frac{1}{2} \tau w''(t_n) + \mathcal{O}(\tau^2), \quad (2.6)$$

provided $w \in C^3[0, T]$.

This smoothness requirement on the exact solution can be relaxed. If $w \in C^1[0, T]$ then

$$w(t_{n+1}) - w(t_n) = \tau \int_0^1 w'(t_n + \sigma \tau) d\sigma,$$

according to the mean-value theorem, see for instance Ortega & Rheinboldt (1970), and hence

$$\rho_n = \int_0^1 (w'(t_n + \sigma\tau) - w'(t_n)) d\sigma.$$

Now if $w \in C^2[0, T]$, we can apply the mean-value theorem to w' to obtain

$$\|w'(t_n + \sigma\tau) - w'(t_n)\| \leq \sigma\tau \max_{t_n \leq s \leq t_{n+1}} \|w''(s)\|,$$

which leads to

$$\|\rho_n\| \leq \frac{1}{2}\tau \max_{t_n \leq s \leq t_{n+1}} \|w''(s)\|. \quad (2.7)$$

The truncation error is only an indication for the actual error that can be expected. To study the *global discretization error*

$$\varepsilon_n = w(t_n) - w_n$$

for $n \geq 0$, we subtract (2.4) from (2.5) to obtain the recursion

$$w(t_{n+1}) - w_{n+1} = w(t_n) - w_n + \tau(F(t_n, w(t_n)) - F(t_n, w_n)) + \tau\rho_n.$$

Using the Lipschitz condition (2.2) it is seen that

$$\|\varepsilon_{n+1}\| \leq (1 + \tau L)\|\varepsilon_n\| + \tau\|\rho_n\|.$$

The error ε_n after n steps thus satisfies

$$\|\varepsilon_n\| \leq \kappa^n \|\varepsilon_0\| + \sum_{j=0}^{n-1} \kappa^j \tau \|\rho_{n-1-j}\|, \quad \kappa = 1 + \tau L.$$

Using $\kappa < e^{\tau L}$ and $\sum_{j=0}^{n-1} \kappa^j = (\kappa^n - 1)/(\kappa - 1)$, this gives the estimate

$$\|\varepsilon_n\| \leq e^{\tau L n} \|\varepsilon_0\| + \frac{1}{L} (e^{\tau L n} - 1) \max_{0 \leq j < n} \|\rho_j\|. \quad (2.8)$$

With this estimate of the global errors, convergence of Euler's method on an interval $[0, T]$ can now be established. Insertion of (2.7) into (2.8) directly gives the following result that shows first-order convergence of Euler's method.

Theorem 2.2 *Consider problem (2.1) with solution $w \in C^2[0, T]$ and assume that F satisfies a Lipschitz condition (2.2) on $[0, T] \times \mathbb{R}^m$. Let $w_0 = w(0)$. Then the errors of the Euler approximations satisfy*

$$\|w(t_n) - w_n\| \leq \frac{1}{2}\tau K \max_{0 \leq t \leq T} \|w''(t)\|$$

for $\tau \rightarrow 0$, uniformly for $t_n \in [0, T]$, with constant $K = \frac{1}{L}(e^{\tau L T} - 1)$. □

This convergence result remains valid if the function F satisfies a Lipschitz condition on some tube around the exact solution, instead of the whole $[0, T] \times \mathbb{R}^m$; see for instance Hairer et al. (1993, Sect.I.7). If we are only interested in convergence and not in the order, then the smoothness assumption $w \in C^2[0, T]$ can be relaxed. What we need is an estimate $\|\rho_n\| = o(1)$ as $\tau \rightarrow 0$, uniformly for $0 \leq t_n \leq T$. If $w \in C^1[0, T]$, for which continuity of F is sufficient, then w' is uniformly continuous on $[0, T]$, and from the derivation leading to (2.7) it then follows that the truncation error $\|\rho_n\|$ can be bounded by an arbitrary small number if we take τ small enough.

When proving convergence of a numerical scheme we will often not specify precise smoothness assumptions. Usually it will be tacitly assumed that the solution is smooth enough for occurring derivatives to exist and that these derivatives are bounded by some moderate constant.

The derivation of Theorem 2.2 relies on two ingredients: firstly, we have an estimate for the truncation errors (*consistency*), and secondly, we have an estimate that tells us that small local errors will lead to small global errors (*stability*). The framework

$$\text{consistency \& stability} \implies \text{convergence}$$

is the standard way to prove convergence of numerical approximations and we will see many more instances in this chapter.

The constant K in Theorem 2.2 contains the exponential of LT and therefore the error bound becomes practically meaningless if LT is very large. The following example shows that this is not due to an over-estimation in the analysis of the errors of Euler's method, but that indeed the errors themselves can become large with increasing Lipschitz constant L .

Example 2.3 Consider the two-way chemical reaction $W_1 \xrightarrow{k_1} W_2 \xrightarrow{k_2} W_1$, giving an ODE system

$$\begin{aligned} w'_1(t) &= -k_1 w_1(t) + k_2 w_2(t), \\ w'_2(t) &= k_1 w_1(t) - k_2 w_2(t), \end{aligned} \tag{2.9}$$

with reaction constants $k_1, k_2 > 0$. In vector form we write this as $w'(t) = Aw(t)$ with matrix

$$A = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

The exact solution is given by ⁷⁾

$$\begin{aligned} w_1(t) &= \frac{k_2}{k_1 + k_2} (w_1(0) + w_2(0)) + \frac{e^{-(k_1+k_2)t}}{k_1 + k_2} (k_1 w_1(0) - k_2 w_2(0)), \\ w_2(t) &= \frac{k_1}{k_1 + k_2} (w_1(0) + w_2(0)) - \frac{e^{-(k_1+k_2)t}}{k_1 + k_2} (k_1 w_1(0) - k_2 w_2(0)). \end{aligned}$$

⁷⁾ Exercise: Derive this formula from the decomposition $A = U \Lambda U^{-1}$ with diagonal Λ by setting $v(t) = U^{-1}w(t)$.

The initial values are taken as $w_1(0) = 0.1$, $w_2(0) = 0.9$ and the end point is $T = 1$. Initially there will be a rapid change in the solution, the so-called transient phase, but after a short while the term $e^{-(k_1+k_2)t}$ becomes negligible and then the solution settles in a steady state. In the max-norm it is easy to show that the Lipschitz constant for $F(t, v) = Av$ can be taken as $L = k_1 + k_2$.

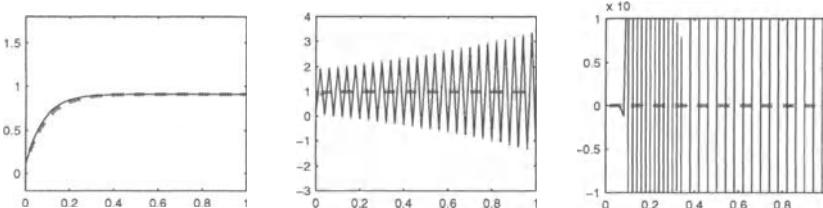


Fig. 2.1. Results for Euler's method with step size $\tau = 1/50$ and $k_2 = 10$ (left), $k_2 = 100$ (middle), $k_2 = 1000$ (right). The exact solution is indicated with dashed grey lines.

To illustrate the deterioration of the numerical result with Euler's method if L becomes large, in Figure 2.1 the numerical approximations for the first component are plotted with step size $\tau = \frac{1}{50}$, $k_1 = 1$ and with increasing k_2 . Note that the vertical scaling in the three pictures is different; the actual range of the numerical solution in the right picture is 10^{127} , close to overflow. It is clear from these pictures that Euler's method cannot be used with this step size if the Lipschitz constant L is too large. ◇

Before we consider other numerical methods that are suited for problems of the above type with large Lipschitz constants, we first take a look at properties of the differential equations. Particular attention will be given to norm estimates for linear equations $w'(t) = Aw(t) + g(t)$ with $m \times m$ matrix A .

2.2 Norms and Matrices

In this subsection we list some definitions and basic properties of normed vector spaces and matrices that are used throughout the text. We will consider the complex vector space \mathbb{C}^m with complex matrices in $\mathbb{C}^{m \times m}$; the corresponding real spaces \mathbb{R}^m and $\mathbb{R}^{m \times m}$ are considered as subspaces.⁸⁾ Good references for matrix theory are the textbooks of Horn & Johnson (1985) and Lancaster & Tismenetsky (1985). Formal definitions and proofs of the statements made below can be found in these books. The topic of matrix computations, also called numerical linear algebra, will not be treated here. A well-known textbook in this field is Golub & van Loan (1996).

⁸⁾ In our applications to linear differential equations $w'(t) = Aw(t)$, the matrix A will usually be real. The eigenvalues and eigenvectors however will be complex in general, so a decomposition $A = U\Lambda U^{-1}$ will lead to $U, \Lambda \in \mathbb{C}^{m \times m}$.

Consider the vector space \mathbb{C}^m and let $h = 1/m$. Vector norms that will be frequently used are the discrete L_p -norms with $p = 1, 2$ or ∞ ,

$$\|v\|_1 = h \sum_{j=1}^m |v_j|, \quad \|v\|_2 = \left(h \sum_{j=1}^m |v_j|^2 \right)^{1/2}, \quad \|v\|_\infty = \max_{1 \leq j \leq m} |v_j| \quad (2.10)$$

for $v = (v_1, v_2, \dots, v_m)^T \in \mathbb{C}^m$.⁹⁾ Note that $\|v\|_1 \leq \|v\|_2 \leq \|v\|_\infty$ for any vector v . The L_2 -norm is generated by the inner product

$$\langle u, v \rangle_2 = h \sum_{j=1}^m \bar{u}_j v_j, \quad \|v\|_2^2 = \langle v, v \rangle_2. \quad (2.11)$$

Given a vector norm, the induced matrix norm for $m \times m$ matrices B is defined as

$$\|B\| = \max_{v \neq 0} \frac{\|Bv\|}{\|v\|}. \quad (2.12)$$

Throughout this text, matrix norms are supposed to be induced norms. Then, by definition, the identity matrix I always has norm equal to one. Obviously, for any matrix $A \in \mathbb{C}^{m \times m}$ and vector $v \in \mathbb{C}^m$ it holds that $\|Av\| \leq \|A\| \|v\|$. Further, for any two $m \times m$ matrices A and B the inequality $\|AB\| \leq \|A\| \|B\|$ is valid. The *condition number* of $B \in \mathbb{C}^{m \times m}$ is defined as $\text{cond}(B) = \|B\| \|B^{-1}\|$.

If $B = (b_{jk})$, with row index j and column index k , then $B^* = (\bar{b}_{kj})$ denotes the Hermitian adjoint of B . If B is real this is the same as the transpose B^T . The set of eigenvalues of B is denoted by $\sigma(B)$ and is called the *spectrum* of B . The *spectral radius* of B , $\max\{|\lambda| : \lambda \in \sigma(B)\}$, is denoted by $\rho(B)$. For any norm we have $\rho(B) \leq \|B\|$. The matrix norms associated to the vector norms (2.10) are

$$\begin{aligned} \|B\|_1 &= \max_{1 \leq k \leq m} \sum_{j=1}^m |b_{jk}|, \\ \|B\|_2 &= \sqrt{\rho(B^*B)}, \\ \|B\|_\infty &= \max_{1 \leq j \leq m} \sum_{k=1}^m |b_{jk}|. \end{aligned} \quad (2.13)$$

The L_2 -norm of a matrix is also called the *spectral norm*. As for vectors, we also have for matrices the Hölder inequality

$$\|B\|_2 \leq \sqrt{\|B\|_1 \|B\|_\infty}, \quad (2.14)$$

⁹⁾ If $v_j = \varphi(jh)$ with a given function $\varphi : [0, 1] \rightarrow \mathbb{R}$, then $\|v\|_p$ approximates the L_p function-norm of φ . If we use the scaling $h = 1$ in (2.10), then the corresponding norms are called the l_p -norms on \mathbb{C}^m .

which follows from

$$\|B\|_2^2 = \rho(B^*B) \leq \|B^*\|_\infty \|B\|_\infty = \|B\|_1 \|B\|_\infty.$$

A vector norm is called *absolute* if $\|u\| = \|v\|$ for any two vectors whose components have equal modulus, that is, $|u_i| = |v_i|$ for all i . This is equivalent with the property

$$\|\Lambda\| = \max_{1 \leq j \leq m} |\lambda_j| \quad \text{for any diagonal matrix } \Lambda = \text{diag}(\lambda_j). \quad (2.15)$$

The above L_p -norms are absolute.

If $B^* = B$ the matrix is an *Hermitian* matrix. A real Hermitian matrix, $B^T = B$, is called *symmetric*. The matrix B is *skew-Hermitian* if $B^* = -B$. If B is real and $B^T = -B$ it is called *skew-symmetric*.

The Hermitian matrix G is said to be *positive definite* if $\langle v, Gv \rangle_2 > 0$ for all non-zero $v \in \mathbb{C}^m$; likewise G is called non-negative definite if $\langle v, Gv \rangle_2 \geq 0$ for all $v \in \mathbb{C}^m$. Any inner product on \mathbb{C}^m can be associated with a positive definite matrix $G = H^*H$, H nonsingular,

$$\langle u, v \rangle = u^*Gv.$$

The corresponding vector norm $\|v\| = \sqrt{\langle v, v \rangle} = \sqrt{(Hv)^*Hv} = \|Hv\|_2$ is absolute if G is diagonal. For any inner product we have the Cauchy-Schwarz inequality

$$|\langle u, v \rangle| \leq \|u\| \|v\| \quad \text{for all } u, v \in \mathbb{C}^m.$$

The matrix B is said to be *unitary* if $B^*B = I$. A unitary matrix B is nonsingular and $B^{-1} = B^*$. This implies that

$$\|Bv\|_2 = \|v\|_2 = \|B^{-1}v\|_2$$

for any vector $v \in \mathbb{C}^m$, and in particular we have

$$\|B\|_2 = 1, \quad \text{cond}_2(B) = \|B\|_2 \|B^{-1}\|_2 = 1.$$

A real unitary matrix, $B^T B = I$, is called *orthogonal*.

The matrix B is said to be *normal* if $BB^* = B^*B$. A matrix is normal iff it has a complete set of orthogonal eigenvectors, see Horn & Johnson (1985, p. 101). So a normal matrix B can be decomposed as

$$B = U\Lambda U^{-1} \quad (2.16)$$

with unitary U and diagonal $\Lambda = \text{diag}(\lambda_j)$. The columns of U are the eigenvectors of B , that is, $Bu_j = \lambda_j u_j$ if $U = [u_1, u_2, \dots, u_m]$. The spectral norm $\|B\|_2$ of a normal matrix B equals its spectral radius $\rho(B)$.

Examples of normal matrices are the unitary (orthogonal) matrices, the Hermitian (symmetric) matrices, and the skew-Hermitian (skew-symmetric) matrices. The eigenvalues of a unitary matrix are all on the unit circle, the eigenvalues of an Hermitian matrix are all real and those of a skew-Hermitian matrix are all purely imaginary.

2.3 Perturbations on ODE Systems

The numerical methods for solving initial value problems will always be formulated for arbitrary nonlinear problems (2.1). The analysis of the methods however will usually be restricted to linear problems. In this section we take a look at properties of linear systems of ODEs and in particular at the influence of perturbations on such systems.

Linear Systems of ODEs

Consider the initial value problem (2.1) for a linear ODE system

$$w'(t) = Aw(t) + g(t), \quad t > 0, \quad (2.17)$$

with given matrix $A \in \mathbb{R}^{m \times m}$ and continuous source term $g : [0, \infty) \rightarrow \mathbb{R}^m$. Setting $F(t, v) = Av + g(t)$, the Lipschitz condition (2.2) will hold with $L = \|A\|$ and with cylinder \mathcal{C}_0 being the whole $\mathbb{R} \times \mathbb{R}^m$. Consequently, for any given initial value $w(0) = w_0 \in \mathbb{R}^m$ there is a unique solution on the interval $[0, \infty)$.

The solution of the homogeneous differential equation $w'(t) = Aw(t)$ can be written as

$$w(t) = e^{tA}w(0), \quad (2.18)$$

where the exponential of the matrix tA is defined by the power series ¹⁰⁾

$$e^{tA} = I + tA + \frac{1}{2}t^2A^2 + \cdots + \frac{1}{k!}t^kA^k + \cdots. \quad (2.19)$$

Since the individual terms in the power series are bounded by $\frac{1}{k!}t^k\|A\|^k$, it follows that the power series converges and that $\|e^{tA}\| \leq e^{t\|A\|}$. We note that from the observation

$$\frac{1}{\tau}(e^{(t+\tau)A} - e^{tA}) = \frac{1}{\tau}(e^{\tau A} - I)e^{tA} = Ae^{tA} + \mathcal{O}(\tau), \quad \tau \rightarrow 0,$$

it can be seen that $w(t) = e^{tA}w(0)$ is indeed the solution of the problem.

For the inhomogeneous problem (2.17) the solution is given by the *variation of constants formula*¹¹⁾

$$w(t) = e^{tA}w(0) + \int_0^t e^{(t-s)A}g(s) ds. \quad (2.20)$$

Generalizations for problems with non-constant coefficients can be found for instance in Hairer et al. (1993) or Coppel (1965).

¹⁰⁾ Exercise: Suppose that A is diagonalizable, $A = U\Lambda U^{-1}$ with diagonal matrix $\Lambda = \text{diag}(\lambda_j)$. Show that $e^{tA} = Ue^{t\Lambda}U^{-1}$ with $e^{t\Lambda} = \text{diag}(e^{t\lambda_j})$.

¹¹⁾ Exercise: Derive this formula by writing the equation (2.17) as $\frac{d}{dt}[e^{-tA}w(t)] = e^{-tA}w'(t) - e^{-tA}Aw(t) = e^{-tA}g(t)$.

Stability for Linear Systems

Consider along with

$$w'(t) = Aw(t) + g(t), \quad w(0) = w_0,$$

also a perturbed problem

$$\tilde{w}'(t) = A\tilde{w}(t) + g(t) + \delta(t), \quad \tilde{w}(0) = \tilde{w}_0.$$

Then for $\varepsilon(t) = \tilde{w}(t) - w(t)$ we find by the variation of constants formula that

$$\varepsilon(t) = e^{tA}\varepsilon(0) + \int_0^t e^{(t-s)A}\delta(s) ds,$$

which leads to the norm estimate

$$\|\varepsilon(t)\| \leq \|e^{tA}\| \|\varepsilon(0)\| + \int_0^t \|e^{(t-s)A}\| ds \max_{0 \leq s \leq t} \|\delta(s)\|.$$

Consequently, if we have the following stability inequality

$$\|e^{tA}\| \leq K e^{t\omega} \quad \text{for all } t \geq 0, \quad (2.21)$$

with constants $K > 0$, $\omega \in \mathbb{R}$, then we obtain

$$\|\varepsilon(t)\| \leq K e^{t\omega} \|\varepsilon(0)\| + \frac{K}{\omega} (e^{t\omega} - 1) \max_{0 \leq s \leq t} \|\delta(s)\|, \quad (2.22)$$

with convention $(e^{t\omega} - 1)/\omega = t$ in case $\omega = 0$. This inequality shows that the overall error $\|\varepsilon(t)\|$ can be bounded in terms of the initial error $\|\varepsilon(0)\|$ and the perturbations $\|\delta(s)\|$, $0 \leq s \leq t$.

In general, the term *stability* will be used to indicate that small perturbations give a small overall effect. This is just what is implied by (2.21) provided $t\omega$ and tK are bounded from above by a constant of moderate size. The term ‘moderate’ will be used throughout this text to indicate something of order of magnitude one, but this must be understood in an operational sense. For example, if we have perturbations with order of magnitude $\sim 10^{-6}$ and these perturbations are amplified with a factor $\sim 10^3$, then this factor might still be considered as ‘moderate enough’ if one is only interested in 3 digits accuracy.

Motivated by the above, we now take a closer look at bounds for $\|e^{tA}\|$. We already saw in the previous subsection that $\|e^{tA}\| \leq e^{t\|A\|}$. However, in general this gives a large over-estimation. As a simple example, consider a scalar problem with $A = \lambda \ll -1$, in which case $\|e^{tA}\| = e^{t\lambda} \ll e^{t|\lambda|} = e^{t\|A\|}$. The reason for this over-estimation is the fact that in the norm $\|A\|$, and consequently also in the Lipschitz constant L , no distinction is made between eigenvalues λ with $\operatorname{Re} \lambda \gg 1$, which lead to exponential growth with a large factor, and the eigenvalues with $\operatorname{Re} \lambda \ll -1$ which lead to exponential decay

with a large factor. Since $\|A\| \geq \rho(A)$, any eigenvalue with a large modulus will lead to a large norm of A .

Suppose that A is diagonalizable, $A = U \Lambda U^{-1}$, and that the vector norm is absolute. Then it follows that

$$\|e^{tA}\| \leq \|U\| \|e^{t\Lambda}\| \|U^{-1}\| = \text{cond}(U) \max_{1 \leq k \leq m} |e^{t\lambda_k}|. \quad (2.23)$$

Consequently if we know that $\text{cond}(U) = \|U\| \|U^{-1}\| \leq K$ and $\text{Re } \lambda_k \leq \omega$, then (2.21) follows with

$$\omega = \max_{1 \leq k \leq m} \text{Re } \lambda_k.$$

In particular, if A is a normal matrix, then the matrix of eigenvectors U is unitary. Since $e^{tA} = U e^{t\Lambda} U^{-1}$ the matrix e^{tA} is also normal. Thus with normal matrices A we have

$$\|e^{tA}\|_2 = \max_{1 \leq k \leq m} |e^{t\lambda_k}|. \quad (2.24)$$

In general, if the matrix A is not normal, an estimate of $\text{cond}(U)$ in some suitable norm may be difficult to obtain. For this reason we will look at a more general concept to obtain bounds for $\|e^{tA}\|$.

The Logarithmic Norm of Matrices

A useful concept for stability results with non-normal matrices is the *logarithmic norm* of a matrix A in $\mathbb{R}^{m \times m}$ or $\mathbb{C}^{m \times m}$, defined as

$$\mu(A) = \lim_{\tau \downarrow 0} \frac{\|I + \tau A\| - 1}{\tau}. \quad (2.25)$$

For $\tau > 0$ the difference ratio on the right-hand side is easily seen to be in the interval $[-\|A\|, \|A\|]$. Moreover, it is monotonically non-decreasing in τ : if $0 < \sigma < 1$ then

$$\frac{1}{\sigma\tau} (\|I + \sigma\tau A\| - 1) \leq \frac{1}{\sigma\tau} (\|\sigma I + \sigma\tau A\| + |1 - \sigma| - 1) = \frac{1}{\tau} (\|I + \tau A\| - 1).$$

Hence the limit in (2.25) exists. Note that the logarithmic norm is not a matrix norm; it can be negative. The importance of this logarithmic norm lies in the following result.

Theorem 2.4 *Let $A \in \mathbb{C}^{m \times m}$ and $\omega \in \mathbb{R}$. We have*

$$\mu(A) \leq \omega \iff \|e^{tA}\| \leq e^{t\omega} \text{ for all } t \geq 0.$$

Proof. Suppose that $\mu(A) \leq \omega$. Then

$$\|I + \tau A\| \leq 1 + \omega\tau + o(\tau), \quad \tau \downarrow 0,$$

$$\|(I + \tau A)^n\| \leq (1 + \omega\tau + o(\tau))^n \rightarrow e^{t\omega} \quad \text{as } \tau \downarrow 0, t = n\tau \text{ fixed.}$$

According to convergence of Euler's method, see Theorem 2.2, we have

$$e^{tA} = \lim_{\tau \downarrow 0} (I + \tau A)^n \quad \text{as } \tau \downarrow 0, t = n\tau \text{ fixed,}$$

and hence $\|e^{tA}\| \leq e^{t\omega}$.

On the other hand, suppose that $\|e^{tA}\| \leq e^{t\omega}$ for all $t > 0$. Since $I + \tau A = e^{\tau A} + \mathcal{O}(\tau^2)$ it follows that

$$\|I + \tau A\| \leq 1 + \tau\omega + \mathcal{O}(\tau^2) \quad \text{for } \tau \downarrow 0,$$

from which we obtain $\mu(A) \leq \omega$. □

It is easy to show that the logarithmic matrix norm has the properties

$$\mu(sI + tA) = s + t\mu(A) \quad \text{for } s \in \mathbb{R}, t \geq 0, \quad (2.26)$$

$$\mu(A + B) \leq \mu(A) + \mu(B). \quad (2.27)$$

We already noted that $-\|A\| \leq \mu(A) \leq \|A\|$. Using the latter inequality it also easily follows that

$$|\mu(A) - \mu(B)| \leq \|A - B\|, \quad (2.28)$$

which shows continuity of the logarithmic norm. Finally we note that ¹²⁾

$$\mu(B) \geq -\|Bv\| / \|v\| \quad \text{for any } v \neq 0 \in \mathbb{C}^m. \quad (2.29)$$

For the L_p vector norms, the corresponding logarithmic norms of a matrix $A \in \mathbb{C}^{m \times m}$ are given by

$$\begin{aligned} \mu_2(A) &= \max_{v \neq 0} \frac{\operatorname{Re} \langle Av, v \rangle_2}{\langle v, v \rangle_2} = \max \left\{ \lambda : \lambda \in \sigma \left(\frac{1}{2}(A + A^*) \right) \right\}, \\ \mu_1(A) &= \max_j \left(\operatorname{Re} a_{jj} + \sum_{i \neq j} |a_{ij}| \right), \\ \mu_\infty(A) &= \max_i \left(\operatorname{Re} a_{ii} + \sum_{j \neq i} |a_{ij}| \right). \end{aligned} \quad (2.30)$$

These expressions can be derived directly from the formulas for the matrix norms (2.13) and the definition (2.25) of $\mu(A)$. In particular, for real matrices we have $\mu_2(A) \leq 0$ iff $\langle v, Av \rangle_2 \leq 0$ for all $v \in \mathbb{R}^m$. Assuming the diagonal elements of A to be negative, we have $\mu_\infty(A) \leq 0$ if A is row-wise diagonally

¹²⁾ This property is easily derived from $\|v\| = \|(I + \tau B)v - \tau Bv\| \leq \|(I + \tau B)v\| + \tau \|Bv\|$, and hence $-\|Bv\| \leq \frac{1}{\tau} (\|(I + \tau B)v\| - \|v\|) \leq \frac{1}{\tau} (\|I + \tau B\| - 1) \|v\|$.

dominant, and $\mu_1(A) \leq 0$ if A is column-wise diagonally dominant. Hence in these cases we have

$$\|e^{tA}\| \leq 1 \quad \text{for all } t \geq 0.$$

Clearly, for large $\|A\|$ this is a much better estimate than $\|e^{tA}\| \leq e^{t\|A\|}$.

For applications it is important to notice that the inequality $\|e^{tA}\| \leq e^{t\mu(A)}$ is in general only sharp for $t \downarrow 0$. The extent to which the inequality will be adequate for larger t may depend crucially on the choice of a suitable norm.

Example 2.5 For the ODE system $w'(t) = Aw(t)$ in (2.9) we directly see that $\mu_1(A) = 0$ and thus

$$\|e^{tA}\|_1 \leq 1 \quad \text{for any } t \geq 0.$$

If we consider the max-norm, then $\mu_\infty(A) = |k_2 - k_1|$. The resulting inequality $\|e^{tA}\|_\infty \leq e^{t|k_2 - k_1|}$ still suggests a large growth if $k_1 = 1, k_2 \gg 1$, but using the norm equivalence

$$\|v\|_1 \leq \|v\|_\infty \leq 2\|v\|_1 \quad \text{for all } v \in \mathbb{C}^2,$$

it also follows easily that

$$\|e^{tA}\|_\infty \leq \min(2, e^{t|k_2 - k_1|}) \quad \text{for all } t \geq 0.$$

In the spectral norm a similar estimate can be obtained.

It is clear from these results that the error propagation in the ODE system (2.9) is favourable, also for large k_2 . The fact that Euler's method did give very poor results is therefore not caused by ill-posedness of the problem. \diamond

The concept of logarithmic norm of a matrix was introduced independently in 1958 by G. Dahlquist and S.M. Lozinskij. References and further properties of logarithmic norms can be found in Coppel (1965) or Dekker & Verwer (1984). Some generalizations of this concept for nonlinear equations can be found at the end of this section. There we will also use the following technical result.

Lemma 2.6 Suppose $A(\sigma) \in \mathbb{R}^{m \times m}$ depends continuously on $\sigma \in [0, 1]$ and $\mu(A(\sigma)) \leq \omega$ for all $0 \leq \sigma \leq 1$. Then $B = \int_0^1 A(\sigma) d\sigma$ satisfies $\mu(B) \leq \omega$.

Proof. Note that the integral is the limit of a Riemann sum,

$$B_k = \frac{1}{k} \sum_{j=1}^k A(j/k) \rightarrow B \quad \text{as } k \rightarrow \infty.$$

From (2.27) it follows that $\mu(B_k) \leq \frac{1}{k} \sum_1^k \mu(A(j/k)) \leq \omega$. The continuity property (2.28) thus implies that $\mu(B) \leq \omega$. \square

As an application we mention here already that a solution of the linear equation $w'(t) = A(t)w(t) + g(t)$, with variable coefficients, will satisfy

$$\|w(t)\| \leq e^{\omega t} \|w(0)\| + \int_0^t e^{\omega(t-s)} \|g(s)\| ds$$

provided that $\mu(A(s)) \leq \omega$ for all $0 \leq s \leq t$. In fact, we will see that related stability results even hold for nonlinear problems. This will be done by studying suitable numerical discretizations.

2.4 The θ -Method and Stiff Problems

In Chapter II numerical ODE methods will be reviewed in some detail. To introduce the main concepts it is for the moment sufficient to consider the following method

$$w_{n+1} = w_n + (1 - \theta)\tau F(t_n, w_n) + \theta\tau F(t_{n+1}, w_{n+1}) \quad (2.31)$$

with parameter $\theta \in [0, 1]$. For any $\theta > 0$ this method is *implicit* since the new approximation w_{n+1} is given by an implicit algebraic relation. This method is known under the fancyless name of ‘ θ -method’. If we take $\theta = 0$ this is just Euler’s method which will be called from now on the *explicit Euler* method or the *forward Euler* method. Other choices that will often be considered are $\theta = \frac{1}{2}$ and $\theta = 1$. The method with $\theta = \frac{1}{2}$ is called the *trapezoidal rule*, on the analogy of the quadrature trapezoidal rule for integrals,

$$\frac{\tau}{2} \left(f(t_n) + f(t_{n+1}) \right) \approx \int_{t_n}^{t_{n+1}} f(t) dt.$$

The method with $\theta = 1$ is called the *implicit Euler* method or *backward Euler* method.

Due to the implicitness, the trapezoidal rule and backward Euler method are more expensive to use than the forward Euler method. However for problems with large Lipschitz constants these implicit methods offer distinct advantages.

Example 2.7 Consider again the two-way reaction problem (2.9) with $k_1 = 1$ and various $k_2 > 0$. In the Figures 2.2, 2.3 the first component of numerical results up to $T = 1$ are plotted for the trapezoidal rule and the implicit Euler method with step size $\tau = 1/50$. The same initial values were used as in Example 2.3.

In comparison with the explicit Euler method, see Figure 2.1, we obtain for large values of k_2 much better results with these implicit methods. With the implicit trapezoidal rule there are for large τL some initial oscillations. The results with the implicit Euler scheme are qualitatively correct, although the errors in the transient phase are also quite large. In particular, in the left

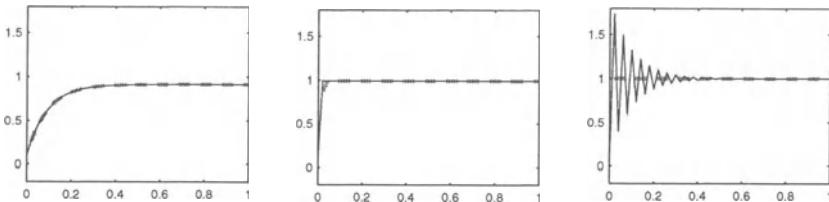


Fig. 2.2. Results for the trapezoidal rule with $\tau = 1/50$ and $k_2 = 10$ (left), $k_2 = 100$ (middle), $k_2 = 1000$ (right). The exact solution is indicated by dashed grey lines.

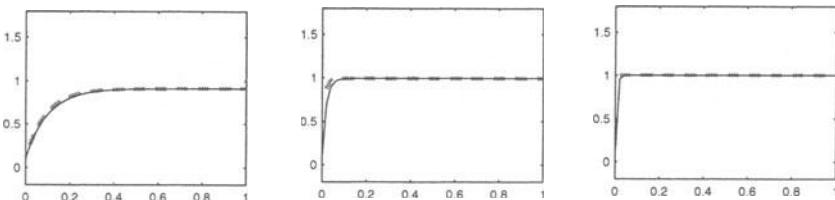


Fig. 2.3. Results for the implicit Euler method with $\tau = 1/50$ and $k_2 = 10$ (left), $k_2 = 100$ (middle), $k_2 = 1000$ (right). The exact solution is indicated by dashed grey lines.

picture ($k_2 = 10$) the results are seen to be less accurate than with the trapezoidal rule.

From a practical point of view it is important to notice that the oscillations with the trapezoidal rule and the inaccuracy with the implicit Euler method can be avoided by using in the transient phase a smaller step size; once the smooth region is reached larger step sizes can be used. With the explicit Euler method this approach will not work since the error propagation will always be unfavourable if τL is not sufficiently small. ◇

The above results are common in the sense that with the explicit Euler method the product τL has to be sufficiently small to obtain a reasonable approximation, whereas with the trapezoidal rule or implicit Euler method large values of τL can often be allowed.

An initial value problem with a smooth, stable solution but with a large Lipschitz constant L is called *stiff*. Stiffness is not a mathematical definition, since no quantification is given for ‘large’ or ‘moderate’. Instead it is an operational concept, indicating the class of problems for which implicit methods can perform (much) better than explicit methods. More examples of stiff equations will follow in this chapter which will make this concept of stiffness more clear.

2.5 Stability of the θ -Method

In order to understand the different behaviour of the explicit Euler method and the implicit methods in the above example, we will consider the stability properties of these methods for linear problems (2.17).

The Scalar Test Equation

To begin with we consider the scalar, complex test equation

$$w'(t) = \lambda w(t) \quad (2.32)$$

with $\lambda \in \mathbb{C}$. Application of the θ -method to this test equation gives approximations

$$w_{n+1} = R(\tau\lambda) w_n, \quad R(z) = \frac{1 + (1 - \theta)z}{1 - \theta z}. \quad (2.33)$$

If we perturb the initial value w_0 to \tilde{w}_0 , we get the recursion $\tilde{w}_{n+1} = R(\tau\lambda)\tilde{w}_n$ for the resulting perturbed sequence $\tilde{w}_n, n \geq 0$. Likewise, for the differences $\tilde{w}_n - w_n$ we find the same recurrence expression, revealing that $R(\tau\lambda)$ determines how for evolving time the initial perturbation develops. This R is therefore called the *stability function* of the method. Near $z = 0$ we have $R(z) = 1 + z + \theta z^2 + \mathcal{O}(z^3)$ and therefore

$$R(z) = e^z + \mathcal{O}(z^{p+1}), \quad z \rightarrow 0,$$

with $p = 2$ if $\theta = \frac{1}{2}$ and $p = 1$ for the other values of θ .

The *stability region* of the method is the set

$$\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$$

in the complex plane. An ODE method that has the property that \mathcal{S} contains the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$ is called *A-stable*. A-stability is an important property for stiff problems.¹³⁾ Using the maximum modulus theorem¹⁴⁾ it is easy to show that the θ -method is A-stable for $\theta \geq \frac{1}{2}$. The stability regions for $\theta = 0, \frac{1}{2}, 1$ are plotted in Figure 2.4. If $\theta = \frac{1}{2}$ then \mathcal{S} is precisely the closed left half-plane \mathbb{C}^- .

It is clear that having $\tau\lambda \in \mathcal{S}$ is sufficient to have stability of the recursion in (2.33). Consequently, if $\operatorname{Re} \lambda \in \mathbb{C}^-$ and the method is A-stable, then we have *unconditional stability*, that is, stability without any condition on the step size.

¹³⁾ Along with the test equation (2.32), this concept has been introduced by Dahlquist (1963). In spite of its simplicity, equation (2.32) was readily acknowledged of being of major importance for the stability analysis of numerical ODE methods. The scalar test equation, the stability region and the A-stability concept are therefore very often used.

¹⁴⁾ Let φ be a non-constant complex function which is analytic on a set $\mathcal{D} \subset \mathbb{C}$ and continuous on its closure. The maximum modulus theorem states that the maximum of $|\varphi(z)|$ on \mathcal{D} is assumed on the boundary of \mathcal{D} and not in the interior, see for instance Marsden (1973). In particular, if φ is rational without poles in \mathbb{C}^- then we know that $\max_{z \in \mathbb{C}^-} |\varphi(z)| = \max_{y \in \mathbb{R}} |\varphi(iy)|$.

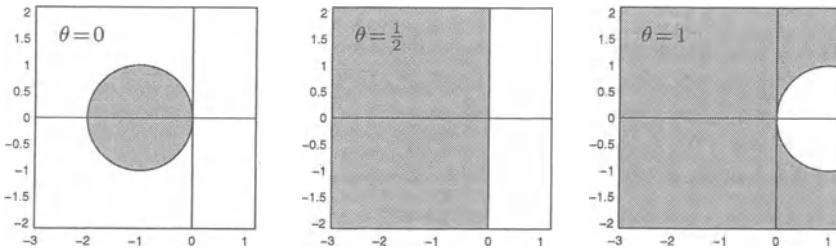


Fig. 2.4. Stability regions (shaded) for the θ -method with $\theta = 0, \frac{1}{2}, 1$.

Note that A -stability mimics the property $|e^z| \leq 1$ for $z \in \mathbb{C}^-$. Of course, the exponential function also satisfies

$$|e^z| = 1 \quad \text{if } \operatorname{Re} z = 0, \quad (2.34)$$

$$|e^z| \leq e^\nu < 1 \quad \text{if } \operatorname{Re} z \leq \nu < 0, \quad |e^z| \rightarrow 0 \quad \text{as } \operatorname{Re} z \rightarrow -\infty. \quad (2.35)$$

An ODE method with stability function R is said to be *strongly A-stable* if it is A -stable with $|R(\infty)| < 1$, and it is said to be *L-stable* if we have in addition $|R(\infty)| = 0$. For strongly A -stable methods we have by the maximum modulus theorem $|R(z)| \leq \max_y |R(\nu + iy)| < 1$ whenever $\operatorname{Re} z \leq \nu < 0$, which is in accordance with the first statement of (2.35).¹⁵⁾ On the other hand, strong A -stability also implies $|R(iy)| < 1$ for $y \in \mathbb{R}$, $|y| \rightarrow \infty$ and this is at odds with (2.34).

The θ -method is strongly A -stable if $\theta > \frac{1}{2}$ and it is L -stable for $\theta = 1$. If $\theta = \frac{1}{2}$ then $R(\infty) = -1$ so then the method is ‘just’ A -stable; with this method we have $|R(iy)| = 1$ on the imaginary axis.

If we consider the test equation $w'(t) = \lambda w(t)$ with real negative λ , then strong A -stability is a favourable property since the numerical approximations will then satisfy $|w_{n+1}| < |w_n|$ which reflects the damping of the exact solution. On the other hand, if λ is purely imaginary then having a conservation property $|w_{n+1}| = |w_n|$ is more natural.

Stability for Linear Systems

For linear systems $w'(t) = Aw(t) + g(t)$ with $A \in \mathbb{R}^{m \times m}$, application of the θ -method gives

$$w_{n+1} = R(\tau A)w_n + (I - \theta\tau A)^{-1}\tau g_{n+\theta}, \quad (2.36)$$

where

$$R(\tau A) = (I - \theta\tau A)^{-1}(I + (1 - \theta)\tau A) \quad (2.37)$$

¹⁵⁾ Also A -stability implies $|R(z)| < 1$ if $\operatorname{Re} z < 0$. However, now $|R(z)| < 1$ is not valid uniformly for $\operatorname{Re} z \leq \nu < 0$ because A -stability allows that $|R(z)| \rightarrow 1$ as $z \rightarrow -\infty$.

and $g_{n+\theta} = (1 - \theta)g(t_n) + \theta g(t_{n+1})$. By elaborating the recursion we obtain

$$w_n = R(\tau A)^n w_0 + \tau \sum_{j=0}^{n-1} R(\tau A)^{n-j-1} (I - \theta \tau A)^{-1} g_{j+\theta}, \quad (2.38)$$

which is a discrete counterpart of the variation of constants formula. With a perturbed initial value \tilde{w}_0 , we get the same formula for the perturbed sequence \tilde{w}_n , $n \geq 0$, so that

$$\tilde{w}_n - w_n = R(\tau A)^n (\tilde{w}_0 - w_0).$$

Hence the powers $R(\tau A)^n$ determine the growth of the initial error $\tilde{w}_0 - w_0$. Stability of the recursion requires a moderate bound for these powers.

If we consider a fixed matrix A , then

$$R(\tau A) = I + \tau A + \mathcal{O}(\tau^2), \quad \tau \downarrow 0,$$

and hence

$$\|R(\tau A)^n\| \leq (1 + \tau \|A\| + \mathcal{O}(\tau^2))^n \leq e^{2t_n\|A\|} \quad \text{for } t_n = n\tau \leq T$$

provided $\tau \|A\|$ is sufficiently small. For stiff systems, however, the norm of A will be very large and consequently this estimate is then useless. Better bounds can be derived by invoking the stability region.

Theorem 2.8 Suppose $\|\cdot\|$ is an absolute vector norm and

$$A = U \Lambda U^{-1} \quad \text{with} \quad \Lambda = \text{diag}(\lambda_j), \quad \text{cond}(U) \leq K.$$

Then

$$\tau \lambda_j \in \mathcal{S}, \quad 1 \leq j \leq m \quad \Rightarrow \quad \|R(\tau A)^n\| \leq K \quad \text{for all } n \geq 1.$$

Proof. From $A = U \Lambda U^{-1}$ it is easily seen that $R(\tau A) = U R(\tau \Lambda) U^{-1}$ and therefore also

$$R(\tau A)^n = U R(\tau \Lambda)^n U^{-1}.$$

The proof now follows from the observation that $R(\tau \Lambda) = \text{diag}(R(\tau \lambda_j))$. \square

As an immediate consequence we obtain the following result in the spectral norm that will be very frequently used in subsequent sections.

Corollary 2.9 Suppose the matrix A is normal. Then

$$\tau \lambda_j \in \mathcal{S}, \quad 1 \leq j \leq m \quad \Rightarrow \quad \|R(\tau A)\|_2 \leq 1$$

and hence $\|R(\tau A)^n\|_2 \leq 1$ for all $n \geq 1$. \square

Example 2.10 Consider once more the matrix

$$A = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix} \quad \text{with } k_1 = 1 \text{ and } k_2 \gg 1.$$

The eigenvalues are $\lambda_1 = 0$, $\lambda_2 = -(k_1 + k_2)$. Setting $\kappa_j = k_j/(k_1 + k_2)$, the matrix of eigenvectors is given by

$$U = \begin{pmatrix} \kappa_2 & 1 \\ \kappa_1 & -1 \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} 1 & 1 \\ \kappa_1 & -\kappa_2 \end{pmatrix}.$$

Let $r = R(-\tau(k_1 + k_2))$. A direct calculation of $R(\tau A)^n = UR(\tau A)^nU^{-1}$ then gives

$$R(\tau A)^n = \begin{pmatrix} \kappa_2 + r^n \kappa_1 & (1 - r^n) \kappa_2 \\ (1 - r^n) \kappa_1 & \kappa_1 + r^n \kappa_2 \end{pmatrix}.$$

It is clear that having $|r^n| \leq K$ for $0 \leq n\tau \leq T$, with a moderate K , is the condition for stability on $[0, T]$ with any of the discrete L_p -norms (2.10). For fixed T this stability condition reads $|r| \leq 1 + C\tau$, $C \sim T^{-1} \log K$, and for large k_2 or large T it can be verified that this condition is essentially the same as $|r| \leq 1$. The explicit Euler method is therefore only stable if the step size satisfies $\tau \leq 2/(k_1 + k_2)$. The trapezoidal rule and implicit Euler method are stable for all $\tau > 0$. Note that with the trapezoidal rule we have $r \approx -1$ if $\tau(k_1 + k_2) \gg 1$, which explains the oscillatory behaviour in Figure 2.2. With the implicit Euler method we have $r \approx 0$ if $\tau(k_1 + k_2) \gg 1$. \diamond

For this simple example in \mathbb{R}^2 , where an explicit expression for $R(\tau A)^n$ can be easily found, the above general results of Theorem 2.8 or Corollary 2.9 are redundant. These will be useful in later sections for linear ODE systems in \mathbb{R}^m with large dimension m .

For a large number of applications Theorem 2.8 gives a verifiable sufficient condition for stability, but there are also many applications for which the condition number of U is very large or for which the matrix A is not diagonalizable. In such a situation the following result, based on the logarithmic norm with inner products, can be helpful.

Theorem 2.11 Suppose the vector norm is generated by an inner product and that¹⁶⁾

$$\operatorname{Re} \langle v, Av \rangle \leq \omega \|v\|^2 \quad \text{for all } v \in \mathbb{C}^m.$$

Then

$$\|R(\tau A)\| \leq \max_{\operatorname{Re} z \leq \tau\omega} |R(z)| = \max(|R(\tau\omega)|, |R(\infty)|)$$

provided that $1 - \theta\tau\omega > 0$.

¹⁶⁾ Exercise: Show that with an inner product norm on \mathbb{C}^m we have $\operatorname{Re} \langle v, Av \rangle \leq \omega \|v\|^2$ for all v iff $\mu(A) \leq \omega$.

Proof. Let $Z = \tau A$ and consider $w_1 = R(Z)w_0$, which we can also write as

$$w_1 = (I + (1 - \theta)Z)(I - \theta Z)^{-1}w_0.$$

By introducing $u = (I - \theta Z)^{-1}w_0$, $v = u/\|u\|$, we have

$$w_1 = u + (1 - \theta)Zu, \quad w_0 = u - \theta Z u,$$

from which it follows that

$$\frac{\|w_1\|^2}{\|w_0\|^2} = \frac{1 + 2(1 - \theta)\langle v, Zv \rangle + (1 - \theta)^2 \|Zv\|^2}{1 - 2\theta\langle v, Zv \rangle + \theta^2 \|Zv\|^2}.$$

This relation can also be written as

$$\frac{\|w_1\|^2}{\|w_0\|^2} = |R(\zeta)|^2, \quad \zeta = \langle v, Zv \rangle + i\sqrt{\|Zv\|^2 - \langle v, Zv \rangle^2}.$$

Since

$$\operatorname{Re} \zeta = \langle v, Zv \rangle \leq \tau\omega,$$

it follows that $\|R(Z)\|$ is bounded by $C = \max\{|R(z)| : \operatorname{Re} z \leq \tau\omega\}$. The equality $C = \max(|R(\tau\omega)|, |1 - 1/\theta|)$ follows directly from the maximum modulus theorem. \square

Corollary 2.12 Suppose that $\mu(A) \leq 0$ in an inner product norm, and consider the θ -method with $\theta \geq \frac{1}{2}$. Then

$$\|R(\tau A)\| \leq 1 \quad \text{for any } \tau > 0.$$

\square

This result is a direct consequence of Theorem 2.11. It shows that for the A -stable θ -methods (that is, $\theta \geq \frac{1}{2}$) we will often have unconditional stability. In contrast, if $\theta < \frac{1}{2}$ then stability will always impose a restriction on the allowable step size, since the stability region is bounded for the methods with $\theta < \frac{1}{2}$. We note that Theorem 2.11 is a special case, for the θ -method, of an important general result of J. von Neumann from 1951 for arbitrary rational functions R . References and a proof of this general result can be found in Hairer & Wanner (1996, Sect. IV.11).

Theorem 2.11 is valid only with inner product norms. For the implicit Euler method we have the following generalization which is valid in any norm.

Theorem 2.13 Let $A \in \mathbb{C}^{m \times m}$ and $\tau > 0$, $\omega \in \mathbb{R}$. We have

$$\mu(A) \leq \omega \iff \|(I - \tau A)^{-1}\| \leq \frac{1}{1 - \tau\omega} \quad \text{whenever } 1 - \tau\omega > 0.$$

Proof. Suppose $\mu(A) \leq \omega$ and consider the relation $w_0 = (I - \tau A)w_1$. By using (2.26) and (2.29) with $B = \tau A - I$, it follows that

$$\|w_0\| \geq -\mu(\tau A - I)\|w_1\| \geq (1 - \tau\omega)\|w_1\|.$$

Thus we see that if $1 - \tau\omega > 0$ then $I - \tau A$ is nonsingular and

$$\|(I - \tau A)^{-1}\| \leq (1 - \tau\omega)^{-1}.$$

On the other hand, assume the latter inequality holds for small τ . Then by using the series expansion

$$(I - \tau A)^{-1} = I + \tau A + \tau^2 A^2 + \dots \quad \text{if } \|\tau A\| < 1,$$

it follows that $\|I + \tau A\| \leq 1 + \omega\tau + \mathcal{O}(\tau^2)$, which implies $\mu(A) \leq \omega$. \square

We note that the result of this theorem does not hold for the θ -methods with $\theta \neq 1$. Also if we consider the L_∞ -norm or L_1 -norm instead of an arbitrary norm, it has to be required that $\theta = 1$ in order to obtain an estimate $\|R(\tau A)\| \leq 1$, see Hairer & Wanner (1996, Sect. IV.11). In this respect, the implicit Euler method is very special. We will return to this matter in later sections in connection with positivity properties.

2.6 Consistency and Convergence of the θ -Method

In this subsection we consider different types of local errors and convergence of the global errors for the θ -method. Inserting the exact solution $w(t)$ of the initial value problem (2.1) into the scheme (2.31) gives

$$w(t_{n+1}) = w(t_n) + (1 - \theta)\tau w'(t_n) + \theta\tau w'(t_{n+1}) + \tau\rho_n \quad (2.39)$$

with (local) truncation error ρ_n . By Taylor expansion around $t = t_n$ it follows that

$$\rho_n = \frac{1}{2}(1 - 2\theta)\tau w''(t_n) + \frac{1}{6}(1 - 3\theta)\tau^2 w'''(t_n) + \mathcal{O}(\tau^3).$$

Thus, for $w(t)$ sufficiently smooth, we get $\|\rho_n\| = \mathcal{O}(\tau)$ if $\theta \neq \frac{1}{2}$ and $\|\rho_n\| = \mathcal{O}(\tau^2)$ if $\theta = \frac{1}{2}$. The constants involved in these $\mathcal{O}(\tau^p)$ terms are dependent on bounds for derivatives of the exact solution but not on the size of the Lipschitz constant.

For further analysis, assume as before that the problem is linear, $F(t, v) = Av + g(t)$, and let $\varepsilon_n = w(t_n) - w_n$, $n \geq 0$, stand for the global discretization errors. We want to find an upper bound for $\|\varepsilon_n\|$. Subtraction of (2.31) from (2.39) leads to the recursion

$$\varepsilon_{n+1} = \varepsilon_n + (1 - \theta)\tau A\varepsilon_n + \theta\tau A\varepsilon_{n+1} + \tau\rho_n.$$

It follows that

$$\varepsilon_{n+1} = R(\tau A)\varepsilon_n + \delta_n \quad \text{for } n \geq 0, \quad \varepsilon_0 = w(0) - w_0, \quad (2.40)$$

with $R(\tau A)$ given in (2.37) and

$$\delta_n = (I - \theta\tau A)^{-1} \tau \rho_n. \quad (2.41)$$

The error recursion (2.40) clearly shows how the global error is built up. The matrix $R(\tau A)$ determines how an error already present at time level t_n is propagated to the next time level. On the other hand, during this time step also a new error δ_n is introduced. This δ_n is called the *local discretization error* because it is the error which would originate with one step from the true solution, that is, from $w_n = w(t_n)$. The method is said to be *consistent of order p* if $\|\delta_n\| = \mathcal{O}(\tau^{p+1})$ whenever the exact solution is sufficiently smooth. Note that we do have

$$\|\delta_n\| \leq C\tau \|\rho_n\|$$

provided that

$$\|(I - \theta\tau A)^{-1}\| \leq C.$$

The existence of this inverse simply means that the implicit relation in the θ -method has a unique solution and a bound on the inverse will always hold if we can bound $\|R(\tau A)\|$, since

$$(I - \theta\tau A)^{-1} = \theta R(\tau A) + (1 - \theta)I.$$

We then find consistency with order $p = 2$ if $\theta = \frac{1}{2}$ and $p = 1$ for the other values of θ .

To relate the local discretization errors to the global errors we need *stability*. Consider a time interval $[0, T]$ and assume

$$\|R(\tau A)^n\| \leq K \quad \text{for } n \geq 0, n\tau \leq T. \quad (2.42)$$

Elaboration of the error recursion (2.40) gives

$$\varepsilon_n = R(\tau A)^n \varepsilon_0 + R(\tau A)^{n-1} \delta_0 + \cdots + R(\tau A) \delta_{n-2} + \delta_{n-1},$$

which leads directly to

$$\|\varepsilon_n\| \leq K \|\varepsilon_0\| + K \sum_{j=0}^{n-1} \|\delta_j\| \quad \text{for } n\tau \leq T. \quad (2.43)$$

Convergence now follows easily: if we have $\|\delta_j\| \leq C\tau^{p+1}$ for all j and $w_0 = w(0)$, then the global errors satisfy the bound

$$\|w(t_n) - w_n\| \leq C' \tau^p \quad \text{for } n\tau \leq T, \quad (2.44)$$

with constant $C' = KTC$, and thus we have *convergence of order p*. Obviously, stability is the crucial point here. To establish (2.42) we can use the sufficient conditions that were presented in the theorems of the previous subsection. For the A -stable methods, $\theta \geq \frac{1}{2}$, this will not involve the Lipschitz constant $\|A\|$ and therefore these methods are convergent for problems with arbitrary stiffness.

2.7 Nonlinear Results for the θ -Method

Generalizations of the θ -method to Runge-Kutta methods or linear multi-step methods are treated in Chapter II. As a rule, stability and convergence properties for ODE methods will only be discussed for linear problems. However, for the θ -method nonlinear results can be obtained without involving too many technical issues. Here we will outline a nonlinear result for the θ -method with $\theta \geq \frac{1}{2}$ that complements the Theorem 2.2 for the explicit Euler method.

We assume here that the function F is continuously differentiable and

$$\mu(J(t, v)) \leq 0 \quad \text{for all } t \in \mathbb{R}, v \in \mathbb{R}^m, \quad (2.45)$$

where $J(t, v)$ stands for the Jacobian matrix $(\partial F_i(t, v)/\partial v_j)$. If $\theta \neq 1$ we will employ Corollary 2.12 for which it must also be assumed that the underlying vector norm on \mathbb{R}^m is generated by an inner product. With $\theta = 1$ we can use Theorem 2.13 which is valid in any norm.

We will demonstrate stability of the θ -method with $\theta \geq \frac{1}{2}$ under assumption (2.45) with appropriate norms. Consider once more

$$w_{n+1} = w_n + (1 - \theta)\tau F(t_n, w_n) + \theta\tau F(t_{n+1}, w_{n+1}), \quad (2.46)$$

along with a perturbed version

$$\tilde{w}_{n+1} = \tilde{w}_n + (1 - \theta)\tau F(t_n, \tilde{w}_n) + \theta\tau F(t_{n+1}, \tilde{w}_{n+1}) + \tau\rho_n. \quad (2.47)$$

Let $Z_n \in \mathbb{R}^{m \times m}$ be defined for $n \geq 0$ by

$$Z_n = \int_0^1 \tau J(t_n, \sigma\tilde{w}_n + (1 - \sigma)w_n) d\sigma, \quad (2.48)$$

so that we have, according to the mean-value theorem for vector functions,

$$\tau F(t_n, \tilde{w}_n) - \tau F(t_n, w_n) = Z_n(\tilde{w}_n - w_n). \quad (2.49)$$

From Lemma 2.6 and (2.48) it is seen that

$$\mu(Z_n) \leq 0. \quad (2.50)$$

By subtraction of (2.46) from (2.47) it follows that the global errors $\varepsilon_n = \tilde{w}_n - w_n$ satisfy the recursion

$$\varepsilon_{n+1} = \varepsilon_n + (1 - \theta)Z_n\varepsilon_n + \theta Z_{n+1}\varepsilon_{n+1} + \tau\rho_n. \quad (2.51)$$

First consider $\theta = 1$, the implicit Euler method. Then (2.51) gives

$$\varepsilon_{n+1} = (I - \theta Z_{n+1})^{-1}(\varepsilon_n + \tau\rho_n) = R(Z_{n+1})(\varepsilon_n + \tau\rho_n). \quad (2.52)$$

Since $\|(I - \theta Z_{n+1})^{-1}\| \leq 1$ according to Theorem 2.13, we obtain

$$\|\varepsilon_{n+1}\| \leq \|\varepsilon_n\| + \tau \|\rho_n\|,$$

which leads to the global result

$$\|\varepsilon_n\| \leq \|\varepsilon_0\| + \sum_{j=0}^{n-1} \tau \|\rho_j\| \leq \|\varepsilon_0\| + t_n \max_{0 \leq j < n} \|\rho_j\|. \quad (2.53)$$

Next we consider inner product norms and $\frac{1}{2} \leq \theta < 1$, with $\theta = \frac{1}{2}$, the trapezoidal rule, as case of special interest. Here (2.51) cannot be used directly since now Z_j at two time levels $j = n, n+1$ are involved. If we define for $n \geq 0$

$$\tilde{\varepsilon}_n = (I - \theta Z_n) \varepsilon_n,$$

then we see that these transformed errors satisfy the recursion

$$\tilde{\varepsilon}_{n+1} = (I + (1 - \theta)Z_n)(I - \theta Z_n)^{-1} \tilde{\varepsilon}_n + \tau \rho_n = R(Z_n) \tilde{\varepsilon}_n + \tau \rho_n. \quad (2.54)$$

Corollary 2.12 implies that $\|R(Z_n)\| \leq 1$. Application to (2.54) thus shows

$$\|\tilde{\varepsilon}_{n+1}\| \leq \|\tilde{\varepsilon}_n\| + \tau \|\rho_n\|,$$

which gives the global result

$$\|\tilde{\varepsilon}_n\| \leq \|\tilde{\varepsilon}_0\| + \sum_{j=0}^{n-1} \tau \|\rho_j\| \leq \|\tilde{\varepsilon}_0\| + t_n \max_{0 \leq j < n} \|\rho_j\|.$$

Since $\|(I - \theta Z_n)^{-1}\| \leq 1$ we have $\|\varepsilon_n\| \leq \|\tilde{\varepsilon}_n\|$. In terms of our original errors ε_n we therefore obtain

$$\|\varepsilon_n\| \leq \|(I - \theta Z_0) \varepsilon_0\| + t_n \max_{0 \leq j < n} \|\rho_j\|. \quad (2.55)$$

Note that $\|Z_0\| \leq \tau L$ with L the Lipschitz constant of F . Insertion of $\|(I - \theta Z_0) \varepsilon_0\| \leq (1 + \tau L) \|\varepsilon_0\|$ into (2.55) leads to a result which is for stiff problems weaker than the corresponding stability result (2.43) for linear problems with respect to the initial error ε_0 . It is sufficient however to prove convergence of the numerical approximations.

To demonstrate convergence, let $\tilde{w}_n = w(t_n)$, $n \geq 0$, so that ε_n becomes the global discretization error and ρ_n the truncation error,

$$\rho_n = \frac{1}{2}(1 - 2\theta)\tau w''(t_n) + \mathcal{O}(\tau^2).$$

Then application of (2.53), (2.55) shows that if the solution $w(t)$ is sufficiently smooth and $\varepsilon_0 = w(0) - w_0 = 0$, we have first-order convergence for $\theta > \frac{1}{2}$ and second-order convergence for $\theta = \frac{1}{2}$ on a bounded time interval $[0, T]$. In this convergence result only bounds for the derivatives of the exact solution $w(t)$ are involved, and thus it is valid for problems with arbitrary stiffness.

We note that the above convergence result for the trapezoidal rule is essentially due to Dahlquist (1963). Convergence of the implicit Euler method for arbitrary norms was first derived by Desoer & Haneda (1972).

2.8 Concluding Remarks

Some Additional Results on Nonlinear Stability

From the above nonlinear stability and convergence results, a stability result for the ODE system itself can also be derived. If we consider $w'(t) = F(t, w(t))$ and $\tilde{w}'(t) = F(t, \tilde{w}(t))$ with different initial values $w(0)$ and $\tilde{w}(0)$, respectively, then assumption (2.45) implies

$$\|\tilde{w}(t) - w(t)\| \leq \|\tilde{w}(0) - w(0)\| \quad \text{for all } t \geq 0.$$

This follows directly from the corresponding stability result for the backward Euler method together with convergence for $\tau \rightarrow 0$. More direct derivations can be found in Dekker & Verwer (1984) and Hairer et al. (1993).

Stability and convergence with the θ -method, $\theta \geq \frac{1}{2}$, remains valid if condition (2.45) is replaced by

$$\mu(J(t, v)) \leq \omega \quad \text{for all } t \geq 0, v \in \mathbb{R}^m, \quad (2.56)$$

with constant $\omega \in \mathbb{R}$. In fact, if $\omega < 0$ and $\theta > \frac{1}{2}$ the above stability results can be sharpened. For example, if we consider the backward Euler method with $\omega < 0$, then we obtain from Theorem 2.13 and (2.52),

$$\|\varepsilon_{n+1}\| \leq \kappa (\|\varepsilon_n\| + \tau \|\rho_n\|), \quad \kappa = \frac{1}{1 - \tau\omega}.$$

This leads to the global estimate

$$\|\varepsilon_n\| \leq \kappa^n \|\varepsilon_0\| + \tau \frac{1 - \kappa^n}{1 - \kappa} \max_{0 \leq j < n} \|\rho_j\|.$$

Since $\omega < 0$ we have $\kappa = (1 + \tau|\omega|)^{-1} < 1$ and $\tau/(1 - \kappa) = |\omega|^{-1} + \tau$, and consequently

$$\|\varepsilon_n\| \leq (1 + \tau|\omega|)^{-n} \|\varepsilon_0\| + (|\omega|^{-1} + \tau) \max_{0 \leq j < n} \|\rho_j\|.$$

We thus see that the initial error will be damped and the contribution of the local errors can be bounded uniformly for $n \geq 0$. So for problems with a truly negative logarithmic norm, the implicit Euler method is not only nonlinearly stable, it has also a favourable local error build-up over arbitrarily long time intervals. With $\theta \in (\frac{1}{2}, 1)$ a similar result can be obtained in inner product norms. For the trapezoidal rule this does not apply, since this method is not strongly A -stable and we always get a growth factor $\kappa \geq 1$.

With inner product norms, assumption (2.56) can also be expressed as

$$\langle \tilde{v} - v, F(t, \tilde{v}) - F(t, v) \rangle \leq \omega \|\tilde{v} - v\|^2 \quad \text{for all } t \geq 0, \tilde{v}, v \in \mathbb{R}^m. \quad (2.57)$$

Equivalence of this formulation can be shown by using an argument similar to (2.48)-(2.50). Inequality (2.57) is called a *one-sided Lipschitz condition* with one-sided Lipschitz constant ω . This condition has been introduced in the analysis of numerical ODE methods by Dahlquist (1975) to generalize his concept of A -stability for linear systems to stiff nonlinear systems.

The Importance of the Linear Stability Theory

Given a problem $w'(t) = F(t, w(t))$, it is in general not an easy task to establish property (2.56) in some suitable norm with a logarithmic norm estimate ω which is independent of the classical Lipschitz constant L . As pointed out earlier, for stiff problems this independence is a necessity. For specific stiff ODE systems originating from spatial discretization of certain simple PDE problems, verification of the assumption is feasible as we will see in subsequent sections using the standard L_p -norms.

When the nonlinear theory based on property (2.56) is not applicable, a heuristic stability consideration for nonlinear problems $w'(t) = F(t, w(t))$ is still possible by using the principle of local linearization and to apply the linear stability theory. Suppose at time \bar{t} the solution value $w(\bar{t})$ is perturbed to $\tilde{w}(\bar{t})$, and let $\tilde{w}(t)$ be the resulting perturbed solution of the ODE system in a right neighbourhood of \bar{t} . Using the mean-value theorem for vector valued functions, it follows that

$$\frac{d}{dt} (\tilde{w}(t) - w(t)) = \left(\int_0^1 J(t, w(t) + \sigma(\tilde{w}(t) - w(t))) d\sigma \right) (\tilde{w}(t) - w(t)) \quad (2.58)$$

for $t \geq \bar{t}$. A linearization along the sought solution $w(t)$ is still not tractable for a numerical stability analysis. However, by ‘freezing’ the integrated Jacobian matrix at $t = \bar{t}$, the system (2.58) is for small perturbations well approximated near \bar{t} by the tractable constant coefficient linear system

$$v'(t) = Av(t), \quad A = J(\bar{t}, w(\bar{t})), \quad (2.59)$$

describing the evolution of $v(t) \approx \tilde{w}(t) - w(t)$. Because a numerical discretization of the general nonlinear problem $w'(t) = F(t, w(t))$ can be linearized in a similar way, the constant coefficient linear system (2.59) is often used to carry out a numerical stability analysis for the general nonlinear problem $w'(t) = F(t, w(t))$.

In many nonlinear applications the linear analysis indeed gives valid information on the stability behaviour which makes the linear stability theory important for practical purposes. The linear analysis can be expected to be successful if the instability in a numerical calculation evolves from small sized perturbations or errors – which is usually valid – and if the step from (2.58) to (2.59) is sufficiently precise. Then, for t in a right neighbourhood of \bar{t} , system (2.59) describes how small initial perturbations $v(\bar{t})$ will develop. Some care should be exercised here since there exist problems for which the ‘freezing’ step does not work in the sense that the spectrum of the frozen Jacobian matrix in (2.59) does not give the correct information concerning the qualitative solution behaviour of the linear system (2.58), see for instance Dekker & Verwer (1984, Sect. 1.1).

The Cumulative Effect of Iteration and Round-off Errors

In the above we have discussed discretization errors assuming exact arithmetic and exact solution of the implicit algebraic relations. In actual applications the algebraic relations will be solved by a Newton type iteration with a prescribed tolerance. Also large linear systems are usually solved iteratively. In addition there will be rounding errors due to finite precision arithmetic.

The cumulative effect of iteration and round-off can be estimated in the same manner as the cumulative effect of local discretization errors. Consider an implemented θ -scheme

$$\tilde{w}_{n+1} = \tilde{w}_n + (1 - \theta)\tau F(t_n, \tilde{w}_n) + \theta\tau F(t_{n+1}, \tilde{w}_{n+1}) + r_n,$$

with actually computed values \tilde{w}_n and with r_n representing errors due to rounding and inexact algebraic solutions. A typical stability estimate such as (2.53) will now read

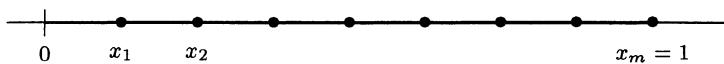
$$\|\tilde{w}_n - w_n\| \leq \|\tilde{w}_0 - w_0\| + \frac{1}{\tau} t_n \max_{0 \leq j < n} \|r_j\|.$$

With fixed t_n and a fixed upper bound for $\max \|r_j\|$, this reveals an unfavourable error propagation when τ becomes too small. It should be noted however that such deterministic estimates can be somewhat pessimistic. Both rounding errors and iteration errors are best viewed as random quantities and a statistical analysis is more appropriate. For such an analysis we refer to the classical textbook of Henrici (1962).

3 Basic Spatial Discretizations

In the construction and analysis of numerical methods for time-dependent PDEs one often considers the discretization of spatial operators like ∂_x and ∂_{xx} separately from that of the temporal derivative operator. By first discretizing the spatial operators on a chosen space grid, the PDE is converted into a system of ODEs, called the *semi-discrete system*, to which an appropriate time integration method is applied to obtain a fully discrete numerical solution.

Time stepping issues will be considered in Section 6. In this section we will introduce some simple spatial discretizations for the periodic constant-coefficient advection-diffusion problem (1.9), (1.10) on a uniform grid $\Omega_h = \{x_1, x_2, \dots, x_m\}$ with grid points $x_j = jh$ and mesh width $h = 1/m$.



On this space grid, approximations $w_j(t)$ to $u(x_j, t)$ are found by replacing the spatial derivatives by difference quotients. This gives a *finite difference* spatial discretization scheme, which is still time-continuous. For the linear periodic advection-diffusion problem the semi-discrete system will be of the form

$$w'_j(t) = \sum_k c_k w_{j+k}(t),$$

or in vector notation, with $w = (w_1, w_2, \dots, w_m)^T \in \mathbb{R}^m$,

$$w'(t) = Aw(t).$$

In this section we will use, in a somewhat heuristic manner, the notion *order* q for the spatial discretization scheme if insertion of exact PDE values yields an $\mathcal{O}(h^q)$ residual,

$$u_t(x_j, t) = \sum_k c_k u(x_{j+k}, t) + \mathcal{O}(h^q),$$

where the solution $u(x, t)$ is assumed to be sufficiently often differentiable. A full discussion of spatial errors, and orders of consistency and convergence is postponed until the next section. First, several examples of finite difference spatial discretizations will be introduced and discussed.

3.1 Discrete Fourier Decompositions

Discrete Fourier decomposition is an important tool to analyze linear difference schemes with spatial periodicity conditions. It can be used to study fundamental properties like stability, convergence, dissipation and dispersion.

Recall from Section 1.2 that the Fourier modes

$$\varphi_k(x) = e^{2\pi i k x}, \quad k \in \mathbb{Z},$$

form an orthonormal basis for the function space $L_2[0, 1]$. Here we consider the *discrete Fourier modes*

$$\phi_k = (\varphi_k(x_1), \varphi_k(x_2), \dots, \varphi_k(x_m))^T \in \mathbb{C}^m, \quad k \in \mathbb{Z}. \quad (3.1)$$

These vectors in \mathbb{C}^m can also be regarded as a grid function defined on Ω_h . In this sense ϕ_k is just the restriction of $\varphi_k(x)$ to the grid. For vectors¹⁷⁾ $v, w \in \mathbb{C}^m$ we will consider the discrete L_2 -inner product and corresponding norm,

$$\langle v, w \rangle_2 = h \sum_{j=1}^m \bar{v}_j w_j, \quad \|v\|_2 = \sqrt{\langle v, v \rangle_2}.$$

¹⁷⁾ As a rule, the j th component of a vector v in \mathbb{C}^m will be denoted by v_j . The discrete Fourier mode ϕ_k itself is a vector in \mathbb{C}^m and its j th component will be denoted by $(\phi_k)_j$. The context in which these notations are used should prevent confusion.

For the discrete Fourier modes we have

$$\langle \phi_k, \phi_l \rangle_2 = h \sum_{j=1}^m e^{2\pi i(l-k)x_j} = h \sum_{j=1}^m \rho^j, \quad \rho = e^{2\pi i(l-k)h}.$$

If $k = l \bmod m$, then $\rho = 1$ and $\langle \phi_k, \phi_l \rangle_2 = 1$. Otherwise

$$\langle \phi_k, \phi_l \rangle_2 = h\rho \frac{1 - \rho^m}{1 - \rho} = 0 \quad \text{since } \rho^m = e^{2\pi i(l-k)m} = 1.$$

It follows that

$$\{\phi_1, \phi_2, \dots, \phi_m\} \text{ is an orthonormal basis for } \mathbb{C}^m. \quad (3.2)$$

Hence any vector $v \in \mathbb{C}^m$ can be written as

$$v = \sum_{k=1}^m \alpha_k \phi_k, \quad (3.3)$$

and the coefficients are given by

$$\alpha_k = \langle \phi_k, v \rangle_2 = h \sum_{j=1}^m (\overline{\phi_k})_j v_j. \quad (3.4)$$

Note that (3.3) and (3.4) are the discrete analogues of the formulas that have been considered in Section 1.2 for functions in $L_2[0, 1]$. We will therefore refer to the α_k as Fourier coefficients and the inversion formula (3.3) will be called a (discrete) Fourier decomposition. The relation

$$\|v\|_2^2 = \langle v, v \rangle_2 = \sum_{k,l} \overline{\alpha_k} \alpha_l \langle \phi_k, \phi_l \rangle_2 = \sum_{k=1}^m |\alpha_k|^2 \quad (3.5)$$

is the discrete counterpart of Parseval's identity.

Let us next consider the initial value problem for the linear system of ODEs

$$w'(t) = Aw(t), \quad w(0) = \sum_{k=1}^m \alpha_k \phi_k, \quad (3.6)$$

where the constant coefficient matrix A is a real $m \times m$ *circulant* matrix defined as

$$A = \begin{pmatrix} c_0 & c_1 & \cdot & \cdot & c_{m-1} \\ c_{m-1} & c_0 & c_1 & \cdot & c_{m-2} \\ \cdot & c_{m-1} & c_0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & c_1 \\ c_1 & c_2 & \cdot & c_{m-1} & c_0 \end{pmatrix}. \quad (3.7)$$

As we will see below, circulant matrices arise in the discretization of periodic PDE problems with constant coefficients. A special property of circulant

matrices A is that every discrete Fourier mode ϕ_k is an eigenvector with corresponding eigenvalue ¹⁸⁾

$$\lambda_k = \sum_{j=0}^{m-1} c_j e^{2\pi i k x_j}. \quad (3.8)$$

A simple calculation then shows that we may express the solution $w(t)$ of system (3.6) by the discrete Fourier decomposition

$$w(t) = \sum_{k=1}^m \alpha_k e^{\lambda_k t} \phi_k. \quad (3.9)$$

Usually we will deal with circulant matrices where all λ_k have non-positive real part, and then (3.5) shows

$$\|w(t)\|_2^2 = \sum_{k=1}^m |\alpha_k e^{\lambda_k t}|^2 \leq \sum_{k=1}^m |\alpha_k|^2 = \|w(0)\|_2^2, \quad t \geq 0.$$

Since $w(t) = e^{tA}w(0)$, this reveals that

$$\|e^{tA}\|_2 \leq 1 \quad \text{for all } t \geq 0,$$

and thus the system (3.6) is stable in the L_2 -norm. If A is skew-symmetric all eigenvalues are purely imaginary, and then we have $\|w(t)\|_2 = \|w(0)\|_2$ for all $t \geq 0$, showing that the solutions are L_2 -norm invariant.

It is instructive to reformulate the above in terms of matrices as in the Sections 2.2, 2.3. Let V be the scaled Fourier matrix

$$V = \sqrt{h} [\phi_1, \phi_2, \dots, \phi_m] \in \mathbb{C}^{m \times m} \quad (3.10)$$

and denote $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ with λ_k the eigenvalues of the circulant matrix A . The matrix V is unitary ¹⁹⁾ and we have

$$A = V \Lambda V^{-1}.$$

Consequently, $e^{tA} = V e^{t\Lambda} V^{-1}$ and

$$\|e^{tA}\|_2 = \|e^{t\Lambda}\|_2 = \max_{1 \leq k \leq m} |e^{\lambda_k t}|. \quad (3.11)$$

Remark 3.1 Because $\phi_k = \phi_l$ if $k = l \bmod m$, the numbering for the basis is not unique. For example, one could also use

$$\{\phi_{-k}, \phi_{-k+1}, \dots, \phi_{m-k-1}\} \quad \text{with } k = \begin{cases} \frac{1}{2}m & \text{if } m \text{ is even,} \\ \frac{1}{2}(m-1) & \text{if } m \text{ is odd.} \end{cases}$$

¹⁸⁾ Exercise: Show that $A\phi_k = \lambda_k \phi_k$ for any k .

¹⁹⁾ Exercise: Show that $(V^*V)_{jk} = h\phi_j^*\phi_k = \langle \phi_j, \phi_k \rangle_2$.

This basis reveals the important fact that on the uniform space grid Ω_h we can actually only represent Fourier modes with wave numbers

$$|k| \leq \lfloor \frac{1}{2}m \rfloor$$

with $\lfloor x \rfloor$ denoting downward integer rounding. Mostly we will retain the numbering $1, 2, \dots, m$, but it should be noted that

$$\phi_m = \phi_0 = (1, 1, \dots, 1)^T, \quad \phi_{m-j} = \phi_{-j} = \overline{\phi_j},$$

and so in the sequence $\phi_1, \phi_2, \dots, \phi_m$ the high wave number modes are in fact the ones with $k \approx \lfloor m/2 \rfloor$. \diamond

Remark 3.2 Component-wise, (3.3) and (3.4) read, respectively,

$$\begin{aligned} v_j &= \sum_{k=1}^m \alpha_k e^{2\pi i k j / m}, \quad j = 1, \dots, m, \\ \alpha_k &= \frac{1}{m} \sum_{j=1}^m v_j e^{-2\pi i k j / m}, \quad k = 1, \dots, m. \end{aligned}$$

The vector $\alpha = (\alpha_1, \dots, \alpha_m)^T$ is called the *discrete Fourier transform* of the vector v . When implemented in a straightforward way this transformation and its inverse involves a matrix-vector multiplication requiring m^2 multiplications with exponentials. However it can be done in a recursive manner by the so-called Fast Fourier Transform (FFT) algorithm, which reduces this number to about $m \log_2 m$ multiplications. The invention of the FFT, by Cooley & Tukey (1965), has increased the use of Fourier transforms in numerical mathematics enormously. The FFT is nowadays the computational workhorse of many numerical codes. The basic principle of the FFT is explained in a number of textbooks, for instance Strang (1986, Sect. 5.5). \diamond

3.2 The Advection Equation

Consider the constant coefficient advection equation

$$u_t + au_x = 0,$$

with periodicity in space $u(x \pm 1, t) = u(x, t)$ and a given initial function $u(x, 0)$. We will replace the spatial derivative u_x by a finite difference approximation to arrive at a semi-discrete system where $w_j(t) \approx u(x_j, t)$ on Ω_h .

The difference formula

$$\frac{1}{h} (u(x-h) - u(x)) = -u_x(x) + \mathcal{O}(h), \quad (3.12)$$

leads to the *first-order upwind* advection discretization

$$w'_j(t) = \frac{a}{h} (w_{j-1}(t) - w_j(t)), \quad j = 1, 2, \dots, m, \quad (3.13)$$

with $w_0(t) = w_m(t)$ by periodicity. For stability reasons, to be discussed later, this formula should only be used for $a > 0$. If $a < 0$ the first-order upwind scheme reads

$$w'_j(t) = \frac{a}{h} (w_j(t) - w_{j+1}(t)), \quad j = 1, 2, \dots, m, \quad (3.14)$$

with $w_{m+1}(t) = w_1(t)$. This spatial discretization thus takes the form of a system of m linear ODEs. The initial values can simply be taken as the restriction to Ω_h of the exact solution, $w_j(0) = u(x_j, 0)$. In vector notation the semi-discrete system (3.13) reads $w'(t) = Aw(t)$ with $m \times m$ circulant matrix A given by

$$A = \frac{a}{h} \begin{pmatrix} -1 & & & & 1 \\ 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & 1 & -1 & \\ & & & 1 & -1 \end{pmatrix}.$$

Likewise, the difference formula

$$\frac{1}{2h} (u(x-h) - u(x+h)) = -u_x(x) + \mathcal{O}(h^2) \quad (3.15)$$

leads to the *second-order central* advection discretization

$$w'_j(t) = \frac{a}{2h} (w_{j-1}(t) - w_{j+1}(t)), \quad j = 1, 2, \dots, m, \quad (3.16)$$

with $w_0(t) = w_m(t)$ and $w_{m+1}(t) = w_1(t)$. Here we have $w'(t) = Aw(t)$ with

$$A = \frac{a}{2h} \begin{pmatrix} 0 & -1 & & & 1 \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ -1 & & & 1 & 0 \end{pmatrix}.$$

Observe that A is here a skew-symmetric circulant matrix.

Since it is based on a second-order accurate difference formula, this central scheme is expected to be more accurate than the first-order upwind scheme, and this is true for smooth solutions. However, consider $a = 1$ and initial profile $u(x, 0) = (\sin(\pi x))^{100}$. Solutions at $t = 1$ are given in Figure 3.1 for $h = 1/50$, with dotted lines for the exact solution and solid lines for the numerical approximation. The first-order upwind scheme is not accurate. At $t = 1$ the initial pulse has nearly collapsed. However, the result of the second-order central scheme is also far from satisfactory: it gives oscillations, negative values and there is a significant phase error. With the second-order central scheme the numerical oscillations are mainly produced after steep gradients. The wiggles in front of the pulse in this picture are due to the fact that we consider here periodic problems and the numerical pulse is in fact catching up its own oscillations.

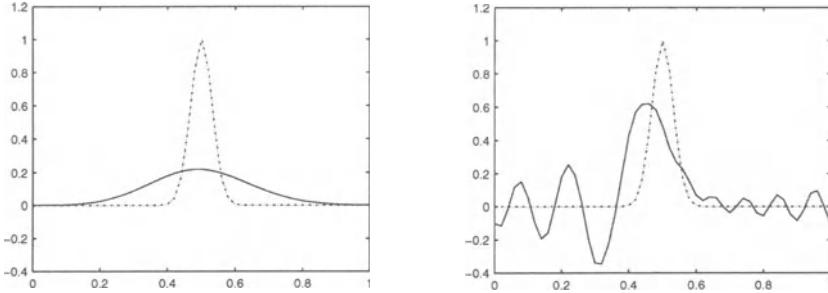


Fig. 3.1. Advection test, $h = 1/50$, with the first-order upwind scheme (left) and the second-order central scheme (right).

Modified Equations

For a sufficiently smooth profile,²⁰⁾ the upwind and central approximation will both give numerical solutions that are qualitatively correct. However, as revealed by the above test, these schemes behave truly differently for profiles which are non-smooth. Insight in the qualitative behaviour can be obtained by regarding the so-called modified equations of the discretizations.²¹⁾ We will use modified equations only in a heuristic fashion, with the aim of understanding the qualitative behaviour of approximations.

Consider the first-order upwind scheme which is based on the difference formula (3.12). Further expansion in (3.12) gives

$$\frac{1}{h} (u(x-h) - u(x)) = -u_x(x) + \frac{1}{2} hu_{xx}(x) + \mathcal{O}(h^2),$$

indicating that the numerical upwind solution will in fact be closer to the solution of the advection-diffusion equation

$$\tilde{u}_t + a\tilde{u}_x = \frac{1}{2} ah\tilde{u}_{xx} \quad (3.17)$$

than to the solution of the original advection equation $u_t + au_x = 0$. This explains the diffusive nature of the first-order upwind scheme: although we are seeking a solution to the advection equation, we are actually generating a solution that is close to an advection-diffusion equation with diffusion coefficient $\frac{1}{2}ah$. The advection-diffusion equation (3.17) is called a *modified equation* for the first-order upwind scheme. The solution of the modified equation reflects quite accurately the nature and quality of the upwind approximation.

²⁰⁾ The notions *smooth* and *non-smooth* are used in an operational numerical sense. Here smoothness is to be related to the mesh width. If h is made sufficiently small, a rapidly varying function is well represented on the grid, and so numerically we then do have a smooth case.

²¹⁾ Modified equations to study the behaviour of discretizations were introduced by Warming & Hyett (1974). A more general discussion on the subject can be found in Griffiths & Sanz-Serna (1986).

It will be demonstrated in the next section that the first-order upwind scheme provides in fact an $\mathcal{O}(h^2)$ approximation to this modified equation whereas the accuracy with respect to the original advection equation is only $\mathcal{O}(h)$. This can already be understood heuristically by expressing the first-order upwind difference formula as

$$\frac{1}{h} \left(u(x-h) - u(x) \right) = \frac{1}{2h} \left(u(x-h) - u(x+h) \right) + \frac{\epsilon}{h^2} \left(u(x-h) - 2u(x) + u(x+h) \right),$$

where $\epsilon = \frac{1}{2}h$. The first difference quotient at the right approximates u_x with second-order accuracy. Similarly, the second difference quotient at the right approximates u_{xx} with order two. So the first-order upwind scheme can be viewed as the second-order advection central scheme plus an *artificial diffusion* term. This artificial diffusion, also called numerical diffusion, has diffusion coefficient $\frac{1}{2}ah$. Hence it will disappear if $h \rightarrow 0$, but for practical choices of h it is not negligible and this is the cause for the severe smoothing and near collapse of the initial profile in Figure 3.1.

In the same way we can consider the second-order central scheme. A further expansion in the difference formula (3.15) gives

$$\frac{1}{2h} \left(u(x-h) - u(x+h) \right) = -u_x(x) - \frac{1}{6} h^2 u_{xxx}(x) + \mathcal{O}(h^4),$$

revealing that the second-order central discretization leads to a fourth-order approximation for the modified equation

$$\tilde{u}_t + a\tilde{u}_x = -\frac{1}{6} ah^2 \tilde{u}_{xxx}. \quad (3.18)$$

A proof of fourth-order convergence towards this modified equation will be given in the next section. The term \tilde{u}_{xxx} gives rise to *dispersion*: Fourier modes $\varphi_k(x)$ travel with a velocity that depends on the wave number k . With initial value $\tilde{u}(x,0) = \varphi_k(x) = e^{2\pi i k x}$ the solution of this modified equation is given by

$$\tilde{u}(x,t) = e^{2\pi i k(x-a_k t)} = \varphi_k(x - a_k t), \quad a_k = a \left(1 - \frac{2}{3}\pi^2 k^2 h^2 \right),$$

and we see that all Fourier modes move with different speeds. An initial pulse such as $u(x,0) = (\sin(\pi x))^{100}$ in the above experiment can be viewed as a finely tuned sum of Fourier modes. If the individual Fourier modes travel with different velocities then this fine-tuning is lost and oscillations will occur; this is somewhat related to the Gibbs phenomenon discussed in Remark 1.2. The resulting oscillations are very clearly visible in Figure 3.1.

Notable is that the advection-dispersion equation (3.18) has right and left traveling Fourier modes as solution, even if the advective velocity a is positive. Solutions of equation (3.18) may consist of Fourier modes $\varphi_k(x)$ with arbitrary wave numbers $k \in \mathbb{Z}$, whereas on our spatial grid we can in fact only represent modes with wave number $|k| \leq m/2$, see Remark 3.1.

When considering a modified equation for a difference scheme this should always be kept in mind and modified equations should be regarded with some care when examining the qualitative behaviour of a difference scheme.

If the initial profile is very smooth then the Fourier coefficients for the high-frequency modes will be extremely small, see Remark 1.2, so then the dispersive nature will be hardly felt and no large oscillations will be seen. However, if the initial profile has large gradients, some high-frequency modes will be significant and the dispersive nature of the central approximation will cause quite large oscillations (wiggles) as has been illustrated in Fig. 3.1. The same holds for the artificial diffusion with the upwind scheme. The dilemma of either accepting artificial diffusion by upwinding or oscillations and negative values due to a central discretization will reappear on numerous occasions.

Fourier Analysis

Since we are considering a constant-coefficient equation with periodicity in space, the upwind scheme (3.13) and the central scheme (3.16) yield a semi-discrete system $w'(t) = Aw(t)$ with circulant matrix A . Starting with a single discrete Fourier mode $w(0) = \phi_k$ the solution is given by $w(t) = e^{\lambda_k t} \phi_k$ with λ_k the corresponding eigenvalue of A . Component-wise this reads

$$w_j(t) = e^{\lambda_k t} e^{2\pi i k x_j}. \quad (3.19)$$

By comparing this numerical evolution of a Fourier mode with the corresponding exact PDE evolution

$$u(x_j, t) = e^{-2\pi i k a t} e^{2\pi i k x_j}, \quad (3.20)$$

we can further unravel the properties of the finite difference schemes for the periodic advection problem.

The eigenvalues λ_k of A are thus to be compared with the corresponding eigenvalues $-2\pi i k a$ of the PDE operator $a \partial_x$. A simple computation yields

$$\lambda_k = \frac{|a|}{h} \left(\cos(2\pi k h) - 1 \right) - \frac{ia}{h} \sin(2\pi k h), \quad k = 1, \dots, m, \quad (3.21)$$

for the first-order upwind scheme and

$$\lambda_k = -\frac{ia}{h} \sin(2\pi k h), \quad k = 1, \dots, m, \quad (3.22)$$

for the second-order central scheme. In Figure 3.2 these eigenvalues are plotted for $a = 1$ and $m = 100$.

The eigenvalues λ_k in these schemes all have non-positive real part, and consequently the solution operator e^{At} satisfies $\|e^{At}\|_2 \leq 1$ for $t \geq 0$. Hence the semi-discrete system is stable in the L_2 -norm for both discretizations. For the second-order central scheme all λ_k are purely imaginary (the matrix A is skew-symmetric), so there we have L_2 -norm invariance, $\|w(t)\|_2 = \|w(0)\|_2$.

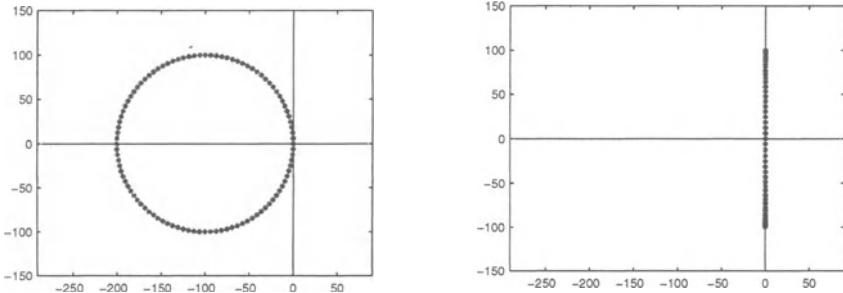


Fig. 3.2. Eigenvalues λ_k for the first-order upwind advection discretization (left) and second-order central advection discretization (right) with $m = 100$ and $a = 1$.

Note that if the formula (3.13) were used for $a < 0$ the eigenvalues would be in the right half-plane with real part as large as $|a| h^{-1}$, and thus this scheme becomes *unstable* when $h \rightarrow 0$. For fixed small $h > 0$ the high frequency Fourier modes are very strongly amplified leading to totally wrong numerical solutions or even computational overflow. If $a < 0$ the correct upwind scheme is given by (3.14). This is also obvious from a physical interpretation by characteristics of the advection equation: if $a < 0$ the time evolution of the exact solution $u(x_j, t + \tau)$ for $\tau > 0$ at the point x_j is determined by the values $u(x, t)$ with $x > x_j$. This will be discussed in greater detail in Section 6 in connection with the so-called CFL condition.

Remark 3.3 We can also relate the order of the spatial discretizations to the difference between $-2\pi i a k$ and λ_k for fixed k and $h \rightarrow 0$. For the first-order upwind and second-order central schemes we find, respectively,

$$\lambda_k = -2\pi i a k - \frac{1}{2} |a|(2\pi k)^2 h + \mathcal{O}(h^2),$$

$$\lambda_k = -2\pi i a k + \frac{1}{6} i a (2\pi k)^3 h^2 + \mathcal{O}(h^4).$$

For h sufficiently small, these expansions are accurate for the low wave numbers k . However, for the greater part of the wave numbers the true PDE eigenvalues $-2\pi i a k$ are badly approximated. When associated Fourier coefficients α_k in the Fourier decomposition are significant, inaccurate approximations will result. We have already encountered this in Figure 3.1. On the other hand, if the initial function is very smooth, that is, α_k is sufficiently small for larger wave numbers, the inaccurate eigenvalue approximation will not be felt.

These accuracy considerations can be put in a mathematical framework by considering the error vector $\varepsilon(t) = (\varepsilon_j(t)) \in \mathbb{C}^m$ with components

$$\varepsilon_j(t) = u(x_j, t) - w_j(t), \quad j = 1, 2, \dots, m.$$

If we assume for example that the initial function can be well represented on the grid,

$$w(0) = \sum_{k=1}^m \alpha_k \phi_k, \quad \varepsilon(0) = 0,$$

then we have in the L_2 -norm, by Parseval's identity,

$$\|\varepsilon(t)\|_2^2 = \sum_{k=1}^m |\alpha_k|^2 |e^{-2\pi i a_k t} - e^{\lambda_k t}|^2.$$

A full convergence analysis based on these Fourier decompositions becomes quite complicated and technical. An easier approach, based on local truncation errors, will be presented in the next section. \diamond

Artificial Dissipation and Dispersion

Comparison of λ_k with its exact PDE counterpart is also useful for getting further insight in *artificial dissipation* and *artificial dispersion* properties. For this purpose we rewrite (3.19) and (3.20) as

$$w_j(t) = e^{t \operatorname{Re} \lambda_k} e^{2\pi i k (x_j - a_k t)}, \quad a_k = -\frac{1}{2\pi k} \operatorname{Im} \lambda_k, \quad (3.23)$$

$$u(x_j, t) = e^{2\pi i k (x_j - at)}. \quad (3.24)$$

The factor $e^{t \operatorname{Re} \lambda_k}$ determines the amount of numerical damping or dissipation for the k th Fourier mode. A finite difference scheme for the advection equation is called *dissipative* if $\operatorname{Re} \lambda_k < 0$ for all $k \neq m$. The scheme is called *non-dissipative* if $\operatorname{Re} \lambda_k = 0$ for all k . Obviously, the second-order central scheme is non-dissipative whereas the first-order upwind scheme is dissipative. The velocity a_k for the k th Fourier mode is called the numerical *phase velocity*. When the phase velocities differ from a this will lead to a phase error. If they are different from each other we will have dispersion. As discussed before, dispersion may give rise to oscillations.

The first-order upwind and second-order central advection discretizations have the same phase velocities $a_k \neq a$. So the upwind scheme is also dispersive, but oscillations will not show up in an actual calculation because the damping factor $e^{t \operatorname{Re} \lambda_k}$ suppresses all high-frequency Fourier modes.

It is obvious that both the first-order upwind scheme (3.13) and the second-order central scheme (3.16) scheme have drawbacks, the first being too dissipative and the second too dispersive. All finite difference advection schemes suffer from either one of these two artefacts. However, by increasing the order of the difference schemes the adverse effects of dissipation and dispersion can be diminished.

Higher-Order Schemes

A general spatial finite difference scheme for the periodic advection problem can be written as

$$w'_j(t) = \frac{a}{h} \sum_{k=-r}^s \gamma_k w_{j+k}(t), \quad j = 1, 2, \dots, m, \quad (3.25)$$

with $w_i(t) = w_{i+m}(t)$ to impose the periodicity condition. Here we use r grid points to the left and s to the right of x_j so that in total $r + s + 1$ points are used. The set $\{x_{j-r}, \dots, x_{j+s}\}$ is called the *stencil* of the discretization around x_j . With wider stencils higher orders can be achieved. The conditions for order q are found by substituting the exact PDE values into the scheme and then using a Taylor expansion to estimate the residual,

$$\begin{aligned} u_t(x, t) - \frac{a}{h} \sum_k \gamma_k u(x + kh, t) \\ = -au_x - \frac{a}{h} \sum_k \gamma_k (u + kh u_x + \frac{1}{2} k^2 h^2 u_{xx} + \dots) \Big|_{(x,t)} \\ = -\frac{a}{h} \sum_k \gamma_k u - a \left(1 + \sum_k k \gamma_k \right) u_x - \frac{a}{2} h \sum_k k^2 \gamma_k u_{xx} - \dots \Big|_{(x,t)}. \end{aligned}$$

The conditions for order q thus read

$$\sum_k \gamma_k = 0, \quad \sum_k k \gamma_k = -1, \quad \sum_k k^2 \gamma_k = 0, \dots, \quad \sum_k k^q \gamma_k = 0. \quad (3.26)$$

These equations for the coefficients $\gamma_{-r}, \dots, \gamma_s$ can be satisfied for $q \leq r+s$.²²⁾ Schemes with order $q = r+s$ are called *optimal-order* schemes. For each r and s there is precisely one such scheme.

There exists a fundamental result on the L_2 -stability of these optimal schemes, due to Iserles & Strang (1983):

If $a > 0$, the optimal-order schemes with $q = r+s$ are stable for $s \leq r \leq s+2$ and unstable otherwise.

A similar result can be formulated for $a < 0$ with the condition $r \leq s \leq r+2$. A proof for the stability of the schemes for $a > 0$ with $r = s, s+1, s+2$ can be found in Iserles & Nørsett (1991, Sect. 6.1). In the same book also the instability of the other schemes is demonstrated, but this is much more complicated and relies on the theory of *order stars*. We note that the sufficiency for stability was proved already by Strang (1962) for fully discrete schemes (see Section III.2).

²²⁾ The order conditions form a linear algebraic system for the coefficients with a Vandermonde type matrix. Such matrices have full rank, see Horn & Johnson (1985).

For $a > 0$ and $r = 2, s = 1$ we obtain the *third-order upwind-biased*²³⁾ advection scheme

$$w'_j(t) = \frac{a}{h} \left(-\frac{1}{6} w_{j-2}(t) + w_{j-1}(t) - \frac{1}{2} w_j(t) - \frac{1}{3} w_{j+1}(t) \right). \quad (3.27)$$

For $a < 0$ the scheme reads

$$w'_j(t) = \frac{a}{h} \left(\frac{1}{3} w_{j-1}(t) + \frac{1}{2} w_j(t) - w_{j+1}(t) + \frac{1}{6} w_{j+2}(t) \right), \quad (3.28)$$

which is a reflection of (3.27) around the point x_j . The modified equation for this discretization, which is approximated with $\mathcal{O}(h^4)$ accuracy, reads

$$\tilde{u}_t + a\tilde{u}_x = -\frac{1}{12} |a| h^3 \tilde{u}_{xxxx}.$$

The term $-\tilde{u}_{xxxx}$ is a higher-order dissipation term giving damping of the high-frequency Fourier modes, but it still allows some oscillations and over/undershoot. It should be noted that with an equation of the type $u_t = -u_{xxxx}$ all Fourier modes are damped, but, unlike the diffusion equation $u_t = u_{xx}$, this equation does not satisfy a maximum principle. For instance, if $u(x, 0) = 1 - \cos(2\pi x)$ then $u(0, 0) = 0$, $u_t(0, 0) = -(2\pi)^4 < 0$ and consequently $u(0, t)$ will be negative for small $t > 0$.

Figure 3.3 gives the numerical solution at $t = 1$ for $h = 1/50$ with initial profile $u(x, 0) = (\sin(\pi x))^{100}$, the same as in Figure 3.1. We see that the third-order upwind-biased discretization indeed gives some oscillations, but these are rather small. Moreover, the numerical solution is good in phase with the exact solution, which is in accordance with the modified equation.

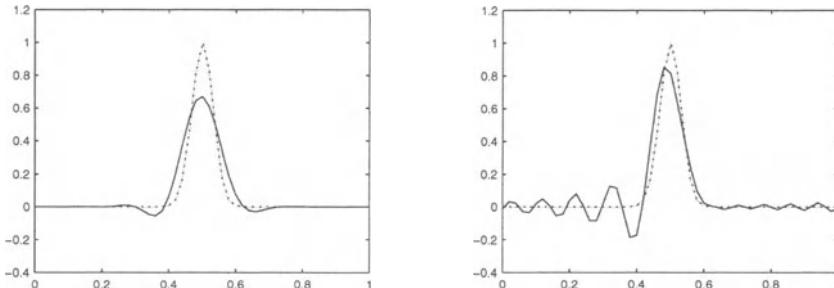


Fig. 3.3. Advection test, $h = 1/50$, with the third-order upwind-biased scheme (left) and the fourth-order central scheme (right).

For $r = s = 2$ we get the *fourth-order central* advection scheme

$$w'_j(t) = \frac{a}{h} \left(-\frac{1}{12} w_{j-2}(t) + \frac{2}{3} w_{j-1}(t) - \frac{2}{3} w_{j+1}(t) + \frac{1}{12} w_{j+2}(t) \right). \quad (3.29)$$

²³⁾ We use the term upwind-biased to distinguish the scheme from the third-order *fully-upwind* scheme, which has $r = 3, s = 0$. Note that this fully-upwind scheme is not stable. Stability is also the reason why (3.27) should not be used for $a < 0$.

The modified equation will now only contain dispersion terms and no damping is built in. For non-smooth solutions this gives strong oscillations. Figure 3.3 illustrates this (same initial profile and mesh width as before). The pattern is typical for central schemes. Note that comparison with Figure 3.1 shows some improvement over the second-order scheme. With smoother profiles or smaller mesh widths this improvement would be more pronounced.

Substitution of the discrete Fourier modes ϕ_k , $1 \leq k \leq m$, into these discretizations gives the eigenvalues

$$\lambda_k = -\frac{4}{3} \frac{|a|}{h} \sin^4(\pi kh) - \frac{i}{3} \frac{a}{h} \sin(2\pi kh)(4 - \cos(2\pi kh)) \quad (3.30)$$

for the third-order upwind-biased scheme and

$$\lambda_k = -\frac{i}{3} \frac{a}{h} \sin(2\pi kh)(4 - \cos(2\pi kh)) \quad (3.31)$$

for the fourth-order central scheme. In both cases we have $\operatorname{Re} \lambda_k \leq 0$ for all k , showing stability of the discretizations. The upwind eigenvalues all have a negative real part, except $\lambda_m = 0$, and the central eigenvalues are all purely imaginary, due to the skew-symmetry. Figure 3.4 gives a plot of the eigenvalues for $m = 100$ and $a = 1$, similar to Figure 3.2 for first-order upwind and second-order central. Note that, although there is damping, many eigenvalues of the third-order upwind-biased scheme are very close to the imaginary axis.

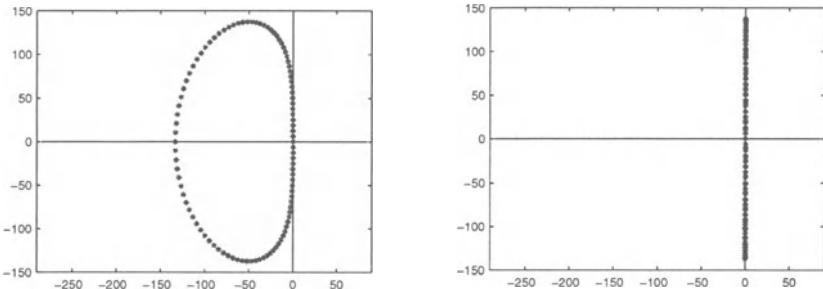


Fig. 3.4. Eigenvalues λ_k for the third-order upwind-biased advection discretization (left) and the fourth-order central advection discretization (right) with $m = 100$ and $a = 1$.

Interestingly, the imaginary parts of the λ_k are equal for these two schemes. This means that they have in a sense the same phase error properties. The damping in the upwind-biased scheme, however, reduces the high frequency dispersion errors of the central scheme that give rise to strong oscillations. Moreover this damping is applied to those Fourier modes that have the largest phase errors, and therefore the total pulse in Figure 3.2 is

more in phase with the upwind-biased scheme. Both schemes are often used in practice, in particular the third-order upwind-biased scheme.

3.3 The Diffusion Equation

We next consider the constant coefficient diffusion equation

$$u_t = du_{xx}$$

with $d > 0$, again assuming periodicity $u(x \pm 1, t) = u(x, t)$ in space. It will become evident that the spatial discretization of this diffusion equation is more straightforward than with the advection equation.

The spatial derivative can be approximated by the difference formula

$$\frac{1}{h^2} (u(x-h) - 2u(x) + u(x+h)) = u_{xx}(x) + \mathcal{O}(h^2). \quad (3.32)$$

Using this second-order accurate formula on the uniform space grid Ω_h leads to the *second-order central* diffusion discretization

$$w'_j(t) = \frac{d}{h^2} (w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)), \quad j = 1, 2, \dots, m, \quad (3.33)$$

with $w_0(t) = w_m(t)$ and $w_{m+1}(t) = w_1(t)$ to impose the periodicity condition. As for the advection problem, we introduce the vector $w = (w_1, \dots, w_m)^T$ and rewrite (3.33) as an ODE system $w'(t) = Aw(t)$ for $t \geq 0$ with given initial value $w(0)$ and with A given by the symmetric circulant matrix

$$A = \frac{d}{h^2} \begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{pmatrix}.$$

Due to the symmetry the matrix A has real eigenvalues. This semi-discrete system represents the most simple spatial discretization for the linear diffusion equation $u_t = du_{xx}$.

Modified Equations

Similar as for the advection discretizations we can look for a modified equation of the scheme (3.33). A further expansion in (3.32) gives

$$\frac{1}{h^2} (u(x-h) - 2u(x) + u(x+h)) = u_{xx}(x) + \frac{1}{12} h^2 u_{xxxx}(x) + \mathcal{O}(h^4),$$

revealing

$$\tilde{u}_t = d\tilde{u}_{xx} + \frac{1}{12} dh^2 \tilde{u}_{xxxx}$$

as a *modified equation* which is approximated with order four by the second-order central difference scheme. However, heuristic insight by this modified equation into the qualitative behaviour of the diffusion discretization scheme is awkward. This modified equation is not stable: if we choose as initial function the Fourier mode $\tilde{u}(x, 0) = e^{2\pi i k x}$ we find as solution

$$\tilde{u}(x, t) = e^{-4d\pi^2 k^2 (1 - \frac{1}{3}\pi^2 h^2 k^2)t} e^{2\pi i k x},$$

which grows exponentially for $h^2 k^2 > 3/\pi^2$. The fourth derivative \tilde{u}_{xxxx} renders the modified equation unstable. This instability is an artefact in the sense that the modified equation admits solutions composed of Fourier modes $\varphi_k(x)$ with arbitrary wave number k , whereas a difference scheme on Ω_h can only represent modes with wave number $|k| \leq \frac{1}{2}m$, see Remark 3.1. Under this restriction $|hk| \leq \frac{1}{2}$ all modes for the modified equation decay like those of the original diffusion problem. Hence under this restriction the qualitative behaviour of the modified equation corresponds with that of the exact solution.

One could also include higher-order terms into the modified equation, for example leading to

$$\tilde{u}_t = d\tilde{u}_{xx} + \frac{1}{12} dh^2 \tilde{u}_{xxxx} + \frac{1}{360} dh^4 \tilde{u}_{xxxxxx}$$

as an alternative modified equation. This equation is approximated with order six by (3.33) and it is stable which can be seen by again inserting Fourier modes. It is clear, however, that the modified equation approach, which gave easy insight in the advection discretizations, is less instructive for the diffusion problem. For advection-diffusion equations with non-zero diffusion coefficients we will therefore rely on Fourier analysis to gain insight into the properties of the discretizations.

Fourier Analysis

Starting with a single discrete Fourier mode $w(0) = \phi_k$ as initial condition, the time evolution in the difference scheme is given by

$$w_j(t) = e^{\lambda_k t} e^{2\pi i k x_j}, \quad (3.34)$$

with eigenvalue λ_k , which is to be compared to the exact PDE evolution

$$u(x_j, t) = e^{-4\pi^2 k^2 dt} e^{2\pi i k x_j}. \quad (3.35)$$

Substitution of ϕ_k into the discretization (3.33) gives the real eigenvalues

$$\lambda_k = \frac{2d}{h^2} (\cos(2\pi kh) - 1) = -\frac{4d}{h^2} \sin^2(\pi kh), \quad k = 1, \dots, m. \quad (3.36)$$

The eigenvalues λ_k are negative, showing stability of the discretization. Moreover we see that they lie between $-4dh^{-2}$ and 0, and some eigenvalues become large negative for small h . Due to the periodicity there is always a zero eigenvalue. Figure 3.5 shows the distribution of λ_k for $m = 100$ and $d = 1$.

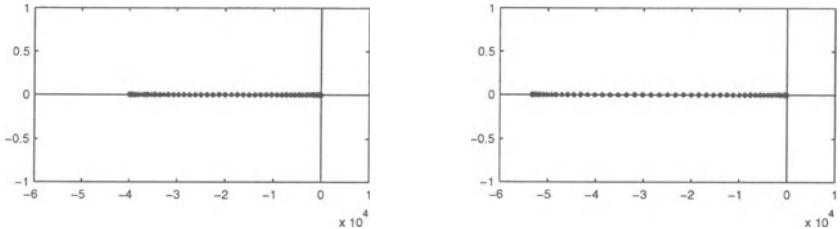


Fig. 3.5. Eigenvalues λ_k for the second-order diffusion discretization (left) and the fourth-order diffusion discretization (right) with $m = 100$ and $d = 1$.

The occurrence of large eigenvalues is typical for diffusion problems. Whereas for the advection problem the spectral radius of the difference matrix A is proportional to h^{-1} , for the diffusion problem the spectral radius is proportional to h^{-2} . Semi-discrete diffusion problems are therefore usually classified as *stiff* problems, see Section 2. With regard to time integration, this means that implicit ODE methods are used in general for diffusion problems.

For k fixed and $h \rightarrow 0$, the eigenvalue λ_k will approach its exact counterpart $-4d\pi^2k^2$ with order $\mathcal{O}(h^2)$. However, for any fixed $h > 0$ most of the λ_k are not close to their exact counterparts, no matter how small h is. While for advection problems this discrepancy readily leads to large errors with non-smooth profiles and even to wrong qualitative behaviour such as oscillations or loss of shape, for diffusion problems this discrepancy is rather harmless. We owe this to the strong decay for the higher harmonics which occurs for the true PDE solution and the finite difference approximation, see (3.34)-(3.36). This smoothing property distinguishes the diffusion problem from the advection problem and makes the spatial discretization of diffusion problems much easier. In spite of its simplicity, the central second-order discretization is therefore often successfully used in actual applications.

Figure 3.6 illustrates that the qualitative behaviour of the numerical solutions is satisfactory, even on rather coarse grids. In this figure numerical solutions are plotted at time $t = 10^{-3}$ and $t = 10^{-2}$, respectively, for the diffu-

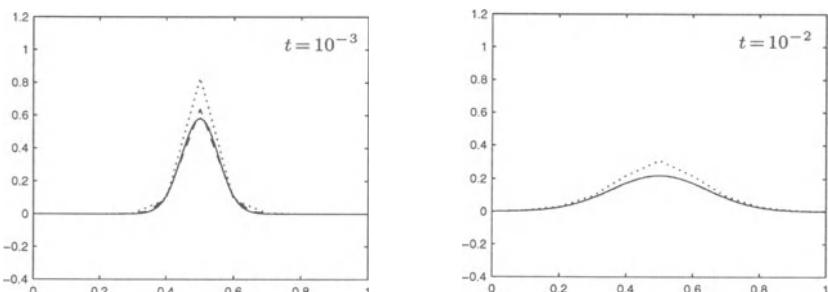


Fig. 3.6. Diffusion test for the second-order scheme at $t = 10^{-3}, 10^{-2}$ with $d = 1$ and $h = 1/10$ (dotted), $1/20$ (dashed). The solid line is the exact solution.

sion equation with $d = 1$ and initial profile $u(x, 0) = (\sin(\pi x))^{100}$. The dotted line is the numerical solution for $h = 1/10$, the dashed line for $h = 1/20$ and the solid line is the exact solution (also found numerically but with a very small h ; the numerical solution with $h = 1/40$ is already virtually the same). Note that even on the coarsest grid with $h = 1/10$ the qualitative behaviour is correct. For a better quantitative behaviour we just need a slightly smaller grid size h .

Higher-Order Schemes

A general spatial finite difference scheme for the periodic diffusion problem can be written as

$$w'_j(t) = \frac{d}{h^2} \sum_{k=-r}^s \gamma_k w_{j+k}(t), \quad j = 1, 2, \dots, m, \quad (3.37)$$

with $w_i(t) = w_{i+m}(t)$. We assume that $r = s$ and $\gamma_{-k} = \gamma_k$, that is, symmetry in space. The conditions for order q are found as before by Taylor expansions,

$$\begin{aligned} u_t(x, t) &= \frac{d}{h^2} \sum_k \gamma_k u(x + kh, t) \\ &= du_{xx} - \frac{d}{h^2} \sum_k \gamma_k \left(u + kh u_x + \frac{1}{2} k^2 h^2 u_{xx} + \dots \right) \Big|_{(x,t)} \\ &= -\frac{d}{h^2} \sum_k \gamma_k u + d \left(1 - \sum_k \frac{1}{2} k^2 \gamma_k \right) u_{xx} - \frac{d}{4!} h^2 \sum_k k^4 \gamma_k u_{xxxx} + \dots \Big|_{(x,t)}. \end{aligned}$$

Note that all odd derivatives cancel due to the symmetry. So the order will be even, and the conditions for order q are

$$\sum_k \gamma_k = 0, \quad \sum_k k^2 \gamma_k = 2, \quad \sum_k k^4 \gamma_k = 0, \dots, \quad \sum_k k^q \gamma_k = 0, \quad (3.38)$$

which can be satisfied for $q \leq 2s$.

For $s = 2$ we obtain the *fourth-order central* diffusion discretization

$$w'_j = \frac{d}{h^2} \left(-\frac{1}{12} w_{j-2} + \frac{4}{3} w_{j-1} - \frac{5}{2} w_j + \frac{4}{3} w_{j+1} - \frac{1}{12} w_{j+2} \right), \quad (3.39)$$

where $w_j = w_j(t)$. Insertion of discrete Fourier modes shows that the eigenvalues for this discretization are all on the negative real axis, between $-\frac{32}{6} dh^{-2}$ and 0, and thus the discretization is stable. The distribution of the eigenvalues is shown in Figure 3.5 for $m = 100$ and $d = 1$.

We note that the restriction to symmetric discretizations is natural since there is no preference of spatial direction in the diffusion equation. Moreover,

if $w'(t) = Aw(t)$ is a non-symmetrical semi-discrete system (3.37), then the symmetrical system

$$w'(t) = \frac{1}{2}(A + A^T)w(t)$$

can be shown to be more accurate (no dispersion terms) and at least as stable. Elaboration of this statement is left as an exercise.

3.4 The Advection-Diffusion Equation

Let us next consider the advection-diffusion problem

$$u_t + au_x = du_{xx}$$

with constant coefficients $a \in \mathbb{R}$, $d > 0$ and the spatial periodicity condition $u(x \pm 1, t) = u(x, t)$. The space discretizations we have discussed for the advection and diffusion terms can of course be combined to obtain spatial discretizations for this problem. With second-order central differences we then obtain the spatial discretization scheme

$$w'_j(t) = \left(\frac{d}{h^2} + \frac{a}{2h} \right) w_{j-1}(t) - \frac{2d}{h^2} w_j(t) + \left(\frac{d}{h^2} - \frac{a}{2h} \right) w_{j+1}(t), \quad (3.40)$$

where $j = 1, \dots, m$ and $w_0(t) = w_m(t)$, $w_{m+1}(t) = w_1(t)$. With first-order upwind advection discretization a similar formula is obtained, except that then d should be replaced by $d + \frac{1}{2}|a|h$. Fourier analysis gives the eigenvalues

$$\lambda_k = \frac{2d}{h^2} \left(\cos(2\pi kh) - 1 \right) - \frac{ia}{h} \sin(2\pi kh), \quad k = 1, \dots, m. \quad (3.41)$$

These eigenvalues are located on an ellipse in the left half-plane \mathbb{C}^- . The loci of the eigenvalues is plotted in Figure 3.7 for $m = 100$, $a = 1$ and $d = 10^{-3}, 10^{-2}$. In the right figure a similar plot is given for the fourth-order discretization which combines (3.29) and (3.39).

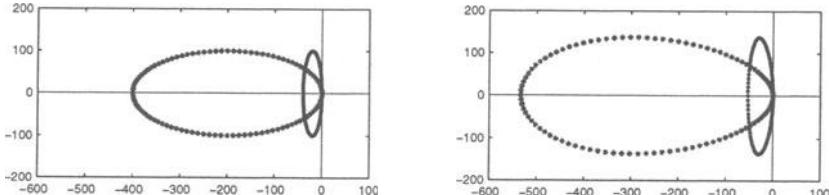


Fig. 3.7. Eigenvalues λ_k for the second-order (left) and fourth-order (right) central advection-diffusion discretizations with $m = 100$, $a = 1$ and $d = 10^{-3}$ (dots), $d = 10^{-2}$ (stars).

The advection-diffusion problem is said to be *advection-dominated* if $d \ll |a|$. Such problems are often studied in numerical analysis as they are

found in many different applications. For such problems we encounter almost the same numerical behaviour as for the pure advection problem. Use of the second-order central scheme will readily lead to numerical oscillations. When using the first-order upwind scheme for the advection term, the diffusion coefficient d is numerically enlarged to $d + \frac{1}{2}|a|h$ (in first approximation). This of course is unrealistic if d is of nearly the same size as $\frac{1}{2}|a|h$ or smaller.

A central scheme will not produce oscillatory solutions if the sought solution is smooth. For arbitrary non-smooth solutions, the only remedy with a central scheme is to take the grid size h small enough if oscillations and the associated negative values are to be avoided. In this regard an important parameter is the so-called *cell Péclet number* ah/d . It is easily seen that the non-diagonal coefficients in the second-order central discretization (3.40) are non-negative iff

$$\frac{|a|h}{d} \leq 2. \quad (3.42)$$

With this restriction the numerical approximations will be free from oscillations and negative values. The non-oscillatory and positivity properties will be discussed in a more general setting in Section 7.

It should be emphasized that for many interesting practical cases condition (3.42) requires extremely small grid sizes, leading to very expensive numerical solutions in terms of CPU time. Fortunately, there exist numerical discretizations that guarantee non-oscillatory approximations without this restriction on the Péclet number. Such schemes will be discussed in Chapter III.

Spatial Scaling

In the above the interval of spatial periodicity was taken with length one, but we could also consider

$$u_t + au_x = du_{xx}, \quad u(x \pm L, t) = u(x, t)$$

with a periodicity interval of arbitrary length L . Then, setting $\bar{x} = x/L$, the equation can be transformed to

$$u_t + \bar{a} u_{\bar{x}} = \bar{d} u_{\bar{x}\bar{x}}, \quad u(\bar{x} \pm 1, t) = u(\bar{x}, t)$$

with scaled coefficients $\bar{a} = a/L$ and $\bar{d} = d/L^2$. Likewise, spatial discretization of the equation on $[0, L]$ with $x_j = jh$, $j = 1, \dots, m$ and $h = L/m$ gives the same result as for the scaled equation with $\bar{x}_j = j\bar{h}$, $\bar{h} = h/L = 1/m$. Hence, in view of (3.41), the eigenvalues of the second-order central discretization are given by

$$\lambda_k = \frac{2d}{h^2} \left(\cos(2\omega_k) - 1 \right) - \frac{ia}{h} \sin(2\omega_k), \quad \omega_k = \frac{\pi kh}{L}, \quad k = 1, \dots, m.$$

Thus the eigenvalues are located on an ellipse with half axes $2d/h^2$ and $|a|/h$. Note that the ellipse itself does not depend on L . If we consider fixed a, d, h and varying L , then it is the number of eigenvalues that is influenced by L : for larger L there are more discrete Fourier modes if the mesh width h is kept constant.

If we let $L \rightarrow \infty$ then we are approaching a pure initial value problem with the whole real line \mathbb{R} as spatial domain and $u(\cdot, t) \in L_2(\mathbb{R})$. For such problems the Fourier series are to be replaced by Fourier integrals in which all frequencies $\omega \in [0, \pi]$ arise, see for instance Strikwerda (1989). Problems on a finite spatial interval with periodicity lead to discrete wave numbers k . The main reason for us to consider mainly this case is the ease of presentation and the fact that it enables directly numerical illustrations.

The eigenvalues of the spatial discretization (3.40) on $[0, L]$ can be included in a wedge

$$\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : \zeta = 0 \text{ or } |\arg(-\zeta)| \leq \alpha\}$$

in the left half-plane, with angle α such that $\tan(\alpha) \approx |a|L/(2\pi d)$; this angle is determined by the location of $\lambda_1 = -4\pi^2 d/L^2 - 2\pi i a/L + \mathcal{O}(h^2)$. For finite L and $h > 0$ sufficiently small we can take $\alpha < \frac{1}{2}\pi$.

Finally we note that, like the spatial variable x , also the time variable t can be scaled of course. This will affect both a and d by a same factor. Moreover, we usually take the initial time equal to 0, but this is just a shift for the time interval $[t_0, t_0 + T]$.

Concluding Remarks and Figures

Finding good advection discretizations is a difficult task. There are many advection schemes around, often of the type discussed above with some modifications. Note that none of the above schemes achieves at the same time good accuracy and positivity. Only the first-order upwind scheme did produce non-negative numerical results in the above tests with initial profile $u(x, 0) = (\sin(\pi x))^{100}$. The oscillations and negative values are relatively small with the third-order upwind-biased scheme. Even better results can be obtained with higher-order upwind-biased schemes, but for these a large stencil is necessary which makes such methods impractical in many real-life simulations with boundary conditions.

As a general rule it can be said that for smooth solutions the central discretizations are to be preferred. However with pure advection or advection-dominated problems steep solution gradients are common. With non-smooth solutions some upwinding is necessary to avoid numerical oscillations. It should be noted that upwind-type schemes are somewhat more difficult to implement and more expensive than the central schemes since the cases $a > 0$ and $a < 0$ are somehow to be considered separately.

The quality of discretizations can be deduced to some extent from the phase velocities and damping factors. These quantities should always be

viewed in combination. The first-order upwind scheme and second-order central scheme have the same phase velocities, but due to the damping of the first-order upwind scheme the numerical results are completely different.

Let λ_k be the eigenvalues of an advection discretization in the Fourier analysis, where we can let the index k run from $-[m/2]$ till $[m/2]$, instead of the customary $k = 1, \dots, m$, to indicate more clearly that

$$\lambda_0 = \lambda_m = 0 \quad \text{and} \quad \lambda_{-j} = \lambda_{m-j} = \overline{\lambda_j},$$

see also Remark 3.1. We consider as in (3.23) the scaled phase velocities and damping factors

$$a_k = -\frac{1}{2\pi ka} \operatorname{Im} \lambda_k, \quad b_k = \frac{1}{ma} \operatorname{Re} \lambda_k, \quad k = 1, \dots, [m/2].$$

Note that the scaling in b_k is taken (somewhat arbitrarily) such that the maximal damping factor for the first-order upwind scheme becomes -2 . The scaling in a_k yields the phase velocity for $a = 1$.

In Figure 3.8 these factors are plotted for the standard optimal-order advection schemes with order $q = 1, 2, 3, 4$ corresponding to (3.13), (3.16), (3.27) and (3.29), respectively. The scaled phase velocity for the exact solution

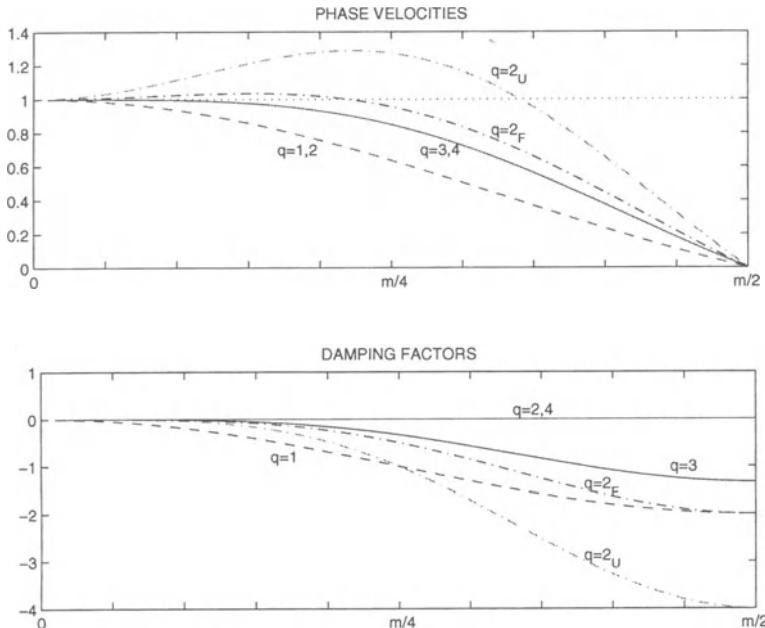


Fig. 3.8. Phase velocities a_k and scaled damping factors b_k as function of the wave number k for the optimal-order advection schemes with $q = 1, 2, 3, 4$, the second-order upwind scheme (indicated by $q = 2_U$) and the Fromm scheme (indicated by $q = 2_F$).

is $a_k = 1$. The plots have been obtained with $m = 100$, but this is not relevant; with any m and $h = 1/m$ we get the same figures.

For future reference we have included also the values for the second-order upwind scheme

$$w'_j(t) = \frac{a}{h} \left(-\frac{1}{2}w_{j-2}(t) + 2w_{j-1}(t) - \frac{3}{2}w_j(t) \right), \quad (3.43)$$

and the second-order upwind-biased scheme

$$w'_j(t) = \frac{a}{h} \left(-\frac{1}{4}w_{j-2}(t) + \frac{5}{4}w_{j-1}(t) - \frac{3}{4}w_j(t) - \frac{1}{4}w_{j+1}(t) \right). \quad (3.44)$$

These formulas are intended for $a > 0$ only; if $a < 0$ one should use their reflection around the central grid point x_j . Scheme (3.44) is obtained by averaging the second-order central scheme and the second-order upwind scheme and is often called the Fromm scheme, after a related fully discrete scheme of Fromm (1968). It was derived to reduce the phase errors of these two second-order schemes. Note however that the Fromm scheme uses the same stencil as third-order upwind-biased and in numerical tests the third-order scheme usually performs slightly better. The second-order upwind scheme (3.43) gives in numerical tests considerable dispersion with oscillations ahead of steep gradients, in contrast to the second-order central scheme where the main oscillations trail the sharp gradients, see Figure 3.3.

For the central diffusion schemes there is no phase velocity, only damping. In Figure 3.9 we have plotted for the second- and fourth-order schemes the relative damping factors

$$c_k = -\frac{1}{4\pi^2 k^2 d} \operatorname{Re} \lambda_k, \quad k = 1, \dots, [m/2].$$

Here the scaling is such that $c_k = 1$ for the exact PDE solution.

Both the second-order and fourth-order diffusion schemes perform well in practice. For situations where a very high accuracy is required one can also consider diffusion schemes with order six or eight, but then the stencil

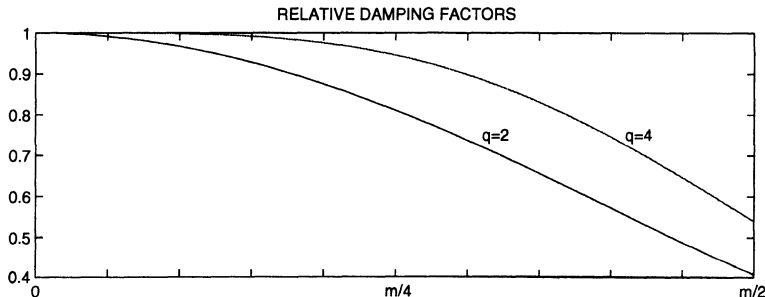


Fig. 3.9. Relative damping factors c_k as function of the wave number k for the optimal-order diffusion schemes with $q = 2$ and $q = 4$.

becomes large and a correct, accurate implementation of boundary conditions becomes quite difficult.

In numerical tests for advection-diffusion problems we will mainly consider the second-order central and third-order upwind-biased advection discretizations and the second-order central diffusion discretization.

4 Convergence of Spatial Discretizations

In the previous section we already came across the concept of stability of a semi-discrete difference scheme $w'(t) = Aw(t)$ with circulant matrix A . In this section stability will be linked with consistency in the general convergence framework

$$\text{stability} \quad \& \quad \text{consistency} \implies \text{convergence}.$$

This type of argument has been used also in Section 2 for ODE methods. For the time being we will discuss it for linear semi-discretizations only, but we drop the restriction to circulant matrices. This is important in the sense that we can consider linear PDEs with variable coefficients and boundary conditions, but standard Fourier analysis is then no longer applicable to demonstrate stability. Considerations on boundary conditions are found in the next section. In this section we will mainly discuss some basic schemes for advection-diffusion problems with variable coefficients, together with the general convergence framework.

4.1 Stability, Consistency and Convergence

Consider a PDE solution $u(x, t)$ for $t \geq 0$ and $x \in \Omega$, with Ω a bounded interval in \mathbb{R} . Without loss of generality we may take $\Omega = (0, 1)$. The solution u may belong to a PDE problem with periodicity conditions or with appropriate boundary conditions given at the end points of Ω , in which case we have an initial-boundary value problem. Discretization on a grid Ω_h consisting of m grid points x_j ($1 \leq j \leq m$) yields a *semi-discrete* system in \mathbb{R}^m

$$w'(t) = F(t, w(t)) \tag{4.1}$$

for $t > 0$, with given initial value $w(0) \in \mathbb{R}^m$. Boundary values are supposed to be included in F . Of importance is that we are not considering a single ODE system, but a *family* of systems parameterized by the grid parameter h .

As before, vectors in \mathbb{R}^m will be identified with grid functions on Ω_h . We want to assess the accuracy of the numerical semi-discrete solution $w(t)$ of (4.1) defined on the discrete grid Ω_h as an approximation to the PDE solution $u(x, t)$ defined on the continuous interval Ω . To enable this assessment, we assume that $u_h(t)$ is a suitable representation of the exact PDE solution

$u(x, t)$ on the grid.²⁴⁾ With the finite difference schemes considered so far, $u_h(t)$ will simply be the restriction of $u(x, t)$ to Ω_h .

The *spatial (discretization) error* is defined by the difference vector

$$\varepsilon(t) = u_h(t) - w(t). \quad (4.2)$$

In order to estimate this global quantity, we consider a suitable norm $\|\cdot\|$ on \mathbb{R}^m and we define the *spatial truncation error*

$$\sigma_h(t) = u'_h(t) - F(t, u_h(t)), \quad (4.3)$$

which is the residual obtained by substituting the exact PDE solution into the difference scheme. Usually, bounds for $\|\sigma_h(t)\|$ are easily found by Taylor expansion, for which it will always be assumed that the PDE solution is sufficiently often differentiable.

We will consider the semi-discretizations on a fixed time interval $t \in [0, T]$. The difference scheme is called *consistent* of order q if we have, for $h \rightarrow 0$,

$$\|\sigma_h(t)\| = \mathcal{O}(h^q) \quad \text{uniformly for } 0 \leq t \leq T. \quad (4.4)$$

The scheme is said to be *convergent* of order p if

$$\|\varepsilon(t)\| = \mathcal{O}(h^p) \quad \text{for } 0 \leq t \leq T. \quad (4.5)$$

For pure initial value problems these orders p and q are usually equal. With boundary conditions we can have $p > q$, as will be discussed in Section 5. As announced above, these concepts of consistency and convergence will be linked through stability.

For stability we will restrict the analysis to linear semi-discrete systems

$$w'(t) = Aw(t) + g(t), \quad (4.6)$$

with an $m \times m$ matrix A and with $g(t) \in \mathbb{R}^m$ representing a source term in the PDE or emanating from boundary conditions (a concrete example will be provided below). The matrix A has constant entries, but need not to be circulant. For Ω_h one may think of a uniform or non-uniform space grid. For the time being we restrict our attention to uniform grids as used before. Then we have the sequence of grids Ω_h with $h = 1/m$, $m \in \mathbb{N}$. If we consider $h \rightarrow 0$ this actually means $m \rightarrow \infty$ so the dimension of the semi-discrete ODE system becomes increasingly large.

The difference scheme and the semi-discrete system (4.6) are called *stable* if we have on all grids Ω_h

$$\|e^{tA}\| \leq Ke^{\omega t} \quad \text{for } 0 \leq t \leq T, \quad (4.7)$$

with constants $K \geq 1$ and $\omega \in \mathbb{R}$ both independent of h .

²⁴⁾ In the finite element literature u_h often denotes a numerical approximation; here it represents the exact PDE solution. Finite element solutions will be denoted by w^h in Chapter III.

Theorem 4.1 Consider the linear semi-discrete system (4.6) and assume the stability condition (4.7) is valid. Suppose further that $\|\sigma_h(t)\| \leq Ch^q$ for $0 \leq t \leq T$ (consistency of order q) and $\|\varepsilon(0)\| \leq C_0 h^q$ with constants $C, C_0 > 0$. Then we have convergence of order $p = q$ with the error bounds

$$\|\varepsilon(t)\| \leq K C_0 e^{\omega t} h^q + \frac{KC}{\omega} (e^{\omega t} - 1) h^q \quad \text{if } \omega \neq 0, \quad 0 \leq t \leq T,$$

and

$$\|\varepsilon(t)\| \leq K C_0 h^q + K C t h^q \quad \text{if } \omega = 0, \quad 0 \leq t \leq T.$$

Proof. By subtraction of $w(t)$ from $u_h(t)$ we find from (4.1) and (4.3) the global error equation

$$\varepsilon'(t) = A\varepsilon(t) + \sigma_h(t). \quad (4.8)$$

By the variation of constants formula (2.20), the solution of this linear system is given by

$$\varepsilon(t) = e^{tA}\varepsilon(0) + \int_0^t e^{(t-s)A} \sigma_h(s) ds.$$

Hence

$$\|\varepsilon(t)\| \leq \|e^{tA}\| \|\varepsilon(0)\| + \int_0^t \|e^{(t-s)A}\| ds \max_{0 \leq s \leq t} \|\sigma_h(s)\|. \quad (4.9)$$

Using the stability assumption (4.7) the error bound now follows. \square

The spatial discretization error $\varepsilon(t)$ and truncation error $\sigma_h(t)$ are related by equation (4.8). The truncation error is determined by local properties of $u(x, t)$ at time point t .²⁵⁾ On the other hand, the spatial discretization error at this point has a time history. It depends on all $\sigma_h(s)$, $0 \leq s \leq t$. For this reason the spatial truncation error $\sigma_h(t)$ is also occasionally called the *local* spatial error and likewise $\varepsilon(t)$ is sometimes called the *global* spatial error.

The constants K and ω in (4.7) can depend on several properties of the PDE and the discretization, and are tacitly assumed to be of ‘moderate’ size. Crucial is that K and ω are independent of h . In other words, we require stability *uniformly* in the mesh width. Note that the arguments used here are similar to the ODE considerations in Section 2.3, but we now consider families of ODEs.

The above result is valid on any given time interval $[0, T]$ with fixed $T > 0$. If $\omega < 0$, the global error bound is even valid for $T = \infty$ and it also indicates that for all times the global error will remain of approximately the same size in the sense that its size will be determined mainly by that of the local error. The error bound for $\omega = 0$ predicts linear error growth in time. In such situations the end point T must be finite for the global error bound to

²⁵⁾ With finite difference spatial discretizations, $\sigma_{h,j}(t)$ will also only depend on nearby values $u(x_{j+k}, t)$, $|k| \leq K_0$, so in this sense $\sigma_h(t)$ is local in space as well.

make sense. If $\omega > 0$, the bound even predicts exponential growth in time. Of course, the end point T must then also be considered finite. It should be stressed that with regard to the temporal behaviour, these bounds will not be sharp if the step from the global error equation (4.8) to the inequality (4.9) is too crude. If, in addition, the stability estimate (4.7) is crude in the sense that a too large value of ω is used, the quality of the prediction further diminishes. For example, if an estimate $\omega > 0$ is used, whereas the matrix A has eigenvalues with non-positive real parts, then the prediction in time will be far from sharp.

In general, verification of the stability condition (4.7) is the difficult step in the analysis of a discretization. To obtain a correct qualitative error bound it is important to find an estimate for ω as small as possible. For problems with constant coefficients and periodicity conditions the matrix A will be circulant and for the L_2 -norm we then can take $\omega = \max_k \operatorname{Re}(\lambda_k)$. In fact this holds for any normal matrix A , see Section 2.3. A possible choice with non-normal matrices is provided by the logarithmic norm, see Theorem 2.4. There the choice of vector norm is important. This norm should always be reasonable in the sense that the error bounds of Theorem 4.1 should be useful in assessing the quality of the discretization. Here we will consider the discrete L_p -norms (2.10) with $p = 1, 2, \infty$.

Below some examples for linear problems with non-constant coefficients will be presented. First we briefly review the convergence results for the advection-diffusion discretizations of the previous section with constant coefficients and periodicity in space.

4.2 Advection-Diffusion with Constant Coefficients

In Section 3 we already used the notion of order of a spatial discretization for advection-diffusion equations with constant coefficients and spatial periodicity. This coincides with the order of consistency defined here by (4.4). For the schemes introduced in Section 3 we did find by Fourier analysis the stability estimate $\|e^{tA}\|_2 \leq 1$ and thus convergence of the discretizations now follows directly from Theorem 4.1. Also convergence towards the modified equations for these discretizations can be obtained this way. In that case we consider the solution \tilde{u} of the modified equation as the exact solution with \tilde{u}_h as restriction to the grid Ω_h , so then we work with a modified spatial truncation error

$$\tilde{\sigma}_h(t) = \tilde{u}'_h(t) - F(t, \tilde{u}_h(t))$$

instead of our original spatial truncation error. The fact that \tilde{u} itself depends on the mesh width h does not matter, the above framework is still valid.

Summarizing, we can formulate the following results in the L_2 -norm for the most important discretizations for advection and diffusion (elaboration of the modified truncation error is left as exercise):

- The first-order upwind advection scheme (3.13), (3.14) is convergent with order 1 for $u_t + au_x = 0$, and it is convergent with order 2 for $\tilde{u}_t + a\tilde{u}_x = \frac{1}{2}h|a|\tilde{u}_{xx}$.
- The second-order central advection scheme (3.16) is convergent with order 2 for $u_t + au_x = 0$, and it is convergent with order 4 for $\tilde{u}_t + a\tilde{u}_x = -\frac{1}{6}h^2a\tilde{u}_{xxx}$.
- The third-order upwind-biased advection scheme (3.27), (3.28) is convergent with order 3 for $u_t + au_x = 0$, and it is convergent with order 4 for $\tilde{u}_t + a\tilde{u}_x = -\frac{1}{12}|a|h^3\tilde{u}_{xxxx}$.
- The second-order central diffusion scheme (3.33) is convergent with order 2 for $u_t = du_{xx}$, and it is convergent with order 6 for $\tilde{u}_t = d\tilde{u}_{xx} + \frac{1}{12}dh^2\tilde{u}_{xxxx} + \frac{1}{360}dh^4\tilde{u}_{xxxxxx}$.

The above results give a good prediction for the accuracy that can be expected on relatively short time intervals. For larger t one should look more closely at the error propagation. For the advection-diffusion model problems with spatial periodicity we have $\|e^{tA}\| = e^{t\omega}$ with $\omega = 0$, since the eigenvalue λ_0 corresponding with the constant Fourier mode $\phi_0 = (1, \dots, 1)^T$ is always equal to zero. Hence Theorem 4.1 predicts a linear growth of the errors in time. Indeed, for advection the error growth will be close to linear initially. Linear growth cannot persist for all $t > 0$, because

$$\|\varepsilon(t)\|_2 = \|u_h(t) - w(t)\|_2 \leq \|u_h(t)\|_2 + \|w(t)\|_2 \lesssim \|u_h(0)\|_2 + \|w(0)\|_2,$$

where the latter inequality only holds approximately because $\|u_h(t)\|_2$ may vary in time somewhat, but it will be bounded. Hence at a certain time the error growth must come to an end because the global error is bounded. Below we will illustrate the error growth with advection numerically.

Numerical Illustration

The main point we want to illustrate here is that the errors for pure advection problems do grow in time, and therefore the time interval should not be too long. We solve $u_t + u_x = 0$ for $t > 0$ and $0 < x < 1$ with a periodic boundary condition and initial function $u(x, 0) = (\sin(\pi x))^{100}$ on the uniform grid Ω_h , using the second-order central and the first-order upwind scheme. Figure 4.1 shows the time evolution of the errors $\|\varepsilon(t)\|_2$ for $h = 1/m$ with $m = 100, 200, 400, 800$. In the plots, the errors were sampled at the time points $t = 0.5, 1, 1.5, \dots, 10$. The horizontal dashed line in the plots represents the value of $\|u_h(0)\|_2$, which is identical up to plotting accuracy on the four grids. This line indicates a 100% relative error.

With the central scheme the linear growth is clearly visible for initial times after which the growth slows down. For $h = 1/800$ the growth remains linear on the whole of the interval $0 \leq t \leq 10$. This behaviour is in complete accordance with the linear error bound in Theorem 4.1.

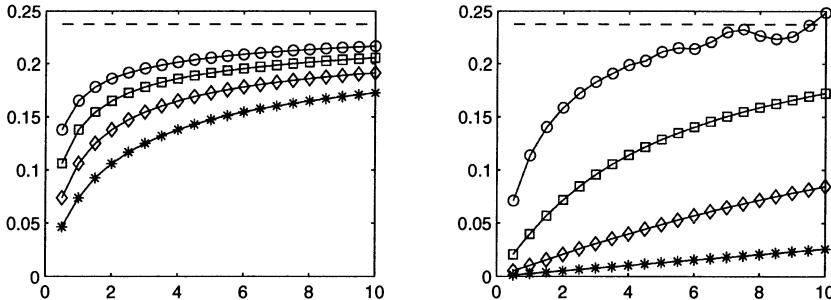


Fig. 4.1. Time evolution of the global error $\|\varepsilon(t)\|_2$, $0 \leq t \leq 10$ for the first-order upwind scheme (left) and the second-order central scheme (right). The markers \circ , \square , \diamond , $*$ correspond with $h = 1/100, 1/200, 1/400, 1/800$, respectively. The dashed horizontal top line represents the value of $\|u_h(0)\|_2$.

For $h = 1/800$ the accuracy of the central scheme could be viewed as acceptable on our time interval; the relative error at $t = 10$ is approximately 10%. The first-order upwind scheme is obviously too inaccurate by any standard, even on the finest grid. It should be emphasized that the grids used in this example are very fine and readily impractical for real (multi-space dimensional) calculations. We use small values here to emphasize that with non-smooth initial functions a fine grid is needed to get a reasonable accuracy with advection over long time intervals. Also it should be noted that the convergence behaviour in space (decrease of factor 4 for central and factor 2 for upwind upon halving h) is not yet well visible. For smoother initial functions the convergence behaviour will be seen of course and also the errors then will become smaller.

Because in general the use of still finer grids is out of the question, we have to resort to methods with a higher order of consistency for getting a better accuracy. Higher-order methods will also show an almost linear growth, but with a smaller slope Ch^p enabling *longer* time intervals.

For pure advection problems on long time intervals the numerical approach by semi-discretization is not advocated. For such problems one can better use a scheme based on characteristics, see Chapter III. Further we note that in applications there are usually some diffusion and source-sink terms, and this leads to a different long-term behaviour.

Fourier Analysis

With a non-zero diffusion coefficient the error evolution will be different from the above pure advection case, even though Theorem 4.1 with $\omega = 0$ still predicts a linear growth. To gain more insight into the evolution we take a brief look once more at discrete Fourier decompositions. Consider the advection-diffusion equation

$$u_t + au_x = du_{xx} + s(x, t)$$

for $t > 0$, $0 < x < 1$, with the periodicity condition $u(x \pm 1, t) = u(x, t)$, constant coefficients a, d and a (space-periodic) source term $s(x, t)$.

For the global spatial discretization error and the local spatial truncation error we consider the discrete Fourier transformations

$$\varepsilon(t) = \sum_k \hat{\varepsilon}_k(t) \phi_k, \quad \sigma_h(t) = \sum_k \hat{\sigma}_{h,k}(t) \phi_k.$$

Here we let the wave number k vary between $-\lfloor m/2 \rfloor$ and $\lfloor m/2 \rfloor$ to distinguish more clearly the high frequency and low frequency modes, see Remark 3.1. Let λ_k be the eigenvalue of the discretization corresponding to the discrete Fourier mode ϕ_k . The transformed error relation $\varepsilon'(t) = A\varepsilon(t) + \sigma_h(t)$ then reads

$$\hat{\varepsilon}'_k(t) = \lambda_k \hat{\varepsilon}_k(t) + \hat{\sigma}_{h,k}(t), \quad t > 0. \quad (4.10)$$

The constant Fourier mode $\phi_0 = (1, 1, \dots, 1)^T$ does not add to the spatial truncation error: we will have $\hat{\sigma}_{h,0}(t) = 0$ for all t . Hence the eigenvalue $\lambda_0 = 0$ plays no role in the error propagation.

As a consequence, instead of $\omega = 0$, we may consider the error propagation bound in Theorem 4.1 with

$$\omega = \max_{k \neq 0} \operatorname{Re} \lambda_k. \quad (4.11)$$

This still may give some over-estimation; the true error evolution is given by formula (4.10). However, using (4.11) already gives some more insight than taking simply $\omega = 0$. We find for example

$$\omega = \begin{cases} -4dh^{-2} \sin^2(\pi h) \approx -4d\pi^2 & \text{for second-order central diffusion,} \\ -2|a|h^{-1} \sin^2(\pi h) \approx -2|a|\pi^2 h & \text{for first-order upwind advection.} \end{cases}$$

With central advection discretizations we still have $\omega = 0$, since all eigenvalues λ_k are purely imaginary.

We see that with upwind discretization for $u_t + au_x = 0$, the error growth will not be truly linear for $h > 0$. It is more important however to note that for the diffusion problem $u_t = u_{xx} + s$ we can use an estimate for the error propagation with $\omega < 0$ uniformly in h , reflecting the damping in the PDE itself, and therefore numerical schemes for the diffusion equation can remain accurate over arbitrarily long time intervals.

4.3 Advection with Variable Coefficients

As pointed out in Section 1.3, there are two forms of advection with variable coefficients: the advective and the conservative form, both of which have a physical relevance. The two forms will be considered separately. In these advection examples we will still deal with a 1D problem and with the periodicity

condition in space. Boundary conditions are considered in the next section and multi-dimensional problems in later chapters. The advective velocity a is supposed to be constant in time to obtain a semi-discrete equation of the form (4.6), but the results easily carry over to time-dependent velocities.

The Conservative Form

Consider a variable-coefficient advection problem in the *conservative form*,

$$u_t + (a(x)u)_x = 0.$$

Assume, as in Section 3, the spatial periodicity condition $u(x \pm 1, t) = u(x, t)$ and let $u(x, 0)$ be an appropriate initial function. The velocity $a(x)$ is allowed to be of either sign, but we take it also 1-periodic and assume it to be differentiable. As in Section 3, we discretize on the uniform grid Ω_h , but now introduce auxiliary points $x_{j \pm 1/2} = \frac{1}{2}(x_{j \pm 1} + x_j)$ midway between the grid points. Further we consider the cells $\Omega_j = [x_{j-1/2}, x_{j+1/2}]$ and the cell averages

$$\bar{u}(x_j, t) = \frac{1}{h} \int_{\Omega_j} u(x, t) dx = u(x_j, t) + \frac{1}{24} h^2 u_{xx}(x_j, t) + \dots.$$

Due to the conservative form, on any cell Ω_j there holds

$$h \frac{d}{dt} \bar{u}(x_j, t) = a(x_{j-\frac{1}{2}}) u(x_{j-\frac{1}{2}}, t) - a(x_{j+\frac{1}{2}}) u(x_{j+\frac{1}{2}}, t). \quad (4.12)$$

This equation tells us that the rate of change of mass over Ω_j is equal to the differences of the in-going and out-going fluxes over the cell boundaries. Instead of conservative form, one therefore also speaks of *flux form*. Over the whole interval $[0, 1]$ the total mass is conserved due to the spatial periodicity conditions,

$$M(t) = \int_0^1 u(x, t) dx = \text{constant}. \quad (4.13)$$

This is called global mass conservation. Relation (4.12) expresses the fact that we have a local mass balance on each cell, and this is sometimes referred to as local mass conservation.

It is of course natural to mimic the conservation property in the spatial discretization. To that end we consider numerical approximations given in the *semi-discrete flux form*

$$w'_j(t) = \frac{1}{h} \left(a(x_{j-\frac{1}{2}}) w_{j-\frac{1}{2}}(t) - a(x_{j+\frac{1}{2}}) w_{j+\frac{1}{2}}(t) \right), \quad j = 1, \dots, m, \quad (4.14)$$

where the $w_{j \pm 1/2}(t)$ are approximate values defined at the cell boundaries. The choice of $w_{j \pm 1/2}(t)$ in terms of neighbouring points $w_i(t)$ determines the actual discretization.

We can interpret the solutions $w_j(t)$ as approximations for the point values $u(x_j, t)$, but it is often more natural to view them as approximations to the cell averages $\bar{u}(x_j, t)$. Note that the difference between $u(x_j, t)$ and $\bar{u}(x_j, t)$ is $\mathcal{O}(h^2)$, and therefore for numerical schemes of order less than three the interpretation matters little. Finite difference schemes like (4.14) whose derivation starts from a conservation law are also called *finite volume* schemes. The finite volume here is the cell Ω_j upon which the conservation law (4.12) is discretized. The periodicity condition will lead to $w_{m+1/2}(t) = w_{1/2}(t)$ and consequently the flux form (4.14) mimics the global conservation rule (4.13) in the sense that the total semi-discrete mass is conserved,

$$h \sum_{j=1}^m w_j(t) = \text{constant}.$$

The First-Order Upwind Scheme in Flux Form

Suppose that $a(x) > 0$ for all x , and consider the choice

$$w_{j+\frac{1}{2}}(t) = w_j(t)$$

defining the upwind difference scheme in flux form,

$$w'_j(t) = \frac{1}{h} \left(a(x_{j-\frac{1}{2}}) w_{j-1}(t) - a(x_{j+\frac{1}{2}}) w_j(t) \right), \quad j = 1, \dots, m, \quad (4.15)$$

with $w_m(t) = w_0(t)$. It is easy to establish first-order consistency. With respect to stability, let

$$\omega = \frac{1}{h} \max_j \left(a(x_{j-\frac{1}{2}}) - a(x_{j+\frac{1}{2}}) \right) = \mathcal{O}(1). \quad (4.16)$$

Note that $\omega = \mathcal{O}(1)$ uniformly in h since $a(x)$ is assumed to be differentiable. We can write (4.15) as a system $w'(t) = Aw(t)$. Using formula (2.30) for the logarithmic norms it is seen directly that $\mu_1(A) = 0$ and $\mu_\infty(A) = \omega$. Hence

$$\|e^{tA}\|_1 \leq 1, \quad \|e^{tA}\|_\infty \leq e^{\omega t} \quad \text{for all } t \geq 0.$$

Using the Hölder inequality for matrices it follows also that

$$\|e^{tA}\|_2 \leq e^{\frac{1}{2}\omega t} \quad \text{for all } t \geq 0.$$

Consequently we have established stability for (4.15). Hence scheme (4.15) is convergent with order one on finite intervals $[0, T]$ in the L_p -norms with $p = 1, 2$ and $p = \infty$.

Note that we obtain different stability estimates for the three norms. This is in agreement with properties of the advection equation in conservative

form. The exact solutions are mass conservative and non-negative²⁶⁾ but there is no maximum principle. For example if we start with $u(x, 0) \equiv 1$ then $u(x, t)$ will increase whenever $a_x(x) < 0$. Mass conservation and $u(x, t) \geq 0$ imply that $\|u(\cdot, t)\|_{L_1[0,1]}$ is conserved, but we can have growth of $\|u(\cdot, t)\|_{L_\infty[0,1]}$.

As we already saw in the previous section, the spatial bias in the scheme (4.15) should be changed if $a(x) < 0$ to maintain the upwind character. Otherwise the scheme makes little physical sense and it becomes unstable. For an arbitrary velocity field, the first-order upwind scheme is correctly defined by the flux form (4.14) with numerical fluxes

$$a(x_{j+\frac{1}{2}}) w_{j+\frac{1}{2}}(t) = a^+(x_{j+\frac{1}{2}}) w_j(t) + a^-(x_{j+\frac{1}{2}}) w_{j+1}(t),$$

where $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$. Verification of the above stability result for this general situation is left as exercise.²⁷⁾

The Second-Order Central Scheme in Flux Form

The choice

$$w_{j+\frac{1}{2}}(t) = \frac{1}{2}(w_j(t) + w_{j+1}(t))$$

determines the central difference scheme in flux form,

$$w'_j(t) = \frac{1}{2h} \left(a(x_{j-\frac{1}{2}})(w_{j-1}(t) + w_j(t)) - a(x_{j+\frac{1}{2}})(w_j(t) + w_{j+1}(t)) \right), \quad (4.17)$$

where $j = 1, \dots, m$ and $w_0(t) = w_m(t), w_{m+1}(t) = w_1(t)$. For a constant velocity a the second-order central scheme (3.16) is regained. We leave it as exercise to establish second-order consistency of (4.17).

Next, writing the difference system (4.17) as $w'(t) = Aw(t)$ it can be shown quite easily that $\langle Av, v \rangle_2 \leq \frac{1}{2}\omega \langle v, v \rangle_2$ for all $v \in \mathbb{R}^m$ with ω given by (4.16).²⁸⁾ Consequently

$$\|e^{tA}\|_2 \leq e^{\frac{1}{2}\omega t} \quad \text{for all } t \geq 0,$$

establishing stability and convergence of order two in the L_2 -norm on finite intervals $[0, T]$ according to Theorem 4.1.

The Advection Form

We next consider the variable-coefficient advection problem in the *advective* form

$$u_t + a(x)u_x = 0,$$

²⁶⁾ Provided the initial profile satisfies $u(x, 0) \geq 0$, of course. The positivity property will be demonstrated in Section 7.

²⁷⁾ Here the *stagnation points*, where $a(x) = 0$, should be examined with care.

²⁸⁾ Exercise: Demonstrate the inequality. For this, write A as a skew-symmetric matrix plus a diagonal matrix.

and assume the same periodicity conditions for a and the solution u as before. This form allows a simple characteristic solution, $u(\xi(t), t)$ is constant along the characteristic $(\xi(t), t)$ in the (x, t) -plane defined by the differential equation $\xi'(t) = a(\xi(t))$. Therefore we have here not only non-negative solutions if $u(x, 0) \geq 0$, but even the maximum principle (1.11) holds. On the other hand, for the variable-coefficient advective form there is no mass conservation. Spatial difference formulas are therefore not in flux form.

For this equation the first-order upwind scheme reads

$$w'_j(t) = \begin{cases} h^{-1}a(x_j)(w_{j-1}(t) - w_j(t)) & \text{if } a(x_j) \geq 0, \\ h^{-1}a(x_j)(w_j(t) - w_{j+1}(t)) & \text{if } a(x_j) \leq 0, \end{cases} \quad (4.18)$$

and the second-order central scheme is given by

$$w'_j(t) = \frac{1}{2h}a(x_j)(w_{j-1}(t) - w_{j+1}(t)), \quad j = 1, \dots, m. \quad (4.19)$$

Again, verification of consistency is a straightforward exercise. Let

$$\omega = \frac{1}{h} \max_j |a(x_j) - a(x_{j-1})| = \mathcal{O}(1).$$

Writing the discretizations in system form, we can now show directly, from the logarithmic norm estimates (2.30), that for the upwind scheme

$$\|e^{tA}\|_1 \leq e^{\omega t}, \quad \|e^{tA}\|_\infty \leq 1 \quad \text{for all } t \geq 0,$$

and consequently, by the Hölder inequality, we also have

$$\|e^{tA}\|_2 \leq e^{\frac{1}{2}\omega t} \quad \text{for all } t \geq 0.$$

With the central scheme the same L_2 -norm stability estimate can be derived.²⁹⁾ Convergence of the schemes in the corresponding norms thus follows from Theorem 4.1.

4.4 Diffusion with Variable Coefficients

Consider the variable-coefficient diffusion problem

$$u_t = (d(x)u_x)_x + s(x, t)$$

for $t > 0$ on the space interval $(0, 1)$. Here $d(x) \geq d_0 > 0$ and $s(x, t)$ represents a source term. Instead of periodic conditions, in this example we assume Dirichlet boundary conditions prescribing the solution for $t > 0$ at $x = 0$ and $x = 1$,

$$u(0, t) = \gamma_0(t), \quad u(1, t) = \gamma_1(t),$$

²⁹⁾ Exercise: Show that for the central scheme $\mu_2(A) \leq \omega$ by writing A as a skew-symmetric matrix plus a lower diagonal remainder.

with γ_0 and γ_1 given functions. Assuming smoothness of the boundary functions and of $d(x)$ and $s(x, t)$, the solution will be smooth for all $t \geq 0$ if the boundary conditions are consistent with the initial function $u(x, 0)$, that is, $u(0, 0) = \gamma_0(0)$ and $u(1, 0) = \gamma_1(0)$.

Since the solution in $x = 0, 1$ is given we consider the grid $\Omega_h = \{x_j\}$ with nodes $x_j = jh$, $j = 1, \dots, m$ where now $h = (m+1)^{-1}$. As before, let $x_{j \pm 1/2} = \frac{1}{2}(x_{j \pm 1} + x_j)$ with $x_0 = 0$, $x_{m+1} = 1$. We discretize this initial-boundary value problem with the conservative central scheme

$$w'_j(t) = \frac{1}{h^2} \left(d(x_{j-\frac{1}{2}})(w_{j-1}(t) - w_j(t)) - d(x_{j+\frac{1}{2}})(w_j(t) - w_{j+1}(t)) \right) + s(x_j, t),$$

where $j = 1, \dots, m$ and

$$w_0(t) = \gamma_0(t), \quad w_{m+1}(t) = \gamma_1(t).$$

This difference scheme fits in the ODE form $w'(t) = Aw(t) + g(t)$ in \mathbb{R}^m , with symmetric matrix A and forcing term $g(t)$ given by

$$A = \frac{1}{h^2} \begin{pmatrix} a_1 & b_1 & & & 0 \\ b_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b_{m-1} \\ 0 & & & b_{m-1} & a_m \end{pmatrix}, \quad g = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{m-1} \\ s_m \end{pmatrix} + \frac{1}{h^2} \begin{pmatrix} b_0 \gamma_0 \\ 0 \\ \vdots \\ 0 \\ b_m \gamma_1 \end{pmatrix},$$

where $a_j = -(d(x_{j-1/2}) + d(x_{j+1/2}))$, $b_j = d(x_{j+1/2})$ and $s_j(t) = s(x_j, t)$. Observe that not only the source function s occurs in g , but also the boundary functions γ_0 and γ_1 divided by h^2 . This discretization is standard for the current parabolic initial-boundary value problem.

Assuming smoothness of $d(x)$ and $s(x, t)$ it is easy to prove second-order consistency. From formula (2.30) it follows directly that $\mu_1(A) \leq 0$ and $\mu_\infty(A) \leq 0$. Thus we have here in the L_p -norms, $p = 1, \infty$,

$$\|e^{tA}\|_p \leq 1 \quad \text{for all } t \geq 0,$$

and using the Hölder inequality for matrices this now also follows for the L_2 -norm. According to Theorem 4.1, we thus can conclude convergence in these norms of order two on any finite interval $[0, T]$.

For the parabolic problem this L_2 -stability result, being based on estimates in the L_1 - and L_∞ -norm, is not optimal in the sense that it does not reflect the dissipative nature of the parabolic problem. To recover this dissipative nature, a more refined estimate is needed as discussed in Section 4.2 for constant coefficients and periodicity. In fact, if we assume that the diffusion coefficient $d(x)$ is constant, then the eigenvalues of A are known to be³⁰⁾

$$\lambda_k = -\frac{4d}{h^2} \sin^2\left(\frac{1}{2}\pi kh\right), \quad k = 1, \dots, m,$$

³⁰⁾ The corresponding eigenvectors are $v_k = \sqrt{2} (\sin(\pi kh), \sin(2\pi kh), \dots, \sin(m\pi kh))^T$.

which are all negative (compare this with expression (3.36) for the periodic case). Since A is symmetric, it follows that $\mu_2(A) \approx -d\pi^2$, revealing a rather strong dissipation.

If the positive diffusion coefficient $d(x)$ is non-constant, it can still be shown that $\mu_2(A) < 0$ uniformly in h .³¹⁾ Hence for the parabolic problem the spatial global error bound from Theorem 4.1 is valid for $t \rightarrow \infty$. Consequently, for evolving time the spatial error behaviour will be notably better than for the advection problems in the sense that the global spatial error will be determined mainly by the size of the local spatial truncation error and a structural growth will not take place.

4.5 Variable Coefficients and Higher-Order Schemes

With the simple discretizations of order one and two discussed above it was relatively easy to obtain consistency and stability results. With higher-order advection-diffusion schemes pertinent stability results are absent in general and consistency has to be examined more carefully.

As an important example we consider here the third-order upwind-biased advection scheme for $u_t + (a(x)u)_x = 0$, which will be used in later experiments. Written in flux form (4.14), the scheme is defined by the cell-boundary values

$$w_{j+\frac{1}{2}}(t) = \begin{cases} \frac{1}{6} (-w_{j-1}(t) + 5w_j(t) + 2w_{j+1}(t)) & \text{if } a(x_{j+\frac{1}{2}}) \geq 0, \\ \frac{1}{6} (2w_j(t) + 5w_{j+1}(t) - w_{j+2}(t)) & \text{if } a(x_{j+\frac{1}{2}}) < 0, \end{cases} \quad (4.20)$$

with numerical fluxes $a(x_{j\pm 1/2}) w_{j\pm 1/2}(t)$. If we substitute exact point values $u(x_j, t)$ for $w_j(t)$ into (4.14), (4.20), it follows by Taylor expansion that the order of consistency is only two. However, the scheme is still third order with respect to cell-average values $\bar{u}(x_j, t)$; this cell-average interpretation fits more naturally with the flux form. Elaboration of the Taylor series expansions is left as a (rather tedious) exercise. Consistency should always be regarded with respect to a specific interpretation of the exact values.

Simple stability estimates are not available with this scheme. We will however encounter many experiments where this spatial discretization is used, and from the results it will be clear that the scheme is stable indeed. The same considerations hold for the advective form and for fourth-order advection and diffusion discretizations with variable coefficients.

The absence of pertinent stability results with variable coefficients for higher-order schemes is in general not a matter of great concern as long as the coefficients are smooth. Theoretical results are very valuable to ascertain

³¹⁾ Exercise: Show that $h\langle v, Av \rangle_2 = -\sum_{j=0}^m b_j(v_j - v_{j+1})^2 \leq -d_0 \sum_{j=0}^m (v_j - v_{j+1})^2$ for any $v \in \mathbb{R}^m$ with $d_0 = \min_x d(x)$ and $v_0 = v_{m+1} = 0$, and compare this with the constant coefficient case.

stability in model situations, such as linear problems with constant coefficients and no boundary conditions, especially as a tool to distinguish ‘good’ basic schemes from ‘bad’ ones. In more difficult, practical situations stability considerations are often more heuristic, but always under the assumption of stability for model situations.

For example, with smooth variable coefficients, it can be argued that instabilities arise at first only locally. However, locally the coefficients can be considered as nearly constant if the mesh width is small enough. This indicates that when the scheme is stable for constant coefficients, it will also be so for non-constant coefficients. This heuristic argument by saying that locally the coefficients are (almost) constant is called an argument by *frozen coefficients*. A mathematical justification of the frozen coefficients argument is outside the scope of this text. For some classical results in this direction we refer to Richtmyer & Morton (1967, Sect. 5.4, 5.5).

5 Boundary Conditions and Spatial Accuracy

The discussion in Section 4.4 for the diffusion equation $u_t = (d(x)u_x)_x + s(x, t)$ subjected to Dirichlet boundary conditions, might give the impression that replacing periodicity for other types of boundary conditions is numerically straightforward. The reverse is true: *boundary conditions nearly always lead to complications* of one sort of another, even in the relatively simple case of one spatial dimension. For example, consider the simple constant-coefficient advection-diffusion equation

$$u_t + au_x = du_{xx}, \quad t > 0, \quad 0 < x < L,$$

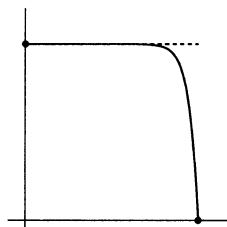
with given initial profile $u(x, 0)$. If $d > 0$ we need boundary conditions at $x = 0$ and $x = L$, such as Dirichlet conditions. On the other hand, for the pure advection problem we need only to prescribe the solution at the *inflow* boundary, that is, at $x = 0$ if $a > 0$ and at $x = L$ if $a < 0$. If $d > 0$ but $d \approx 0$, or more precisely if the *Péclet number* $|aL/d|$ is large, then the Dirichlet condition at the outflow boundary will give rise to a *boundary layer*, as the following example illustrates.

Example 5.1 Let $u(0, t) = 1$, $L = 1$ and $a > 0$. Then the advection equation $u_t + au_x = 0$ gives the stationary solution

$$u(x, t) = 1.$$

On the other hand, the advection-diffusion equation $u_t + au_x = du_{xx}$ with the Dirichlet conditions $u(0, t) = 1$ and $u(1, t) = 0$ has the stationary solution

$$u(x, t) = (e^{ax/d} - e^{ax/d}) / (e^{ax/d} - 1). \quad \diamond$$



If the Péclet number $|aL/d|$ is large, the problem is called *singularly perturbed*. Boundary layer behaviour is characteristic for singularly perturbed problems. Resolving boundary layers on a uniform grid may require a very large number of grid points. A remedy is found in solution adapted grids or special fitted schemes that are suited to deal with boundary layers, see for instance Morton (1996) and Roos, Stynes & Tobiska (1996). A boundary layer as shown in the above picture will be absent if at the outflow boundary the Neumann condition $u_x = 0$ is imposed. Then rapid changes may still occur in the spatial derivatives of u , but u itself will not show the nearly discontinuous behaviour that arises with a Dirichlet condition at the outflow boundary.

Singularly perturbed problems and solution adapted grids are out of the scope of this section; they will be considered in some detail in Chapter III. Here we return to problems where boundary layers are absent and the solution is (relatively) smooth.

5.1 Refined Global Error Estimates

Even for smooth solutions without boundary layers, the presence of boundary conditions can complicate the numerical treatment, simply due to the fact that at a boundary point a different (one-sided) spatial discretization has to be used which may lead to a lower order of consistency. This is expected to have an adverse effect on the global accuracy. However, this effect is often not as large as expected in the sense that the (global) order of convergence p can be greater than the (local) order of consistency q . Even if the discretization is locally inconsistent, that is $q = 0$, we can still have $p > 0$, that is, convergence. To reveal such phenomena an error analysis is required which is more refined than the one behind Theorem 4.1.

We will use here the same notation as in the previous section. Hence $\varepsilon(t) = u_h(t) - w(t)$ stands for the global spatial discretization error, $\sigma_h(t)$ is the local spatial truncation error and for the linear semi-discrete system $w'(t) = Aw(t) + g(t)$ we will use the stability assumption $\|e^{tA}\| \leq Ke^{\omega t}$ with $K > 0$ and $\omega \in \mathbb{R}$ independent of h .

Theorem 5.2 Consider the linear semi-discrete system (4.6) and assume (4.7) is valid (stability). Suppose that for $0 \leq t \leq T$ we can decompose the truncation error $\sigma_h(t)$ as

$$\sigma_h(t) = A\xi(t) + \eta(t) \quad \text{with} \quad \|\xi(t)\|, \|\xi'(t)\|, \|\eta(t)\| \leq Ch^r, \quad (5.1)$$

and suppose that $\|\varepsilon(0)\| \leq C_0 h^r$, where $C, C_0 > 0$ are constants. Then we have convergence of order $p = r$ with the error bounds

$$\|\varepsilon(t)\| \leq KC_0 e^{\omega t} h^r + \left(1 + Ke^{\omega t} + \frac{2K}{\omega}(e^{\omega t} - 1)\right) Ch^r \quad \text{if } \omega \neq 0, 0 \leq t \leq T,$$

and

$$\|\varepsilon(t)\| \leq KC_0 h^r + (1 + K + 2Kt) Ch^r \quad \text{if } \omega = 0, 0 \leq t \leq T.$$

Proof. The global error $\varepsilon(t)$ satisfies

$$\varepsilon'(t) = A\varepsilon(t) + \sigma_h(t) = A(\varepsilon(t) + \xi(t)) + \eta(t).$$

Hence $\tilde{\varepsilon}(t) = \varepsilon(t) + \xi(t)$ satisfies

$$\tilde{\varepsilon}'(t) = A\tilde{\varepsilon}(t) + \xi'(t) + \eta(t), \quad \tilde{\varepsilon}(0) = \varepsilon(0) + \xi(0).$$

We can now proceed in the same way as in Theorem 4.1 to obtain the error bound

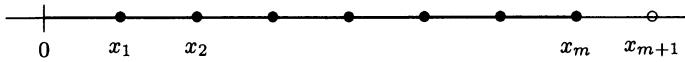
$$\|\varepsilon(t)\| \leq \|\xi(t)\| + Ke^{\omega t} \|\varepsilon(0) + \xi(0)\| + \frac{K}{\omega} (e^{\omega t} - 1) \max_{0 \leq s \leq t} \|\xi'(s) + \eta(s)\|,$$

with the usual convention that $\omega^{-1}(e^{\omega t} - 1) = t$ in case $\omega = 0$. The stated result now follows immediately from this error bound. \square

The idea behind the above local error decomposition is that due to *negative* powers of the mesh size h contained in A , it can happen that r is greater than the order of consistency q . For example, if $\|A\| \sim h^{-k}$ then from (5.1) it can only be concluded that $\|\sigma_h\| = \mathcal{O}(h^{r-k})$. The remainder of this section is devoted to a number of examples to illustrate this phenomenon in connection with boundary conditions. It should be emphasized that the low orders of consistency in these examples are not caused by lack of smoothness of the solution. It will always be assumed that the solution is sufficiently often differentiable.

5.2 Outflow with Central Advection Discretization

Consider the advection equation $u_t + u_x = 0$ for $t > 0$ on the spatial interval $x \in (0, 1)$ with given inflow condition $u(0, t) = \gamma_0(t)$ and given initial profile $u(x, 0)$. Second-order central spatial discretization on the grid $\Omega_h = \{x_j\}$



with nodes $x_j = jh$, $j = 1, \dots, m$ and mesh width $h = 1/m$ gives the semi-discrete system

$$w'_j(t) = \frac{1}{2h} (w_{j-1}(t) - w_{j+1}(t)), \quad j = 1, 2, \dots, m,$$

where $w_0(t) = \gamma_0(t)$ and $w_{m+1}(t)$ represents an approximation at the *virtual point* $x_{m+1} = 1 + h$. This value can be found by extrapolation, for example,

$$w_{m+1}(t) = \theta w_m(t) + (1 - \theta)w_{m-1}(t).$$

We consider $\theta = 1$ (constant extrapolation) and $\theta = 2$ (linear extrapolation). This last choice seems more natural because we then apply in fact first-order upwind discretization at the outflow boundary.

For the spatial truncation error $\sigma_h(t) = (\sigma_{h,1}(t), \dots, \sigma_{h,m}(t))^T$ we find that $\sigma_{h,j}(t) = \mathcal{O}(h^2)$ for $j < m$, whereas at the outflow point $x_m = 1$ we have

$$\begin{aligned}\sigma_{h,m}(t) &= \frac{d}{dt}u(x_m, t) - \frac{1}{2h}(\theta u(x_{m-1}, t) - \theta u(x_m, t)) \\ &= -\frac{1}{2}(2-\theta)u_x(x_m, t) - \frac{1}{4}\theta h u_{xx}(x_m, t) + \mathcal{O}(h^2).\end{aligned}$$

So, for the spatial truncation error we have the bounds

$$\|\sigma_h(t)\|_\infty = \mathcal{O}(h^s), \quad \|\sigma_h(t)\|_2 = \mathcal{O}(h^{s+\frac{1}{2}}), \quad \|\sigma_h(t)\|_1 = \mathcal{O}(h^{s+1}),$$

with $s = 0$ if $\theta = 1$ and $s = 1$ if $\theta = 2$. Assuming stability, Theorem 4.1 thus predicts different convergence orders, the highest in the L_1 -norm. Numerical experiments, however, indicate that $\|\varepsilon(t)\| = \mathcal{O}(h^{s+1})$ not only holds for the L_1 -norm but also for the other two norms.

This observation is in accordance with the local error decomposition (5.1). We have

$$A = \frac{1}{2h} \begin{pmatrix} 0 & -1 & & \\ 1 & 0 & -1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 0 & -1 \\ & & & \theta & -\theta \end{pmatrix}, \quad \sigma_h(t) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} ch^s + \mathcal{O}(h^{s+1}),$$

with $c = -\frac{1}{2}u_x(1, t)$ if $\theta = 1$ and $c = -\frac{1}{2}u_{xx}(1, t)$ if $\theta = 2$. Ignoring for the moment the $\mathcal{O}(h^{s+1})$ term in σ_h , then putting $A\xi = \sigma_h$ gives

$$\xi_{j-1} - \xi_{j+1} = 0 \quad (j = 1, \dots, m-1) \quad \text{with} \quad \xi_0 = 0,$$

$$\theta \xi_{m-1} - \theta \xi_m = 2ch^{s+1}.$$

Hence we have $\xi = (\xi_1, 0, \xi_1, 0, \dots)^T$ with $\xi_1 = (-1)^m 2\theta^{-1} ch^{s+1}$. This implies $\|\xi\| = \mathcal{O}(h^{s+1})$ in any of the above L_p -norms. In this way we have established (5.1) with $r = s + 1$, where η contains the higher-order terms in σ_h . Consequently, from Theorem 5.2 we then may conclude a global convergence order $p = s + 1$ in all three norms, provided we have stability.

In the present example stability is easy to prove in the L_2 -norm. Consider on \mathbb{R}^m the inner product

$$\langle v, w \rangle = h \left(\sum_{j=1}^{m-1} v_j w_j + \frac{1}{\theta} v_m w_m \right),$$

and corresponding norm $\|v\| = \langle v, v \rangle^{1/2}$. We have for any $v \in \mathbb{R}^m$,

$$\langle v, Av \rangle = -\frac{1}{2}v_m^2 \leq 0.$$

Hence, for any solution of $w'(t) = Aw(t)$,

$$\frac{d}{dt} \|w(t)\|^2 = 2\langle w(t), w'(t) \rangle = 2\langle w(t), Aw(t) \rangle \leq 0,$$

showing that $\|w(t)\|$ is non-increasing and thus $\|e^{tA}\| \leq 1$ for $t \geq 0$.³²⁾ If $\theta = 1$, the norm $\|\cdot\|$ is the standard L_2 -norm so that the stability estimate (4.7) holds in the L_2 -norm with $K = 1$ and $\omega = 0$. For $\theta = 2$ the norm $\|\cdot\|$ is equivalent to the L_2 -norm,

$$\|v\|_2^2 \geq \|v\|^2 = \|v\|_2^2 - \frac{1}{2}hv_m^2 \geq \frac{1}{2}\|v\|_2^2.$$

So in this case the stability estimate holds in the L_2 -norm with $K = \sqrt{2}$ and $\omega = 0$. This concludes the proof of convergence in the L_2 -norm with convergence order $p = s + 1$.³³⁾ In particular, by using the first-order upwind scheme at the outflow point, we maintain the second-order convergence of the central discretization even though the scheme is consistent with order $\frac{3}{2}$ only.

Stability results in the L_1 - and L_∞ -norm are lacking. Since $\|v\|_1 \leq \|v\|_2$ for any v , we can conclude second-order convergence in the L_1 -norm. In experiments the same order of convergence was observed with the L_∞ -norm.

5.3 Boundary Conditions with the Heat Equation

The Neumann Boundary Condition for Diffusion

Consider the diffusion model problem $u_t = u_{xx}$ for $t > 0$ on the spatial interval $x \in (0, 1)$ with at the left boundary point the Dirichlet condition $u(0, t) = \gamma_0(t)$ and at the right boundary point the homogeneous Neumann condition $u_x(1, t) = 0$. Again we use the uniform grid Ω_h with nodes $x_j = jh$, $j = 1, \dots, m$ and $h = 1/m$. The standard second-order central differencing now gives

$$w'_j(t) = \frac{1}{h^2} (w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)), \quad j = 1, 2, \dots, m,$$

where $w_0(t) = \gamma_0(t)$ and the value $w_{m+1}(t)$ at the virtual point $x_{m+1} = 1 + h$ is to be determined by the Neumann condition. Let us consider the difference formulas

$$\frac{1}{h} (w_{m+1}(t) - w_m(t)) = 0 \quad \text{and} \quad \frac{1}{2h} (w_{m+1}(t) - w_{m-1}(t)) = 0,$$

³²⁾ This is just a repetition of the argument $\mu(A) \leq 0 \Rightarrow \|e^{tA}\| \leq 1$ with inner products.

³³⁾ The L_2 -convergence result is basically due to Gustafsson (1975). The results in that paper are much more general (hyperbolic systems with multistep time integration). For further general results, see also Gustafsson, Kreiss & Oliger (1995, Sect. 11.4, 12.7).

which approximate the condition $u_x(1, t) = 0$ with order one and two, respectively. Thus we set, with parameter $\theta = 0$ or 1,

$$w_{m+1}(t) = \theta w_m(t) + (1 - \theta)w_{m-1}(t).$$

For a smooth solution it can be assumed that both the differential equation and the Neumann condition are valid at $x_m = 1$. This implies that $u_{xxx}(1, t) = u_{tx}(1, t) = 0$. Inserting the exact solution in the difference scheme, we find local truncation errors $\sigma_{h,j}(t) = \mathcal{O}(h^2)$ for $1 \leq j < m$, while at the right boundary point x_m we have

$$\sigma_{h,m}(t) = \frac{1}{2}\theta u_{xx}(1, t) + \mathcal{O}(h^2).$$

Hence if $\theta = 0$ we have the usual $\mathcal{O}(h^2)$ local error everywhere on the grid. On the other hand, if $\theta = 1$ we have an inconsistency at $x_m = 1$.

Yet, by means of the decomposition formula (5.1) for the local error, first-order convergence can be proven if $\theta = 1$. Ignoring the $\mathcal{O}(h^2)$ term in σ_h and putting $A\xi = \sigma_h$ gives

$$\xi_{j-1} - 2\xi_j + \xi_{j+1} = 0 \quad (j = 1, \dots, m-1) \quad \text{with} \quad \xi_0 = 0,$$

$$\xi_{m-1} - \xi_m = (2 - \theta)^{-1}ch^2,$$

where $c = \frac{1}{2}\theta u_{xx}(1, t)$. The components ξ_j are easily found to be

$$\xi_j = -j(2 - \theta)^{-1}ch^2, \quad j = 1, \dots, m.$$

It follows that $\|\xi\| = \mathcal{O}(h)$, giving $r = 1$ in (5.1) in the L_1 , L_2 and L_∞ -norms.

There remains to prove stability. In the present example we have the bound $\|e^{tA}\|_\infty \leq 1$ due to diagonal dominance in the rows, see (2.30). Hence we can immediately conclude first-order convergence in the maximum norm, and therefore also in the L_1 - and L_2 -norm, on any interval $[0, T]$. For actual applications, first-order convergence is rather poor of course and one better selects the second-order implementation.

Remark 5.3 The choice $w_{m+1} = w_{m-1}$ corresponding to $\theta = 0$ in this example presents itself in a natural way if we consider, instead of $u(x, t)$ for $0 \leq x \leq 1$, the function $\tilde{u}(x, t)$ for $0 \leq x \leq 2$, defined by

$$\tilde{u}(x, t) = \begin{cases} u(x, t) & \text{for } 0 \leq x \leq 1, \\ u(1-x, t) & \text{for } 1 \leq x \leq 2. \end{cases}$$

For \tilde{u} we then have $\tilde{u}_t = \tilde{u}_{xx}$, $0 < x < 2$ with $\tilde{u}(0, t) = \tilde{u}(2, t) = \gamma_0(t)$, and the homogeneous Neumann condition at $x = 1$ is automatically fulfilled due to symmetry around the point $x = 1$. So this Neumann condition can be seen here as a mirror condition. Discretizing this extended problem with central differences will give the same symmetry in the semi-discrete system, and hence $\tilde{w}_{m+j}(t) = \tilde{w}_{m-j}(t)$. \diamond

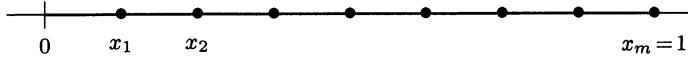
Boundaries with Cell/Vertex Centered Grids

In the previous examples the grid was chosen such that the boundaries did coincide with grid points. If fluxes are prescribed on the boundaries it would be more natural to let the boundaries coincide with cell vertices. As an example we consider

$$u_t = u_{xx}, \quad u(0, t) = \gamma_0(t), \quad u_x(1, t) = \gamma_1(t)$$

for $t > 0$, $0 < x < 1$ with given initial profile $u(x, 0)$, and we will present numerical results for the second-order central discretizations on three different uniform grids.

First, consider the *vertex centered grid* with nodes $x_j = jh$ and $h = 1/m$,



which coincides with the standard finite difference grid. Using $x_{m+1} = 1 + h$ as virtual point with value w_{m+1} determined by $(2h)^{-1}(w_{m+1} - w_{m-1}) = \gamma_1$, we obtain the semi-discrete system

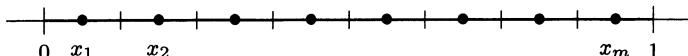
$$\left. \begin{aligned} w'_1(t) &= \frac{1}{h^2} (-2w_1(t) + w_2(t)) + \frac{1}{h^2} \gamma_0(t), \\ w'_j(t) &= \frac{1}{h^2} (w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)), \quad 2 \leq j \leq m-1, \\ w'_m(t) &= \frac{1}{h^2} (2w_{m-1}(t) - 2w_m(t)) + \frac{2}{h} \gamma_1(t). \end{aligned} \right\} \quad (5.2)$$

The truncation error $\sigma_{h,j}(t)$ is $\mathcal{O}(h^2)$ at all nodes x_j except at the right boundary where we find

$$\sigma_{h,m}(t) = \frac{1}{3}hu_{xxx}(1, t) + \mathcal{O}(h^2).$$

Note that in the previous example with a homogeneous Neumann condition it could be concluded that $u_{xxx}(1, t) = 0$, but if $\gamma'_1(t) \neq 0$ we will have $\sigma_{h,m}(t) = \mathcal{O}(h)$ only.

Next, consider the *cell centered grid* with nodes $x_j = (j - \frac{1}{2})h$, $h = 1/m$.



Now the right boundary condition fits in a natural way if we interpret the difference formula as a flux form, but implementing the Dirichlet condition at the left becomes more cumbersome. Using the virtual point $x_0 = -\frac{1}{2}h$

with virtual value w_0 such that $\frac{1}{2}(w_0 + w_1) = \gamma_0$, we obtain the semi-discrete system

$$\left. \begin{aligned} w'_1(t) &= \frac{1}{h^2} \left(-3w_1(t) + w_2(t) \right) + \frac{2}{h^2} \gamma_0(t), \\ w'_j(t) &= \frac{1}{h^2} \left(w_{j-1}(t) - 2w_j(t) + w_{j+1}(t) \right), \quad 2 \leq j \leq m-1, \\ w'_m(t) &= \frac{1}{h^2} \left(w_{m-1}(t) - w_m(t) \right) + \frac{1}{h} \gamma_1(t). \end{aligned} \right\} \quad (5.3)$$

In this case we even have an inconsistency at the left,

$$\sigma_{h,1}(t) = \frac{1}{4} u_{xx}(0, t) + \mathcal{O}(h^2),$$

at the right boundary we get

$$\sigma_{h,m}(t) = \frac{1}{24} h u_{xxx}(1, t) + \mathcal{O}(h^2),$$

and elsewhere the truncation errors are $\mathcal{O}(h^2)$.

We can also combine the grids by taking $h = 1/(m + \frac{1}{2})$ and $x_j = jh$, so that now on both sides the boundary conditions fit naturally. On the left we get the vertex centered discretization and on the right the cell centered. We will refer to this grid as *hybrid*.

In Table 5.1 the errors on the three grids are given in the max-norm and L_2 -norm for the solution $u(x, t) = 1 + e^{-\frac{1}{4}\pi^2 t} \cos(\frac{1}{2}\pi x)$ at output time $T = \frac{1}{4}$. Initial and boundary conditions are derived from this prescribed solution. It is obvious from the results that there is no reduction in accuracy. In all cases the errors decrease with a factor ≈ 4 upon halving the mesh width h . Even with the cell centered grid, where the truncation error is only $\mathcal{O}(h^0)$ at the left boundary, we find second-order convergence in the max-norm. This can be explained just as in the previous examples; elaboration of this result is left as exercise.³⁴⁾ Further it should be noted that although the rates of convergence are the same in the three cases, the error constants are smallest on the hybrid grid.

The terminology vertex/cell centered stems from finite volume schemes that will again be encountered in Chapter III. If the interval is divided into cells, then with a cell centered scheme approximations in the middle of the cells are found, whereas with a vertex centered scheme the nodes x_j primarily define the grid and cell boundaries or vertices are placed half-way between the nodes.

Recall from Section 3 that $w_j(t)$ can be viewed either as an approximation to a point value or as cell average. In the latter case the difference formula $u_{xx}(x_j) \approx h^{-2}(w_{j-1} - 2w_j + w_{j+1})$ is more naturally written in the flux form

³⁴⁾ Exercise: Prove the second-order convergence of the difference schemes on the various grids.

m	vertex centered		cell centered		hybrid	
	L_2 -error	L_∞ -error	L_2 -error	L_∞ -error	L_2 -error	L_∞ -error
10	.11 10^{-2}	.21 10^{-2}	.12 10^{-2}	.17 10^{-2}	.11 10^{-3}	.19 10^{-3}
20	.26 10^{-3}	.52 10^{-3}	.32 10^{-3}	.42 10^{-3}	.29 10^{-4}	.56 10^{-4}
40	.63 10^{-4}	.13 10^{-3}	.79 10^{-4}	.10 10^{-3}	.76 10^{-5}	.15 10^{-4}
80	.16 10^{-4}	.33 10^{-4}	.20 10^{-4}	.26 10^{-4}	.19 10^{-5}	.39 10^{-5}

Table 5.1. Space discretization errors with vertex/cell centered grids.

$u_{xx}(x_j) \approx h^{-1}(f_{j-1/2} - f_{j+1/2})$ with $f_{j+1/2} = h^{-1}(w_j - w_{j+1})$ approximating the diffusive fluxes $-u_x(x_{j+1/2})$. Since the difference between point values and cell averages is $\mathcal{O}(h^2)$, the interpretation makes no difference here in terms of order of convergence. However, the Dirichlet condition fits more easily in the point value interpretation if a grid point is located at the boundary. Likewise, with the Neumann condition the flux is given at the boundary, so this fits well within a cell average interpretation if the boundary of the domain coincides with a cell boundary. Finally it should be mentioned that the results in Table 5.1 are for point values. The results with cell averages would be very much the same.

5.4 Boundary Conditions and Higher-Order Schemes

The above procedure to implement boundary conditions by considering virtual points outside the spatial region Ω can also be applied to higher-order schemes. More general procedures are possible; one could consider at all points near the boundaries special discretizations, but then the number of possibilities is overwhelming. Therefore we will regard here only simple extrapolations.

As an important example we consider the third-order upwind-biased advection scheme for $u_t + au_x = 0$ on $\Omega = (0, 1)$ with $a > 0$ constant and with inflow condition $u(0, t) = \gamma_0(t)$. We consider as before the uniform (vertex centered) grid with nodes $x_j = jh$, $j = 1, \dots, m$ with $h = 1/m$, and we set $w_0 = \gamma_0$. Then we can apply the discretization (3.27) at the interior points x_2, \dots, x_{m-1} . At the point x_1 adjacent to the left boundary and at the right boundary point x_m we use the same discretization formula, but with virtual values found by linear ($q = 1$) and quadratic extrapolation ($q = 2$),

$$q = 1 : \quad w_{-1} = 2w_0 - w_1, \quad w_{m+1} = 2w_m - w_{m-1},$$

$$q = 2 : \quad w_{-1} = 3w_0 - 3w_1 + w_2, \quad w_{m+1} = 3w_m - 3w_{m-1} + w_{m-2}.$$

In the resulting difference formulas q equals the local order of consistency near the boundaries. It should be noted that with $q = 2$ we are actually

applying the second-order central discretization (3.16) at the inflow and the second-order upwind discretization (3.43) at the outflow boundary. For stability considerations we set $\gamma_0 = 0$ and we consider $a = 1$. The resulting semi-discrete system is written as $w'(t) = Aw(t)$.

Stability estimates for these choices are absent. The following numerical test result indicates stability in the L_2 - and L_∞ -norm. In Figure 5.1 the time evolution of $\|e^{tA}\|$ is given for $0 \leq t \leq 1.5$ with $h = \frac{1}{100}$. This result for one value of h does not mean that much, but it has been verified numerically that with other values of h we get very similar pictures with the same upper bounds for $\|e^{tA}\|$, and this clearly points to stability.

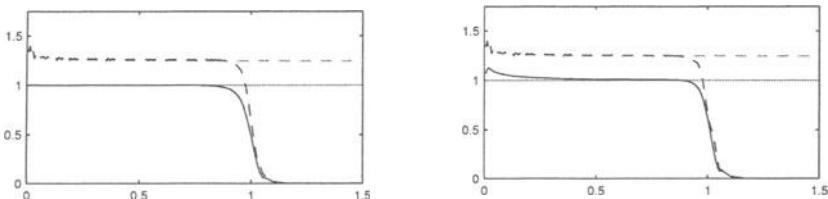


Fig. 5.1. Evolution of $\|e^{tA}\|$ for $t \in [0, 1.5]$, $h = 1/100$ with linear extrapolation (left) and quadratic extrapolation (right). Results in L_2 -norm (solid) and L_∞ -norm (dashed). The corresponding results for the periodic case are indicated by grey lines.

As a reference the corresponding values for $\|e^{tA}\|$ with periodicity in space are also plotted, with grey lines, where we know by Fourier analysis that the scheme is stable in the L_2 -norm. We see that up to $t = 1$, approximately, the results for the periodic case and the case with boundary conditions nearly coincide. After $t = 1$ the value of $\|e^{tA}\|$ drops quickly for the case with boundary conditions. This is as it should be since with $a = 1$ and $\gamma_0 = 0$ the exact solution becomes identically equal to zero for $t > 1$.

With respect to the L_2 -norm it should be mentioned that with linear extrapolation at the boundaries we do *not* have $\|e^{tA}\|_2 \leq 1$. Although it is hardly visible in Figure 5.1, calculations for $t > 0$ close to 0 reveal that $\|e^{tA}\|_2$ is then slightly larger than 1. Further it should be noted that in the L_∞ -norm we can no longer employ the logarithmic norm to prove stability due to lack of diagonal dominance. Numerical experiments on accuracy have revealed that the order of convergence of the schemes is $q + 1$ both in the L_2 - and L_∞ -norm. So with quadratic extrapolation we retrieve the third-order convergence of the interior discretization, although we then have local consistency of order two near the boundaries. In view of Theorem 5.2 this does not come as a surprise (although verification of the bound for $\|\xi\|$ is not easy anymore).

Relying only on some experiments in model situations with boundary conditions seems more precarious than with the variable coefficients considered in Section 4.5. It is known, for non-pathological cases, that boundary

conditions can ruin the stability of the overall scheme. This, however, seems to occur only with non-dissipative schemes. For theory and experiments with such schemes we refer to Trefethen (1983, 1984). The above third-order advection discretization is dissipative. A well-known theoretical treatise on this subject is due to Gustafsson, Kreiss & Sundström (1972), and this theory is usually called GKS-theory. It is not an easy theory, but a good account is given in Strikwerda (1989) (where it is called the GKS-theory to honour previous contributions of Osher) and in Gustafsson, Kreiss & Oliger (1995).

As a practical rule it can be said that the implementation of boundary conditions is not so critical for dissipative schemes. Heuristically it can be argued that any onset of instability at the boundaries will be damped as soon as it reaches the interior domain. Still this should always remain a matter of concern in practical computations, and some numerical testing is advisable.

6 Time Stepping for PDEs

6.1 The Method of Lines and Direct Discretizations

The Method of Lines (MOL)

So far we have only discussed the spatial discretization of advection-diffusion problems, leading to semi-discrete systems of ODEs. In this section we will direct our attention to some basic time stepping methods. First the *method of lines* (MOL) approach is discussed, where the semi-discrete system is integrated with an appropriate ODE method.³⁵⁾

As in Section 4 we suppose that a PDE problem with solution $u(x, t)$, possibly subjected to boundary conditions, has been discretized in space on a certain grid Ω_h with mesh width $h > 0$ to yield a semi-discrete system

$$w'(t) = F(t, w(t)), \quad 0 < t \leq T, \quad w(0) \text{ given}, \quad (6.1)$$

with $w(t) = (w_j(t))_{j=1}^m \in \mathbb{R}^m$, m being proportional to the number of grid points in space. Discretized boundary conditions are supposed to be contained in F . According to the MOL approach, *fully discrete* approximations $w_j^n \approx u(x_j, t_n)$ are now obtained by applying some suitable ODE method with step size τ for the time levels $t_n = n\tau$, $n = 1, 2, \dots$. As a standard example for the ODE method we consider the θ -method (2.31),

$$w_{n+1} = w_n + \tau(1 - \theta)F(t_n, w_n) + \tau\theta F(t_{n+1}, w_{n+1}). \quad (6.2)$$

³⁵⁾ It is emphasized that the method of lines is not a ‘method’ in the numerical sense, it is a way or approach to construct and analyze methods, and ‘lines’ is a metaphor for lines $(x_i, t), t \geq 0$ in the (x, t) -domain, x_i fixed, along which the approximations to the PDE solutions are studied.

Here and in the following, $w_n = (w_j^n)_{j=1}^m \in \mathbb{R}^m$ denotes the vector containing the fully discrete numerical solution at time level $t = t_n$ (also viewed as grid function). More sophisticated methods are discussed in the next chapter.

A typical MOL reasoning goes as follows. Let, as before, $u_h(t)$ denote the restriction of $u(x, t)$ to Ω_h . If we know that the spatial discretization is convergent of order p_1 , i.e., $\|u_h(t) - w(t)\| \leq C_1 h^{p_1}$, and the ODE theory tells us that the integration method is convergent of order p_2 , i.e., $\|w(t_n) - w_n\| \leq C_2 \tau^{p_2}$, then we immediately obtain the error bound

$$\|u_h(t_n) - w_n\| \leq \|u_h(t_n) - w(t_n)\| + \|w(t_n) - w_n\| \leq C_1 h^{p_1} + C_2 \tau^{p_2}. \quad (6.3)$$

However, there does exist an important difference with ODE theory where (6.1) is considered as a single, fixed system. Here, in the setting of PDEs, the system (6.1) stands for a semi-discrete problem and hence it represents a *family* of ODE systems parameterized by the mesh width h . For a proper interpretation of the error bound, saying that we have temporal convergence order p_2 *uniformly* in h , we must assure that C_2 and p_2 are independent of the mesh width h . If for example $C_2 \sim h^{-1}$ for $h \rightarrow 0$, the error bound will not reflect temporal convergence with order p_2 when τ and h tend to zero simultaneously. Consistency and stability of the ODE method should thus be verified for all $h > 0$.

The stability requirement may impose a *restriction* on the temporal step size τ in terms of the mesh width h . With explicit time stepping methods this usually leads to a step size restriction of the form $\tau \leq Ch$ for hyperbolic problems and $\tau \leq Ch^2$ for parabolic problems, where the constant C is determined by the actual problem and discretization.

Direct Space-Time Discretizations

The MOL approach, where space and time discretizations are considered separately, is conceptually simple and flexible. The popularity of this approach is mainly due to the fact that it is easy to combine various discretizations for advection and diffusion with reaction terms. Another attractive, practical point is that there exist nowadays many well developed ODE methods (to be discussed in Chapter II) and for these methods sophisticated software is freely available.

However, if we apply a standard ODE method to a semi-discrete system (6.1), information about the underlying PDE problem might be neglected. This holds in particular for advection problems where knowledge about the characteristics can be used to obtain combined space-time discretizations which can be more efficient in special cases, see for instance the Lax-Wendroff scheme in Example 6.3 below. Unfortunately, with such a scheme the flexibility of the MOL approach of easily combining advection, diffusion and reaction discretizations is strongly diminished.

Even if a scheme can be viewed as a MOL scheme, it can be advantageous to consider space and time errors simultaneously. In this section we

will therefore also look at time stepping with the so-called *direct* approach where the separation between space and time discretization is not a priori assumed.

The difference schemes for linear equations considered in this section can all be formulated as a two-level finite difference scheme

$$B_0 w_{n+1} = B_1 w_n + G(t_n, t_{n+1}), \quad n = 0, 1, \dots, \quad (6.4)$$

obtained from combining discretizations in space and time for an initial-boundary value problem for a linear PDE with constant coefficients. The matrices $B_0, B_1 \in \mathbb{R}^{m \times m}$ and the inhomogeneous term $G(t_n, t_{n+1}) \in \mathbb{R}^m$ depend on the mesh width h and the temporal step size τ . If the scheme is explicit, B_0 is the identity operator. Otherwise we have to ‘invert’ B_0 to find w_{n+1} and for this we assume that B_0 is nonsingular.

Example 6.1 Two-level MOL schemes also belong to class (6.4). For example, applying the θ -method (6.2) to a linear semi-discrete system $w'(t) = Aw(t) + g(t)$ yields

$$w_{n+1} = w_n + (1 - \theta)\tau(Aw_n + g(t_n)) + \theta\tau(Aw_{n+1} + g(t_{n+1})),$$

which fits in the form (6.4) with

$$\begin{aligned} B_0 &= I - \theta\tau A, & B_1 &= I + (1 - \theta)\tau A, \\ G(t_n, t_{n+1}) &= (1 - \theta)\tau g(t_n) + \theta\tau g(t_{n+1}). \end{aligned}$$

Stability issues for these schemes are considered in the next section.

We note that, whereas in the PDE literature a scheme that can be formulated in terms of the time levels t_n and t_{n+1} is called a two-level scheme, in the ODE literature such a scheme is usually called a *one-step* scheme (for multistep ODE schemes additional past levels t_{n-k} are used). All Runge-Kutta methods, to be considered in Chapter II, fall in the one-step category. Such methods use in addition auxiliary time levels intermediate between t_n and t_{n+1} , but still the schemes basically fit in the formulation (6.4). \diamond

Example 6.2 : Courant-Isaacson-Rees scheme. Conceptual differences between the direct space-time discretization and MOL approach are best illustrated by simple examples. Let us discretize the advection model equation $u_t + au_x = 0$, $a > 0$, by means of first-order upwind in space and first-order forward differencing in time. Assuming periodicity in space, this gives the fully discrete scheme

$$w_j^{n+1} = w_j^n + \frac{a\tau}{h} (w_{j-1}^n - w_j^n), \quad j = 1, \dots, m, \quad (6.5)$$

for $n = 0, 1, \dots$ with $h = 1/m$ and $w_0^n = w_m^n$. This scheme, with its upwind counterpart for $a < 0$, is also known as the *Courant-Isaacson-Rees* scheme

(Courant, Isaacson & Rees, 1952). Writing it in the form (6.4) gives a zero inhomogeneous term and operators

$$B_0 = I, \quad B_1 = I + \tau A,$$

where A is the upwind matrix belonging to the semi-discrete scheme (3.13). Trivially, this scheme also belongs to the class of MOL schemes with the forward Euler method as time integrator. So one can analyze it as a MOL scheme, and under the appropriate time step restriction for stability we then obtain a bound (6.3) with $p_1 = p_2 = 1$. However, a direct analysis where space and time discretizations are considered simultaneously leads to a more refined error bound.

Within the direct approach one compares the numerical approximations w_j^n directly with the exact solution values $u(x_j, t_n)$. Substitution of the exact values into (6.5) gives, for $j = 1, 2, \dots, m$,

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \frac{a\tau}{h} (u(x_{j-1}, t_n) - u(x_j, t_n)) + \tau \rho_j^n,$$

with a *local truncation error* $\rho_n = (\rho_j^n)$ whose components are given by

$$\begin{aligned} \rho_j^n &= \left[(u_t + \frac{1}{2}\tau u_{tt} + \dots) + a(u_x - \frac{1}{2}hu_{xx} + \dots) \right] (x_j, t_n) \\ &= -\frac{1}{2}ah \left(1 - \frac{a\tau}{h} \right) u_{xx}(x_j, t_n) + \mathcal{O}(h^2) + \mathcal{O}(\tau^2). \end{aligned}$$

Using the vector notation, we have

$$u_h(t_{n+1}) - w_{n+1} = B_1(u_h(t_n) - w_n) + \tau \rho_n.$$

From (2.13) it is directly seen that $\|B_1\| \leq 1$ in the L_1, L_2 and L_∞ -norm as long as

$$\frac{a\tau}{h} \leq 1, \tag{6.6}$$

and then it follows by the usual stability-consistency argument – see Section 6.2 – that the global space-time error satisfies

$$\|u_h(t_n) - w_n\| \leq \frac{1}{2} t_n ah \left(1 - \frac{a\tau}{h} \right) \max_{x,t} \|u_{xx}(x, t)\| + \mathcal{O}(h^2).$$

This result reveals the conceptual difference with the MOL analysis. If we let $\tau \rightarrow 0$ with h fixed, we just regain the bound for the spatial error. We see, however, that the error for (6.5) will actually *decrease* for $\tau > 0$ and it will be *less* than the error of the semi-discrete system with exact time integration. Apparently, the error of the explicit Euler time stepping counteracts the error of first-order upwind space discretization, something which is not found with the error bound (6.3) derived in the MOL fashion. ◇

Example 6.3 : Lax-Wendroff scheme. In the above example the fully discrete scheme still could be viewed within the MOL framework, only a more refined analysis was needed to obtain the true error behaviour. There are also schemes which cannot be regarded as an ODE method applied to a certain space discretization. A well-known example is the *Lax-Wendroff* scheme for advection equations (Lax & Wendroff, 1960). For $u_t + au_x = 0$ it reads

$$w_j^{n+1} = w_j^n + \frac{a\tau}{2h} (w_{j-1}^n - w_{j+1}^n) + \frac{1}{2} \left(\frac{a\tau}{h} \right)^2 (w_{j-1}^n - 2w_j^n + w_{j+1}^n). \quad (6.7)$$

This scheme uses second-order central differencing in space combined with the Taylor series expansion

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \tau u_t(x_j, t_n) + \frac{1}{2} \tau^2 u_{tt}(x_j, t_n) + \mathcal{O}(\tau^3).$$

Replacing the temporal derivatives u_t and u_{tt} by the corresponding spatial derivatives $-au_x$ and $a^2 u_{xx}$, insertion of the central difference formulas yields (6.7). This construction does not fit in the MOL framework. If we specify boundary conditions, it obviously does fit in the two-level formulation (6.4).

The local truncation error in space and time, defined as in the previous example, is given by

$$\rho_j^n = \frac{1}{6} ah^2 \left(1 - \left(\frac{a\tau}{h} \right)^2 \right) u_{xxx}(x_j, t_n) + \mathcal{O}(h^3) + \mathcal{O}(\tau^3).$$

We thus see that similar as for the Courant-Isaacson-Rees scheme, the truncation error gets smaller when τ approaches $h/|a|$. As we will see in Section 6.4, the time step restriction for stability is $\tau|a| \leq h$.

Note that for $h > 0$ fixed and $\tau \rightarrow 0$ we get the same truncation error as for the semi-discrete system

$$w'_j(t) = \frac{a}{2h} (w_{j-1}(t) - w_{j+1}(t))$$

obtained with second-order central spatial differences. This is not surprising: if we divide (6.7) by τ , the Lax-Wendroff scheme can be written as

$$\frac{1}{\tau} (w_j^{n+1} - w_j^n) = \frac{a}{2h} (w_{j-1}^n - w_{j+1}^n) + \frac{\tau a^2}{2h^2} (w_{j-1}^n - 2w_j^n + w_{j+1}^n).$$

For $\tau \rightarrow 0$ the left-hand side will approach the time derivative $w'_j(t_n)$ and the diffusive term on the right-hand side vanishes. Consequently, for $\tau \rightarrow 0$ the Lax-Wendroff solution tends to the solution of the semi-discrete system (3.16) for the advection equation with second-order central differences.³⁶⁾ \diamond

³⁶⁾ Writing the semi-discrete system as $w'(t) = Aw(t)$, application of a standard explicit one-step ODE method from Chapter II will yield a scheme of the form $w_{n+1} = P(\tau A)w_n$ with polynomial $P(z) = \sum_{j=0}^s c_j z^j$. This class of schemes does not contain the Lax-Wendroff scheme.

Additional schemes that fall outside the MOL framework will be discussed in Chapter III for advection problems. There also exist some special schemes for the heat equation or more general parabolic problems that fall outside the MOL framework; a well-known example is the Du Fort-Frankel scheme, see Section IV.3.1. However, for the actual application these special parabolic schemes are less interesting.

The Courant-Isaacson-Rees and Lax-Wendroff schemes for the advection equation can be interpreted in terms of the *characteristics*. From the characteristic solution we know that $u(x_j, t_{n+1}) = u(x_j - \tau a, t_n)$. Let $\nu = \tau a/h$ and suppose $-1 \leq \nu \leq 1$. Then $x_j - \tau a = x_j - \nu h$ lies in between the grid points x_{j-1} and x_{j+1} . The Courant-Isaacson-Rees scheme is obtained if we apply linear interpolation on the grid at time level t_n to find an approximation for $u(x_j, t_{n+1}) = u(x_j - \nu h, t_n)$,

$$u(x_j - \nu h) \approx (1 - \nu)u(x_j) + \nu u(x_{j-1}), \quad a > 0,$$

$$u(x_j - \nu h) \approx (1 + \nu)u(x_j) - \nu u(x_{j+1}), \quad a < 0,$$

where we have omitted the argument $t = t_n$ for convenience of notation. In the same way the Lax-Wendroff scheme is obtained by quadratic interpolation from the grid points x_{j-1}, x_j, x_{j+1} ,

$$u(x_j - \nu h) \approx \frac{1}{2}\nu(\nu + 1)u(x_{j-1}) + (1 - \nu^2)u(x_j) + \frac{1}{2}\nu(\nu - 1)u(x_{j+1}).$$

For this formula the sign of a plays no role. It is obvious that if $\nu = \pm 1$ there is no interpolation error, which explains the technical fact that the truncation errors vanish. This combination of characteristics and interpolation can be directly generalized to the equation $u_t + a(x)u_x = 0$, also in more dimensions; a further discussion is found in Chapter III.

Advection schemes which are derived through interpolation of characteristic solutions on a fixed grid are often called *semi-Lagrangian* schemes. Schemes which are not based on a fixed grid and only use the characteristic solution are called *Lagrangian* schemes. Semi-Lagrangian schemes can benefit from their characteristic bias as the above local error expressions nicely show. However, addition of diffusion and reaction terms cannot be done in a straightforward way as with MOL schemes.

6.2 Stability, Consistency and Convergence

Without mentioning it explicitly, for the Courant-Isaacson-Rees scheme we already used the fundamental notions of stability, consistency and convergence. In this subsection we will outline these concepts for the general two-level scheme (6.4).

The (*space-time*) truncation error ρ_n for (6.4) at time $t = t_n$ is defined by the relation

$$B_0 u_h(t_{n+1}) = B_1 u_h(t_n) + G(t_n, t_{n+1}) + \tau \rho_n. \quad (6.8)$$

The error term $\tau\rho_n$ is the residue of the scheme obtained with the true PDE solution, and in this sense it measures how well the difference scheme approximates the PDE problem. Subtracting (6.4) from (6.8) shows that the *global (space-time discretization) errors* $\varepsilon_n = u_h(t_n) - w_n$ satisfy

$$B_0\varepsilon_{n+1} = B_1\varepsilon_n + \tau\rho_n.$$

This relation is rewritten to

$$\varepsilon_{n+1} = B\varepsilon_n + \delta_n, \quad (6.9)$$

with an amplification operator

$$B = B_0^{-1}B_1 \quad (6.10)$$

and *local (space-time discretization) error*

$$\delta_n = \tau B_0^{-1}\rho_n. \quad (6.11)$$

To avoid a possible confusion, we emphasize the subtle difference between truncation error and local error. Whereas the truncation error is a residue, the local error is the error committed in one step starting on the exact solution ($\varepsilon_n = 0$). For explicit schemes they only differ by a factor τ .

Elaborating (6.9) gives the familiar global error expression for one-step or two-level schemes

$$\varepsilon_n = B^n\varepsilon_0 + B^{n-1}\delta_0 + \cdots + B\delta_{n-2} + \delta_{n-1},$$

which adds all previously committed local errors, weighted with powers of B , to the global error at the current time level. This expression is identical to its counterpart found for one-step ODE methods; see for example the discussion in Section 2.6 for the θ -method. Consequently, similar techniques and reasoning can be used to establish convergence through the general law

$$\text{stability} \quad \& \quad \text{consistency} \implies \text{convergence}. \quad (6.12)$$

Considering the time interval $[0, T]$, a convergence analysis amounts to assess the global errors ε_n for all $n \geq 0$, $\tau \rightarrow 0$ such that $t_n = n\tau \leq T$ and for $h \rightarrow 0$. Often a certain dependence of τ on h is assumed a priori, say $\tau = \tau_h$. The two-level scheme (6.4) is *stable* if a positive constant K exists of moderate size, independent of τ , h , such that

$$\|B^n\| \leq K \quad \text{for } n \geq 0, n\tau \leq T. \quad (6.13)$$

Stability will in general depend on the relation between h and $\tau = \tau_h$. For the Courant-Isaacson-Rees scheme a linear relation satisfying (6.6) should be assumed.

Consistency can be defined through the truncation error ρ_n . A natural assumption is

$$\|B_0^{-1}\| \leq C,$$

with C a constant independent of τ and h . The existence and boundedness of the inverse means that the scheme has a numerically well-defined, unique solution. With this assumption we have

$$\|\delta_n\| \leq C\tau\|\rho_n\|. \quad (6.14)$$

The two-level scheme (6.4) is *consistent* if $\|\rho_n\| \rightarrow 0$ for $\tau, h \rightarrow 0$, possibly with different orders in time and space. Putting all this together gives the global error estimate

$$\|\varepsilon_n\| \leq K\|\varepsilon_0\| + K \sum_{k=0}^{n-1} \|\delta_k\| \leq K\|\varepsilon_0\| + KCt_n \max_{0 \leq k \leq n-1} \|\rho_k\|, \quad (6.15)$$

proving convergence for $t_n \in [0, T]$ from stability and consistency. In general, consistency is easily found by Taylor series expansions. In the remainder of this section we will discuss some possibilities to establish stability.

Remark 6.4 The material of this section is standard and has been discussed before in textbooks and review papers, most notably in the classic book of Richtmyer & Morton (1967). Some other good references of a more recent date are the introductory book of Mitchell & Griffiths (1980), the review paper of Thomée (1990) and the books of Strikwerda (1989) and Gustafsson, Kreiss & Oliger (1995). ◇

The Lax Equivalence Theorem

We have shown that if the two-level scheme (6.4) is consistent, then stability is sufficient for convergence. The famous Lax equivalence theorem – also known as the Lax-Richtmyer theorem – says that for a consistent linear finite difference scheme, stability is also necessary. The proof of this theorem with its precise assumptions and technicalities, in the setting of Banach or Hilbert spaces, is beyond this introductory chapter. Good references are Richtmyer & Morton (1967) and Strikwerda (1989). In this theory the underlying PDE problem is assumed to be well-posed and initial functions of arbitrary non-smoothness are considered.³⁷⁾

³⁷⁾ Occasionally one also encounters slightly weaker stability concepts where $\|B^n\|$ is allowed to grow (at most) polynomially in n or h^{-1} , and this is still sufficient for convergence if certain smoothness assumptions on the solution are satisfied; see for example van Dorsselaer et al. (1993) and the references therein. Such weak (in-)stabilities are sometimes found with the maximum norm in the presence of boundary conditions.

The Courant-Friedrichs-Lowy (CFL) Condition

Explicit schemes for the advection equation $u_t + au_x = 0$ give rise to step size restrictions (stability conditions) of the form

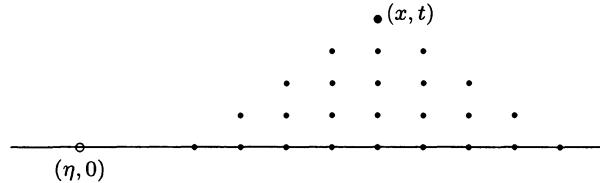
$$\frac{\tau|a|}{h} \leq C, \quad (6.16)$$

with C an appropriate positive constant independent of τ and h . In the numerical literature this stability condition (6.16) is called a CFL condition, after Courant, Friedrichs & Lewy (1928), and the number $\nu = \tau a/h$ is usually called the Courant number (or CFL number). In this classic paper, written long before computers were available, finite difference approximations were used to prove existence of PDE solutions. The authors showed that for convergence of finite difference approximations to certain classes of hyperbolic PDEs, conditions like (6.16) are necessary in order to guarantee that the mathematical *domain of dependence* of a PDE problem lies within the numerical counterpart of the finite difference method. For our simple advection equation this important observation is easily explained.

An explicit two-level scheme for $u_t + au_x = 0$ will be of the form

$$w_j^{n+1} = \sum_{k=-r}^s \gamma_k w_{j+k}^n$$

with coefficients γ_k depending on $\nu = a\tau/h$. Consider for a fixed point (x, t) approximations $w_j^n \approx u(x, t)$ with $x_j = x$ and $t_n = t$. This w_j^n depends on the initial data w_i^0 from the grid points $x_i = ih$ with $i = j - nr, \dots, j + ns$.



If we consider $\tau, h \rightarrow 0$ with constant ratio ν , then $x_{j-nr} \rightarrow x - (r/\nu)at$ and $x_{j+ns} \rightarrow x + (s/\nu)at$, and thus we see that the numerical approximations to $u(x, t)$ are determined by the initial data in the interval

$$\mathcal{D} = \left[x - \frac{r}{\nu}at, x + \frac{s}{\nu}at \right].$$

This is the numerical domain of dependence. The exact solution on the other hand is determined by the value $u(x - at, 0)$, and hence the point $\{x - at\}$ is the mathematical domain of dependence. If $x - at = \eta$ is not contained in the interval \mathcal{D} , then the numerical result cannot be correct because it has no knowledge about the initial data at $(x - at, 0)$. A local change in the initial value $u(x - at, 0)$ will not affect the numerical approximations to $u(x, t)$.

When we combine this necessary condition for convergence with the Lax equivalence theorem, the Courant-Friedrichs-Lowy condition for an explicit, consistent two-level scheme can be formulated as follows:

A necessary condition for stability is that the mathematical domain of dependence of the PDE is contained in the numerical domain of dependence.

The mathematical domain of dependence for the advection equation consists of a single point. For more general hyperbolic systems it can be a finite interval. For parabolic problems such as the heat equation $u_t = du_{xx}$ it will be the whole real line. It is now obvious that a condition like $d\tau/h \leq C$ cannot be sufficient for stability of explicit schemes for parabolic problems and that we need stricter conditions like $d\tau/h^2 \leq C$.

6.3 Stability for MOL – Stability Regions

We will illustrate the MOL stability considerations for linear problems (6.1),

$$F(t, v) = Av + g(t),$$

where A stands for a discretized advection-diffusion operator. Since stability concerns the difference of two solutions, we can disregard the inhomogeneous term. Application of the θ -method, or any other one-step scheme, then gives for the homogeneous problem the recurrence

$$w_{n+1} = R(\tau A)w_n, \quad (6.17)$$

where R is the *stability function*. Any possible step size restriction for τ in terms of h is to be recovered from the stability function and the properties of A . For the θ -method, the important roles of the *stability region* defined by R and the *eigenvalues* of A has been discussed in Section 2. All what has been said there applies to the current situation.

Example 6.5 As standard example we will consider the θ -method (6.2) with $\theta \in [0, 1]$. For $\theta = 0$ we get the explicit Euler method. In view of numerical experiments at the end of this chapter we also consider the *explicit trapezoidal rule*

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_n + \tau F(t_n, w_n)). \quad (6.18)$$

This method is obtained by inserting the explicit Euler prediction $\bar{w}_{n+1} = w_n + \tau F(t_n, w_n)$ into the implicit trapezoidal rule, i.e., (6.2) with $\theta = \frac{1}{2}$. Method (6.18) has order two like its implicit counterpart. The stability function of this method is

$$R(z) = 1 + z + \frac{1}{2}z^2.$$

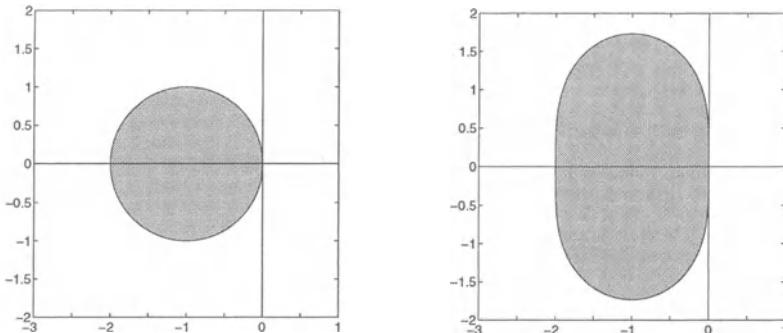


Fig. 6.1. Stability regions for the explicit Euler method (left) and the explicit trapezoidal rule (right).

The stability region is plotted in Figure 6.1. For comparison the stability region of the explicit Euler method is also added. Shortly we will see that the stability region of the explicit trapezoidal rule has an important advantage over the explicit Euler region when we consider third-order advection discretizations. ◇

Normal Matrices

Let us first suppose that A is a normal matrix. Recall that circulant matrices, which arise when we have spatial periodicity and constant coefficients in the PDE, are normal. Let λ_k be the k th eigenvalue of A and \mathcal{S} the stability region of R . Owing to the normality, in the L_2 -norm we then have

$$\|R(\tau A)^n\|_2 = \max_{1 \leq k \leq m} |R(\tau \lambda_k)^n|,$$

and thus we will have stability if we satisfy the *eigenvalue criterion*

$$\tau \lambda_k \in \mathcal{S} \quad \text{for all } k. \quad (6.19)$$

This criterion is actually a bit more strict than necessary as it implies

$$\|R(\tau A)^n\|_2 \leq K \quad \text{for } n \geq 0, n\tau \leq T$$

with $K = 1$. Sufficient for stability is to have a positive constant K of moderate size and independent of n and h (power boundedness, uniformly in h). If we consider a fixed time interval $[0, T]$, this holds with the eigenvalue criterion

$$|R(\tau \lambda_k)| \leq 1 + K' \tau$$

for $K' > 0$. Loosely speaking, $\tau \lambda_k$ is then allowed to lie a distance $\mathcal{O}(\tau)$ outside \mathcal{S} . In the ODE literature one usually encounters (6.19) whereas in the PDE literature also the slightly more relaxed condition is used (von Neumann's criterion, to be discussed later).

Step Size Restrictions for Advection-Diffusion

Table 6.1 lists the step size restrictions based on (6.19) for the θ -method with the standard first- and second-order discretizations for advection and diffusion discussed earlier, see the eigenvalue expressions (3.21), (3.22), (3.36) and the Figures 3.2 and 3.5. Remember that for $\theta \geq \frac{1}{2}$ we have A-stability and hence there are no step size restrictions (*unconditional stability*). For $\theta < \frac{1}{2}$ the stability region is the closed disc³⁸⁾

$$\mathcal{S} = \{z \in \mathbb{C} : |z + \alpha| \leq \alpha\} \quad \text{with} \quad \alpha = 1/(1 - 2\theta). \quad (6.20)$$

With this bounded disc we clearly do find step size restrictions (*conditional stability*), whereas for central advection we even have *instability* for all $\tau > 0$ due to the fact that the disc has no intersection with the imaginary axis and all eigenvalues of the central advection discretization are purely imaginary.

	$\theta < \frac{1}{2}$	$\theta \geq \frac{1}{2}$
Upw. Advection (3.13),(3.14)	$ a \tau/h \leq 1/(1 - 2\theta)$	$\tau \leq \infty$
Central Advection (3.16)	<i>instability</i>	$\tau \leq \infty$
Central Diffusion (3.33)	$d\tau/h^2 \leq 1/(2 - 4\theta)$	$\tau \leq \infty$

Table 6.1. Step size restrictions for stability of the θ -method obtained with the eigenvalue criterion (6.19).

When we compare the stability regions in Figure 6.1 with the eigenvalue plots in the Figures 3.2, 3.5, it is obvious that we obtain for the explicit trapezoidal rule (6.18) the same step size restrictions as for the explicit Euler method ($\theta = 0$) in Table 6.1. However, in more general cases the explicit trapezoidal rule does have a clear advantage over the explicit Euler method.

Considering the third-order advection discretization, with eigenvalues given by (3.30), we see from Figure 3.4 that many of these eigenvalues stay close to the imaginary axis. These eigenvalues do not fit into the stability region of the explicit Euler method if the Courant number $\nu = a\tau/h$ is fixed: the stability condition reads $|1 + z| \leq 1$ with

$$z = \tau\lambda = -\frac{4}{3}|\nu|\sin^4(\omega) - \frac{1}{3}i\nu\sin(2\omega)(4 - \cos(2\omega)),$$

where $\omega \in [0, \pi]$.³⁹⁾ For small values of ω it is seen that this condition cannot hold, and consequently the explicit Euler method gives here an unstable

³⁸⁾ Exercise: Prove that for $\theta < \frac{1}{2}$ the stability region of the θ -method is given by (6.20).

³⁹⁾ To be more precise: this condition should hold for all values $\omega = \pi kh$ where $k = 1, \dots, m$, $h = 1/m$, see formula (3.30). For small $h > 0$ and ν fixed we may consider all $\omega \in [0, \pi]$.

scheme. On the other hand, with the third-order advection discretization we get for the explicit trapezoidal rule a positive step size restriction, given by

$$\frac{\tau|a|}{h} \leq 0.87, \quad (6.21)$$

approximately. Here the value 0.87 was found experimentally.⁴⁰⁾ Derivation of a precise bound is quite complicated and technical. Since this third-order advection discretization is often used to avoid the excessive damping of the first-order upwind scheme and the oscillations of the second-order central scheme, the fact that the explicit trapezoidal rule does provide a stable integration scheme (under the step size restriction) is important for practice. At the end of this chapter numerical examples will be presented where this scheme is used.

As we saw, both the explicit Euler method and the explicit trapezoidal rule are unstable for second-order central advection discretization. This changes if some diffusion is added, and also then the explicit trapezoidal rule has an advantage. For the equation $u_t + au_x = du_{xx}$ with second-order central discretization the condition $|R(z)| \leq 1$, $z = \tau\lambda$, must be verified for

$$z = \tau\lambda = 2\mu(\cos(2\omega) - 1) - i\nu\sin(2\omega) \quad \text{with} \quad \mu = \frac{d\tau}{h^2}, \quad \nu = \frac{a\tau}{h}.$$

For the explicit Euler method it can be shown that⁴¹⁾

$$\nu^2 \leq 2\mu \leq 1 \quad (6.22)$$

is a necessary and sufficient condition for stability. Precise conditions for the explicit trapezoidal rule are not so easy to obtain. From Figure 6.1 it can be deduced⁴²⁾ by geometrical considerations that the condition

$$\frac{1}{3}\nu^2 \leq 2\mu \leq 1 \quad (6.23)$$

is sufficient, and also that having $2\mu \leq 1$ and $\frac{1}{3}\nu^2 \leq 1$ is necessary. The condition on ν in terms of μ is not strictly necessary. In particular for more complicated spatial discretizations and ODE methods, finding strict conditions for stability is cumbersome. Some technical derivations of sufficient conditions can be found in Wesseling (2001, Chap. 5).

Damping and Qualitative Behaviour

As for spatial discretizations, we can also consider the qualitative effect of temporal discretization for the advection equation $u_t + au_x = 0$. This is in

⁴⁰⁾ Simply by numerical computations of $\max|1 + z + \frac{1}{2}z^2|$ on the relevant z -curve with various values of ν .

⁴¹⁾ Exercise: Prove that the stability condition (6.22) is necessary and sufficient.

⁴²⁾ Exercise: Prove that (6.23) is sufficient. Use the observation that the ellipse-shaped region $\{z = x + iy \in \mathbb{C} : (x + 1)^2 + \frac{1}{3}y^2 \leq 1\}$ fits in the stability region.

particular of interest for implicit schemes which are used in practice with (relatively) large time steps in order to benefit from their favourable stability properties.

In Figure 6.2 numerical results are plotted for the backward Euler method and the implicit trapezoidal rule. The lay-out is as in Section 3, with the exact solution indicated by a dotted line and the numerical result given by a solid line. As before we consider the advection model equation with $a = 1$, space interval $x \in (0, 1)$ with a periodic boundary condition and initial profile $u(x, 0) = (\sin(\pi x))^{100}$. The spatial discretization is performed with the third-order upwind-biased differences on a fine grid with $h = 10^{-3}$. On this fine grid the spatial errors are small (not visible in the plots), so the errors here can be completely attributed to the time discretizations. The time step was taken as $\tau = \frac{1}{50}$.

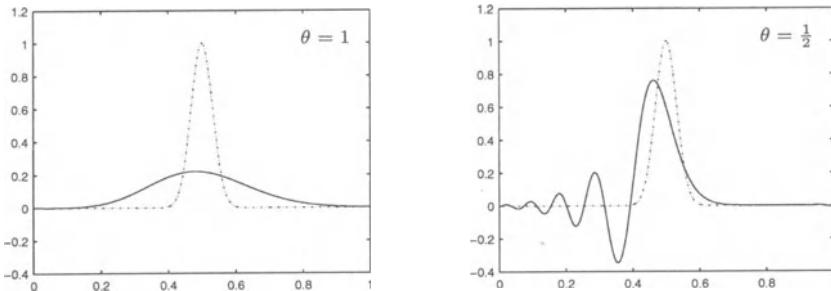


Fig. 6.2. Advection test for the implicit Euler method ($\theta = 1$, left) and the implicit trapezoidal rule ($\theta = 1/2$, right) with time step $\tau = 1/50$ and fine spatial grid.

The results in Figure 6.2 show damping for the implicit Euler method and oscillatory dispersion errors for the trapezoidal rule. In fact, this result for the implicit Euler method is almost identical to the result of the first-order upwind spatial discretization for $h = \frac{1}{50}$ with accurate time integration that was presented in Figure 3.1. This qualitative behaviour can be explained by a consideration of the truncation errors and modified equations.

If we write the semi-discrete system as

$$w'(t) = Aw(t),$$

where A is the discretized advection operator, the θ -method reads

$$w_{n+1} = w_n + (1 - \theta)\tau Aw_n + \theta\tau Aw_{n+1}, \quad (6.24)$$

with $w_0 = w(0)$, and the temporal truncation error equals

$$\begin{aligned} \rho_n &= \frac{1}{\tau}(w(t_{n+1}) - w(t_n)) - (1 - \theta)Aw(t_n) - \theta Aw(t_{n+1}) \\ &= \left(\frac{1}{2} - \theta\right)\tau w''(t_n) + \left(\frac{1}{6} - \frac{1}{2}\theta\right)\tau^2 w'''(t_n) + \dots \end{aligned}$$

Now we can also consider a *modified equation*

$$\tilde{w}'(t) = \tilde{A}\tilde{w}(t), \quad \tilde{A} = A + (\theta - \frac{1}{2})\tau A^2 + (\frac{1}{2}\theta - \frac{1}{6})\tau^2 A^3. \quad (6.25)$$

Assuming this equation to have a stable smooth solution with $\tilde{w}(0) = w(0)$, we can study the error of the θ -method (6.24) with respect to this solution. Then we get a modified truncation error

$$\tilde{\rho}_n = \frac{1}{\tau}(\tilde{w}(t_{n+1}) - \tilde{w}(t_n)) - (1 - \theta)A\tilde{w}(t_n) - \theta A\tilde{w}(t_{n+1}).$$

Using a Taylor expansion and the relations $\tilde{w}' = \tilde{A}\tilde{w}$, $\tilde{w}'' = \tilde{A}^2\tilde{w}$, it follows that

$$\tilde{\rho}_n = \mathcal{O}(\tau^2) \quad \text{if } \theta \neq \frac{1}{2}, \quad \tilde{\rho}_n = \mathcal{O}(\tau^4) \quad \text{if } \theta = \frac{1}{2}.$$

Hence the solution of the modified equation is approximated by the θ -method with a higher order of accuracy than the solution of the original semi-discrete equation.

If we consider sufficiently fine spatial grids, so that the space error is negligible, we can also replace the difference operator A by its continuous counterpart, where $A = -a\partial_x$ for the advection model equation. Hence, neglecting space errors, the modified equation of the θ -method for $u_t + au_x = 0$ is given by

$$\begin{aligned} \tilde{u}_t + a\tilde{u}_x &= \tau a^2(\theta - \frac{1}{2})\tilde{u}_{xx} && \text{if } \theta \neq \frac{1}{2}, \\ \tilde{u}_t + a\tilde{u}_x &= -\frac{1}{12}\tau^2 a^3 \tilde{u}_{xxx} && \text{if } \theta = \frac{1}{2}. \end{aligned} \quad (6.26)$$

For $\theta = 1$ we have artificial diffusion and for $\theta = \frac{1}{2}$ artificial dispersion, similar to (3.17) and (3.18). The numerical results in Figure 6.2 are in close agreement with the exact solutions of these equations.

Remark 6.6 Due to the factor $\theta - \frac{1}{2}$ in front of the diffusive term in (6.26), the modified equation is well posed iff $\theta \geq \frac{1}{2}$, which corresponds nicely with the (in)stability of the θ -method if we use second-order central or higher-order spatial differences, see also Table 6.1. This suggests that stability or instability of a numerical scheme can also be deduced from its modified equation. This type of argument, however, is not theoretically sound and it can be misleading.

For example, the same argument as above can be repeated for the diffusion model equation $u_t = du_{xx}$. In analogy to (6.26) we then get a modified equation

$$\tilde{u}_t = d\tilde{u}_{xx} + \tau d^2(\theta - \frac{1}{2})\tilde{u}_{xxxx}$$

for $\theta \neq \frac{1}{2}$. This equation however is only well posed if $\theta \leq \frac{1}{2}$, and this is totally at odds with the fact that the θ -method is unconditionally stable for the diffusion equation iff $\theta \geq \frac{1}{2}$. Thus it should be emphasized that the considerations based on a modified equation are only sound if the numerical

scheme is stable and the modified equation is well posed. If one of these two assumptions is missing no conclusions can be drawn! If, on the other hand, the assumptions are satisfied, then convergence to the solution of the modified equation follows by the usual stability-consistency argument that was presented in the Sections 2.6 and 6.2.

A modified equation for the θ -method that is well posed for the diffusion equation if $\theta \geq \frac{1}{2}$ is given by

$$\tilde{w}'(t) = (I - (\theta - \frac{1}{2})\tau A)^{-1} A \tilde{w}(t).$$

Neglecting spatial errors we can insert here $A = d\partial_{xx}$. This modified equation however lacks the simplicity and transparency of (6.25). For a further discussion on the scope of modified equations we refer to Griffiths & Sanz-Serna (1986).

Finally we note that insight in qualitative properties can also be obtained by studying the propagation of individual Fourier modes. The diffusive character of the implicit Euler method in Figure 6.2 is related to the fact that $|R(iy)| < 1$ if $y \neq 0$ which gives damping of modes. This will be discussed in more detail in the next section in connection with von Neumann stability analysis. \diamond

Non-normal Matrices and the Eigenvalue Criterion

If we consider, instead of stability for normal matrices, the somewhat wider class of matrices A which are diagonalizable,

$$A = U \Lambda U^{-1}, \quad \Lambda = \text{diag}(\lambda_k),$$

then, with an absolute vector norm, we obtain

$$\|R(\tau A)^n\| \leq \text{cond}(U) \max_{1 \leq k \leq m} |R(\tau \lambda_k)^n|.$$

If the condition number does not grow as $h \rightarrow 0$ and takes on modest values, then the eigenvalue criterion (6.19) can still be applied to assess stability, similar as with normal matrices.

In general, however, one must be very careful with the eigenvalue criterion if A is not normal because it may lead to *wrong* conclusions. A notorious example is given by the forward Euler method combined with first-order upwind discretization (Courant-Isaacson-Rees scheme) for the advection problem

$$u_t + u_x = 0, \quad t > 0, \quad 0 < x < 1,$$

with a given initial function $u(x, 0)$ and the Dirichlet inflow boundary condition $u(0, t) = 0$ for $t \geq 0$. The semi-discrete system reads

$$w'_j(t) = \frac{1}{h} (w_{j-1}(t) - w_j(t)), \quad w_0(t) = 0,$$

where $h = 1/m$, $j = 1, 2, \dots, m$. In vector form we have $w'(t) = Aw(t)$ with

$$A = \frac{1}{h} \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}.$$

This matrix consists of a single Jordan block and it has only one eigenvalue, namely $\lambda = -1/h$. So, with the explicit Euler method we have $\tau\lambda \in \mathcal{S}$ iff $\tau/h \leq 2$. This step size restriction contrasts sharply with the condition $\tau/h \leq 1$ given in Table 6.1 (case $\theta = 0$) for periodic boundary conditions. These findings indicate that for $1 < \tau/h \leq 2$ we get wrong results with the eigenvalue criterion, revealing a failure of this concept for non-normal matrices.

The failure of the eigenvalue criterion for the forward Euler - upwind example is well-known in the literature, see for instance Morton (1980) and van Dorsselaer, Kraaijevanger & Spijker (1993). It is instructive to illustrate it numerically. Using $\tau/h = \frac{3}{2}$, in Figure 6.3 the L_2 -norm growth factors $\|R(\tau A)^n\|_2$ are plotted versus n for $h = \frac{1}{10}$ and $h = \frac{1}{20}$. Clearly, with this ratio the scheme is not stable (with moderate constants). Although we see that $\|R(\tau A)^n\| \rightarrow 0$ for $n \rightarrow \infty$, before this happens $\|R(\tau A)^n\|_2$ can become very large, leading to an unacceptable error propagation. Precise bounds for this example are given in van Dorsselaer et al. (1993).

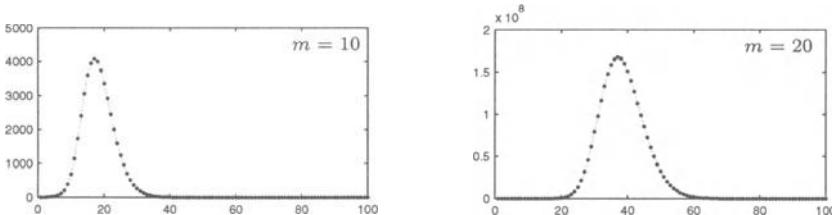


Fig. 6.3. Growth factors $\|R(\tau A)^n\|_2$ as function of n for $\tau/h = 3/2$ with $m = 10$ (left) and $m = 20$ (right).

A further illustration, comparing a stable solution with an unstable one, is given in Figure 6.4. These two pictures show results of a time integration from $t = 0$ till $t = 0.4$, using the initial condition $u(x, 0) = (\sin(\pi x))^{100}$ and a mesh width $h = \frac{1}{50}$. In the left plot we have the Courant number $\nu = \tau/h = \frac{5}{6}$ (stable) and in the right plot $\nu = \tau/h = \frac{5}{4}$ (unstable) was used. The numerical solutions are plotted for $t = 0$ (dotted), $t = 0.2$ (dashed) and $t = 0.4$ (solid lines). For smaller values of h the result with $\nu = \frac{5}{6}$ will approach the exact solution, whereas $\nu = \frac{5}{4}$ will eventually lead to blow-up.

Remark 6.7 The property $\|R(\tau A)^n\| \rightarrow 0$ for $n \rightarrow \infty$ in Figure 6.3 is due to the fact that with ratio $\tau/h = \frac{3}{2}$ the spectral radius is $\rho(R(\tau A)) = \frac{1}{2}$.

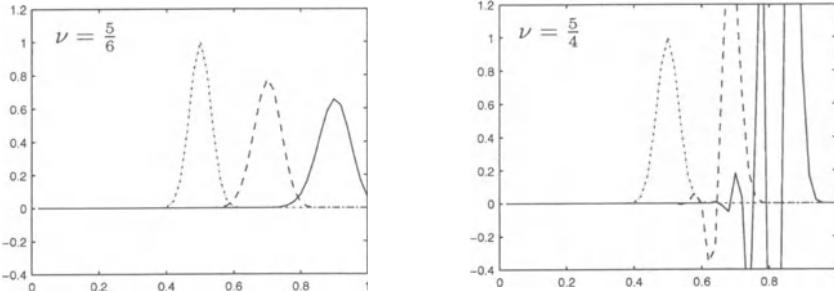


Fig. 6.4. Time evolution with the forward Euler-upwind example for $h = 1/50$ with $\tau/h = 5/6$ (left) and $\tau/h = 5/4$ (right).

If $\rho(R(\tau A)) < 1$, then the matrix $R^n(\tau A)$ approaches the null matrix as $n \rightarrow \infty$, see for example Varga (1962, Thms. 1.4, 3.1). However, the asymptotic behaviour gives no indication about the magnitude of w_{n+1} for finite n . For actual computations this property of asymptotic stability is therefore of little use. \diamond

6.4 Von Neumann Stability Analysis

At various places in our previous sections on spatial discretization we have used the discrete Fourier analysis from Section 3.1 to prove stability. In the 1940's, John von Neumann introduced Fourier analysis in the theory of finite difference schemes for time-dependent PDEs. A first systematic account of von Neumann's ideas has been given by O'Brien, Hyman & Kaplan (1951). After this publication the von Neumann stability analysis has been successfully used throughout the literature on numerical PDEs.

Example 6.8 As a simple example, let us consider the standard diffusion problem $u_t = du_{xx}$, $t > 0$, $0 < x < 1$ with a periodic boundary condition and $d > 0$ constant. Second-order central differences on the uniform grid Ω_h and first-order forward differencing in time gives the fully discrete scheme

$$w_j^{n+1} = w_j^n + \frac{\tau d}{h^2} (w_{j-1}^n - 2w_j^n + w_{j+1}^n), \quad j = 1, 2, \dots, m, \quad (6.27)$$

where $h = 1/m$, $n = 0, 1, \dots$ and $w_0^n = w_m^n$, $w_{m+1}^n = w_1^n$ for all n . As initial value for w_j^0 we first consider a discrete Fourier mode with wave number $1 \leq k \leq m$,

$$w_j^0 = \varphi_k(x_j) = e^{2\pi i k x_j},$$

and we make the ansatz (compare with (1.15))

$$w_j^{n+1} = r_k w_j^n, \quad n \geq 0,$$

or, equivalently,

$$w_j^n = (r_k)^n e^{2\pi i k x_j}, \quad n \geq 0. \quad (6.28)$$

The number r_k acts as *amplification factor* for the k th Fourier mode. Insertion in (6.27) yields

$$r_k^{n+1} e^{2\pi i k x_j} = r_k^n e^{2\pi i k x_j} \left(1 + \frac{\tau d}{h^2} (e^{-2\pi i k h} - 2 + e^{2\pi i k h}) \right),$$

and hence

$$r_k = 1 + \frac{\tau d}{h^2} (e^{-2\pi i k h} - 2 + e^{2\pi i k h}) = 1 - \frac{4\tau d}{h^2} \sin^2(\pi h k).$$

Imposing

$$|r_k| \leq 1 \quad \text{for all } k = 1, 2, \dots, m,$$

gives the stability condition

$$\frac{\tau d}{h^2} \leq \frac{1}{2}. \quad (6.29)$$

Denoting, as in Section 3.1,

$$\phi_k = (\varphi_k(x_1), \varphi_k(x_2), \dots, \varphi_k(x_m))^T \in \mathbb{C}^m,$$

we can express arbitrary starting vectors $w_0 \in \mathbb{R}^m$ as $w_0 = \sum_{k=1}^m \alpha_k \phi_k$ and the approximations w_n as

$$w_n = \sum_{k=1}^m \alpha_k (r_k)^n \phi_k.$$

For the discrete L_2 -norm we then find

$$\|w_n\|_2^2 = \sum_{k=1}^m |\alpha_k|^2 |r_k|^{2n} \leq \sum_{k=1}^m |\alpha_k|^2 = \|w_0\|_2^2,$$

proving L_2 -stability for the homogeneous difference scheme (6.27) under condition (6.29).⁴³⁾ ◇

For the (homogeneous) two-level difference scheme (6.4), component-wise written as

$$w_j^{n+1} = (w_{n+1})_j = (Bw_n)_j, \quad B = B_0^{-1} B_1,$$

von Neumann analysis goes entirely similar. Assuming that B_0 and B_1 are circulant matrices, substitution of the Fourier mode (6.28) into the difference

⁴³⁾ Exercise: The stability condition $\tau d \leq \frac{1}{2}h^2$ found for the explicit diffusion scheme (6.27) is, not surprisingly, the same as in Table 6.1 for $\theta = 0$. For practical applications this time step restriction is very severe. Replace the first-order forward time-stepping by the θ -method, $\theta > 0$. Apply to this scheme a von Neumann stability analysis, determine the stability condition and compare this with Table 6.1.

scheme leads to an amplification factor r_k from which the step size restriction can be found. The amplification factor r_k is the eigenvalue of B associated with the eigenvector ϕ_k . This implies that for the discrete L_2 -norm

$$\|B\|_2 = \max_{1 \leq k \leq m} |r_k|, \quad (6.30)$$

and we see that stability in the sense of (6.13) holds if τ and h are such that

$$|r_k| \leq 1 + \mathcal{O}(\tau). \quad (6.31)$$

This condition is called the *von Neumann condition*. Elaboration of this condition yields the sought step size restriction.

Instead of (6.31) one can also use the somewhat simpler *strict von Neumann condition*

$$|r_k| \leq 1. \quad (6.32)$$

However, the strict condition is not always applicable. For example, with the spatial periodic diffusion-reaction type problem

$$u_t = d u_{xx} + c u,$$

with $c > 0$, $d > 0$, the reaction term leads to exponentially growing solutions. The numerical solution should mimic this growth and hence we cannot require the amplification factors to be bounded by one in modulus.

If the two-level difference scheme (6.4) can be interpreted as a MOL scheme, and B is a circulant, it should be clear now that the strict von Neumann condition is identical to the eigenvalue criterion (6.19) for ODE methods. If B is a normal matrix, but not circulant, (6.30) is still valid with the r_k the eigenvalues of B . Consequently, the von Neumann criteria are then still applicable to assess stability.

Example 6.9 As a final example, consider the Lax-Wendroff scheme (6.7) for $u_t + au_x = 0$, $t > 0$, $0 < x < 1$ with a periodic boundary condition. With $\nu = a\tau/h$, we find amplification factors

$$\begin{aligned} r_k &= 1 + \frac{1}{2}\nu(e^{-2\pi i kh} - e^{2\pi i kh}) + \frac{1}{2}\nu^2(e^{-2\pi i kh} - 2 + e^{2\pi i kh}) \\ &= 1 - 2\nu^2 \sin^2(\omega) - i\nu \sin(2\omega), \quad \omega = \pi kh. \end{aligned}$$

Hence

$$|r_k|^2 = 1 - 4\nu^2(1 - \nu^2) \sin^4(\omega),$$

and we see that the scheme is stable in the L_2 -norm if

$$\frac{\tau|a|}{h} \leq 1.$$

Since the amplification factor r_k determines the numerical propagation of individual Fourier modes, the precise form of r_k can also be used to study

the artificial (numerical) dissipation and dispersion properties, similar as in Section 3 for spatial discretizations. The numerical and exact Fourier modes are

$$w_j^n = (r_k)^n e^{2\pi i k x_j}, \quad u(x_j, t_n) = e^{2\pi i k (x_j - a t_n)}.$$

To study the difference between r_k and $\exp(-2\pi i k a \tau)$, we write

$$r_k = |r_k| e^{-2\pi i k a_k \tau},$$

in which a_k is the numerical phase velocity. With the substitution $\tau = \nu h/a$ it follows that ⁴⁴⁾

$$a_k = \frac{a}{2\nu\omega} \arctan\left(\frac{\nu \sin(2\omega)}{1 - 2\nu^2 \sin^2(\omega)}\right), \quad \omega = \pi k h.$$

If $-1 < \nu < 1$, we see that there will be a damping error, due to $|r_k| \neq 1$, and a phase error, due to $a_k \neq a$. For the qualitative behaviour the phase error will be the dominating contribution since

$$|r_k| = 1 - 2\nu^2(1 - \nu)^2\omega^4 + \mathcal{O}(\omega^8), \quad \omega \rightarrow 0,$$

$$a_k = a\left(1 - \frac{2}{3}(1 - \nu^2)\omega^2\right) + \mathcal{O}(\omega^4), \quad \omega \rightarrow 0.$$

Therefore, we have a damping error $1 - |r_k| = \mathcal{O}(\omega^4)$ whereas the phase error equals $a - a_k = \mathcal{O}(\omega^2)$. Numerical experiments with the Lax-Wendroff scheme indeed show a somewhat dispersive behaviour, similar to the trapezoidal rule in time or central differences in space, but the oscillations are less pronounced due to damping.

Finally we note that as in Section 6.3 for the θ -method, we can also look at the modified equation of the Lax-Wendroff scheme. The equation

$$\tilde{u}_t + a\tilde{u}_x = -\frac{1}{6}ah^2(1 - \nu^2)\tilde{u}_{xxx},$$

is approximated by the Lax-Wendroff scheme with fourth-order accuracy, see also Richtmyer & Morton (1967, Sect. 12.14), once more revealing that the main error contribution is dispersive. \diamond

Remark 6.10 Occasionally we have denoted $\omega = \pi kh$, taken in the interval $[0, \pi]$. It is obvious that the interval could also be chosen as $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$. This is similar to shifting the indices of the Fourier modes as discussed in Remark 3.1.

Further we note that in the literature, frequencies of Fourier modes are often taken to be continuous. This is natural when considering pure initial value problems with the whole real line \mathbb{R} as spatial domain. Then the Fourier series are to be replaced by Fourier integrals in which all frequencies arise, see for instance Strikwerda (1989, Chap. 2) and the remarks on spatial scaling in Section 3.4. Many additional examples of the von Neumann analysis are found in Hirsch (1988) and Iserles (1996), for example. \diamond

⁴⁴⁾ Exercise: Derive the expression for a_k from $\zeta = |\zeta|e^{i\psi}$, $\tan \psi = \text{Im}(\zeta)/\text{Re}(\zeta)$.

Beyond the Standard Fourier Analysis

Von Neumann's stability analysis based on Fourier modes as presented here applies only to problems without boundary conditions and constant coefficients. In practice, however, this type of analysis is also used when these restrictions are not met. In general, a heuristic von Neumann analysis leads to reliable step size criteria. In this heuristic approach boundary conditions are ignored and one 'freezes' the coefficients in the linear difference scheme, which can even be the result of linearization of a nonlinear scheme. The von Neumann analysis can be used in multiple dimensions, for systems of PDEs (instead of an amplification factor r we then get an amplification matrix), and with all kinds of time-stepping formulas. Stability estimates with amplification matrices are usually based on a fundamental result of Kreiss from 1962, known as the Kreiss matrix theorem, which states that power boundedness of a matrix B is equivalent to the resolvent condition

$$\| (B - zI)^{-1} \| \leq \frac{C}{|z| - 1} \quad \text{for all } |z| > 1$$

with constant $C > 0$. An excellent account of the von Neumann stability analysis, including multiple dimensions and systems, can be found in Richtmyer & Morton (1967) and Strikwerda (1989).

With von Neumann's analysis the power boundedness property (6.13) easily follows for scalar problems with constant coefficients and spatial periodicity. To prove that property for problems with boundary conditions or non-constant coefficients is in general very complicated. Many references to early research on this topic can be found in Richtmyer & Morton (1967), see also the survey Thomée (1990). Part of these results concern extensions of Fourier analysis for linear problems with smooth, variable coefficients, for instance the results of John from 1952 on parabolic equations and of Lax & Nirenberg from 1966 on hyperbolic problems, see loc. cit. The classical theory on stability with boundary conditions is the so-called GKS theory, see Gustafsson, Kreiss & Sundström (1972), which was already mentioned in Section 5.4.

More recent progress on the stability issue has been made by Spijker and co-workers using resolvent conditions, see for instance van Dorsselaer, Kraaijevanger & Spijker (1993) and Borovych, Drissi & Spijker (2000), and by Trefethen and co-workers using so-called pseudo-spectra, see Reddy & Trefethen (1992), Trefethen (1997).

When the trapezoidal or implicit Euler method is used for time stepping, the nonlinear results presented in Section 2.7 can be considered. However, establishing the logarithmic matrix norm inequality (2.45) is in general only possible for first- and second-order spatial discretizations for which results have been given in Sections 4 and 5.

7 Monotonicity Properties

As we saw from the numerical illustrations in Section 3, many spatial discretizations produce oscillations and negative values. In Section 1 it was seen that with ODEs describing chemical reactions negative values may lead to instabilities. Here we take a closer look at ODE systems with the aim of identifying those systems for which negative values and spatial oscillations can be expected.

7.1 Positivity and Maximum Principle

Positivity

For advection-diffusion-reaction equations whose solutions are concentrations of chemical species, physical interpretation tells us that

$$u(x, 0) \geq 0 \quad \text{for all } x \implies u(x, t) \geq 0 \quad \text{for all } x \text{ and } t > 0.$$

As we have seen, there is no guarantee that spatial discretizations maintain this property. This of course is undesirable and therefore we would like to have a criterion that tells us when non-negativity is preserved.

In the following we will write $v \geq 0$ for a vector $v \in \mathbb{R}^m$ if all its components are non-negative. Consider an ODE system in \mathbb{R}^m for $t \geq 0$,

$$w'(t) = F(t, w(t)). \tag{7.1}$$

This system will be called *positive* (short for ‘non-negativity preserving’) if

$$w(0) \geq 0 \implies w(t) \geq 0 \quad \text{for all } t > 0.$$

The following theorem provides a simple criterion on F that tells us whether the system is positive.

Theorem 7.1 *Suppose that $F(t, v)$ is continuous and satisfies a Lipschitz condition (2.2) with respect to v . Then system (7.1) is positive iff for any vector $v \in \mathbb{R}^m$ and all $i = 1, \dots, m$ and $t \geq 0$,*

$$v \geq 0, \quad v_i = 0 \implies F_i(t, v) \geq 0. \tag{7.2}$$

Proof. Necessity of the above criterion follows immediately by considering the solution $w(t)$ of (7.1) with $w(0) = v$ and small $t > 0$. As for sufficiency, note that the criterion is equivalent with

$$w(t) \geq 0, \quad w_i(t) = 0 \implies w'_i(t) \geq 0.$$

This is in itself not enough to prove positivity, we also need the Lipschitz condition. (A counterexample is given in Example 2.1.)

It would be enough to have

$$w(t) \geq 0, \quad w_i(t) = 0 \implies w'_i(t) \geq \varepsilon > 0,$$

because then it is obvious that $w(t)$ cannot cross the hyperplanes $\mathcal{H}_i = \{w \in \mathbb{R}^m : w_i = 0\}$. This will hold for the perturbed ODE system with

$$\tilde{F}_i(t, w) = F_i(t, w) + \varepsilon, \quad i = 1, 2, \dots, m.$$

Using the Lipschitz condition, we can apply a standard stability argument for ODEs to show that the solution of the unperturbed system with given $w(0)$ will be approximated with any precision by solutions of the perturbed system if we let $\varepsilon \rightarrow 0$, see for instance Coppel (1965, Thm. 3). \square

The requirement of a Lipschitz condition in this theorem can be slightly relaxed. It was shown by Horváth (1998) that (7.2) is sufficient if it is assumed in addition that the initial value problem has a unique solution for any $w(0) \geq 0$. On the other hand, establishing a Lipschitz condition is the standard way to show uniqueness.

Theorem 7.2 *The linear system $w'(t) = Aw(t)$ is positive iff*

$$a_{ij} \geq 0 \quad \text{for all } j \neq i. \quad (7.3)$$

Proof. This is a consequence of Theorem 7.1. A more direct proof for linear systems follows from the relation

$$e^{\tau A} = I + \tau A + \mathcal{O}(\tau^2)$$

to show necessity and

$$e^{t_n A} = \lim_{n \rightarrow 0} (I + \tau A)^n \quad \text{with } t_n = n\tau \text{ fixed}$$

to show sufficiency. Hence sufficiency is concluded from convergence of the forward Euler method. \square

Remark 7.3 Related to positivity, one can also study a *comparison principle*

$$w(0) \leq \tilde{w}(0) \implies w(t) \leq \tilde{w}(t) \quad \text{for all } t > 0,$$

for any two solutions of (7.1). Suppose that F is continuously differentiable and that for all $v \in \mathbb{R}^m$, $t \geq 0$ we have

$$\frac{\partial F_i(t, v)}{\partial v_j} \geq 0 \quad \text{for } j \neq i.$$

Then the difference $v(t) = \tilde{w}(t) - w(t)$ of two solutions will satisfy $v'(t) = A(t)v(t)$ where $A(t)$ is the integrated Jacobian matrix as in (2.58), for which we have $a_{ij}(t) \geq 0$ if $j \neq i$. By Theorem 7.1 we thus can conclude that $v(t) \geq 0$ provided $v(0) \geq 0$. \diamond

Maximum Principle

Positivity may also imply a *maximum principle*,

$$\min_j w_j(0) \leq w_i(t) \leq \max_j w_j(0) \quad \text{for all } t \geq 0, \quad (7.4)$$

which we can associate with the absence of unwanted *global* overshoots and undershoots. For linear PDEs without boundary conditions, the semi-discrete system will often satisfy the property

$$F(t, \alpha v + \beta e) = \alpha F(t, v) \quad \text{for all } \alpha, \beta \in \mathbb{R} \text{ and } v \in \mathbb{R}^m, \quad (7.5)$$

where $e = (1, 1, \dots, 1)^T \in \mathbb{R}^m$. This means that if $w(t)$ is a solution of (7.1) and $v(0) = \alpha w(0) + \beta e$, then $v(t) = \alpha w(t) + \beta e$ is also a solution of (7.1). So, in particular, if $0 \leq w_i(0) \leq 1$ and $v_i(0) = 1 - w_i(0)$ for all components i , then $v_i(t) \geq 0$ implies $w_i(t) \leq 1$. More general, if property (7.5) holds, then with choosing β appropriately for shifting initial profiles upward or downward and α to turn over the profile, we see that positivity implies the maximum principle (7.4), and thus global over- and undershoots cannot arise. Of course, (7.4) also implies $\|w(t)\|_\infty \leq \|w(0)\|_\infty$, so for linear systems we can conclude as well maximum-norm stability, i.e., $\|e^{tA}\|_\infty \leq 1$ for all $t \geq 0$.

There are many related monotonicity properties, most notably the so-called TVD property which frequently appears in the literature on computational fluid dynamics. This will be discussed in some detail in Chapter III.

7.2 Positive Semi-discrete Systems

Linear Spatial Advection Schemes

Consider the advection equation $u_t + au_x = 0$ with a constant and periodicity in space. The general linear spatial advection scheme (3.25) was written as

$$w'_j(t) = \frac{a}{h} \sum_{k=-r}^s \gamma_k w_{j+k}(t), \quad j = 1, 2, \dots, m,$$

with $w_i(t) = w_{i+m}(t)$. From Theorem 7.2 it follows that the requirement for positivity is

$$a \gamma_k \geq 0 \quad \text{for all } k \neq 0. \quad (7.6)$$

We can now immediately conclude that the first-order upwind scheme (3.13), (3.14) is positive. However, the second-order central scheme (3.16), the third-order upwind-biased scheme (3.27), (3.28) and the fourth-order central scheme (3.29) are not. Figures 3.1 and 3.3 confirm that from these schemes only first-order upwind is free from oscillations and negative values. These schemes satisfy the property (7.5), and thus the maximum principle applies to the first-order upwind scheme.

Apparently, positivity is a very stringent condition. Indeed, there exists an *order barrier* of one for positivity with advection schemes. Despite its inaccuracy and diffusive nature, the first-order upwind discretization even turns out to be ‘optimal’ under the positive advection discretizations:

Consider the order conditions (3.26) for the general advection discretization (3.25). For order $q \geq 2$ we need $\sum_k k^2 \gamma_k = 0$ and therefore

$$(7.6) \implies q \leq 1.$$

Furthermore, if $q = 1$ then the leading error term is proportional to $|\sum_k k^2 \gamma_k|$. Since we have $\sum_k k \gamma_k = -1$, it follows that

$$(7.6) \implies \left| \sum_k k^2 \gamma_k \right| \geq 1,$$

while the minimal error coefficient $|\sum_k k^2 \gamma_k| = 1$ is achieved by the first-order upwind discretization.

From (7.6) we can not only conclude positivity, but also a comparison principle and the TVD property, which will be discussed in Chapter III, see also the monograph of LeVeque (1992, Sect. 15). The order barrier $q \leq 1$ for positive or monotone advection schemes is due to Godunov (1959). We will refer to this result as *the Godunov barrier*.

It is possible to overcome this order barrier and to derive positive advection schemes significantly more accurate than first-order upwind. We then have to leave the class of linear schemes (3.25) and turn our attention to *nonlinear discretizations* for linear PDEs. This issue will also be taken up in Chapter III.

Linear Spatial Diffusion Schemes

For the diffusion equation $u_t = du_{xx}$ with constant $d > 0$ we can proceed as above. The general linear spatial diffusion scheme (3.37) was written as

$$w'_j(t) = \frac{d}{h^2} \sum_{k=-r}^r \gamma_k w_{j+k}(t), \quad j = 1, 2, \dots, m,$$

with $\gamma_{-k} = \gamma_k$ and $w_i(t) = w_{i+m}(t)$ for spatial periodicity. Thus again the positivity requirement is equivalent with

$$d \gamma_k \geq 0 \quad \text{for all } k \neq 0. \tag{7.7}$$

We see that the second-order central discretization (3.33) is positive, whereas the fourth-order scheme (3.39) fails due to the $-1/12$ coefficients in that formula. Indeed, the requirement of positivity again leads to an order barrier:

Consider the order conditions (3.38) for the general diffusion discretization (3.37). For order $q > 2$ we need $\sum_k k^4 \gamma_k = 0$ and therefore

$$(7.7) \implies q \leq 2.$$

Furthermore, if $q = 2$ then the leading error term is proportional to $\sum_k k^4 \gamma_k$. Since we have $\sum_k k^2 \gamma_k = 2$, it follows that

$$(7.7) \implies \sum_k k^4 \gamma_k \geq 2,$$

while the minimal error coefficient $\sum_k k^4 \gamma_k = 2$ is achieved by the standard second-order central discretization (3.33).

Although this is again somewhat disappointing, the situation is not as bad as for the advection equation, since for many practical purposes the second-order discretization is sufficiently accurate. Moreover, since solutions of linear parabolic problems are often quite smooth, the positivity property is here not as critical as for advection problems. Therefore, in general also the fourth-order discretization can be used for parabolic problems.

Linear Spatial Advection-Diffusion Schemes

Consider the variable coefficient advection-diffusion equation

$$u_t + (a(x, t)u)_x = (d(x, t)u_x)_x$$

with $d(x, t) > 0$ and assume spatial periodicity. Discretization in space by means of the second-order central discretization in flux-form gives

$$\begin{aligned} w'_j &= \frac{1}{2h} \left(a_{j-\frac{1}{2}} (w_{j-1} + w_j) - a_{j+\frac{1}{2}} (w_j + w_{j+1}) \right) \\ &\quad + \frac{1}{h^2} \left(d_{j-\frac{1}{2}} (w_{j-1} - w_j) - d_{j+\frac{1}{2}} (w_j - w_{j+1}) \right), \end{aligned}$$

for $j = 1, \dots, m$, where $w_j = w_j(t)$, $w_0 = w_m$, $w_{m+1} = w_1$ and

$$a_{j \pm \frac{1}{2}} = a(x_{j \pm \frac{1}{2}}, t), \quad d_{j \pm \frac{1}{2}} = d(x_{j \pm \frac{1}{2}}, t).$$

From Theorem 7.1 follows, after an elementary calculation, that this discretization is positive iff the cell Péclet numbers ah/d satisfy the inequality

$$\max_{x,t} \frac{|a(x,t)| h}{d(x,t)} \leq 2. \quad (7.8)$$

If we discretize the advection part with first-order upwind in flux form and the diffusion part as above, we get

$$\begin{aligned} w'_j &= \frac{1}{h} \left(a_{j-\frac{1}{2}}^+ w_{j-1} + (a_{j-\frac{1}{2}}^- - a_{j+\frac{1}{2}}^+) w_j - a_{j+\frac{1}{2}}^- w_{j+1} \right) \\ &\quad + \frac{1}{h^2} \left(d_{j-\frac{1}{2}} (w_{j-1} - w_j) - d_{j+\frac{1}{2}} (w_j - w_{j+1}) \right), \end{aligned} \quad (7.9)$$

where $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$. This semi-discrete system is positive without a cell Péclet restriction.

If the advective velocity a is constant in space, both schemes satisfy property (7.5) and hence we can also conclude stability and convergence in the max-norm. The claim made in the beginning of this section that solutions of the advection-diffusion equation with non-negative initial profiles stay non-negative ('by physical interpretation') can be proven mathematically this way.

The positivity property will hold with spatial periodicity, but also with prescribed non-negative Dirichlet boundary conditions, for example. Examination of other types of boundary conditions, such as homogeneous Neumann at the outflow, is left as exercise. With prescribed boundary conditions or with velocities a that vary in space, the invariance property (7.5) will not hold anymore, but stability in the max-norm can then still be proven by using the logarithmic matrix norm and Lemma 2.6.

General Reaction Equations

A generic form of vector functions for ODE systems derived from the mass-action law of chemical kinetics is the production-loss form

$$F(t, v) = p(t, v) - L(t, v) v, \quad (7.10)$$

where $p(t, v)$ is a vector and $L(t, v)$ a diagonal matrix, whose components $p_i(t, v)$ and $L_i(t, v)$ are of polynomial type with non-negative coefficients, see Section 1.1.

Trivially, the vector functions (7.10) then satisfy the criterion (7.2); in fact, having $p(t, v)$ non-negative suffices. Since $F(t, v)$ depends polynomially on v , it will satisfy a Lipschitz condition on any ball $\|v\| \leq K$, $K > 0$, and hence we can conclude that the associated ODE system $w'(t) = F(t, w(t))$ of chemical reaction rate equations is positive.

Note that these reaction equations describe the evolution of chemical concentrations. The genuine, physical concentrations are non-negative of course. Establishing the positivity property thus means that in this respect the mathematical model is correct.

7.3 Positive Time Stepping Methods

Positivity properties also put restrictions on time integration methods. In this subsection we will discuss this for the forward and backward Euler method, the θ -method, and the explicit trapezoidal rule. Conditions for more general methods will be considered in the next chapter.

Linear Positivity

Consider a linear semi-discrete system $w'(t) = Aw(t)$ where A satisfies

$$a_{ij} \geq 0 \quad \text{for } i \neq j \quad \text{and} \quad a_{ii} \geq -\alpha \quad \text{for all } i, \quad (7.11)$$

with $\alpha > 0$ a fixed number. As we saw from Theorem 7.2, positivity of the system is guaranteed irrespective of the value of α . Naturally we would like to maintain positivity when time integration is performed. It turns out that guaranteeing this for any starting vector $w(0) \geq 0$ is much more restrictive towards the step size τ than stability.

Let us first consider the forward and backward Euler methods. Application of forward Euler to the linear system gives

$$w_{n+1} = (I + \tau A)w_n$$

and we see that $I + \tau A \geq 0$ (inequality component-wise) provided we have $1 + \tau a_{ii} \geq 0$ for all i . This will hold if the step size is restricted such that

$$\alpha\tau \leq 1.$$

The backward Euler method gives

$$w_{n+1} = (I - \tau A)^{-1}w_n.$$

Suppose that

$$A \text{ has no eigenvalues on the positive real axis.} \quad (7.12)$$

Then $I - \tau A$ is invertible for all $\tau > 0$ and thus the implicit relation in the backward Euler method has a unique solution. In fact, this solution is also positive because the conditions (7.11), (7.12) imply

$$(I - \tau A)^{-1} \geq 0 \quad \text{for all } \tau > 0.$$

The proof of this statement will be given in Lemma 7.4 for a nonlinear case.

With these results for forward and backward Euler, results for the θ -method

$$w_{n+1} = w_n + (1 - \theta)\tau Aw_n + \theta\tau Aw_{n+1}$$

follow immediately, since this method can be viewed as a combination of a step with forward Euler using a step size $(1 - \theta)\tau$ followed by a step with backward Euler using $\theta\tau$. Thus positivity is guaranteed if the step size is restricted such that

$$\alpha\tau \leq 1/(1 - \theta).$$

In other words, for the stability function $R(z) = (1 + (1 - \theta)z)/(1 - \theta z)$ of the θ -method we have

$$R(\tau A) \geq 0 \quad \text{if } \alpha\tau \leq 1/(1 - \theta). \quad (7.13)$$

For linear systems with a non-negative source term,

$$w'(t) = Aw(t) + g(t),$$

with A satisfying (7.11) and $g(t) \geq 0$ for all $t \geq 0$, the above results for the Euler methods and the θ -method are easily seen to remain unchanged. Note that the step size conditions given here are sufficient conditions for positivity. In Chapter II a general theory will be presented from which it can be seen that the conditions are also necessary for the class of linear problems satisfying (7.11).

Nonlinear Positivity

For the general nonlinear system $w'(t) = F(t, w(t))$, the counterpart of the linear positivity condition (7.11) reads: there is an $\alpha > 0$ such that

$$v + \tau F(t, v) \geq 0 \quad \text{for all } t \geq 0, v \geq 0 \text{ and } \alpha\tau \leq 1. \quad (7.14)$$

Obviously, this condition guarantees positivity for forward Euler. For linear systems $w'(t) = Aw(t)$ we also assumed that A has no eigenvalues on the positive real axis to ensure that the implicit relations for the backward Euler method have a unique solution. As nonlinear counterpart we introduce the assumption that

for any $v \geq 0$, $t \geq 0$ and $\tau > 0$ the equation $u = v + \tau F(t, u)$ has a unique solution that depends continuously on τ and v . (7.15)

This means that the backward Euler relation is well defined and, according to the following lemma, it implies unconditional positivity as well.

Lemma 7.4 *Conditions (7.14)-(7.15) imply positivity for backward Euler for any step size $\tau > 0$.*

Proof. For given t, v and with τ variable we consider the equation $u = v + \tau F(t, u)$ and we call its solution $u(\tau)$. We have to show that $v \geq 0$ implies $u(\tau) \geq 0$ for all positive τ . By continuity it is sufficient to show that $v > 0$ implies $u(\tau) \geq 0$. This is true (even $u(\tau) > 0$) because if we assume that $u(\tau) > 0$ for $\tau \leq \tau_0$, except for the i th component $u_i(\tau_0) = 0$, then

$$0 = u_i(\tau_0) = v_i + \tau_0 F_i(t, u(\tau_0)).$$

According to (7.14) we have $F_i(t, u(\tau_0)) \geq 0$ and thus $v_i + \tau_0 F_i(t, u(\tau_0)) > 0$ which is a contradiction. □

Condition (7.14) can often be verified directly, but condition (7.15) is not very transparent. Sufficient for it to hold is that F is continuously differentiable and that

$$\|(I - \tau J(t, v))^{-1}\| \leq C \quad \text{for any } v \in \mathbb{R}^m, t \geq 0 \text{ and } \tau > 0,$$

with C some positive constant and $J(t, v)$ the Jacobian matrix of derivatives of $F(t, v)$ with respect to v . Existence and uniqueness of the solution then

follows from Hadamard's theorem and by the implicit function theorem this solution depends continuously on τ, t and v , see Ortega & Rheinboldt (1970, pp. 128, 137).

As for the linear case, these results for the forward and backward Euler method can be directly used with the θ -method to obtain the following result.

Corollary 7.5 *Conditions (7.14)-(7.15) imply positivity for the θ -method for any step size $\tau > 0$ satisfying $\alpha\tau \leq 1/(1 - \theta)$. \square*

In particular, the trapezoidal rule ($\theta = \frac{1}{2}$) is nonlinearly positive if $\alpha\tau \leq 2$. Note that in spite of the A -stability of the method, the positivity requirement imposes a step size restriction.

Next consider the explicit trapezoidal rule

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_n + \tau F(t_n, w_n)),$$

see (6.18). If we write this method as

$$\begin{aligned} \bar{w}_{n+1} &= w_n + \tau F(t_n, w_n), \\ w_{n+1} &= \frac{1}{2}w_n + \frac{1}{2}(\bar{w}_{n+1} + \tau F(t_{n+1}, \bar{w}_{n+1})), \end{aligned} \tag{7.16}$$

it follows from (7.14) that we have nonlinear positivity if $\alpha\tau \leq 1$. Although this restriction is worse than what we get for the implicit method, it is incomparable to the difference in stability between this implicit and explicit method. Thus we see that stability and positivity are not directly related, and when positivity is a strict requirement, unconditionally stable implicit methods may lose their advantage, of course with the backward Euler method as the favourable exception.

7.4 Numerical Illustrations

The difference between the implicit trapezoidal rule and the backward Euler method with respect to positivity for the standard advection equation has already been illustrated in Figure 6.2. As we will see in later chapters, the requirement of positivity and related monotonicity properties is of particular interest with *nonlinear* advection discretizations. Here we briefly consider two examples for diffusion and reaction equations.

A Diffusion Problem with Initial Discontinuity

Consider the parabolic initial-boundary value problem

$$u_t = u_{xx}, \quad t > 0, \quad 0 < x < 1,$$

with boundary condition $u(0, t) = u(1, t) = 0$ for $t \geq 0$ and initial function

$$u(x, 0) = \begin{cases} 0 & \text{for } 0 < x < \frac{1}{2}, \\ 1 & \text{for } \frac{1}{2} \leq x < 1, \end{cases}$$

giving discontinuities at $x = \frac{1}{2}, 1$ for $t = 0$. Space discretization with second-order central differences gives positive approximations $w_i(t) \approx u(x_i, t)$ by

$$w'(t) = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} w(t), \quad w_i(0) = \begin{cases} 0 & \text{for } 1 \leq i < \frac{1}{2}m, \\ 1 & \text{for } \frac{1}{2}m \leq i \leq m, \end{cases}$$

with $x_i = ih$ and $h = 1/(m + 1)$. Application with $\tau = h = 1/50$ of the backward Euler method and the trapezoidal rule – also known for parabolic problems as the Laasonen and Crank-Nicolson scheme, respectively⁴⁵⁾ – gives the approximate solutions shown in Figure 7.1. The trapezoidal rule can be seen to give temporal oscillations and negative values near the points $x = \frac{1}{2}$ and $x = 1$.

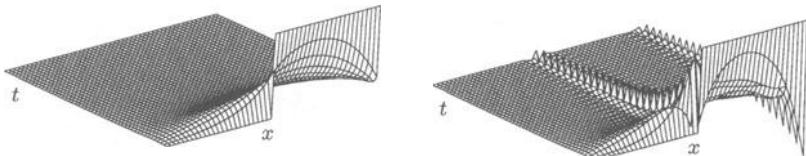


Fig. 7.1. Discontinuous diffusion solutions with backward Euler (left) and the trapezoidal rule (right). Time evolution for increasing t in upper-left direction.

A sufficient condition for positivity of the trapezoidal rule is $\tau/h^2 \leq 1$. This inequality follows from (7.11), (7.13) with $\alpha = 2/h^2$ and $\theta = \frac{1}{2}$. Obviously, it is not satisfied here. It should be noted that the behaviour of the trapezoidal rule in this example is actually determined by several properties. Firstly, due to lack of positivity, the scheme produces large over- and undershoots in the initial phase. Secondly, due to the poor damping properties of the trapezoidal rule, these over- and undershoots persist for a long time. Moreover, due to the fact that $R(z) = -1$ for $z = -\infty$, high-frequency spatial Fourier modes are amplified by a factor close to -1 , and this causes the oscillatory behaviour in the figure.

In practice, problems with positivity are not very often encountered with linear parabolic equations. The solutions for such problems are in general rather smooth and then negative values will not show up due to sufficient accuracy. Also in the discontinuous example presented here negative values could have been avoided by starting the trapezoidal rule with small τ and then gradually increasing the time step.

⁴⁵⁾ In the classic numerical PDE literature, backward Euler and the trapezoidal rule are also known under the names Laasonen scheme and Crank-Nicolson scheme, respectively. This goes back to Laasonen (1949) and Crank & Nicolson (1947).

Condition (7.13) was formulated for arbitrary matrices A . For the above matrix $A = h^{-2} \text{tridiag}(1, -2, 1)$ this condition is a bit too strict. MATLAB experiments show that for this particular matrix the actual upper bound on τ/h^2 is approximately 1.17 for the Crank-Nicolson scheme, which is still close to the general bound (7.13). With spatial periodicity conditions it can be proven that the Crank-Nicolson scheme will be positive if $\tau/h^2 \leq \frac{3}{2}$, see Dautray & Lions (1993, p. 50).

A Stiff ODE System from Atmospheric Chemistry

As a second illustration we consider the stiff system defined by the atmospheric reaction set (1.8). The solution has large jumps over the whole time interval due to the photochemical reaction which is switched on at sunrise and switched off at sunset, see Figure 1.1. The system has been integrated with the trapezoidal rule and the backward Euler method. Figure 7.2 shows the trapezoidal rule solution for a 1 hour step size, where the implicitly defined unknowns were obtained by modified Newton iteration in high accuracy.

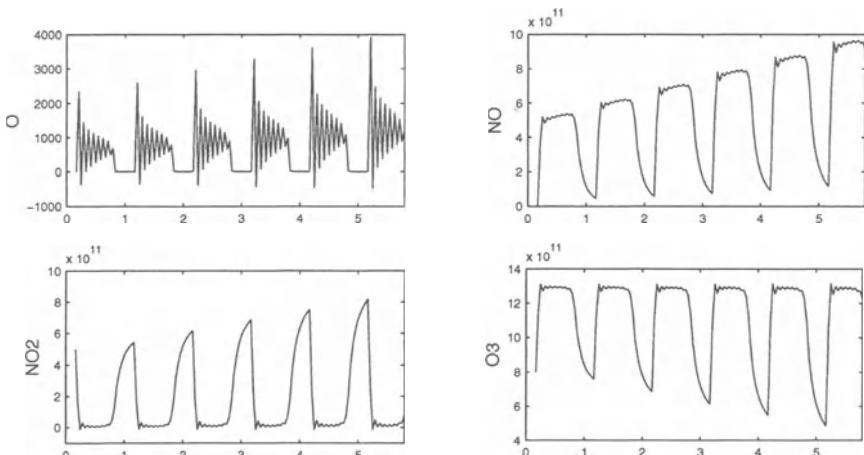


Fig. 7.2. Time evolution with the trapezoidal rule for the concentrations of the atmospheric chemistry problem (1.8), 6 day period with step size of 1 hour.

The lack of positivity is evident here. The approximation for the atomic oxygen O suffers from severe oscillations and becomes negative after sunrise. The concentrations for NO_2, O_3, NO , being much larger in size, show relatively small oscillations. The backward Euler solution (not pictured) is positive and completely free from oscillations. It is, however, somewhat inaccurate in O and NO_2 at the peaks. For a 1/2 hour step size both methods perform satisfactorily. For the trapezoidal rule there are some minor oscillations left in O but oscillations in NO_2, O_3, NO are no longer visible.

A $1/2$ to 1 hour step size is not uncommon in real life atmospheric chemistry computations and stiff ODEs from this field often have Jacobian matrices with a real negative spectrum, see Verwer et al. (1999, 2002). In the present example, the 4×4 Jacobian has two eigenvalues equal to zero (two mass conservation laws), one eigenvalue close to zero, and one approximately equal to -10^5 . With such a structure the L -stability property (damping) of the backward Euler method is to be preferred over the marginal A -stability property (no damping) of the trapezoidal rule, as it better deals with the second reaction in the system which introduces the stiff eigenvalue. These observations are in accordance with the findings for the simple two-way reaction in Figure 2.3.

We stress that in other applications the damping property of the backward Euler method might be non-physical, turning it into a disadvantage. The examples of the next section will illustrate this. Finally we note that an interesting approach to enforce positivity for chemical kinetics, while obeying the mass conservation, has been obtained by Sandu (1999).

8 Numerical Test Examples

The basic discretizations introduced in this chapter will be discussed here in some detail for two 1D test examples. We will consider the behaviour of various spatial discretizations and time stepping schemes. For the latter we consider the explicit trapezoidal rule (6.18) and the implicit θ -method (6.2) with $\theta = 1$ (backward Euler) and $\theta = \frac{1}{2}$ (trapezoidal rule). Comparing the performance of explicit and implicit methods depends to a large extent on the implementation of implicit relations and the programming environment. Although we do not want to descend into details, some general comments are in order here.

Solving implicit relations: Application of the θ -method to a semi-discrete system $w'(t) = F(w(t))$ leads in each time step to an algebraic system

$$w_{n+1} - w_n - (1 - \theta)\tau F(w_n) - \theta\tau F(w_{n+1}) = 0$$

with unknown vector w_{n+1} . This is solved by a modified Newton iteration

$$v^{k+1} = v^k - (I - \theta\tau A_n)^{-1}(v^k - w_n - (1 - \theta)\tau F(w_n) - \theta\tau F(v^k)) \quad (8.1)$$

for $k \geq 0$. The matrix A_n is an approximation to the Jacobian matrix $F'(w_n)$. The starting value can be chosen as $v^0 = w_n$ or $v^0 = w_n + \tau F(w_n)$. Usually the latter choice is a bit faster, but the first choice is sometimes more robust in transient phases. In the present tests the difference was marginal.

Note that per step, from t_n to t_{n+1} , the Jacobian approximation is held fixed during the iteration (modified Newton iteration) rather than updated at each iteration step (standard Newton iteration). These forms of Newton

iteration are discussed in many numerical analysis textbooks, see for example Ortega & Rheinboldt (1970). Modified Newton iteration is the common technique in the stiff ODE field. Usually the iteration is terminated if a certain convergence criterion is satisfied, for example if the displacement $\|v^{k+1} - v^k\|$ comes below a certain tolerance (10^{-6} in the tests here). In order to obtain a simple linear problem in the iteration (8.1), the matrix $A_n \approx F'(w_n)$ may be suitably chosen. For example, certain contributions in the exact Jacobian matrix $F'(w_n)$ may be neglected to obtain a simple banded matrix. Of course, this must be done with some care; the convergence speed of the iteration should not suffer too much, otherwise each individual iteration can become easier to compute, but more iterations might be needed.

The programming environment: The programs for the 1D test problems in this chapter have been written in MATLAB. The arising matrices were declared as sparse, and standard *LU*-decomposition was used to solve linear systems. All function evaluations were done directly with vectors, to avoid loops over the components which are not efficient in MATLAB.

When comparing the performance of the various methods, it should be kept in mind that with another programming language, such as FORTRAN, the CPU times would be very different. For example, solving a linear system with a banded matrix, say tridiagonal, is a highly recursive procedure with many memory operations, and in general recursions are treated more efficiently in FORTRAN than in MATLAB. Consequently, the implementation of implicit methods considered here can certainly not be considered as optimal. With these 1D examples specialized subroutines might give a considerable speed-up. It should be realized, however, that most problems in practice are multi-dimensional and then solving the linear systems is always very time consuming, no matter what programming environment or subroutines are used. In this sense, comparisons for 1D problems with non-optimized linear algebra solvers are fair, since it gives an indication for practical problems in more space dimensions, but we emphasize that all CPU timings should be considered merely as rough indications.

8.1 The Nonlinear Schrödinger Equation

The first test example is the nonlinear Schrödinger equation

$$u_t = iu_{xx} + i\gamma|u|^2u, \quad t > 0, \quad x \in \mathbb{R}, \quad (8.2)$$

where $u(x, t)$ is complex scalar, $i = \sqrt{-1}$ and γ is a positive constant. This equation, from quantum mechanics, is often used in numerical tests since it provides a simple case of an equation with *soliton* solutions, examples of which are given below. Some background and references can be found in Sanz-Serna (1984) and Sanz-Serna & Verwer (1986). The examples and discussion here are based on the latter paper.

Some Analytic Properties

Dispersion and nonlinearity: The linear homogeneous Schrödinger equation

$$u_t = iu_{xx}$$

provides a model for the propagation of dispersive waves. It possesses Fourier solutions

$$u(x, t) = e^{2\pi i k(x - a_k t)}, \quad a_k = k,$$

corresponding to translation of the initial profile $u(x, 0) = e^{2\pi i k x}$ with a speed a_k depending on the wave number k . An initial profile consisting of a superposition of such modes will therefore quickly loose its shape by the dispersion.

As we will see, the cubic term in (8.2) opposes dispersion and it is possible for the nonlinear Schrödinger equation to possess solutions where the competing forces of dispersion and nonlinearity balance each other, giving rise to traveling wave solutions with a pulse-like shape which can collide with each other without loosing shape, the so-called solitons.

Spatially homogeneous solutions: Solutions of the ODE

$$v_t = i\gamma|v|^2 v,$$

with general solution $v(t) = b e^{i\gamma|b|^2 t}$, provide spatially homogeneous solutions $u(x, t) = v(t)$ to (8.2).⁴⁶⁾ Linearization of (8.2) around a spatially homogeneous solution with $b \neq 0$ exhibits *growing* Fourier modes (instability with respect to long-wave perturbations, Yuen & Ferguson, 1978).

Conservation laws: The pure initial value problem (8.2) possesses an infinite set of conservation laws (Zakharov & Shabat, 1972). The conservation of the squared L_2 -norm, often called the *energy* of the solution,

$$E(u) = \int_{\mathbb{R}} |u(x, t)|^2 dx,$$

is of particular importance to the discussion here. This energy conservation implies that the growth of the Fourier modes predicted by the linearization around spatially homogeneous solutions cannot persist. The initially unstable modes draw energy from the stable modes, but due to the above conservation property we know that after a while this process will come to an end. The nonlinear Schrödinger equation does have stable solutions that will be used here for numerical tests.

Soliton solutions: The single soliton solution of (8.2) is given by

$$u(x, t) = \sqrt{\frac{2\alpha}{\gamma}} e^{i(\frac{1}{2}cx - (\frac{1}{4}c^2 - \alpha)t)} \operatorname{sech}(\sqrt{\alpha}(x - ct)),$$

⁴⁶⁾ Exercise: Prove that $v(t) = b e^{i\gamma|b|^2 t}$, $b \in \mathbb{C}$, is the general solution of the ODE by first showing that $\frac{d}{dt}|v(t)|^2 = 0$.

with real parameters α, c and with $\operatorname{sech}(x) = 2(e^x + e^{-x})^{-1}$ denoting the hyperbolic secant function. The soliton travels with speed c and its amplitude is governed by α . Obviously, the initial condition should be prescribed as

$$u(x, 0) = \sqrt{\frac{2\alpha}{\gamma}} e^{\frac{1}{2}icx} \operatorname{sech}(\sqrt{\alpha}x).$$

For our numerical tests we consider the superposition of two solitons, a slower one ahead of a faster one, such that initially they are well separated. As time progresses the faster soliton catches the slower one and passes through it in such a way that the shape and velocity remain unchanged, while their phases are shifted. The initial profile is given by

$$u(x, 0) = \sqrt{\frac{2\alpha}{\gamma}} \left(e^{\frac{1}{2}ic_1 x} \operatorname{sech}(\sqrt{\alpha}x) + e^{\frac{1}{2}ic_2(x-\delta)} \operatorname{sech}(\sqrt{\alpha}(x-\delta)) \right) \quad (8.3)$$

with constants $\alpha = \frac{1}{2}$, $\gamma = 1$, $c_1 = 1$, $c_2 = \frac{1}{10}$ and $\delta = 25$. We take the time interval $0 \leq t \leq T = 44$. An illustration of the time evolution, found by accurate numerical simulation, is given in the Figures 8.1 and 8.2.

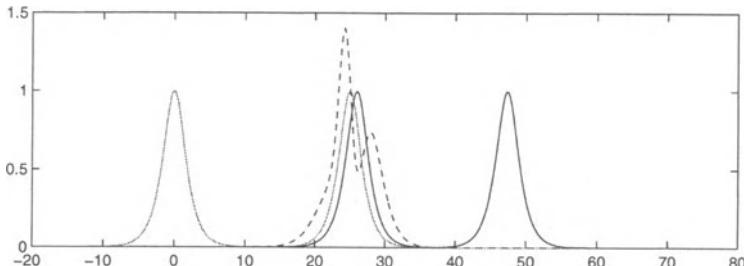


Fig. 8.1. Time evolution with the soliton collision: snapshots of $|u|$ for $x \in [-20, 80]$ at time $t = 0$ (grey), $t = 23$ (dashed) and $t = 44$ (solid).

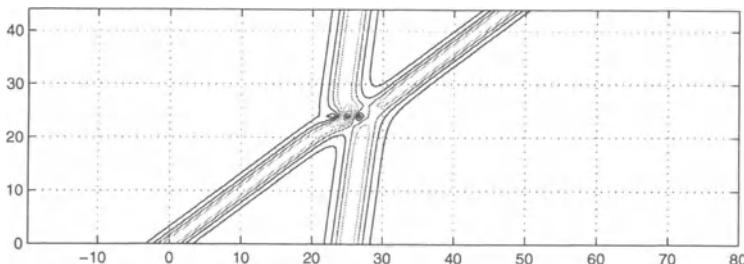


Fig. 8.2. Time evolution with the soliton collision: contour plot of $|u|$ with time $0 \leq t \leq 44$ vertical and space $-20 \leq x \leq 80$ horizontal. The contours levels are $0.2j$, $j \in \mathbb{N}$.

Spatial Discretization

For the numerical simulation of the soliton collision problem (8.2), (8.3) we consider the spatial domain $x_L = -20 < x < x_R = 80$. On the time interval $0 \leq t \leq T = 44$ the solution is negligibly small outside this spatial domain. At the boundaries we prescribe homogeneous Neumann conditions,

$$u_x(x, t) = 0 \quad \text{for } t \geq 0, x = x_L, x_R.$$

This choice of boundary conditions is not essential in the present test. Very similar results are obtained with homogeneous Dirichlet conditions or with periodicity conditions.

Spatial discretization of (8.2) is performed here with second- and fourth-order differences for the u_{xx} term on uniform grids $x_j = x_L + jh$, $j = 0, 1, \dots, m$ with $h = (x_R - x_L)/m$. The semi-discrete solutions are denoted by $w_j(t) \approx u(x_j, t)$. The homogeneous Neumann boundary conditions are enforced by using symmetric virtual values $w_{-j} = w_j$ and $w_{m+j} = w_{m-j}$ in the spatial difference scheme, see Section 5.

We consider grids with $m = 100 \cdot 2^j$ with integers $j \geq 1$. For both discretizations the results were qualitatively wrong with $m = 100$. The results for the other grids are given in Table 8.1. The errors have been estimated by comparing the results with a numerical reference solution calculated on a grid with $m = 3200$ and fourth-order differences. To obtain these results a time integration with very small time steps was performed so that temporal errors play no role here.

m	200	400	800	1600
$q = 2$	1.313	0.4718	0.1154	0.0286
$q = 4$	0.3985	0.0236	0.0015	0.0001

Table 8.1. Relative L_2 -errors in $u(x, T)$ with q th order spatial discretizations.

The asymptotic convergence rate is clearly visible with both the second- and fourth-order discretizations. It should be noted that although the errors with the second-order discretization are very large on the coarser grids, the qualitative behaviour of $|u|$ is correct; the errors are mainly phase errors that arise during the collision of the solitons. In contrast, the results for both discretizations with $m = 100$ are qualitatively incorrect. It is mainly the collision phase that needs a sufficiently fine representation; up to the collision a coarser grid could be used.

Discrete Energy Conservation

As mentioned, energy conservation is important for the dynamics of the nonlinear Schrödinger equation. In a discrete fashion, the energy (squared L_2 -

norm) is also conserved with the spatial discretizations. To see this, we first consider the second-order discretization which gives the semi-discrete system

$$w'_j(t) = \frac{i}{h^2} (w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)) + i\gamma|w_j(t)|^2 w_j(t),$$

with $j = 0, 1, \dots, m$ and $w_{-1}(t) = w_1(t)$, $w_{m+1}(t) = w_{m-1}(t)$. In vector form we can write this as

$$w'(t) = A w(t) + D(w(t)) w(t),$$

where $w(t) = (w_0(t), w_1(t), \dots, w_m(t))^T$,

$$A = \frac{i}{h^2} \begin{pmatrix} -2 & 2 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 2 & -2 \end{pmatrix}, \quad D(w) = i\gamma \operatorname{diag}(|w_j|^2).$$

If we consider the inner product

$$\langle u, v \rangle = \frac{1}{2} h \bar{u}_0 v_0 + h \sum_{j=1}^{m-1} \bar{u}_j v_j + \frac{1}{2} h \bar{u}_m v_m$$

on \mathbb{C}^{m+1} and corresponding norm $\|v\| = \langle v, v \rangle^{1/2}$, it easily follows that

$$\operatorname{Re} \langle v, Av \rangle = 0, \quad \operatorname{Re} \langle v, D(u)v \rangle = 0$$

for all complex vectors u, v . Consequently we have

$$\frac{d}{dt} \|w(t)\|^2 = 2 \operatorname{Re} \langle w(t), Aw(t) + D(w(t))w(t) \rangle = 0,$$

and thus the discrete energy $E_h(w(t)) = \|w(t)\|^2$ is a conserved quantity. Note that the scaling in the inner product is necessary here because of the Neumann boundary conditions. With periodicity one could take the standard L_2 inner product. For the fourth-order discretization a similar conservation property can be shown.⁴⁷⁾

Temporal Discretization

The matrix A arising in the semi-discrete approximation has purely imaginary eigenvalues. This means that the first- and second-order explicit methods considered so far (forward Euler and explicit trapezoidal rule) should

⁴⁷⁾ Exercise: Demonstrate the discrete energy conservation for the fourth-order discretization.

not be considered for this problem due to their lack of stability for imaginary eigenvalues (suitable explicit methods will be considered in the next chapter). Here we will consider only the θ -method (2.31) with the implicit trapezoidal rule ($\theta = \frac{1}{2}$) and the implicit Euler method ($\theta = 1$) as cases of interest. As we will see, the results of the trapezoidal rule are superior to those of the implicit Euler method in the present test. Partly this can be attributed to the higher order, but the main reason is a qualitative property of the implicit Euler method which is unfavourable here, namely damping. It should be emphasized that damping is not an unfavourable property if the equation itself has a dissipative nature, as has been illustrated in Section 7.4. However, with the Schrödinger equation it is unfavourable because it destroys the energy conservation.

As shown above, for this semi-discrete system $w'(t) = F(w(t))$ with $F(w) = Aw + D(w)w$, we have

$$\operatorname{Re} \langle v, F(v) \rangle = 0$$

for any vector v . It follows that the θ -method satisfies

$$\begin{aligned} \|w_{n+1}\|^2 + \theta^2\tau^2\|F(w_{n+1})\|^2 &= \|w_{n+1} - \theta\tau F(w_{n+1})\|^2 \\ &= \|w_n + (1 - \theta)\tau F(w_n)\|^2 = \|w_n\|^2 + (1 - \theta)^2\tau^2\|F(w_n)\|^2. \end{aligned}$$

Therefore, with the trapezoidal rule we have conservation of the quantity

$$\tilde{E}_h(w_n) = \|w_n\|^2 + \frac{1}{4}\tau^2\|F(w_n)\|^2, \quad (8.4)$$

which is an $\mathcal{O}(\tau^2)$ approximation to the discrete energy. On the other hand, with the implicit Euler method we see that

$$\|w_{n+1}\|^2 + \tau^2\|F(w_{n+1})\|^2 = \|w_n\|^2, \quad (8.5)$$

so here the discrete energy is strictly diminishing. This is of course very much related to the dissipative character of the implicit Euler method that was already illustrated numerically in Section 6.3 for the standard advection test problem and in Section 7.4 for the standard parabolic test problem. The damping leads for the present problem to soliton approximations with diminishing magnitude and large temporal errors.

The relative temporal errors in the L_2 -norm are given in Table 8.2 for the second-order spatial differences on a fixed mesh with $m = 800$ and increasing number of time steps N . A reference solution on this mesh was computed with the classical explicit fourth-order Runge-Kutta method, which will be discussed in Chapter II. We see that whereas the trapezoidal rule quickly attains a temporal error on the level of the spatial error, the implicit Euler method gives here very inaccurate results. Even with N up to 6400 there still is no indication of convergence. This is caused by the fact that the damping of the implicit Euler method changes the shape of the solitons and during and after the collision this leads to large phase errors in the numerical solution.

N	50	100	200	400	800	1600
$\theta = \frac{1}{2}$	1.5208	1.1822	0.4535	0.1250	0.0320	0.0081
$\theta = 1$	0.9151	1.4258	1.0176	1.3268	1.2816	1.1486

Table 8.2. Relative temporal L_2 -errors $\|w(t_N) - w_N\|/\|w(t_N)\|$ versus the number of steps N with $\tau = T/N$, $m = 800$, second-order spatial discretization and $\theta = \frac{1}{2}, 1$.

Remark 8.1 In the above we used the complex formulation of equation (8.2). It is also possible to rewrite the equation as a real system

$$\phi_t = -\psi_{xx} - \gamma(\phi^2 + \psi^2)\psi, \quad \psi_t = \phi_{xx} + \gamma(\phi^2 + \psi^2)\phi,$$

corresponding to $u(x, t) = \phi(x, t) + i\psi(x, t)$, see Sanz-Serna & Verwer (1986). This leads to a different numerical implementation, with a semi-discrete system in \mathbb{R}^{2m+2} instead of \mathbb{C}^{m+1} . Here the complex form was chosen for ease of programming. There is one peculiarity that should be mentioned: in the complex form the nonlinear term $i\gamma|u|^2u$ is not differentiable. For the Newton iteration (8.1), the matrix A_n was simply taken as $A + i\gamma\text{diag}(|w_j|^2)$ with A the difference approximation for $i\partial_{xx}$. It was observed in the tests that the iteration (8.1) did give fast convergence with this choice. \diamond

8.2 The Angiogenesis Model

Spatial Discretization

As a second test example we consider the 1D angiogenesis model (1.32). First we focus on spatial discretization aspects. Temporal aspects are discussed later. The spatial discretization of the tumour angiogenesis factor (TAF) equation

$$c_t = \delta c_{xx} - \lambda c - \frac{\alpha\rho c}{\gamma + c}, \quad c(0, t) = 1, \quad c(1, t) = 0,$$

gives little difficulties. In this test we applied standard second-order differences for the diffusion term on the grid points $x_j = jh$, $j = 1, \dots, m$ with $h = (m+1)^{-1}$ and with boundary points $x_0 = 0$, $x_{m+1} = 1$. The discretization for the equation of the endothelial cell density

$$\rho_t = \varepsilon\rho_{xx} - \kappa(c_x\rho)_x + \mu\rho(1-\rho)\max(0, c - c^*) - \beta\rho,$$

$$\rho(0, t) = 0, \quad \rho(1, t) = 1,$$

is more complicated. Again the diffusion term is discretized with second-order central differences. For the advective term $(c_x\rho)_x$ we consider conservative discretization on cells $[x_{j-1/2}, x_{j+1/2}]$ by

$$-(c_x\rho)_x \Big|_{x=x_j} \approx \frac{1}{h} (a_{j-\frac{1}{2}}\rho_{j-\frac{1}{2}} - a_{j+\frac{1}{2}}\rho_{j+\frac{1}{2}}), \quad a_{j+\frac{1}{2}} = \frac{1}{h} (-c_j + c_{j+1}),$$

where the values $\rho_{j+1/2}$ are chosen according to the first-order upwind, second-order central and third-order upwind-biased schemes; see (4.15), (4.17), (4.20). With the third-order scheme linear extrapolation is used to find values outside the domain, as described in Section 5.4. In the following we refer to the resulting schemes simply as first-order upwind, second-order central and third-order upwind-biased, although this actually only applies to the choice of the cell-boundary values $\rho_{j+1/2}$; the approximations for c_x and for the diffusion terms are all second-order accurate.

The parameter values and initial conditions are the same as used earlier in (1.32). If $\delta = 1$ we take the output time as $T = 0.7$ whereas with $\delta = 10^{-3}$ we put $T = 0.5$. In its present form the reaction term $\mu\rho(1 - \rho)$ quickly leads to instabilities when small negative values for ρ are encountered. To avoid these instabilities, this reaction term is replaced by $\mu|\rho|(1 - \rho)$. Even then the problem is quite difficult to solve numerically and a relatively large number of grid points is needed. As we will see, the numerical difficulties are caused mainly by a subtle interplay of the spatial differential operators with the reaction terms.

In Figure 8.3 the L_2 -errors of ρ at time T are given for the three schemes with mesh width $h = 1/25, 1/50, 1/100, 1/200, 1/400$. The errors in the concentrations c are much smaller and therefore not considered here. To obtain these results a time integration has been performed with a very small time step, so that no temporal errors are visible. For the second- and third-order schemes the results on the coarse grids are inaccurate, but with small h the convergence behaviour is satisfactory. In contrast, the first-order upwind scheme gives slow convergence and large errors on all grids. Note that here absolute errors are given. For the larger mesh widths the relative errors of the first-order upwind scheme are approximately 100%.

With $\delta = 1$, where we have smooth solutions, the third-order upwind-biased scheme gives an $\mathcal{O}(h^3)$ convergence rate for these mesh widths in spite of the fact that the diffusion and c_x approximations are only second-order

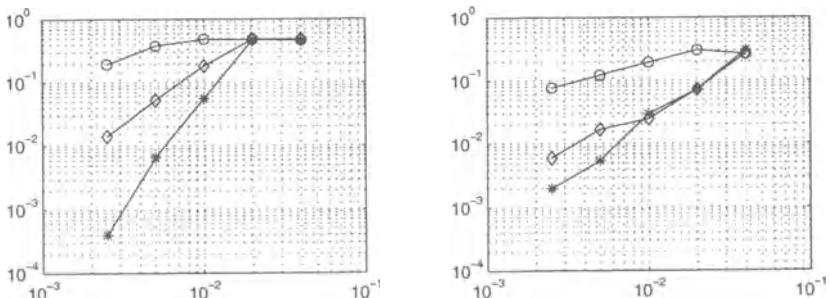


Fig. 8.3. L_2 -errors in ρ -component as function of h with $\delta = 1$, $T = 0.7$ (left) and $\delta = 10^{-3}$, $T = 0.5$ (right). The markers \circ , \diamond , $*$ represent the first-order upwind, second-order central and third-order upwind-biased schemes, respectively.

accurate. As can be observed from Figure 1.4, the solution c is very smooth and therefore the error constants for the $\mathcal{O}(h^2)$ error will be very small. With $\delta = 10^{-3}$ both the third-order and second-order central scheme give an $\mathcal{O}(h^2)$ convergence approximately. The reference solution, by which the errors are measured, has been calculated on a very fine mesh, $h = 1/1600$.

An illustration of the numerical solutions with the three schemes is given in Figure 8.4, where the solutions are plotted at the output time $t = T$ together with the reference solution. To discuss the numerical behaviour some comments on the exact solutions are in order. For both values of δ the cell density solution ρ has formed at time T a front traveling to the left; this front is rather sharp for the small diffusion coefficient $\delta = 10^{-3}$. If the front enters a region where $c > c^*$ then ρ quickly grows due to the reaction term $\mu|\rho|(1 - \rho)(c - c^*)$. Therefore the speed of the front not only depends on the advection term but also on the size of the diffusion coefficient ε ; increasing ε gives a larger speed due to the interaction with the nonlinear reaction term. This effect is most pronounced with $\delta = 1$. For the smaller diffusion coefficient $\delta = 10^{-3}$ the transport of c , which is purely diffusive, is slower. This causes larger gradients of c in the front, and consequently the advection term becomes more dominant with a larger velocity than for $\delta = 1$.

The numerical results in Figure 8.4 with $\delta = 10^{-3}$ are more or less what can be expected for a convection dominated problem with large gradients. Here the results with $h = 1/50$ are displayed, for which the errors are quite large so that the individual lines for the different schemes are easy to distinguish in the plots. We see that the first-order upwind scheme gives a solution that is somewhat too smooth and for which the front travels much too fast. Also this last property is a consequence of the fact that the scheme adds numerical diffusion, which leads to an artificial increase of ε and this gives an increase of the front velocity through the reaction term. The second-

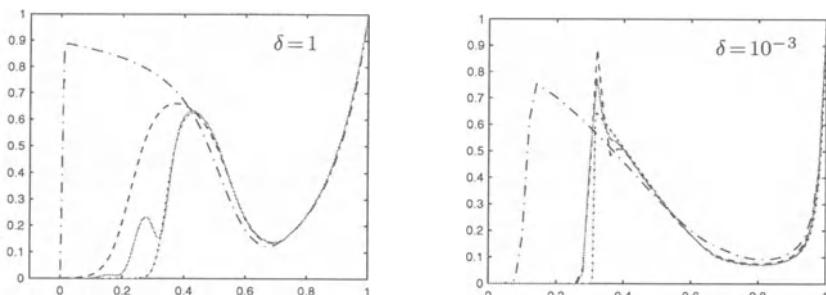


Fig. 8.4. Solutions $\rho(x, T)$ with $\delta = 1$, $T = 0.7$, $h = 1/100$ (left) and $\delta = 10^{-3}$, $T = 0.5$, $h = 1/50$ (right). Numerical solutions for first-order upwind (dash-dots), second-order central (dashed) and the third-order upwind-biased scheme (solid, grey). The reference solution is the dotted line.

third-order scheme produce relatively accurate solutions but with oscillations trailing the front.

The results for $\delta = 1$ are more surprising. The results with $h = 1/100$ are given in the left picture of Figure 8.4. Although the solution is reasonably smooth, all three schemes produce inaccurate approximations. The very bad result of the first-order scheme is again due to its numerical diffusion which leads to a front speed that is much too large. The third-order scheme by nature produces (relatively small) oscillations ahead of fronts, see Figure 3.3, and by the reaction term $\mu|\rho|(1 - \rho)$ these oscillations are amplified towards larger positive values. It should be mentioned that in the original form, without the absolute sign in the reaction term for ρ , the results with the third-order scheme are much worse with large negative values. The result for the second-order central scheme can be understood as follows. The dispersive character of this scheme for advection usually leads to some smearing ahead of fronts and oscillations behind it. Its seems that the smearing ahead of the front, in combination with the reaction term, leads here to a front velocity that is too large. Oscillations behind the front are not visible because of the smoothness of the profile.

It is clear from these results that none of the three schemes considered here gives good results on coarse grids. The main reason seems to be lack of accuracy for the first-order upwind scheme and lack of monotonicity of the other two schemes. In Chapter III advection discretizations will be considered where monotonicity is combined with high accuracy and there we will return to this test problem.

Temporal Discretization

In order to discuss temporal discretizations, we consider a fixed spatial grid with $h = 1/200$ and third-order upwind-biased advection discretization in space. From Figure 8.3 we see that the L_2 -errors in the ρ -component are approximately 0.006 if $\delta = 1$ and 0.005 if $\delta = 10^{-3}$. Here we consider the effect of time integration for the explicit trapezoidal rule (6.18) and the implicit θ -method (2.31) for $\theta = \frac{1}{2}, 1$. We will compare the methods on this fixed spatial grid with respect to accuracy and computational work (CPU) for MATLAB programs on a workstation. As mentioned in the introduction of this section, such comparisons are to be considered as indicative only. The time integration will be considered as ‘accurate’ if the temporal error is not larger than the spatial error.

First consider the problem with $\delta = 1$. Then stability with the explicit method is governed by the discretized diffusion term δc_{xx} . In view of (3.36) and Figure 6.1 we know that the explicit trapezoidal rule requires a step size restriction $\tau \leq \frac{1}{2}h^2$. This leads to a very large number of time steps and consequently to a large CPU time, approximately 10 minutes. On the other hand, with such time steps the method is accurate and there is hardly a temporal contribution to the errors.

The implicit trapezoidal rule is much more efficient for this problem. Considering time steps such that $\nu = \tau/h = 2^{-k}$ with integers k , this method produces accurate result for $\nu \leq \frac{1}{2}$, and $\nu = \frac{1}{2}$ requires about 20 seconds CPU time. The situation with the implicit Euler method is much less favourable. As for the Schrödinger equation, implicit Euler introduces here a significant amount of artificial diffusion (effectively enlarging ε) and this leads for $\nu = \frac{1}{2}$ to an L_2 -error 0.23 in the ρ -component. Moreover halving the step size only halves the error since the method is first-order accurate. To achieve a temporal error comparable with the spatial error the implicit Euler scheme therefore needs ν in the range $[\frac{1}{64}, \frac{1}{32}]$, leading to CPU times between 3 and 6 minutes.

Next we consider the test with $\delta = 10^{-3}$. Here the explicit trapezoidal rule produces accurate results with $\nu \leq \frac{1}{8}$, with a CPU time of approximately 8 seconds. The implicit trapezoidal rule allows larger step sizes with respect to stability, but in order to achieve good accuracy with temporal errors of the same size as the spatial errors we still need a step size such that $\nu \leq \frac{1}{2}$, resulting in a CPU time of approximately 30 seconds. The implicit Euler method is also for this problem very inefficient. As before, this method requires $\nu \in [\frac{1}{64}, \frac{1}{32}]$ to achieve temporal errors smaller than the spatial ones, giving large CPU times in the range 3 – 6 minutes.

As mentioned already, the explicit trapezoidal rule gives good results here with $\nu \leq \frac{1}{8}$. This step size restriction is due to stability for the advective term. This term is nonlinear and therefore simple considerations based on Fourier decompositions are no longer possible. Still, in a heuristic fashion an indication for the allowable step size can be obtained. The advective velocity is $a = c_x$, which is equal to -1 initially. Ignoring the fact that we are dealing with a nonlinear problem with boundary conditions, a von Neumann analysis thus predicts stability for $\nu \leq 0.87$ for small times $t > 0$, see (6.21). As time evolves, the gradient c_x becomes larger, especially near $x = 0$. For that reason ν should be taken smaller; we observed numerically that on the present grid $\nu = \frac{1}{8}$ is sufficiently small.

For actual nonlinear applications, the heuristic fashion to obtain a practical indication for allowable step sizes is quite standard. Very often such a practical indication works well, but some knowledge about the solution might be needed in advance. In practice one therefore often uses variable step sizes τ based on local error estimates, which are then automatically adapted to such varying stability constraints. This will be discussed in Chapter II.

II Time Integration Methods

For the numerical solution of initial value problems for systems of ODEs there are many methods available, such as Runge-Kutta methods and linear multistep methods. In this chapter we give examples of methods which are of interest in the discretization of time-dependent PDEs. We will confine ourselves to methods having a low to moderate order. Further we pay attention to properties of specific interest to PDEs, namely the positivity property and the accuracy behaviour of Runge-Kutta methods for initial-boundary value problems. Excellent general references on ODE methods are Lambert (1991), Hairer, Nørsett & Wanner (1993) and Hairer & Wanner (1996).

In Section I.2 we have already discussed some simple ODE methods and their main properties. As in that section, we consider the general non-autonomous formulation of an initial value problem for a system of ODEs,

$$w'(t) = F(t, w(t)), \quad t > 0, \quad w(0) = w_0,$$

with given $F : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ and $w_0 \in \mathbb{R}^m$. The exact solution $w(t)$ will be approximated in the points $t_n = n\tau$, $n = 0, 1, 2, \dots$ with $\tau > 0$ being the step size. The numerical approximations are denoted by $w_n \approx w(t_n)$. To keep the presentation short we will first focus on fixed, constant step sizes τ . It should be emphasized however that in many PDE applications variable step sizes are crucial to obtain an efficient code. This issue is discussed in Section 5.

1 Runge-Kutta Methods

Runge-Kutta methods are one-step methods (or two-level methods in PDE nomenclature, see Section I.6). That means that they step forward from computed approximations w_n at times t_n to new approximations w_{n+1} at the forward times t_{n+1} using only w_n as input. During a step, however, auxiliary intermediate approximations are computed, denoted by $w_{ni} \approx w(t_n + c_i \tau)$, $i = 1, 2, \dots, s$, where the integer s is called the number of stages. These intermediate approximations serve to obtain a sufficiently high accuracy for the approximations w_n at the main step points $t_n = n\tau$. The general form of a Runge-Kutta method is

$$\begin{aligned} w_{n+1} &= w_n + \tau \sum_{i=1}^s b_i F(t_n + c_i \tau, w_{ni}), \\ w_{ni} &= w_n + \tau \sum_{j=1}^s \alpha_{ij} F(t_n + c_j \tau, w_{nj}), \quad i = 1, \dots, s. \end{aligned} \tag{1.1}$$

Here α_{ij}, b_i are coefficients defining the particular method and $c_i = \sum_{j=1}^s \alpha_{ij}$.

The method is called *explicit* if $\alpha_{ij} = 0$ for $j \geq i$, since then the internal approximations w_{ni} can be computed one after another from an explicit relation. Otherwise the method is called *implicit* due to the fact that the w_{ni} must be retrieved from a system of linear or nonlinear algebraic relations. Following Butcher (1987), a Runge-Kutta method is often represented in a compact way by the Butcher-array

$$\frac{c}{b^T} \left| \begin{array}{c} \mathcal{A} \\ \hline \end{array} \right. = \left| \begin{array}{c|ccc} c_1 & \alpha_{11} & \cdots & \alpha_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & \alpha_{s1} & \cdots & \alpha_{ss} \\ \hline b_1 & \cdots & & b_s \end{array} \right.$$

Remark 1.1 The formula $c_i = \sum_{j=1}^s \alpha_{ij}$ for the abscissa c_i is a convention. It is natural since it implies that the Runge-Kutta method will give the same approximation values for the non-autonomous system $w'(t) = F(t, w(t))$ as for the augmented autonomous form

$$\begin{pmatrix} w(t) \\ t \end{pmatrix}' = \begin{pmatrix} F(t, w(t)) \\ 1 \end{pmatrix}, \tag{1.2}$$

in which the variable t is treated as an independent variable, similarly as the components of w . \diamond

Remark 1.2 Let $v = (v_1, \dots, v_m)^T$ be a given weight vector and suppose that F satisfies

$$v^T F(t, w) = \sum_{j=1}^m v_j F_j(t, w) = 0 \quad \text{for all } t \geq 0, w \in \mathbb{R}^m. \tag{1.3}$$

Trivially, it implies that any solution $w(t)$ of the ODE system satisfies the linear *conservation law* $v^T w(t) = \text{constant}$. This applies for example to chemical reaction systems, see Section I.1. It is a simple exercise to prove that all Runge-Kutta methods (1.1) mimic this conservation property, i.e., $v^T w_{n+1} = v^T w_n$ for all n . \diamond

1.1 The Order Conditions

Consider a fixed time point $t_N = N\tau$. The *global error* at t_N is the difference $w(t_N) - w_N$ between the sought solution starting at the given initial value

$w(0) = w_0$ and the approximation w_N . Clearly, the global error at t_N must depend on errors present in all preceding approximations w_n ($1 \leq n \leq N-1$). By means of stability analysis the global error can be interpreted as being built up from *local* errors, much the same as we have shown for the explicit Euler method and the θ -method in Section I.2. Assuming stability, the local errors thus determine the size of the global error.

Let w_{n+1}^* , w_n^* be the Runge-Kutta approximations obtained by starting at time level t_n on the exact solution, $w_n^* = w(t_n)$. Then $w(t_{n+1}) - w_{n+1}^*$ can be viewed as the error committed in one step; we call this the *local error*. The method is called *consistent* of order p , or more shortly to have order p , if the local error satisfies

$$w(t_{n+1}) - w_{n+1}^* = \mathcal{O}(\tau^{p+1}) \quad (1.4)$$

whenever F is sufficiently differentiable. Considering the fixed point t_N in a convergence analysis, the N local errors are added up so that one power of $\tau = t_N/N$ is lost and the global error becomes $\mathcal{O}(\tau^p)$. Thus a method with order of consistency p will be convergent with order p when applied to a smooth ODE problem.¹⁾

The order p of a Runge-Kutta method is determined by its coefficients α_{ij} , b_i , c_i . By making Taylor developments of $w(t_{n+1})$ and w_{n+1}^* in powers of τ , starting from $w(t_n)$, and requiring that these developments are identical up to $\mathcal{O}(\tau^p)$, one obtains the order conditions for the coefficients. The conditions for $p = 1, 2, 3, 4$ are summarized in Table 1.1 where $\mathcal{C} = \text{diag}(c_i)$ and $c^k = \mathcal{C}^k e$, $e = (1, 1, \dots, 1)^T$.

order p	order conditions	
1	$b^T e = 1$	
2	$b^T c = 1/2$	
3	$b^T c^2 = 1/3$	$b^T \mathcal{A} c = 1/6$
4	$b^T c^3 = 1/4$	$b^T \mathcal{C} \mathcal{A} c = 1/8$
	$b^T \mathcal{A} c^2 = 1/12$	$b^T \mathcal{A}^2 c = 1/24$

Table 1.1. Order conditions of Runge-Kutta methods for $p = 1, 2, 3, 4$.

Setting up the Taylor developments is rather technical. For orders up to four say, it can still be accomplished relatively easily by hand, but the

¹⁾ For stiff ODEs and semi-discrete PDEs this standard rule no longer applies. For these problems we encounter the phenomenon of order reduction. Convergence and order reduction for semi-discrete PDEs will be discussed in some detail in Section 2. For a discussion of these issues for stiff ODEs we refer to the B-convergence chapter of Dekker & Verwer (1984).

derivation of higher-order Runge-Kutta methods is complicated and involves many order conditions. A systematic approach consists of the use of Butcher trees, see Butcher (1987) or Hairer et al. (1993).

The *stage order* q is the minimal order over all internal stages, that is, q is such that

$$w(t_n + c_i \tau) - w_{ni}^* = \mathcal{O}(\tau^{q+1})$$

for $i = 1, \dots, s$, whenever F is sufficiently smooth. Although we are not interested in accuracy of the intermediate approximations, this stage order has some relevance for the accuracy of the main step approximations w_n for semi-discrete systems arising from PDEs with boundary conditions. For any reasonable method it holds that $q \leq p$ and for many methods q is substantially smaller than p .

1.2 Examples

Example 1.3 The three most simple explicit Runge-Kutta methods are given by the following Butcher arrays:

$$\begin{array}{c|c} 0 & 0 \\ \hline 1 & \end{array} \quad \begin{array}{c|cc} 0 & & \\ \hline 1 & 1 & \\ \hline 1/2 & 1/2 & \end{array} \quad \begin{array}{c|cc} 0 & & \\ \hline 1/2 & 1/2 & \\ \hline 0 & 0 & 1 \end{array}$$

The first method is the familiar first-order forward Euler method

$$w_{n+1} = w_n + \tau F(t_n, w_n). \quad (1.5)$$

The second method has two stages and is in fact the second-order explicit trapezoidal rule already encountered in Section I.6,

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_n + \tau, w_n + \tau F(t_n, w_n)). \quad (1.6)$$

It is also called the modified Euler method. The third method also has two stages and is also of second-order. It is sometimes called the one-step explicit midpoint rule and can be written as

$$w_{n+1} = w_n + \tau F(t_n + \frac{1}{2}\tau, w_n + \frac{1}{2}\tau F(t_n, w_n)). \quad (1.7)$$

Typical examples of explicit methods of order three and four are given by the following Butcher arrays:

$$\text{a)} \quad \begin{array}{c|cc} 0 & & \\ \hline 1/3 & 1/3 & \\ \hline 2/3 & 0 & 2/3 \\ \hline 1/4 & 0 & 3/4 \end{array} \quad \text{b)} \quad \begin{array}{c|cccc} 0 & & & & \\ \hline 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ \hline 1 & 0 & 0 & 1 & \\ \hline 1/6 & 1/3 & 1/3 & 1/6 & \end{array} \quad (1.8)$$

Method (1.8.a) is Heun's third-order method. Method (1.8.b) used to be called *the* method of Runge-Kutta. We will refer to it as the classical fourth-order method.

Observe that for the above methods the order equals the number of stages, $p = s$. For $p \geq 5$ no explicit method exists of order p with $s = p$ stages, see Butcher (1987) and Hairer et al. (1993) for more details.

Any explicit Runge-Kutta method has stage order $q = 1$, since the second stage is the forward Euler method with step size $\tau\alpha_{21}$ (the first stage is trivial because $w_{n1} = w_n$). \diamond

Example 1.4 Three low-order implicit Runge-Kutta methods are:
the backward Euler method with $p = q = 1$,

$$w_{n+1} = w_n + \tau F(t_{n+1}, w_{n+1}), \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \quad (1.9)$$

the implicit midpoint rule with $p = 2, q = 1$,

$$w_{n+1} = w_n + \tau F\left(t_n + \frac{1}{2}\tau, \frac{1}{2}w_n + \frac{1}{2}w_{n+1}\right), \quad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \quad (1.10)$$

and the implicit trapezoidal rule with $p = q = 2$,

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_{n+1}). \quad \begin{array}{c|cc} 0 & 0 & 0 \\ \hline 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \quad (1.11)$$

The trapezoidal rule and backward Euler method have been discussed quite extensively in Section I.2 as the θ -method with $\theta = \frac{1}{2}$ and $\theta = 1$.

Generalizations of the above methods, with higher order, are based on quadrature or collocation formulas.²⁾ Two examples are given by the following arrays:

$$\text{a) } \begin{array}{c|cc} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} \\ \hline \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} + \frac{1}{6}\sqrt{3} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \quad \text{b) } \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ \hline 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \quad (1.12)$$

Method (1.12.a) is the 2-stage *Gauss method* of order four. The higher-order Gauss methods, based on Gauss-Legendre quadrature, have order $p = 2s$.

²⁾ With collocation, for given $w_n \approx w(t_n)$, we search for a polynomial v of degree s such that $v(t_n) = w_n$ and $v'(t_n + c_i\tau) = F(t_n + c_i\tau, v(t_n + c_i\tau))$, $i = 1, \dots, s$, giving $w_{n+1} = v(t_{n+1})$. This is equivalent, via $w_{ni} = v(t_n + c_i\tau)$, to a Runge-Kutta method with coefficients

$$\alpha_{ij} = \int_0^{c_i} L_j(t) dt, \quad b_j = \int_0^1 L_j(t) dt, \quad L_j(t) = \prod_{k \neq j} \frac{t - c_k}{c_j - c_k}.$$

Method (1.12.b) is the 2-stage *Radau method* of order three. Radau quadrature (with $c_s = 1$) leads to a class of methods of order $p = 2s - 1$. Other well-known methods are based on Lobatto quadrature (with $c_1 = 0, c_s = 1$), with the trapezoidal rule as example. Properties of collocation methods are found in Hairer et al. (1993, Sect. II.7). All collocation methods have stage order $q = s$. For more comprehensive discussions on general implicit Runge-Kutta methods we refer to Butcher (1987), Dekker & Verwer (1984) and Hairer & Wanner (1996). \diamond

Implicit methods are more expensive per step than explicit ones because the intermediate approximations w_{ni} have to be solved from a system of non-linear algebraic equations, usually by a Newton type iteration. Yet, implicit methods are often used, for instance for parabolic PDEs and stiff chemistry problems, because of their superior stability properties. If the Runge-Kutta matrix \mathcal{A} is full the dimension of the system of algebraic equations is ms and the w_{ni} must be solved simultaneously. With larger s -values this can become very costly. A compromise is found in the *diagonally implicit* methods where \mathcal{A} is lower triangular, so that the internal approximations w_{ni} can be solved subsequently for $i = 1, 2, \dots, s$.

Example 1.5 Two diagonally implicit methods, with a parameter $\gamma > 0$, are

$$\text{a)} \quad \begin{array}{c|cc} \gamma & \gamma \\ \hline 1-\gamma & 1-2\gamma & \gamma \\ \hline & 1/2 & 1/2 \end{array} \quad \text{b)} \quad \begin{array}{c|ccc} 0 & 0 \\ \hline 2\gamma & \gamma & \gamma \\ 1 & b_1 & b_2 & \gamma \\ \hline & b_1 & b_2 & \gamma \end{array} \quad (1.13)$$

with $b_1 = \frac{3}{2} - \gamma - \frac{1}{4\gamma}$ and $b_2 = -\frac{1}{2} + \frac{1}{4\gamma}$. If $\gamma = \frac{1}{2} \pm \frac{1}{6}\sqrt{3}$ both methods have order $p = 3$, whereas we have $p = 2$ for other γ -values. The first method has stage order $q = 1$ since the first stage consists of a backward Euler step, whereas the second method has $q = 2$ due to the fact that its first nontrivial stage is a trapezoidal rule step. Method (1.13.a) is one of the diagonally implicit methods that have been proposed independently by Nørsett (1974) and Crouzeix (1975). According to Crouzeix & Raviart (1980), method (1.13.b) is to be attributed to R. Alt, 1973. \diamond

1.3 The Stability Function

In Section I.2.5 a stability analysis for the θ -method was presented for linear ODE systems. A central role was played by the scalar, complex test equation

$$w'(t) = \lambda w(t).$$

In spite of its simplicity, this test equation is of major importance for predicting the stability behaviour of numerical ODE methods. Let in the following

$z = \tau\lambda$. Application of (1.1) to the test equation gives, similarly as for the θ -method, the scalar recursion

$$w_{n+1} = R(z) w_n,$$

with R the *stability function* of the method. This function can be found to be

$$R(z) = 1 + z b^T (I - z \mathcal{A})^{-1} e, \quad (1.14)$$

where $e = (1, 1, \dots, 1)^T$. By considering $(I - z \mathcal{A})^{-1}$ in terms of determinants, it follows that for explicit methods $R(z)$ is a polynomial of degree $\leq s$. For implicit methods it is a rational function with degree of both denominator and numerator $\leq s$.

The stability function of an explicit method with $p = s \leq 4$ is given by the polynomial

$$R(z) = 1 + z + \frac{1}{2} z^2 + \dots + \frac{1}{s!} z^s. \quad (1.15)$$

For degree $s = 1$ this polynomial is the stability function of the forward Euler method, for $s = 2$ of the explicit trapezoidal and midpoint rule (which have the same stability function), for $s = 3$ of Heun's method and for $s = 4$ of the classical fourth-order method. The stability regions $\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ are pictured in Figure 1.1 for $s = 3, 4$. These regions are to be compared with Figure I.6.1 where the stability regions for the methods with $s = 1, 2$ have been plotted.

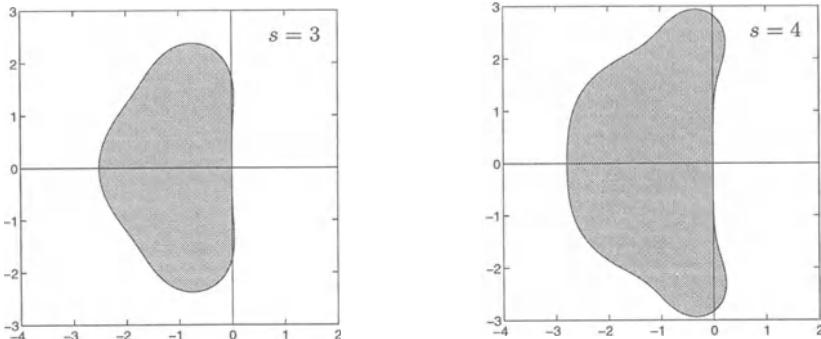


Fig. 1.1. Stability regions \mathcal{S} for the stability functions (1.15) of degree $s = 3, 4$.

An ODE method that has the property that \mathcal{S} contains the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$ is called *A-stable*. Recall, from Section I.2, that A-stability mimics the property $|e^z| \leq 1$ for $z \in \mathbb{C}^-$ and that A-stability is an important property for stiff problems. The exponential function also satisfies

$$|e^z| < 1 \quad \text{if } \operatorname{Re} z < 0, \quad |e^z| \rightarrow 0 \quad \text{as } \operatorname{Re} z \rightarrow -\infty.$$

An ODE method with stability function R is said to be *strongly A-stable* if it is *A-stable* with $|R(\infty)| < 1$, and it is said to be *L-stable* if we have in addition $|R(\infty)| = 0$. For strongly *A*-stable methods we have by the maximum modulus theorem $|R(z)| < 1$ uniformly for $\operatorname{Re} z < -\mu$, whenever $\mu > 0$. On the other hand, strong *A*-stability also implies $|R(iy)| < 1$ for $y \in \mathbb{R}$, $|y| \rightarrow \infty$ and this is at odds with the conservation property

$$|e^z| = 1 \quad \text{if } \operatorname{Re} z = 0.$$

Consequently, strong *A*-stability and *L*-stability are favourable properties for parabolic PDEs or stiff chemical systems with damping, whereas methods that are *A*-stable but not strongly *A*-stable will often give better results for hyperbolic PDEs.

The implicit trapezoidal and midpoint rule share the stability function

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}. \quad (1.16)$$

The stability region for this R is precisely the left-half of the complex plane and hence the implicit trapezoidal and midpoint rule are *A*-stable. Notice that $|R(z)| = 1$ if $\operatorname{Re} z = 0$ and that there is no damping at infinity. The stability function of the backward Euler method is

$$R(z) = \frac{1}{1 - z}, \quad (1.17)$$

hence this method is *L*-stable. The two diagonally implicit methods of Example 1.5 share the stability function

$$R(z) = \frac{1 + (1 - 2\gamma)z + (\frac{1}{2} - 2\gamma + \gamma^2)z^2}{(1 - \gamma z)^2}, \quad (1.18)$$

and both methods are *A*-stable iff $\gamma \geq \frac{1}{4}$. Thus for the two γ -values leading to order three only $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ gives *A*-stability. Further, $R(\infty) = 0$ iff $\gamma = 1 \pm \frac{1}{2}\sqrt{2}$.

Stability functions $R(z)$ are approximations to the exponential function, and for a method of order p we have

$$R(z) = e^z + \mathcal{O}(z^{p+1}), \quad z \rightarrow 0. \quad (1.19)$$

If $R(z)$ is rational with degree s_0 for the numerator and degree s_1 for the denominator, then the maximal order equals $p = s_0 + s_1$. Rational functions with this maximal order are called *Padé approximations*. For explicit methods, with polynomial R , this leads to (1.15). Rational Padé approximations are obtained from the implicit methods considered in Remark 1.4. In particular, for the Gauss methods we have $s_0 = s_1 = s$, $p = 2s$, and for the Radau methods we have $s_0 = s - 1$, $s_1 = s$, $p = 2s - 1$.

Padé approximations and other stability functions have been studied extensively in the numerical ODE literature, see the textbooks mentioned in the beginning of this chapter; in particular we mention the theory of *order stars*³⁾ developed by Wanner, Hairer & Nørsett (1978). An important result of the order star theory is the proof of a conjecture of Ehle (1969):

The Padé approximations are A-stable iff $s_1 - 2 \leq s_0 \leq s_1$.

Hence, the Gauss and Radau methods are all A-stable. In fact, the Radau methods are even L-stable, whereas the Gauss methods possess the conservation property $|R(iy)| = 1$ for all $y \in \mathbb{R}$.

Linear ODE Systems

For a rational function R given by

$$R(z) = \frac{p_0 + p_1 z + \cdots + p_s z^s}{q_0 + q_1 z + \cdots + q_s z^s}, \quad z \in \mathbb{C},$$

we define $R(Z)$ for matrix arguments $Z \in \mathbb{R}^{m \times m}$ as

$$R(Z) = (p_0 I + p_1 Z + \cdots + p_s Z^s)(q_0 I + q_1 Z + \cdots + q_s Z^s)^{-1}. \quad (1.20)$$

Application of the Runge-Kutta method (1.1) to the linear ODE system

$$w'(t) = Aw(t) + g(t)$$

will yield a recursion

$$w_{n+1} = R(\tau A) w_n + \sum_{j=1}^s Q_j(\tau A) \tau g(t_n + c_j \tau), \quad (1.21)$$

where R is the stability function (1.14) and the rational functions Q_j are given by

$$Q_j(z) = b^T(I - z\mathcal{A})^{-1}e_j \quad (1.22)$$

with e_j being the j th unit vector (all components 0 except the j th one which equals 1). Note that the Runge-Kutta coefficient matrix $\mathcal{A} = (\alpha_{ij})$ is an $s \times s$ matrix, and this is to be distinguished from the matrix $A \in \mathbb{R}^{m \times m}$ of the problem (the close resemblance in notation is unfortunate but historical). The derivation of (1.21) is easily obtained for scalar problems by writing the intermediate approximations w_{ni} in a s -dimensional vector. Then for systems, the same derivation can be used by replacing scalar quantities z by matrices $Z = \tau A$.

³⁾ In the order star theory one considers the set $\{z \in \mathbb{C} : |R(z)| > |e^z|\}$, or its complement, instead of the stability region. These sets are star-shaped near the origin, and relations between stability and the order of a method can be derived in a surprisingly simple fashion.

When considering stability, we are interested in the difference $v_n = \tilde{w}_n - w_n$ for two sequences $\{\tilde{w}_n\}$, $\{w_n\}$ starting from initial vectors \tilde{w}_0 and w_0 , respectively. According to (1.21) we have $v_{n+1} = R(\tau A)v_n$, or

$$v_n = R(\tau A)^n v_0.$$

Stability is thus determined by power boundedness of $R(\tau A)$.

Assuming that A is diagonalizable, $A = U\Lambda U^{-1}$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$, it is seen from (1.20) that

$$R(\tau A) = U \text{diag}(R(z_1), \dots, R(z_m)) U^{-1}$$

with $z_j = \tau\lambda_j$. It follows that in an absolute vector norm we have

$$z_j \in \mathcal{S}, j = 1, \dots, m \implies \|R(\tau A)^n\| \leq \text{cond}(U) \text{ for all } n \geq 1. \quad (1.23)$$

In particular, if A is normal then $\|R(\tau A)\| \leq 1$ in the L_2 -norm. This is completely similar to the results for the θ -method that were discussed in the Sections I.2 and I.6 of the previous chapter.

With respect to generalizations for non-normal matrices, we mention a theorem of von Neumann from 1951 which states that $\|R(\tau A)\| \leq 1$ whenever $R(z)$ is A -stable and $\langle v, Av \rangle \leq 0$ for all vectors v , with an inner product norm $\|v\| = \langle v, v \rangle^{1/2}$. A proof of this theorem was given in Section I.2 in the special case of the θ -method; for a general proof we refer to Hairer & Wanner (1996, Sect. IV.11). This theorem of von Neumann should not be confused with the von Neumann stability analysis for PDEs. The analysis deals with Fourier transformations for circulant matrices A , whereas the theorem is applicable to non-normal matrices.⁴⁾

Remark 1.6 In Section I.2.7 stability results for the θ -method were presented for *nonlinear* problems $w'(t) = F(t, w(t))$ satisfying the one-sided Lipschitz condition

$$\langle \tilde{v} - v, F(t, \tilde{v}) - F(t, v) \rangle \leq \omega \|\tilde{v} - v\|^2 \text{ for all } t \geq 0, \tilde{v}, v \in \mathbb{R}^m. \quad (1.24)$$

This condition has been introduced in the numerical ODE field by Dahlquist (1975) to extend his concept of A -stability to stiff nonlinear systems satisfying (1.24) with $\omega = 0$. Dahlquist's focus was on multistep methods. Shortly thereafter, Butcher (1975) introduced the related property B -stability for implicit Runge-Kutta methods, where it is required that

$$\|\tilde{w}_{n+1} - w_{n+1}\| \leq \|\tilde{w}_n - w_n\|$$

under (1.24) with $\omega = 0$, for any two sequences $\{\tilde{w}_n\}$, $\{w_n\}$ computed by the Runge-Kutta method with arbitrary step size τ . This gives a generalization

⁴⁾ Both the von Neumann theorem and the von Neumann stability analysis refer to John von Neumann [1903-1957]. The Neumann boundary conditions for PDEs refer to Carl G. Neumann [1832-1925].

of A -stability to nonlinear ODE systems. Algebraic characterizations were obtained by Burrage & Butcher (1979) and Crouzeix (1979). Many fully implicit Runge-Kutta methods turn out to be B -stable, in particular the Gauss and Radau methods. For a thorough treatment of B -stability, see for instance Dekker & Verwer (1984) or Hairer & Wanner (1996). \diamond

1.4 Step Size Restrictions for Advection-Diffusion

The von Neumann analysis for PDE discretizations is valid for problems with constant coefficients and periodicity in space. If the temporal and spatial discretization can be separated (Method of Lines), this leads to the eigenvalue criterion $z = \tau\lambda \in \mathcal{S}$ with λ representing the eigenvalues of the spatial difference operator and \mathcal{S} being the stability region of the ODE method, see Section I.6. From this eigenvalue criterion one then can deduce the step size restrictions that are needed for stability. Here we list such restrictions for the advection problem $u_t + au_x = 0$ and the diffusion problem $u_t = du_{xx}$ for explicit Runge-Kutta methods having the Padé polynomials (1.15) as stability function.

Stability Restrictions for Advection

First consider the advection problem $u_t + au_x = 0$ with $a > 0$. Step size restrictions for this problem are often called CFL conditions, and they are formulated in terms of the Courant number $\nu = \tau a / h$ where h stands for the mesh width in space. For the spatial advection discretizations that were discussed in Section I.3 the relevant expressions for $z = \tau\lambda$ are

$$\begin{aligned} \text{first-order upwind: } z_{a,1} &= \nu(e^{-2i\omega} - 1), \\ \text{second-order central: } z_{a,2} &= \frac{1}{2}\nu(e^{-2i\omega} - e^{2i\omega}), \\ \text{third-order upwind-biased: } z_{a,3} &= \frac{1}{6}\nu(-e^{-4i\omega} + 6e^{-2i\omega} - 3 - 2e^{2i\omega}), \\ \text{fourth-order central: } z_{a,4} &= \frac{1}{12}\nu(-e^{-4i\omega} + 8e^{-2i\omega} - 8e^{2i\omega} + e^{4i\omega}), \end{aligned}$$

where $\omega \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$. If $a < 0$, we can put $\nu = \tau|a|/h$ and replace ω by $-\omega$. The maximal Courant numbers are given in Table 1.2. These numbers were

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$z_{a,1}$	1.00 (1.00)	1.00 (0.50)	1.25 (0.41)	1.39 (0.34)
$z_{a,2}$	0.00 (0.00)	0.00 (0.00)	1.73 (0.57)	2.82 (0.70)
$z_{a,3}$	0.00 (0.00)	0.87 (0.43)	1.62 (0.54)	1.74 (0.43)
$z_{a,4}$	0.00 (0.00)	0.00 (0.00)	1.26 (0.42)	2.05 (0.51)

Table 1.2. Maximal Courant numbers $\nu = \tau a / h$ for stability, with scaled numbers ν/s in parentheses.

determined experimentally, accurate up to the second digit. The degree s equals the number of function evaluations in the corresponding Runge-Kutta method, and therefore the proper measure for efficiency is the scaled Courant number ν/s . We see that with respect to stability, the forward Euler method ($s = 1$) is the most efficient with first-order upwind, Heun's method ($s = 3$) with third-order upwind-biased, and the classical fourth-order method ($s = 4$) with second- and fourth-order spatial central discretizations.

Stability Along the Imaginary Axis

Central spatial discretization schemes for advection problems with constant coefficients give rise to purely imaginary eigenvalues. Consequently, the largest Courant numbers are then determined by the *imaginary stability boundary* β_I . By definition, $[-i\beta_I, i\beta_I]$ is the largest segment of the imaginary axis contained in \mathcal{S} . For stability one would like the scaled boundary β_I/s to be as large as possible:

For explicit, consistent Runge-Kutta methods (1.1), the imaginary stability boundary $\beta_I = s - 1$ can be attained.

In fact, it was shown by van der Houwen (1977) that the upper bound $\beta_I \leq 2 \lfloor s/2 \rfloor$ is valid for consistent explicit methods (with $\lfloor x \rfloor$ denoting downward integer rounding), and that equality $\beta_I = s - 1$ is obtained if s is odd for a certain polynomial having order $p = 2$. Independently, Kinnmark & Gray (1984) and Sonneveld & van Leer (1985) later showed that with the same polynomial the optimal limit $\beta_I = s - 1$ is also reached for s even, except that then we have order $p = 1$. For $s = 2, 3, 4$ these polynomials are given by

$$\begin{aligned} P_2(z) &= 1 + z + z^2, & \beta_I &= 1, \\ P_3(z) &= 1 + z + \frac{1}{2}z^2 + \frac{1}{4}z^3, & \beta_I &= 2, \\ P_4(z) &= 1 + z + \frac{5}{9}z^2 + \frac{4}{27}z^3 + \frac{4}{81}z^4, & \beta_I &= 3. \end{aligned}$$

For the general definition of the optimal polynomials we refer to the cited literature and to van der Houwen (1996), Hairer & Wanner (1996, Sect. IV.2).

For practical purposes it is interesting to compare this optimal fourth-degree polynomial $P_4(z)$ with the fourth-order Padé polynomial

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4, \quad \beta_I = 2\sqrt{2},$$

generated by the classical fourth-order explicit Runge-Kutta method (1.8.b). For this method we have $\beta_I/(s-1) \approx 0.94$, so this is quite close to the optimal value 1. The classical Runge-Kutta method thus combines high order and good stability on the imaginary axis, and it is therefore popular for solving hyperbolic problems with central differences in space.

Remark 1.7 For the advection problem with first-order upwind spatial discretization, the stability region should contain the disc $\{z \in \mathbb{C} : |z + \nu| \leq \nu\}$

with Courant number ν . By using the CFL-criterion for stability with domains of dependences, it is easily shown, see Section I.6,⁵⁾ that any explicit Runge-Kutta method will give a restriction $\nu \leq s$. In terms of the scaled values ν/s , it follows that the forward Euler method is optimal. In fact, we also saw in Section I.6 that with respect to accuracy the forward Euler method combines favourably with first-order upwind, since spatial and temporal errors cancel. \diamond

Stability Restrictions for Diffusion

Stability restrictions for the diffusion equation $u_t = du_{xx}$ with $d > 0$ are obtained in a similar way. The relevant expressions for $z = \tau\lambda$ with the standard diffusion discretizations of order two and four are

$$\text{second-order central: } z_{d,2} = \mu(e^{-2i\omega} - 2 + e^{2i\omega}),$$

$$\text{fourth-order central: } z_{d,4} = \frac{1}{12}\mu(-e^{-4i\omega} + 16e^{-2i\omega} - 30 + 16e^{2i\omega} - e^{4i\omega}),$$

where $\omega \in [-\frac{1}{2}\pi, \frac{1}{2}\pi]$ and $\mu = \tau d/h^2$. The maximal μ -values, together with the scaled quantities μ/s , are given in Table 1.3 for the Padé polynomials up to degree four. Because the eigenvalues are real negative, these maximal values are determined by the *real stability boundary* β_R of the polynomials, where, by definition, $[-\beta_R, 0]$ is the largest segment of the negative real axis contained in the stability region \mathcal{S} .

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$z_{d,2}$	0.50 (0.50)	0.50 (0.25)	0.62 (0.20)	0.69 (0.17)
$z_{d,4}$	0.37 (0.37)	0.37 (0.18)	0.47 (0.11)	0.52 (0.13)

Table 1.3. Maximal values $\mu = \tau d/h^2$ for stability, with corresponding scaled values μ/s in parentheses.

Due to the factor h^2 , a restriction on μ may lead to a very small step size τ . Therefore the conventional explicit Runge-Kutta methods are in general of little practical use for diffusion equations. Diffusion equations are often better solved with A -stable implicit methods which provide unconditional stability, or with special explicit methods which possess very long stability intervals along the negative real line. Such special methods will be discussed in Chapter V.

1.5 Rosenbrock Methods

Rosenbrock methods are Runge-Kutta type methods for stiff ODEs which are linearly implicit. They are named after Rosenbrock (1963) who proposed

⁵⁾ Note that an s -stage explicit Runge-Kutta method with first-order upwind in space for $u_t + au_x = 0$, $a > 0$, leads to a fully discrete scheme of the form $w_j^{n+1} = \sum_{k=-s}^0 \gamma_k w_{j+k}^n$.

the first methods of this kind. In literature different forms have been used. Nowadays a Rosenbrock method is understood to solve an autonomous ODE system $w'(t) = F(w(t))$ by means of the s -stage one-step formula

$$\begin{aligned} w_{n+1} &= w_n + \sum_{i=1}^s b_i k_i, \\ k_i &= \tau F(w_n + \sum_{j=1}^{i-1} \alpha_{ij} k_j) + \tau A \sum_{j=1}^i \gamma_{ij} k_j, \end{aligned} \quad (1.25)$$

where $A = A_n$ equals the Jacobian matrix $F'(w_n)$. As for the Runge-Kutta method (1.1), the formula coefficients b_i , α_{ij} and γ_{ij} are chosen to obtain a desired order of consistency and A -stability or L -stability.

At each stage these methods solve a system of linear algebraic equations with the matrix $I - \gamma_{ii}\tau A$. In this respect they bear a close relationship with diagonally implicit Runge-Kutta methods, because solving the implicit relations in these diagonally implicit methods by means of modified Newton iteration results in formulas of the above kind if a fixed number of iterations is performed. The crucial consideration put forth by Rosenbrock (1963) was to no longer use the iterative Newton method, but instead to derive stable formulas by working the Jacobian matrix directly into the integration formula. This way one gets a somewhat simpler implementation since no stopping strategy on the Newton iteration is needed.

Rosenbrock methods have proven successful for many different stiff ODE and PDE applications, especially in the low to moderate accuracy range. To economize on the linear system solution the coefficients γ_{ii} are usually taken constant, $\gamma_{ii} = \gamma$, so that per time step only linear systems with the same matrix $I - \gamma\tau A$ are to be dealt with. For large dimensions this saves computing time when using LU -decompositions or preconditioned iterative solvers. In actual implementations the matrix-vector multiplications in (1.25) are avoided by a simple transformation, see the examples below. Finally we note that if we would take in (1.25) for A the zero matrix, a standard explicit Runge-Kutta method results.

The definition of order of consistency is the same as for the Runge-Kutta methods (1.1). For a maximum of four stages and assuming $\gamma_{ii} = \gamma$, the conditions for order $p \leq 3$ are given in Table 1.4 where

$$\beta_{ij} = \alpha_{ij} + \gamma_{ij}, \quad c_i = \sum_{j=1}^{i-1} \alpha_{ij}, \quad d_i = \sum_{j=1}^{i-1} \beta_{ij}.$$

The conditions for $p \leq 5$ for an arbitrary number of stages, complementing Table 1.4, can be found in Hairer & Wanner (1996, Sect. IV.7). The derivation of these conditions by Taylor expansions is technical and requires considerable effort.

order p	order conditions
1	$b_1 + b_2 + b_3 + b_4 = 1$
2	$b_2 d_2 + b_3 d_3 + b_4 d_4 = 1/2 - \gamma$
3	$b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = 1/3$ $b_3 \beta_{32} d_2 + b_4 (\beta_{42} d_2 + \beta_{43} d_3) = 1/6 - \gamma + \gamma^2$

Table 1.4. Order conditions of Rosenbrock methods with $\gamma_{ii} = \gamma$ for $s \leq 4$, $p \leq 3$.

Similar as for diagonally implicit Runge-Kutta methods, Rosenbrock methods with $\gamma_{ii} = \gamma$ possess stability functions of the form

$$R(z) = \frac{P(z)}{(1 - \gamma z)^s},$$

where P is a polynomial of degree $\leq s$. General results on the stability and accuracy for such rational functions can be found for example in the monograph of Iserles & Nørsett (1991).

Remark 1.8 By using for non-autonomous systems $w'(t) = F(t, w(t))$ the transformation (1.2), for such systems the expression for k_i changes to

$$k_i = \tau F(t_n + c_i \tau, w_n + \sum_{j=1}^{i-1} \alpha_{ij} k_j) + \gamma_i \tau^2 F_t(t_n, w_n) + \tau A \sum_{j=1}^i \gamma_{ij} k_j,$$

where $\gamma_i = \gamma_{i1} + \dots + \gamma_{i,i-1} + \gamma_{ii}$ and $A = F_w(t_n, w_n)$. It is also possible to simply omit the partial derivative F_t , but then the order conditions change and readily become more complicated for $p \geq 3$. This is the main reason why Rosenbrock methods are mostly written in the autonomous form.

Rosenbrock methods also maintain linear conservation properties discussed in Remark 1.2. This easily follows by observing that (1.3) implies $v^T F_t(t, w) = 0$ and $v^T F_w(t, w) = 0$. \diamond

Example 1.9 The 1-stage method

$$w_{n+1} = w_n + k_1, \quad k_1 = \tau F(w_n) + \gamma \tau A k_1, \quad (1.26)$$

is of order two if $\gamma = \frac{1}{2}$. Otherwise the order is one. The stability function is

$$R(z) = \frac{1 + (1 - \gamma)z}{1 - \gamma z},$$

and hence the method is A -stable for any $\gamma \geq \frac{1}{2}$ and L -stable for $\gamma = 1$. The order remains unchanged if we use for A an $\mathcal{O}(\tau)$ approximation to the exact Jacobian matrix,

$$A = F'(w_n) + \mathcal{O}(\tau).$$

The method is also called the linearized θ -method because with $\gamma = \theta$ it results from the implicit θ -method through one Newton iteration using w_n as starting point. \diamond

Example 1.10 The 2-stage method

$$\begin{aligned} w_{n+1} &= w_n + b_1 k_1 + b_2 k_2, \\ k_1 &= \tau F(w_n) + \gamma \tau A k_1, \\ k_2 &= \tau F(w_n + \alpha_{21} k_1) + \gamma_{21} \tau A k_1 + \gamma \tau A k_2, \end{aligned} \quad (1.27)$$

with

$$b_1 = 1 - b_2, \quad \alpha_{21} = 1/(2b_2), \quad \gamma_{21} = -\gamma/b_2,$$

is of second-order for any choice of γ and $b_2 \neq 0$. With these coefficients the method is in fact of order two for any matrix A , and it retains this property for non-autonomous problems also without the F_t contribution, see Dekker & Verwer (1984, p. 233). Of course, to obtain good stability properties, A should be related to the exact Jacobian matrix.

The stability function is the rational function (1.18), just as for the two diagonally implicit Runge-Kutta methods of Example 1.5. Hence the method is A -stable for $\gamma \geq \frac{1}{4}$ and it is L -stable if $\gamma = 1 \pm \frac{1}{2}\sqrt{2}$.

Setting $b_2 = 1/2$ we get the second-order scheme

$$\begin{aligned} w_{n+1} &= w_n + \frac{1}{2} k_1 + \frac{1}{2} k_2, \\ k_1 &= \tau F(w_n) + \gamma \tau A k_1, \\ k_2 &= \tau F(w_n + k_1) - 2\gamma \tau A k_1 + \gamma \tau A k_2. \end{aligned} \quad (1.28)$$

If A is taken as the zero matrix we now regain the second-order explicit trapezoidal rule. To avoid the matrix-vector multiplication Ak_1 in the second stage, the equivalent form

$$\begin{aligned} w_{n+1} &= w_n + \frac{3}{2} \tilde{k}_1 + \frac{1}{2} \tilde{k}_2, \\ \tilde{k}_1 &= \tau F(w_n) + \gamma \tau A \tilde{k}_1, \\ \tilde{k}_2 &= \tau F(w_n + \tilde{k}_1) - 2\tilde{k}_1 + \gamma \tau A \tilde{k}_2, \end{aligned}$$

is usually implemented. This method has been used for atmospheric transport-chemistry problems in Verwer et al. (1999) with $\gamma = 1 + \frac{1}{2}\sqrt{2}$. \diamond

Example 1.11 With two stages one can also achieve order of consistency three provided A is the exact Jacobian matrix or at least satisfies $A = F'(w_n) + \mathcal{O}(\tau)$. If we allow $\mathcal{O}(\tau)$ perturbations, then the conditions for order three become

$$b_1 + b_2 = 1, \quad b_2(\alpha_{21} + \gamma_{21}) = \frac{1}{2} - \gamma, \quad b_2 \alpha_{21}^2 = \frac{1}{3}, \quad \gamma^2 - \gamma + \frac{1}{6} = 0, \quad b_2 \gamma_{21} = -\gamma.$$

The last of these is extra, see Table 1.4, and serves to allow $\mathcal{O}(\tau)$ perturbations. With $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ these conditions have a unique solution defining the third-order, strongly A -stable scheme

$$\begin{aligned} w_{n+1} &= w_n + \frac{1}{4}k_1 + \frac{3}{4}k_2, \\ k_1 &= \tau F(w_n) + \gamma\tau Ak_1, \\ k_2 &= \tau F(w_n + \frac{2}{3}k_1) - \frac{4}{3}\gamma\tau Ak_1 + \gamma\tau Ak_2. \end{aligned} \quad (1.29)$$

As in the previous example, one matrix-vector multiplication can be saved by implementing a form with $\tilde{k}_1 = k_1$ and $\tilde{k}_2 = k_2 - \frac{4}{3}k_1$. For non-autonomous problems the partial derivative F_t is needed to retain its third-order, see Remark 1.8. The methods (1.28), (1.29) will be again discussed in Chapter IV in connection with splitting methods. \diamond

2 Convergence of Runge-Kutta Methods

2.1 Order Reduction

In Section I.6.1 we introduced the Method of Lines (MOL) concept for PDEs and it was argued that, within the MOL framework, properties of the time stepping method always should be considered in conjunction with the spatial discretization and mesh width h . In this section we will take a closer look at consistency and convergence properties of Runge-Kutta methods within the MOL framework. In particular our attention will be directed to the awkward phenomenon of *order reduction* which shows up in high accuracy calculations for initial-boundary value problems.

Numerical Illustration

We begin with an example to illustrate this order reduction phenomenon numerically. Consider the advection problem with source term

$$u_t + u_x = u^2, \quad 0 < t \leq \frac{1}{2}, \quad 0 < x < 1, \quad (2.1)$$

with solution $u(x, t) = \sin^2(\pi(x - t)) / (1 - t \sin^2(\pi(x - t)))$. We solve this problem with given initial function $u(x, 0)$, and with boundary condition either the inflow Dirichlet condition

$$u(0, t) = \sin^2(\pi t) / (1 - t \sin^2(\pi t)), \quad (2.2)$$

or the periodicity condition

$$u(x \pm 1, t) = u(x, t). \quad (2.3)$$

	Dirichlet condition (2.2)				Periodicity condition (2.3)			
h	L_2 -error	order	L_∞ -error	order	L_2 -error	order	L_∞ -error	order
$\frac{1}{40}$	$0.18 \cdot 10^{-3}$		$0.30 \cdot 10^{-3}$		$0.17 \cdot 10^{-3}$		$0.21 \cdot 10^{-3}$	
$\frac{1}{80}$	$0.13 \cdot 10^{-4}$	3.80	$0.25 \cdot 10^{-4}$	3.57	$0.11 \cdot 10^{-4}$	3.98	$0.14 \cdot 10^{-4}$	3.98
$\frac{1}{160}$	$0.86 \cdot 10^{-6}$	3.90	$0.19 \cdot 10^{-5}$	3.75	$0.67 \cdot 10^{-6}$	3.99	$0.85 \cdot 10^{-6}$	3.99
$\frac{1}{320}$	$0.56 \cdot 10^{-7}$	3.96	$0.13 \cdot 10^{-6}$	3.91	$0.42 \cdot 10^{-7}$	4.00	$0.53 \cdot 10^{-7}$	4.00
$\frac{1}{640}$	$0.35 \cdot 10^{-8}$	3.98	$0.79 \cdot 10^{-8}$	3.96	$0.26 \cdot 10^{-8}$	4.00	$0.33 \cdot 10^{-8}$	4.00

Table 2.1. Spatial errors for the test example (2.1) - (2.3).

For the spatial discretization a uniform grid $\{x_j = jh\}$ is used with mesh width h and with fourth-order central differences for u_x . With the periodic boundary condition these can be used in all grid points. For the Dirichlet condition we need an adjustment at grid points $x = h, 1 - h, 1$, and at these points we use third-order differences with 4-point stencil. In view of the results presented in Section I.5, we then still expect a fourth-order spatial error provided this discretization is stable.

Table 2.1 gives for a sequence of decreasing mesh widths relative spatial discretization errors at time $t = 1/2$ in the L_2 -norm and L_∞ -norm together with estimated orders (defined as the \log_2 of the ratio of consecutive errors). For both boundary conditions the expected fourth-order convergence indeed shows up. These results were found numerically using the classical fourth-order explicit Runge-Kutta method with very small step size τ , to render temporal errors negligible.

	Dirichlet condition (2.2)				Periodicity condition (2.3)			
τ	L_2 -error	order	L_∞ -error	order	L_2 -error	order	L_∞ -error	order
$\frac{1}{20}$	$0.76 \cdot 10^{-3}$		$0.13 \cdot 10^{-2}$		$0.75 \cdot 10^{-3}$		$0.11 \cdot 10^{-2}$	
$\frac{1}{40}$	$0.68 \cdot 10^{-4}$	3.48	$0.16 \cdot 10^{-3}$	2.96	$0.56 \cdot 10^{-4}$	3.76	$0.87 \cdot 10^{-4}$	3.72
$\frac{1}{80}$	$0.95 \cdot 10^{-5}$	2.84	$0.46 \cdot 10^{-4}$	1.83	$0.37 \cdot 10^{-5}$	3.90	$0.59 \cdot 10^{-5}$	3.88
$\frac{1}{160}$	$0.17 \cdot 10^{-5}$	2.52	$0.12 \cdot 10^{-4}$	1.98	$0.24 \cdot 10^{-6}$	3.95	$0.38 \cdot 10^{-6}$	3.95
$\frac{1}{320}$	$0.30 \cdot 10^{-6}$	2.48	$0.29 \cdot 10^{-5}$	1.99	$0.15 \cdot 10^{-7}$	3.98	$0.24 \cdot 10^{-7}$	3.97

Table 2.2. Full space-time errors, with $\tau = 2h$, for the test example (2.1) - (2.3).

We next take a look at the full relative discretization errors and estimated orders given in Table 2.2. These results were obtained by applying the classical fourth-order Runge-Kutta method with a realistic step size $\tau = 2h$. For the periodic boundary condition we again observe order four, approximately. However, quite surprisingly, for the Dirichlet condition there is a clear *order reduction*. Instead of order 4, we get approximately order 2.5 in the L_2 -norm and order 2 in the L_∞ -norm. Since the spatial error behaves as expected, this observed reduction for the full error must emanate from its temporal part.

Preliminaries for the Error Analysis

Below we will discuss and explain the cause for the order reduction. This will be done in a general framework for explicit and implicit Runge-Kutta methods, based on the derivations of Brenner, Crouzeix & Thomée (1982). For ease of presentation we hereby restrict ourselves to linear problems. The linear case gives the insight needed to understand the phenomenon, while for nonlinear problems the explanations readily become highly technical.

Consider a linear semi-discrete system in \mathbb{R}^m ,

$$w'(t) = Aw(t) + g(t), \quad t > 0, \quad w(0) = w_0. \quad (2.4)$$

If the underlying PDE problem has non-homogeneous boundary conditions, such as non-zero Dirichlet conditions, these are incorporated in $g(t)$ together with genuine source terms. Consequently, as for the matrix A , the time-dependent term $g(t)$ will often contain negative powers of the mesh width h .

The exact PDE solution restricted to the space grid is denoted by u_h . It is assumed throughout that u is sufficiently smooth, i.e., all spatial and temporal derivatives encountered in the analysis exist and are of moderate size. The task we have set ourselves is to examine convergence of full Runge-Kutta approximations w_n to $u_h(t_n)$ under simultaneous refinement in space $h \rightarrow 0$ and time $\tau \rightarrow 0$, the latter possibly obeying a certain dependence $\tau = \tau(h)$, as, for example, imposed by a CFL condition.

Throughout this section a given norm $\|\cdot\|$ on \mathbb{R}^m is considered and the following notation will be used: for $v \in \mathbb{R}^m$ depending on τ and h we write $v = \mathcal{O}(\tau^\alpha h^\beta)$ if $\|v\| \leq C\tau^\alpha h^\beta$ with $C > 0$ independent of τ and h . In particular, $v = \mathcal{O}(\tau^\alpha)$ means that no negative powers of h are hidden in the bound. The same notation is used for matrices. If it is necessary to specify the norm we write $\|v\| = \mathcal{O}(\tau^\alpha h^\beta)$.

The global space-time discretization error $\varepsilon_n = u_h(t_n) - w_n$ can be interpreted as

$$\varepsilon_n = \underbrace{u_h(t_n) - w(t_n)}_{\text{spatial error}} + \underbrace{w(t_n) - w_n}_{\text{temporal error}}. \quad (2.5)$$

As shown in Section I.4, the spatial error is to a great extent determined by the spatial truncation error

$$\sigma_h(t) = u'_h(t) - Au_h(t) - g(t). \quad (2.6)$$

To understand the order reduction phenomenon, our attention will be mainly directed to the temporal error behaviour for $\tau, h \rightarrow 0$.

2.2 Local Error Analysis

Application of the Runge-Kutta method (1.1) to (2.4) gives a recursion

$$w_{n+1} = R(\tau A) w_n + \sum_{j=1}^s Q_j(\tau A) \tau g(t_n + c_j \tau), \quad (2.7)$$

with stability function R and polynomial or rational functions Q_j determined by the method, see (1.14), (1.21) and (1.22). Inserting the exact PDE solution $u_h(t)$ into (2.7) gives

$$u_h(t_{n+1}) = R(\tau A) u_h(t_n) + \sum_{j=1}^s Q_j(\tau A) \tau g(t_n + c_j \tau) + \delta_n, \quad (2.8)$$

with δ_n the local (space-time) discretization error made in one single step from $u_h(t_n)$, that is, if we put $w_n = u_h(t_n)$ then $\delta_n = u_h(t_{n+1}) - w_{n+1}$. Using the local space error expression (2.6), we can eliminate the source terms $g(t_n + c_j \tau)$ to obtain

$$\begin{aligned} \delta_n &= u_h(t_{n+1}) - R(\tau A) u_h(t_n) \\ &- \sum_{j=1}^s Q_j(\tau A) \tau [u'_h(t_n + c_j \tau) - A u_h(t_n + c_j \tau) - \sigma_h(t_n + c_j \tau)]. \end{aligned} \quad (2.9)$$

We will now first focus on the temporal part and for simplicity we therefore assume $\sigma_h \equiv 0$; the spatial error contribution will be considered in Section 2.4.

The first step is to express δ_n as a Taylor series in terms of the exact solution u_h and its derivatives,

$$\delta_n = \sum_{k \geq 0} \frac{1}{k!} H_k(\tau A) \tau^k u_h^{(k)}(t_n), \quad (2.10)$$

with the functions H_k defined by

$$H_0(z) = 1 - R(z) + z \sum_{j=1}^s Q_j(z),$$

$$H_k(z) = 1 + \sum_{j=1}^s (z c_j^k - k c_j^{k-1}) Q_j(z), \quad k \geq 1.$$

The Taylor expansion can of course be truncated at any level τ^k with a remainder term proportional to τ^{k+1} , involving derivatives $u_{h,i}^{(k+1)}$ of the components $u_{h,i}$ ($i = 1, \dots, m$) at intermediate points in $[t_n, t_{n+1}]$.

Let p be the order of the Runge-Kutta method according to the standard consistency concept for ODEs. For any fixed matrix A , i.e., for h fixed, we then have $\delta_n = \mathcal{O}(\tau^{p+1})$. Consequently, for $z \rightarrow 0$ the functions H_k satisfy

$$H_k(z) = \mathcal{O}(z^{p+1-k}), \quad 0 \leq k \leq p. \quad (2.11)$$

On the other hand, if $h \rightarrow 0$, the norm of the finite difference matrix A grows with its negative power of h rendering the asymptotics for $z \rightarrow 0$ not applicable. Hence we need a different asymptotics. For that purpose let $q \in \mathbb{N}$ be such that

$$H_k(z) = 0 \quad \text{for } k = 0, 1, \dots, q. \quad (2.12)$$

This means that the Runge-Kutta method is exact if $u_h(t)$ is a polynomial of degree q or less. The integer q is in fact the stage order mentioned in Section 1.1, see also Dekker & Verwer (1984) or Hairer & Wanner (1996). Because all Runge-Kutta methods have at least a stage order of one, it always follows that H_0 and H_1 are identically equal to zero.

Stability usually implies $H_k(\tau A)$ to be bounded, but this only gives the estimate $\delta_n = \mathcal{O}(\tau^{q+1})$. If $q < p$, as it often is, this means a reduction of the order of the local error. We will now further elaborate this estimate through two examples related to the numerical illustration given for problem (2.1).

Example 2.1 The classical explicit fourth-order method has $p = 4$, $q = 1$ and the polynomials Q_j are given by

$$\begin{aligned} Q_1(z) &= \frac{1}{6} + \frac{1}{6}z + \frac{1}{12}z^2 + \frac{1}{24}z^3, \\ Q_2(z) &= \frac{1}{3} + \frac{1}{6}z + \frac{1}{12}z^2, \quad Q_3(z) = \frac{1}{3} + \frac{1}{6}z, \quad Q_4(z) = \frac{1}{6}. \end{aligned} \quad (2.13)$$

It follows that

$$H_2(z) = \frac{1}{48}z^3, \quad H_3(z) = \frac{1}{96}z^3 - \frac{1}{48}z^2, \quad H_4(z) = \frac{1}{192}z^3 - \frac{1}{48}z^2 + \frac{1}{24}z,$$

leading to the local error expression

$$\begin{aligned} \delta_n &= \frac{1}{96}\tau^2 [\tau^3 A^3] u_h^{(2)}(t_n) + \frac{1}{576}\tau^3 [\tau^3 A^3 - 2\tau^2 A^2] u_h^{(3)}(t_n) \\ &\quad + \frac{1}{4608}\tau^4 [\tau^3 A^3 - 4\tau^2 A^2 + 8\tau A] u_h^{(4)}(t_n) + \mathcal{O}(\tau^5). \end{aligned} \quad (2.14)$$

Note that the error constants are quite small, which is common for a higher-order method.

Stability for this explicit method requires $\tau A = \mathcal{O}(1)$. Instead of the classical estimate $\mathcal{O}(\tau^5)$ for δ_n this gives only an estimate $\mathcal{O}(\tau^2)$. However, if $Au_h^{(2)}(t_n) = \mathcal{O}(1)$ then we can write the leading term as

$$\frac{1}{96}\tau^3 (\tau A)^2 [Au_h^{(2)}(t_n)] = \mathcal{O}(\tau^3), \quad (2.15)$$

giving $\delta_n = \mathcal{O}(\tau^3)$. More general we have

$$\begin{aligned}\delta_n &= \mathcal{O}(\tau^2) && \text{if } \tau A = \mathcal{O}(1), \\ \delta_n &= \mathcal{O}(\tau^3) && \text{if } Au_h^{(2)} = \mathcal{O}(1), \\ \delta_n &= \mathcal{O}(\tau^4) && \text{if } A^2u_h^{(2)} = \mathcal{O}(1), Au_h^{(3)} = \mathcal{O}(1), \\ \delta_n &= \mathcal{O}(\tau^5) && \text{if } A^3u_h^{(2)} = \mathcal{O}(1), A^2u_h^{(3)} = \mathcal{O}(1), Au_h^{(4)} = \mathcal{O}(1),\end{aligned}$$

where the conditions are of course accumulative.

Apparently, if we let simultaneously $\tau, h \rightarrow 0$ such that $\tau A = \mathcal{O}(1)$, the classical $\mathcal{O}(\tau^5)$ estimate only holds if a substantial number of additional conditions is satisfied. The next example serves to illustrate that these conditions emanate from the *boundaries*. They are always satisfied in case of spatial periodicity but not with genuine boundary conditions. \diamond

Example 2.2 Consider the familiar example, arising from first-order upwind advection discretization with an inflow Dirichlet condition,

$$A = \frac{1}{h} \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

and consider a vector $v = (v_j) \in \mathbb{R}^m$ with $v_j = \psi(x_j), x_j = jh$, for some fixed, smooth function ψ , for instance $\psi = u_{tt}$. Then

$$Av = -\frac{1}{h}\psi(0) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \psi_x(x_1) \\ \psi_x(x_2) \\ \vdots \\ \psi_x(x_m) \end{pmatrix} + \dots,$$

and therefore $Av = \mathcal{O}(1)$ in the L_2 - or L_∞ -norm iff $\psi(0) = 0$. Otherwise we will have $\|Av\|_2 \sim h^{-1/2}$ and $\|Av\|_\infty \sim h^{-1}$. In case that $\psi(0) = 0$, we have

$$A^2v = -\frac{1}{h}\psi_x(0) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \psi_{xx}(x_1) \\ \psi_{xx}(x_2) \\ \vdots \\ \psi_{xx}(x_m) \end{pmatrix} + \dots,$$

and thus we see that for having $A^2v = \mathcal{O}(1)$ in the L_2 - or L_∞ -norm we need $\psi(0) = \psi_x(0) = 0$. Likewise we can proceed for higher powers of A : if $\psi^{(j)}(0) = 0$ for $j < k$ and $\psi^{(k)}(0) \neq 0$ then we have

$$Av = \mathcal{O}(1), \quad A^2v = \mathcal{O}(1), \dots, \quad A^kv = \mathcal{O}(1),$$

whereas

$$\|A^{k+1}v\|_2 = \mathcal{O}(h^{-1/2}), \quad \|A^{k+1}v\|_\infty = \mathcal{O}(h^{-1}).$$

With a spatial periodicity condition the situation is different. If we consider

$$A = \frac{1}{h} \begin{pmatrix} -1 & & & 1 \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{m \times m}$$

and ψ a smooth periodic function, we get simply

$$Av = (\psi_x(x_j)) + \mathcal{O}(h), \quad A^2v = (\psi_{xx}(x_j)) + \mathcal{O}(h), \quad \dots$$

and there will be no order reduction.

Obviously, also for problems where all spatial derivatives vanish at the boundary there will be no order reduction. This holds for homogeneous problems (no source term) with homogeneous Dirichlet boundary conditions.

Higher-order discretizations of advection and diffusion, with Dirichlet or Neumann boundary conditions, or with periodicity conditions, can be considered in a similar way. \diamond

In view of the above, the result of Table 2.2 for the periodic case (2.3) does not come as a surprise. With periodicity conditions we get local errors of $\mathcal{O}(\tau^{p+1})$ and thus global errors of $\mathcal{O}(\tau^p)$.

Also the fact that with Dirichlet conditions a lower order of convergence was observed is no longer surprising. Note however that the results for the Dirichlet case (2.2) are still not well explained. Since (2.2) gives an inhomogeneous Dirichlet condition, the above suggests that we will have $\|\delta_n\|_\infty = \mathcal{O}(\tau^2)$ and $\|\delta_n\|_2 = \mathcal{O}(\tau^{2.5})$ for the local errors, and then we expect one power less for the global errors. However, the global errors given in Table 2.2 converge with the same power of τ . To explain this we next take a closer look at the propagation of local errors.

2.3 Global Error Analysis

Having studied the local error in some detail we now examine the temporal part of the global error (2.5), still assuming a zero space error. Subtracting (2.7) from (2.8) gives the global error recurrence

$$\varepsilon_{n+1} = R(\tau A)\varepsilon_n + \delta_n, \quad (2.16)$$

with δ_n given by (2.10). For a given time interval $[0, T]$, we make the common stability hypothesis

$$\|R(\tau A)^n\| \leq K \quad \text{for } h \rightarrow 0 \text{ and } n \geq 0, n\tau \leq T. \quad (2.17)$$

With stability at hand, elaboration of recursion (2.16) in the usual manner gives convergence for the global time error with one power of τ less than for the local error. As just mentioned, one power less is in contradiction with observed numerical results. The following analysis explains this.

Consider first the general recursion

$$\varepsilon_{n+1} = S\varepsilon_n + \delta_n \quad (n = 0, \dots, N-1), \quad \varepsilon_0 = 0,$$

with stability assumption $\|S^n\| \leq K$ for all $n = 1, \dots, N$.

Lemma 2.3 Suppose δ_n can be written as

$$\delta_n = (I - S)\xi_n + \eta_n,$$

with $\|\xi_n\| \leq C\tau^r$, $\|\eta_n\| \leq C\tau^{r+1}$ and $\|\xi_{n+1} - \xi_n\| \leq C\tau^{r+1}$ for all n . Then there is a $C' > 0$, depending on C, K and $T = N\tau$, such that $\|\varepsilon_n\| \leq C'\tau^r$ for $n\tau \leq T$.

Proof. We have

$$\varepsilon_{n+1} = S\varepsilon_n + (I - S)\xi_n + \eta_n, \quad \varepsilon_0 = 0.$$

Introducing $\hat{\varepsilon}_n = \varepsilon_n - \xi_n$ we can write

$$\hat{\varepsilon}_{n+1} = S\hat{\varepsilon}_n + \eta_n - (\xi_{n+1} - \xi_n), \quad \hat{\varepsilon}_0 = -\xi_0.$$

This gives in the standard way $\hat{\varepsilon}_n = \mathcal{O}(\tau^r)$ and thus also $\varepsilon_n = \mathcal{O}(\tau^r)$ with constants determined by C, K and T . \square

We note that the decomposition of the local error δ_n used in this lemma can also be shown to be necessary for having $\varepsilon_n = \mathcal{O}(\tau^r)$ in case the δ_n are constant, see Hundsdorfer (1992). In this sense, the result is sharp.

The above lemma will be applied to the global error recursion (2.16). To understand the behaviour of the global error it then suffices to consider only the leading error term in δ_n . The contribution of the other terms to the global error is found in a similar way. So we consider

$$\delta_n = \frac{1}{(q+1)!} H_{q+1}(\tau A) \tau^{q+1} u_h^{(q+1)}(t_n) \quad (2.18)$$

and define for given $\alpha \geq 0$ the function

$$\varphi_\alpha(z) = (1 - R(z))^{-1} H_{q+1}(z) z^{-\alpha}.$$

Theorem 2.4 Consider for $\tau, h \rightarrow 0$ the recursion (2.16) with local error (2.18) and $\varepsilon_0 = 0$. Let the stability condition (2.17) be satisfied. If

$$\varphi_\alpha(\tau A) = \mathcal{O}(1) \quad \text{and} \quad A^\alpha u_h^{(q+j)}(t) = \mathcal{O}(1), \quad j = 1, 2, \quad (2.19)$$

uniformly in $t \in [0, T]$, the global error satisfies $\varepsilon_n = \mathcal{O}(\tau^{q+1+\alpha})$ for $n\tau \leq T$.

Proof. The result is an immediate consequence of Lemma 2.3. Take $\eta_n = 0$ and

$$\xi_n = \frac{1}{(q+1)!} \tau^{q+1+\alpha} \varphi_\alpha(\tau A) \left(A^\alpha u_h^{(q+1)}(t_n) \right).$$

Then $\delta_n = (I - R(\tau A))\xi_n$ and the proof follows. \square

Thus with this theorem we have the estimate $\mathcal{O}(\tau^{q+1+\alpha})$ rather than $\mathcal{O}(\tau^q)$ found with the standard error analysis. To study the assumptions of the theorem, let \mathcal{S} be the stability region of the Runge-Kutta method and let $\mathcal{D} \subset \mathcal{S}$. We consider the L_2 -norm. Assume that A is diagonalizable and write $A = V\Lambda V^{-1}$ with $\text{cond}(V) = K = \mathcal{O}(1)$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ such that

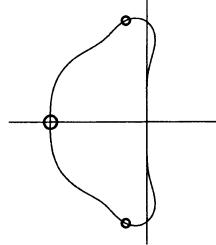
$$\tau \lambda_j \in \mathcal{D} \subset \mathcal{S}. \quad (2.20)$$

Then we have stability, $\|R(\tau A)^n\| \leq K$ for all n . If we assume in addition that

$$|\varphi_\alpha(z)| \leq C \quad \text{for all } z \in \mathcal{D}, \quad (2.21)$$

then $\|\varphi_\alpha(\tau A)\| = \mathcal{O}(1)$.

To apply this result we have to chose some suitable region $\mathcal{D} \subset \mathcal{S}$. We want the point 0 to be on its boundary, so that the result can be applied for a step size interval $(0, \tau_0]$. For this we need boundedness of $\varphi_\alpha(z)$ near $z = 0$. This holds for $\alpha \leq p - q - 1$, due to (2.11). Further we can take \mathcal{D} arbitrary in \mathcal{S} , except for points $z \neq 0$ on the boundary of \mathcal{S} where $R(z) = 1$.



Example 2.5 For the classical Runge-Kutta method we have $q = 1$ and

$$\varphi_\alpha(z) = -\frac{1}{48} \frac{z^{2-\alpha}}{1 + \frac{1}{2}z + \frac{1}{6}z^2 + \frac{1}{24}z^3}$$

which is bounded near 0 if $\alpha \leq 2$. The order 2.5 result in Table 2.2 for the L_2 -norm follows if we can show that $A^\alpha u_h^{(2)}(t) = \mathcal{O}(1)$ and $A^\alpha u_h^{(3)}(t) = \mathcal{O}(1)$ for α up to 0.5. Although probably true, this seems difficult to prove. An easier alternative is to take $\alpha = 1$ and to write the local error as

$$\delta_n = (I - R(\tau A))\xi_n, \quad \xi_n = \frac{1}{2}\tau^{2.5} \varphi_1(\tau A) [\tau^{0.5} Au_h^{(2)}].$$

Then, with τ/h constant and $\tau A = \mathcal{O}(1)$ we have $\|\tau^{0.5} Au_h^{(k)}\|_2 = \mathcal{O}(1)$, $k = 2, 3$, which leads to the observed global error estimate. \diamond

Example 2.6 The second-order implicit midpoint rule

$$w_{n+1} = w_n + \tau F(t_n + \frac{1}{2}\tau, \frac{1}{2}w_n + \frac{1}{2}w_{n+1})$$

gives the form (2.7) with $s = 1$, $c_1 = \frac{1}{2}$ and $Q_1(z) = (1 - \frac{1}{2}z)^{-1}$. We have $p = 2$, $q = 1$ and $H_2(z) = -\frac{1}{4}z(1 - \frac{1}{2}z)^{-1}$. Consequently, unless $Au_h^{(2)}(t) = \mathcal{O}(1)$, the local error

$$\delta_n = -\frac{1}{8}(I - \frac{1}{2}\tau A)^{-1}\tau A [\tau^2 u_h^{(2)}(t_n)] + \mathcal{O}(\tau^3)$$

suffers from order reduction with standard estimate $\delta_n = \mathcal{O}(\tau^2)$. This result applies for example to parabolic problems.⁶⁾ However, by noting that

$$\delta_n = -\frac{1}{8}(I - R(\tau A)) \tau^2 u_h^{(2)}(t_n) + \mathcal{O}(\tau^3),$$

direct application of Lemma 2.3 reveals that the global error will show nicely the common $\mathcal{O}(\tau^2)$ behaviour in any norm for which we have stability, even if $Au_h^{(2)}(t_n) \neq \mathcal{O}(1)$. \diamond

Example 2.7 For the second-order trapezoidal rule

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_{n+1})$$

we get $s = 2$, $c_1 = 0$, $c_2 = 1$ and $Q_1(z) = Q_2(z) = \frac{1}{2}(1 - \frac{1}{2}z)^{-1}$, $H_2(z) = 0$. Hence we have here $p = q = 2$ and thus no order reduction will take place. \diamond

Numerical Illustration

For a numerical example, with results mostly taken from Verwer (1986), we consider Burgers' equation

$$u_t + uu_x = du_{xx}, \quad 0 < t \leq T = 1, \quad 0 < x < 1, \quad (2.22)$$

which is the standard 1D model for nonlinear advection coupled to linear diffusion. For $0 < d \ll 1$ solutions have steep gradients. To get a test problem with smooth solutions we use the relatively large value $d = 0.1$, and the initial function $u(x, 0)$ and Dirichlet boundary values are taken to match the exact solution given by Whitham (1974, Ch. 4),

$$u(x, t) = 1 - 0.9 \frac{r_1}{r_1 + r_2 + r_3} - 0.5 \frac{r_2}{r_1 + r_2 + r_3}, \quad (2.23)$$

where

$$r_1 = e^{\frac{-1}{400d}(20(x-\frac{1}{2})+99t)}, \quad r_2 = e^{\frac{-1}{16d}(4(x-\frac{1}{2})+3t)}, \quad r_3 = e^{\frac{-1}{2d}(x-\frac{3}{8})}.$$

The solution $u(x, t)$ is shown in Figure 2.1 for $t = 0, 0.5, 1.0$. In space a uniform grid $\{x_j = jh\}_{j=1}^m$ is used, $h = 1/(m+1)$, and the advective form (2.22)

⁶⁾ For the model problem $u_t = u_{xx} + f(t)$, discretized in space with second-order differences and subjected to time-dependent Dirichlet conditions, a more refined local error can be obtained by using $\|A^\gamma u_h^{(k)}\|_2 = \mathcal{O}(1)$ for $\gamma < 1/4$, see Lemma 6.4 in Chapter III, which leads in fact to $\|\delta_n\|_2 = \mathcal{O}(\tau^{2.25})$.

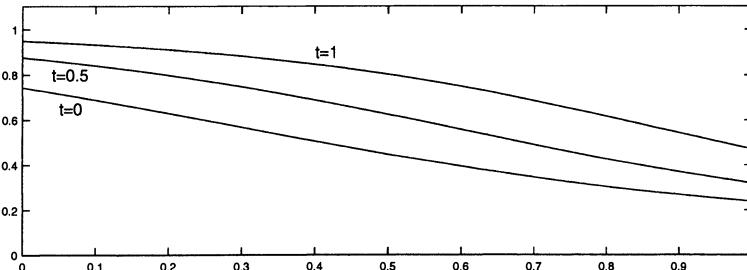


Fig. 2.1. Exact solution $u(x, t)$ of the 1D Burgers equation for $t = 0, 0.5, 1.0$.

is discretized with fourth-order central differences at the interior points using a 5-point stencil. At the grid points x_1 and x_m , adjacent to the boundaries, we use a 4-point stencil with third-order formulas for u_x and second-order central differences for u_{xx} . The global accuracy of this spatial discretization still has order four. Although the problem is nonlinear, this is in line with the results of Section I.5.

For time integration we apply the implicit midpoint rule, which has order $p = 2$ and stage order $q = 1$. Further we use the diagonally implicit methods (1.13.a) ($p = 3, q = 1$) and (1.13.b) ($p = 3, q = 2$) with $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. As an example of a higher-order diagonally implicit method we also consider the A -stable, 3-stage method given by the array

$$\begin{array}{c|ccc} \gamma & \gamma & & \\ \frac{1}{2} & \frac{1}{2}-\gamma & \gamma & \gamma = \frac{1}{2} + \frac{1}{3}\sqrt{3} \cos\left(\frac{1}{18}\pi\right), \\ 1-\gamma & 2\gamma & 1-4\gamma & \delta = \frac{1}{6}(2\gamma-1)^{-2}, \\ \hline & \delta & 1-2\delta & \delta \end{array} \quad (2.24)$$

which has $p = 4, q = 1$. Method (2.24) is due to Nørsett (1974) and Crouzeix (1975).

Table 2.3 gives full absolute L_2 -errors at time $T = 1$ for a sequence of step sizes τ_B . The spatial mesh is refined simultaneously, $h = \tau_B$. To enable a more or less fair comparison of accuracy and costs, we have put $\tau = \tau_B$ for the implicit midpoint rule, $\tau = 2\tau_B$ for the methods (1.13.a), (1.13.b) and $\tau = 3\tau_B$ for the 3-stage method (2.24). Hence we suppose that the amount of work is proportional to the number of implicit relations to be solved per step. As in Table 2.2, we also list the observed orders.

Due to the smoothness of the solution we readily get small errors. The order behaviour of the implicit midpoint rule, with the classical order $p = 2$, is in accordance with the cancellation of local order reductions, see Example 2.6. The same holds for method (1.13.b), for which we observe a third-order convergence. On the other hand, the other diagonally implicit methods apparently converge with $\mathcal{O}(\tau^{2.25})$ and hence show reduction from level p to level $q + 1$ as predicted by the above theory, with an extra factor 0.25 as in the footnote in Example 2.6.

	Impl. Mpt. Rule		Method (1.13.a)		Method (1.13.b)		Method (2.24)	
τ_B	L_2 -error	order						
$\frac{1}{24}$	$4.6 \cdot 10^{-5}$		$5.6 \cdot 10^{-5}$		$1.8 \cdot 10^{-5}$		$8.8 \cdot 10^{-5}$	
$\frac{1}{48}$	$1.2 \cdot 10^{-5}$	2.00	$9.8 \cdot 10^{-6}$	2.52	$2.4 \cdot 10^{-6}$	2.91	$1.4 \cdot 10^{-5}$	2.63
$\frac{1}{96}$	$2.9 \cdot 10^{-6}$	2.00	$1.8 \cdot 10^{-6}$	2.43	$3.1 \cdot 10^{-7}$	2.96	$2.8 \cdot 10^{-6}$	2.35
$\frac{1}{192}$	$7.3 \cdot 10^{-7}$	2.00	$3.6 \cdot 10^{-7}$	2.34	$3.9 \cdot 10^{-8}$	2.98	$5.9 \cdot 10^{-7}$	2.25
$\frac{1}{384}$	$1.8 \cdot 10^{-7}$	2.00	$7.4 \cdot 10^{-8}$	2.28	$4.9 \cdot 10^{-9}$	2.99	$1.3 \cdot 10^{-7}$	2.23

Table 2.3. Accuracy test with the Burgers equation for the implicit midpoint rule ($p = 2, q = 1$), method (1.13.a) ($p = 3, q = 1$), method (1.13.b) ($p = 3, q = 2$) and method (2.24) ($p = 4, q = 1$).

It is obvious in this numerical example that the higher stage order $q = 2$ of method (1.13.b) leads to more accuracy than the higher classical order $p = 4$ of method (2.24). It should be noted however that in this example the solution is very smooth, and for this reason the terms emanating from the boundary conditions are dominant in the total error. For problems where the interior solution is not so smooth the classical order also will have importance: in general an error behaviour of the form $C_\Gamma \tau^{q+1} + C_\Omega \tau^p$ can be expected where usually the error constant C_Γ , corresponding to the boundary conditions, will be smaller than the error constant C_Ω determined by the interior solution. Consequently, for large step sizes the term $C_\Omega \tau^p$ may still be dominant but for high accuracy calculations the extra term from the boundaries will be felt if $q + 1 < p$.

2.4 Concluding Notes

The Total Space-Time Error

For simplicity of presentation, we have neglected the spatial part

$$\delta_{h,n} = \tau \sum_{j=1}^s Q_j(\tau A) \sigma_h(t_n + c_j \tau)$$

of the local space-time error (2.9) in the preceding analysis. The contribution of $\delta_{h,n}$ to the global space-time error $\varepsilon_n = u_h(t_n) - w_n$ through recurrence (2.16) can be studied in the same way as the temporal error.

In the stable transition from local to global errors, based on assumption (2.17), the spatial contributions build up such that at the end time T all previous local space errors $\sigma_h(t_n + c_j \tau)$ are present in the global space-time error.

Stability implies in general boundedness of $Q_j(\tau A)$ and hence for $\tau, h \rightarrow 0$ the order of convergence of the global error is determined by the consistency order of the truncation errors $\|\sigma_h\|$.

A more refined spatial error analysis akin to Theorem I.5.2 is again applicable without smoothness assumptions at the semi-discrete level. Suppose

$$\sigma_h(t) = A\xi_h(t) + \eta_h(t) \quad \text{with} \quad \xi_h(t), \xi'_h(t), \eta_h(t) = \mathcal{O}(h^\beta),$$

where β may be larger than the order of the truncation error $\|\sigma_h\|$ due to boundary conditions, see Section I.5 for some examples. Then

$$\delta_{h,n} = \sum_{j=1}^s \tau A Q_j(\tau A) \xi_h(t_n) + \mathcal{O}(\tau h^\beta)$$

assuming boundedness of the rational expressions $Q_j(\tau A)$ and $\tau A Q_j(\tau A)$. Comparing the rational functions Q_j in (1.21) with R in (1.14), we see that

$$\sum_{j=1}^s z Q_j(z) = R(z) - 1,$$

and consequently we have

$$\delta_{h,n} = (R(\tau A) - I) \xi_h(t_n) + \mathcal{O}(\tau h^\beta).$$

Application of Lemma 2.3 with $r = 1$ shows that the local spatial errors $\delta_{h,n}$, $n = 0, \dots, N-1$ will give an $\mathcal{O}(h^\beta)$ contribution to the full global error $u_h(t_N) - w_N$.

Avoiding Order Reduction

Understanding the mechanism of order reduction leads to manners by which it can be diminished or completely avoided. Here we briefly illustrate this for explicit Runge-Kutta methods and the linear model problem

$$u_t + u_x = f(x, t), \quad 0 < t \leq T, \quad 0 < x < 1, \quad (2.25)$$

with given initial function $u(x, 0)$ and the Dirichlet condition $u(0, t) = \gamma_0(t)$.

A detailed study of order reduction for explicit methods and hyperbolic equations was given by Sanz-Serna, Verwer & Hundsdorfer (1987), see also Sanz-Serna & Verwer (1989). In the first paper an example was given showing how order reduction can be diminished by a transformation of the problem. A problem with inhomogeneous, time-dependent Dirichlet boundary conditions can often be transformed to a problem with homogeneous Dirichlet conditions and this will increase the temporal PDE order by one, as indicated in Example 2.2.

To illustrate this increase in order for the test problem (2.1)-(2.2), we consider $v(x, t) = u(x, t) - \gamma_0(t)$, $\gamma_0(t) = u(0, t)$, which will give a homogeneous inflow condition. The resulting problem

$$v_t + v_x = (v + \gamma_0(t))^2 - \gamma_0'(t), \quad v(0, t) = 0 \quad (2.26)$$

was solved with the same spatial discretization and step sizes as used in Table 2.2 (case Dirichlet condition). The results are presented in Table 2.4. Apparently, in the L_2 -norm the remaining order reduction manifests itself only on very fine grids. By a further transformation of the problem – subtracting a polynomial in x with suitable time-dependent coefficients – the order reduction could be completely avoided here, also in the max-norm; see Sanz-Serna et al. (1987).

A different technique was proposed by Carpenter, Gottlieb, Abarbanel & Don (1995). This consists of differentiating the boundary function $\gamma_0(t)$, which is then integrated along with the Runge-Kutta method. Let us consider the model problem (2.25). Instead of inserting $\gamma_0(t)$ directly in the semi-discrete system (the conventional approach), we can perform the time integration on the semi-discrete system augmented by the scalar problem

$$w'_0(t) = \gamma_0'(t), \quad 0 < t \leq T, \quad w(0) = \gamma_0(0). \quad (2.27)$$

This amounts to solving the semi-discrete vector $\bar{w} = (w_0, w_1, \dots, w_m)^T$ from

$$\bar{w}'(t) = \bar{A} \bar{w}(t) + \bar{g}(t), \quad (2.28)$$

where the finite difference matrix \bar{A} is the matrix of order $\bar{m} = m + 1$ whose first row is zero and whose remaining rows are the common ones associated with the grid points x_1, \dots, x_m . For example, with first-order spatial upwind discretization, we would have

	Transformation (2.26)				Differentiation (2.27)			
τ	L_2 -error	order	L_∞ -error	order	L_2 -error	order	L_∞ -error	order
$\frac{1}{20}$	$0.56 \cdot 10^{-3}$		$0.13 \cdot 10^{-2}$		$0.71 \cdot 10^{-3}$		$0.13 \cdot 10^{-2}$	
$\frac{1}{40}$	$0.37 \cdot 10^{-4}$	3.93	$0.95 \cdot 10^{-4}$	3.74	$0.47 \cdot 10^{-4}$	3.92	$0.95 \cdot 10^{-4}$	3.74
$\frac{1}{80}$	$0.29 \cdot 10^{-5}$	4.00	$0.63 \cdot 10^{-5}$	3.91	$0.23 \cdot 10^{-5}$	4.00	$0.63 \cdot 10^{-5}$	3.90
$\frac{1}{160}$	$0.14 \cdot 10^{-6}$	3.99	$0.42 \cdot 10^{-6}$	3.90	$0.19 \cdot 10^{-6}$	3.97	$0.50 \cdot 10^{-6}$	3.66
$\frac{1}{320}$	$0.93 \cdot 10^{-8}$	3.96	$0.47 \cdot 10^{-7}$	3.17	$0.12 \cdot 10^{-7}$	3.92	$0.59 \cdot 10^{-7}$	3.09

Table 2.4. Test example (2.1) - (2.2). Full space-time errors obtained with boundary corrections.

$$\bar{A} = \frac{1}{h} \begin{pmatrix} 0 & 0 \\ 1 & -1 \\ & 1 & -1 \\ & & \ddots & \ddots \\ & & & 1 & -1 \end{pmatrix}, \quad \bar{g}(t) = \begin{pmatrix} \gamma'_0(t) \\ f(x_1, t) \\ f(x_2, t) \\ \vdots \\ f(x_m, t) \end{pmatrix}.$$

Note that applying a Runge-Kutta method to (2.27) means that the boundary values $\gamma_0(t_n + c_j \tau)$ are approximated by quadrature rules applied to

$$u(0, t_n + c_j \tau) = u(0, t_n) + \int_{t_n}^{t_n + c_j \tau} \gamma'_0(t) dt.$$

This introduces an error of course, but the point is that this error is more or less the same as the temporal error on the interior grid points. Finite difference approximations for the spatial derivatives near the boundary will therefore be better balanced than with the exact boundary values.

Considering the local error for the augmented system (2.28), similar to the Examples 2.1 and 2.2, reveals that

$$\bar{A} \bar{u}_h^{(2)} = \mathcal{O}(1), \quad \bar{u}_h^{(2)} = (u_0^{(2)}, u_1^{(2)}, \dots, u_m^{(2)})^T. \quad (2.29)$$

Consequently, for the local error of the fourth-order explicit Runge-Kutta method we will have $\|\delta_n\|_\infty = \mathcal{O}(\tau^3)$, $\|\delta_n\|_2 = \mathcal{O}(\tau^{3.5})$, and we then expect that also for the global error we get these temporal orders, due to cancellations in the transition to global errors.

To illustrate this numerically, the test problem (2.1)-(2.2) was again solved with the same spatial discretization and step sizes as used in Table 2.2, except that here the scalar problem (2.27) was integrated simultaneously. The results collected in Table 2.4 confirm the analysis, in spite of the nonlinearity of the source term. For even smaller values of τ , orders close to 3 in the L_∞ -norm and 3.5 in the L_2 -norm are found. Observe also that the accuracies are very close to those in the left block of the table for the transformed problem (2.26).

A further correction technique, based on continued differentiation of the Dirichlet boundary function, has been proposed in Carpenter et al. (1995). It was shown that for the model problem $u_t + u_x = 0$, $u(0, t) = \gamma_0(t)$, the order four of the classical Runge-Kutta method can be restored this way. However, nonlinearities, or even simple source terms as in (2.25), seem to necessitate additional measures.

For more results on hyperbolic problems, including numerical examples for nonlinear hyperbolic problems and systems, we refer to Abarbanel, Gottlieb & Carpenter (1996) and Pathria (1997). A recent paper on avoiding order reduction for parabolic problems is Calvo & Palencia (2002).

Final Remarks

Order reduction for semi-discrete PDEs does not stand on its own. Implicit Runge-Kutta methods also suffer from it when applied to common stiff ODEs.

This was observed first by Prothero & Robinson (1974) for the simple scalar test model $w'(t) = \lambda(w(t) - g(t)) + g'(t)$. The order of convergence for nonlinear ODE systems was examined in detail by Frank, Schneid & Ueberhuber (1985); see also Dekker & Verwer (1984) for a review of the subject of *B-convergence*. A comprehensive nonlinear convergence analysis along these lines for the implicit midpoint rule and the trapezoidal rule was obtained by Kraaijevanger (1985).

The first analysis on order reduction for PDEs is due to Crouzeix (1975) in connection with implicit Runge-Kutta methods applied to parabolic problems, see also Brenner, Crouzeix & Thomée (1982). For the local error analysis presented here the latter paper was closely followed. Using ideas from *B-convergence*, Verwer (1986) studied diagonally implicit Runge-Kutta methods for linear and nonlinear parabolic problems. Further results on nonlinear equations can be found in Lubich & Ostermann (1993, 1995a). For linear parabolic problems and strongly *A*-stable methods it was shown in Lubich & Ostermann (1995b) that the classical order of convergence p will still be valid in the interior of the spatial domain.

Also Runge-Kutta-Rosenbrock methods suffer from order reduction. The local and global error analysis presented here for the linear problem class (2.4) can be redone for the Rosenbrock formulas of Section 1.7, giving similar results. Further see Ostermann & Roche (1993), Lubich & Ostermann (1995c), Steinebach (1995) and Lang & Verwer (2001). All these papers deal with parabolic problems; the latter paper gives numerical results for three different Rosenbrock methods.

In conclusion, any explicit or implicit Runge-Kutta method of classical order $p \geq 3$ and stage order $q < p - 1$ may suffer from order reduction to the level $q + 1$. The question arises whether the extra computational work of adding more stages to get a higher classical order pays off. A general answer cannot be given as this is problem dependent. However, the understanding of the order reduction phenomenon and the use of correction techniques is essential in situations where one is interested in higher order methods. It goes without saying that in practice also the spatial accuracy must be taken into account.

3 Linear Multistep Methods

Runge-Kutta methods are one-step methods, which means that for computing w_{n+1} only the last approximation w_n is needed. Linear multistep methods use additional past approximations from the previous time levels. The linear k -step method is defined by the formula

$$\sum_{j=0}^k \alpha_j w_{n+j} = \tau \sum_{j=0}^k \beta_j F(t_{n+j}, w_{n+j}), \quad n = 0, 1, \dots, \quad (3.1)$$

which uses the k past values w_n, \dots, w_{n+k-1} to compute w_{n+k} . Notice that the most advanced time level here is t_{n+k} instead of t_{n+1} . This is only for convenience of notation. The method is explicit if $\beta_k = 0$ and implicit otherwise. Formula (3.1) can be scaled since multiplication of all coefficients α_j, β_j with a same factor will leave the computational scheme unchanged. Usually, scaling is used to set $\alpha_k = 1$ or $\beta_0 + \beta_1 + \dots + \beta_k = 1$. We will assume $\alpha_k > 0$.

For $k = 1$ there is an overlap with Runge-Kutta methods: the θ -method belongs to class (3.1) and can also be written as a Runge-Kutta method. However, for $k \geq 2$ multistep methods are essentially different. The linear multistep methods form an important class and are often used in practice, both for ODEs and PDEs. Most of the theory behind linear multistep methods started from the seminal paper Dahlquist (1956).

The computational advantage of a linear k -step method over a one-step s -stage Runge-Kutta method is that in case of explicit methods only one F -evaluation is needed against s for the Runge-Kutta method, while in case of implicit methods only one system of nonlinear algebraic equations, of the ODE dimension m , has to be solved. On the other hand, there are some additional features that need consideration when implementing a multistep method (3.1). Firstly, it needs k starting values w_0, w_1, \dots, w_{k-1} for the first step while only the initial value $w_0 = w(0)$ is given. The other starting values can be computed with a one-step Runge-Kutta method. Another possibility is to use a linear 1-step method to compute w_1 , then a linear 2-step method for w_2 , and so on, with small initial step size, until all necessary starting values have been found. Secondly, changing the step size τ will change all coefficients α_j, β_j . To keep our presentation short and as simple as possible, in this section the step size τ is always supposed to be fixed in time so that also α_j, β_j will remain fixed. Variable step sizes will be discussed in Section 5.

Finally we note that, like Runge-Kutta and Rosenbrock methods, any linear multistep method mimics the mass conservation property discussed in Remark 1.2, provided this holds for the starting values.

3.1 The Order Conditions

Consider a fixed time point $t_N = N\tau \leq T$. The *global* error at t_N is the difference $w(t_N) - w_N$. This error depends on errors present in all preceding approximations w_n ($1 \leq n \leq N-1$) and can be interpreted as being built up from *local* errors. This local error build-up is subject to stability, as it is for Runge-Kutta methods. Assuming stability, the local errors will determine the size of the global error.

Let w_{n+k}^* be the linear multistep approximation which is obtained if the k past values w_{n+k-1}, \dots, w_n are taken equal to $w(t_{n+k-1}), \dots, w(t_n)$, i.e.,

$$\sum_{j=0}^{k-1} \alpha_j w(t_{n+j}) + \alpha_k w_{n+k}^* = \tau \sum_{j=0}^{k-1} \beta_j F(t_{n+j}, w(t_{n+j})) + \tau \beta_k F(t_{n+k}, w_{n+k}^*).$$

The method is then called *consistent* of order p , or more shortly to have order p , if the *local error* satisfies

$$w(t_{n+k}) - w_{n+k}^* = \mathcal{O}(\tau^{p+1}) \quad (3.2)$$

whenever F is sufficiently differentiable. Consistency of order p thus means that the local error introduced in a single step from the exact solution is $\mathcal{O}(\tau^{p+1})$. Considering the point t_N in a convergence analysis, the preceding local errors are built up such that one power of $\tau = t_N/N$ is lost and the global error becomes $\mathcal{O}(\tau^p)$. Hence a stable linear multistep method of consistency order p converges with order p when applied to a smooth ODE problem. A requirement is that the $k-1$ additional starting values w_1, \dots, w_{k-1} must also have been computed within convergence order p . We will pay attention to convergence analysis in Section 3.5 after having dealt with consistency and stability.

To find the conditions for order p consistency, it is customary to insert exact solution values in (3.1) at all points t_{n+j} , including t_{n+k} . This yields

$$\sum_{j=0}^k \alpha_j w(t_{n+j}) = \tau \sum_{j=0}^k \beta_j F(t_{n+j}, w(t_{n+j})) + \tau \rho_{n+k-1}, \quad (3.3)$$

with $\tau \rho_{n+k-1}$ as defect. The local error and this defect are related by the equation

$$w(t_{n+k}) - w_{n+k}^* = (\alpha_k I - \tau \beta_k A_{n+k})^{-1} \tau \rho_{n+k-1}, \quad (3.4)$$

$$A_{n+k} = \int_0^1 \frac{\partial F}{\partial w}(t_{n+k}, \sigma w(t_{n+k}) + (1-\sigma)w_{n+k}^*) d\sigma.$$

Note that here the mean value theorem has been used, see Ortega & Rheinboldt (1970, p. 71). A similar relation was given in Section I.2 for the θ -method.

For a fixed ODE system, we see from (3.4) that (3.2) is equivalent with $\|\tau \rho_{n+k-1}\| = \mathcal{O}(\tau^{p+1})$. Consequently, for deriving the order conditions we may proceed with this defect. The scaled quantity $(\alpha_k)^{-1} \rho_{n+k-1}$ is usually called the (*local*) *truncation error* of the method. Assuming the solution to be sufficiently often differentiable, Taylor series expansion around $t = t_n$ yields

$$\tau \rho_{n+k-1} = C_0 w(t_n) + \tau C_1 w'(t_n) + \tau^2 C_2 w''(t_n) + \dots,$$

where

$$C_0 = \sum_{j=0}^k \alpha_j, \quad C_i = \frac{1}{i!} \left(\sum_{j=0}^k \alpha_j j^i - i \sum_{j=0}^k \beta_j j^{i-1} \right) \quad \text{for } i \geq 1.$$

Thus the method has order p iff it satisfies the *order conditions*

$$\sum_{j=0}^k \alpha_j = 0, \quad \sum_{j=0}^k \alpha_j j^i = i \sum_{j=0}^k \beta_j j^{i-1} \quad \text{for } i = 1, 2, \dots, p. \quad (3.5)$$

3.2 Examples

Example 3.1 The method

$$w_{n+2} - w_n = 2\tau F(t_{n+1}, w_{n+1}) \quad (3.6)$$

is called the 2-step *explicit midpoint rule*. Its order is equal to two and it is often used for special classes of problems arising from hyperbolic PDEs and other wave type equations. When combined with central differences it gives the well-known *leap-frog* scheme. We will see that the explicit midpoint rule has rather poor stability properties for general ODE systems which restricts its use. \diamond

Example 3.2 *Adams methods* are characterized by

$$\alpha_k = 1, \quad \alpha_{k-1} = -1, \quad \alpha_j = 0 \quad (0 \leq j \leq k-2)$$

with β_j chosen such that the order is optimal. Explicit Adams methods, also called *Adams-Basforth* methods, have order k . The method with $k=1$ is forward Euler. The 2-step and 3-step methods are

$$w_{n+2} - w_{n+1} = \frac{3}{2}\tau F_{n+1} - \frac{1}{2}\tau F_n, \quad (3.7)$$

$$w_{n+3} - w_{n+2} = \frac{23}{12}\tau F_{n+2} - \frac{16}{12}\tau F_{n+1} + \frac{5}{12}\tau F_n, \quad (3.8)$$

where F_j stands for $F(t_j, w_j)$. Implicit Adams methods, also known as *Adams-Moulton* methods, have order $k+1$. The method with $k=1$ is the trapezoidal rule and for $k=2, 3$ we have

$$w_{n+2} - w_{n+1} = \frac{5}{12}\tau F_{n+2} + \frac{8}{12}\tau F_{n+1} - \frac{1}{12}\tau F_n, \quad (3.9)$$

$$w_{n+3} - w_{n+2} = \frac{9}{24}\tau F_{n+3} + \frac{19}{24}\tau F_{n+2} - \frac{5}{24}\tau F_{n+1} + \frac{1}{24}\tau F_n. \quad (3.10)$$

Implicit Adams methods are usually applied in a *predictor-corrector* fashion, that is, first a predictor \bar{w}_{n+k} is computed from the explicit k -step method and its function value is then inserted in the right-hand side of the implicit k -step method for the most advanced time level. The method thus obtained is explicit and has order $k+1$. It is, however, no longer a genuine linear k -step method. It falls in the wider class of so-called multistep Runge-Kutta methods, with the prediction \bar{w}_{n+k} playing the role of an internal vector. For $k=1$ this procedure gives the explicit trapezoidal rule (1.6). \diamond

Example 3.3 *Backward differentiation formulas*, usually called *BDFs* or *BDF methods*, are implicit and defined by

$$\beta_k = 1, \quad \beta_j = 0 \quad (0 \leq j \leq k-1)$$

with α_j chosen such that the order is optimal, namely k . The 1-step BDF method is implicit Euler. The 2-step and 3-step BDF methods are

$$\frac{3}{2}w_{n+2} - 2w_{n+1} + \frac{1}{2}w_n = \tau F_{n+2}, \quad (3.11)$$

$$\frac{11}{6}w_{n+3} - 3w_{n+2} + \frac{3}{2}w_{n+1} - \frac{1}{3}w_n = \tau F_{n+3}. \quad (3.12)$$

Due to their favourable stability properties for stiff problems, the BDF methods are well suited to solve stiff chemical reaction equations and parabolic problems. In fact, in chemistry applications they belong to the most widely used methods. The BDF methods were introduced in the numerical ODE field by Curtiss & Hirschfelder (1952), and their popularity can be attributed to a large extent to Gear (1971).

In Chapter IV we will combine implicit methods with closely related explicit ones, obtained by replacing the implicit function evaluation F_{n+k} by a linear combination of explicit terms using k th-order extrapolation. These explicit counterparts of the implicit BDF methods for $k = 2, 3$ are given by

$$\frac{3}{2}w_{n+2} - 2w_{n+1} + \frac{1}{2}w_n = 2\tau F_{n+1} - \tau F_n, \quad (3.13)$$

$$\frac{11}{6}w_{n+3} - 3w_{n+2} + \frac{3}{2}w_{n+1} - \frac{1}{3}w_n = 3\tau F_{n+2} - 3\tau F_{n+1} + \tau F_n. \quad (3.14)$$

These methods also have order k ; they are often called *extrapolated BDF methods*. Being explicit, they lack of course the favourable stability properties of the original implicit BDF methods. \diamond

3.3 Stability Analysis

As for Runge-Kutta methods, stability of linear multistep methods is studied mainly for linear ODEs and in particular for the familiar scalar test equation $w'(t) = \lambda w(t)$ with λ complex valued.

Linear Recursions

Let us first recall some properties of the scalar linear recursion formula

$$\sum_{j=0}^k \gamma_j w_{n+j} = 0, \quad n = 0, 1, \dots, \quad (3.15)$$

with constant coefficients $\gamma_j \in \mathbb{C}$. For any set of starting values w_0, \dots, w_{k-1} the behaviour of the sequence $\{w_n\}$, in particular boundedness for $n \rightarrow \infty$, is determined by the zeros of the so-called *characteristic polynomial*

$$\pi(\zeta) = \sum_{j=0}^k \gamma_j \zeta^j.$$

Let $\zeta_1, \zeta_2, \dots, \zeta_k$ denote these zeros with multiple zeros repeated. The general solution of (3.15) can then be verified to be

$$w_n = c_1 n^{\nu_1} \zeta_1^n + c_2 n^{\nu_2} \zeta_2^n + \dots + c_k n^{\nu_k} \zeta_k^n, \quad n = 0, 1, \dots,$$

where the constants c_i are determined by the starting values and ν_i are non-negative integers defined as follows: $\nu_i = 0$ if ζ_i is a simple zero and $\nu_i = 0$, $\nu_{i+1} = 1, \dots, \nu_{i+l} = l$ if $\zeta_i = \dots = \zeta_{i+l}$ is a zero with multiplicity $l + 1$. This is a basic result in the theory of linear difference schemes found in most numerical analysis textbooks.

The characteristic polynomial is said to satisfy the *root condition* if

$$|\zeta_i| \leq 1 \quad \text{for all } i, \text{ and} \quad |\zeta_i| < 1 \quad \text{if } \zeta_i \text{ is not simple.} \quad (3.16)$$

From the formula for the general solution one then deduces that this root condition is equivalent with *boundedness* of the sequence $\{w_n\}$ for $n \rightarrow \infty$ for arbitrary starting values.

In the following these results are used for the multistep method (3.1). The coefficients of the method will be contained in the polynomials

$$\rho(\zeta) = \sum_{j=0}^k \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^k \beta_j \zeta^j. \quad (3.17)$$

Zero-Stability

For $F = 0$ the linear multistep formula (3.1) reduces to the linear recursion

$$\sum_{j=0}^k \alpha_j w_{n+j} = 0, \quad n = 0, 1, \dots, \quad (3.18)$$

with characteristic polynomial $\rho(\zeta)$. The linear multistep method is said to be *zero-stable* if its characteristic polynomial $\rho(\zeta)$ satisfies the root condition. Zero-stability is a prerequisite for a method to be useful. Since recursion (3.18) is obtained for the trivial equation $w'(t) = 0$, a method which fails to be zero-stable cannot even solve this trivial equation properly (in a stable manner). Note that zero-stability is a typical multistep requirement; the corresponding requirement for consistent one-step methods holds trivially.

Remark 3.4 Inspection of the order conditions and counting the number of coefficients α_j, β_j indicates that the maximal attainable order of an implicit and explicit k -step method is $2k$ and $2k - 1$, respectively. However, zero-stability reduces the attainable order to $p = k$ for explicit methods and to $p = 2\lfloor(k+2)/2\rfloor$ for implicit ones. This is known as the *first Dahlquist barrier*, Dahlquist (1956).

As an example, consider the 2-step explicit method

$$w_{n+2} - (1 + \alpha_0)w_{n+1} + \alpha_0 w_n = \frac{1}{2}(3 - \alpha_0)\tau F_{n+1} - \frac{1}{2}(1 + \alpha_0)\tau F_n.$$

For $\alpha_0 = 0$ the explicit Adams method with $p = 2$ results. For $\alpha_0 = -5$ we obtain a method of order three, but this method is not zero-stable because $\rho(\zeta)$ has as roots 1 and -5 . The unstable behaviour of this third-order method is discussed and illustrated for instance in Hairer et al. (1993) and Lambert (1991). \diamond

The Stability Region

For the scalar stability test equation $w'(t) = \lambda w(t)$, $\lambda \in \mathbb{C}$, we get the recursion

$$\sum_{j=0}^k (\alpha_j - z\beta_j) w_{n+j} = 0, \quad n = 0, 1, \dots, \quad (3.19)$$

where $z = \tau\lambda$. This recursion has the characteristic polynomial

$$\pi_z(\zeta) = \rho(\zeta) - z\sigma(\zeta) = \sum_{j=0}^k (\alpha_j - z\beta_j) \zeta^j. \quad (3.20)$$

The *stability region* $\mathcal{S} \subset \mathbb{C}$ of the linear multistep method is defined as the set consisting of all z such that the sequence $\{w_n\}$ is bounded for any choice of starting vectors w_0, \dots, w_{k-1} . Obviously,

$$z \in \mathcal{S} \iff \pi_z \text{ satisfies the root condition}$$

and the linear multistep method is zero-stable iff $0 \in \mathcal{S}$.

One of the roots of $\pi_z(\zeta)$ approximates e^z up to $\mathcal{O}(z^{p+1})$ for $z \rightarrow 0$.⁷⁾ This particular root is called the *principal root*. The remaining $k-1$ roots are called *spurious* or *parasitic* roots. These spurious roots have no relation to accuracy of the method but they are often a cause for instability.

Example 3.5 For the 2-step explicit midpoint rule (3.6), with given starting values w_0, w_1 , the recursion (3.19) reads

$$w_{n+2} = w_n + 2zw_{n+1}, \quad n = 0, 1, \dots. \quad (3.21)$$

The characteristic polynomial

$$\pi_z(\zeta) = \zeta^2 - 2z\zeta - 1$$

has the two roots $\zeta_{\pm} = z \pm \sqrt{1+z^2}$, where ζ_+ is the principal root, approximating e^z up to $\mathcal{O}(z^3)$ for $z \rightarrow 0$, and ζ_- is the spurious one. Their product equals -1 and thus the root condition is satisfied iff ζ_{\pm} lie on the unit circle and are simple. Considering $\zeta_+ = e^{i\theta}$ and $\zeta_- = -e^{-i\theta}$, the z -values leading to these roots satisfy

$$z = (\zeta_{\pm}^2 - 1)/(2\zeta_{\pm}) = \frac{1}{2}(e^{i\theta} - e^{-i\theta}) = i \sin \theta,$$

and thus the stability region is

$$\mathcal{S} = \{z \in \mathbb{C} : \operatorname{Re} z = 0, |z| < 1\}.$$

The form of this stability region is unusual since it is merely the line segment on the imaginary axis between $-i$ and $+i$ (the points $\pm i$ are excluded since

⁷⁾ This can be shown by inserting $w(t) = e^{\lambda t}$, $F(t, w) = \lambda w$, into (3.3).

there the two roots coincide). This restricted stability confines the application of the explicit midpoint rule to problems with a purely imaginary spectrum.

With $z = i \sin \theta$, $|\theta| < \frac{1}{2}\pi$, $w_0 = 1$ and forward Euler to generate the second starting value $w_1 = 1 + z$, recursion (3.21) has the general solution

$$w_n = \frac{1 + \cos \theta}{2 \cos \theta} e^{ni\theta} + (-1)^n \frac{-1 + \cos \theta}{2 \cos \theta} e^{-ni\theta}, \quad n = 2, 3, \dots, \quad (3.22)$$

whose first term is associated to the principal root and the second one to the spurious root. Observe that this second term gives rise to an oscillatory solution behaviour. Except for z close to 0, that is, $\cos \theta$ close to 1, the oscillatory behaviour will show up.

This means that in hyperbolic PDE applications where the given initial solution is non-smooth, i.e., rich in high spatial frequencies, the oscillatory behaviour will readily be felt. Since the oscillations are not damped, the only remedy to reduce them is to lower the step size τ . On the other hand, for wave type problems damping may be unwanted, so in such circumstances the 2-step explicit midpoint rule can have an advantage over methods with damping. Finally, note that stability of the forward Euler method plays no role in (3.22) as it is applied only once. \diamond

To determine \mathcal{S} for a general method we observe that on the boundary $\partial\mathcal{S}$ one of the roots of the characteristic polynomial $\pi_z(\zeta)$ must have modulus 1. Since $\pi_z(\zeta) = 0$ iff $z = \rho(\zeta)/\sigma(\zeta)$, it follows that any point on the boundary $\partial\mathcal{S}$ is of the form

$$z = \rho(e^{i\theta})/\sigma(e^{i\theta}), \quad 0 \leq \theta \leq 2\pi. \quad (3.23)$$

This image of the unit circle under ρ/σ is called the *root locus curve*.

The stability regions of the explicit Adams-Bashforth and explicit BDF-type methods with $k = 2, 3$ are given in the Figures 3.1, 3.2. Observe that the regions are rather small. The Adams-Moulton methods are implicit but

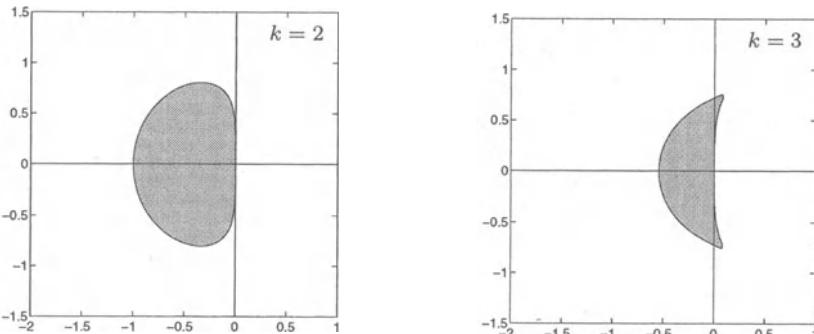


Fig. 3.1. Stability regions \mathcal{S} for the 2-step and 3-step Adams-Bashforth methods (3.7), (3.8).

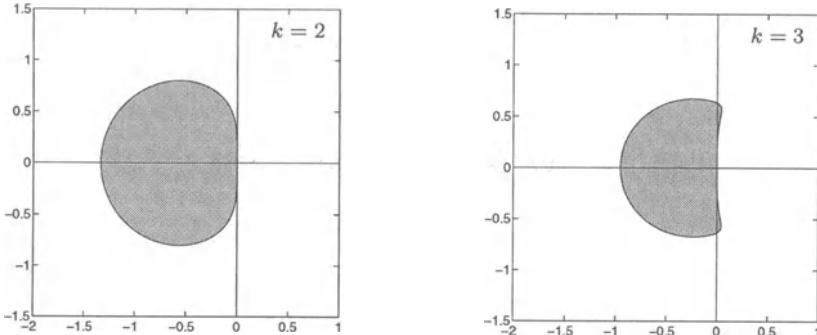


Fig. 3.2. Stability regions \mathcal{S} for the 2-step and 3-step extrapolated BDF methods (3.13), (3.14).

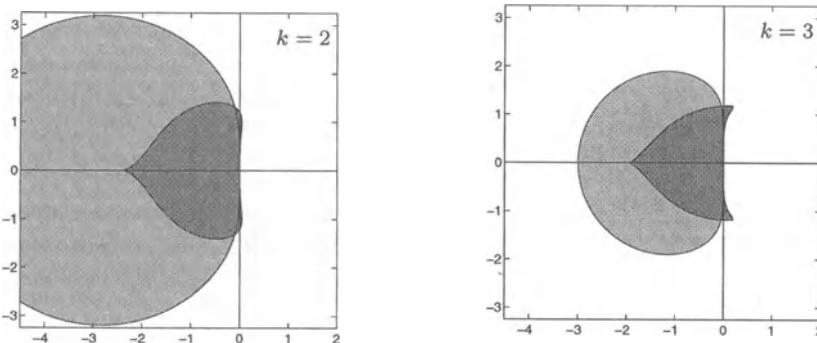


Fig. 3.3. Stability regions \mathcal{S} for the 2-step and 3-step Adams-Moulton methods (3.9), (3.10) (light shaded) and corresponding predictor-corrector methods (dark shaded).

still have bounded stability regions for $k > 1$; the stability region of these methods intersects the negative real axis at $z = -6$ if $k = 2$ and at $z = -3$ if $k = 3$, see Figure 3.3. These implicit methods are therefore not suitable for stiff problems. For application to non-stiff problems the Adams methods are usually implemented in a predictor-corrector fashion, where the explicit formula is once inserted into the right-hand side of the implicit formula. The stability regions of these 2-stage predictor-corrector methods for $k = 2, 3$ are also given in Figure 3.3. Notice that the scale in Figure 3.3 is larger than in the Figures 3.1 and 3.2. Pictures of the stability regions of some other Adams methods can be found in Lambert (1991), Hairer & Wanner (1996).

A-Stability

As for Runge-Kutta methods, the *A*-stability property is desirable for stiff problems. For multistep methods it is convenient to include also the point $z = \infty$ into our considerations. Let $\overline{\mathbb{C}} = \mathbb{C} \cup \infty$. We say that $z = \infty$ belongs to

the stability region \mathcal{S} if the polynomial $\sigma(\zeta)$ with leading coefficient $\beta_k \neq 0$ satisfies the root condition. This makes sense because the roots of π_z tend to the roots of σ for $z \rightarrow \infty$ (divide $\pi_z(\zeta)$ by z). A linear multistep method is then called *A-stable* if

$$\mathcal{S} \supset \{z \in \overline{\mathbb{C}} : \operatorname{Re} z \leq 0 \text{ or } z = \infty\}.$$

For most methods it suffices to simply require $\mathcal{S} \supset \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\}$, but it is possible that with this weaker assumption the case $z \rightarrow \infty$ will give a weak instability, see also Lemma 3.7.

In contrast to Runge-Kutta methods there are not many *A*-stable linear multistep methods. Dahlquist (1963) proved that the consistency order of an *A*-stable multistep method can be at most two. This fundamental result is called the *second Dahlquist barrier* and it might seem an obstacle for solving stiff ODE problems. However, for many stiff ODEs, such as chemical kinetics problems and semi-discrete parabolic PDEs, the *A*-stability property is not needed since for such problems the eigenvalues λ with large modulus stay away from the imaginary axis.

In this respect a less demanding property is *A(α)-stability* with angles $\alpha \in [0, \frac{1}{2}\pi]$. For $z = re^{i\phi}$ with $r > 0$, $\phi \in (-\pi, \pi]$, let $\arg(z) = \phi$. A method is said to be *A(α)*-stable if

$$\mathcal{S} \supset \{z \in \overline{\mathbb{C}} : z = 0, \infty \text{ or } |\arg(-z)| \leq \alpha\}.$$

Hence the eigenvalues are supposed to lie in the left half-plane within an infinite wedge with angle α with respect to the negative real axis. This concept is useful for all ODE methods, but particularly so for multistep methods since it allows methods of consistency order higher than two.

Example 3.6 The implicit BDF methods from Example 3.3 are *A*-stable for $k = 1, 2$ and *A(α)*-stable for $3 \leq k \leq 6$ with angle α depending on k :

k	1	2	3	4	5	6	7
α	90°	90°	86°	73°	51°	17°	—

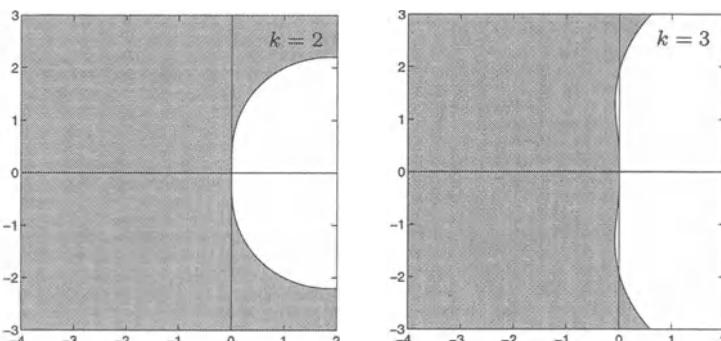


Fig. 3.4. Stability regions \mathcal{S} for the 2-step and 3-step BDF methods (3.11), (3.12).

For $k \geq 7$ the methods are no longer zero-stable and hence not convergent (Gear, 1971). Since the angle α for the 6-step method is rather small, the BDF methods are mostly used for $k \leq 5$. Figure 3.4 shows the stability regions for $k = 2, 3$. Notice that the BDF method for $k = 1$ is just the backward Euler method. Pictures for the larger k can be found in Hairer & Wanner (1996) and Gear (1971). \diamond

Stability for Linear Systems

For the stability analysis it may be convenient to write the k -step recursion (3.19) for the scalar test equation in a one-step form. Observe that (3.19) is equivalent to

$$w_{n+k} = - \sum_{j=0}^{k-1} \frac{\alpha_j - z\beta_j}{\alpha_k - z\beta_k} w_{n+j}.$$

By introducing the vector $W_n = (w_{n+k-1}, \dots, w_n)^T \in \mathbb{C}^k$, the scalar k -step recursion can then be written in the one-step form in \mathbb{C}^k ,

$$W_{n+1} = R(z) W_n, \quad (3.24)$$

where

$$R(z) = \begin{pmatrix} r_1(z) & r_2(z) & \cdots & r_k(z) \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}, \quad r_i(z) = -\frac{\alpha_{k-i} - z\beta_{k-i}}{\alpha_k - z\beta_k}.$$

The matrix $R(z)$ is usually called the *companion matrix* of the multistep method. Obviously, it is the counterpart of the stability function for Runge-Kutta methods. From the equivalence of (3.19) and (3.24) it is clear that

$$z \in \mathcal{S} \iff R(z) \text{ is power bounded.}$$

For linear ODE systems $w'(t) = Aw(t) \in \mathbb{R}^m$ we obtain in the same way $W_{n+1} = R(Z)W_n$ with $Z = \tau A \in \mathbb{R}^{m \times m}$ and

$$R(Z) = \begin{pmatrix} r_1(Z) & r_2(Z) & \cdots & r_k(Z) \\ I & O & & \\ & \ddots & \ddots & \\ & & I & O \end{pmatrix} \in \mathbb{R}^{mk \times mk}.$$

Let $\|\cdot\|$ be the given norm on \mathbb{R}^m . We then define the norm $|||V|||$ for vectors $V = (V_1^T, \dots, V_k^T)^T \in \mathbb{R}^{mk}$ by

$$|||V||| = \max_{1 \leq i \leq k} \|V_i\|. \quad (3.25)$$

Other, equivalent norms can be taken as well but the above norm is convenient to relate the norms of the W_n to those of the w_n . It holds with these norms that $\{w_n\}$ is a bounded sequence iff $\{W_n\}$ is bounded. Note that $W_n = R(\tau A)^n W_0$ and thus

$$|||W_n||| \leq |||R(\tau A)^n||| \cdot |||W_0|||.$$

Lemma 3.7 *Let $\mathcal{D} \subset \mathcal{S}$ be closed in $\overline{\mathbb{C}}$ and suppose $A = UAU^{-1}$ with $\tau\lambda_i \in \mathcal{D}$. Then there exists a constant $C > 0$, only depending on \mathcal{D} and the coefficients of the method, such that*

$$\sup_{n \geq 0} |||R(\tau A)^n||| \leq C \operatorname{cond}(U).$$

Due to the fact that A is assumed to be diagonalizable, the proof of the lemma can be reduced to the scalar complex case. With this, however, the proof does not become trivial, since we then have to show that $|||R(z)^n||| \leq C'$ uniformly for $z \in \mathcal{D}$. A proof of this can be found in Hairer & Wanner (1996, Sect. V.7). Here the condition that \mathcal{D} is closed in $\overline{\mathbb{C}}$ is needed to ensure that points z on the boundary of \mathcal{S} that do not belong to \mathcal{S} itself are avoided (this is relevant if π_z has a double root with modulus 1 on the boundary). If $z_i = \tau\lambda_i$ gets too close to such a point then the constant C will become large. For most known methods \mathcal{S} itself is closed in $\overline{\mathbb{C}}$ and then we can simply take $\mathcal{D} = \mathcal{S}$.

The above lemma can be used, for example, for $A(\alpha)$ -stable methods if we deal with the L_2 -norm and the matrix A is normal with eigenvalues λ_i such that $|\arg(-\lambda_i)| \leq \alpha$. This is applicable to show unconditional stability for advection-diffusion equations, see Section I.3.4.

3.4 Step Size Restrictions for Advection-Diffusion

In Section 1.4 step size restrictions for explicit Runge-Kutta methods have been discussed emerging from a von Neumann analysis for the advection problem $u_t + au_x = 0$ and the diffusion problem $u_t = du_{xx}$. Here similar results are given for explicit multistep methods. The methods considered are

	EBD2	AB2	ABM2	EBD3	AB3	ABM3
$z_{a,1}$	0.66	0.50	0.98 (0.49)	0.47	0.27	0.79 (0.39)
$z_{a,2}$	0.00	0.00	1.20 (0.60)	0.63	0.72	1.17 (0.58)
$z_{a,3}$	0.46	0.58	1.02 (0.51)	0.49	0.39	0.80 (0.40)
$z_{a,4}$	0.00	0.00	0.87 (0.43)	0.46	0.52	0.85 (0.42)

Table 3.1. Maximal Courant numbers $\nu = \tau a/h$ for stability, with scaled values $\nu/2$ for ABM in parentheses.

	EBD2	AB2	ABM2	EBD3	AB3	ABM3
$z_{d,2}$	0.33	0.25	0.60 (0.30)	0.23	0.13	0.48 (0.24)
$z_{d,4}$	0.25	0.18	0.44 (0.22)	0.17	0.10	0.36 (0.18)

Table 3.2. Maximal values $\mu = \tau d/h^2$ for stability, with scaled values $\mu/2$ for ABM in parentheses.

2- and 3-step extrapolated BDF-type methods (EBD) and Adams-Bashforth (AB) methods, and the explicit predictor-corrector combinations of Adams-Moulton using AB as predictor (ABM), see Examples 3.2, 3.3. The layout of the tables is the same as in Tables 1.2 and 1.3 for the Runge-Kutta methods, with $z_{a,q}$ and $z_{d,q}$ the relevant expressions of $z = \tau \lambda$ for the common advection and diffusion finite difference discretizations of order q . The ABM methods require per step two function evaluations, and therefore we also list for these methods the corresponding scaled values.

An interesting observation is that compared to Runge-Kutta methods the maximal Courant numbers and μ -values are not that different if they are scaled so as to incorporate the computational effort per step. Compare for example the scaled Courant numbers for $s = 4$ in the Runge-Kutta table with those for the fourth-order ABM3 method. Yet, for advection problems this comparison slightly favours the classical explicit Runge-Kutta method since it is easier to implement due to its one-step nature. For diffusion problems none of these conventional explicit methods is advocated due to the very severe step size constraints.

Remark 3.8 For the advection problem with spatial central differences, the maximal step size for stability is determined by the segment $[-i\beta_I, i\beta_I]$ of the imaginary axis that fits in the stability domain. For general results on the maximal size of the imaginary stability boundary β_I we refer to Jeltsch & Nevanlinna (1981). As for Runge-Kutta methods, problems with purely imaginary eigenvalues are mostly solved with conventional methods, such as the 2-step explicit midpoint rule or higher-order Adams predictor-corrector methods. ◇

3.5 Convergence Analysis

Once consistency and stability has been established, convergence easily follows. We will sketch this for linear systems $w'(t) = Aw(t) + g(t)$ in \mathbb{R}^m as we did in Section I.2.6 for the θ -method. Let $\varepsilon_n = w(t_n) - w_n$ be the global error. Subtracting (3.1) from (3.3), we obtain the error recursion

$$\sum_{j=0}^k (\alpha_j I - \tau \beta_j A) \varepsilon_{n+j} = \tau \rho_{n+k-1}, \quad n = 0, 1, \dots . \quad (3.26)$$

Next introduce the following vectors in \mathbb{R}^{km} ,

$$E_n = \begin{pmatrix} \varepsilon_{n+k-1} \\ \varepsilon_{n+k-2} \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad D_n = \begin{pmatrix} (\alpha_k I - \tau \beta_k A)^{-1} \tau \rho_{n+k-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and write (3.26) in the one-step form

$$E_{n+1} = R(\tau A) E_n + D_n, \quad n = 0, 1, \dots \quad (3.27)$$

in the space \mathbb{R}^{km} where $R(\tau A)$ is the block-structured $km \times km$ companion matrix acting as stability matrix. The first vector component of D_n is just the local error $w(t_{n+k}) - w_{n+k}^*$ that was introduced in Section 3.1. Repeated application gives

$$E_n = R(\tau A)^n E_0 + \sum_{j=0}^{n-1} R(\tau A)^{n-j-1} D_j. \quad (3.28)$$

To relate local errors to the global error, we consider on \mathbb{R}^{km} the norm (3.25) starting from a given norm $\|\cdot\|$ on \mathbb{R}^m . The following logical step is to assume that $R(\tau A)$ is power bounded. More precisely, consider a time interval $[0, T]$ and suppose a moderately sized constant K exists such that

$$\|R(\tau A)^n\| \leq K \quad \text{for } n \geq 0, n\tau \leq T, \quad (3.29)$$

see Lemma 3.7 for a sufficient condition. Then for all $n \geq 0, n\tau \leq T$ we have

$$\|E_n\| \leq K \|E_0\| + K \sum_{j=0}^{n-1} \|D_j\|.$$

From $\|R(\tau A)\| \leq K$, it follows in particular that $\|r_i(\tau A)\| \leq K$ for $i = 1, \dots, k$. Choosing numbers γ_j such that $\sum_{j=0}^k \alpha_j \gamma_j = 1$ and $\sum_{j=0}^k \beta_j \gamma_j = 0$, we obtain

$$(\alpha_k I - \tau \beta_k A)^{-1} = \gamma_k I - \sum_{j=0}^{k-1} \gamma_j r_{k-j}(\tau A) = \mathcal{O}(1).$$

Hence we have the local error bound $\|D_j\| \leq C \|\rho_{j+k-1}\|$. Now an elementary calculation yields the global error estimate

$$\|\varepsilon_n\| \leq K \max_{0 \leq j \leq k-1} \|\varepsilon_j\| + K C T \max_{k-1 \leq j \leq n-1} \|\rho_j\| \quad (3.30)$$

for $n \geq 0, n\tau \leq T$. Assuming the method to be consistent of order p and the starting values to be sufficiently accurate, $\|\varepsilon_j\| \leq C_0 \tau^p$, $j = 0, 1, \dots, k-1$, we thus obtain the global result $\|\varepsilon_n\| = \mathcal{O}(\tau^p)$, that is, convergence of order p .

Noteworthy is that the residuals ρ_j only depend on powers of τ multiplying higher derivatives of the exact solution. Hence if the solution is smooth, in the sense that the solution derivatives are of moderate size, we retain this smoothness property in the estimate for the global error. For stiff ODE problems and semi-discrete PDEs this is a favourable situation because it implies that linear multistep methods do not suffer from order reduction and retain their classical ODE convergence order when applied to semi-discrete PDEs.

The convergence analysis for PDEs in the Method of Lines framework, as discussed in Section 2 for Runge-Kutta methods, therefore hardly differs from the ODE analysis given above. We leave it as an exercise to incorporate the local space error $\sigma_h(t) = u'_h(t) - F(t, u_h(t))$ in ρ_n and to derive for the linear case an estimate for the space-time global error along the above lines.

When a linear multistep method is implemented in the predictor-corrector fashion, the resulting scheme can be seen as one using intermediate stages as in Runge-Kutta methods. If the order of the intermediate stages is lower than the order of the corrector formula, which usually holds, we then again have to face order reduction for semi-discrete PDEs, similar as for Runge-Kutta methods. The same happens for the so-called *one-leg formulation* of linear multistep methods

$$\sum_{j=0}^k \alpha_j w_{n+j} = \tau F\left(\sum_{j=0}^k \beta_j t_{n+j}, \sum_{j=0}^k \beta_j w_{n+j}\right), \quad (3.31)$$

with scaling $\sum_{j=0}^k \beta_j = 1$. These one-leg methods were introduced by Dahlquist (1975). The order conditions for $p \leq 2$ are the same as for linear multistep methods, but for higher order additional conditions arise. For these methods there may be a local order reduction, but globally the classical order will be retrieved in the same way as for the implicit midpoint rule in Example 2.6, see Hairer & Wanner (1996) for details.

Remark 3.9 In the above convergence analysis, stability through the power boundedness property (3.29) is the crucial consideration. If we are in a non-stiff situation, say $\|A\| = \mathcal{O}(1)$, then we have

$$R(\tau A) = R(O) + \mathcal{O}(\tau),$$

with O the $m \times m$ zero matrix, and then power boundedness can be deduced from zero-stability by a perturbation argument. This approach can even be extended to general non-stiff nonlinear systems $w'(t) = F(t, w(t))$. Hence for non-stiff problems zero-stability is the essential stability property needed to establish convergence. See for instance Lambert (1991) and Hairer et al. (1993) for a detailed treatment.

In stiff situations, where $\|\tau A\| \gg 1$, this perturbation argument cannot be used, even if τ is small. Zero-stability remains necessary but one then needs stronger stability concepts, such as $A(\alpha)$ -stability, to derive sensible global

error estimates. Lemma 3.7 is essentially applicable to normal matrices only. A generalization for A -stable methods and non-normal matrices satisfying $\max_v \langle v, Av \rangle \leq 0$ was obtained by Nevanlinna (1985), similar to the results in Section I.2.7 for the θ -method. Related nonlinear results for stiff systems were obtained by Dahlquist (1975) in the G -stability theory. Again we refer to Hairer & Wanner (1996) for a comprehensive presentation of these results. More recent linear stability results for $A(\alpha)$ -stable methods, formulated in a general Banach-space setting, can be found in González & Palencia (1998). \diamond

4 Monotone ODE Methods

In Section I.7 we have discussed positivity⁸⁾ properties for ODE systems and a few simple integration methods: the implicit and explicit Euler method, the θ -method and the explicit trapezoidal rule. For these methods sufficient conditions on the time step were derived. In the current section we continue with positivity for more general time integration methods. We will learn that the positivity requirement usually imposes a severe step size restriction, also for implicit methods.

Positivity of a numerical scheme is closely related to other monotonicity properties, such as maximum principles, max-norm contractivity and diminution of total variation. For the moment we concentrate on positivity which is conceptually the most simple. These related monotonicity concepts will be discussed afterwards.

4.1 Linear Positivity for One-Step Methods

For one-step methods there exists a complete theory developed by Bolley & Crouzeix (1978) for linear systems $w'(t) = Aw(t)$ in which the matrix $A = (a_{ij})$ satisfies

$$a_{ij} \geq 0 \quad \text{for } i \neq j \quad \text{and} \quad a_{ii} \geq -\alpha \quad \text{for all } i, \quad (4.1)$$

with problem parameter $\alpha > 0$, and

$$A \text{ has no eigenvalues on the positive real axis.} \quad (4.2)$$

Recall from Section I.7 that (4.1) implies $w(t) \geq 0$ for all $t > 0$, $w(0) \geq 0$, irrespective of α . Also, the conditions (4.1) and (4.2) together imply that $(I - \tau A)^{-1} \geq 0$ for any step size $\tau > 0$. With fixed $\alpha > 0$, condition (4.1) is obviously equivalent to $I + \tau A \geq 0$ for $\alpha\tau \leq 1$.

Application to $w'(t) = Aw(t)$ by a one-step method of order p with stability function R gives

$$w_{n+1} = R(\tau A) w_n.$$

⁸⁾ As before ‘positivity’ actually means ‘preservation of non-negativity’.

The function R is said to be *absolutely monotonic* on an interval $[-\gamma, 0]$ if R and all its derivatives are non-negative on this interval. Let γ_R be the largest γ for which this holds. If there is no $\gamma > 0$ such that R is absolutely monotonic on $[-\gamma, 0]$ we set $\gamma_R = 0$.

In the following we consider the class \mathcal{M}_α consisting of all matrices satisfying (4.1) and (4.2) with a fixed $\alpha > 0$ and arbitrary dimension m . The next theorem is due to Bolley & Crouzeix (1978). We elaborate the proof because it gives insight in the occurrence of the derivatives of R in the requirement of absolute monotonicity; the fact that R itself should be non-negative is obvious from the scalar case $m = 1$.

Theorem 4.1 $R(\tau A) \geq 0$ for all $A \in \mathcal{M}_\alpha$ iff $\alpha\tau \leq \gamma_R$.

Proof. We write $\mu = \alpha\tau$, $N = \mu I + \tau A$ and

$$R(\tau A) = \sum_{j \geq 0} \frac{1}{j!} R^{(j)}(-\mu) N^j. \quad (4.3)$$

First the validity of this series expansion for $\mu \leq \gamma_R$ has to be demonstrated. For this, consider the scalar series

$$\sum_{j \geq 0} \frac{1}{j!} R^{(j)}(-\mu) \zeta^j, \quad \zeta \in \mathbb{C},$$

and let r be its radius of convergence. Since all coefficients are non-negative, this series diverges at $\zeta = r$. Hence R has a pole in $z = -\mu + r$. On the other hand we know that R and all its derivatives exist on $[-\gamma_R, 0]$, and thus $r > \mu$. We will show that the spectral radius $\rho(N) \leq \mu$. This implies that $\|N\| < r$ in some norm, and consequently the expansion (4.3) is valid (see Horn & Johnson (1991, p. 412), for example).

To show that $\rho(N) \leq \mu$, note that $N = \mu I + \tau A \geq 0$. According to the Perron-Frobenius theorem for non-negative matrices, see for instance Horn & Johnson (1985, p. 503), there is a non-zero vector $v \geq 0$ in \mathbb{R}^m such that

$$(\mu I + \tau A)v = \rho(N)v.$$

Hence, for any $\kappa > 0$,

$$(\kappa I - \tau A)v = (\kappa - \rho(N) + \mu)v.$$

Since $(\kappa I - \tau A)^{-1} \geq 0$, it follows that $\kappa - \rho(N) + \mu \geq 0$ for all $\kappa > 0$, and thus

$$\rho(N) \leq \mu.$$

Having established the validity of (4.3), sufficiency of the condition $\alpha\tau \leq \gamma_R$ for having $R(\tau A) \geq 0$ now follows directly from the fact that $N \geq 0$.

To prove necessity, consider the first-order upwind discretization for the advection model problem $u_t + u_x = 0$ with inflow boundary condition $u(0, t) = 0$, giving the semi-discrete system $w'(t) = Aw(t)$ with

$$A = \frac{1}{h} \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} = \frac{1}{h}(E - I) \in \mathbb{R}^{m \times m},$$

where E denotes the backward shift operator on \mathbb{R}^m . Taking $\mu = \tau/h$ gives $\tau A = -\mu I + \mu E$ and therefore

$$\begin{aligned} R(\tau A) &= R(-\mu)I + \mu R'(-\mu)E + \frac{1}{2}\mu^2 R''(-\mu)E^2 + \dots \\ &\quad \dots + \frac{1}{(m-1)!}\mu^{m-1}R^{(m-1)}(-\mu)E^{m-1}. \end{aligned}$$

This matrix is lower triangular with equal entries along the lower diagonals and the sign is determined by $R^{(j)}(-\mu)$, $j = 0, 1, \dots, m-1$. Thus we see that in order to have $R(\tau A) \geq 0$, for arbitrarily large m , it is necessary to have $R^{(j)}(-\mu) \geq 0$ for all $j \geq 0$. \square

Remark 4.2 Assumption (4.2), or rather the consequence $(\kappa I - \tau A)^{-1} \geq 0$ for $\kappa, \tau > 0$, was only used in the proof to establish the expansion (4.3). In case R is a polynomial this expansion always holds, and thus assumption (4.2) can be omitted for explicit methods. \diamond

Condition (4.1) is formulated for arbitrary matrices $A \in \mathcal{M}_\alpha$. In special practical cases this condition can be too strict, see for instance the comments in Section I.7.4 on the Crank-Nicolson scheme with the heat equation. Bolley & Crouzeix (1978) showed the condition of Theorem 4.1 to be necessary for matrices of the type $A = -cI + \epsilon \text{tridiag}(1, -2, 1)$ with $c > 0$ and with $\epsilon > 0$ sufficiently small. The use of the matrix $A = h^{-1}(E - I)$ in the above proof was taken from Spijker (1983) where contractivity in the maximum norm was studied.

The Threshold Factor γ_R

In view of the above theorem, we are interested in methods that have a large threshold factor γ_R , preferably $\gamma_R = \infty$ in which case we will have unconditional positivity, that is, positivity for any starting vector and any step size $\tau > 0$. This can only be true for implicit methods and in Section I.7 it has been proved that it holds for the backward Euler method. One might hope to find more accurate methods with this property. However, this hope is dashed by the following result, also due to Bolley & Crouzeix (1978).

Theorem 4.3 *Any method with $\gamma_R = \infty$ has order $p \leq 1$.*

A proof of this theorem, based on a characterization given by Bernstein in 1928, can be found in Bolley & Crouzeix (1978) and Hairer & Wanner (1996, Sect. IV.11). As a consequence of this theorem we see that the backward Euler method is the only well-known method having $\gamma_R = \infty$.

For the θ -method it is easily derived that $\gamma_R = 1/(1-\theta)$, see also (I.7.13). As a further result, for the Padé polynomials we have

$$R(z) = 1 + z + \frac{1}{2} z^2 + \cdots + \frac{1}{s!} z^s \quad \Rightarrow \quad \gamma_R = 1. \quad (4.4)$$

For the calculation of this bound one can use in a repeated fashion the fact that if $0 < \gamma \leq 1$ and P is a polynomial with $P(0) = 1$, $0 \leq P'(z) \leq 1$ for $z \in [-\gamma, 0]$, then also $0 \leq P(z) \leq 1$ for $z \in [-\gamma, 0]$. This result is relevant to the well-known explicit Runge-Kutta methods having $s = p \leq 4$. A table of values of γ_R for rational Padé approximations is presented in Hairer & Wanner (1996, Sect. IV.11), based on work of Kraaijevanger (1986) and van de Griend & Kraaijevanger (1986).

In the latter references results on threshold factors for many other methods can be found. For example, for explicit methods Kraaijevanger (1986) has shown that

$$p = s - 1 \quad \Rightarrow \quad \gamma_R \leq 2, \quad (4.5)$$

and equality $\gamma_R = 2$ is attained by the polynomial

$$R(z) = 1 + z + \frac{1}{2} z^2 + \cdots + \frac{1}{(s-1)!} z^{s-1} + \frac{1}{2 s!} z^s. \quad (4.6)$$

Hence in comparison with explicit $(s-1)$ -stage methods of order $s-1$, the allowable step size with respect to positivity on the class \mathcal{M}_α is then doubled at the cost of one extra evaluation. A further result of this type (with $p = 2$) will be given in Example 4.8.

Inhomogeneous Linear Systems

As a generalization we consider linear systems with a non-negative source term,

$$w'(t) = A w(t) + g(t), \quad (4.7)$$

where $A \in \mathcal{M}_\alpha$ and $g(t) \geq 0$ for all $t \geq 0$. Application of a one-step Runge-Kutta or Rosenbrock method will then lead to a recursion of the type

$$w_{n+1} = R(\tau A)w_n + \sum_{j=1}^s Q_j(\tau A) \tau g(t_n + c_j \tau), \quad (4.8)$$

see also (1.21). Hence, positivity is ensured if

$$R(\tau A) \geq 0 \quad \text{and} \quad Q_j(\tau A) \geq 0, \quad j = 1, \dots, s.$$

When we consider arbitrary non-negative source terms this may lead to a further restriction on the step size.

Example 4.4 From Section I.7.3 we know that for systems (4.7) the implicit and explicit trapezoidal rule are positive if $\alpha\tau \leq 2$ and $\alpha\tau \leq 1$, respectively. With the implicit midpoint rule (1.10) we obtain

$$w_{n+1} = (I - \frac{1}{2}\tau A)^{-1}(I + \frac{1}{2}\tau A)w_n + (I - \frac{1}{2}\tau A)^{-1}\tau g(t_{n+\frac{1}{2}}),$$

again leading to the restriction $\alpha\tau \leq 2$. On the other hand, for the one-step explicit midpoint rule (1.7) we now get

$$w_{n+1} = (I + \tau A + \frac{1}{2}\tau^2 A^2)w_n + \frac{1}{2}\tau^2 A g(t_n) + \tau g(t_{n+\frac{1}{2}}).$$

giving the step size restriction $\alpha\tau \leq 0$, meaning that positivity for arbitrary $g(t) \geq 0$ cannot be achieved. The restriction comes from the term $\frac{1}{2}\tau^2 A g(t_n)$. Note however that under the mild extra condition

$$2g(t_{n+\frac{1}{2}}) - g(t_n) \geq 0$$

on the source term, it holds that

$$\frac{1}{2}\tau^2 A g(t_n) + \tau g(t_{n+\frac{1}{2}}) \geq \frac{1}{2}(I + \tau A)\tau g(t_n),$$

and hence we will then again have positivity for $\alpha\tau \leq 1$. \diamond

Example 4.5 For linear problems $w'(t) = Aw(t)$ with $A \in \mathcal{M}_\alpha$, both the classical explicit fourth-order Runge-Kutta method (1.8.b) and the third-order Heun method (1.8.a) are subject to the restriction $\tau\alpha \leq 1$, see (4.4).

For the inhomogeneous system (4.7) the classical Runge-Kutta method results in (4.8) with the functions $Q_j(z)$ given by (2.13). Note that Q_2 and Q_3 can be taken together since $c_2 = c_3$. It follows by some straightforward calculations that positivity is ensured for $\alpha\tau \leq \frac{2}{3}$. This bound is determined by Q''_1 .

On the other hand, with Heun's method positivity is not ensured due to a zero threshold in Q_2 . We leave it as an exercise to show that this is caused by the zero weight coefficient b_2 . \diamond

4.2 Nonlinear Positivity for One-Step Methods

In Section I.7.3 nonlinear systems $w'(t) = F(t, w(t))$ were considered with F satisfying

$$v + \tau F(t, v) \geq 0 \quad \text{for all } t \geq 0, v \geq 0 \text{ and } \alpha\tau \leq 1. \quad (4.9)$$

Under this assumption positivity of forward Euler and the explicit trapezoidal rule was demonstrated. We also saw that for implicit methods we need a nonlinear counterpart of condition (4.2),

for any $v \geq 0$, $t \geq 0$ and $\tau > 0$ the equation $u = v + \tau F(t, u)$ has a unique solution that depends continuously on τ and v . $\quad (4.10)$

Recall from Section I.7.3 that uniform boundedness of $(I - \tau J(t, v))^{-1}$, with Jacobian matrix $J(t, v) = (\partial F(t, v)/\partial v)$, is sufficient for this.

Following an idea of Shu & Osher (1988) for explicit Runge-Kutta methods, it is possible to derive nonlinear positivity results for a class of diagonally implicit Runge-Kutta methods (1.1). We write the diagonally implicit methods in a special form: let $v_0 = w_n$,

$$v_i = \sum_{j=0}^{i-1} \left(p_{ij} v_j + q_{ij} \tau F(t_n + \tilde{c}_j \tau, v_j) \right) + q_i \tau F(t_n + \tilde{c}_i \tau, v_i) \quad (4.11)$$

for $i = 1, \dots, s$, and finally set $w_{n+1} = v_s$. If $\sum_{j=0}^{i-1} p_{ij} = 1$ and $q_s = 0$, this is just another way of writing the s -stage diagonally implicit form of (1.1) with $w_{ni} = v_{i-1}$ and $c_i = \tilde{c}_{i-1}$. The form (4.11) is theoretically convenient because the whole process is written in terms of linear combinations of scaled forward and backward Euler steps.

Theorem 4.6 *If all parameters p_{ij}, q_{ij}, q_i with $0 \leq j < i \leq s$ are non-negative, then method (4.11) will be positive for any F satisfying (4.9)-(4.10) under the step size restriction*

$$\alpha \tau \leq \min_{0 \leq j < i \leq s} (p_{ij}/q_{ij}),$$

with convention $p_{ij}/0 = +\infty$ for $p_{ij} \geq 0$. For explicit methods, having all $q_i = 0$, condition (4.10) can be omitted.

Proof. If $s = 1$ the proof follows from the results obtained in Section I.7.3 for the explicit and implicit Euler method. For larger s the proof follows by induction with respect to i in (4.11). \square

Example 4.7 With this theorem we regain the results on nonlinear positivity given in Section I.7.3 for the θ -method and the explicit trapezoidal rule. Let us here again consider the implicit midpoint rule (1.10) and its explicit counterpart (1.7). The implicit rule can be written as

$$\begin{aligned} w_{n+\frac{1}{2}} &= w_n + \frac{1}{2} \tau F(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}), \\ w_{n+1} &= w_{n+\frac{1}{2}} + \frac{1}{2} \tau F(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}). \end{aligned} \quad (4.12)$$

The first backward Euler stage gives unconditional positivity. However, the second stage consists of a forward Euler step with $\frac{1}{2}\tau$ as (internal) step size, leading to the restriction $\tau\alpha \leq 2$.

For the explicit midpoint rule (1.7), the second stage fits in the format (4.11) by writing it as

$$w_{n+1} = (1 - \theta) w_n - \frac{1}{2} \theta \tau F(t_n, w_n) + \theta w_{n+\frac{1}{2}} + \tau F(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}})$$

with $\theta \in \mathbb{R}$ arbitrary and $w_{n+\frac{1}{2}} = w_n + \frac{1}{2}\tau F(t_n, w_n)$. However, here we cannot achieve a form (4.11) with a positive threshold. In fact, we already saw for the linear inhomogeneous system (4.7) that it is not possible to have positivity if only (4.9) and (4.10) are assumed; additional regularity assumptions are required. \diamond

Example 4.8 It was shown by Kraaijevanger (1986, 1991) that for explicit s -stage Runge-Kutta methods

$$p = 2 \quad \Rightarrow \quad \gamma_R \leq s - 1, \quad (4.13)$$

and the optimal value $\gamma_R = s - 1$ for the interval of absolute monotonicity is attained by the polynomial

$$R(z) = \frac{1}{s} + \frac{s-1}{s} \left(1 + \frac{z}{s-1}\right)^s. \quad (4.14)$$

Moreover, the corresponding Runge-Kutta method (1.1) is given by the coefficients

$$\alpha_{ij} = (s-1)^{-1}, \quad b_i = s^{-1} \quad \text{for } 1 \leq i \leq s, \quad 1 \leq j < i.$$

This method is nonlinearly positive under assumption (4.9) for $\alpha\tau \leq s-1$. This positivity result follows immediately by writing the method in the form (4.11) with $v_0 = w_n$,

$$\begin{aligned} v_j &= v_{j-1} + \frac{1}{s-1} \tau F\left(t_n + \frac{j-1}{s-1} \tau, v_{j-1}\right), \quad j = 1, \dots, s-1, \\ w_{n+1} &= \frac{1}{s} w_n + \frac{s-1}{s} \left(v_{s-1} + \frac{1}{s-1} \tau F(t_{n+1}, v_{s-1})\right). \end{aligned} \quad (4.15)$$

This form is also convenient for actual implementation in view of its low storage demand.

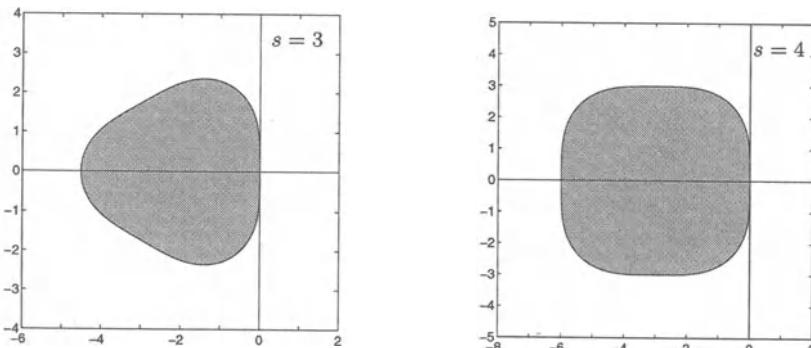


Fig. 4.1. Stability regions S for the second-order s -stage Runge-Kutta methods (4.15) with $s = 3, 4$.

The stability regions for $s = 3, 4$ are illustrated in Figure 4.1. The apparent s -fold symmetry of these regions can be deduced from formula (4.14) with $\zeta = 1 + z/(s - 1)$. The method with $s = 3$ has been used by Gerisch & Weiner (2003) for explicit integration of flux-limited advection semi-discretizations within operator splittings, where positivity is important. These issues will be discussed in later chapters. \diamond

To conclude this section, we recall that for Heun's explicit third-order method (1.8.a) we have an empty positivity interval for inhomogeneous linear systems (4.7). Obviously this method cannot be written in the form (4.11) with non-negative coefficients p_{ij}, q_{ij} . The same holds for the classical explicit fourth-order Runge-Kutta method, although this method has a non-empty interval for linear positivity. A proof for the non-existence of coefficients $p_{ij}, q_{ij} \geq 0$ is given in Gottlieb & Shu (1998); this also follows in a roundabout way from the contractivity results of Kraaijevanger (1991). Related general results on nonlinear positivity of Runge-Kutta methods can be found in Horváth (1998). For higher-order methods (with $p = 3, s = 3, 4, 5$ and $p = 4, s = 5$) with optimal nonlinear positivity properties we refer to Kraaijevanger (1991, Sect. 9) and Spiteri & Ruuth (2002). Finally we mention a recent result of Ferracina & Spijker (2003), showing that with an optimal choice of the parameters p_{ij}, q_{ij} and q_i in Theorem 4.6, the resulting step size restriction is not only sufficient but also necessary.

4.3 Positivity for Multistep Methods

In the fundamental paper of Bolley & Crouzeix (1978) also positivity results were derived for linear multistep methods applied to linear problems with arbitrary non-negative starting values.

In the following we consider the linear multistep method (3.1) with $\alpha_k > 0, \beta_k \geq 0$. For linear systems $w'(t) = Aw(t)$, the method can be written as

$$w_{n+k} = \sum_{j=0}^{k-1} r_{k-j}(\tau A) w_{n+j}, \quad r_i(z) = -\frac{\alpha_{k-i} - z\beta_{k-i}}{\alpha_k - z\beta_k}, \quad (4.16)$$

see also the form (3.24) with companion matrix R . It easily follows that the functions r_i are all absolutely monotonic on the interval $[-\gamma_R, 0]$ with

$$\gamma_R = \min_{0 \leq j \leq k-1} (-\alpha_j/\beta_j) \quad \text{if } \alpha_j \leq 0, \beta_j \geq 0, j = 1, \dots, k-1, \quad (4.17)$$

where we use again the convention $a/0 = +\infty$ for $a \geq 0$. We could define $\gamma_R = 0$ if all $\alpha_j \leq 0$ but some $\beta_j < 0$ ($1 \leq j \leq k-1$), but we will only consider methods with $\gamma_R > 0$. Consequently, $r_i(\tau A) \geq 0$ whenever $A \in \mathcal{M}_\alpha$ and $\alpha\tau \leq \gamma_R$. It thus follows that method (4.16) will give linear positivity under this step size restriction for arbitrary starting values $w_0, \dots, w_{k-1} \geq 0$.

The maximal size of the threshold factor γ_R for explicit k -step methods of order p has been studied by Lenferink (1989). For explicit methods of order $p = 1$ we have $\gamma_R \leq 1$, which is already attained by Euler's method. For explicit methods with $k \geq 2$ Lenferink showed that

$$\gamma_R \leq \frac{k-p}{k-1}. \quad (4.18)$$

With implicit multistep methods the result of Theorem 4.3 again applies, see Bolley & Crouzeix (1978). For implicit methods of order $p \geq 2$ we have $\gamma_R \leq 2$, see Lenferink (1991), where the optimal $\gamma_R = 2$ is attained by the trapezoidal rule. Hence, as for Runge-Kutta methods, the requirement of positivity does place a severe step size restriction on implicit multistep methods.

Generalization of the above linear positivity result to nonlinear systems is directly obtained by writing the method as

$$w_{n+k} - q_k \tau F(t_{n+k}, w_{n+k}) = \sum_{j=0}^{k-1} (p_j w_{n+j} + q_j \tau F(t_{n+j}, w_{n+j})) \quad (4.19)$$

with $p_j = -\alpha_j/\alpha_k$ and $q_j = \beta_j/\alpha_k$. If $\gamma_R > 0$ we have $p_j, q_j \geq 0$, and then the terms on the right are scaled forward Euler steps. The left-hand side is similar to a backward Euler step. Thus, assuming (4.9), (4.10), we again obtain positivity whenever $\alpha\tau \leq \gamma_R$. The idea of writing the method as a combination of Euler steps was employed by Shu (1988) for a related monotonicity property (TVD) that will be discussed in the next chapter.

Example 4.9 Optimal multistep methods were constructed by Shu (1988) and Lenferink (1989, 1991). The explicit 3-step method

$$w_{n+3} - \frac{3}{4}w_{n+2} - \frac{1}{4}w_n = \frac{3}{2}\tau F(t_{n+2}, w_{n+2}) \quad (4.20)$$

has order $p = 2$ and we have $\gamma_R = \frac{1}{2}$. By the bound (4.18) we know that this value of the threshold factor is optimal for second-order explicit 3-step

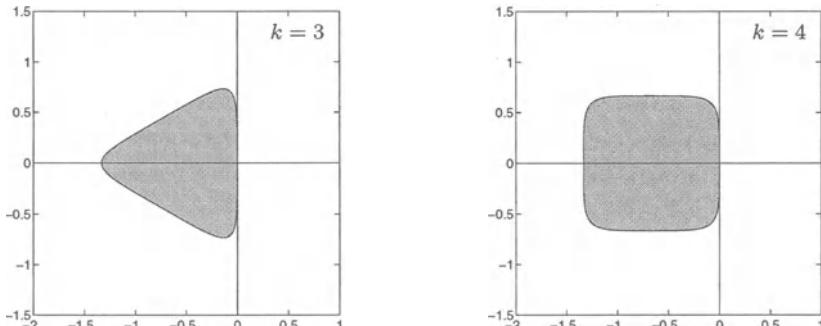


Fig. 4.2. Stability regions \mathcal{S} for the second-order k -step methods (4.20), (4.21).

methods. Similarly, the second-order 4-step method

$$w_{n+4} - \frac{8}{9}w_{n+3} - \frac{1}{9}w_n = \frac{4}{3}\tau F(t_{n+3}, w_{n+3}) \quad (4.21)$$

has threshold factor $\gamma_R = \frac{2}{3}$. The stability regions of these methods are displayed in Figure 4.2. Higher-order methods of this type can be found in the above references as well as in the review paper of Gottlieb, Shu & Tadmor (2001). \diamond

Inclusion of Starting Procedures

The conditions in the above for positivity with linear multistep methods are very restrictive. None of the standard Adams or BDF methods of Section 3.2 has $\gamma_R > 0$, due to the requirement $\alpha_j \leq 0, \beta_j \geq 0$ for $j = 1, \dots, k-1$. For example, with the well-known implicit BDF2 method

$$w_{n+2} = \frac{4}{3}w_{n+1} - \frac{1}{3}w_n + \frac{2}{3}\tau F(t_{n+2}, w_{n+2})$$

it is seen immediately that one does not have $w_2 \geq 0$ for arbitrary $w_0, w_1 \geq 0$, due to the negative coefficient $-\frac{1}{3}$. In fact, this already happens for the trivial equation $w'(t) = 0$. Of course, for this trivial equation it does not make sense to consider arbitrary w_0, w_1 since only the choice $w_1 = w_0$ is consistent. By including starting procedures in the considerations more realistic step size restrictions for positivity can be obtained.

Example 4.10 The BDF2 method is positive for $\alpha\tau \leq \frac{1}{2}$ provided w_1 is computed from w_0 by a suitable starting procedure, say by the backward Euler method $w_1 = w_0 + \tau F(t_1, w_1)$. We illustrate this for the linear inhomogeneous system (4.7) with non-negative source term and with $(I - \theta\tau A)^{-1} \geq 0$ whenever $\theta, \tau > 0$.⁹⁾ References to more general results are given below.

Instead of a recursion for the w_n , we consider the equivalent recursion in terms of w_n and $v_n = 2w_n - w_{n-1}$,

$$(I - \frac{2}{3}\tau A) v_{n+2} = \frac{2}{3}v_{n+1} + \frac{1}{3}(w_{n+1} + 2\tau Aw_{n+1}) + \frac{4}{3}\tau g(t_{n+2}), \quad n \geq 0,$$

with the starting formula

$$(I - \tau A) v_1 = w_0 + \tau Aw_0 + 2\tau g(t_1)$$

obtained from backward Euler. Using $w_0 \geq 0$, it is seen that $v_1 = 2w_1 - w_0 \geq 0$ if $\alpha\tau \leq 1$. By induction to $n = 0, 1, \dots$, it now follows that $2w_{n+2} \geq w_{n+1} \geq 0$ under the step size restriction $\alpha\tau \leq \frac{1}{2}$. \diamond

⁹⁾ The arguments used here originate from discussions with M. van Loon (1996, private communications) for chemical systems.

For the more general class of BDF2-type methods, with parameter $\theta \geq 0$,

$$\frac{3}{2}w_{n+2} - 2w_{n+1} + \frac{1}{2}w_n = \theta\tau F_{n+2} + 2(1-\theta)\tau F_{n+1} - (1-\theta)\tau F_n, \quad (4.22)$$

similar linear results were derived in Hundsdorfer (2001). For $\theta = 1$ and $\theta = 0$ we regain the implicit BDF2 method (3.11) and its explicit counterpart (3.13), respectively. Method (4.22) has order two and the method is A -stable whenever $\theta \geq \frac{3}{4}$. If $\theta = \frac{3}{4}$ the stability region consists precisely of the left half complex plane; in fact this method can then be shown to be equivalent to the implicit trapezoidal rule, in the sense of Dahlquist (1975).

By considering suitable linear combinations $v_n = w_n - \epsilon w_{n-1}$ with $\epsilon = \epsilon(\theta) \in [\frac{1}{3}, 1]$, it can be shown that if $w_0 \geq 0$ and $w_1 - \epsilon w_0 \geq 0$, the method is positive for linear systems under the step size restriction $\alpha\tau \leq \gamma_R(\theta)$ with function $\gamma_R(\theta)$ displayed in Figure 4.3; the precise analytical form can be found in Hundsdorfer (2001). The requirement $w_1 - \epsilon w_0 \geq 0$ can be viewed as a condition on the starting procedure. We have

$$\gamma_R(0) = \frac{5}{8}, \quad \gamma_R\left(\frac{3}{4}\right) = 2, \quad \gamma_R(1) = \frac{1}{2}.$$

Hence the threshold for the explicit method, $\theta = 0$, is better than with the fully implicit method $\theta = 1$. Further note the peculiar discontinuity in $\theta = \frac{3}{4}$, where we have $\gamma_R(\frac{3}{4}) = 2$ and $\gamma_R(\theta) \rightarrow 1$ as $\theta \downarrow \frac{3}{4}$.

A similar result can be derived for the class of 2-step Adams type methods

$$w_{n+2} - w_{n+1} = \theta\tau F_{n+2} + \left(\frac{3}{2} - 2\theta\right)\tau F_{n+1} + \left(\theta - \frac{1}{2}\right)\tau F_n. \quad (4.23)$$

Here the method is A -stable if $\theta \geq \frac{1}{2}$. If $\theta = \frac{1}{2}$ the method reduces again to the implicit trapezoidal rule. The corresponding threshold function $\gamma_R(\theta)$ is also illustrated in Figure 4.3. Here we have $\gamma_R(0) = \frac{4}{9}$.

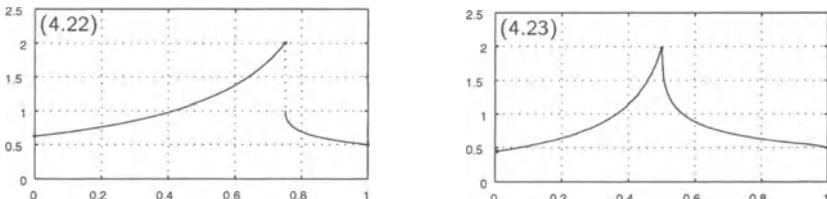


Fig. 4.3. Threshold functions $\gamma_R(\theta)$ versus θ , for the 2-step BDF2-type methods (4.22) and the 2-step Adams methods (4.23).

Generalizations to nonlinear problems have been obtained by Hundsdorfer, Ruuth & Spiteri (2003). With nonlinear problems satisfying (4.9), (4.10), the threshold values for the 2-step BDF and Adams type methods remain virtually the same as in Figure 4.3; only for larger values of θ significant differences between the conditions for linear and nonlinear positivity were found. In that

paper also some numerical experiments can be found which give a verification of the curves in Figure 4.3 for the advection problem $u_t + u_x = 0$ using first-order upwind differences in space. Moreover, in a nonlinear test for Burgers' equation $u_t + (u^2)_x = 0$, the standard explicit Adams-Basforth methods and the extrapolated BDF schemes of order two and three turned out to be preferable over some specially constructed explicit methods, like the ones in Example 4.9, for which monotonicity with arbitrary starting values hold.

4.4 Related Monotonicity Results

The term ‘monotonicity’ covers a range of related concepts. In the above we concentrated on positivity since the relevance of this concept is obvious when the unknowns represent concentrations or densities that are to be non-negative by definition. Related concepts include contractivity in arbitrary norms and the total variation diminishing (TVD) property, see LeVeque (1992, Sect. 15) for other concepts of similar type. The TVD property will be discussed in the next chapter, here we briefly mention some relations between positivity and contractivity for linear problems.

The threshold factors γ_R for absolute monotonicity also occur in the study of the contractivity property $\|R(\tau A)\| \leq 1$ of one-step methods if the underlying vector norm is not generated by an inner product, such as the max-norm. Contractivity implies of course power boundedness and hence stability. The contractivity property is relevant for linear systems $w'(t) = Aw(t)$ in which the matrix A satisfies

$$\|e^{tA}\| \leq 1 \quad \text{for all } t \geq 0 .$$

Comprehensive research on contractivity properties has been carried out by Spijker and co-workers. As an example we mention the following result of Spijker (1983):

$$\|A + \alpha I\| \leq \alpha, \quad \alpha\tau \leq \gamma_R \quad \implies \quad \|R(\tau A)\| \leq 1 . \quad (4.24)$$

This can be viewed as a counterpart of Theorem 4.1. The results of Kraaijevanger (1986, 1991) for one-step methods and of Lenferink (1989, 1991) for multistep methods, that we used in the above for positivity, were derived in the framework of contractivity.

In certain situations the relation between positivity and contractivity is even more obvious. It was mentioned already in Section I.7.1 that if $Ae = 0$, with $e = (1, 1, \dots, 1)^T \in \mathbb{R}^m$, then positivity implies a maximum principle and thus also L_∞ -contractivity. Similarly, if $e^T A = 0$ then positivity is easily shown to imply L_1 -contractivity, and vice versa. Notice that the latter assumption $e^T A = 0$ is directly related to mass conservation.

5 Variable Step Size Control

In modern practice, ODE problems are solved by means of codes which integrate with variable step sizes. Users have to specify a tolerance and norm, and then the code automatically adjusts τ to the local variation in the computed solution to meet a certain local error criterion in that norm. This approach leads to smaller (larger) τ in regions of rapid (resp. slow) variation and normally results in efficient computations, in terms of CPU versus accuracy. Complicated PDE problems are still often solved with constant τ . However, through Method of Lines software, variable step sizes gain popularity. In this section we will therefore briefly discuss the basic ideas behind variable step sizes. For a more comprehensive treatment we refer to Hairer et al. (1993, 1996) and Shampine (1994).

5.1 Step Size Selection

Consider an attempted step from t_n to $t_{n+1} = t_n + \tau_n$ with step size τ_n . Suppose the method has order p and we have an available estimate D_n for the norm of the local error, together with a tolerance Tol which is specified by the user. Roughly speaking, variable step sizes are based on the following rule: if $D_n \leq Tol$ this step is accepted and the next step size τ_{n+1} will be slightly increased, whereas if $D_n > Tol$ the step is rejected and redone with a smaller step size τ_n . In both cases the new step size is chosen such that D_n (with rejection) or D_{n+1} (upon acceptance) will be close to Tol .

To be more specific, suppose a smooth error function $C(t)$ exists such that

$$D_n = C(t_n) \tau_n^{\tilde{p}+1} + \mathcal{O}(\tau_n^{\tilde{p}+2}),$$

with $\tilde{p} \leq p$. Here $\tilde{p} = p$ if D_n is an estimate of the genuine local error of the method, but often the estimate D_n is quite rough and \tilde{p} may be less than p . For example, the estimate may be obtained by comparing w_{n+1} with the result \tilde{w}_{n+1} obtained from a lower-order method, say of order $p-1$, in which case $\tilde{p} = p-1$. Having the estimate D_n two cases can occur: $D_n > Tol$ or $D_n \leq Tol$. In the first case we decide to reject this step and to redo it with a smaller step size τ_{new} , where we aim at $D_{new} = Tol$. With τ_{new} we get

$$D_{new} = C(t_n) \tau_{new}^{\tilde{p}+1} + \mathcal{O}(\tau_{new}^{\tilde{p}+2}).$$

Neglecting higher-order terms and using the available estimate D_n enables us to eliminate the unknown $C(t_n)$ giving

$$\tau_{new} = r \tau_n, \quad r = (Tol / D_n)^{1/(\tilde{p}+1)}. \quad (5.1)$$

In the second case, $D_n \leq Tol$, we decide to accept the step and to continue integration from t_{n+1} to t_{n+2} with a new step size τ_{n+1} . There holds

$$D_{n+1} = C(t_{n+1}) \tau_{n+1}^{\tilde{p}+1} + \mathcal{O}(\tau_{n+1}^{\tilde{p}+2}),$$

and the aim is to choose τ_{n+1} such that $D_{n+1} \approx Tol$. Neglecting higher-order terms and using the fact that $C(t_{n+1}) = C(t_n) + \mathcal{O}(\tau_n)$, then yields again the expression (5.1) for $\tau_{n+1} = \tau_{new}$.

Because estimates are used and additional control on decrease and increase of step sizes is desirable, the expression for the new trial step size found in most codes has the form

$$\tau_{new} = \min(r_{max}, \max(r_{min}, \vartheta r)) \tau_n, \quad (5.2)$$

where r_{max} and r_{min} are a maximal and minimal growth factor, respectively, and $\vartheta < 1$ serves to make the estimate conservative so as to avoid repeated rejections. The choices for these parameters depend somewhat on the type of integration method; one-step methods are more flexible in this respect than multistep methods. Typical values are $\vartheta \in [0.7, 0.9]$, $r_{min} \in [0.1, 0.5]$ and $r_{max} \in [1.5, 10]$.

A starting step size for the first step on scale with the initial solution variation must be prescribed. Sophisticated ODE codes do this automatically. Step sizes can be further constrained by a minimum and maximum value. Chosen values should reflect the smallest and largest time constants of the problem, at least if one wishes to resolve all scales in the solution. The maximum can also be used to prevent conditionally stable codes from taking step sizes beyond stability limits. If a stability limit is available it is recommended to impose a maximum since this enhances robustness. For implicit methods an additional decrease in step size might be imposed if the convergence rate of the Newton process is deemed insufficient. Finally we note that in most codes one has to specify an absolute tolerance Tol_A and a relative tolerance Tol_R , which are then combined to give a total tolerance such as $Tol = Tol_A + Tol_R \|w_n\|$, for example.

Local Error Estimation and Asymptotic Expansions

In practice, estimates D_n for the local errors are often found on heuristical grounds. There is some relevant theory on which estimators can be based. This will be briefly illustrated here for the backward Euler method. For notational convenience the differential equation $w'(t) = F(w(t))$ is assumed to be in autonomous form and we also assume that F is sufficiently smooth. General theoretical results for higher-order methods can be found in Hairer et al. (1993) or Shampine (1994).

First consider the backward Euler method with fixed step size,

$$w_{n+1} = w_n + \tau F(w_{n+1}), \quad n \geq 0, \quad w_0 = w(0).$$

The local error satisfies

$$\delta_n = -\frac{1}{2}\tau^2 w''(t_n) + \mathcal{O}(\tau^3),$$

see (3.4), where it is tacitly assumed that we are dealing with a fixed, non-stiff ODE system. Then it can be shown (Gragg, 1965) that there exists an expansion for the global error of the form

$$w_n = w(t_n) + \tau e_1(t_n) + \tau^2 e_2(t_n) + \dots$$

with principal error function e_1 defined by

$$e'_1(t) = F'(w(t))e_1(t) + \frac{1}{2}w''(t), \quad e_1(0) = 0.$$

Using this expansion, it follows that the local error δ_n can be estimated by

$$d_n = -\frac{1}{2} \left(w_{n+1} - w_n - \tau F(w_n) \right). \quad (5.3)$$

This can be verified directly from

$$\begin{aligned} d_n &= -\frac{1}{2}\tau \left[w' + \frac{1}{2}\tau w'' + \tau e'_1 - F(w) - \tau F'(w)e_1 + \dots \right]_{t=t_n} \\ &= -\frac{1}{2}\tau^2 \left[\frac{1}{2}w'' + e'_1 - F'(w)e_1 + \dots \right]_{t=t_n} = -\frac{1}{2}\tau^2 w''(t_n) + \mathcal{O}(\tau^3). \end{aligned}$$

Remark 5.1 Note that the justification of the estimator (5.3) is based on the global error expansion. The same expressions would have been obtained if we had assumed $w_n = w(t_n)$ with $w_{n+1} = w(t_{n+1}) + \frac{1}{2}\tau^2 w''(t_n) + \mathcal{O}(\tau^3)$. On the other hand, the simple substitution $w_k = w(t_k)$, $k = n, n+1$, in this estimator (5.3) would yield an extra, incorrect, factor $\frac{1}{2}$. \diamond

A similar global error expansion is known to hold also for variable step sizes that are generated by

$$\tau_n = t_{n+1} - t_n = s(t_n)\bar{\tau},$$

where $s(t) \in (0, 1]$ is a smooth function and $\bar{\tau}$ the maximal step size, see for instance Shampine (1994). Notice that for such theoretical step size sequences we have

$$\tau_{n+1} / \tau_n = 1 + \mathcal{O}(\tau_n). \quad (5.4)$$

In practice the step size sequence can have many jumps where (5.4) is not reasonable. Moreover, the above derivation is only valid for non-stiff equations. Still, estimates as derived here are used with success in situations that are not covered by the theory.

Finally we note that the actual local error of the backward Euler method for stiff systems is given by

$$\delta_n = -\frac{1}{2}\tau^2 (I - \tau A_n)^{-1} w''(t_n) + \mathcal{O}(\tau^3)$$

with A_n an integrated Jacobian matrix, see (3.4). This could be approximated by

$$(I - \tau \tilde{A}_n)^{-1} d_n, \quad \tilde{A}_n \approx F'(w_n),$$

with d_n as in (5.3). The solution of a linear system with matrix $(I - \tau \tilde{A}_n)$ can be relatively cheap if an LU -decomposition or good preconditioner is available. Since for the solution of the implicit relation by Newton's method the same matrix is involved, this is often so. If not, one usually settles for the simpler estimate d_n , which gives in practice also good results.

Remark 5.2 Asymptotic error expansions, which have been illustrated here for the backward Euler method, can be derived as well for more general methods, say $w_{n+1} = S_\tau(w_n)$. Such asymptotic expansions are also the basis for *extrapolation methods*. Extrapolation becomes in particular interesting for *(time) symmetric* methods, for which $S_{-\tau} = S_\tau^{-1}$. For such methods we have an expansion of the form

$$w_n = w(t_n) + \tau^2 e_2(t_n) + \tau^4 e_4(t_n) + \dots$$

with only even terms. With one extrapolation step we then gain two orders of accuracy; see for instance Hairer et al. (1993, Sect. II.8 and II.9). Examples of symmetric methods are the implicit midpoint rule and trapezoidal rule. \diamond

5.2 An Explicit Runge-Kutta Example

Variable step sizes are easily implemented for Runge-Kutta and Rosenbrock methods due to their one-step nature. As an example we will provide the explicit trapezoidal rule

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_n + \tau F(t_n, w_n))$$

with a simple step size control and illustrate the resulting solver for Burgers' equation. Since the forward Euler result $w_{n+1}^* = w_n + \tau F(t_n, w_n)$ is available already in this step, we can use

$$D_n = \|w_{n+1} - w_{n+1}^*\| \tag{5.5}$$

as an estimator for the norm of the local error. Notice that for $w_n = w(t_n)$ this estimator satisfies

$$D_n = \frac{1}{2}\tau^2 \|w''(t_n)\| + \mathcal{O}(\tau^3).$$

It is obvious that this estimator is rather crude. Actually, it provides an accurate estimator for the local error of Euler's method, but we will use this nevertheless for the second-order explicit trapezoidal rule. Then for \tilde{p} used in formula (5.1) we have the value 1. By choosing an appropriate norm for computing D_n and by making a choice for (5.2) the step size control can now be used.

Numerical Illustration

As a numerical example we consider Burgers' equation

$$u_t + \frac{1}{2}(u^2)_x = d u_{xx}, \quad 0 < t \leq 1, \quad 0 < x < 1, \quad (5.6)$$

see also (2.22), now using for the diffusion coefficient d the value 10^{-3} . Again we use the exact solution (2.23) with corresponding Dirichlet conditions for test purposes. Figure 5.1 shows this exact solution $u(x, t)$ at $t = 0, 0.5, 1$. Initially u is composed of two right moving fronts of which the upper one has a larger speed. At about $t = 0.5$ the faster front overtakes the slower one and thereafter a single right moving front remains.

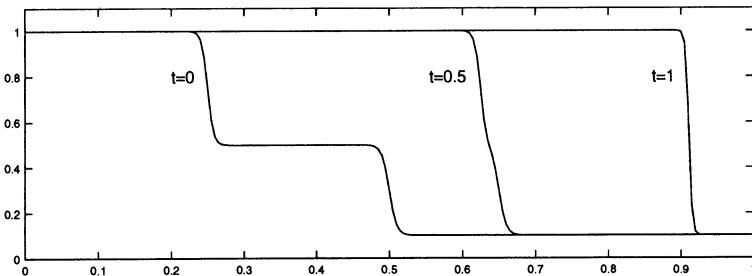


Fig. 5.1. Exact solution $u(x, t)$ of the 1D Burgers equation for $t = 0, 0.5, 1$.

Because d is small we discretize the conservative form (5.6) with the third-order upwind-biased scheme for the nonlinear advection term on a uniform grid with mesh width $h = 1/200$. The diffusion term is discretized with the second-order central scheme. Since Dirichlet boundary conditions are prescribed and $u > 0$, for the first grid point $x_1 = h$ a virtual value at $x_{-1} = -h$ is needed, and this is found by linear extrapolation.

The resulting semi-discrete system $w'(t) = F(w(t))$ is now integrated with the explicit trapezoidal rule provided with the above step size control, using the L_1 -norm,

$$\tau_{new} = \min(1.5, \max(0.5, 0.8 r)) \tau_n$$

and (5.1) with $\tilde{p} = 1$. Since the order is low, only modest values of Tol should be used, say $10^{-2} \leq Tol \leq 10^{-4}$. With very small Tol a higher-order scheme would be much more efficient. On the other hand, in the present test, values of Tol larger than 10^{-2} lead to numerical solutions that drift away from the exact solution and become unstable. This is avoided by smaller values of Tol , but we also could use a norm which is more receptive to local variations such as the L_2 - or L_∞ -norm.

Due to the conditional stability, robustness could be enhanced by imposing a maximum for τ derived from a stability estimate. For the test model $u_t + au_x = 0$ with a constant, the combination explicit trapezoidal rule

and third-order upwind-biased spatial discretization gives the CFL condition $\tau \leq 0.87 h / |a|$, see Table 1.2. From this CFL condition a stability estimate for the nonlinear Burgers equation $u_t + uu_x = du_{xx}$ is obtained by neglecting the diffusion term (d is sufficiently small to allow this) and by ‘freezing’ the velocity $a = u$ in $u_t + uu_x = 0$ to obtain the test model. The maximal velocity $a = 1$ then should be substituted, where we use the a priori knowledge that $0 \leq u(x, t) \leq 1$. With $h = 1/200$ this leads to the CFL bound

$$\tau \leq 0.87 h / |a| \approx 4.4 \cdot 10^{-3}.$$

For $Tol = 10^{-2}$ and $\tau_0 = 10^{-4}$ as starting step size, Figure 5.2 shows the numerical results at times $t = 0.5, 1$ with solid lines for the numerical solution and dashed grey lines for the exact solution; these dashed lines are plotted on top of the solid ones to make them visible. This run took 181 steps with 3 rejections. In the small frame the step sizes τ_n are plotted versus the step number n , and rejected steps are indicated by *-marks. We see that step sizes are selected beyond the linear CFL limit $4.4 \cdot 10^{-3}$, indicated by dashes, which results in a zigzag pattern for τ . Still the accuracy is quite good with a global error at time $t = 1$ of 0.0026 in the L_1 -norm; some wiggles arise but for most part the numerical and exact solutions coincide visually. It should be noted that these wiggles are not caused by the spatial discretization, but they arise because of local step sizes that are significantly larger than the CFL bound. Observe also that the wiggles occur near the value $u = 1$, because that is where the linear CFL bound will be violated mostly. Finally we note that there are smaller step sizes in the middle and at the end of the integration interval. This is because the output points $t = 0.5$ and $t = 1$ were required to be step points t_n . Such behaviour can be avoided by using interpolation to obtain solutions at prescribed output points.

For a second test we used $Tol = 10^{-3}$ and initial step size $\tau_0 = 0.01$. This gave 186 steps, with 1 rejection in the beginning because the initial step was too large. The numerical solutions in Figure 5.3 are now close to plot-accuracy and the wiggles have disappeared. The total global error here is close to the

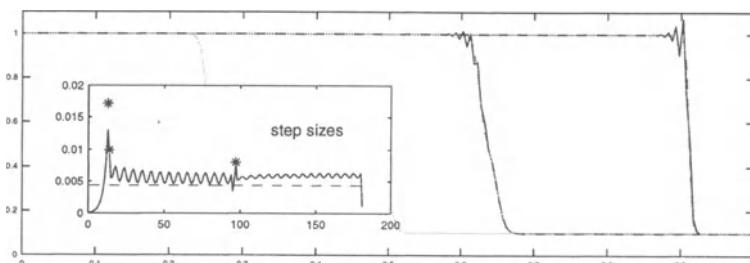


Fig. 5.2. Solution of the Burgers equation for $t = 0.5, 1.0$ with $Tol = 10^{-2}$. Exact solutions are dashed grey. In small frame: step sizes τ_n versus step number n .

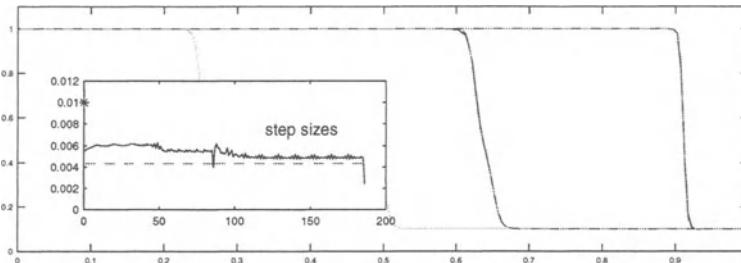


Fig. 5.3. Solution of the Burgers equation for $t = 0.5, 1.0$ with $Tol = 10^{-3}$. Exact solutions are dashed grey. In small frame: step sizes τ_n versus step number n .

spatial error. A very similar solution is also obtained with $Tol = 10^{-2}$ when the CFL bound $\tau_n \leq 4.4 \cdot 10^{-3}$ is enforced.

From this experiment we can conclude that the step size control is able to detect the instabilities for this conditionally stable scheme, and in this way numerical solutions are obtained without a priori knowledge of the solution. On the other hand, if a reliable stability estimate is available, then this can be used to enhance the robustness of the variable step size code. In general, the code should be run with different values of Tol to obtain confidence in the quality of the solutions.

Remark 5.3 The selection of a pair of formulas (w_{n+1}^*, w_{n+1}) such that w_{n+1}^* is just one order lower ($\tilde{p} = p - 1$) and available at little costs is standard practice nowadays. Then, as in the above example, $D_n = \|w_{n+1} - w_{n+1}^*\|$ does not necessarily estimate the genuine local error of w_{n+1} . Instead, the local error of the lower-order method serves as estimator. Most Runge-Kutta solvers use such formula pairs for step size control.

A clarifying discussion is given in Shampine (1994, Sect. 7.4), where the above control is called error per step control with local extrapolation (case XEPS) and where it is shown that this control asymptotically yields tolerance proportionality: the global error behaves like $w(t_n) - w_n \sim Tol$ for $Tol \rightarrow 0$.

If we use an estimator with $\tilde{p} = p$, this proportionality will not hold since we then bound approximations to the local error which will add up to the global error. Proportionality can be achieved by requiring that $D_n \leq (\tau_n/T) Tol$, which is often called error control per unit step. This control, however, may lead to difficulties if the solution is non-smooth because then the requirement on D_n may be impossible to fulfil even for $\tau_n \downarrow 0$. ◇

5.3 An Implicit Multistep Example

Genuine multistep methods use information from at least two previous time levels. Consequently, when variable step sizes are used with a multistep method, the formula coefficients need to be adjusted for maintaining the

order of consistency. Consider, as an example, the implicit BDF2 method

$$w_{n+1} = \frac{4}{3}w_n - \frac{1}{3}w_{n-1} + \frac{2}{3}\tau F(t_{n+1}, w_{n+1}),$$

introduced in Example 3.3. We denote here the most advanced time level by t_{n+1} , to have a closer resemblance in notation with one-step methods.

With variable steps, let $\tau_n = t_{n+1} - t_n$ with step size ratio $r = \tau_n/\tau_{n-1}$. The variable step size version of the BDF2 method, with the coefficients adjusted for maintaining second-order consistency, then reads

$$w_{n+1} = \frac{(1+r)^2}{1+2r} w_n - \frac{r^2}{1+2r} w_{n-1} + \frac{1+r}{1+2r} \tau_n F(t_{n+1}, w_{n+1}). \quad (5.7)$$

The consistency is easily verified by inserting the exact solution and using Taylor expansion, see Section 3.1. Letting $\tau_n, \tau_{n-1} \rightarrow 0$ yields a residual in the right-hand side of

$$\tau_n \rho_n = -\frac{(1+r)^2}{6r(1+2r)} \tau_n^3 w'''(t_n) + \mathcal{O}(\tau_n^4), \quad (5.8)$$

in which ρ_n is the truncation error of the variable step size formula.

Providing (5.7) with step size control can be done in a manner similar to the previous example. A low-order estimator is given by

$$D_n^* = \frac{r}{1+r} \| w_{n+1} - (1+r)w_n + r w_{n-1} \| \approx \frac{1}{2} \tau_n^2 \| w''(t_n) \|, \quad (5.9)$$

which approximates the local error of the implicit Euler (BDF1) method. Alternatively, we can use

$$D_n = \frac{1+r}{1+2r} \| w_{n+1} + (r^2 - 1)w_n - r^2 w_{n-1} - (1+r)\tau_n F(t_n, w_n) \| \quad (5.10)$$

to approximate the norm of (5.8) in a heuristic manner (not based on an asymptotic expansion). For the first step, $n = 0$, we will use the implicit Euler method to calculate $w_1 = w_0 + \tau_0 F(t_1, w_1)$ with

$$D_0^* = \frac{1}{2} \| w_1 - w_0 - \tau_0 F(t_0, w_0) \| \approx \frac{1}{2} \tau_0^2 \| w''(t_0) \| \quad (5.11)$$

as initial estimator, see (5.3). In Section 6 both estimators (5.9) and (5.10) will be used in a numerical test.

For higher-order multistep methods similar implementations can be used, but derivation of suitable coefficients becomes more complicated. Much effort has been spent on code development, and with great success, especially for BDF methods for stiff ODEs and Adams methods for non-stiff ODEs. There exist efficient high-order, variable step size BDF and Adams codes which are self-starting and even provided with variable order mechanisms to optimize their performance. Some of these codes that are freely available will be discussed next.

5.4 General Purpose ODE Codes

For the numerical solution of ODE systems many efficient codes have been developed. Here we briefly mention some of these codes that are available for stiff and for non-stiff equations. All of the codes are applicable to ODE systems $w'(t) = F(t, w(t))$ in \mathbb{R}^m . Some of the stiff solvers are also applicable to more general systems, like $Mw'(t) = F(t, w(t))$ with matrix M possibly singular, so that algebraic constraints can be incorporated. Systems of such type are called *differential algebraic equations* (DAEs). The numerical solution of DAE systems is reminiscent of the numerical solution of stiff ODEs. To a great extent, however, the subject stands on its own and has been explored extensively in the last two decades. Introductions to the subject are found in Brenan, Campbell & Petzold (1989), Hairer & Wanner (1996, Chap. VII) and Ascher & Petzold (1998).

Runge-Kutta codes: For non-stiff problems the code DOPRI5 of Hairer, Nørsett & Wanner (1993) has become quite popular. DOPRI5 is based on explicit Runge-Kutta methods from Dormand & Prince (1980). The method has order five and it is equipped with an embedded formula of order four for error control. For very high accuracy the code DOP853 of order eight is also available; see Hairer et al. (1993) for details and related codes.

For stiff problems, there is the implicit Runge-Kutta code RADAU5 of Hairer & Wanner (1996). It uses the fully implicit L -stable 3-stage Runge-Kutta method of order five, based on collocation with Radau quadrature, see Example 1.4. Error control is based on an embedded fourth-order formula. See Hairer & Wanner (1996) for details and tests on a large set of stiff ODE problems.

In these tests also the linearly implicit Rosenbrock code RODAS showed good results, in particular for accuracies not too high. This code is based on a Rosenbrock method of order four with an embedded third-order formula for error control. Compared to the Radau codes, this one often gives faster results for the somewhat lower – by ODE standards – accuracy ranges. See Hairer & Wanner (1996) for details.

The source code for these Runge-Kutta and Rosenbrock codes can be obtained from

<http://www.unige.ch/math/folks/hairer/software.html>

Other efficient and well-documented codes for stiff and non-stiff problems are found in the software collection RKSUITE, which is described in Brankin, Gladwell & Shampine (1992). The source code can be obtained from

<http://www.netlib.org/ode/index.html>

Multistep codes: Variable-order Adams methods, for non-stiff problems, and implicit BDF schemes, for stiff problems, are all incorporated in the code VODE of Brown, Byrne & Hindmarsh (1989). This code and related ones, like LSODE, were the result of a long development, which started with Gear (1971). Apart from step sizes, also the orders are selected during the time

integration, based on efficiency considerations of nearby formulas; see, for example, Shampine (1994) for details. A variant, that will be used in Chapter IV, is VODPK (Byrne, 1992) where the Krylov method GMRES is used for the solution of linear systems. Other well-known solvers employing BDF formulas, also suited for DAE systems, are DASSL (Petzold, 1982), described in detail in Brenan et al. (1989), and SPRINT (Berzins & Furzeland, 1986). The latter package is included in the NAG software library. The source code for the other multistep codes can be obtained from

<http://www.netlib.org/ode/index.html>

Comparisons: General recommendations are difficult. Each of the above codes has a specific problem class where it outperforms the others if comparable implementations (in particular for the linear algebra) are used. Extensive tests for many ODE systems can be found in Hairer et al. (1993, 1996) and Lioen et al. (1996). These tests also cover some PDEs, after spatial discretization with finite differences. It should be noted, however, that the main accent in these tests was on high accuracy in time, and it is not so clear whether the spatial error is sufficiently small to justify these high accuracies from the PDE point of view. Comparisons for PDE problems, in particular for multi-dimensional ones, are not easy since the choice of proper linear algebra subroutines and good preconditioners will then often be the decisive factor for efficiency. Moreover, storage demands then may also become of prime importance and for that reason very simple schemes (implicit Euler, Crank-Nicolson) are still often used. However, for many problems it will be beneficial to use schemes with higher order.

6 Numerical Examples

In this section we present some numerical results on time integration for two examples with one-dimensional PDEs. The first one is parabolic with local discontinuities and is used to illustrate local step size selections and temporal error estimates. The methods used here are the implicit Euler method and the BDF2 scheme. In the second example we consider the hyperbolic Schrödinger equation to illustrate the performance of some higher-order methods, for which we have chosen several Gauss and Radau Runge-Kutta formulas.

6.1 A Model for Antibodies in Tumorous Tissue

As a first numerical test example we consider the parabolic problem

$$u_t = u_{ss} - \kappa uv, \quad v_t = -\kappa uv,$$

for time $t \in (0, T]$ and space variable $s \in (0, \infty)$, with initial conditions $u(s, 0) = 0$, $v(s, 0) = 1$ and boundary condition

$$u(0, t) = \begin{cases} 2 & \text{if } 0 < t \leq 5, \\ 0 & \text{otherwise.} \end{cases}$$

This is a simple model for the penetration of radio-labeled antibodies with concentration u into a tumorous tissue with concentration v . A theoretical analysis and background references can be found in Hilhorst et al. (1996). For the numerical solution we follow Lioen et al. (1996) where the semi-discrete problem was used as a test example for ODE solvers. Let

$$x = \frac{s}{s+4},$$

so that the infinite domain $s \in (0, \infty)$ is transformed to $x \in (0, 1)$. The differential equation then becomes

$$u_t = \frac{1}{16}(1-x)^4 u_{xx} - \frac{1}{8}(1-x)^3 u_x - \kappa u v, \quad v_t = -\kappa u v. \quad (6.1)$$

Note that at $x = 1$ the spatial operators vanish, and hence no additional boundary conditions are required at this point. The transformation introduces an advective term, but the diffusion is still dominating.

Illustrations of the solutions are given in the Figures 6.1 and 6.2 with reaction constant $\kappa = 100$ and $T = 12$. The problem has discontinuities at $(x, t) = (0, 0)$ and $(0, 5)$ whereas for $t = T$ the solution has almost reached a steady state. Therefore this problem is well suited for variable time stepping, where the local time steps should be small for $t = 0, 5$ and large for t approaching the final time T .

The equation (6.1) is discretized in space using standard second-order central differences on a uniform grid with grid points $x_j = jh$, $1 \leq j \leq m$, $h = 1/m$. We will take $m = 200$, which gives a spatial L_2 -error of $3 \cdot 10^{-4}$ approximately. The value 10^{-4} will be considered as a lower bound for relevant temporal accuracies.

For the time discretization we consider variable step size implementations of the backward Euler method (BDF1) and the implicit BDF2 method (5.7),

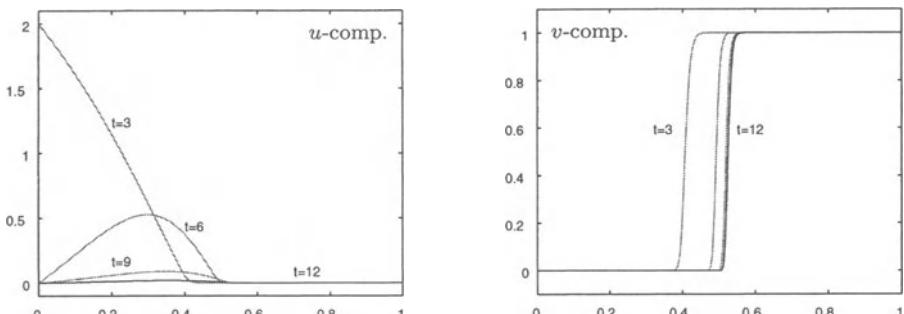


Fig. 6.1. Time evolution for (6.1): snapshots of u and v for $x \in [0, 1]$ at time $t = 3, 6, 9$ (grey) and $t = T = 12$.

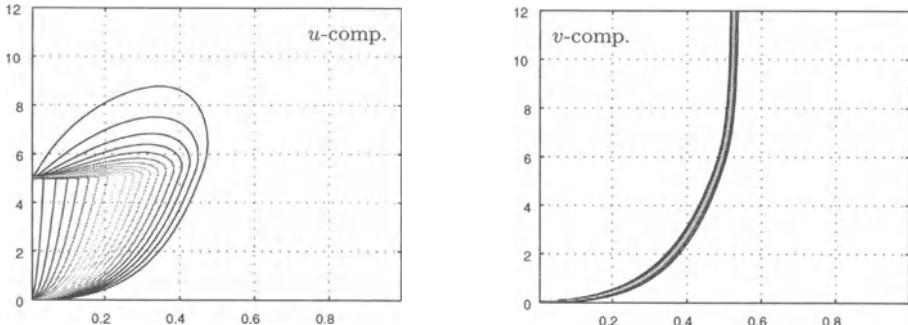


Fig. 6.2. Time evolution for (6.1): contour plots of u and v with time $0 \leq t \leq 12$ vertical and space $0 \leq x \leq 1$ horizontal. Contours on levels $0.1 \cdot j$, $j \in \mathbb{N}$.

which also starts with BDF1, with initial time step $\tau_0 = 10^{-2}$. As local error estimators we use (5.9), $\tilde{p} = 1$, for both methods, and for the BDF2 method also the estimator (5.10), $\tilde{p} = 2$. In the initial step, estimator (5.11) is employed. Further we use (5.2) with $r_{max} = 1.5$, $r_{min} = 0.5$ and safety factor $\vartheta = 0.8$, and the underlying norm was taken as the L_2 -norm. The actual step size behaviour for both estimators is rather similar in this test. Typical step size sequences are displayed in Figure 6.3 for both estimators with $Tol = 10^{-3}$. Note that the final step in the right plot shows a slight decrease in step size because T is required to be a step point. Rejected steps are indicated by *-marks. The number of rejections at the discontinuities could be reduced by using a smaller value for r_{min} , but the aim of this test is to illustrate some simple step size strategies, not to optimize them for this particular problem.

For an additional comparison we also use the very simple error estimator

$$D_n = \|w_{n+1} - w_n\|_2 \approx \tau_n \|w'(t_n)\|_2 \quad (6.2)$$

with $\tilde{p} = 0$. The results are given in Figure 6.4 where the absolute global errors at the final time $T = 12$ are plotted as function of $1/N$ with N being the total number of time steps, that is, rejected plus accepted steps. The tolerance Tol was taken as 10^{-j} , $j \geq 0$, but the results are only displayed for global errors $\geq 10^{-4}$ since this gives a relevant temporal error in comparison to the

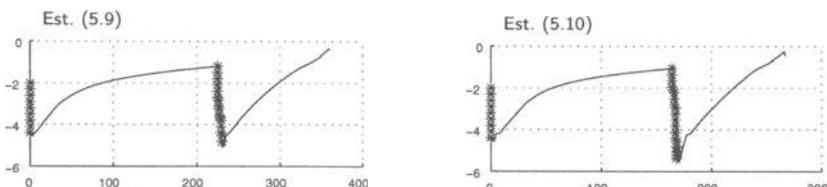


Fig. 6.3. Plots of $\log_{10}(\tau_n)$ versus step number n for the BDF2 method with $Tol = 10^{-3}$ and estimators (5.9), (5.10).

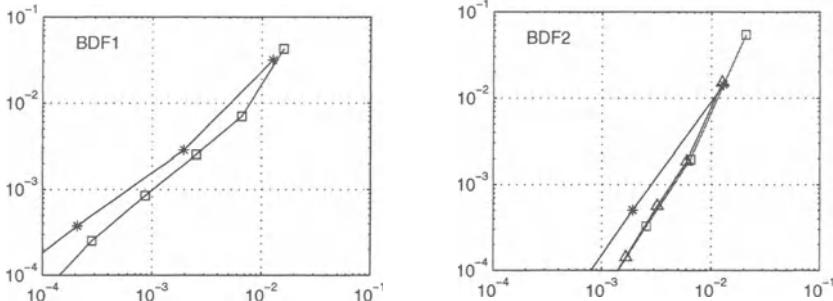


Fig. 6.4. Errors versus $1/N$ for BDF1 and BDF2 methods, using the error estimators (6.2) (*-marks) and (5.9) (□-marks), and the estimator (5.10) (△-marks) for BDF2.

spatial error. We note that the actual number of steps for a given Tol is very dependent on the choice of estimator, but in terms of efficiency the results are close. The Newton iteration for the implicit relations was terminated if the L_2 -displacement was less than $Tol/10$. This did give usually 2 iterations and therefore N can be considered as a fair measure for the amount of work.

We see from Figure 6.4 that the actual estimator is not crucial in this test. All estimators give a sufficient step size reduction near the discontinuities, although the simple estimator (6.2) is somewhat less efficient than the others. Still, the most important observation is the fact that all estimators give satisfactory results for this problem, for which fixed step sizes would be inefficient due to the strong variations in temporal smoothness.

In this test we used the rather tight interval $[r_{min}, r_{max}] = [0.5, 1.5]$. With wider intervals the step size behaviour becomes a bit more irregular after the discontinuities, but in terms of efficiency (error versus work) the results are similar. Error proportionality in this test is found for BDF1 with (6.2) and BDF2 with (5.9), in accordance with Remark 5.3. Finally we mention that on this problem also the trapezoidal rule was used, with the same step selections as for BDF1, and this gave results comparable in efficiency to the BDF2 results in Figure 6.4.

6.2 The Nonlinear Schrödinger Equation

In Section I.8 of the previous chapter, a soliton collision problem was considered for the nonlinear Schrödinger equation

$$u_t = iu_{xx} + i|u|^2u, \quad 0 < t \leq T = 44, \quad x \in (-20, 80), \quad (6.3)$$

with initial value

$$u(x, 0) = e^{\frac{1}{2}ix} \operatorname{sech}(x/\sqrt{2}) + e^{\frac{1}{20}i(x-25)} \operatorname{sech}((x-25)/\sqrt{2}) \quad (6.4)$$

and homogeneous Neumann boundary conditions $u_x = 0$ at $x = -20, 80$. Recall that $u(x, t)$ is complex scalar and $i = \sqrt{-1}$. For this problem some test results were already presented with time stepping based on the implicit Euler method and the trapezoidal rule and with second-order differences in space. Here some additional methods are considered for fourth-order spatial differences on a fine grid with $m = 800$ points. From Table I.8.1 it is seen that the relative spatial error in the L_2 -norm is approximately $1.5 \cdot 10^{-3}$.

On this fixed grid, several one-step ODE methods have been tested with a sequence of constant step sizes. For this soliton collision problem variable steps would be useful during the actual collision. Before and after, the solitons move at a uniform velocity in the spatial region. In the following only temporal errors are considered; an accurate reference solution was computed with the classical fourth-order explicit Runge-Kutta method using very small time steps.

In Figure 6.5 the results are presented for a number of implicit methods. In these pictures the relative L_2 -errors at time $t = T$ are plotted as function of the inverse of the quantity C , where C is a measure of work or costs. We have chosen C to be the number of mega-flops (1 flop being one floating point operation as measured in MATLAB, version 5). Because very different methods are considered this gives a better measure than the number of time steps or CPU timings. The time steps were taken as $\tau = T/N$ with $N = 25 \cdot 2^k$ for integer values $k \geq 0$. In these plots, horizontal dashed lines are drawn at the error levels 10^0 and $1.5 \cdot 10^{-3}$. Errors larger than 10^0 indicate numerical solutions which are qualitatively wrong. On the other hand, temporal errors less than the spatial error on this grid, $1.5 \cdot 10^{-3}$, are not very relevant either.

For the implicit methods a modified Newton iteration (I.8.1) was used with the choice $A_n = A + i \operatorname{diag}(|w_j|^2)$, as described in Section I.8, where A is the fourth-order difference approximation of $i\partial_{xx}$. For the implicit midpoint rule the slightly different choice $A_n = A$ was made, see below. As stopping criterion for the iteration we used the requirement that the displacement in the L_2 -norm must be less than 10^{-4} (10^{-5} for the higher-order methods). The results were not very sensitive with respect to this criterion.

Low-order implicit methods: The left picture in Figure 6.5 contains the results for the following low-order implicit methods:

- BE, the backward Euler method. The results for this method are very disappointing. Due to the inaccuracy and damping of the backward Euler method very small step sizes have to be taken. In the plot, results in the range $100 \leq N \leq 12800$ are represented and only for a larger number of time steps we start to see convergence. A similar behaviour was observed already in Section I.8 with second-order spatial differences.
- MR, the implicit midpoint rule. This method gives results that are much better than for the Euler method. With this method the Jacobian approximation in the Newton iteration was only based on the linear operator $i\partial_{xx}$ in the equation, following a suggestion of Sanz-Serna & Verwer (1986). For the

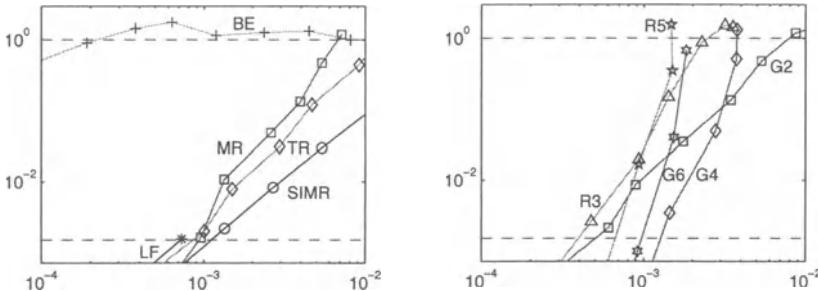


Fig. 6.5. Relative L_2 -errors versus C^{-1} , with measure of work C the number of mega-flops. Left figure: first- and second-order methods. Right figure: higher-order implicit Runge-Kutta methods of Gauss and Radau type. See text for description of the methods.

overall performance, this implementation aspect did not have a large impact, see also scheme G2 below.

- TR, the implicit trapezoidal rule. The results for this method are very similar to the implicit midpoint rule. Although here the Jacobian approximation A_n for the Newton iteration is recomputed each step (together with LU-decompositions), this is more or less balanced by faster convergence of the iteration.
- SIMR, the semi-implicit midpoint rule. This is an implementation of the implicit midpoint rule where in each time step only one Newton iteration is performed. It corresponds to the 1-stage Rosenbrock method (1.26) with $\gamma = \frac{1}{2}$. For large step sizes this method is here more efficient than the implicit midpoint rule and implicit trapezoidal rule, but to reach the error level $1.5 \cdot 10^{-3}$ almost the same amount of work C is required.

In the left picture of Figure 6.5 also results for the explicit leap-frog method can be seen; this method is discussed later. Further we note that the convergence behaviour in the figure seems somewhat irregular for the implicit midpoint and trapezoidal rule; this is mainly due to the fact that the horizontal line represents the inverse of the amount of work and not the time step. When the time step τ is decreased the number of Newton iterations per step diminishes in average, leading to less work per step.

High-order implicit methods: The right picture in Figure 6.5 contains the results for some fully implicit Runge-Kutta methods that were introduced in Example 1.4. We consider here the Gauss methods of order $p = 2, 4, 6$, indicated in the picture as G2, G4 and G6, and the Radau methods of order $p = 3, 5$, indicated as R3 and R5. Recall that in terms of number of stages s , we have $p = 2s$ for the Gauss methods and $p = 2s - 1$ for the Radau methods. The Gauss methods are in this test more efficient than the Radau methods due to the damping of the latter ones. This fact is not surprising, since this was observed already for the backward Euler method (Radau, order 1) and

the implicit midpoint rule (Gauss, order 2). It should also be noted that the results here with the second-order Gauss method (G2) are slightly different from the implicit midpoint rule (MR) results in the left plot of Figure 6.5; this is due to a somewhat different implementation of the Newton iteration.

Explicit methods: The same test has also been performed with the explicit midpoint 2-step method (leap-frog, LF) and the classical fourth-order Runge-Kutta method (RK4). Their results are essentially different from those of the implicit methods. With number of steps $N = 25 \cdot 2^k$, $k \geq 0$, the leap-frog method requires $N = 25600$ steps for stability and the Runge-Kutta method needs $N = 6400$ steps. This is in accordance with the linear stability restrictions for $u_t = iu_{xx}$ with fourth-order differences in space.¹⁰⁾ With this small step size, the leap-frog method gives a temporal error less than the spatial error, see the left picture in Figure 6.5 and the mega-flop count C is comparable to the implicit methods. With the fourth-order Runge-Kutta method the results do not fit in the plots; for $N = 6400$ we get approximately the same amount of work as for the leap-frog method but the corresponding error is $0.2 \cdot 10^{-7}$, which is very much less than with the other methods.

Taking the number of mega-flops as measure of work is somewhat arbitrary, but it was observed that the CPU time is much more sensitive to implementation issues. For a further comparison of implicit and explicit methods we have listed in Table 6.1 for several methods the number of steps, mega-flops and corresponding CPU time, needed to reach a temporal error level that is lower than the spatial discretization error $1.5 \cdot 10^{-3}$. Again the number of steps N is taken here from the sequence $\{25 \cdot 2^k : k \geq 0, \text{ integer}\}$. This division is rather coarse, and therefore – even more than in the above – the results should be regarded as indicative only.

	LF	RK4	TR	SIMR	G4	R5	G6
N^*	25600	6400	6400	6400	400	200	100
C	1375	1464	1929	1484	1075	1735	1050
CPU	109	81	278	270	698	584	341

Table 6.1. Minimal number of steps N^* needed to achieve a temporal error less than the spatial error, with corresponding mega-flop count C and CPU timings (seconds).

¹⁰⁾ To be more precise: the eigenvalues for the linear problem with spatial periodicity are in between $-i(32/6)h^{-2}$ and 0 on the imaginary axis. This gives a prediction of the necessary time step of $\tau \leq (6/32)h^2\beta_I$, where β_I is the imaginary stability boundary. In the present test this prediction turned out to be accurate for the fourth-order Runge-Kutta method ($\beta_I = 2\sqrt{2}$). For the leap-frog method ($\beta_I = 1$) a step size slightly (10%) smaller than predicted was required to avoid instabilities at the boundaries.

With the present implementations the explicit methods compare quite well to the implicit schemes, in spite of the stability restrictions. As mentioned in Section I.8, comparisons between implicit and explicit methods do depend for a great part on implementation issues. The programs for this 1D test were written in MATLAB with standard linear algebra solvers based on *LU*-decompositions. More sophisticated implementations, with specialized linear algebra subroutines, would probably give results more favourable for the implicit schemes, in particular for the high-order Gauss schemes. On the other hand, for multi-dimensional Schrödinger equations efficient implementation of the implicit schemes, especially the higher-order ones, will be much more complicated than for this 1D test problem. In that respect the explicit methods certainly have a clear advantage.

Conservation Properties

As mentioned in Section I.8, the dynamics of the Schrödinger equation is to a large extent determined by conservation properties, such as energy conservation. For all Gauss methods, energy is a conserved quantity, see Sanz-Serna & Verwer (1986) and Sanz-Serna & Calvo (1994). We illustrate this fact here for the implicit midpoint rule.

Recall from Section I.8 that the spatial discretization of the Schrödinger equation leads to a semi-discrete system $w'(t) = F(w(t))$ where

$$\operatorname{Re}\langle v, F(v) \rangle = 0$$

for all vectors $v \in \mathbb{C}^m$ with $\langle \cdot, \cdot \rangle$ a suitable inner product, and consequently it holds that

$$E_h(w(t)) = E_h(w(0)) \quad \text{for all } t > 0,$$

where $E_h(w) = \langle w, w \rangle = \|w\|^2$ denotes the discrete energy.

Writing the implicit midpoint rule as

$$w_{n+1} - w_n = \tau F\left(\frac{1}{2}w_n + \frac{1}{2}w_{n+1}\right),$$

we obtain, by taking the inner product with $\frac{1}{2}(w_n + w_{n+1})$, directly the conservation relation

$$E_h(w_{n+1}) = E_h(w_n) \quad \text{for all } n \geq 0.$$

Note that this property holds irrespective of the step size τ .

In Section I.8 it was shown that for the trapezoidal rule the related quantity $\tilde{E}_h(w_n) = \|w_n\|^2 + \frac{1}{4}\tau^2\|F(w_n)\|^2$ is conserved. Although this is not strictly energy conservation, the above numerical test did give very comparable results for the trapezoidal rule and the implicit midpoint rule.

As we saw from the results with the fourth-order explicit Runge-Kutta method, energy conservation is not needed in the present test as long as the scheme is sufficiently accurate. This would change however if a good

qualitative solution is required for arbitrarily long time intervals. With this Runge-Kutta method, stability for the linear Schrödinger equation implies that all Fourier modes will be (slightly) damped, because $|R(z)| < 1$ if $z = iy$, $0 < y < \beta_I$, so eventually for time t large enough the numerical solution will vanish.

In the recent numerical ODE literature schemes with inherent conservation properties, such as symplectic methods and time-reversible methods, have obtained much attention, see for example Sanz-Serna & Calvo (1994) and Hairer, Lubich & Wanner (2002). At present, the practical significance of these concepts for PDEs is still somewhat unclear, but it is a very active research field where new developments may be expected.

III Advection-Diffusion Discretizations

In this third chapter we return to the discretization of advection-diffusion problems. We will treat here a number of special subjects that are supplementary to the more introductory material of Chapter I.

1 Non-oscillatory MOL Advection Discretizations

As we saw in Chapter I, the common spatial advection discretizations show an oscillatory behaviour. An exception is the first-order upwind scheme, but that scheme is very diffusive. In this section we focus on the question how to achieve a non-oscillatory spatial discretization that has better accuracy than the first-order scheme. We start with positivity for the linear test problem $u_t + au_x = 0$, a constant. After that, the related TVD property and more general equations are treated.

1.1 Spatial Discretization for Linear Advection

Consider the advection test problem $u_t + au_x = 0$ with a constant velocity a . As in Section I.4 we discretize in space on uniformly distributed grid points $x_j = jh$ by means of the flux form

$$w'_j(t) = \frac{1}{h} \left(f_{j-\frac{1}{2}}(t, w(t)) - f_{j+\frac{1}{2}}(t, w(t)) \right), \quad f_{j\pm\frac{1}{2}}(t, w) = a w_{j\pm\frac{1}{2}}, \quad (1.1)$$

where the values $w_{j\pm\frac{1}{2}}$ are defined at the cell boundaries $x_{j\pm\frac{1}{2}}$. These approximate values determine the actual discretization in terms of neighbouring values w_i .

With first-order upwind fluxes

$$f_{j+\frac{1}{2}}(t, w) = \max(a, 0) w_j + \min(a, 0) w_{j+1},$$

positivity of the scheme can be immediately concluded from Theorem I.7.2. However, the first-order upwind scheme is in most cases too inaccurate and therefore not generally advocated. On the other hand, common higher-order

discretizations produce oscillatory solutions with the possibility of negative values that may be unphysical; for example, for densities or concentrations. Positivity can always be obtained by simply ‘cutting off’ negative approximations w_j . A disadvantage then is that we are adding mass and do not eliminate undershoot and overshoot. To impose mass conservation we will keep the schemes in the flux form (1.1). Spatial discretizations in this flux form that will give better accuracy than first-order upwind, as well as positive solutions without under- and overshoot, can be achieved by modifying the fluxes of a higher-order discretization by a technique called *limiting*.

Flux Limiting

Higher-order fluxes can be written as the first-order flux plus a correction. This correction mitigates the diffusive character of the first-order upwind scheme, and therefore it is sometimes called ‘anti-diffusion’. For example, with the second-order central flux we have

$$f_{j+\frac{1}{2}}(t, w) = \frac{1}{2}a(w_j + w_{j+1}) = a\left(w_j + \frac{1}{2}(w_{j+1} - w_j)\right),$$

and with the third-order upwind-biased flux we have, for $a > 0$,

$$f_{j+\frac{1}{2}}(t, w) = \frac{1}{6}a(-w_{j-1} + 5w_j + 2w_{j+1}) = a\left(w_j + \left(\frac{1}{3} + \frac{1}{6}\theta_j\right)(w_{j+1} - w_j)\right),$$

where θ_j is the ratio

$$\theta_j = \frac{w_j - w_{j-1}}{w_{j+1} - w_j}. \quad (1.2)$$

Since these schemes produce oscillations it can be concluded that the corrections are sometimes too large and therefore their value needs to be limited.

In the following we consider the more general form

$$f_{j+\frac{1}{2}}(t, w) = a\left[w_j + \psi(\theta_j)(w_{j+1} - w_j)\right], \quad a \geq 0, \quad (1.3)$$

with *limiter function* ψ . This limiter function is to be chosen such that we have better accuracy than first-order upwind, but still positivity. The choice of the limiter function ψ defines the actual discretization.

The general discretization (1.1), (1.3) written out in full gives

$$\begin{aligned} w'_j &= \frac{a}{h}\left(w_{j-1} + \psi(\theta_{j-1})(w_j - w_{j-1}) - w_j - \psi(\theta_j)(w_{j+1} - w_j)\right) \\ &= \frac{a}{h}\left(1 - \psi(\theta_{j-1}) + \frac{1}{\theta_j}\psi(\theta_j)\right)(w_{j-1} - w_j), \end{aligned} \quad (1.4)$$

with $w_j = w_j(t)$. In view of positivity, as in Theorem I.7.1, we require

$$1 - \psi(\theta_{j-1}) + \frac{1}{\theta_j}\psi(\theta_j) \geq 0.$$

Here θ_{j-1} and θ_j can assume any value in \mathbb{R} independent of each other. A sufficient condition on the limiter function is

$$0 \leq \psi(\theta) \leq 1, \quad 0 \leq \frac{1}{\theta} \psi(\theta) \leq \mu \quad \text{for all } \theta \in \mathbb{R}, \quad (1.5)$$

with μ a positive parameter free to choose. The larger μ , the less restricted the choice of ψ is. However, time integration puts a limit on μ (according to formula (1.26) below) and it turns out that $\mu = 1$ is a reasonable value.

With negative velocity a we get, by reflection around the point $x_{j+1/2}$, the flux expression

$$f_{j+\frac{1}{2}}(t, w) = a \left[w_{j+1} + \psi\left(\frac{1}{\theta_{j+1}}\right)(w_j - w_{j+1}) \right], \quad a < 0, \quad (1.6)$$

which is the same formula as (1.3), only seen from the ‘backside’. It will be shown below that (1.5) ensures positivity of the scheme also for variable velocities, including the possibility of sign changes in a .

By limiting we obtain a *nonlinear* spatial discretization for the *linear* advection problem. The general flux expression (1.3) or (1.6) inserted in (1.1) gives the invariance property (I.7.5) that was discussed with the maximum principle in Section I.7.1. Consequently, if we achieve positivity, then we are also sure that there will be no *global* under- and overshoot. In Section 1.3 it will be shown that in this case we also have the TVD (Total Variation Diminishing) property by which *local* under- and overshoot is avoided as well.

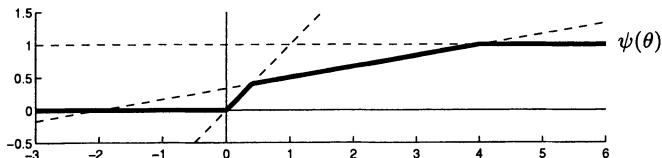
Choices of Limiter Functions

Example 1.1 For a smooth profile we have $\theta_j \approx 1$, except near extrema. Therefore we will take

$$\psi(\theta) = \frac{1}{3} + \frac{1}{6}\theta$$

near $\theta = 1$, so that the accuracy of the third-order scheme will be maintained in smooth regions away from extrema. An example of a limiter function satisfying (1.5) with $\mu = 1$, based on the third-order upwind-biased scheme is

$$\psi(\theta) = \max \left(0, \min \left(1, \frac{1}{3} + \frac{1}{6}\theta, \theta \right) \right). \quad (1.7)$$



This limiter function was introduced by Koren (1993) and can be seen to coincide with the original third-order upwind-biased function $\psi(\theta) = \frac{1}{3} + \frac{1}{6}\theta$ for $\frac{2}{5} \leq \theta \leq 4$.

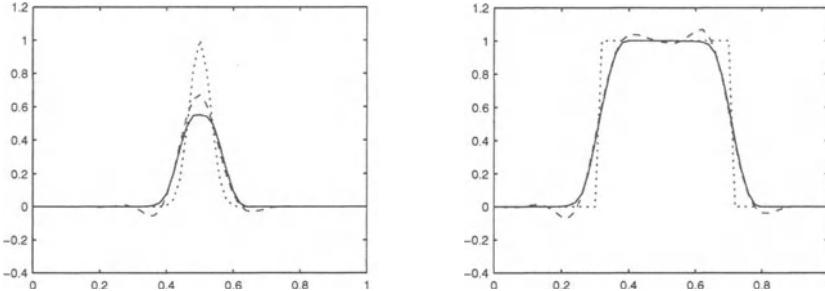


Fig. 1.1. Results for the third-order upwind-biased scheme, $h = 1/50$. Exact solutions are dotted, non-limited solutions dashed, limited solutions solid.

Some numerical results for this limiter are shown in Figure 1.1 on the test model $u_t + u_x = 0$ with $t > 0$, $0 < x < 1$ and a periodic boundary condition. The plots give numerical solutions at time $t = 1$ with $h = 1/50$ and two initial profiles, viz. the peaked function $u(x, 0) = (\sin(\pi x))^{100}$ and the block function $u(x, 0) = 1$ for $0.3 \leq x \leq 0.7$ and 0 otherwise. Note that the block function is discontinuous and hence not differentiable. Consequently, the characteristic solution $u(x, t) = u(x - at, 0)$ is not a solution of the differential equation in the usual, classical sense. It is, however, a solution of the underlying integral conservation law (see Section 1.4 for some more details). Further we note that time-stepping was performed here with an explicit Runge-Kutta method, using a step size small enough to render temporal errors negligible.

The limiting can be seen to provide positive solutions free from oscillations. The result for the peaked initial function $u(x, 0) = (\sin(\pi x))^{100}$ shows that the limited discretization has a very good phase speed, but it has further increased the amplitude error of the third-order upwind-biased scheme near the extremum. At the extremum we have $\theta_j \leq 0$ and thus the limiter will switch to the first-order upwind flux $f_{j+\frac{1}{2}} = aw_j$ causing additional damping. However, the inaccuracy remains confined to a relatively small region near

h	L_1 -error	order	L_2 -error	order	L_∞ -error	order
$\frac{1}{10}$	$7.18 \cdot 10^{-2}$		$8.96 \cdot 10^{-2}$		$1.56 \cdot 10^{-1}$	
$\frac{1}{20}$	$1.73 \cdot 10^{-2}$	2.05	$2.25 \cdot 10^{-2}$	1.99	$4.98 \cdot 10^{-2}$	1.65
$\frac{1}{40}$	$3.82 \cdot 10^{-3}$	2.18	$5.96 \cdot 10^{-3}$	1.92	$1.67 \cdot 10^{-2}$	1.58
$\frac{1}{80}$	$8.33 \cdot 10^{-4}$	2.20	$1.57 \cdot 10^{-3}$	1.92	$5.63 \cdot 10^{-3}$	1.57
$\frac{1}{160}$	$1.66 \cdot 10^{-4}$	2.33	$4.00 \cdot 10^{-4}$	1.97	$1.88 \cdot 10^{-3}$	1.58

Table 1.1. Errors and estimated orders for $u_t + u_x = 0$, $u(x, 0) = \sin^2(\pi x)$ with limiter (1.7).

the extremum. The result for the block function shows that limiting can also have an overall favourable effect on the accuracy.

Formal statements on accuracy near extrema are difficult to obtain due to the switches in the discretization. In Table 1.1 errors are given for the smooth initial function $u(x, 0) = \sin^2(\pi x)$ in the L_1, L_2, L_∞ -norm together with the estimated order upon halving h . Limiting now has an *adverse* effect on accuracy with an order drop in all three norms, due to damping of the peak value. For this smooth solution third-order accuracy is found, of course, if we do not limit. \diamond

Example 1.2 The κ -scheme is a one-parameter family of advection schemes proposed by van Leer (1985). For $a \geq 0$ it is defined by the linear flux

$$f_{j+\frac{1}{2}}(t, w) = a \left[w_j + \frac{1-\kappa}{4} (w_j - w_{j-1}) + \frac{1+\kappa}{4} (w_{j+1} - w_j) \right], \quad (1.8)$$

giving the scheme

$$w'_j = \frac{a}{4h} \left(-(1-\kappa) w_{j-2} + (5-3\kappa) w_{j-1} - (3-3\kappa) w_j - (1+\kappa) w_{j+1} \right). \quad (1.9)$$

The κ -scheme unifies several advection schemes. The values $\kappa = 1, -1$ and $\frac{1}{3}$ give the second-order central, the second-order upwind and the third-order upwind-biased scheme, respectively. With $\kappa = 0$ we get the Fromm scheme (I.3.44). Scheme (1.9) has order three for $\kappa = \frac{1}{3}$ and order two otherwise.

This κ -scheme can be considered with limiter

$$\psi(\theta) = \max \left(0, \min \left(1, \frac{1}{4}(1+\kappa) + \frac{1}{4}(1-\kappa)\theta, \theta \right) \right). \quad (1.10)$$

Note that the original, non-limited scheme can be written as (1.3) with

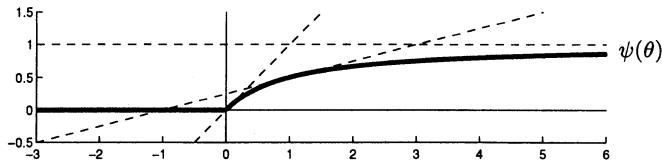
$$\psi(\theta) = \frac{1}{4}(1+\kappa) + \frac{1}{4}(1-\kappa)\theta.$$

For all κ -values considered, the limited scheme coincides with the linear scheme in a neighbourhood of $\theta = 1$. Such limiters are sometimes called *target limiters*, see Zalesak (1987), meaning that the nonlinear limiter is as close as possible to the linear scheme that serves as target, within the constraints (1.5). \diamond

Limiters of the form (1.3) were originally introduced for Lax-Wendroff type methods by Sweby (1984), mainly based on ideas of van Leer developed in the 1970's; see van Leer (1974, '77, '79), for instance. More references and a more general discussion can be found in the monographs of LeVeque (1992, 2002). The limiter (1.10) for semi-discretizations is just one of many possibilities. It does provide good results for linear advection.

Example 1.3 An example of a smoother limiter, due to van Leer (1974), is

$$\psi(\theta) = \frac{1}{2} \frac{\theta + |\theta|}{1 + |\theta|}. \quad (1.11)$$



This limiter is not a target limiter and introduces a nonlinear advection scheme which stands on its own. Linearizing this limiter function around $\theta = 1$, i.e., replacement of $\psi(\theta)$ by $\psi(1) + \psi'(1)(\theta - 1) = \frac{1}{4} + \frac{1}{4}\theta$, gives the Fromm scheme (I.3.44). The van Leer limiter is somewhat cheaper to implement than (1.10) as it involves no max-min operations. However, it is slightly more diffusive and hence in general it is slightly less accurate than (1.10). Table 1.2 illustrates this, showing the results obtained with van Leer's limiter for the test problem with smooth solution of Table 1.1. For sharp peaks or block functions the plots for this limiter are close to Figure 1.1, but slightly more diffused.

h	L_1 -error	order	L_2 -error	order	L_∞ -error	order
$\frac{1}{10}$	$1.03 \cdot 10^{-1}$		$1.21 \cdot 10^{-1}$		$1.96 \cdot 10^{-1}$	
$\frac{1}{20}$	$3.18 \cdot 10^{-2}$	1.70	$3.71 \cdot 10^{-2}$	1.71	$6.84 \cdot 10^{-2}$	1.52
$\frac{1}{40}$	$1.05 \cdot 10^{-2}$	1.60	$1.25 \cdot 10^{-2}$	1.57	$2.59 \cdot 10^{-2}$	1.40
$\frac{1}{80}$	$2.87 \cdot 10^{-3}$	1.87	$3.75 \cdot 10^{-3}$	1.74	$9.69 \cdot 10^{-3}$	1.42
$\frac{1}{160}$	$7.05 \cdot 10^{-4}$	2.03	$1.11 \cdot 10^{-3}$	1.76	$3.60 \cdot 10^{-3}$	1.43

Table 1.2. Errors and estimated orders for $u_t + u_x = 0$, $u(x, 0) = \sin^2(\pi x)$ obtained with van Leer's limiter (1.11).

◇

Many more examples and pictures can be found in the review paper Zalesak (1987). In that paper and in Zalesak (1979) one also finds results on a related limiting procedure called Flux Corrected Transport (FCT), originally developed by Boris & Book (1973). With FCT high-order fluxes are mixed with first-order upwind fluxes taking the inflow and outflow over cells simultaneously into account. The FCT procedure is more complicated from a computational point of view. An advantage of FCT is that it can be applied to any discretization in flux form. A detailed description with comparisons can also be found in Durran (1999). A review of some more recent spatial discretizations, such as the so-called essentially non-oscillatory (ENO) schemes, can be found in Shu (1999).

In practice, higher-order schemes provided with limiting perform very well compared to first-order upwind. In the literature such schemes are therefore

often called *high-resolution schemes* to distinguish them from the first-order upwind scheme. In actual implementations one usually adds a small number ϵ to the denominator of the ratio θ_j to prevent division by 0. This may result in small negative values of approximately size ϵ .

Variable Coefficients

For linear variable coefficient problems

$$u_t + (a(x, t)u)_x = 0$$

the above formulas are easily adjusted. We then take the flux expression

$$f_{j+\frac{1}{2}}(t, w) = \max(a_{j+\frac{1}{2}}, 0) w_{j+\frac{1}{2}}^R + \min(a_{j+\frac{1}{2}}, 0) w_{j+\frac{1}{2}}^L,$$

where $a_{j+1/2} = a(x_{j+1/2}, t)$ and $w_{j+1/2}^R, w_{j+1/2}^L$ are given by the bracketed terms in (1.3) and (1.6), respectively. With variable coefficients maximum principles no longer hold, see Section I.4.3, but limiting will still prevent negative solution values and local oscillations.

To prove the positivity property with (1.5), we show this to hold for the explicit Euler method under the step size restriction

$$\frac{\tau}{h} |a_{j+\frac{1}{2}}| \leq \frac{1}{1+\mu}, \quad \frac{\tau}{h} |a_{j+\frac{1}{2}} - a_{j-\frac{1}{2}}| \leq \frac{1}{1+\mu}. \quad (1.12)$$

For smooth velocity fields it will be the first condition that determines the maximal step size. To demonstrate sufficiency the various possibilities of signs of $a_{j\pm 1/2}$ can be considered individually. Let in the following $\nu_{j\pm 1/2} = \tau a_{j\pm 1/2}/h$.

First suppose $a_{j\pm 1/2} \geq 0$. Then we obtain, similar as for (1.4),

$$\begin{aligned} w_j + \frac{\tau}{h} \left(f_{j-\frac{1}{2}}(t, w) - f_{j+\frac{1}{2}}(t, w) \right) &= \left[1 + \nu_{j-\frac{1}{2}} \psi(\theta_{j-1}) \right. \\ &\quad \left. - \nu_{j+\frac{1}{2}} \left(1 + \frac{\psi(\theta_j)}{\theta_j} \right) \right] w_j + \left[\nu_{j-\frac{1}{2}} (1 - \psi(\theta_{j-1})) + \nu_{j+\frac{1}{2}} \frac{\psi(\theta_j)}{\theta_j} \right] w_{j-1}. \end{aligned}$$

It follows that in this case the first condition in (1.12) is sufficient for positivity. Suppose next $a_{j-1/2} \leq 0$ and $a_{j+1/2} \geq 0$. Then we obtain in a similar way

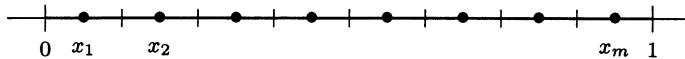
$$\begin{aligned} w_j + \frac{\tau}{h} \left(f_{j-\frac{1}{2}}(t, w) - f_{j+\frac{1}{2}}(t, w) \right) &= \left[1 - |\nu_{j-\frac{1}{2}}| \left(1 + \theta_j \psi\left(\frac{1}{\theta_j}\right) \right) \right. \\ &\quad \left. - \nu_{j+\frac{1}{2}} \left(1 + \frac{\psi(\theta_j)}{\theta_j} \right) \right] w_j + |\nu_{j-\frac{1}{2}}| \theta_j \psi\left(\frac{1}{\theta_j}\right) w_{j+1} + \nu_{j+\frac{1}{2}} \frac{\psi(\theta_j)}{\theta_j} w_{j-1}. \end{aligned}$$

It follows from (1.5) by some straightforward calculations that in this case the second condition in (1.12) ensures positivity. For the remaining possibilities of signs of $a_{j\pm 1/2}$ verification can be done in the same way.

Positivity of the semi-discrete form (1.1) now follows by convergence of the explicit Euler method for $\tau \rightarrow 0$. For this, Lipschitz continuity of the fluxes has to be established, which is a somewhat tedious exercise. In fact, positivity of fully discrete schemes using explicit Runge-Kutta or linear multistep methods for time stepping will be shown directly using the corresponding property of the explicit Euler method.

Boundary Conditions

A general procedure to implement boundary conditions for advection problems is to use virtual points and simple extrapolation formulas. Several examples were given in Section I.5; in particular in Section I.5.4 this procedure was illustrated for the third-order upwind-biased scheme with a vertex centered grid. Here we will illustrate this procedure again for the cell centered grid $\{x_j : x_j = (j - \frac{1}{2})h, j = 1, \dots, m, h = 1/m\}$ and the semi-discretization (1.1), (1.3) with given inflow boundary value at $x = 0$ and outflow at $x = 1$.



A simple inspection shows that three fluxes need to be adjusted, viz. $f_{1/2}, f_{3/2}$ and $f_{m+1/2}$. If $a > 0$ the flux $f_{1/2}$ is just the inflow boundary flux, so it is natural to define $w_{1/2}(t) = \gamma_0(t)$ where $\gamma_0(t)$ is the prescribed boundary value. For $f_{3/2}$ and $f_{m+1/2}$ the virtual values w_0 and w_{m+1} are introduced, respectively. Both can be computed by constant, linear or quadratic extrapolation. Constant extrapolation guarantees positivity, but linear or quadratic extrapolation usually gives more accuracy. At the outflow boundary one also often takes the first-order upwind fluxes, in particular if it is not known in advance whether the solution is smooth at the boundary, such as with outflow boundary layers for problems with small diffusive terms.

1.2 Numerical Examples

An Adsorption Model

As a test example we consider an adsorption model from soil mechanics. Consider a flow through a porous medium with macroscopic velocity a and a chemical species that dissolves in the fluid but which can also be adsorbed by the solid medium. Let u be the dissolved concentration and v the adsorbed concentration. The conservation law for the total concentration $u + v$ then reads

$$(u + v)_t + (au)_x = 0.$$

The local balance between u and v is given by

$$v_t = -k(v - \phi(u)),$$

where $k > 0$ is the reaction rate and

$$\phi(u) = \frac{k_1 u}{1 + k_2 u}$$

describes the steady state ratio between u and v with $k_1, k_2 > 0$. In soil mechanics ϕ is known as a Langmuir isotherm. These equations can be written as a system of advection-reaction equations

$$\begin{aligned} u_t + (au)_x &= k(v - \phi(u)), \\ v_t &= -k(v - \phi(u)). \end{aligned} \quad (1.13)$$

Having $u, v \geq 0$ is necessary for the model to make physical sense. Moreover, $\phi(u)$ has a singularity at $u = -1/k_2$. As a consequence, non-limited higher-order advection discretizations cannot be used here if there are steep gradients near a state $u = 0$, since this will lead to negative values or even divergence by the singularity in ϕ .

We will take the values $k = 1000$, $k_1 = k_2 = 100$ and solve the equations as a stiff advection-reaction system for $t > 0$, $0 < x < 1$. The velocity a is taken spatially homogeneous, given by

$$a(t) = \begin{cases} 1 & \text{if } t \leq 1, \\ -1 & \text{if } t > 1. \end{cases}$$

The initial functions are $u(x, 0) = v(x, 0) = 0$ and as boundary condition for u we have $u(0, t) = 1$ for $t \leq 1$ and $u(1, t) = 0$ for $t > 1$. As time interval we take $0 \leq t \leq \frac{5}{4}$. An illustration of the exact solution is given in Figure 1.2 where the concentrations u , v and the total concentration $u + v$ are plotted as a function of x at time $t = 1$ and $t = \frac{5}{4}$.

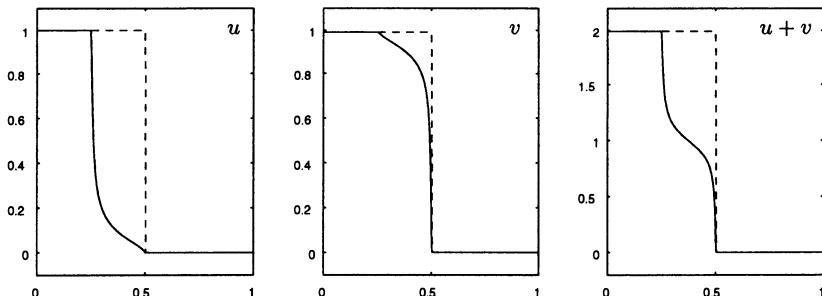


Fig. 1.2. Solution of the adsorption problem (1.13). Dissolved concentration u (left), adsorbed concentration v (middle) and total concentration $u + v$ (right) at time $t = 1$ (dashed lines) and $t = \frac{5}{4}$ (solid lines).

For $0 \leq t \leq 1$ there is a shock front traveling to the right. The speed of the front is not equal to the advective velocity $a = 1$ but equals only half of it, approximately, since the propagation is slowed down by adsorption. After $t = 1$ the advective velocity is reversed and then a rarefaction wave is formed due to advection of u to the left and dissolution of adsorbed concentration. Note that the top of this rarefaction wave, where u becomes 1, now travels with speed $a = -1$.

For the spatial discretization we consider the first-order upwind and limited third-order upwind-biased discretization, with limiter (1.7). A cell centered grid has been used with $h = 1/100$ and the inflow and outflow boundary have been dealt with as described in the paragraph above on boundary conditions, using quadratic extrapolation for the third-order scheme. Time integration was performed with very small time steps so that no temporal errors are visible. As said before, non-limited discretizations cannot be used due to negative values. As an illustration for the need of mass conservation, we have included in the experiment the non-limited third-order discretization where in each time step the numerical approximation w_n is replaced by $\max(w_n, 0)$. Hence we simply cut-off negative values whenever they arise, which is a rather primitive way to avoid negative solutions. We refer to this as *clipping*. The results are given in Figure 1.3.

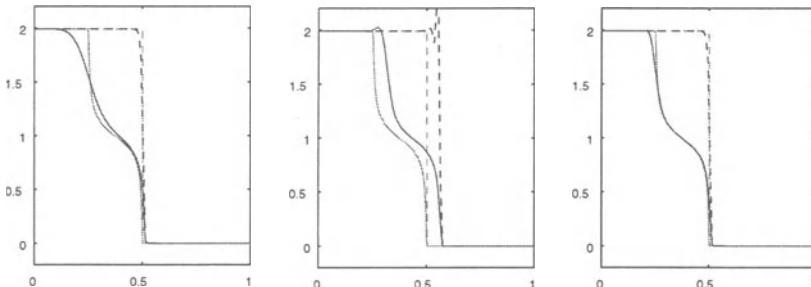


Fig. 1.3. Numerical solutions $u + v$ of (1.13) at $t = 1$ (dashed lines) and $t = \frac{5}{4}$ (solid lines) with $h = 1/100$ for first-order upwind (left), third-order upwind-biased with clipping (middle) and third-order upwind-biased with limiting (right). The exact solutions are indicated by grey lines.

Comparing the first-order and limited third-order discretization, we see little difference up to $t = 1$. This can be expected since in the shock the limiter will also switch to first-order upwind and outside the shock the solution is constant, so there any consistent discretization gives the same result. Enlargement of the plot would show a small difference: with first-order upwind the shock is a bit more smeared. For $t > 1$ we clearly see that the first-order discretization gives larger errors due to numerical diffusion. Even on the short interval $t \in [1, \frac{5}{4}]$ the error has become significant. The limited third-order discretization also adds some diffusion but much less. The difference between

these two discretizations would be more pronounced if the time interval were larger.

For the non-limited third-order discretization with clipping of negative values the front speed is clearly too large for $0 \leq t \leq 1$. This is caused by the fact that by clipping we are effectively *adding mass* in the front which results in a faster adsorption process. This, in turn, speeds up the total solution and gives large errors. Note that there is also some overshoot visible.

Remark 1.4 Incorrect shock speeds for schemes in non-conservative form are typical for nonlinear hyperbolic equations. In connection to (1.13) the following can be noted. In the above experiments the reaction constant was given by $k = 1000$ and an increase of k hardly changes the solution. In the limit $k \rightarrow \infty$ we have $v = \phi(u)$, or

$$(u + \phi(u))_t + au_x = 0,$$

which can be formulated as a nonlinear conservation law for $\bar{u} = u + \phi(u)$,

$$\bar{u}_t + af(\bar{u})_x = 0, \quad (1.14)$$

where f is defined by the relation

$$\bar{u} = u + \phi(u) \implies u = f(\bar{u}).$$

We can discretize (1.14) by a flux form similar to (1.1), which leads to a solution that is virtually the same as in Figure 1.3 for $k = 1000$. Semi-discrete flux forms for nonlinear conservation laws such as (1.14) will be discussed in Section 1.4. \diamond

The Angiogenesis Model with Limiting and Clipping

The adsorption test suggests that limiting is a significantly better way to avoid negative solutions than clipping. This is not always true. An example is provided by the angiogenesis model (I.1.32) that was used for numerical tests in Section I.8. An important difference with the above adsorption model is that for the angiogenesis model there is no strict mass conservation underlying the equations due to the source terms in (I.1.32).

Here the third-order upwind-biased scheme is considered without limiting, with limiting and with clipping. The limiter is again (1.7). Figure 1.4 contains numerical solution plots for the endothelial cell density ρ on a single grid, chosen as in Figure I.8.4. In case of the smooth solution ($\delta = 1$, left plot), both clipping and limiting do remove the oscillation ahead of the front and hence give a significantly better qualitative approximation. In the non-smooth case ($\delta = 10^{-3}$, right plot) the three approximations are of comparable quality on this grid (consequently, the individual lines are hard to distinguish in the plot). The limited solution does prevent overshoot after the front, but

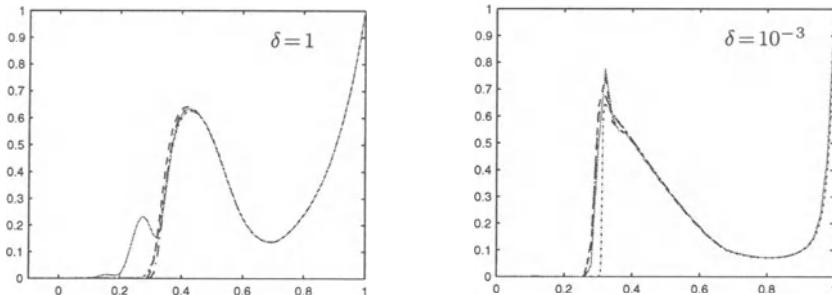


Fig. 1.4. Solutions $\rho(x, T)$ with $\delta = 1$, $T = 0.7$, $h = 1/100$ (left) and $\delta = 10^{-3}$, $T = 0.5$, $h = 1/50$ (right). Numerical solutions for the third-order upwind-biased scheme (grey), with limiting (dash-dot) and with clipping (dashed). The dotted line is an accurate reference solution.

the overall error is mainly caused by numerical diffusion which gives a front velocity that is too large, see also the discussion in Section I.8.

Also on other grids the L_2 -errors of the schemes with limiting and clipping were close together, and only for $\delta = 1$ a considerable improvement over the non-limited scheme was found. It should be noted however that for this non-limited scheme the reaction term for the endothelial cell density was replaced by its absolute value to prevent instabilities due to negative values, see Section I.8. For the present simple model this modification was straightforward, but for large systems with many components such a modification may not be possible or it may require detailed insight in the properties of the system.

1.3 Positivity and the TVD Property

On several occasions it has been mentioned that positivity together with the invariance property (I.7.5) implies the maximum principle (I.7.4), which assures that global undershoot and overshoot cannot occur. However, (I.7.4) does not guarantee the absence of oscillations which are localized under- and overshoots. In this section we will discuss the TVD (*Total Variation Diminishing*) property which does prevent localized under- and overshoots. The conditions giving the TVD property will turn out to be the same as for positivity. We note that the study of the TVD property originates with nonlinear hyperbolic equations, which will be treated in the next section. It is instructive to consider first linear advection-diffusion equations.

Linear TVD Systems

To discuss the TVD property, we first consider linear systems $w'(t) = Aw(t)$ with A an $m \times m$ circulant matrix (I.3.7) and let $h = 1/m$. So the ODE system

is here supposed to be a semi-discrete system and to originate from a linear PDE with constant coefficients and spatial periodicity, for instance, from the advection-diffusion test problem $u_t + au_x = du_{xx}$. For vectors $v \in \mathbb{R}^m$ we introduce

$$|v|_{TV} = h \sum_{j=1}^m |v_{j-1} - v_j| \quad \text{where } v_0 = v_m. \quad (1.15)$$

Viewing v as a grid function on the uniform space grid with nodes $x_j = jh$, (1.15) is seen to approximate the *total variation* of a periodic function $u(x)$ on $[0, 1]$ when $v_j = u(x_j)$. As in (I.3.7), the circulant system is denoted with

$$(Av)_i = \sum_{j=0}^{m-1} c_j v_{i+j} \quad \text{where } v_{i\pm m} = v_i, \quad (1.16)$$

and thus the positivity condition (I.7.3) holds iff

$$c_j \geq 0 \quad \text{for all } j \neq 0. \quad (1.17)$$

Furthermore, for any consistent spatial discretization we have

$$\sum_{j=0}^{m-1} c_j = 0. \quad (1.18)$$

This implies the invariance property so that with (1.17)–(1.18) the maximum principle (I.7.4) is valid.

With (1.17)–(1.18) it can also be shown that $|w(t)|_{TV}$ is *non-increasing* for evolving time. The semi-discretization is then said to be TVD. It will prevent local oscillations to arise because these would lead to an increase of the total variation $|w(t)|_{TV}$.

To demonstrate that the total variation is non-increasing, we first prove that the explicit Euler method maintains the TVD property if its step size τ satisfies $|\tau c_0| \leq 1$, which is just the positivity constraint. We have

$$\begin{aligned} |(I + \tau A)v|_{TV} &= h \sum_{i=1}^m \left| (v_{i-1} - v_i) + \tau \sum_{j=0}^{m-1} c_j (v_{i-1+j} - v_{i+j}) \right| \\ &\leq (1 + \tau c_0) |v|_{TV} + \tau \sum_{j=1}^{m-1} c_j |v|_{TV} = |v|_{TV} \quad \text{if } |\tau c_0| \leq 1. \end{aligned}$$

Through convergence of the explicit Euler method for $\tau \rightarrow 0$, we thus find that $|w(t)|_{TV}$ will be non-increasing if (1.17) and (1.18) hold.

Nonlinear TVD Systems

Next consider the semi-discrete system (1.1) for the linear advection problem $u_t + au_x = 0$ with constant $a \geq 0$ and spatial periodicity. Suppose that the

fluxes are defined by (1.3) and limited such that (1.5) holds. The semi-discrete system is then nonlinear and according to (1.4) this system can be written in the form

$$w'_j(t) = \alpha_j(w(t)) (w_{j-1}(t) - w_j(t)), \quad j = 1, \dots, m, \quad (1.19)$$

with the nonlinear function $\alpha_j(w)$ satisfying

$$0 \leq \alpha_j(w) \leq \frac{a}{h} (1 + \mu). \quad (1.20)$$

Assuming Lipschitz continuity, positivity for this nonlinear system can then be concluded from Theorem I.7.1. Since a is constant, also the invariance property (I.7.5) holds, guaranteeing the maximum principle (I.7.4). Furthermore, similar as in the linear case above, this nonlinear system is TVD.

To demonstrate that $|w(t)|_{TV}$ is non-increasing, we again use the explicit Euler method

$$w_j^{n+1} = w_j^n + \tau \alpha_j(w_n) (w_{j-1}^n - w_j^n), \quad (1.21)$$

where $w_n = (w_j^n) \in \mathbb{R}^m$. By spatial periodicity we have $w_{j\pm m}^n = w_j^n$ and $\alpha_{j\pm m} = \alpha_j$. A somewhat more general space-periodic scheme is

$$w_j^{n+1} = w_j^n + \tau \alpha_j(w_n) (w_{j-1}^n - w_j^n) - \tau \beta_j(w_n) (w_j^n - w_{j+1}^n). \quad (1.22)$$

The TVD property of this scheme is based on the following lemma, due to Harten (1983).

Lemma 1.5 *Sufficient for (1.22) to be TVD are the inequalities $\alpha_j(w) \geq 0$, $\beta_j(w) \geq 0$ and $\tau(\alpha_{j+1}(w) + \beta_j(w)) \leq 1$ for all indices j and $w \in \mathbb{R}^m$.*

Proof. Denoting $\alpha_j^n = \alpha_j(w_n)$, $\beta_j^n = \beta_j(w_n)$, the proof follows directly by summing the absolute values of

$$\begin{aligned} w_{j-1}^{n+1} - w_j^{n+1} &= (1 - \tau \alpha_j^n - \tau \beta_{j-1}^n) (w_{j-1}^n - w_j^n) \\ &\quad + \tau \alpha_{j-1}^n (w_{j-2}^n - w_{j-1}^n) + \tau \beta_j^n (w_j^n - w_{j+1}^n) \end{aligned}$$

and using the non-negativity of each coefficient and the space periodicity. \square

In view of (1.20) we find that the explicit Euler method (1.21) is TVD if

$$\frac{\tau a}{h} \leq \frac{1}{1 + \mu}. \quad (1.23)$$

By assuming Lipschitz continuity and letting $\tau \rightarrow 0$, we can conclude that the semi-discrete system (1.19) is TVD. Using (1.22) it is now easy to derive the TVD property also for the semi-discrete system with second-order central diffusive terms included.

The TVD Condition and the Explicit Euler Method

The TVD property for the explicit Euler method (1.21) will prove useful for deriving results for higher-order integration methods. It should be stressed however, that explicit Euler itself is not recommended for actual integration. Although the method satisfies the maximum principle and is TVD for step sizes τ satisfying (1.23), it may not be stable. When this occurs the method can give very peculiar results, such as turning smooth slopes into blocks or staircases, a behaviour known as *compression*.

An illustration is given in Figure 1.5 for $u_t + u_x = 0$ with spatial periodicity $u(x \pm 1, t) = u(x, t)$ and a cone shaped initial profile with output times $t = 1, 5$. These results have been obtained with limiter (1.7), number of grid points $m = 100$ and Courant number $\nu = \frac{1}{3}$. For reference, also results with a second-order Runge-Kutta method are included, for which we have taken here the explicit trapezoidal rule, with the same Courant number. This solution gives some smearing at the top and bottom of the cone, but apart from this the solution is accurate. In the figure this RK2 solution (indicated by dashes) largely coincides with the exact solution (dots).

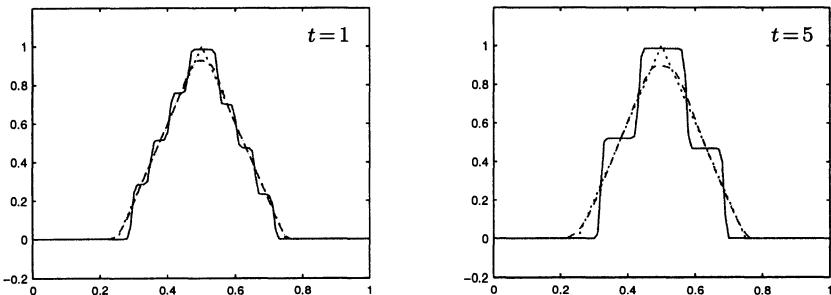


Fig. 1.5. Illustration of compression with limiting and forward Euler (solid lines), $h = 1/100$, $\nu = \frac{1}{3}$ at $t = 1, 5$. The exact solutions are dotted, RK2 solutions dashed.

For a heuristic explanation of the compression, we note that the limiter (1.7) is based on the third-order upwind-biased scheme. The explicit Euler method is not stable for this underlying linear scheme, see Section I.6.3, leading initially to a growth of the Fourier coefficients of the low-frequency modes. But then the limiter interferes with this growth, preventing an increase of the total variation. With limiting we thus get the interesting *nonlinear* phenomenon of instability combined with boundedness due to the maximum principle and the TVD property. This illustrates in particular that for nonlinear systems boundedness is not sufficient for having stability.

Finally we note that compression can be very beneficial if maintenance of sharp profiles or discontinuities is important. Then the compression will counteract the numerical diffusion which usually gives smearing. With a block

profile the forward Euler method would give excellent results in the above test, with a numerical solution hard to distinguish from the exact one.

TVD Results for ODE Methods

System (1.19), (1.20) is an example of an equation $w'(t) = F(t, w(t))$ with function F such that

$$|v + \tau F(t, v)|_{TV} \leq |v|_{TV} \quad \text{for all } v \in \mathbb{R}^m, t \geq 0 \text{ and } 0 < \tau \leq \tau^*. \quad (1.24)$$

Here τ^* can be viewed as the maximal step size allowed with forward Euler. Alternatively, for any given $\tau^* > 0$ we can regard (1.24) as a condition on F .

The backward Euler method $w_{n+1} = w_n + \tau F(t_{n+1}, w_{n+1})$ will be TVD under assumption (1.24) without any time step restriction. This is easily seen from

$$\begin{aligned} \left(1 + \frac{\tau}{\tau^*}\right) w_{n+1} &= w_n + \frac{\tau}{\tau^*} \left(w_{n+1} + \tau^* F(t_{n+1}, w_{n+1})\right), \\ \left(1 + \frac{\tau}{\tau^*}\right) |w_{n+1}|_{TV} &\leq |w_n|_{TV} + \frac{\tau}{\tau^*} |w_{n+1}|_{TV}, \end{aligned}$$

showing $|w_{n+1}|_{TV} \leq |w_n|_{TV}$ for any $\tau > 0$.

As for the forward Euler method, the result for backward Euler is mainly of use to derive TVD results for higher-order methods for advection equations. In spite of being unconditionally TVD and possessing excellent stability properties, this implicit method is not advocated for advection integration because of its excessive damping; see Figure I.6.2 for an illustration. Note however that the derivation can be easily extended to the advection-diffusion problem $u_t + au_x = (du_x)_x$ if the diffusion term, which may be non-constant, is discretized with second-order central differences. This shows in particular that the backward Euler method is unconditionally TVD for the purely parabolic problem $u_t = (du_x)_x$, which is a useful result in its own right.

Sufficient conditions for the TVD property can now be easily derived for explicit and diagonally implicit Runge-Kutta methods that fit in the form (II.4.11), by using properties known for the forward and backward Euler method. This is completely similar to positivity in Theorem II.4.6. Likewise, results for linear multistep methods are obtained from the form (II.4.19). In fact, the forms (II.4.11) and (II.4.19) were originally introduced by Shu & Osher (1988), Shu (1988) to study TVD properties.

Consider a Runge-Kutta method (II.4.11) with non-negative parameters p_{ij} , q_{ij} , q_i . Then it follows that the method is TVD for general systems $w'(t) = F(w(t))$ satisfying (1.24), under the step size restriction

$$\tau \leq \tau^* \min_{0 \leq j < i \leq s} \left(\frac{p_{ij}}{q_{ij}} \right). \quad (1.25)$$

In particular, for nonlinear systems defined by the space periodic, limited, advection discretization (1.19), (1.20), the TVD property holds under the

restriction

$$\nu = \frac{\tau a}{h} \leq \frac{1}{1 + \mu} \min_{0 \leq j < i \leq s} \left(\frac{p_{ij}}{q_{ij}} \right). \quad (1.26)$$

Here we have adopted the convention $p_{ij}/0 = +\infty$ for $p_{ij} \geq 0$. Observe that the CFL restriction (1.26) suggests to take μ smaller than 1. This, however, would restrict the domain (1.5) too much, affecting the quality of the spatial discretization. As mentioned there, the choice $\mu = 1$ is therefore a good compromise.

The necessity of the step size restriction (1.26) was experimentally studied for several Runge-Kutta methods in Hundsdorfer et al. (1995) and Gerisch & Weiner (2002) for the model equation $u_t + u_x = 0$ with the limiter (1.7). In general it was found that (1.26) is somewhat too strict. This is not surprising since near extrema the limiter will switch to first-order upwind fluxes. The 1D test results were not conclusive. Closer resemblance with the theoretical restriction (1.26) was found with 2D advection tests. For example, in 2D the classical Runge-Kutta method (II.1.8.b) then always did give negative solution values, in accordance to the bound (1.26) and the positivity result of Theorem II.4.6.

From a practical point of view we advocate the second-order 2-stage explicit trapezoidal rule (II.1.6) and the explicit second-order 3-stage method (II.4.15). These two methods satisfy the theoretical TVD condition (1.26) for $\nu \leq \frac{1}{2}$ and $\nu \leq 1$, respectively, if $\mu = 1$. Further, with third-order upwind-biased (non-limited) spatial discretization the trapezoidal rule and the 3-stage method are von Neumann stable for $\nu \leq 0.87$ and $\nu \leq 1.25$, respectively.

Results for multistep methods of the form (II.4.19) are obtained in a similar manner. However, as noted already in Section II.4.3, there are not that many multistep methods satisfying (II.4.19) with coefficients $p_j, q_j \geq 0$. To get around this restriction, starting procedures are to be taken into account. In Hundsdorfer, Ruuth & Spiteri (2003) results of the type $|w_n|_{TV} \leq M|w_0|_{TV}$ were derived with constant $M \geq 1$ depending on the starting procedure. Instead of TVD, such a property is usually called *total variation boundedness* or TVB; in practice this also suffices. Assuming (1.26) with $\mu = 1$, it was shown that the second-order Adams-Basforth method (II.3.7) and the extrapolated BDF2 scheme (II.3.13) are TVB for $\nu \leq \frac{5}{18}$ and $\nu \leq \frac{1}{4}$, respectively. If the amount of work per step is taken into account, this step size restriction is comparable to the restriction for the second-order Runge-Kutta methods. For the implicit BDF2 scheme (II.3.11) the restriction $\nu \leq \frac{1}{4}$ was found.

With respect to positivity and the TVD property, implicitness seems to have little added value, in clear contrast with stability. For example, the implicit trapezoidal rule satisfies (1.26) with $\nu \leq 1$ if $\mu = 1$, and the implicit Euler method is the only common implicit method that satisfies (1.26) unconditionally. As we know, the implicit Euler method is not appropriate for advection integration. All diagonally implicit Runge-Kutta methods mentioned before have rather small TVD intervals. The poor performance of some implicit methods will be illustrated numerically below.

Tests with Implicit Advection Schemes

The virtue of implicit methods is stability and many implicit methods have attractive unconditional stability properties for advection, thus allowing large Courant numbers as far as stability is concerned. However, as outlined above, with regard to positivity and TVD implicit methods allow only marginally larger Courant numbers than explicit methods. Consequently, their accuracy quickly deteriorates when applied with larger and larger Courant numbers.

Figures 1.6 and 1.7 illustrate this for the trapezoidal rule (II.1.11) and the implicit BDF2 scheme (II.3.11) started with the implicit Euler method. These are among the most common second-order implicit ODE methods. The test problem is $u_t + u_x = 0$ with $t > 0$, $0 < x < 1$ and a periodic boundary condition, using third-order upwind-biased discretization in space. Both methods are unconditionally von Neumann stable in this case. We show results without and with limiting. The limiter is (1.7). The initial function is $u(x, 0) = (\sin(\pi x))^{100}$ and the pictures presented here are for $t = 1$ and a uniform grid with grid size $h = 1/50$.

The pictures for both methods are self-evident. For the limited solution the qualitative behaviour and temporal accuracy is good with both methods for the smallest displayed Courant numbers $\nu = \tau/h$. This can be seen by comparison with the left picture of Figure 1.1 which shows the same solution for negligible time errors. The trapezoidal rule performs well up to $\nu = 1$ but gives oscillations and negative values for $\nu = 2$. The BDF2 scheme already produces some oscillation for $\nu = \frac{1}{2}$. Note that for the trapezoidal rule positivity and the TVD property are ensured for $\nu \leq 1$ by (1.26), and, as mentioned above, the BDF2 scheme is TVB for Courant numbers $\nu \leq \frac{1}{4}$.

With other implicit methods and other limiter functions a similar behaviour was observed. An obvious conclusion is that for scalar advection problems implicit time stepping is not attractive if positivity or the TVD property are to be maintained. Furthermore, the implementation of implicit advection with limiters is difficult. It gives rise to nonlinear implicit equations that must be solved by a Newton process to allow large Courant numbers. Due to the switches of the limiter this turns out to be troublesome and

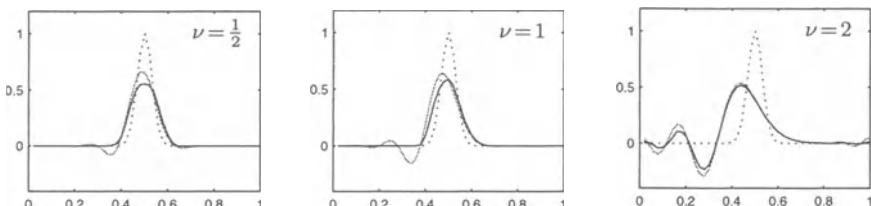


Fig. 1.6. Advection test for the implicit trapezoidal rule with Courant numbers $\nu = \frac{1}{2}, 1, 2$. Exact solutions are dotted, limited solutions solid and non-limited solutions grey.

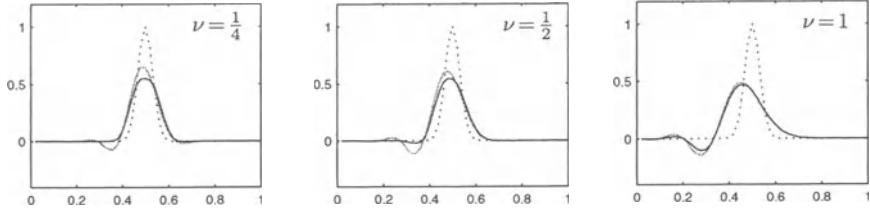


Fig. 1.7. Advection test for the implicit BDF2 scheme with Courant numbers $\nu = \frac{1}{4}, \frac{1}{2}, 1$. Exact solutions are dotted, limited solutions solid and non-limited solutions grey.

computationally expensive, even for smooth solutions and smooth limiters. Compared to explicit advection, implicit treatment of advection can only be expected to be more efficient for very smooth solutions without limiting and fast linear solvers at hand.

For advection-diffusion problems with non-smooth solutions where the diffusion part requires an implicit treatment, implicit-explicit or splitting techniques are appropriate. Such techniques will be addressed in the next chapter.

1.4 Nonlinear Scalar Conservation Laws

We next consider the scalar nonlinear advection problem

$$u_t + f(u)_x = 0, \quad (1.27)$$

with a differentiable, nonlinear flux function f independent of x and t . Equations of this type are called nonlinear *conservation laws*, indicating that the total amount of u in any interval or grid cell $[x - \frac{1}{2}h, x + \frac{1}{2}h]$ only changes by the transport or flux through its boundaries,

$$\frac{d}{dt} \int_{x-\frac{1}{2}h}^{x+\frac{1}{2}h} u(s, t) ds = f(u(x - \frac{1}{2}h, t)) - f(u(x + \frac{1}{2}h, t)). \quad (1.28)$$

In contrast to linear advection equations, solutions of (1.27) may form discontinuities even if we start with a smooth initial profile. In this section we will briefly discuss some of the main features of solutions with shocks and appropriate spatial discretizations.

The scalar nonlinear advection-diffusion problem

$$u_t + f(u)_x = d u_{xx}, \quad (1.29)$$

with $d \geq 0$, is said to be in *conservation form*. Solutions of this equation will be differentiable for $d > 0$, due to the smoothing of the diffusion term. In that case we can differentiate $f(u)_x$ to obtain the *advective form*

$$u_t + a(u)u_x = d u_{xx}, \quad a(u) = f'(u). \quad (1.30)$$

In Section II.2.3 this form was used for Burgers' equation, $f(u) = \frac{1}{2}u^2$, with a relatively large diffusion coefficient $d = 10^{-1}$, leading to very smooth solutions. In Section II.5.2 this equation was considered with 10^{-3} , giving solutions with steep fronts, and there the discretization was based on the conservative form. For very smooth solutions it does not matter much whether the conservative or the advective form is used, but for non-smooth solutions it is essential to discretize the conservation form in order to get correct propagations of steep fronts. For the hyperbolic problem (1.27), correct discretizations, based on the conservative form, are even more important to have a correct propagation of the discontinuities.

Solutions with Shocks

The constant coefficient advection problem $u_t + au_x = 0$ has the general solution $u(x, t) = u(x - at, 0)$, revealing that initial profiles, including discontinuous ones, are transported without change of shape along parallel straight characteristic lines.

In the nonlinear case $u_t + a(u)u_x = 0$ the characteristics in the (x, t) -plane are defined by $\xi'(t) = a(u(\xi(t), t))$ where u is considered for the moment as given. Along these characteristics we have

$$\frac{d}{dt}u(\xi(t), t) = u_t + \xi'u_x = 0.$$

Hence u is constant along the characteristics, and this implies in its turn that the characteristics are straight lines. The slope of the characteristic that passes through a point $(x_0, 0)$ is given by $\xi'(t) = a(u(x_0, 0))$. Hence, unlike the linear case, the characteristics are no longer parallel and in general some of them will cross after a finite time, even if the initial profile is smooth. When this happens the solution becomes discontinuous and at that moment the advective form is no longer equivalent to the conservative form. The advective form would lead to unphysical multi-valued solutions. With the conservative form, on the other hand, the solution will continue with a shock that has a finite propagation speed. Since the solution is discontinuous the differential equation (1.27) and its advective form are not valid anymore in the classical sense. What remains valid is the integral relation

$$\int_{x_L}^{x_R} [u(x, t_2) - u(x, t_1)] dx + \int_{t_1}^{t_2} [f(u(x_R, t)) - f(u(x_L, t))] dt = 0 \quad (1.31)$$

for all $x_L < x_R$ and $t_1 < t_2$, which is obtained from equation (1.27) in conservation form. It is the integral solution (1.31) of (1.27) that is to be approximated by the numerical scheme. Numerical discretizations should therefore be based on conservation forms and not on advective forms.

The development of a discontinuity from a smooth initial function is a typical property of nonlinear hyperbolic equations. For example, the formation of *shocks* in fluid flows is described by such equations.

Conservative Space Differencing

A discontinuous solution of (1.27) can be interpreted as a vanishing viscosity ($d \rightarrow 0$) solution of (1.29). In fact, numerically little difference exists between solving (1.29) with d small positive or $d = 0$. With $d > 0$, no matter how small, solutions are differentiable and the advective form does have a meaning. Still, the conservative form should be used for discretization and this even is advocated for d moderately small. With a conservative semi-discrete scheme

$$w'_j(t) = \frac{1}{h} \left(f(w_{j-\frac{1}{2}}(t)) - f(w_{j+\frac{1}{2}}(t)) \right), \quad (1.32)$$

replacing (1.28), we can expect that a steep traveling front or shock is computed in the correct location, whereas a non-conservative scheme does not guarantee this at all. In LeVeque (1992, Fig. 12.1) this is illustrated for the inviscid Burgers equation and our Figure 1.3 illustrates it for the adsorption problem (1.14).

It is easy to demonstrate that a conservative scheme (1.32) will give the correct shock speed; see also LeVeque (1992, Sect. 3.3, 12.3). Consider a discontinuous solution of (1.27) for $x \in \mathbb{R}$, $t \geq 0$, with shock speed s ,

$$u(x, t) = \begin{cases} u_L & \text{for } x < st, \\ u_R & \text{for } x > st. \end{cases} \quad (1.33)$$

Due to the finite speed of propagation and consistency of the flux approximations, we may assume that $w_{j-1/2}(t) = u_L$, $j \leq -J$ and $w_{j+1/2}(t) = u_R$, $j \geq J$ for all $t \in [0, T]$, with index J sufficiently large. Summation of (1.32) yields

$$\frac{d}{dt} \sum_{j=-J}^J h w_j(t) = f(u_L) - f(u_R). \quad (1.34)$$

On the other hand, from the integral relation (1.28) we get

$$\frac{d}{dt} \int_{-x_{J+\frac{1}{2}}}^{x_{J+\frac{1}{2}}} u(x, t) dx = f(u_L) - f(u_R). \quad (1.35)$$

The conservation property implies that the semi-discrete scheme puts a shock in the right location; otherwise the sum in (1.34) would vary in time with a wrong rate. The shock might be smeared by the semi-discrete method, but will be in the right position for evolving time. Since we have, in view of (1.33),

$$\frac{d}{dt} \int_{-x_{J+\frac{1}{2}}}^{x_{J+\frac{1}{2}}} u(x, t) dx = \frac{d}{dt} \left((-x_{J+\frac{1}{2}} + st) u_L + (x_{J+\frac{1}{2}} - st) u_R \right) = s(u_L - u_R),$$

it also follows that the shock speed is given by the so-called *Rankine-Hugoniot relation*

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R}. \quad (1.36)$$

The fact that the semi-discrete solution puts shocks in the right location gives confidence that it will converge to the exact discontinuous integral solution. It was shown by Lax & Wendroff (1960) that if a numerical scheme in conservation form converges to a function $v(x, t)$, then this is an integral solution of the conservation law. However, the integral form allows multiple solutions so that the physically relevant solution must be identified. This is done with so-called *entropy conditions* and the numerical solution thus should converge to the *entropy solution*, which corresponds to the vanishing viscosity solution mentioned above. In Harten, Hyman & Lax (1976) it was shown that convergence towards the correct solution is guaranteed if the numerical scheme satisfies certain monotonicity conditions. In practice the TVD property is usually sufficient for this. An account of this theory can be found in the books of Godlewski & Raviart (1996), LeVeque (1992, 2002) and Kröner (1997), and the review paper of Harten, Lax & van Leer (1983).

The choice of the cell-boundary approximations $w_{j\pm 1/2}$ in (1.32) determines the actual semi-discrete scheme, similar as for the flux form (1.1) for the linear problem. Assuming the flux function f to be monotone, that is, monotonically non-decreasing or non-increasing, we take, as in (1.3), (1.6),

$$w_{j+\frac{1}{2}} = \begin{cases} w_{j+\frac{1}{2}}^L = w_j + \psi(\theta_j)(w_{j+1} - w_j) & \text{if } f'(w_{j+\frac{1}{2}}^L) \geq 0, \\ w_{j+\frac{1}{2}}^R = w_{j+1} + \psi(1/\theta_{j+1})(w_j - w_{j+1}) & \text{if } f'(w_{j+\frac{1}{2}}^R) < 0, \end{cases} \quad (1.37)$$

with limiter function ψ satisfying (1.5). One can choose for example (1.7) or (1.11). As in the linear case, the limiters provide a suitable balance between the first-order upwind flux and higher-order fluxes.

With limiting we also have $w(t) \geq 0$ whenever $w(0) \geq 0$, together with monotonicity properties such as the TVD property. With nonlinear conservation laws this property is often crucial to obtain convergence towards the physically relevant solution. The TVD property can also be derived for the time-stepping schemes similar as for the linear problem, based on TVD for the explicit Euler method.¹⁾

Remark 1.6 For a non-monotone flux function f the numerical fluxes are not well defined by (1.37). In this more general case the early scheme of Godunov (1959) is applicable, but this reduces to first-order upwind if f is a monotone function. Modern schemes for conservation laws aim at a higher order in smooth regions. A well-known example generalizing (1.37) is due to Engquist & Osher (1980). Originally this was also formulated as a first-order scheme, but we can use the left and right states from the expressions in (1.37) to get higher order. For scalar problems this scheme can be regarded

¹⁾ For example, if f is monotonically increasing we have $w_{j\pm 1/2} = w_{j\pm 1/2}^L$ and $w_j + \frac{\tau}{h} \left(f(w_{j-\frac{1}{2}}) - f(w_{j+\frac{1}{2}}) \right) = w_j + \frac{\tau}{h} f'(v_j) \left(1 - \psi(\theta_{j-1}) + \theta_j^{-1} \psi(\theta_j) \right) (w_{j-1} - w_j)$ with v_j some intermediate point between $w_{j-1/2}$ and $w_{j+1/2}$, and then apply Lemma 1.5.

as a splitting of $f(u)$ into $f(u) = f^+(u) + f^-(u)$ with

$$f^+(u) = f(u^*) + \int_{u^*}^u \max(f'(v), 0) dv, \quad f^-(u) = \int_{u^*}^u \min(f'(v), 0) dv,$$

where u^* is arbitrary, for example $u^* = 0$. The numerical flux $f_{j+1/2}(w)$, to be inserted in (1.32) instead of $f(w_{j+1/2})$, can then be chosen as

$$f_{j+\frac{1}{2}}(w) = f^+(w_{j+\frac{1}{2}}^L) + f^-(w_{j+\frac{1}{2}}^R),$$

where $w_{j+1/2}^L$ and $w_{j+1/2}^R$ denote a left state and right state, respectively, which can be taken as in (1.37). An equivalent form is given by

$$f_{j+\frac{1}{2}}(w) = \frac{1}{2} \left(f(w_{j+\frac{1}{2}}^L) + f(w_{j+\frac{1}{2}}^R) \right) - \frac{1}{2} \int_{w_{j+\frac{1}{2}}^L}^{w_{j+\frac{1}{2}}^R} |f'(v)| dv.$$

For a monotone flux function we simply regain (1.37).

The extension of such forms to systems of conservation laws is an important but specialized subject, outside our scope. Good introductions are provided by LeVeque (1992, 2002) and Godlewski & Raviart (1996). \diamond

Numerical Example: the Buckley-Leverett Equation

We will use the Buckley-Leverett equation, defined by the conservation law (1.27) with

$$f(u) = \frac{cu^2}{cu^2 + (1-u)^2}, \quad (1.38)$$

for a numerical illustration. This equation provides a simple model for the flow of two immiscible fluids in a porous medium and has applications in oil-reservoir simulation. The unknown u then represents the saturation of water in an oil reservoir and lies between 0 and 1. The constant $c > 0$ gives the mobility ratio of the two fluid components.

We consider this problem for $c = 3$, $0 < t \leq \frac{1}{4}$ and $0 < x < 1$ with a periodic boundary condition. The initial function is given by the block-profile

$$u(x, 0) = \begin{cases} 0 & \text{for } 0 < x \leq \frac{1}{2}, \\ 1 & \text{otherwise.} \end{cases}$$

The flux function is monotonically increasing for $0 \leq u \leq 1$, and thus the flow is from left to right. The solution consists of two shocks followed by rarefaction waves. A description of the analytic form can be found in LeVeque (1992, Sect. 4.2).

A uniform grid has been used with grid points $x_j = jh$ for $j = 1, \dots, m$, $h = 1/m$. For spatial discretization we used the limited conservative scheme (1.32) based on (1.37) with limiter (1.7). In this example limiting is crucial;

the use of the non-limited standard third-order upwind-biased scheme leads here to qualitatively wrong solutions with large negative values in the wake of the right shock. With limiting the numerical solutions are accurate, only the shocks are somewhat diffused over 3 grid cells.

The resulting semi-discrete system $w'(t) = F(w(t))$ is integrated by the implicit 2-step BDF method (II.3.11) and its explicit counterpart (II.3.13), which is usually called the extrapolated BDF2 scheme. To solve the algebraic system for the implicit method, a Newton-type iteration was used with Jacobian matrix corresponding to first-order upwind discretization; see Hundsdorfer (2001) for some implementation tests. It is stressed that per step the implicit method is much more costly than the explicit one. Hence the implicit method can only be efficient if it allows much larger time steps than the explicit method. The implicit method is unconditionally stable in the sense of von Neumann, whereas stability of the explicit method imposes a maximal Courant number of 0.5, approximately.

We use this example to illustrate once more the importance of monotonicity. Due to the monotonicity restriction, the implicit BDF2 method cannot be used with large time steps if undershoots and overshoots are to be avoided. In the Figures 1.8 and 1.9 the numerical solutions at time $t = \frac{1}{4}$ are plotted as a function of x with solid lines. Dashed lines indicate a time-accurate reference solution on the same grid; this corresponds to the exact solution of the semi-discrete system. In Figure 1.8, where $\tau = 1/400$ and $h = 1/100$, we observe little difference between the implicit and explicit solutions and both are close to the reference solution. However, if the step size is increased to $\tau = 1/200$ we see from Figure 1.9 that now the explicit solution becomes unstable, but at the same time the implicit solution becomes very inaccurate: both the shock speed and the shock height are no longer correct. This is due to under- and overshoots after the shocks.

With linear convection, $f(u) = u$, the same phenomenon was already observed for the implicit trapezoidal rule and the implicit BDF2 method. Apparently, if the solution has steep gradients, then an implicit method gives

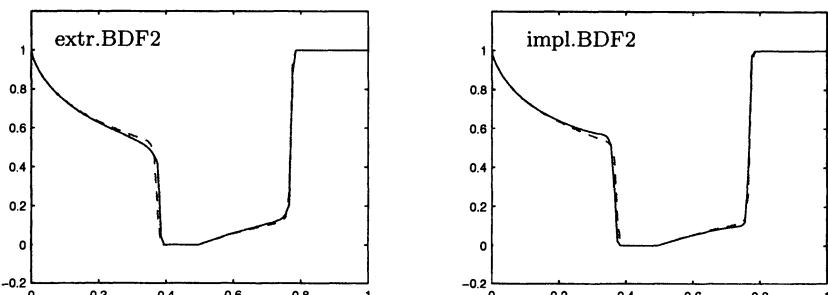


Fig. 1.8. BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $h = 1/100$ and $\tau = 1/400$. The dashed line is a time-accurate semi-discrete solution.

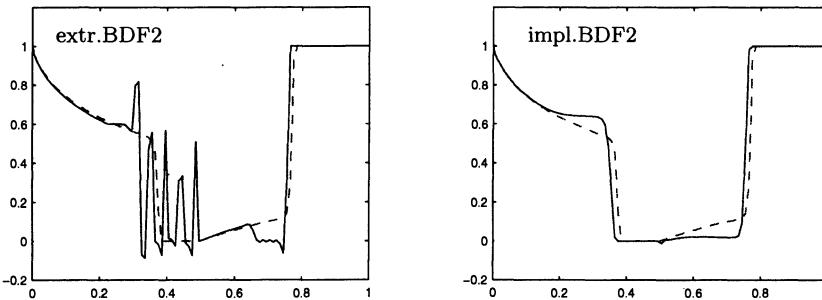


Fig. 1.9. BDF2 solutions at $t = 1/4$ for the Buckley-Leverett equation with $h = 1/100$ and $\tau = 1/200$. The dashed line is a time-accurate semi-discrete solution.

poor results whenever the step sizes are significantly larger than those that can be taken with an explicit method. This disappointing qualitative behaviour of implicit methods is due to their loss of monotonicity for large step sizes.

2 Direct Space-Time Advection Discretizations

In Section I.6 we already briefly considered some fully discrete *direct space-time* (DST) discretizations for advection problems. With such schemes space and time are discretized simultaneously instead of separately as in the MOL approach. The advantage of DST discretization is that we can give schemes a proper characteristic bias which enhances accuracy. The explicit Courant-Isaacson-Rees scheme (I.6.5) and Lax-Wendroff scheme (I.6.7) are well-known examples. In this section general schemes of this type will be discussed first for the advection model problem with constant coefficients. Further we consider explicit DST schemes in combination with flux limiting and we will briefly discuss unconditional stability for explicit schemes obtained by shifting of stencils.

2.1 Optimal-Order DST Schemes

Consider once more the model equation $u_t + au_x = 0$ with constant velocity a and given initial value $u(x, 0)$. Boundary conditions will be treated as for the semi-discrete schemes; for the moment assume that we have a pure initial value problem, for example with spatial periodicity. Taking grid points $x_j = jh$ and $t_n = n\tau$, we will obtain fully discrete approximations $w_j^n \approx u(x_j, t_n)$. The general explicit two-level DST scheme, using r grid points to the left and s to the right, is given by

$$w_j^{n+1} = \sum_{k=-r}^s \gamma_k w_{j+k}^n, \quad (2.1)$$

with coefficients γ_k depending on the Courant number $\nu = a\tau/h$. Examples are the first-order Courant-Isaacson-Rees scheme (I.6.5) with $r = 1, s = 0$, for $a > 0$, and the second-order Lax-Wendroff scheme (I.6.7) with $r = s = 1$. The local (space-time) truncation errors ρ_j^n are defined by

$$u(x_j, t_{n+1}) = \sum_{k=-r}^s \gamma_k u(x_{j+k}, t_n) + \tau \rho_j^n. \quad (2.2)$$

We can consider order conditions to have $\rho_j^n = \mathcal{O}(h^p)$ for smooth solutions with ν fixed. If the scheme is stable this leads to a global error bound $u(x_j, t_n) - w_j^n = \mathcal{O}(h^p)$; see Section I.6.2 for the general framework.

As for the semi-discrete schemes in Section I.3.2, it can be shown that with this stencil an order $p = r + s$ is possible. This can be demonstrated by setting up the order conditions. However, an alternative and easier approach is to reinterpret the method. Since we know that $u(x_j, t_{n+1}) = u(x_j - \tau a, t_n)$, formula (2.1) can also be viewed as an interpolation procedure. The optimal-order schemes correspond to Lagrange interpolation: given function values v_{j+k} at x_{j+k} , $k = -r, \dots, s$, the interpolation polynomial is

$$V(x) = \sum_{k=-r}^s L_{j+k}(x) v_{j+k}, \quad L_{j+k}(x) = \prod_{\substack{l=-r \\ l \neq k}}^s \frac{x - x_{j+l}}{x_{j+k} - x_{j+l}}.$$

For the advection scheme this gives the coefficients

$$\gamma_k(\nu) = L_{j+k}(x_j - \tau a) = \prod_{\substack{l=-r \\ l \neq k}}^s \frac{l + \nu}{l - k}. \quad (2.3)$$

By inserting $u(x_j, t_{n+1}) = u(x_j^*, t_n)$, $x_j^* = x_j - \tau a$ in (2.2), the truncation error follows from the polynomial interpolation error formula,

$$\tau \rho_j^n = \frac{1}{(p+1)!} \left(\prod_{k=-r}^s (x_j^* - x_{j+k}) \right) \frac{\partial^{p+1} u(\hat{x}, t_n)}{\partial x^{p+1}} \quad (2.4)$$

with $p = r + s$ and $\hat{x} \in (x_{j-r}, x_{j+s})$. By taking the factor $x_j^* - x_j = -\tau a$ out of the product, we see that $\rho_j^n = \mathcal{O}(h^p)$. Moreover we see that the error vanishes if ν is an integer between $-r$ and s . This last point is a typical feature of DST formulas, not shared by MOL schemes.

Remark 2.1 A fluid flow can be either considered from the *Lagrangian* point of view where one moves along with the fluid, or from the *Eulerian* point of view where the position of the observer is fixed in space. Schemes obtained by tracing characteristics and using interpolation are often called *semi-Lagrangian* schemes. Here ‘semi’ indicates that after each time step the solution is mapped back onto the Eulerian grid. \diamond

Stability Restrictions

The advection scheme (2.1) is stable, in the strict sense of von Neumann (see Section I.6.4), if the condition

$$\left| \sum_{k=-r}^s \gamma_k(\nu) e^{2i\omega k} \right| \leq 1 \quad \text{for all } -\frac{1}{2}\pi \leq \omega \leq \frac{1}{2}\pi,$$

is satisfied. This condition implies L_2 -stability for pure initial value problems with spatial region $\Omega = \mathbb{R}$ and for problems on a bounded interval Ω with periodic boundary conditions. The stability restrictions of the optimal-order schemes are given by the following fundamental result of Strang (1962) (stability for $s \leq r \leq s+2$) and Iserles & Strang (1983) (instability of the other schemes).

Theorem 2.2 Consider for $a > 0$, $\nu = a\tau/h$, the advection scheme (2.1) with order $p = r+s$. For this optimal-order scheme there holds

$$\begin{aligned} r = s &\implies \text{stability for } 0 \leq \nu \leq 1, \\ r = s+1 &\implies \text{stability for } 0 \leq \nu \leq 1, \\ r = s+2 &\implies \text{stability for } 0 \leq \nu \leq 2. \end{aligned}$$

With the other choices, $r \neq s, s+1, s+2$, we have instability for $\nu > 0$ small.

Of course, for $u_t + au_x = 0$ with $a < 0$ the situation is similar; we then get stability for $s = r, r+1, r+2$. Hence the central scheme with $r = s$ is in fact stable for $-1 \leq \nu \leq 1$. The stability for $a > 0$ with $r = s, s+1, s+2$ will be demonstrated below; for the instability result we refer to Iserles & Strang (1983) and Iserles & Nørsett (1991). In these references, and in Jeltsch (1988), generalizations of these results can be found for implicit schemes and multistep schemes.

Remark 2.3 In Section I.3.2 a similar result was formulated for stability of spatial discretizations. Not surprisingly, there is a close connection. If we write the fully discrete advection scheme (2.1) as $w_{n+1} = Bw_n$, with $w_n = (w_j^n)$ on a fixed spatial mesh, then

$$B = I + \tau A + \mathcal{O}(\tau^2), \quad \tau \rightarrow 0, \quad h > 0 \text{ fixed}.$$

Hence for a given $h > 0$ we can view (2.1), for theoretical purposes, as a consistent time stepping formula for the ODE system $w'(t) = Aw(t)$, which is the corresponding semi-discrete system. Stability of (2.1) for $\nu \rightarrow 0$ thus guarantees that the semi-discrete system is stable as well. Likewise, instability of the semi-discrete system implies that the fully discrete scheme is also unstable. \diamond

Proof of stability. To prove the sufficiency in Theorem 2.2, we closely follow the arguments of Strang (1962). For the main part we will concentrate on the central scheme with $r = s$. Stability for $r = s + 1, s + 2$ will follow from this case.

Stability can be reformulated in terms of interpolation. The polynomial interpolating $e^{i\theta x}$, $\theta = 2\omega/h$, with abscissae x_{-r}, \dots, x_s , is given by

$$P_{-r,s}(x, \theta) = \sum_{k=-r}^s L_k(x) e^{i\theta x_k}, \quad L_k(x) = \prod_{\substack{l=-r \\ l \neq k}}^s \frac{x - x_l}{x_k - x_l}. \quad (2.5)$$

For convenience of notation the central point x_j is taken here as x_0 . Moreover we take $h = 1$, $x_j = j$; this is just a matter of scaling. Then to show that the central scheme, with $r = s$, is stable for Courant numbers up to 1, we have to demonstrate that

$$|P_{-s,s}(x, \theta)| \leq 1 \quad \text{whenever } \theta \in [-\pi, \pi] \text{ and } x \in [-1, 1]. \quad (2.6)$$

Because of symmetry we may restrict ourselves to $\theta \in [0, \pi]$ and $x \in [0, 1]$.

So, consider the complex polynomial $P = P_{-s,s}$ with abscissae $-s, \dots, s$. We write this as $P(x, \theta) = C(x, \theta) + iS(x, \theta)$ with C, S real. Then $C(x, \theta)$ is an even polynomial of degree $2s$ in x interpolating $\cos(\theta x)$, and $S(x, \theta)$ is an odd polynomial of degree $2s - 1$ in x interpolating $\sin(\theta x)$. If $\theta = 0$, we have $P = 1$ for all x . The basic idea of the proof is to show that $\frac{\partial}{\partial \theta}(C^2 + S^2) < 0$ for $\theta \in (0, \pi)$ and $x \in (0, 1)$.

As a first step, note that C and S share the following property of $\cos(\theta x)$ and $\sin(\theta x)$:

$$\frac{\partial C}{\partial \theta} = -xS.$$

Because both sides are polynomials of degree $2s$ in x interpolating $x \sin(\theta x)$ on the $2s + 1$ abscissae $-s, \dots, s$, they thus are identical. As a consequence we have

$$\frac{\partial}{\partial \theta}(C^2 + S^2) = 2\left(C \frac{\partial C}{\partial \theta} + S \frac{\partial S}{\partial \theta}\right) = -2S\left(xC - \frac{\partial S}{\partial \theta}\right). \quad (2.7)$$

Next we consider the factors on the right of (2.7) separately, and first show that

$$xC - \frac{\partial S}{\partial \theta} = \frac{(-1)^s}{(2s)!} \left(2 \sin \frac{1}{2}\theta\right)^{2s} \prod_{k=-s}^s (x - k), \quad (2.8)$$

which is positive for $\theta \in (0, \pi)$, $x \in (0, 1)$. To demonstrate (2.8), note that

$$xC - \frac{\partial S}{\partial \theta} = \sum_{k=-s}^s (x - k) L_k(x) \cos(\theta k).$$

Comparing this with the right-hand side of (2.8), we see that both expressions vanish at the points $x = -s, \dots, s$. Since we are dealing with polynomials of

degree $2s + 1$ in x , it suffices to show that the coefficient of x^{2s+1} is equal. For the last expression this coefficient is

$$\sum_k \frac{1}{\prod_{l \neq k} (k-l)} \cos(\theta k) = \sum_k \frac{(-1)^{s-k}}{(s-k)!(s+k)!} \cos(\theta k).$$

Using the binomial theorem it follows that this equals

$$\sum_k \frac{(-1)^{s-k}}{(2s)!} \binom{2s}{s+k} \operatorname{Re} e^{i\theta k} = \frac{1}{(2s)!} \operatorname{Re} (e^{\frac{1}{2}i\theta} - e^{-\frac{1}{2}i\theta})^{2s} = \frac{(-1)^s}{(2s)!} \left(2 \sin \frac{1}{2}\theta\right)^{2s}.$$

This coefficient corresponds to the one on the right-hand side of (2.8).

The other factor in (2.7) is S . Since we have $S(x, \theta) = \sin(\theta x) + \mathcal{O}(\theta^{2s+1})$, it is obvious that $S(x, \theta) > 0$ for all $x \in (0, s]$ if $\theta > 0$ is sufficiently small. Using the interpolation property $S(x, \theta) = \sin(\theta x)$ for $x = 0, 1, \dots, s$, we can now regard for increasing θ the location of the zeros in x for $x > 0$. For small $\theta > 0$ these zeros are located on the right of $x = s$. Since $S(1, \theta) = \sin \theta$ remains positive for $\theta < \pi$, it is clear by continuity that

$$S(x, \theta) > 0 \quad \text{for all } x \in (0, 1] \text{ and } \theta \in (0, \pi). \quad (2.9)$$

With the relations (2.7)–(2.9), stability of the optimal-order central scheme now follows directly. Taking $\theta = 0$, we have $P(x, 0) = 1$ for all x . From the above we see that $|P(x, \theta)|$ is decreasing in θ whenever $0 < \theta < \pi$ and $0 < x < 1$. Using symmetry, we therefore obtain (2.6) for arbitrary $s \geq 1$.

For the upwind formula with $r = s + 2$, we have

$$P_{-s-2,s}(x, \theta) = e^{-i\theta} P_{-s-1,s+1}(x+1, \theta).$$

Hence $|P_{-s-2,s}(x, \theta)| \leq 1$ whenever $\theta \in [-\pi, \pi]$ and $-2 \leq x \leq 0$.

Finally, for $r = s + 1$ we have by the Neville formula, see for instance Stoer & Bulirsch (1980, Sect. 2.1.2),

$$P_{-s-1,s}(x, \theta) = \frac{x - x_{-s-1}}{x_s - x_{-s-1}} P_{-s,s}(x, \theta) + \frac{x - x_s}{x_{-s-1} - x_s} P_{-s-1,s-1}(x, \theta).$$

If $-1 \leq x \leq 0$ then both $|P_{-s,s}(x, \theta)|$ and $|P_{-s-1,s-1}(x, \theta)|$ are bounded by 1 for all θ , and thus also $|P_{-s-1,s}(x, \theta)| \leq 1$.

2.2 A Non-oscillatory Third-Order DST Scheme

Assuming $a > 0$, we now focus on the optimal-order 4-point advection scheme (2.1) with $r = 2$ and $s = 1$, involving two upwind points and one downwind point. First we again take a constant; variable velocities are dealt with later.

The main reason to focus on this stencil is the use of limiting, where it appears in a natural way. We thus consider the third-order scheme

$$\begin{aligned} w_j^{n+1} = & -\frac{1}{6}\nu(1-\nu^2)w_{j-2}^n + \frac{1}{2}\nu(2-\nu)(1+\nu)w_{j-1}^n \\ & + \frac{1}{2}(2-\nu)(1-\nu^2)w_j^n - \frac{1}{6}\nu(2-\nu)(1-\nu)w_{j+1}^n, \end{aligned} \quad (2.10)$$

whose coefficients follow directly from (2.3) and which is stable for $\nu \in [0, 1]$ according to Theorem 2.2. As mentioned above, the scheme can also be interpreted as a semi-Lagrangian scheme where the characteristics are traced backward in time from t_{n+1} to t_n and cubic polynomial interpolation is used.

Semi-Lagrangian schemes can be made non-oscillatory by choosing a monotone interpolation procedure, see e.g. Williamson & Rasch (1989). Traditional semi-Lagrangian schemes are based on the advective form. We will mainly consider the conservative form which ensures mass conservation for variable velocity fields. For (2.10) the conservative flux form reads

$$w_j^{n+1} = w_j^n + \frac{\tau}{h} \left(\bar{f}_{j-\frac{1}{2}}^n - \bar{f}_{j+\frac{1}{2}}^n \right), \quad (2.11)$$

with the fluxes $\bar{f}_{j+\frac{1}{2}}^n$ given by

$$\bar{f}_{j+\frac{1}{2}}^n = a \left[-\frac{1}{6}(1-\nu^2)w_{j-1}^n + \frac{1}{6}(1+\nu)(5-2\nu)w_j^n + \frac{1}{6}(2-\nu)(1-\nu)w_{j+1}^n \right].$$

As indicated in Remark 2.3, there exists a close connection between the DST scheme (2.10) and the semi-discrete third-order upwind-biased scheme defined by (1.9) with $\kappa = 1/3$. For h fixed and ν tending to zero, the DST scheme becomes identical to the corresponding MOL scheme with exact time integration. On the other hand, from the characteristic interpretation it follows that (2.10) becomes more accurate the closer ν is to 1. Putting $\nu = 1$ gives $w_j^{n+1} = w_{j-1}^n$, and in this case we exactly follow the characteristic and hence obtain the exact solution. The advantage can also be illustrated by considering the local truncation error, see (2.4). For the scheme (2.10) it is given by

$$\rho_j^n = \frac{1}{24}(1-\nu)(2-\nu)(1+\nu)a h^3 u_{xxxx}(x_j, t_n) + (1-\nu)\mathcal{O}(\tau h^3),$$

which vanishes as $\nu \rightarrow 1$. Hence the error for the DST scheme will become smaller for larger ν , up to $\nu = 1$, whereas in a corresponding MOL scheme the error will grow with increasing ν due to growing temporal inaccuracy of the numerical ODE method.

Flux Limiting

Like its semi-discrete counterpart, the third-order DST scheme (2.11) is not positive and it may produce under- and overshoot. We consider flux limiting

to remedy this. The derivation here is close to the derivation for the semi-discrete scheme given in Section 1.1. For convenience of notation, we omit the superscript n in the following, thus denoting $w_j = w_j^n$. Let

$$c_0 = \frac{1}{6}(2 - \nu)(1 - \nu), \quad c_1 = \frac{1}{6}(1 - \nu^2).$$

The fluxes for (2.11) then can be written as

$$\bar{f}_{j+\frac{1}{2}} = a[w_j + (c_0 + c_1\theta_j)(w_{j+1} - w_j)],$$

where θ_j denotes the ratio (1.2). As in the semi-discrete case, we next consider for limiting the general form

$$\bar{f}_{j+\frac{1}{2}} = a[w_j + \psi(\theta_j)(w_{j+1} - w_j)], \quad (2.12)$$

with the limiter function ψ yet to be determined and dependent on the ratio θ_j and the Courant number ν . Since $w_{j+1} - w_j = \theta_j^{-1}(w_j - w_{j-1})$, the total scheme can be written as

$$w_j^{n+1} = (1 - \nu\beta_j)w_j + \nu\beta_j w_{j-1}, \quad \beta_j = 1 - \psi(\theta_{j-1}) + \frac{1}{\theta_j}\psi(\theta_j), \quad (2.13)$$

from which we immediately conclude that the scheme is positive iff

$$0 \leq \nu\beta_j \leq 1. \quad (2.14)$$

The limiter function ψ should now be chosen such that this requirement holds for arbitrary θ_j . On the other hand the original non-limited flux should be recovered for ratios θ_j close to 1, since this is the generic situation for a smooth profile. Let us consider

$$\psi(\theta) = \max \left(0, \min \left(1, c_0 + c_1\theta, \mu\theta \right) \right), \quad (2.15)$$

where μ is a positive parameter which is still free. This limiter is the counterpart of (1.7) and it satisfies $0 \leq \psi(\theta) \leq 1$ and $0 \leq \psi(\theta)/\theta \leq \mu$. Hence $0 \leq \beta_j \leq 1 + \mu$ and the condition for positivity thus is

$$\nu(1 + \mu) \leq 1. \quad (2.16)$$

The free parameter μ is still to be chosen. Large values of μ give more accurate results, as more often the original third-order flux will be used. However, larger values for μ result in smaller allowable Courant numbers in (2.16). In the MOL approach $\mu = 1$ has been advocated. The fact that large values of μ are inefficient in the MOL approach is due to the ODE solvers. These require smaller step sizes, that is, smaller ν , to maintain positivity in the time integration if μ is increased, just as with the corresponding positivity condition (1.23) for the explicit Euler method. With the DST scheme we have

the natural flexibility to let μ depend on the Courant number. Taking μ as large as possible within the positivity constraint gives

$$\mu = (1 - \nu)/\nu \quad \text{for } 0 \leq \nu \leq 1. \quad (2.17)$$

Now the DST scheme is positive for all $\nu \leq 1$ and thus the most accurate region for this scheme where ν is near 1 can be included. Moreover, for small Courant numbers, μ gets large which increases the accuracy compared to fixed $\mu = 1$ with less numerical diffusion, leading, for example, to better peak values.

Notice that as long as the positivity condition (2.16) is satisfied, we can apply Lemma 1.5 to (2.13) to prove the TVD property, similar as for the explicit Euler method (1.21).

With a variable velocity $a(x, t)$ the flux formulas for $\bar{f}_{j+1/2}$ will be applied using $a_{j+1/2} \approx a(x_{j+1/2}, t)$. If $a_{j+1/2} < 0$ the stencil $\{x_{j-1}, x_j, x_{j+1}\}$ used in (2.12) should be reflected around the point $x_{j+1/2}$ to maintain the upwind character. We then get, similar as (1.6),

$$\bar{f}_{j+\frac{1}{2}} = a_{j+\frac{1}{2}} \left[w_{j+1} + \psi\left(\frac{1}{\theta_{j+1}}\right)(w_j - w_{j+1}) \right]. \quad (2.18)$$

The order of consistency of the non-limited scheme for variable velocity equals two pointwise and three for cell-average values, see also Section I.4.5. Similar as for the MOL scheme, flux limiting itself might also lead to some order reduction, even when a is constant. However, this will only happen in regions where the solution is not smooth or near extrema. Elsewhere the θ_j will be close to one, so there the limiter is not active. Consequently, the effect of limiting on the global accuracy consists mainly of some numerical diffusion near extrema and steep gradients.

To illustrate the advantage of the limited DST scheme using (2.17) over the MOL approach we consider again the test problem $u_t + u_x = 0$ with $0 < x < 1$ and a periodic boundary condition for the initial profile $u(x, 0) = (\sin(\pi x))^{100}$. In Figure 2.1 limited and non-limited solutions are given at $t = 1$ for $h = 1/50$ with Courant numbers $\nu = \frac{5}{7}, \frac{1}{10}$. Also given are the corresponding semi-discrete (MOL) solutions without limiting and with

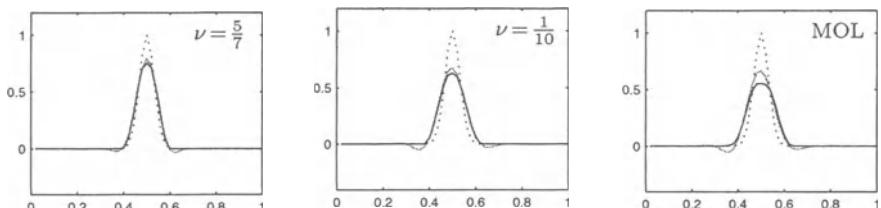


Fig. 2.1. Advection test for the DST scheme with Courant numbers $\nu = \frac{5}{7}, \frac{1}{10}$ and the semi-discrete (MOL) solution, $h = 1/50$. Exact solutions are dotted, limited solutions solid and non-limited solutions grey.

limiter (1.7). Note that the non-limited DST solution with $\nu = \frac{1}{10}$ is close to its semi-discrete counterpart, but with limiting the DST solution is more accurate due to the larger value of μ in the limiter.

Remark 2.4 The above linear DST scheme can be derived in various ways, see for instance Leonard (1988) (the QUICKEST scheme) and Bott (1992). The derivation here with the limiter (2.12), (2.17) is based on Hundsdorfer & Trompert (1994). Two other popular flux limiters, expressed in terms of $\phi(\theta) = 2\psi(\theta)/(1 - \nu)$,²⁾ are the MUSCL limiter

$$\phi(\theta) = \max \left(0, \min \left(2, \frac{1}{2}(1 + \theta), 2\theta \right) \right) \quad (2.19)$$

of van Leer (1979) and the ‘superbee’ limiter

$$\phi(\theta) = \max \left(0, \min \left(1, 2\theta \right), \min \left(\theta, 2 \right) \right) \quad (2.20)$$

of Roe (1985). The MUSCL limiter gives slightly more diffusion and clipping of peaks than (2.15), whereas the superbee limiter gives roughly the same peak values as (2.15) but better results for advection of a square wave. However, the superbee limiter is less accurate for smooth solutions, showing a strong tendency to turn smooth curves into staircase-shaped functions; see Zalesak (1987) for illustrations. This behaviour is similar to the MOL approach with the explicit Euler method, see Section 1.3, indicating an underlying linear instability. ◇

Remark 2.5 For nonlinear scalar advection equations $u_t + f(u)_x = 0$ the above schemes can be generalized in various ways. For schemes with limiters we refer to LeVeque (1992, 2002) and Kröner (1997). Such schemes become computationally complex for systems of conservation laws, but this is unavoidable to resolve complicated solutions with discontinuities.

On the other hand, there are also nonlinear hyperbolic equations where solutions remain smooth. In that case simpler schemes can be used. A well known scheme is the *MacCormack scheme*. This scheme of MacCormack (1969) has two stages, with alternating upwind and downstream approximations to achieve second-order accuracy,

$$\begin{aligned} w_j^* &= w_j^n + \frac{\tau}{h} (f(w_j^n) - f(w_{j+1}^n)), \\ w_j^{n+1} &= \frac{1}{2} (w_j^* + w_j^n) + \frac{\tau}{2h} (f(w_{j-1}^*) - f(w_j^*)). \end{aligned} \quad (2.21)$$

When applied to the linear problem $u_t + au_x = 0$ it reduces to the linear central Lax-Wendroff scheme (I.6.7). Hence for steep solutions it is prone

²⁾ For $\nu = 0$ these limiters expressed in terms of $\psi(\theta)$ are applicable in the MOL setting of Section 1. The MUSCL (Monotone Upwind Schemes for Conservation Laws) limiter then gives (1.10) with $\kappa = 0$. The form with $\phi(\theta) = 2\psi(\theta)/(1 - \nu)$ is traditional, based on the fact that then $\phi \equiv 1$ corresponds to the Lax-Wendroff scheme for which such limiters were derived originally.

to oscillatory behaviour. For smooth solutions the MacCormack scheme is second order in space and time. The scheme fits in the conservation form

$$w_j^{n+1} = w_j^n + \frac{\tau}{h} (\bar{f}_{j-\frac{1}{2}}^n - \bar{f}_{j+\frac{1}{2}}^n), \quad \bar{f}_{j+\frac{1}{2}}^n = \frac{1}{2} f(w_{j+1}^n) + \frac{1}{2} f(w_j^*). \quad (2.21)$$

Moreover this scheme is also directly applicable to systems. \diamond

2.3 Explicit Schemes with Unconditional Stability

As for semi-Lagrangian methods using the advective form, it is possible to obtain an explicit DST conservation form that is unconditionally stable. Again, this is a feature of DST schemes that is not shared by MOL schemes. Consider once more the flux form (2.11), now for the variable coefficient problem $u_t + (a(x, t)u)_x = 0$. For the computation of the fluxes $\bar{f}_{j+1/2}$ we can allow the stencil to vary with the Courant number. With large Courant numbers also the contribution to the flux of non-adjacent cells then should be taken into account. In this way unconditional stability can be achieved while maintaining explicitness and mass conservation of the scheme.

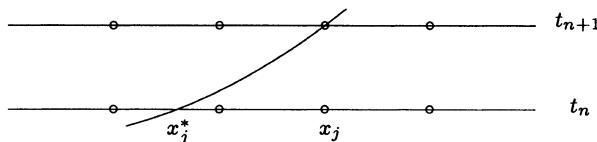
As a preliminary example, let us consider the first-order upwind scheme in advective form

$$w_j^{n+1} = w_j^n + \frac{\tau a_j}{h} (w_{j-1}^n - w_j^n), \quad (2.22)$$

where $a_j \approx a(x_j, t_n) > 0$. This formula can be interpreted in a semi-Lagrangian fashion. Let x_j^* be the departure point at $t = t_n$ of the characteristic that passes through (x_j, t_{n+1}) . We have $u(x_j, t_{n+1}) = u(x_j^*, t_n)$. Using the approximation $x_j^* = x_j - \tau a(x_j, t_n)$, formula (2.22) is obtained by linear interpolation in space at time level t_n . In this form the scheme (2.22) is stable for constant $a > 0$ with Courant number $\nu = \tau a/h \leq 1$. However, this Lagrangian interpretation makes it clear that we can use the scheme for large Courant numbers if the stencil is shifted appropriately,

$$w_j^{n+1} = w_{j-k}^n + \frac{\tau a_j}{h} (w_{j-k-1}^n - w_{j-k}^n),$$

where k is such that $x_j^* \in [x_{j-k-1}, x_{j-k}]$. Also the departure point might be computed more accurately, say by a Runge-Kutta method, leading to replacement of τa_j by $(x_j - x_j^*)$.



The same idea can be used for the first-order upwind scheme in conservation form, also known as the donor cell scheme. Omitting the upper index n ,

for notational convenience, this scheme is given by (2.11) with fluxes

$$\frac{\tau}{h} \bar{f}_{j+\frac{1}{2}} = \begin{cases} \nu_{j+\frac{1}{2}} w_j & \text{if } a_{j+\frac{1}{2}} \geq 0, \\ -\nu_{j+\frac{1}{2}} w_{j+1} & \text{if } a_{j+\frac{1}{2}} < 0. \end{cases}$$

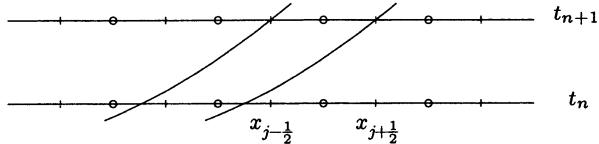
Here $\nu_{j+1/2} = |a_{j+1/2}| \tau / h$ is the local (absolute) Courant number at the cell boundary $x_{j+1/2}$. In this form the scheme is stable under the CFL restriction $\nu \leq 1$ if a is constant; for a variable the practical condition $\max_j \nu_{j+1/2} \leq 1$ is imposed. This stability restriction is avoided by taking

$$\frac{\tau}{h} \bar{f}_{j+\frac{1}{2}} = \begin{cases} \tilde{\nu}_{j+\frac{1}{2}} w_{j-k} + (w_{j-k+1} + \dots + w_j) & \text{if } a_{j+\frac{1}{2}} \geq 0, \\ -\tilde{\nu}_{j+\frac{1}{2}} w_{j+k+1} - (w_{j+1} + \dots + w_{j+k}) & \text{if } a_{j+\frac{1}{2}} < 0, \end{cases}$$

where

$$k = k_{j+\frac{1}{2}} = \lfloor \nu_{j+\frac{1}{2}} \rfloor \quad \text{and} \quad \tilde{\nu}_{j+\frac{1}{2}} = \nu_{j+\frac{1}{2}} - k_{j+\frac{1}{2}},$$

and $\lfloor \nu \rfloor$ denotes the largest integer $\leq \nu$. Inserting this into (2.11), we see that for constant a the same formula is applied as for Courant numbers ≤ 1 , with a shift over k grid points. Therefore the scheme will be unconditionally stable. In a semi-Lagrangian interpretation $\tau a_{j+1/2}$ is seen to approximate the distance $(x_{j+1/2} - x_{j+1/2}^*)$ with $x_{j+1/2}^*$ the departure point at $t = t_n$ of the characteristic that passes through $(x_{j+1/2}, t_{n+1})$.



Instead of the first-order fluxes we can use the third-order fluxes with limiting. Then the terms $\tilde{\nu}_{j+1/2} w_{j-k}$ are replaced by their third-order counterparts. The full formulas are as follows. If $a_{j+1/2} \geq 0$ we have

$$\frac{\tau}{h} \bar{f}_{j+\frac{1}{2}} = \tilde{\nu}_{j+\frac{1}{2}} \left[w_{j-k} + \psi(\tilde{\nu}_{j+\frac{1}{2}}, \theta_{j-k}) (w_{j-k+1} - w_{j-k}) \right] + \sum_{l=j-k+1}^j w_l,$$

with $\tilde{\nu}_{j+1/2}$ and $k_{j+1/2}$ as before. Here the dependence of the limiter on the Courant number is denoted explicitly, so with $\psi(\nu, \theta) = c_0(\nu) + c_1(\nu)\theta$ the original non-limited scheme is obtained shifted over k grid points. For $a_{j+1/2} < 0$ the corresponding flux formula reads

$$\frac{\tau}{h} \bar{f}_{j+\frac{1}{2}} = -\tilde{\nu}_{j+\frac{1}{2}} \left[w_{j+k+1} + \psi\left(\tilde{\nu}_{j+\frac{1}{2}}, \frac{1}{\theta_{j+k+1}}\right) (w_{j+k} - w_{j+k+1}) \right] - \sum_{l=j+1}^{j+k} w_l.$$

For variable velocities one should use sufficiently accurate approximations for the departure points $x_{j+1/2}^*$ and use the above formulas with

$$\nu_{j+\frac{1}{2}} = \frac{1}{h} |x_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}^*|,$$

that is, $\tau a_{j+1/2} = x_{j+1/2} - x_{j+1/2}^*$. Schemes of this type for nonlinear conservation laws were introduced by LeVeque (1982) and Brenier (1984). Applications for linear convection on a sphere, where the longitudinal mesh widths become very small near the poles, can be found in Petersen et al. (1998).

Some Practical Aspects

For pure 1D advection calculations the explicit DST approach is very attractive. The characteristic bias of DST schemes is an advantage for accuracy compared with MOL schemes and in addition DST schemes can be made unconditionally stable without loss of accuracy. However, boundary conditions then may require adjustments more complicated than for MOL schemes. A clear advantage of the MOL approach is the much greater flexibility for combining advection, diffusion and reaction calculations. Use of a DST advection scheme then necessitates the use of time splitting, which will be discussed in the next chapter. A generalization of the DST formula (2.10) to two dimensions in conservation form, with an FCT procedure for limiting, was considered in Rasch (1994). Related schemes can be found in LeVeque (1996, 2002). For two- and three-dimensional advection calculations we will mainly use dimensional time splitting. We therefore postpone giving numerical results until time splitting has been discussed.

3 Implicit Spatial Discretizations

This section contains a review of some implicit spatial discretizations that generalize well-known schemes for stationary advection-diffusion problems. The object here is to select schemes that perform well on the 1D time-dependent advection-diffusion model problem with constant coefficients and uniform grids, without consideration of boundary conditions. The most interesting schemes will be described in later sections in more generality via a finite volume or finite element approach. In view of these extensions we will restrict ourselves here mainly to 3-point schemes.

For the advection-diffusion model equation

$$u_t + a u_x = d u_{xx} + s(x, t), \quad (3.1)$$

with constant $a \in \mathbb{R}$, $d \geq 0$ and source term $s(x, t)$, we consider spatial discretizations of the form

$$\sum_{k=-r}^r \beta_k w'_{j+k}(t) = h^{-2} \sum_{k=-r}^r \alpha_k w_{j+k}(t) + \sum_{k=-r}^r \beta_k g_{j+k}(t), \quad (3.2)$$

where $w_j(t) \approx u(x_j, t)$, $x_j = jh$ and $g_j(t) = s(x_j, t)$. The coefficients β_k, α_k will in general depend on a, d and also on the mesh width h . The coefficients in

front of w'_{j+k} and g_{j+k} are chosen to be the same, so that in the degenerate case $a = d = 0$ we obtain $w'_j(t) = g_j(t)$. With this choice, moreover, the truncation error is the same for time-dependent problems as for stationary problems. In the following we will use the normalization

$$\sum_{k=-r}^r \beta_k = 1. \quad (3.3)$$

The above spatial discretization will be called *implicit* if some β_k ($k \neq 0$) is not equal to zero. Implicit discretizations arise in a natural way from finite elements or certain finite volume derivations. Here the term implicit refers to the fact that the time derivatives are not explicitly given. This is related to the term implicit for ODE methods, where it means that the temporal increments are not explicitly given. Note that an explicit ODE method applied to an implicit spatial discretization will yield a fully discrete scheme with some amount of implicitness (see also Section 3.5). We will review some implicit spatial formulas with $r = 1$ that are discussed in Morton (1996, Ch. 3) for stationary problems, but here the emphasis will be on the time-dependent case, which has some distinct features with respect to stability and monotonicity.

3.1 Order Conditions

In vector notation the above spatial discretization (3.2) will be written as

$$Bw'(t) = Aw(t) + Bg(t), \quad (3.4)$$

with matrices $A = (a_{ij}) = (h^{-2}\alpha_{j-i})$ and $B = (b_{ij}) = (\beta_{j-i})$. Let u_h be the restriction of the exact solution u to the grid. We consider the spatial truncation error

$$\sigma_h(t) = Bu'_h(t) - Au_h(t) - Bg(t). \quad (3.5)$$

Let ξ_k, η_k be defined for $k \geq 0$ as

$$\xi_k = \sum_{j=-r}^r j^k \alpha_j, \quad \eta_k = \sum_{j=-r}^r j^k \beta_j,$$

with convention $0^0 = 1$. By a Taylor series expansion we find that the truncation error in a point (x_j, t) equals

$$\sigma_{h,j}(t) = h^{-2}(C_0 u + h C_1 u_x + h^2 C_2 u_{xx} + h^3 C_3 u_{xxx} + \dots) \Big|_{(x_j, t)} \quad (3.6)$$

with error coefficients

$$\begin{aligned} C_0 &= -\xi_0, & C_1 &= -\xi_1 - ah \eta_0, \\ C_k &= \frac{-1}{k!} \left(\xi_k + k ah \eta_{k-1} - k(k-1) d \eta_{k-2} \right), & k \geq 2. \end{aligned}$$

With the normalization (3.3), the discretization has order q if the truncation error satisfies $\sigma_h = \mathcal{O}(h^q)$ for smooth solutions, that is,

$$C_k = \mathcal{O}(h^{q+2-k}) \quad \text{for } k = 0, 1, \dots, q+2. \quad (3.7)$$

In the following we consider 3-point schemes ($r = 1$), in which case we simply have $\xi_k = \alpha_1 + (-1)^k \alpha_{-1}$, $\eta_k = \beta_1 + (-1)^k \beta_{-1}$ for $k \geq 1$. Further, $\eta_0 = 1$ according to (3.3). Throughout this section it will be required explicitly that $C_0 = C_1 = 0$, so that the discretization is exact if $u_{xx} \equiv 0$. Thus

$$\alpha_{-1} + \alpha_0 + \alpha_1 = 0, \quad \alpha_{-1} - \alpha_1 = ah. \quad (3.8)$$

Both conditions (3.3) and (3.8) will be tacitly assumed in the remainder.

Before dealing with the advection-diffusion equation, let us first consider some simple examples for separate advection or diffusion.

Example 3.1 Consider the diffusion model equation $u_t = du_{xx}$. In this case $C_1 = 0$ implies that $\alpha_{-1} = \alpha_1$. In view of space symmetry, put $\beta_{-1} = \beta_1$. This now implies that $C_k = 0$ for k odd. We have $C_2 = 0$ iff

$$\alpha_{-1} = \alpha_1 = d, \quad \alpha_0 = -2d,$$

giving for instance the standard explicit central scheme with $\beta_{-1} = \beta_1 = 0$. We have order four, with $C_4 = 0$, if in addition $\beta_{-1} = \beta_1 = \frac{1}{12}$, $\beta_0 = \frac{10}{12}$. Written out this fourth-order discretization reads

$$\frac{1}{12}(w'_{j-1} + 10w'_j + w'_{j+1}) = \frac{d}{h^2}(w_{j-1} - 2w_j + w_{j+1}). \quad (3.9)$$

Several of the fully discrete schemes in the list of Richtmyer & Morton (1967, Sect. 8.2) are based on this spatial discretization.

Next, consider the advection model equation $u_t + au_x = 0$. Then $C_2 = 0$ iff

$$\alpha_{-1} = \frac{1}{2}ah + ah(\beta_{-1} - \beta_1), \quad \alpha_1 = -\frac{1}{2}ah + ah(\beta_{-1} - \beta_1).$$

With $\beta_{-1} = \beta_1 = 0$ this gives the standard second-order explicit central scheme. We obtain order four, with $C_3 = C_4 = 0$, $C_5 = \mathcal{O}(h)$, if $\alpha_{-1} = \frac{1}{2}ah$, $\alpha_1 = -\frac{1}{2}ah$ and $\beta_{-1} = \beta_1 = \frac{1}{6}$, thus giving

$$\frac{1}{6}(w'_{j-1} + 4w'_j + w'_{j+1}) = \frac{a}{2h}(w_{j-1} - w_{j+1}). \quad (3.10)$$

Assuming $a > 0$, we can consider pure upwind schemes with $\beta_1 = \alpha_1 = 0$. Here we can achieve order two, with $C_2 = 0$, $C_3 = \mathcal{O}(h)$, by setting $\beta_{-1} = \frac{1}{2}$ and $\alpha_{-1} = ah$, resulting in the scheme

$$\frac{1}{2}(w'_{j-1} + w'_j) = \frac{a}{h}(w_{j-1} - w_j). \quad (3.11)$$

With time discretization by the implicit trapezoidal rule, this is often called the *box scheme*. The formula for $a < 0$ is obtained by the usual reflection around the central grid point x_j . \diamond

3.2 Examples

We now return to the advection-diffusion equation $u_t + au_x = du_{xx}$. The source term $s(x, t)$ in (3.1) will be omitted for convenience of presentation; the coefficients for this source term would be the same as for the temporal derivatives. The familiar second-order explicit central formula is

$$w'_j = \frac{a}{2h}(w_{j-1} - w_{j+1}) + \frac{d}{h^2}(w_{j-1} - 2w_j + w_{j+1}), \quad (3.12)$$

with error coefficients $C_2 = 0$, $C_3 = \frac{1}{6}ah$, $C_4 = -\frac{1}{12}d$. Further we will also often consider the following implicit central formula

$$\frac{1}{6}(w'_{j-1} + 4w'_j + w'_{j+1}) = \frac{a}{2h}(w_{j-1} - w_{j+1}) + \frac{d}{h^2}(w_{j-1} - 2w_j + w_{j+1}). \quad (3.13)$$

The error coefficients are $C_2 = C_3 = 0$, $C_4 = \frac{1}{12}d$ and $C_5 = -\frac{1}{180}ah$. This discretization also has order two, except for the pure advection equation, see (3.10), where it has order four. The reason for considering this formula closely is that it is also obtained by the most simple finite element method (Galerkin with piecewise linear basis functions).

Apart from the central discretizations (3.12), (3.13), the class (3.2) with $r = 1$ contains many more schemes that have appeared in the numerical literature. In the following examples the coefficients of the discretizations are expressed in terms of the cell Péclet number $\mu = ah/d$.

Example 3.2: Compact schemes. For (3.2) with $r = 1$ it is possible to achieve order four. Schemes with such an optimal order are called *compact* or OCI (Operator Compact Implicit) schemes. Here the term compact refers to the fact that with the usual explicit discretizations we need wider stencils to achieve this order; if we require order four then the 3-point stencil is the smallest one possible.

It will be required that $C_2 = 0$. The order conditions then imply

$$\alpha_{-1} = d + \frac{1}{2}ah - ah(\beta_1 - \beta_{-1}), \quad \alpha_1 = d - \frac{1}{2}ah - ah(\beta_1 - \beta_{-1}),$$

with $\alpha_0 = -(\alpha_{-1} + \alpha_1)$.

It is possible to chose the coefficients β_k such that $C_3 = C_4 = 0$. This gives the following fourth-order OCI scheme

$$\beta_{-1} = \frac{1}{\gamma}(6 + 3\mu - \mu^2), \quad \beta_0 = \frac{1}{\gamma}(60 - 4\mu^2), \quad \beta_1 = \frac{1}{\gamma}(6 - 3\mu - \mu^2),$$

with scaling factor $\gamma = 72 - 6\mu^2$ and $C_5 = \mathcal{O}(h)$. This formula is only suited for $\mu^2 < 12$, because otherwise the scaling factor becomes zero or negative.

To make the discretization suitable for all μ , Berger et al. (1980) considered the more relaxed condition $C_k = \mathcal{O}(h^{6-k})$, $k \leq 5$. For $a > 0$ we may choose

$$\beta_{-1} = \frac{1}{\gamma}(6 + 6\mu + 3\mu^2 + \frac{3}{2}\mu^3), \quad \beta_0 = \frac{1}{\gamma}(60 + 30\mu + 9\mu^2 + \frac{3}{2}\mu^3), \quad \beta_1 = \frac{1}{\gamma}(6 - 6\mu - 3\mu^2 - \frac{3}{2}\mu^3),$$

with scaling factor $\gamma = 72 + 36\mu + 12\mu^2 + 3\mu^3$; see also Morton (1996, Sect. 3.2.3). We will refer to it as the generalized OCI scheme. \diamond

Example 3.3 : Adaptive upwinding. Let $a > 0$. The adaptive upwind scheme is then constructed as follows: we select the parameters α_k such that we get a mixture of the standard upwind ($\kappa = 1$) and central ($\kappa = 0$) advection discretization,

$$\sum_{k=-1}^1 \alpha_k w_{j+k} = \frac{1}{2}ah(w_{j-1} - w_{j+1}) + (d + \frac{1}{2}ah\kappa)(w_{j-1} - 2w_j + w_{j+1})$$

with parameter κ as small as possible but such that the off-diagonal elements $\alpha_{\pm 1}$ remain non-negative,

$$\kappa = \max(0, 1 - 2/\mu).$$

For the explicit scheme we simply set $\beta_{-1} = \beta_1 = 0$. This gives formally second order (for $h \rightarrow 0$), but for large μ the formula essentially reduces to first-order upwind advection and hence the numerical dissipation will then also be quite large. The leading error coefficient is given by $C_2 = -\frac{1}{2}\kappa ah$.

An implicit discretization that is more accurate for large cell Péclet numbers μ is obtained by requiring that $C_2 = 0$. For stationary problems with source terms, a simple and quite effective scheme – quoting Morton (1996, p. 81) – is obtained with the choice

$$\beta_{-1} = \frac{1}{2}\kappa, \quad \beta_0 = 1 - \frac{1}{2}\kappa, \quad \beta_1 = 0,$$

and we will consider this choice also for the time-dependent case. For this implicit formula the principal error coefficient is given by $C_3 = \frac{1}{6}ah$ for $\mu \leq 2$, and $C_3 = -\frac{1}{12}ah(1 - 12\mu^{-2})$ for $\mu > 2$. In case $a < 0$ we get similar formulas by reflection around the central grid point x_j . \diamond

Example 3.4 : Exponential fitting. The explicit exponential fitting scheme is defined by the coefficients

$$\alpha_{-1} = ah \frac{e^\mu}{e^\mu - 1}, \quad \alpha_1 = ah \frac{1}{e^\mu - 1}$$

and $\alpha_0 = -(\alpha_1 + \alpha_{-1})$. This corresponds to an upwind factor

$$\kappa = \frac{e^\mu + 1}{e^\mu - 1} - \frac{2}{\mu}$$

in the adaptive upwind formulation of the previous example. Here the formula can be used for $a < 0$ as well; in this respect, note that $\kappa \rightarrow \pm 1$ as $\mu \rightarrow \pm\infty$ and $\kappa = \frac{1}{6}\mu + \mathcal{O}(\mu^3)$ as $\mu \rightarrow 0$. The scheme is of order two with principal error coefficient $C_2 = -\frac{1}{2}\kappa ah$. Discretizations of this type were originally

derived by Allen & Southwell (1955). This discretization is often used in semi-conductor device simulations, where it is usually called the Scharfetter-Gummel discretization. It is also known as the Il'in scheme, after Il'in (1969) who derived convergence results.

The scheme can be derived in the following natural way: we write the equation (3.1) with $s = 0$ as $u_t = -f(u)_x$ with advection-diffusion flux $f(u) = au - du_x$, and discretize this as $w'_j = h^{-1}(f_{j-\frac{1}{2}} - f_{j+\frac{1}{2}})$. Now consider fixed t and assume that the advection-diffusion flux is locally constant in space,

$$f = f_{j+\frac{1}{2}} \quad \text{constant on } [x_j, x_{j+1}].$$

Then, under this assumption, we can derive an expression for the flux by solving the equation $av - dv_x = f_{j+1/2}$ for $x \in [x_j, x_{j+1}]$ with $v(x_j) = w_j$, $v(x_{j+1}) = w_{j+1}$, to obtain

$$f_{j+\frac{1}{2}} = a \frac{e^\mu w_j - w_{j+1}}{e^\mu - 1}. \quad (3.14)$$

This derivation shows that the exponential fitting scheme is exact for stationary solutions of the advection-diffusion equation $u_t = -f(u)_x$ since the derivation is solely based on the assumption that the flux is constant over $[x_j, x_{j+1}]$.

In the implicit formula of El-Mistikawy & Werle (1978) we have the same α_k but now

$$\beta_{-1} = \frac{1}{2} \left(\frac{e^\mu}{e^\mu - 1} - \frac{1}{\mu} \right), \quad \beta_0 = \frac{1}{2}, \quad \beta_1 = \frac{1}{2} \left(\frac{1}{\mu} - \frac{1}{e^\mu - 1} \right).$$

With this choice $C_2 = 0$, $C_3 = -\frac{1}{12}(ah + 6d\kappa)$ and all β_k are non-negative.

For large cell Péclet numbers these exponentially fitted schemes become very close to the adaptive upwind schemes of the previous example. ◇

A Numerical Comparison

Consider the advection-diffusion equation $u_t + au_x = du_{xx}$ with $a = 1$, $d = 10^{-3}$, $x \in (0, 1)$ and a periodic boundary condition. The initial value is taken as $u(x, 0) = (\sin(\pi x))^{100}$ and $h = 1/m$, $m = 50$. In Figure 3.1 the numerical results are presented for the explicit and implicit central discretizations (3.12) and (3.13), indicated by solid lines. Also given are the results for the implicit and explicit adaptive upwind schemes of Example 3.3, indicated by dashed lines. The dotted lines represent the exact solution (reference solution calculated with high accuracy); these lines almost coincide visually with the implicit central numerical solutions.

The implicit central scheme gives very accurate solutions in this test. The accuracy of the implicit adaptive upwind scheme is also much better than for its explicit counterpart, but it is obvious that monotonicity of the numerical solutions is lost. Also the implicit central scheme is not free of oscillations.

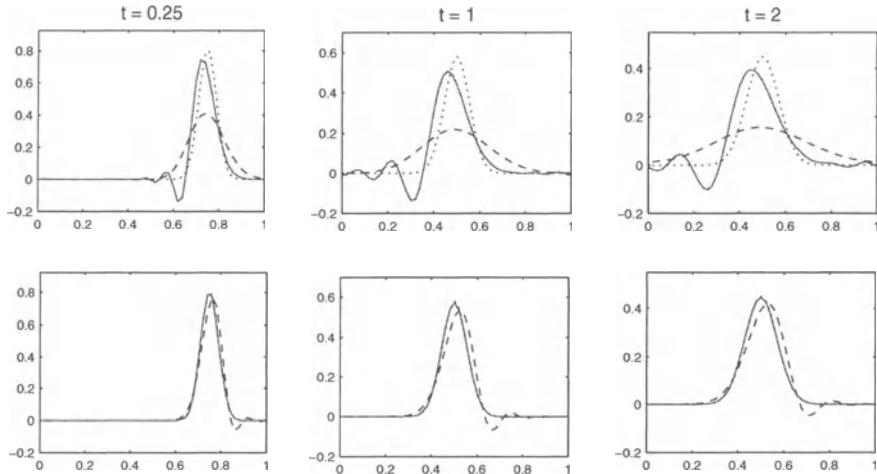


Fig. 3.1. Results for the advection-diffusion equation with spatial periodicity and $a = 1$, $d = 10^{-3}$, $h = 1/50$ at time $t = 0.25, 1, 2$. Top/bottom pictures for explicit/implicit discretizations, respectively. Solid lines for the central schemes, dashed lines for adaptive upwind and dotted lines for the exact solution.

This becomes more prominent with smaller values of d , but in comparison to the explicit central scheme the oscillations are very modest.

In this test the generalized OCI scheme of Example 3.2 produces results that are nearly identical to those of the implicit adaptive upwind discretization, due to the fact that we have a relatively large cell Péclet number $\mu = ah/d = 20$. The explicit exponential fitting scheme of Example 3.4 gives results that are almost the same as for the explicit adaptive upwind discretization. Likewise, the results for the implicit exponential fitting scheme are quite close to those of the implicit adaptive upwind scheme, except that the oscillations are slightly larger. It should be noted that here the matrix B becomes singular with the implicit exponential fitting scheme if m is even, but the scheme is still well defined. We will discuss this in more detail in the next section.

In Figure 3.2 the numerical results are presented for the same problem and initial value but now with Dirichlet conditions $u(0, t) = 1$, $u(1, t) = 0$. Then a front enters from the left, and for $t > 1$ a boundary layer will be formed at $x = 1$. The numerical results given here are for the central and adaptive upwind formulas similar as above. The boundary layer that arises for larger t is well resolved by the adaptive upwind schemes. The central schemes do not approximate this boundary layer correctly; both the implicit and explicit scheme give strong oscillations. For steady state solutions the implicit and explicit schemes are identical, which clearly shows up in the plots for $t = 1.5$. For the smaller output times the implicit central scheme again gives very good results.

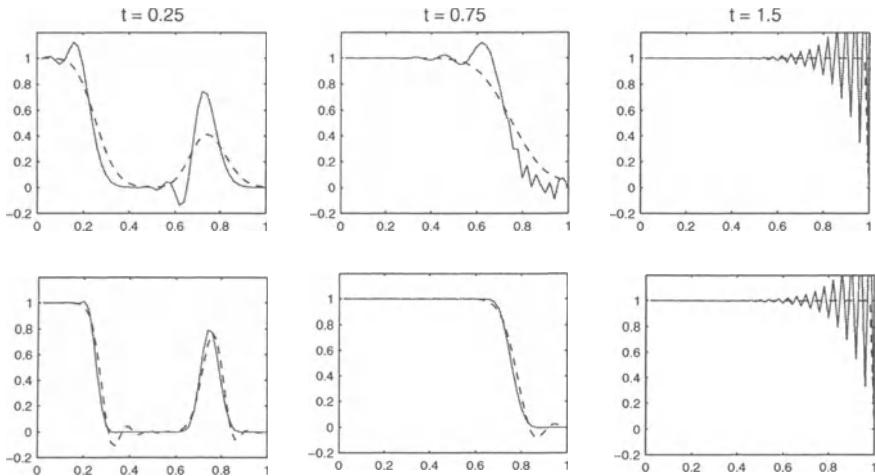


Fig. 3.2. Results for the advection-diffusion equation with Dirichlet conditions and $a = 1$, $d = 10^{-3}$, $h = 1/50$ at time $t = 0.25, 0.75, 1.5$. Top/bottom pictures for explicit/implicit discretizations, respectively. Solid lines for the central schemes, dashed lines for adaptive upwind.

With respect to the other methods considered in this section, it was observed once more that the generalized OCI scheme produces in this test results close to the implicit adaptive upwind scheme. The exponential fitting formulas also give again results close to the adaptive upwind schemes, explicit as well as implicit.

Inventory

Although definite conclusions cannot be drawn from these simple tests, it is clear that the implicit discretizations give quite good results compared to their explicit counterparts in terms of accuracy, but the implicit upwind schemes do produce rather large oscillations with moving fronts. The good behaviour of the implicit central scheme (3.13) is noteworthy: its solutions are not monotone but the oscillations are relatively small, except in the situation where a stationary boundary layer arises and where the implicit scheme reduces to its explicit counterpart (3.12). For the periodic problem also explicit schemes with third- or fourth-order advection discretizations were tested, but the results were more oscillatory than with the implicit central scheme (3.13).

An interesting property of the exponentially fitted schemes for stationary problems is the existence of error bounds in the maximum norm that hold uniformly in the cell Péclet number μ : for $au_x = du_{xx} + g$ the explicit scheme has a maximum error bounded by Ch and the maximum error of the implicit scheme is bounded by Ch^2 , without requirements on the smoothness of u , see Morton (1996, Sect. 3.4). However, in view of the above experiment,

this seems not so relevant for the time-dependent case if resolution of the boundary layer is not the principal goal. The choice of the coefficients β_k with the implicit exponentially fitted and adaptive upwind schemes is a common one for stationary problems with source terms. It is clear from the above test that in the time-dependent case this choice leads to oscillations.

The generalized OCI scheme is more complicated than the other schemes and in the above test it did give virtually the same results as the implicit adaptive upwind and exponential fitting schemes. For this reason the generalized OCI scheme will not be taken along in the subsequent discussions.

Remark 3.5 We will not look in a systematic way at implicit discretizations that use more than three points, but as an interesting example of a 4-point discretization for $u_t + au_x = du_{xx} + s(x, t)$ we mention the scheme

$$\frac{1}{2}(w'_{j-1} + w'_j) = \frac{a}{h}(w_{j-1} - w_j) + \frac{d}{2h^2}(w_{j-2} - w_{j-1} - w_j + w_{j+1}) + \frac{1}{2}(g_{j-1} + g_j).$$

Note that it is the diffusion term that leads here to the 4-point stencil. This scheme arises in a natural way if we consider a finite volume approach with cells $\Omega_j = (x_{j-1}, x_j)$ as control volumes and with $w_j(t) \approx u(x_j, t)$ located at the cell vertices (a so-called *cell-vertex scheme*). Setting up a mass balance on Ω_j , with average mass given by $\frac{1}{2}(w_{j-1} + w_j)$, leads to

$$\frac{1}{2}(w'_{j-1} + w'_j) = \frac{1}{h}(f_{j-1} - f_j) + \frac{1}{2}(g_{j-1} + g_j)$$

with $g_j = s(x_j, t)$ and with central advection-diffusion fluxes

$$f_j = aw_j + \frac{d}{2h}(w_{j-1} - w_{j+1}).$$

In the above test with periodicity this scheme did give results comparable to those of the implicit exponential fitting formula, and again the matrix B becomes singular if the number of grid points is even. The implementation of boundary conditions is not straightforward here. For example, consider Dirichlet conditions at $x = 0, 1$ and grid points $x_j = jh$, $j = 0, 1, \dots, m+1$ with $h = 1/(m+1)$. Then we have m unknowns w_1, \dots, w_m but there are $m+1$ control volumes $\Omega_1, \dots, \Omega_{m+1}$, leading to $m+1$ equations. For details on the proper implementation of this scheme and related ones, see Morton (1996) and Roos, Stynes & Tobiska (1996). \diamond

3.3 Stability and Convergence

Consider the implicit discretization (3.4) with normalization (3.3). Premultiplying the scheme by B^{-1} , it seems that the stability requirement should be

$$\|\exp(tB^{-1}A)\| \leq C \quad \text{for all } t > 0.$$

Indeed, if this holds and B^{-1} is uniformly bounded, convergence easily follows. However, the situation may be more complicated. The inverse of B may not exist, and in fact it does not for some of the schemes with periodic problems as was mentioned in the previous section. Moreover, even if the inverse of B exists, it may be close to singular.

To discuss these aspects in some detail, we look at the test problem (3.1) with a uniform grid on $[0, 1]$ and spatial periodicity, for which Fourier decompositions can be used. Let $x_j = jh$, $j = 1, \dots, m$, $h = 1/m$, and consider

$$A = V \text{diag}(a_k) V^{-1}, \quad B = V \text{diag}(b_k) V^{-1}$$

with V the matrix of discrete Fourier modes (I.3.10) and with a_k, b_k the eigenvalues. Consider the global discretization error $\varepsilon(t) = u_h(t) - w(t)$ and the transformed error $\hat{\varepsilon}(t) = V^{-1}\varepsilon(t)$ together with the transformed truncation error $\hat{\sigma}_h(t) = V^{-1}\sigma_h(t)$. Then we have

$$b_k \frac{d}{dt} \hat{\varepsilon}_k(t) = a_k \hat{\varepsilon}_k(t) + \hat{\sigma}_{h,k}(t), \quad k = 1, \dots, m.$$

We assume that $|a_k| + |b_k| > 0$.³⁾ Let in the following $\lambda_k = a_k/b_k$. Then

$$\begin{aligned} \hat{\varepsilon}_k(t) &= e^{t\lambda_k} \hat{\varepsilon}_k(0) + \int_0^t e^{(t-s)\lambda_k} \frac{1}{b_k} \hat{\sigma}_{h,k}(s) ds \quad \text{if } b_k \neq 0, \\ \hat{\varepsilon}_k(t) &= -\frac{1}{a_k} \hat{\sigma}_{h,k}(t) \quad \text{if } b_k = 0, a_k \neq 0. \end{aligned} \tag{3.15}$$

To prove convergence we will use the stability condition

$$\operatorname{Re} \lambda_k \leq 0 \quad \text{and} \quad |a_k| + |b_k| \geq \delta, \quad k = 1, \dots, m,$$

for some $\delta > 0$ independent of h . If we denote $\phi(z) = h^{-2} \sum_j \alpha_j z^j$ and $\psi(z) = \sum_j \beta_j z^j$, then

$$\lambda_k = \phi(e^{2\pi i kh}) / \psi(e^{2\pi i kh}). \tag{3.16}$$

Instead of regarding only the discrete Fourier modes, it is here more convenient to deal with the continuous case; see also Remark I.6.10. Then the stability condition becomes

$$\operatorname{Re}(\phi(z)/\psi(z)) \leq 0 \quad \text{and} \quad |\phi(z)| + |\psi(z)| \geq \delta \quad \text{for all } |z| = 1. \tag{3.17}$$

This condition is easy to verify in terms of the coefficients.

³⁾ The assumption $|a_k| + |b_k| > 0$ for all k is directly related to regularity of the matrix pencil $A - \lambda B$, that is $\det(A - \lambda B) \neq 0$, and this regularity is a general requirement for well-posedness of implicit differential equations $Bw'(t) = Aw(t)$, see for instance Brenan, Campbell & Petzold (1989).

Lemma 3.6 Consider the discretization (3.2) with $r = 1$ and error coefficients $C_0 = C_1 = 0$, $C_2 = \mathcal{O}(h)$, and assume that

$$h^{-2}|\alpha_0| + |\beta_0 - \frac{1}{2}| \geq \gamma$$

for some $\gamma > 0$ independent of h . Then the stability condition (3.17) holds iff

$$2ah(\beta_1 - \beta_{-1}) \geq \alpha_0 \quad \text{and} \quad \alpha_0(1 - 2\beta_0) \geq 0.$$

Proof. Consider $z = 1 - x + iy$ with $|z| = 1$. Then $z^{-1} = \bar{z}$ and using the conditions (3.3), (3.8) it follows that

$$h^2\phi(z) = \alpha_0x - iahy, \quad \psi(z) = 1 - (1 - \beta_0)x + i(\beta_1 - \beta_{-1})y.$$

From the assumptions on α_0 and β_0 it can be seen by some calculations ⁴⁾ that $|\phi(z)| + |\psi(z)|$ is bounded away from 0 for all $|z| = 1$.

Further we have $\operatorname{Re}(\phi(z)/\psi(z)) \leq 0$ iff

$$h^2\operatorname{Re}(\overline{\psi(z)}\phi(z)) = \alpha_0x - \alpha_0(1 - \beta_0)x^2 - ah(\beta_1 - \beta_{-1})y^2 \leq 0.$$

By substituting $y^2 = 1 - (1 - x)^2 = 2x - x^2$ we obtain the requirement

$$(\alpha_0 - 2ah(\beta_1 - \beta_{-1}))x - (\alpha_0(1 - \beta_0) - ah(\beta_1 - \beta_{-1}))x^2 \leq 0$$

which should hold for all $0 \leq x \leq 2$. The result now directly follows. \square

For explicit schemes the above stability condition simply reads $\alpha_0 \leq 0$. Also for the implicit adaptive upwind and exponential fitting schemes considered here it is easily verified that the schemes are stable.⁵⁾ Note that the stability here, through Fourier decompositions, is stability in the L_2 -norm, just as the von Neumann stability considered in Chapter I.

Having this stability, convergence in the L_2 -norm then easily follows. Assuming that $\varepsilon(0) = 0$ and that the truncation error satisfies

$$\|\sigma_h(t)\|_2 \leq Kh^p \quad \text{uniformly for } 0 \leq t \leq T,$$

the same holds for the Fourier transformed vector $\hat{\sigma}_h(t)$. This is sufficient for convergence in the above framework. Note that we can write $e^{\lambda_k(t-s)}/b_k = (\lambda_k/a_k)e^{\lambda_k(t-s)}$ in case b_k approaches 0. From (3.15) it then follows that

$$|\hat{\varepsilon}_k(t)| \leq C \min(|a_k|^{-1}, |b_k|^{-1}) |\hat{\sigma}_{h,k}(t)|,$$

uniformly on a finite time interval $[0, T]$, and hence convergence with order p is obtained.

⁴⁾ Consider the cases $a = 0$ and $a \neq 0$ separately. Note that if $a = 0$, then we have $\alpha_0 = -2d + \mathcal{O}(h)$.

⁵⁾ Note that $C_2 = 0$ implies $\alpha_0 = 2ah(\beta_1 - \beta_{-1}) - 2d$. Moreover, for the schemes considered here we have $\alpha_0 \leq 0$, and then the remaining stability condition simply reads $\beta_0 \geq \frac{1}{2}$.

3.4 Monotonicity

The implicit adaptive upwind and exponential fitting schemes of Example 3.3 and 3.4 provide good numerical approximations for boundary layers, but we saw that for time-dependent problems positivity is lost and the solutions become oscillatory. This seems a fundamental problem, which already occurs with the underlying advection upwind formula (3.11).

Consider the general scheme $Bw'(t) = Aw(t) + Bg(t)$ with non-negative source term $g(t) \geq 0$. Assume for convenience that A and B are invertible. Then the positivity requirement for the time-dependent case reads

$$\exp(tB^{-1}A) \geq 0 \quad \text{for all } t > 0. \quad (3.18)$$

On the other hand, for stationary solutions we get $w = -A^{-1}Bg$, and thus the positivity requirement then reads

$$-A^{-1}B \geq 0. \quad (3.19)$$

With explicit upwind schemes these two requirements naturally combine: if $\exp(tA) \geq 0$ for all $t > 0$, which is ensured by non-negative off-diagonal entries, then we also have $(I - \tau A)^{-1} \geq 0$ for all $\tau > 0$ (see Section I.7), and by considering $\tau \rightarrow \infty$ it follows that $-A^{-1} \geq 0$. However, implicitness of a scheme seems to exclude the temporal positivity requirement (3.18).

Example 3.7 Consider $\alpha_{-1} = ah$, $\alpha_0 = -ah$, $\alpha_1 = 0$ with $ah > 0$. This is a limit case, for $d \rightarrow 0$, with both adaptive upwinding and exponential fitting, see also (3.11). Further assume $B \geq 0$, which is a natural assumption for the stationary case to ensure that $Bg \geq 0$ whenever $g \geq 0$. Let us suppose for convenience that (3.1) is a pure initial value problem with spatial domain the real line \mathbb{R} . Then the semi-discrete system is defined in \mathbb{R}^∞ . Denoting by E the forward shift operator on \mathbb{R}^∞ , we have $A = ah^{-1}(E^{-1} - I)$ and $B = \beta_{-1}E^{-1} + \beta_0I + \beta_1E$. Put

$$B^{-1} = \sum_{k \in \mathbb{Z}} \gamma_k E^k.$$

Then we have

$$B^{-1}A = ah^{-1} \sum_{k \in \mathbb{Z}} \gamma_k (E^{k-1} - E^k) = ah^{-1} \sum_{k \in \mathbb{Z}} (\gamma_{k+1} - \gamma_k) E^k.$$

Thus we see that all off-diagonal elements of $B^{-1}A$ are non-negative iff

$$\gamma_k \leq \gamma_{k+1} \quad \text{for all } k \neq 0, \quad (3.20)$$

and by Theorem I.7.2 we know that this is equivalent to (3.18). On the other hand,

$$B^{-1}B = \sum_{k \in \mathbb{Z}} (\gamma_{k+1}\beta_{-1} + \gamma_k\beta_0 + \gamma_{k-1}\beta_1) E^k,$$

and thus

$$\gamma_{k+1}\beta_{-1} + \gamma_k\beta_0 + \gamma_{k-1}\beta_1 = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.21)$$

However, with $\beta_k \geq 0$, the conditions (3.20) and (3.21) are contradictory, except for the explicit case $B = I$. \diamond

Remark 3.8 The coefficients β_k in the Examples 3.3 and 3.4 were chosen non-negative to ensure the property (3.19). Apart from positivity this also implies certain maximum principles or comparison principles for stationary problems, see for instance Morton (1996, Lemma 3.2.3). However, as we saw in the above test for the model problem $u_t + au_x = du_{xx}$, the results for time-dependent problems may be quite poor.

Better temporal accuracy with the implicit upwind type schemes is obtained by requiring that $C_2 = C_3 = 0$. Writing

$$\alpha_{-1} = \frac{1}{2}ah(\kappa + 1) + d, \quad \alpha_0 = -ah\kappa - 2d, \quad \alpha_1 = \frac{1}{2}ah(\kappa - 1) + d$$

with upwind factor κ , we have $C_2 = 0$ iff

$$\beta_{-1} = \frac{1}{2}(1 - \beta_0) + \frac{1}{4}\kappa, \quad \beta_1 = \frac{1}{2}(1 - \beta_0) - \frac{1}{4}\kappa.$$

The condition $C_3 = 0$ is then achieved by taking

$$\beta_0 = \frac{2}{3} + \mu^{-1}\kappa.$$

With κ taken as in the adaptive upwind and exponentially fitted schemes, this choice for the β_j was observed to improve the results in the above tests, also with larger cell Péclet numbers, even though (3.18) does not hold. Almost the same improvements were also obtained with the choice $C_2 = 0$, $\beta_0 = \frac{2}{3}$. However, with both these choices the condition $B \geq 0$ is lost and this destroys the stationary positivity property (3.19) with source terms. \diamond

Remark 3.9 An invertible matrix $C = (c_{ij}) \in \mathbb{R}^{m \times m}$ is called an *M-matrix* if $C^{-1} \geq 0$ and $c_{ij} \leq 0$ for all $i \neq j$. A sufficient condition is

$$c_{ij} \leq 0 \quad (i \neq j), \quad c_{ii} > \sum_{j \neq i} |c_{ij}|,$$

see for instance Ortega & Rheinboldt (1970, Sect. 2.4). In the construction of discretization schemes it is common to require that $B \geq 0$ and that $-A$ is an *M-matrix*; the latter condition being ensured by diagonal dominance. Obviously this implies (3.19), that is monotonicity for the stationary case. However, to guarantee in addition that (3.18) holds for the time-dependent case, the matrix $-B^{-1}A$ should be an *M-matrix*. This however is not so easy to verify. \diamond

3.5 Time Integration Aspects

Implicit spatial discretization leads to a semi-discrete system of the form

$$Bw'(t) = F(t, w(t)), \quad (3.22)$$

where $F(t, w) = Aw + Bg(t)$ for the linear problem (3.1). Application of the θ -method for time integration gives

$$Bw_{n+1} = Bw_n + (1 - \theta)\tau F(t_n, w_n) + \theta\tau F(t_{n+1}, w_{n+1}).$$

More general methods, such as Runge-Kutta or linear multistep methods, are modified in a similar way for (3.22). With the θ -method an algebraic system of the type

$$Bw_{n+1} - \theta\tau F(t_{n+1}, w_{n+1}) = v_n$$

with given v_n is to be solved in each time step. Even with the forward Euler method ($\theta = 0$) there is ‘some amount’ of implicitness, in the sense that a linear system with matrix B has to be solved in each step. Note that if the matrix B in (3.22) is constant – as it will be for spatial discretizations by a Galerkin finite element method, even for nonlinear problems – then we can make in the first step an LU or Cholesky factorization which can be reused in all subsequent steps (with iterative linear solvers these factorizations would be taken incomplete and used for preconditioning). This makes the use of explicit ODE methods feasible.

Still it is clear that the use of explicit ODE methods is less attractive than for the case $B = I$. Moreover, the stability requirements become in general more stringent for implicit spatial discretizations than for $B = I$. An illustration is given in Figure 3.3 where the eigenvalues $\lambda_k = a_k/b_k \in \sigma(B^{-1}A)$ (as in Section 3.3) are pictured for the advection diffusion problem with spatial periodicity. The eigenvalues are given for the central and adaptive upwind schemes. With the latter, two choices for the β_k are considered: the choice of Example 3.3 with $\beta_j \geq 0$, indicated by dark grey o-marks, and the

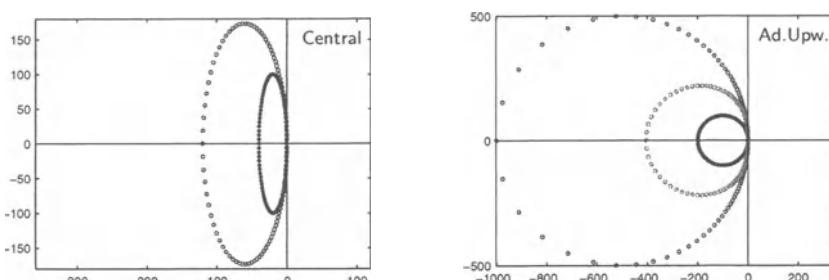


Fig. 3.3. Spectra for the advection-diffusion equation, $a = 1$, $d = 10^{-3}$, $h = 1/100$, with the central and adaptive upwind schemes. Black dots for explicit spatial schemes, grey circles for implicit spatial schemes (see text for details).

choice of Remark 3.8 with $C_2 = C_3 = 0$, indicated by lighter o-marks. In particular for the first choice the eigenvalues become very large in modulus, which makes the use of explicit ODE methods impractical. The precise form of the eigenvalues that are illustrated here can be easily obtained from the formula (3.16) for these central and adaptive upwind schemes.

Finally we note that when using an implicit A -stable ODE method for time stepping, there will be little difference between implementing the explicit or implicit spatial discretizations. A Newton iteration process will involve matrices $B - \theta\tau A$ with $\theta > 0$, $A \approx F_w(t, w)$, and for the schemes considered here these matrices will be in general non-singular for any $\tau > 0$.

4 Non-uniform Grids – Finite Volumes (1D)

So far we have discussed spatial advection-diffusion discretizations on uniform grids only. For resolving solutions that vary strongly in local regions, non-uniform grids are often used. Similar to variable step sizes in time, the mesh width in space then should be adapted to capture the local variations with sufficient resolution. In regions with little variation the mesh width can then be taken relatively large. In particular for multi-dimensional problems this may lead to significantly fewer grid points than with uniform grids.

Some of the main theoretical properties of non-uniform grids already become clear with one-dimensional problems and in this section we restrict ourselves to the 1D advection-diffusion equation in conservation form

$$u_t + (a(x, t)u)_x = (d(x, t)u_x)_x + s(x, t) \quad (4.1)$$

with variable coefficients. The schemes considered here will be written in a conservation form, using local mass balances over a cell Ω_j surrounding x_j , the so-called *control volume*. Approximations $w_j(t) \approx u(x_j, t)$ are thus also associated to an average mass or concentration on Ω_j . From now on we will refer to such schemes as *finite volume schemes*. Schemes that are not based on local mass balances, but where approximations $w_j(t)$ are purely seen as point values will still be called finite difference schemes.

Non-uniform grids can be constructed in two ways. One can define the grid by the points x_j and then set up a mass balance on $\Omega_j = [x_{j-1/2}, x_{j+1/2}]$ where $x_{j\pm 1/2} = \frac{1}{2}(x_j + x_{j\pm 1})$. This leads to so-called *vertex centered schemes* where the vertices of the cells are centered with respect to the sequence $\{x_j\}$.⁶⁾ Alternatively, we can partition the domain into cells Ω_j and then take x_j to be the center of Ω_j . This leads to so-called *cell centered schemes*.

⁶⁾ Actually, terms like ‘cell centered’ and ‘vertex centered’ originate from multi-dimensional problems (see Section 6.5 below). For 1D problems the cell boundaries $x_{j\pm 1/2}$ are also called cell vertices.

In the following, h_j will always denote the width of the cell around x_j . Then in both cases the semi-discrete equation has the conservation form

$$w'_j(t) = \frac{1}{h_j} \left(f_{j-\frac{1}{2}}(t, w(t)) - f_{j+\frac{1}{2}}(t, w(t)) \right) + g_j(t) \quad (4.2)$$

with the fluxes $f_{j+1/2}$ defined at $x_{j+1/2}$. The source term can be taken pointwise, $g_j(t) = s(x_j, t)$, or averaged over Ω_j . The difference between cell centered and vertex centered lies in the location of the cell vertices $x_{j+1/2}$ with respect to the grid points x_j . Neglecting boundary effects and taking $s(x, t) = 0$, this general semi-discrete conservation form implies the global conservation law

$$\sum_j h_j w_j(t) = \text{constant}. \quad (4.3)$$

4.1 Vertex Centered Schemes

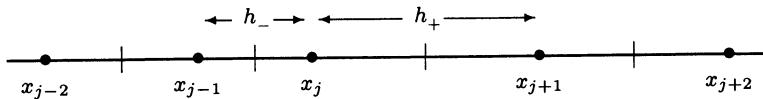
Consider a non-uniform grid defined by unequally spaced grid points $\{x_j\}$. For the moment boundaries of the domain are not taken along in our considerations. For a finite volume discretization we need to prescribe the location of cell boundaries, or vertices, for flux computations. Here we consider the vertex centered grid where the cell vertices $x_{j+1/2}$ are centered between the grid points x_j and x_{j+1} , thus giving

$$x_{j+\frac{1}{2}} = \frac{1}{2}(x_j + x_{j+1}), \quad h_j = \frac{1}{2}(x_{j+1} - x_{j-1}). \quad (4.4)$$

Let for convenience of notation

$$h_- = x_j - x_{j-1}, \quad h_+ = x_{j+1} - x_j$$

when a specified grid point x_j is considered.



For the variable coefficient advection-diffusion problem (4.1) the most simple central flux form on the vertex centered non-uniform grid is

$$f_{j+\frac{1}{2}}(t, w) = a_{j+\frac{1}{2}} \frac{w_j + w_{j+1}}{2} + d_{j+\frac{1}{2}} \frac{w_j - w_{j+1}}{h_+}, \quad (4.5)$$

where $a_{j+1/2}$, $d_{j+1/2}$ are exact or approximate values at time t at the cell boundary $x_{j+1/2}$, for example linear averages. For uniform grids this flux gives the second-order central advection-diffusion discretization discussed in

Section I.4. With the flux (4.5) the central non-uniform vertex centered discretization becomes

$$\begin{aligned} w'_j &= \frac{1}{2h_j} \left(a_{j-\frac{1}{2}} w_{j-1} + (a_{j-\frac{1}{2}} - a_{j+\frac{1}{2}}) w_j - a_{j+\frac{1}{2}} w_{j+1} \right) \\ &\quad + \frac{1}{h_j} \left(\frac{d_{j-\frac{1}{2}}}{h_-} w_{j-1} - \left(\frac{d_{j-\frac{1}{2}}}{h_-} + \frac{d_{j+\frac{1}{2}}}{h_+} \right) w_j + \frac{d_{j+\frac{1}{2}}}{h_+} w_{j+1} \right) + g_j. \end{aligned} \quad (4.6)$$

Upwinding can be introduced in this 3-point scheme by an artificial increase of the diffusion coefficients to

$$\tilde{d}_{j\pm\frac{1}{2}} = d_{j\pm\frac{1}{2}} + \frac{1}{2} \kappa_{\pm} h_{\pm} a_{j\pm\frac{1}{2}},$$

where the first-order upwind scheme (I.4.15) corresponds to the choice $\kappa_{\pm} = \text{sign}(a_{j\pm\frac{1}{2}})$. Intermediate values $|\kappa_{\pm}| < 1$ can be chosen such that the upwinding becomes adaptive or exponentially fitted, see Section 3.⁷⁾ In this section we mainly consider the central scheme assuming that the grid has been properly placed to avoid oscillations, which could be controlled for instance through the cell Péclet numbers.

Consistency, Stability and Convergence Properties of (4.6)

The theoretical complications that are introduced by non-uniform grids are best illustrated by the central scheme (4.6) with constant coefficients. Unless indicated otherwise, it is assumed in the following that a and d are constant.

Consistency: The spatial truncation error $\sigma_h(t) = (\sigma_{h,j}(t))$ of (4.6) is obtained by inserting exact solution values into the scheme. Using Taylor expansions we then obtain

$$\begin{aligned} \sigma_{h,j}(t) &= (h_+ - h_-) \left(\frac{a}{2} u_{xx}(x_j, t) - \frac{d}{3} u_{xxx}(x_j, t) \right) \\ &\quad + (h_+^2 - h_+ h_- + h_-^2) \left(\frac{a}{6} u_{xxx}(x_j, t) - \frac{d}{12} u_{xxxx}(x_j, t) \right) + \dots \end{aligned} \quad (4.7)$$

The leading error terms are proportional to $h_+ - h_-$ both for the advection and diffusion contributions. Hence for arbitrary grid spacings we have only a first-order truncation error for $h_-, h_+ \rightarrow 0$.

If we assume the grid to be *smooth* in the sense that $h_+ - h_- = \mathcal{O}(h^2)$ where h denotes a maximal mesh width, the usual second-order behaviour in h is recovered. For example, if the grid is based on a smooth transformation

⁷⁾ In the derivation of the exponential fitting scheme in Example 3.4 no use was made of the uniformity of the grid. With variable coefficients we would obtain the flux expression (3.14) with $a = a_{j+1/2}$ and $\mu = \mu_{j+1/2} = a_{j+1/2}(x_{j+1} - x_j)/d_{j+1/2}$.

$x = x(\xi)$ with a uniform grid for the underlying variable ξ , we get such a smooth grid and then (4.7) can be further expanded to

$$\sigma_h = a\bar{h}^2 \left(\frac{1}{2}x_{\xi\xi} u_{xx} + \frac{1}{6}x_\xi^2 u_{xxx} \right) - d\bar{h}^2 \left(\frac{1}{3}x_{\xi\xi} u_{xxx} + \frac{1}{12}x_\xi^2 u_{xxxx} \right) + \mathcal{O}(\bar{h}^4)$$

with \bar{h} being the uniform mesh width used for ξ .

Stability: Let us next examine the stability properties of the central scheme (4.6), for which L_2 -stability on uniform grids was established in Section I.4. Consider the advection-diffusion equation (4.1) with a, d constant on the spatial interval $[0, 1]$ with Dirichlet boundary conditions at $x_0 = 0, x_{m+1} = 1$. For stability we can consider homogeneous boundary conditions and source term $s = 0$. Then (4.6) can be written in the following linear system form in \mathbb{R}^m ,

$$w'(t) = Aw(t), \quad A = H^{-1}(B_1 + B_2), \quad (4.8)$$

where $H = \text{diag}(h_j)$ and B_1, B_2 are the contributions of advection and diffusion. The matrices B_1 and B_2 are easily seen to be skew-symmetric and symmetric non-positive definite, respectively. The discrete L_2 -inner product and corresponding norm on the non-uniform grid are defined in a natural way by

$$\langle u, v \rangle = \sum_{j=1}^m h_j u_j v_j = u^T H v, \quad \|v\|^2 = \langle v, v \rangle.$$

With this inner product we find

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|w(t)\|^2 &= \langle w(t), w'(t) \rangle = \langle w(t), H^{-1}(B_1 + B_2)w(t) \rangle \\ &= w(t)^T (B_1 + B_2) w(t) = w(t)^T B_2 w(t) \leq 0. \end{aligned}$$

It follows that $\|w(t)\|$ is non-increasing in t , and thus we have stability in the L_2 -norm.

We note that this L_2 -stability result can be easily extended to variable coefficients provided $a(x, t)$ is smooth in x . Then the diagonal contributions of the advective terms in (4.6) are $\mathcal{O}(1)$ uniformly in h , and stability with moderate growth can be proven as indicated in Section I.4.3 (by considering the diagonal as a perturbation on the skew-symmetric case). Variable diffusion coefficients can be dealt with as in Section I.4.4, without change. With Neumann boundary conditions a similar result can be obtained by considering an inner product that is slightly modified at the boundaries, see Section I.5.2.

In the maximum norm we can establish stability if the cell Péclet numbers

$$\mu_{j+\frac{1}{2}} = a_{j+\frac{1}{2}} (x_{j+1} - x_j) / d_{j+\frac{1}{2}}$$

are at most 2 in modulus. Also this follows easily as in Section I.4.3 by considering the logarithmic norm of the matrix A . For advection dominated

problems such a restriction on the cell Péclet numbers is impractical to impose over the whole spatial domain, but by local adaptation of the mesh we can impose this in those regions where the variation of the solutions is large and where maximum-norm stability will matter most.

Convergence: Let $u_h(t)$ be the restriction of the exact PDE solution to the spatial grid. The global spatial discretization error

$$\varepsilon(t) = u_h(t) - w(t)$$

of (4.6) satisfies the equation $\varepsilon'(t) = A\varepsilon(t) + \sigma_h(t)$, with matrix A given by (4.8) and truncation error σ_h given by (4.7). For the convergence analysis we assume that $\|e^{tA}\| \leq Ke^{t\omega}$ in some suitable discrete L_p -norm with moderate constants $K > 0$ and $\omega \in \mathbb{R}$, for which some sufficient conditions have been listed above. With an estimate for $\|\sigma_h(t)\|$, a global error bound for $\|\varepsilon(t)\|$ can then be obtained by the standard argument as in Theorem I.4.1.

If the grid is smooth in the above sense, second-order convergence will hold similar as for the uniform grid. The truncation error however suggests that on arbitrary grids only a first-order convergence result will hold. This is too pessimistic, as will be shown next.⁸⁾ It is assumed that the advection-diffusion equation is taken on the spatial domain $\Omega = (0, 1)$ with Dirichlet conditions at the boundaries. Let in the following

$$\Delta_j = x_j - x_{j-1}, \quad j = 1, \dots, m+1,$$

where $x_0 = 0$ and $x_{m+1} = 1$. To derive correct global estimates, with second-order convergence, we use Theorem I.5.2. The essential point then is to find a vector $\xi \in \mathbb{R}^m$ such that $A\xi = \rho$ and $\|\xi\| = \mathcal{O}(h^2)$, where h is the maximal grid size and $\rho \in \mathbb{R}^m$ is the leading term of the truncation error $\sigma_h(t)$ for t fixed,

$$\rho_j = (\Delta_{j+1} - \Delta_j)v(x_j), \quad v(x_j) = \frac{1}{2}au_{xx}(x_j, t) - \frac{1}{3}du_{xxx}(x_j, t).$$

The higher-order terms of the truncation error are incorporated in the function η of Theorem I.5.2. These will give an $\mathcal{O}(h^2)$ contribution to the global error.

First consider the pure diffusion case, $a = 0$, and set for convenience $d = 1$.⁹⁾ Then $A\xi = \rho$ reads

$$\frac{1}{\Delta_j}(\xi_{j-1} - \xi_j) - \frac{1}{\Delta_{j+1}}(\xi_j - \xi_{j+1}) = h_j\rho_j, \quad j = 1, \dots, m, \quad (4.9)$$

⁸⁾ Results of this type for stationary advection-diffusion problems are found in Samarskij (1984) and Manteuffel & White (1986). A more transparent derivation for cell centered schemes has been given by Weiser & Wheeler (1988) for the pure diffusion case $a = 0$.

⁹⁾ Taking $d = 1$ is without loss of generality, since d multiplies A as well as ρ if $a = 0$.

with $\xi_0 = \xi_{m+1} = 0$. Define

$$p_j = \sum_{k=1}^{j-1} h_k \rho_k, \quad q_j = \sum_{k=1}^j \Delta_k p_k, \quad j = 0, 1, \dots, m+1, \quad (4.10)$$

where empty sums are taken equal to zero. Then it can be verified directly that the solution of the recursion (4.9) with $\xi_0 = \xi_{m+1} = 0$ is given by

$$\xi_j = q_j - x_j q_{m+1}, \quad j = 0, 1, \dots, m+1. \quad (4.11)$$

Now we can estimate $|p_j|$ and then use $|q_j| \leq \max_k |p_k|$. We have

$$\begin{aligned} p_j &= \sum_{k=1}^{j-1} h_k (\Delta_{k+1} - \Delta_k) v(x_k) = \sum_{k=1}^{j-1} \frac{1}{2} (\Delta_{k+1}^2 - \Delta_k^2) v(x_k) \\ &= \frac{1}{2} \Delta_j^2 v(x_{j-1}) - \frac{1}{2} \sum_{k=2}^{j-1} \Delta_k^2 (v(x_k) - v(x_{k-1})) - \frac{1}{2} \Delta_1^2 v(x_1). \end{aligned}$$

Thus we obtain $|p_j|, |q_j| \leq \frac{1}{2} Ch^2$ for all j , with a constant $C > 0$ determined by bounds on $|v|, |v_x|$, ¹⁰⁾ and thus from (4.11) it follows that $|\xi_j| \leq Ch^2$ for $j = 1, \dots, m$. Hence $\|\xi\| \leq Ch^2$ in any discrete L_p -norm.

With advection terms the proof becomes somewhat more technical. Then the relation $A\xi = \rho$ reads

$$\frac{1}{\Delta_j} (d + \frac{1}{2} a \Delta_j) (\xi_{j-1} - \xi_j) - \frac{1}{\Delta_{j+1}} (d - \frac{1}{2} a \Delta_{j+1}) (\xi_j - \xi_{j+1}) = h_j \rho_j \quad (4.12)$$

for $j = 1, \dots, m$ and $\xi_0 = \xi_{m+1} = 0$. Define

$$c_j = \prod_{k=1}^j \left(\frac{d - \frac{1}{2} a \Delta_k}{d + \frac{1}{2} a \Delta_k} \right), \quad c_{j+\frac{1}{2}} = (d - \frac{1}{2} a \Delta_{j+1}) c_j. \quad (4.13)$$

Then (4.12) can be written as

$$\frac{c_{j-\frac{1}{2}}}{\Delta_j} (\xi_{j-1} - \xi_j) - \frac{c_{j+\frac{1}{2}}}{\Delta_{j+1}} (\xi_j - \xi_{j+1}) = c_j h_j \rho_j.$$

This is of the same form as (4.9), only with $\tilde{\Delta}_j = \Delta_j / c_{j-\frac{1}{2}}$ and $\tilde{h}_j = h_j c_j$ replacing Δ_j, h_j . Defining likewise

$$\tilde{p}_j = \sum_{k=1}^{j-1} \tilde{h}_k \rho_k, \quad \tilde{q}_j = \sum_{k=1}^j \tilde{\Delta}_k \tilde{p}_k,$$

¹⁰⁾ More precisely, we can set $C = \max_{1 \leq j \leq m} (\theta_j^2 \max(|2v(x_j)|, |v_x(x_j)|))$ with $\theta_j = h^{-1} \Delta_j$.

we obtain

$$\xi_j = \tilde{q}_j - \tilde{x}_j \tilde{q}_{m+1}, \quad \tilde{x}_j = \sum_{k=1}^j \tilde{\Delta}_k / \sum_{k=1}^{m+1} \tilde{\Delta}_k.$$

For fixed $a \in \mathbb{R}$, $d > 0$ we have $c_j = e^{-ax_j/d} + \mathcal{O}(h^2)$ and we can estimate the terms $|\xi_j|$ in a similar way as for the pure diffusion case, to arrive at a bound $|\xi_j| \leq Ch^2$ for all j . Hence (4.6) will be convergent with order two in discrete L_p -norms, under the assumption of stability.

The above proof breaks down if we allow $|a|/d \rightarrow \infty$, say $d \rightarrow 0$ with a fixed, since then $\{c_j\}$ will no longer be a smooth sequence. Moreover, then also boundedness assumptions on derivatives of u are no longer justified since a steep layer should be expected at the outflow boundary. For this limit case different type of error bounds are required. Some numerical results and remarks relevant to this case will be presented in Section 4.3.

Remark 4.1 For the pure advection problem $u_t + au_x = 0$, $a > 0$ on $\Omega = [0, \infty)$ with $u(0, t)$ given,¹¹⁾ the central scheme (4.6) may be convergent with order one rather than with order two if the grid is non-smooth. To see why there is no favourable propagation of the truncation error, in contrast to the advection-diffusion case, consider once more the relation $A\xi = \rho$ with leading truncation error terms $\rho_j = \frac{1}{2}a(\Delta_{j+1} - \Delta_j)u_{xx}(x_j, t)$. Then we get

$$\xi_{j-1} - \xi_{j+1} = (\Delta_{j+1}^2 - \Delta_j^2)v(x_j), \quad v(x_j) = \frac{1}{2}u_{xx}(x_j, t),$$

for $j \geq 1$ with $\xi_0 = 0$, see (4.12). Here ξ_1 is free to choose, reflecting singularity of A , and we can take for instance $\xi_1 = 0$. It follows that

$$\xi_j = (\Delta_1^2 - \Delta_2^2)v(x_1) + (\Delta_3^2 - \Delta_4^2)v(x_3) + \cdots + (\Delta_{j-1}^2 - \Delta_j^2)v(x_{j-1})$$

if j is even, and a similar expression is found for j odd. If the grid is unfavourable there will be no error cancellation. For example, consider a theoretical sequence with $\Delta_{2k-1} = h$, $\Delta_{2k} = \frac{1}{2}h$. Then, for j even,

$$\xi_j = \frac{3}{4}h^2(v(x_1) + v(x_3) + \cdots + v(x_{j-1})) = \mathcal{O}(h) \quad \text{for } x_j \text{ fixed, } h \rightarrow 0.$$

Of course, this choice of grid is only of theoretical interest, but it does show that for second-order convergence with the central scheme (4.6) we either need $d > 0$ or some smoothness of the grid. Numerical illustrations are given in Section 4.3. ◇

¹¹⁾ Here Ω is chosen unbounded to the right to avoid additional numerical boundary conditions. For a complete convergence proof in L_1 or L_2 -norm it can then be assumed that the solution has compact support, say $u(x, t) = 0$ for $x \geq x^*$, $0 \leq t \leq T$.

A Finite Difference Alternative

With three arbitrarily spaced grid points it is possible to discretize the advection equation $u_t + (a(x, t)u)_x = 0$ to obtain a finite difference formula with a second-order truncation error.¹²⁾ An elementary calculation shows that this is achieved by the discretization

$$w'_j = \frac{1}{2h_j} \left(\frac{h_+}{h_-} a_{j-1} w_{j-1} - \left(\frac{h_+}{h_-} - \frac{h_-}{h_+} \right) a_j w_j - \frac{h_-}{h_+} a_{j+1} w_{j+1} \right). \quad (4.14)$$

If a is constant, the spatial truncation error satisfies

$$\sigma_{h,j}(t) = \frac{1}{6} ah_+ h_- u_{xxx}(x_j, t) + \dots = \mathcal{O}(h^2).$$

If the grid is smooth, the truncation errors of (4.14) and the conservative advection discretization used in (4.6) are similar. For arbitrary grid spacings scheme (4.14) seems preferable in view of its smaller local truncation error. However, (4.14) is not conservative, and this may lead to a wrong qualitative behaviour for problems where mass conservation is essential. Moreover there is a question of stability. When brought in the linear system form $w'(t) = Aw(t)$, $A = H^{-1}B_1$, the matrix B_1 for this scheme is not skew-symmetric. Theoretical stability results are lacking, but it was found experimentally that the scheme is (weakly) L_2 -stable on finite time intervals $[0, T]$, provided the ratios between largest and smallest mesh widths are not too large. Such grids are often called *quasi-uniform*. On arbitrary grids the scheme can become unstable. For problems with spatial periodicity a relatively wide range of grids are admissible for stability. Instabilities arise more easily for advection-diffusion equations with small diffusion coefficients and Dirichlet conditions at outflow boundaries. This observation was made already in Veldman & Rinzema (1992). A related case for a cell centered scheme will be discussed more extensively in the next section. Numerical comparisons between (4.14) and (4.6) will be given in Section 4.3.

First-Order Upwind Advection

Standard upwinding for the advection equation is achieved by taking the fluxes as

$$f_{j+\frac{1}{2}}(t, w) = \begin{cases} a_{j+\frac{1}{2}} w_j & \text{if } a_{j+\frac{1}{2}} \geq 0, \\ a_{j+\frac{1}{2}} w_{j+1} & \text{if } a_{j+\frac{1}{2}} \leq 0. \end{cases} \quad (4.15)$$

If a is constant, stability for this scheme in the L_1 , L_2 and L_∞ -norm is easily established, see also Section I.4; in fact, as in that section, stability in the L_1 -norm $\|v\|_1 = \sum_j h_j |v_j|$ can be shown to hold for arbitrary variable

¹²⁾ We leave it as an exercise to show that with three arbitrarily spaced grid points a second-order truncation error for the diffusion equation $u_t = du_{xx}$ is impossible. Therefore we consider here only the advection equation.

$a(x, t)$. Convergence however is again more complicated. For constant $a > 0$ the truncation error is found to be

$$\sigma_{h,j}(t) = \frac{a}{2h_j} (\Delta_j - \Delta_{j+1}) u_x(x_j, t) - \frac{a}{2h_j} \Delta_j^2 u_{xx}(x_j, t) + \dots$$

and a similar expression is obtained for $a < 0$. Hence on non-smooth grids the local truncation error gives an inconsistency. However, here we again have a favourable propagation of the truncation error, leading to first-order convergence.

Consider, as in Remark 4.1, $u_t + au_x = 0$ with $a > 0$ on the spatial domain $\Omega = [0, \infty)$ with a Dirichlet condition at $x = 0$. Then setting $A\xi = \sigma_h$ gives

$$\xi_{j-1} - \xi_j = \frac{1}{2} (\Delta_j - \Delta_{j+1}) u_x(x_j, t) + \mathcal{O}(h^2), \quad j \geq 1, \quad \xi_0 = 0,$$

which is satisfied for instance with

$$\xi_j = \frac{1}{2} \Delta_{j+1} u_x(x_{j+1}, t) - \frac{1}{2} \Delta_1 u_x(x_1, t),$$

showing first-order convergence by Theorem I.5.2.

4.2 Cell Centered Schemes

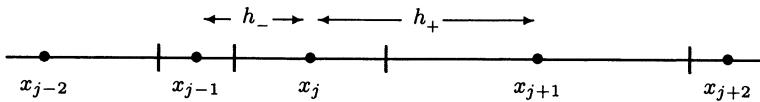
Consider a partitioning of the spatial domain into cells $\Omega_j = [x_{j-1/2}, x_{j+1/2}]$, and let the points x_j be the cell centers. So here the grid is primarily defined by the cells, that is, by the sequence $\{x_{j+1/2}\}$, and we have

$$x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}), \quad h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}. \quad (4.16)$$

As before, we denote

$$h_- = x_j - x_{j-1} = \frac{1}{2}(h_{j-1} + h_j), \quad h_+ = x_{j+1} - x_j = \frac{1}{2}(h_j + h_{j+1})$$

when we are considering a specified grid point x_j .



On this grid the standard choice for a central flux is

$$f_{j+\frac{1}{2}}(t, w) = a_{j+\frac{1}{2}} \left(\frac{h_{j+1}}{2h_+} w_j + \frac{h_j}{2h_+} w_{j+1} \right) + d_{j+\frac{1}{2}} \frac{w_j - w_{j+1}}{h_+}. \quad (4.17)$$

Note that the expression for the diffusive part has a similar form as with the vertex centered scheme, but the location of x_j and $x_{j\pm 1/2}$ is different here. The advective fluxes are obtained by linear interpolation. With this

expression for the fluxes the central non-uniform cell centered advection-diffusion discretization for (4.1) becomes

$$\begin{aligned} w'_j &= \frac{1}{2h_j} \left(\frac{a_{j-\frac{1}{2}}h_j}{h_-} w_{j-1} + \left(\frac{a_{j-\frac{1}{2}}h_{j-1}}{h_-} - \frac{a_{j+\frac{1}{2}}h_{j+1}}{h_+} \right) w_j - \frac{a_{j+\frac{1}{2}}h_j}{h_+} w_{j+1} \right) \\ &\quad + \frac{1}{h_j} \left(\frac{d_{j-\frac{1}{2}}}{h_-} w_{j-1} - \left(\frac{d_{j-\frac{1}{2}}}{h_-} + \frac{d_{j+\frac{1}{2}}}{h_+} \right) w_j + \frac{d_{j+\frac{1}{2}}}{h_+} w_{j+1} \right) + g_j. \end{aligned} \quad (4.18)$$

Upwinding can again be introduced by artificially increasing the diffusion coefficients.

The advective fluxes in (4.17) are obtained by linear interpolation. As an alternative these fluxes can simply be based on averaging

$$f_{j+\frac{1}{2}}(t, w) = \frac{1}{2} a_{j+\frac{1}{2}}(w_j + w_{j+1}), \quad (4.19)$$

with diffusive fluxes taken as in (4.17). This leads to the same formula as in (4.6); again it is the location of x_j with respect to the points $x_{j\pm 1/2}$ which is different than with the vertex centered scheme. As we will see, this modification with averaged fluxes has more favourable stability properties than (4.18). For stationary problems this scheme has been advocated in Wesseling (2001). However, for time-dependent problems with dominating advection the accuracy will turn out to be insufficient.

Consistency, Stability and Convergence Properties of (4.18)

Similar as for the vertex centered scheme, the cell centered scheme (4.18) and the modification (4.19) will be analyzed under the assumption that a and d are constant.

Consistency: Insertion of the exact solution of (4.1) into (4.18) yields for constant coefficients the truncation error

$$\begin{aligned} \sigma_{h,j}(t) &= -\frac{d}{4h_j} \left(h_{j+1} - 2h_j + h_{j-1} \right) u_{xx}(x_j, t) \\ &\quad - \frac{d}{6h_j} \left(h_+^2 - h_-^2 \right) u_{xxx}(x_j, t) + \frac{a}{8} \left(h_{j+1} - h_{j-1} \right) u_{xx}(x_j, t) + \mathcal{O}(h^2), \end{aligned} \quad (4.20)$$

with h being the maximal mesh width. Note that for arbitrary grids, without smoothness, the diffusion discretization now even leads to $\sigma_h = \mathcal{O}(h^0)$, that is inconsistency.

For smooth grids generated by a transformation $x = x(\xi)$, with underlying mesh width \bar{h} , we recover second-order consistency,

$$\begin{aligned} \sigma_h &= ah^2 \left(\frac{1}{4} x_{\xi\xi} u_{xx} + \frac{1}{6} x_\xi^2 u_{xxx} \right) \\ &\quad - d\bar{h}^2 \left(\frac{1}{4} (x_\xi)^{-1} x_{\xi\xi\xi} u_{xx} - \frac{1}{6} x_{\xi\xi} u_{xxx} + \frac{1}{12} x_\xi^2 u_{xxxx} \right) + \mathcal{O}(\bar{h}^3). \end{aligned}$$

Remark 4.2 Since with this cell centered scheme the grid is primarily defined by the cells, and not by the points x_j , it might be argued that consistency should be regarded with respect to cell average values

$$\bar{u}(x_j, t) = \frac{1}{h_j} \int_{\Omega_j} u(s, t) ds , \quad (4.21)$$

instead of point values. However, also upon inserting these cell average values into the scheme, a similar expression for the truncation error is obtained, again with inconsistency for the diffusion term. As far as second-order convergence is concerned it makes little difference whether point values or cell averages are regarded. Since

$$\bar{u}(x_j, t) = u(x_j, t) + \frac{1}{24} h_j^2 u_{xx}(x_j, t) + \mathcal{O}(h_j^4) ,$$

the difference between the two is $\mathcal{O}(h^2)$. \diamond

If we consider the scheme with averaged advective fluxes (4.19) for constant a , the leading advection contributions in the truncation error become

$$\begin{aligned} \sigma_{h,j}(t) &= a \left(\frac{h_+ + h_-}{2h_j} - 1 \right) u_x(x_j, t) + a \frac{h_+^2 - h_-^2}{4h_j} u_{xx}(x_j, t) \\ &= \frac{a}{4} \left(\frac{h_{j-1} - 2h_j + h_{j+1}}{h_j} \right) u_x(x_j, t) + \mathcal{O}(h) , \end{aligned}$$

so here we get an inconsistency on non-smooth grids for the advection problem. The terms due to diffusion remain the same as in (4.20) of course. Therefore, the leading term in the truncation error for the advection-diffusion problem $u_t + au_x = du_{xx}$ becomes

$$\sigma_{h,j}(t) = \left(\frac{h_{j-1} - 2h_j + h_{j+1}}{4h_j} \right) (au_x(x_j, t) - du_{xx}(x_j, t)) + \dots$$

This leading term vanishes for steady state problems, but otherwise we may have inconsistency. As we will see later on, convergence will still be second-order if $d > 0$, also on non-smooth grids. However, for pure advection problems the inconsistency may prevent convergence.

Stability: For the vertex centered scheme (4.6) L_2 -stability for constant coefficients was easily found to hold due to the skew-symmetry in the advective terms. With the cell centered scheme (4.18) this property is lost and there are diagonal contributions in the advective terms. Therefore, L_2 -stability of this scheme can only be demonstrated easily if the grid is smooth, in which case these diagonal contributions are $\mathcal{O}(1)$ uniformly in h . Extensive numerical tests on various non-uniform grids have shown that for problems with spatial periodicity the scheme is L_2 -stable on finite time intervals $[0, T]$ if the

ratios between largest and smallest mesh widths are bounded. For advection-diffusion problems with outflow Dirichlet conditions instabilities can easily occur if the grid is not smooth. This instability can be avoided to some extent by using upwinding in the advection discretization at the outflow boundary. These results are comparable to those of the finite difference scheme (4.14), but it was observed experimentally that the cell centered scheme (4.18) does allow for larger irregularities in the grids.

In this respect the modification (4.19) has an advantage since then again L_2 -stability will hold for Dirichlet conditions without restriction on a, d or the mesh. However, as mentioned already, this modification may lead to non-convergence for time-dependent problems on non-smooth grids, see also Remark 4.4 below.

Finally we note that as for the vertex centered scheme, it easily follows that (4.18) will be stable in the max-norm if the cell Péclet numbers $a_{j\pm 1/2}h_j/d_{j\pm 1/2}$ are bounded by 2 in modulus.

Remark 4.3 The observed difference in stability for (4.18) between the periodic case and the Dirichlet case can be understood to some extent by an eigenvalue analysis for a constant, say $a = 1$, and $d = 0$.

Then with spatial periodicity it follows by some calculations that the matrix A of (4.18) can be written as

$$A = (I + E^T) D^{-1} (E - I)$$

with shift $E(v_1, v_2, \dots, v_m)^T = (v_m, v_1, \dots, v_{m-1})^T$ and diagonal matrix $D = 2 \operatorname{diag}(\Delta_1, \dots, \Delta_m)$, $\Delta_j = x_j - x_{j-1}$. For any two square matrices M, N the eigenvalues of the product MN are the same as for NM , see for instance Horn & Johnson (1985; Thm. 1.3.20). Therefore the eigenvalues of A are the same as for

$$\tilde{A} = D^{-1} (E - I) (I + E^T) = D^{-1} (E - E^T).$$

This matrix however is similar to the skew-symmetric matrix $D^{1/2} \tilde{A} D^{-1/2}$, and thus we see that the eigenvalues of A are purely imaginary.

With Dirichlet conditions the situation is different. Suppose we have homogeneous Dirichlet conditions at $x_0 = 0, x_{m+1} = 1$, which is justified if the advection equation is considered as the limit for $d \rightarrow 0$ of $u_t + u_x = du_{xx}$. Then the matrix A of (4.18) will be tridiagonal with diagonal elements

$$a_{jj} = \frac{1}{2h_j} \left(\frac{h_{j-1}}{\Delta_j} - \frac{h_{j+1}}{\Delta_{j+1}} \right) = -\frac{1}{2\Delta_j} + \frac{1}{2\Delta_{j+1}}.$$

Hence

$$\operatorname{trace}(A) = \sum_{j=1}^m a_{jj} = -\frac{1}{2\Delta_1} + \frac{1}{2\Delta_{m+1}},$$

and since the trace of a matrix equals the sum of its eigenvalues, it is clear that A will have eigenvalues with (large) positive real part if $\Delta_{m+1} \ll \Delta_1$. This is a natural grid choice for advection-diffusion with a boundary layer at the outflow boundary point $x = 1$. To avoid instability with (4.18), upwinding near the outflow boundary will then be needed. \diamond

Convergence: As we already saw with the vertex centered schemes, the truncation error may give incorrect information about convergence. With the cell centered schemes this is even more important in view of the inconsistency of the diffusion discretization. Here we will demonstrate second-order convergence of (4.18) under the assumption of stability in a discrete L_p -norm. To begin with, we consider the pure diffusion case, $a = 0$, with scaling $d = 1$ and with Dirichlet conditions at $x = 0, 1$. For convenience and to obtain a closer resemblance with the results in Section 4.1, we assume that $x_0 = 0$ and $x_{m+1} = 1$, which means that half-cells $[0, \frac{1}{2}h_0]$ and $[1 - \frac{1}{2}h_{m+1}, 1]$ are placed at the boundaries.

First, consider only the leading term of the truncation error,

$$\rho_j = \frac{1}{h_j} (h_{j-1} - 2h_j + h_{j+1}) v(x_j), \quad v(x_j) = -\frac{1}{4} u_{xx}(x_j, t).$$

Recall that for proving second-order convergence by means of Theorem I.5.2, the essential point is to find a vector $\xi \in \mathbb{R}^m$ such that $A\xi = \rho$, $\|\xi\| = \mathcal{O}(h^2)$. Setting $A\xi = \rho$, we can follow (4.9), (4.10) and (4.11), to arrive again at

$$\xi_j = q_j - x_j q_{m+1}, \quad q_j = \sum_{k=1}^j \Delta_k p_k, \quad p_j = \sum_{k=1}^{j-1} h_k \rho_k,$$

where $\Delta_j = x_j - x_{j-1}$. Here the estimation of $|\xi_j|$ requires some care. We have

$$\begin{aligned} p_j &= \sum_{k=1}^{j-1} (h_{k-1} - 2h_k + h_{k+1}) v(x_k) = (h_0 - h_1) v(x_1) \\ &\quad - (h_{j-1} - h_j) v(x_{j-1}) + \sum_{k=2}^{j-1} (h_{k-1} - h_k) (v(x_k) - v(x_{k-1})). \end{aligned}$$

The last sum can be developed as

$$\begin{aligned} \sum_{k=2}^{j-1} (h_{k-1} - h_k) (\Delta_k v_x(x_k) + \mathcal{O}(h^2)) &= \frac{1}{2} \sum_{k=2}^{j-1} (h_{k-1}^2 - h_k^2) v_x(x_k) + \mathcal{O}(h^2) \\ &= \frac{1}{2} h_1^2 v_x(x_2) + \frac{1}{2} \sum_{k=2}^{j-2} h_k^2 (v_x(x_{k+1}) - v_x(x_k)) - \frac{1}{2} h_{j-1}^2 v_x(x_{j-1}) + \mathcal{O}(h^2), \end{aligned}$$

which gives in total an $\mathcal{O}(h^2)$ contribution. Hence

$$p_j = (h_0 - h_1) v(x_1) - (h_{j-1} - h_j) v(x_{j-1}) + \mathcal{O}(h^2).$$

It follows that

$$\begin{aligned} q_j &= \sum_{k=1}^j \Delta_k (h_0 - h_1) v(x_1) - \sum_{k=1}^j \Delta_k (h_{k-1} - h_k) v(x_{k-1}) + \mathcal{O}(h^2) \\ &= x_j (h_0 - h_1) v(x_1) - \frac{1}{2} \sum_{k=1}^j (h_{k-1}^2 - h_k^2) v(x_{k-1}) + \mathcal{O}(h^2) \\ &= x_j (h_0 - h_1) v(x_1) + \mathcal{O}(h^2). \end{aligned}$$

Since all remainder terms are $\mathcal{O}(h^2)$ uniformly in j , we obtain the estimate

$$|\xi_j| = |q_j - x_j q_{m+1}| \leq Ch^2, \quad j = 1, \dots, m,$$

with a constant $C > 0$ determined by bounds on $|v|, |v_x|, |v_{xx}|$, and thus we see that this inconsistent truncation error term will give an $\mathcal{O}(h^2)$ contribution to the global error. The same holds for the $\mathcal{O}(h)$ terms in the truncation error (4.20); this can be demonstrated just as in Section 4.1. Consequently, assuming stability in a discrete L_p -norm, the scheme will be convergent with order two in that norm.

Advection terms can be included in the analysis by using integrating factors as in (4.12), (4.13). For the scheme with averaged advection fluxes (4.19) we can use the integrating factors (4.13); for the standard scheme (4.18) with interpolated fluxes these factors need a little modification.¹³⁾ As for the vertex centered scheme, second-order convergence can then be demonstrated provided d is bounded away from 0, for both the schemes with interpolated or averaged fluxes.

Remark 4.4 Convergence can also be considered for pure advection problems as in Remark 4.1, and then the inconsistency of (4.19) may have a large impact. Consider as before $u_t + au_x = 0$ with $a > 0$ on the spatial domain $\Omega = [0, \infty)$ with a Dirichlet condition at $x = 0$. We assume for convenience that the boundary coincides with the grid point x_0 .

First we examine the standard scheme (4.18). Then setting $A\xi = \sigma_h$ gives

$$\frac{1}{\Delta_j} (\xi_{j-1} - \xi_j) + \frac{1}{\Delta_{j+1}} (\xi_j - \xi_{j+1}) = \frac{1}{4} (h_{j+1} - h_{j-1}) u_{xx}(x_j, t) + \dots$$

for $j \geq 0$ and $\xi_0 = 0$. A solution is given by

$$\xi_j = -\frac{1}{8} h_j^2 u_{xx}(x_j, t) + \frac{1}{8} h_0^2 u_{xx}(0, t),$$

showing second-order convergence without any restriction on the grid.

For the scheme with averaged fluxes (4.19) the situation is completely different. Then $A\xi = \sigma_h$ gives

$$\xi_{j-1} - \xi_{j+1} = \frac{1}{2} (h_{j-1} - 2h_j + h_{j+1}) u_x(x_j, t) + \dots$$

¹³⁾ For (4.18) one can take $c_j = \prod_{k=1}^j (d - \frac{1}{2} ah_k)/(d + \frac{1}{2} ah_k)$, $c_{j+1/2} = (d - \frac{1}{2} ah_j)c_j$ to arrive again at (4.12).

with $\xi_0 = 0$, and here we may set also $\xi_1 = 0$. It easily follows that unfavourable grid choices can be made such that there is no cancellation of truncation errors. For example, with $h_{2k-1} = h$, $h_{2k} = \frac{1}{2}h$ we obtain, omitting higher-order terms,

$$\xi_j = \frac{1}{2}h \left(u_x(x_1, t) + u_x(x_3, t) + \cdots + u_x(x_{j-1}, t) \right) = \mathcal{O}(h^0)$$

if j is even, and likewise for j odd. Numerical experiments have confirmed that indeed with this grid the scheme does not converge at all. Experiments on random grids are presented below and we will see that also then the convergence behaviour is quite poor. \diamond

First-Order Upwind Advection

Standard advection upwinding corresponds to the fluxes (4.15). As for the vertex centered scheme, stability easily follows. Assuming $a > 0$ constant, the truncation error is found to be

$$\sigma_{h,j}(t) = \frac{a}{2h_j} (h_{j-1} - h_j) u_x(x_j, t) - \frac{a}{8h_j} (h_{j-1} + h_j)^2 u_{xx}(x_j, t) + \cdots$$

and a similar expression holds for $a < 0$. Therefore the truncation error has a similar form as with the vertex centered upwind scheme, and again we have a favourable error propagation leading to first-order convergence on any grid. Elaboration of this is the same as for the vertex centered upwind scheme.

4.3 Numerical Illustrations

In this section some simple numerical tests are presented. We consider the vertex centered scheme VC_2 given by formula (4.6), the finite difference scheme FD_2 of (4.14) and the upwind vertex centered scheme VC_1 . With the latter two schemes diffusion terms are taken as in (4.6). Likewise we consider the cell centered scheme CC_2 of (4.18), its modification CC_2^a with averaged fluxes (4.19) and the upwind cell centered scheme CC_1 .

Advection on Random Grids

For a first numerical illustration we consider the pure advection problem $u_t + u_x = 0$ for $t > 0$, $x \in (0, 1)$ with periodic boundary conditions and smooth initial condition $u(x, 0) = \sin^4(\pi x)$. The problem is discretized on a sequence of random grids. Of course, solving this advection problem on random grids makes no practical sense, but it does provide an illustration for the sensitivity of the accuracy of the schemes with respect to jumps in the grid. This sensitivity has some practical relevance for multi-dimensional problems where the choice of the grids may be dictated by complicated geometrical shapes of the spatial domain Ω .

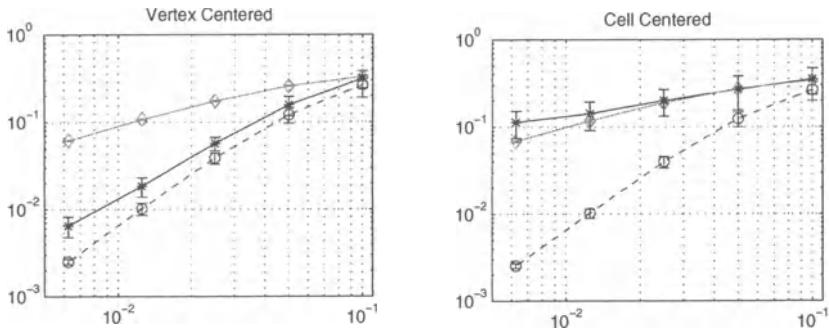


Fig. 4.1. L_2 -errors versus $1/m$ for the advection test on random grids. Solid lines and $*$ -marks for VC_2 , CC_2^a . Dashed lines and \circ -marks for FD_2 , CC_2 . Gray lines and \diamond -marks for VC_1 , CC_1 .

To construct the grids with a given number of grid points, random numbers $\omega_j \in (0, 1)$, $j = 1, \dots, m$, were produced by a random generator and normalized by $\delta_j = \omega_j / \sum_{k=1}^m \omega_k$. For the vertex centered case we took $x_0 = 0$ and $\Delta_j = x_j - x_{j-1} = \delta_j$. For the cell centered case $x_1 = \frac{1}{2}\delta_1$ and $h_j = \delta_j$ was used.

For each value of $m = 10 \cdot 2^k$, $0 \leq k \leq 4$, we performed 50 runs on different random grids and the L_2 -errors at $t = 1$ were measured. In Figure 4.1 the mean of these errors over the 50 runs are presented, together with the standard deviations indicated by error bars. We see that the results for the upwind schemes and for FD_2 , CC_2 are not very sensitive with respect to the grid variation; in particular for the upwind schemes the error bars are very close together and therefore not well visible. For VC_2 the standard deviations are larger, which is to be expected since this scheme is convergent with order two on smooth grids and with order one only on unfavourable grids, see Remark 4.1. The behaviour of the cell centered scheme with averaged advective fluxes CC_2^a is unsatisfactory. The standard deviations are large and the convergence behaviour is worse than for the first-order upwind schemes, which is not surprising in view of Remark 4.4. Thus it can be concluded that CC_2^a is not suited for advection (dominated) problems if the grid is not smooth.

Boundary Layers and Special Grids

We next consider the standard problem $u_t + u_x = du_{xx}$, $0 < x < 1$, with Dirichlet conditions $u(0) = 1$, $u(1) = 0$. This problem has the stationary solution

$$u(x) = \frac{e^{1/d} - e^{x/d}}{e^{1/d} - 1}, \quad 0 \leq x \leq 1,$$

with a boundary layer at $x = 1$ if $d > 0$ is small, see Section I.5 for an illustration. To resolve the boundary layer with the central schemes we use fine grids near $x = 1$. A number of special grids have been constructed for such

problems, see for instance Roos, Stynes & Tobiska (1996, Sect. 2.4.2). Here we consider a so-called *Shishkin grid* which consists of two uniform sub-grids with $m/2$ points on the intervals $[0, 1 - \delta]$ and $[1 - \delta, 1]$, where $\delta = Kd \ln m$ with positive constant K . It is assumed that $md < 1$. Convergence of upwind schemes on such grids, uniformly for $d > 0$, was demonstrated by Shishkin (1990).

In this test we took $d = 10^{-3}, 10^{-6}$ and $K = 2, 4$. The stationary solution has been approximated by central schemes. The L_2 -errors for a various number of grid points m are given in the Figures 4.2 and 4.3. The schemes VC_2 and CC_2^a for which L_2 -stability could be demonstrated are indicated by *-marks, the schemes FD_2 and CC_2 are indicated by o-marks. Since the grid only has one jump and the solution is stationary, the inconsistency of the scheme CC_2^a is absent here.

Also for the two cell centered schemes we used grid points (with half-cells) at the boundaries, and therefore the results for the modified cell centered scheme CC_2^a are nearly identical to those of the vertex centered scheme VC_2 . The use of virtual points and extrapolation to implement the Dirichlet conditions on a grid where the boundaries coincide with cell vertices, did lead to less accurate results for small values of d . In practice a standard cell centered grid might be used with some upwinding at outflow boundaries. In the present test the results would benefit from local upwinding, but the purpose here is to test genuine central schemes.

It is obvious from the figures that the schemes VC_2 and CC_2^a , for which L_2 -stability was established, produce much better results in these tests than the finite difference scheme FD_2 and the cell centered scheme CC_2 . With the latter schemes strongly oscillatory solutions are obtained. If the number of grid points becomes sufficiently large, the diffusion term provides stabilization. Also for larger K the oscillations become less pronounced since then the grid interface is shifted towards the region where the solution is smooth. Further we note that the convergence rates in these tests are actually less than second-order (in the range 1.6–1.7). In the derivation of the second-order results

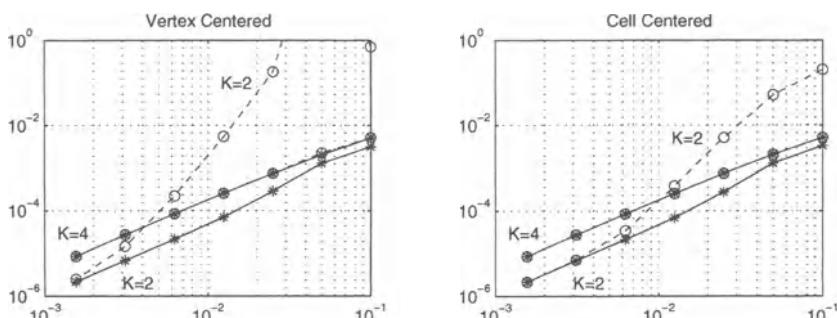


Fig. 4.2. L_2 -errors versus $1/m$ with Shishkin grids, $K = 2, 4$, for $d = 10^{-3}$. Solid lines and *-marks for VC_2 , CC_2^a . Dashed lines and o-marks for FD_2 , CC_2 .

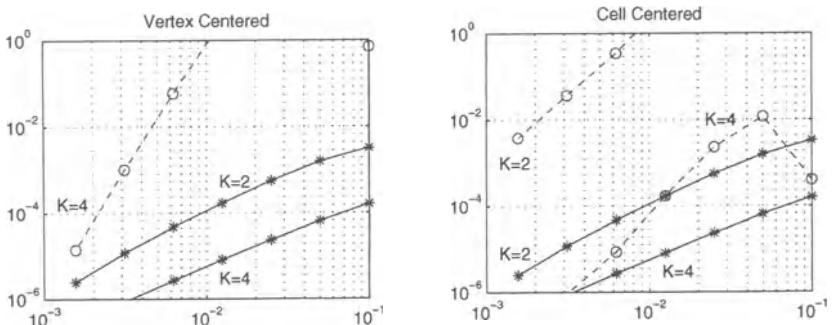


Fig. 4.3. L_2 -errors versus $1/m$ with Shishkin grids, $K = 2, 4$, for $d = 10^{-6}$. Solid lines and *-marks for VC_2 , CC_2^a . Dashed lines and o-marks for FD_2 , CC_2 . Results for FD_2 with $K = 2$ are outside the frame (errors larger than 1).

in Sections 4.1 and 4.2 smoothness of the solutions was assumed. In the boundary layer this is not a reasonable assumption, unless the grid gets very fine ($h \ll d$). Error bounds with first-order convergence, uniformly in $d > 0$, were obtained by Shishkin (1990). For related convergence results, applicable to problems with boundary layers, we refer to the books of Morton (1996) and Roos et al. (1996).

Practical Conclusions

The above analysis and tests show that the use of schemes on arbitrary (non-smooth) non-uniform grids is far from straightforward. Among the four central schemes considered here none performs without problems. The finite difference scheme FD_2 is not mass conservative and it loses stability if there are large jumps in the grid. To a lesser extent this instability is also present with the cell centered scheme CC_2 , in particular with Dirichlet conditions at the outflow. With the standard vertex centered scheme VC_2 irregularities in the grid may lead to an order of convergence less than two, but at least one. The inconsistency of the cell centered scheme CC_2^a with averaged advective fluxes leads to very inaccurate results for time-dependent advection (dominated) problems. In conclusion, the schemes VC_2 and CC_2 are recommended, although with the latter one large jumps in the grid should be avoided.

4.4 Higher-Order Methods and Limiting

The construction of higher-order schemes on non-uniform grids can be obtained by suitable piecewise interpolation for cell centered schemes with cell average values (4.21). This simple and elegant procedure was introduced by Colella & Woodward (1984), see also LeVeque (1992, Sect. 17.3). We consider here the generalization of the third-order upwind biased advection scheme to non-uniform cell centered grids.

The basic problem is as follows: given cell average values $\bar{u}(x_j, t)$ at a given time t , we need point values at the cell boundaries to construct the fluxes. For this, consider the primitive function

$$U(x) = \int_{x_L}^x u(s, t) ds,$$

where the lower limit x_L is arbitrary, for instance $x_{1/2}$. Then the value of $U(x)$ is known at the cell boundaries,

$$U(x_{j+\frac{1}{2}}) = \sum_{k=1}^j h_k \bar{u}(x_k, t).$$

On the interval $\Omega_j = [x_{j-1/2}, x_{j+1/2}]$ we can consider the interpolating cubic polynomial $W(x)$ passing through the points $(x_{k+1/2}, U(x_{k+1/2}))$ for $j-2 \leq k \leq j+1$, to obtain

$$w_{j-\frac{1}{2}}^R = W'(x_{j-1/2}), \quad w_{j+\frac{1}{2}}^L = W'(x_{j+1/2}).$$

Note that the interpolant $W(x)$ itself will be fourth-order accurate on Ω_j but by differentiation one order is lost, leading to third-order approximations at the cell boundaries. For the advective fluxes we then take

$$f_{j+\frac{1}{2}}(t, w) = \max(0, a_{j+\frac{1}{2}}) w_{j+\frac{1}{2}}^L + \min(0, a_{j+\frac{1}{2}}) w_{j+\frac{1}{2}}^R.$$

On uniform grids this procedure returns the third-order upwind-biased advection scheme.

Limiting can be applied to the resulting scheme in the same way as in Section 1, where it is easiest to interpret the limiter as an adjustment of the values $w_{j-1/2}^R, w_{j+1/2}^L$ so as to prevent spurious oscillations.

The above procedure can also be applied to reproduce the lower-order cell centered schemes. If we use a quadratic polynomial through the points $(x_{k+1/2}, U(x_{k+1/2}))$, $k = j-1, j, j+1$, to find $w_{j+1/2} = W'(x_{j+1/2})$, the advective flux of (4.17) will result. Moreover, the second derivative $W''(x_{j+1/2})$ then produces the diffusive flux of (4.17).

Remark 4.5 On vertex centered grids there does not seem to be an obvious way to generalize the third-order upwind-biased advection scheme. Here it should be noted that if we use point-wise interpolation through (x_k, w_k) , $k = j-1, j, j+1$, to find $w_{j-1/2}^R$ and $w_{j+1/2}^L$, then the resulting scheme on uniform grids is not the third-order scheme but rather the second-order scheme (1.9) with $\kappa = \frac{1}{2}$. \diamond

Remark 4.6 By the above reconstruction through primitive functions it is also easy to generalize the DST advection schemes of Section 2 to non-uniform cell centered grids. Let $W(x)$ be the piecewise polynomial interpolant of the

primitive function at time t_n , and let $x_{j-1/2}^*$ be the departure point at t_n of the characteristic that passes through $(t_{n+1}, x_{j+1/2})$. Then the DST advection scheme (2.11) is given by

$$w_j^{n+1} = w_j^n + \frac{\tau}{h_j} \left(\bar{f}_{j-\frac{1}{2}}^n - \bar{f}_{j+\frac{1}{2}}^n \right), \quad \bar{f}_{j+\frac{1}{2}}^n = \frac{1}{\tau} \left(W(x_{j+\frac{1}{2}}) - W(x_{j+\frac{1}{2}}^*) \right).$$

The departure points can be calculated by a Runge-Kutta method. For instance, Euler's method to trace the characteristics backwards gives

$$x_{j+\frac{1}{2}}^* = x_{j+\frac{1}{2}} - \tau a(x_{j+\frac{1}{2}}, t_{n+1}).$$

To achieve a higher accuracy for variable velocities we can replace here Euler's method by the explicit trapezoidal rule or the classical Runge-Kutta method, for example. \diamond

5 Non-uniform Grids – Finite Elements (1D)

As we saw in the previous section, generalizing finite difference and finite volume schemes to non-uniform grids is not a trivial matter. In particular for multi-dimensional problems with complex geometrical regions this becomes a major concern. Adaptation to complex regions is easier with finite element methods.

Some of the main features of finite elements already become clear with one-dimensional problems. In this section a brief description is given of a simple finite element method in 1D. The extension to 2D will be outlined later on. We will not try to describe finite element discretizations in a general way; it is only our intention to point out basic similarities and dissimilarities with the spatial discretizations considered thus far. More complete introductions to finite elements are given in Strang & Fix (1973), Wait & Mitchell (1985) and Johnson (1987). For more advanced and recent material we refer to Brenner & Scott (1994) and Morton (1996), and for time-dependent problems also to Thomée (1984).

5.1 The Basic Galerkin Method

In a finite element method the approximate solution is not defined at discrete points only. Instead, the numerical solution $w^h(x, t) \approx u(x, t)$ will be such that for each fixed t the function $w^h(\cdot, t)$ belongs to a finite dimensional function space. We will select spatial basis functions $\phi_j(x)$ and set $w^h(x, t) = \sum_j w_j(t) \phi_j(x)$. The basis functions are chosen piecewise polynomial with a compact support. For the basic method described here piecewise linear functions will be taken.

Stationary Problems

To introduce the method for advection-diffusion problems it is easiest to start with a time-independent problem

$$a u_x = d u_{xx} - c u + s(x), \quad 0 < x < 1, \quad (5.1)$$

with linear reaction-source term $-cu + s(x)$ and constant coefficients $a \in \mathbb{R}$ and $d, c \geq 0$. For the boundary conditions we take, as an example,

$$u(0) = \gamma_0, \quad u_x(1) = 0. \quad (5.2)$$

With Neumann conditions on both sides we would need $c > 0$ to have a unique solution.

Up to now we have not given much attention to the function space in which a solution is sought. Working with finite elements forces us to be more specific on this point. In the form (5.1) we would require u to be twice differentiable. However, we can also formulate the problem in a so-called *weak form*. If we multiply (5.1) by a test function v satisfying $v(0) = 0$ and use integration by parts, it follows that

$$\int_0^1 (du_x(x)v_x(x) + au_x(x)v(x) + cu(x)v(x)) dx = \int_0^1 s(x)v(x) dx.$$

This can be written compactly as

$$[u, v] = (s, v)$$

with inner product (\cdot, \cdot) and bilinear form $[\cdot, \cdot]$ given by

$$(w, v) = \int_0^1 w v dx, \quad [w, v] = \int_0^1 (dw_x v_x + aw_x v + cwv) dx. \quad (5.3)$$

The natural function space for considering (\cdot, \cdot) is the space $L_2[0, 1]$ consisting of functions which are square integrable, $\int_0^1 v^2 dx < \infty$. These functions need not be continuous, but for convenience we restrict our attention here to functions that are at most discontinuous in a finite number of points. For the bilinear form we consider the space

$$\mathcal{H} = \{v : v \in C^0[0, 1], v_x \in L_2[0, 1]\}.$$

This is the most simple example of a so-called Sobolev space.¹⁴⁾ Now let

$$\mathcal{V} = \{v \in \mathcal{H} : v(0) = \gamma_0\}, \quad \mathcal{V}_0 = \{v \in \mathcal{H} : v(0) = 0\}. \quad (5.4)$$

¹⁴⁾ The description given here of L_2 and \mathcal{H} suffices for our purposes, but it can be made more general and precise. See, for instance, Brenner & Scott (1994) for a rigorous description and details.

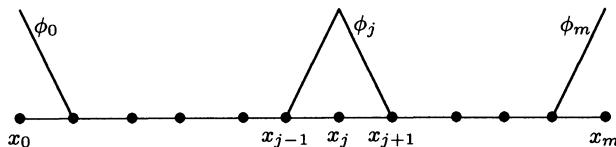
Then the weak formulation of (5.1) is as follows: we look for $u \in \mathcal{V}$ such that

$$[u, v] = (s, v) \quad \text{for all } v \in \mathcal{V}_0. \quad (5.5)$$

Suppose, for convenience, that $s \in L_2[0, 1]$. It is obvious that if u is a solution of (5.1), (5.2), then it will also satisfy (5.5). On the other hand, if u is twice differentiable and satisfies (5.5), then it can be shown that it also satisfies (5.1) and (5.2), see Brenner & Scott (1994), for example. Notice however, with the weak form (5.5) it is not necessary for u to be twice differentiable. This form is well defined if we require $u \in \mathcal{H}$.

The homogeneous Neumann condition at $x = 1$ has been used here only in the integration by parts, leading to the weak formulation (5.5). In the finite element literature a homogeneous Neumann condition is often called a *natural* boundary condition. Dirichlet conditions, which have to be built into the approximation space, are often called *essential* boundary conditions. If we had considered a nonhomogeneous Neumann condition, say $u_x(1) = \gamma_1$, this would have given an additional source term by the integration by parts.

We now come to our numerical approximation, which will be denoted by w^h , and which will be taken here piecewise linear. Let $\{x_0, x_1, \dots, x_m\}$ be a set of grid points with $x_0 = 0$ and $x_m = 1$, and let \mathcal{H}^h be the set of continuous, piecewise linear functions on $[0, 1]$ that are linear on all intervals (x_j, x_{j+1}) . We consider the so-called hat functions $\phi_j(x) \in \mathcal{H}^h$, $j = 0, 1, \dots, m$, which are such that $\phi_j(x_i) = \delta_{ij}$ (that is, 1 if $i = j$ and 0 otherwise). These hat functions span the finite dimensional function space \mathcal{H}^h .



Let \mathcal{V}^h and \mathcal{V}_0^h be the restrictions of $\mathcal{V}, \mathcal{V}_0$ to \mathcal{H}^h . For the numerical approximation we require that $w^h \in \mathcal{V}^h$ is such that

$$[w^h, v^h] = (s, v^h) \quad \text{for all } v^h \in \mathcal{V}_0^h. \quad (5.6)$$

This defines the numerical solution. Here the *trial space* \mathcal{V}^h and the *test space* \mathcal{V}_0^h are the same except for the shift by γ_0 . Finite element methods with this property are commonly called *Galerkin methods*.¹⁵⁾

A more transparent form of (5.6) is obtained by invoking the basis functions ϕ_j . Writing

$$w^h(x) = \sum_{j=0}^m w_j \phi_j(x),$$

¹⁵⁾ Methods where the trial and test space coincide are sometimes also called Bubnov-Galerkin methods, to make a clearer distinction with Petrov-Galerkin methods where these spaces differ.

and taking the test function v^h to be ϕ_j for $j = 1, \dots, m$, we obtain the following linear system for the coefficients w_1, \dots, w_m ,

$$\sum_{k=0}^m [\phi_k, \phi_j] w_k = (s, \phi_j), \quad j = 1, 2, \dots, m, \quad (5.7)$$

in addition to $w_0 = \gamma_0$. The entries $[\phi_k, \phi_j]$ are easily calculated; in particular we have $[\phi_k, \phi_j] = 0$ if $|k - j| > 1$. Let $\Delta_j = x_j - x_{j-1}$ and $h_j = \frac{1}{2}(\Delta_j + \Delta_{j+1})$. Then the discretization on the interior grid points can be written as

$$\begin{aligned} \frac{a}{2h_j}(-w_{j-1} + w_{j+1}) - \frac{d}{h_j}\left(\frac{1}{\Delta_j}w_{j-1} - \left(\frac{1}{\Delta_j} + \frac{1}{\Delta_{j+1}}\right)w_j + \frac{1}{\Delta_{j+1}}w_{j+1}\right) \\ + \frac{c}{6h_j}\left(\Delta_j w_{j-1} + 4h_j w_j + \Delta_{j+1} w_{j+1}\right) = \frac{1}{h_j}(s, \phi_j). \end{aligned} \quad (5.8)$$

If the source term is taken piecewise linear, $s(x) = \sum_k s_k \phi_k(x)$, then also

$$\frac{1}{h_j}(s, \phi_j) = \frac{1}{6h_j}\left(\Delta_j s_{j-1} + 4h_j s_j + \Delta_{j+1} s_{j+1}\right).$$

In (5.8) we have divided by h_j to get an easier recognition of the terms originating from advection and diffusion. These have the same form as with the vertex centered finite volume formula (4.6) on non-uniform grids. With (5.6) the derivation is directly obtained on non-uniform grids. If we assume the grid to be uniform we regain the standard second-order central discretizations for advection and diffusion.

Time-Dependent Problems

For the time-dependent problem

$$u_t + a u_x = d u_{xx} - c u + s(x, t), \quad 0 < x < 1, \quad 0 < t \leq T, \quad (5.9)$$

with given $u(x, 0)$ and boundary conditions (5.2), we can follow the same derivations by viewing the u_t term as a forcing (source) term; this simply means modifying s to $\tilde{s} = s - u_t$. The weak formulation for (5.9), with $u(\cdot, t) \in \mathcal{V}$ for all $t \in [0, T]$, then reads

$$(u_t, v) + [u, v] = (s, v) \quad \text{for all } v \in \mathcal{V}_0, \quad 0 < t \leq T. \quad (5.10)$$

The numerical approximation $w^h(x, t)$ is required to satisfy

$$(w_t^h, v^h) + [w^h, v^h] = (s, v^h) \quad \text{for all } v^h \in \mathcal{V}_0^h, \quad 0 < t \leq T, \quad (5.11)$$

together with the initial condition $w^h(x, 0) = u_0^h(x)$, where $u_0^h \in \mathcal{V}^h$ is some suitable representation of $u(x, 0)$.

In terms of the basis functions ϕ_j we set

$$w^h(x, t) = \sum_{j=0}^m w_j(t) \phi_j(x) \quad (5.12)$$

with time-dependent coefficients $w_j(t)$. The discrete weak form (5.11) leads to a system similar to (5.8), but with on the left-hand side additional time-derivatives

$$\frac{1}{h_j}(w_t^h, \phi_j) = \frac{1}{6h_j} (\Delta_j w'_{j-1}(t) + 4h_j w'_j(t) + \Delta_{j+1} w'_{j+1}(t)). \quad (5.13)$$

The resulting system has a close resemblance to the compact central difference schemes that were considered in Section 3, and for advection-diffusion on uniform grids we regain (3.13), which is second-order accurate in general but of fourth order for pure advection. Due to this property, the finite element method will in general be much more accurate than the corresponding second-order central finite difference scheme for advection dominated equations; see the discussion in Section 3.2 and in particular the Figures 3.1 and 3.2.

If we denote $w(t) = (w_j(t))_{j=1}^m$ in \mathbb{R}^m , then the semi-discrete system for homogeneous boundary conditions can be written in the form

$$Bw'(t) = Aw(t) + Bg(t), \quad (5.14)$$

where

$$A = (a_{jk}) = -([\phi_k, \phi_j])_{j,k=1}^m, \quad B = (b_{jk}) = ((\phi_k, \phi_j))_{j,k=1}^m, \quad (5.15)$$

and $g(t) = (s_j(t))_{j=1}^m$ for a piecewise linear source term. Inhomogeneous boundary conditions will give additional terms in the first and last component of $g(t)$. For finite element methods, the matrix B is usually called the *mass matrix* and A is known as the *stiffness matrix*. This terminology is due to the original applications of finite elements in structural mechanics.

Remark 5.1 Application of an explicit time integration method will give some implicitness due to the multiplication of $w'(t)$ by B ; see also Section 3.5. To avoid this we could use the approximation

$$\frac{1}{h_j}(w_t^h, \phi_j) \approx w'_j(t).$$

Then the semi-discrete system becomes $w'(t) = Aw(t) + Bg(t)$. For the constant-coefficient problem $u_t + au_x = du_{xx}$ the discretization then becomes equivalent to the finite volume scheme (4.6); treatment of the source terms will still be different from (4.6). This procedure, replacing $Bw'(t)$ by $w'(t)$, is called *mass lumping*. It can also be achieved by taking low-order quadrature to approximate the integrals, e.g. the trapezoidal rule. The numerical comparison in Section 3.2 on uniform grids does show that mass lumping may have a negative effect on the accuracy of the scheme. \diamond

Remark 5.2 For problems with variable coefficients and nonlinear reactions, formulation of the finite element method with piecewise linear functions for trial space and test space can be obtained in the same manner, by allowing variable coefficients in the bilinear form in formula (5.3); see also Section 6.5. Then the integrals giving the matrix entries in (5.15) are approximated by some suitable quadrature rule. A natural choice is Simpson quadrature

$$\int_{x_{j-1}}^{x_j} \psi(x) dx \approx \frac{1}{6} \Delta_j (\psi(x_{j-1}) + 4\psi(x_{j-\frac{1}{2}}) + \psi(x_j)),$$

since this gives exact entries in the semi-discrete system for the model problem (5.9) with constant coefficients. For higher-order finite element methods, based on higher-order piecewise polynomials, a corresponding quadrature rule will be used to approximate the entries $[\phi_k, \phi_j]$ and (ϕ_k, ϕ_j) in (5.11). ◇

Remark 5.3 With finite element methods the numerical approximation is given by $w^h(x, t) = \sum_j w_j(t)\phi_j(x)$ with basis functions that have local support. The use of global basis functions, like Legendre or Chebyshev polynomials, leads to *spectral methods*. Such methods can be very accurate for smooth solutions, but they lack the flexibility of finite elements (and finite volumes) for local adaptation, in particular with multi-dimensional problems. Spectral methods are not treated in this text; instead we refer to the books of Canuto et al. (1988), Quarteroni & Valli (1994) and Boyd (2001). ◇

5.2 Standard Galerkin Error Estimates

To analyze the error for the above finite element method we can follow the usual consistency-stability argument applied to (5.14) as in Section 3.3. However, the fact that the numerical approximation w^h is now a function in x , rather than a grid function, allows for different types of error estimates. We give here an illustration essentially based on material from Strang & Fix (1973). It is assumed that $d > 0$, excluding the pure advection case. In the following estimates, some of the constants will depend on the Péclet number $|a|L/d$ where $L=1$ is the length of the spatial interval. If $a < 0$ it is assumed that the Péclet number is less than 2.

Stationary Problems

Let us first consider the time-independent problem (5.1) with $d > 0$ and $c \geq 0$. Then

$$\|v\|_* = \left(\int_0^1 (d v_x^2 + c v^2) dx \right)^{1/2} \quad (5.16)$$

defines a norm on \mathcal{V}_0 , often called the *energy norm*. We will also deal with the familiar L_2 -norm $\|v\| = (v, v)^{1/2}$. It can be shown by some calculations that we have, for all $v, w \in \mathcal{V}_0$,

$$[v, w] \leq C_1 \|v\|_* \|w\|_*, \quad [v, v] \geq C_2 \|v\|_*^2 \quad (5.17)$$

with $C_1, C_2 > 0$ depending on the Péclet number; in case $a = 0$ it is easily verified that these properties hold with $C_1 = C_2 = 1$.¹⁶⁾ The two inequalities in (5.17) are called, respectively the boundedness and *coercivity* property of the bilinear form.

Since we have $\mathcal{V}^h \subset \mathcal{V}$ and $\mathcal{V}_0^h \subset \mathcal{V}_0$ – finite elements with this property are called *conforming* – it follows that $[u, v^h] = (s, v^h)$ for any $v^h \in \mathcal{V}_0^h$, and thus

$$[u - w^h, v^h] = 0 \quad \text{for all } v^h \in \mathcal{V}_0^h. \quad (5.18)$$

Using $u - w^h \in \mathcal{V}_0$, we thus see that

$$C_2 \|u - w^h\|_*^2 \leq [u - w^h, u - w^h] = [u - w^h, u - v^h] \leq C_1 \|u - w^h\|_* \|u - v^h\|_*$$

whenever $v^h \in \mathcal{V}^h$. Hence we get the following error estimate in the energy norm,

$$\|u - w^h\|_* \leq C \min_{v^h \in \mathcal{V}^h} \|u - v^h\|_*, \quad (5.19)$$

with $C = C_1/C_2$. So in this energy norm the error $u - w^h$ is determined by how well u can be approximated in \mathcal{V}^h . Since this norm also involves the spatial derivative, which is approximated by piecewise constants, we can expect no better than first-order convergence. By choosing v^h in (5.19) to be the interpolating function for the solution u , that is $v^h(x) = \sum_j u(x_j) \phi_j(x)$, it follows that

$$\|u - w^h\|_* \leq Kh \|u_{xx}\| \quad (5.20)$$

whenever u is twice differentiable, where $h = \max_j (x_j - x_{j-1})$ denotes the maximal mesh width. Note that, although this is only a first-order convergence result, we do obtain it directly on non-uniform meshes and in the energy norm. In the L_2 -norm we can show second-order convergence, but to establish that some additional analysis is needed.

An elegant way to show this second-order convergence within the finite element framework is to consider the auxiliary problem for $\psi \in \mathcal{V}_0$,

$$[\psi, v] = (u - w^h, v) \quad \text{for all } v \in \mathcal{V}_0,$$

where the error $u - w^h$ plays the role of the source term in the original equation. In the same way as we have $\|u_{xx}\| \leq M \|s\|$ for (5.1) if $\gamma_0 = 0$,¹⁷⁾ it follows that the second derivative of ψ can be bounded by this auxiliary source term as

$$\|\psi_{xx}\| \leq M \|u - w^h\|$$

¹⁶⁾ If $a = 0$ then (5.17) follows directly from $\int_0^1 (dw_x v_x + cvw) dx \leq \|v\|_* \|w\|_*$ which is a consequence of the Cauchy-Schwarz inequality for inner products, see Horn & Johnson (1985, Thm. 5.1.4) for example. If $a \neq 0$ some additional calculations are needed; this is left as a (lengthy) exercise, where one should use $v(x) = \int_0^x v'(y) dy$ for $v \in \mathcal{V}_0$. For $a < 0$ it has to be assumed that the Péclet number is less than 2 to show coercivity. Note however that if $a < 0$ then it will be more natural in general to impose the Neumann condition at the left boundary.

¹⁷⁾ As for the coercivity property (5.17), this easily follows for $a = 0$ by taking the inner product in (5.1) with u_{xx} . If $a \neq 0$ some additional calculations are needed.

with a constant $M > 0$. Moreover, taking $v = u - w^h$, it follows that

$$\|u - w^h\|^2 = [\psi, u - w^h] = [\psi - v^h, u - w^h] \leq C_1 \|\psi - v^h\|_* \|u - w^h\|_*$$

for any $v^h \in \mathcal{V}_0^h$. In view of the estimate (5.20) for our original problem (5.1), it follows that there is a $v^h \in \mathcal{V}_0^h$ such that $\|\psi - v^h\|_* \leq Kh\|\psi_{xx}\|$. Combining the above, including (5.20), we thus obtain the second-order result

$$\|u - w^h\| \leq Ch^2 \|u_{xx}\|. \quad (5.21)$$

Notice that for the derivation of these error bounds only the (orthogonality) property (5.18) and the inequalities (5.17) were used.

Time-Dependent Problems

For the time-dependent problem (5.9) with $d > 0$, $c \geq 0$, consider the weak forms (5.10), (5.11) for the exact solution $u(x, t)$ and the numerical approximation $w^h(x, t)$. We introduce $u^h(\cdot, t) \in \mathcal{V}^h$, for $t \in [0, T]$, such that

$$[u - u^h, v^h] = 0 \quad \text{for all } v^h \in \mathcal{V}_0^h, \quad 0 \leq t \leq T. \quad (5.22)$$

If $a = 0$, then $[\cdot, \cdot]$ is an inner product, in which case u^h can be viewed as an orthogonal projection of u on \mathcal{V}^h . Property (5.22) is similar to (5.18), and hence, by the above results for stationary problems, we can conclude that

$$\|u - u^h\| \leq Ch^2 \|u_{xx}\|, \quad \|u_t - u_t^h\| \leq Ch^2 \|u_{txx}\|.$$

The error $u - w^h$ can be decomposed as

$$u - w^h = (u - u^h) + \varepsilon^h, \quad \varepsilon^h = u^h - w^h.$$

From the weak forms (5.10), (5.11) and (5.22) it follows that

$$(\varepsilon_t^h, \varepsilon^h) + [\varepsilon^h, \varepsilon^h] = (u_t^h - u_t, \varepsilon^h).$$

By coercivity, we know that $-[\varepsilon^h, \varepsilon^h] \leq -C_2 \|\varepsilon^h\|_*^2 \leq \omega \|\varepsilon^h\|^2$ for some constant $\omega < 0$,¹⁸⁾ and consequently

$$\frac{d}{dt} \|\varepsilon^h\| \leq \omega \|\varepsilon^h\| + \|u_t^h - u_t\|.$$

Combining these estimates we finally arrive at

$$\begin{aligned} \|u(\cdot, t) - w^h(\cdot, t)\| &\leq e^{\omega t} \|u^h(\cdot, 0) - w^h(\cdot, 0)\| \\ &+ Ch^2 \|u_{xx}(\cdot, t)\| + \frac{1}{\omega} (e^{\omega t} - 1) Ch^2 \max_{0 \leq s \leq t} \|u_{txx}(\cdot, s)\|, \end{aligned} \quad (5.23)$$

¹⁸⁾ Note that $\|v\|^2 \leq L\|v\|_\infty^2 \leq L^2\|v_x\|^2 \leq (L^2/d)\|v\|_*^2$, for $v \in \mathcal{V}_0$, where $L = 1$ is the length of the spatial interval. Here $\|v\|_\infty^2 \leq L\|v_x\|^2$ is a consequence of Schwarz's inequality,

$$|v(x)|^2 = |\int_0^x v'(y) dy|^2 \leq (\int_0^x 1^2 dy)(\int_0^x |v'(y)|^2 dy).$$

which gives second-order convergence in the L_2 -norm.

It is obvious that on uniform grids the usual estimates, by means of stability and consistency, are more transparent and easier. Moreover, the above proof is valid only for $d > 0$, due to the reliance on the energy norm estimate (5.20), so the pure advection case is excluded. On the other hand, within the finite element framework we can directly obtain results on non-uniform grids. Although we have considered here only the most simple finite element method with piecewise linear functions, these results also extend to higher-order methods and to multi-dimensional problems.

5.3 Upwinding

For advection dominated problems the standard Galerkin schemes will produce spatial oscillations, just as the central finite difference schemes. Upwinding is then the obvious modification. To include upwinding in finite element schemes we consider *Petrov-Galerkin* methods in which the test space is chosen differently from the trial space. As in the rest of this section this is briefly illustrated here for the time-dependent problem (5.9) with boundary conditions (5.2) and with trial space \mathcal{V}^h consisting of the piecewise linear functions.

In Petrov-Galerkin methods test spaces $\tilde{\mathcal{V}}_0^h \neq \mathcal{V}_0^h$ are considered. The numerical solution $w^h \in \mathcal{V}^h$ is required to satisfy

$$(w_t^h, v^h) + [w^h, v^h] = (s, v^h) \quad \text{for all } v^h \in \tilde{\mathcal{V}}_0^h, \quad 0 < t \leq T. \quad (5.24)$$

Instead of the hat functions ϕ_j we will use basis functions ψ_j for $\tilde{\mathcal{V}}_0^h$. The test space will be chosen such that there is a tilt in the upwind direction. A simple choice is

$$\psi_j(x) = \phi_j(x) + \kappa \sigma_j(x), \quad (5.25)$$

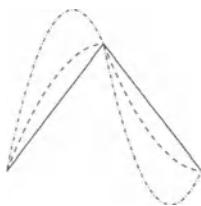
where κ is an upwind parameter and

$$\sigma_j(x) = \begin{cases} (3/\Delta_j^2)(x - x_{j-1})(x_j - x) & \text{for } x_{j-1} \leq x \leq x_j, \\ -(3/\Delta_{j+1}^2)(x - x_j)(x_{j+1} - x) & \text{for } x_j \leq x \leq x_{j+1}. \end{cases}$$

The upwind parameter can be taken as a function of the local cell Péclet number $\mu = ah_j/d$, for instance

$$\kappa = \frac{e^\mu + 1}{e^\mu - 1} - \frac{2}{\mu},$$

similar to Example 3.4, with exponential fitting. An illustration of ψ_j is given on the right for the values $\mu = 2$ (dashed) and $\mu = 20$ (dash-dots).



Elaboration of the integrals in (5.24) reveals that on uniform grids we get the same advection-diffusion terms in the stiffness matrix A as with the exponentially fitted schemes of Example 3.4. With these test functions, however,

the mass matrix B may contain negative entries. A related scheme can be obtained with non-negative test functions ψ_j of exponential type, see Hemker (1977) and Morton (1996, Sect. 5.3). This gives a scheme with a non-negative mass matrix, close to the El-Mistikawy–Werle difference scheme which was considered in Example 3.4. The entries for the mass matrix B can be somewhat manipulated by taking various quadrature rules. Hence with suitable choices of the test space, the various 3-point upwind-type difference schemes of Section 3 are produced. Of course with this Petrov-Galerkin formulation non-uniform grids are directly incorporated. Extension to variable coefficients is also straightforward.

As we saw in Section 3, these common Petrov-Galerkin schemes will not give monotonic solutions for time-dependent problems. In fact the experiments in Section 3 showed quite large oscillations for the implicit adaptive upwind scheme and the results with the exponential type test space are close to this.¹⁹⁾ It should be stressed however that Petrov-Galerkin upwinding is very successful with stationary problems and for such problems it is often used in practice. For time-dependent problems additional limiting of the mass matrix seems necessary, see Berzins (2001).

Remark 5.4 In the above only spatial discretization has been considered. Fully discrete schemes can be found with the method of lines approach. An alternative is provided by the *Taylor-Galerkin methods* where one starts from the truncated Taylor series, for instance with the second-order approximation

$$u(x, t_{n+1}) \approx u(x, t_n) + \tau u_t(x, t_n) + \frac{1}{2} \tau^2 u_{tt}(x, t_n).$$

Then the temporal derivatives u_t and u_{tt} are replaced by spatial derivatives by using the PDE, and only then spatial discretization is performed by finite elements. For pure advection problems with finite differences this procedure will yield the Lax-Wendroff scheme. If we let $\tau \rightarrow 0$, again semi-discrete systems (5.11), (5.24) will result. Due to the occurrence of mass matrices, explicit time stepping with finite elements becomes somewhat less attractive than for finite differences and finite volumes, see also Section 3.5. ◇

6 Multi-dimensional Aspects

Up to now we have mainly restricted our attention to one-dimensional (1D) advection-diffusion equations $u_t + (au)_x = (Du_x)_x + s(x, t)$. Most problems of practical interest are multi-dimensional. Although considerations in one spatial dimension are often the basis for schemes that are also applicable

¹⁹⁾ With the quadratic elements in (5.25) the oscillations in the time-dependent test problems $u_t + au_x = du_{xx}$ of Section 3 are not that pronounced, but then monotonicity for stationary problems is no longer guaranteed due to negative entries in the mass matrix, see Remark 3.8.

to higher dimensions, there are certain features that do not have a 1D analogue. The object of this section is to point out some essential concepts for multi-dimensional advection-diffusion schemes. Some of the more theoretical aspects are discussed for Cartesian grids, on which one-dimensional schemes are naturally extended. In multi-dimensional calculations major issues often originate from complex geometries where standard Cartesian grids are no longer applicable. These aspects will be discussed here only briefly with references to more specialized literature.

6.1 Cartesian Grid Discretizations

The standard form of advection-diffusion equations in multiple spatial dimensions with Cartesian coordinates $\underline{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ reads²⁰⁾

$$u_t + \sum_{k=1}^d (a_k u)_{x_k} = \sum_{k=1}^d (d_k u_{x_k})_{x_k} + s(\underline{x}, t). \quad (6.1)$$

Here the source term s may also contain nonlinear reaction terms, in which case we will write $s(\underline{x}, t, u)$, and further nonlinearities may be present in the advection or diffusion terms. The equation is compactly written as

$$u_t + \nabla \cdot (\underline{a} u) = \nabla \cdot (D \nabla u) + s(\underline{x}, t), \quad (6.2)$$

where $\underline{a} = (a_k) \in \mathbb{R}^d$ and $D = \text{diag}(d_k)$. Often we will have a scalar coefficient D (isotropic diffusion). On the other hand, more general forms with non-diagonal D also occur, in particular after a transformation of the spatial region, see Section 6.4 below.

Working on Cartesian grids, we can simply insert the one-dimensional spatial discretizations for the individual terms $(a_k u)_{x_k}$ and $(d_k u_{x_k})_{x_k}$ in the various directions, to arrive at a semi-discrete system. Reaction terms are then also easily included. It is this possibility of superposition that makes the method of lines approach quite popular. Methods with combined space-time discretizations, such as the Lax-Wendroff method, are much harder to formulate for multi-dimensional advection-diffusion-reaction problems.

In the following we will mainly restrict ourselves to two-dimensional problems. The main concepts for problems with a higher dimension are similar. Of course, technical issues such as data handling and specific implementations become much more prominent then.

Point Values and Cell Averages

Insertion of the one-dimensional discretizations into a multi-dimensional problem on a Cartesian grid is natural in a finite difference setting. In two

²⁰⁾ This notation with x_k being a component of $\underline{x} \in \mathbb{R}^d$ is used only locally in these paragraphs and in Section 6.4. If $d = 2, 3$ we will write the vector \underline{x} as $(x, y)^T$ or $(x, y, z)^T$, and x_i will denote a grid point on the x -axis, as will be clear from the context.

dimensions, the unknowns are then viewed as point values in grid points (x_i, y_j) . Spatial accuracy considerations based on the local truncation errors directly carry over.

Deriving the multi-dimensional discretizations can also be done entirely within the finite volume framework by considering inflow and outflow over cell boundaries. Then the unknowns are primarily viewed as averages over cells

$$\Omega_{ij} = [x_i - \frac{1}{2}\Delta x, x_i + \frac{1}{2}\Delta x] \times [y_j - \frac{1}{2}\Delta y, y_j + \frac{1}{2}\Delta y].$$

For the two-dimensional advection-diffusion problem

$$u_t + f(u, u_x)_x + g(u, u_y)_y = 0 \quad (6.3)$$

in conservation form, with advection-diffusion fluxes f, g , this results in the conservative scheme ²¹⁾

$$w'_{ij}(t) = \frac{1}{\Delta x} \left(f_{i-\frac{1}{2},j}(t) - f_{i+\frac{1}{2},j}(t) \right) + \frac{1}{\Delta y} \left(g_{i,j-\frac{1}{2}}(t) - g_{i,j+\frac{1}{2}}(t) \right). \quad (6.4)$$

The fluxes $f_{i\pm\frac{1}{2},j}$ and $g_{i,j\pm\frac{1}{2}}$ can be calculated from the one-dimensional expressions. In view of the integral form of the equation, they are in fact midpoint quadrature approximations of the line integrals over the cell boundaries. In general this is expected to give a scheme with order not greater than two. Still the fluxes may be calculated using higher-order expressions in view of the shape-preservation properties.

Cell-averaged source terms are simply added to the right-hand side of (6.4). With reactions terms $s(x, y, t, u)$ we often take $s(x_i, y_j, t, w_{ij}(t))$ in the semi-discrete system. Although this is common practice, with space-dependent or nonlinear reactions it is not entirely consistent with the cell-average interpretation, since we then actually again view w_{ij} as a point value; for complete consistency we would need a reaction description that works with cell averages.

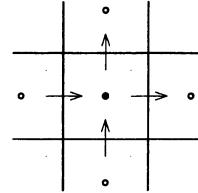
Irrespective of the spatial dimension, the difference between a cell-average value and the value in a cell center is of second order. In the two-dimensional case we have

$$\begin{aligned} & \frac{1}{\Delta x \Delta y} \int_{\Omega_{ij}} u(x, y, t) dx dy - u(x_i, y_j, t) \\ &= \frac{1}{24} \Delta x^2 u_{xx}(x_i, y_j, t) + \frac{1}{24} \Delta y^2 u_{yy}(x_i, y_j, t) + \mathcal{O}(\Delta x^4) + \mathcal{O}(\Delta y^4). \end{aligned}$$

²¹⁾ Recall that for $u_t + \nabla \cdot \underline{f} = 0$ we have, in view of Gauss' divergence theorem, the integral form

$$\int_{\Omega_{ij}} u_t d\Omega = - \int_{\Omega_{ij}} (\nabla \cdot \underline{f}) d\Omega = - \int_{\Gamma_{ij}} (\underline{f} \cdot \underline{n}) d\Gamma$$

with Γ_{ij} the boundary of Ω_{ij} , and \underline{n} the outward normal vector on the boundary.



Therefore, the order p of a discretization may depend on the interpretation, either as finite differences (with point values) or as finite volumes (with cell averages), in case $p \geq 3$. For $p \leq 2$ the interpretation does not matter in this respect.

6.2 Diffusion on Cartesian Grids

The common 2D scalar diffusion equation with a reaction-source term has the form

$$u_t = (d_1(u) u_x)_x + (d_2(u) u_y)_y + s(x, y, t, u), \quad (6.5)$$

where the diffusion coefficients may also depend on x, y and t . Standard second-order discretization, on a uniform Cartesian grid, then gives the semi-discrete system

$$\begin{aligned} w'_{ij} &= \frac{1}{\Delta x^2} \left(d_{1,i-\frac{1}{2},j} (w_{i-1,j} - w_{ij}) - d_{1,i+\frac{1}{2},j} (w_{ij} - w_{i+1,j}) \right) \\ &\quad + \frac{1}{\Delta y^2} \left(d_{2,i,j-\frac{1}{2}} (w_{i,j-1} - w_{ij}) - d_{2,i,j+\frac{1}{2}} (w_{ij} - w_{i,j+1}) \right) + s_{ij}, \end{aligned} \quad (6.6)$$

where $w_{ij} = w_{ij}(t)$. We can take, for instance, $s_{ij} = s(x_i, y_j, t, w_{ij})$ and

$$d_{1,i+\frac{1}{2},j} = d_1(x_{i+\frac{1}{2}}, y_j, t, \frac{1}{2}w_{ij} + \frac{1}{2}w_{i+1,j}),$$

and likewise for the coefficients $d_{2,i,j+1/2}$ in the vertical direction.

The standard tool to investigate stability is the von Neumann analysis, just as for the 1D case. For that we first linearize, and then freeze the coefficients with omission of constant terms. This means that instead of the varying d_k we insert a constant coefficient d_k^* which should represent the variable coefficient at specific states. Constant terms are omitted since stability actually deals with the difference between solutions. In a similar way the reaction term is replaced by a linear term c^*u with constant c^* (for systems this would be a matrix). Then for this constant coefficient problem we can perform an analysis of the eigenvalues of the linearized difference operator, which will be a normal matrix for Dirichlet boundary conditions or spatial periodicity conditions.

The eigenvalues of the two-dimensional difference operator will be of the form

$$\lambda = \lambda_1 + \lambda_2$$

where λ_1 will stand for the eigenvalues of the operator in the x -direction and λ_2 in the y -direction. These are of the same form as in 1D. The semi-discrete scheme will be stable if λ_1 and λ_2 are in the left half (complex) plane. If we also consider a time discretization method, with stability region \mathcal{S} , the stability requirement for step size τ will read

$$\tau\lambda = \tau\lambda_1 + \tau\lambda_2 \in \mathcal{S}.$$

If the reaction is stiff, we get a large frozen coefficient $c^* \ll 0$ for that term, and then this contribution should also be incorporated in the stability requirement. With non-stiff reactions we may neglect this contribution; for genuine source terms $s(x, y, t)$ we even have $c^* = 0$.

With the standard second-order discretization we will get eigenvalues in the range

$$\lambda_1 \in [-4d_1^*/\Delta x^2, 0], \quad \lambda_2 \in [-4d_2^*/\Delta y^2, 0],$$

see the elaborated example below for the heat equation. If the time stepping method has the real stability boundary β_R , that is, the line segment $[-\beta_R, 0]$ is precisely contained in its stability region, then the corresponding requirement will be

$$4 \frac{\tau d_1^*}{\Delta x^2} + 4 \frac{\tau d_2^*}{\Delta y^2} \leq \beta_R. \quad (6.7)$$

Hence the stability requirement in 2D can be viewed as an addition of the 1D requirements. The same will hold if we use higher-order discretizations, for which the one-dimensional formulas were discussed in Section I.3.

The 2D Heat Equation with Dirichlet Conditions

The above general principles will be described here in detail for the two-dimensional heat equation with source term and Dirichlet boundary conditions on the unit square,

$$\begin{aligned} u_t &= u_{xx} + u_{yy} + s(x, y, t) \quad \text{on } \Omega, \\ u(x, y, t) &= u_\Gamma(x, y, t) \quad \text{on } \Gamma = \partial\Omega, \\ u(x, y, 0) &= u_0(x, y) \quad \text{on } \Omega, \end{aligned} \quad (6.8)$$

where $t > 0$ and $(x, y) \in \Omega = (0, 1) \times (0, 1)$. Using a Cartesian grid and standard second-order spatial discretization, we get a linear semi-discrete system

$$w'(t) = Aw(t) + g(t).$$

The matrix A consists of two parts, $A = A_1 + A_2$, where A_1 acts in the x -direction and A_2 in the y -direction. Both A_1 and A_2 are essentially one-dimensional.

To be more specific, consider the Cartesian grid

$$\Omega_h = \{(x_i, y_j) : x_i = i\Delta x, y_j = j\Delta y, 1 \leq i \leq m_1, 1 \leq j \leq m_2\}$$

with mesh widths $\Delta x = (m_1 + 1)^{-1}$, $\Delta y = (m_2 + 1)^{-1}$. We identify grid-functions on Ω_h and vectors in \mathbb{R}^M , $M = m_1 m_2$, in a natural way assuming a row-wise ordering on Ω_h . Thus a grid function $w : \Omega_h \rightarrow \mathbb{R}^M$ is identified with

$$w = (w_1^T, \dots, w_{m_2}^T)^T \in \mathbb{R}^M \quad \text{with} \quad w_j = (w_{1j}, \dots, w_{m_1 j})^T \in \mathbb{R}^{m_1} \quad (6.9)$$

and $w_{ij}(t) \approx u(x_i, y_j, t)$.

Let I_m be the $m \times m$ identity matrix. Denote by B_m the discretized 1D diffusion operator based on m (internal) grid points and Dirichlet conditions,

$$B_m = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & \ddots & & \\ & \ddots & \ddots & 1 & \\ & & 1 & -2 & \end{pmatrix} \in \mathbb{R}^{m \times m}, \quad h = \frac{1}{m+1}.$$

The matrices A and A_1, A_2 can then be written as

$$A = A_1 + A_2, \quad A_1 = I_{m_2} \otimes B_{m_1}, \quad A_2 = B_{m_2} \otimes I_{m_1}, \quad (6.10)$$

where \otimes is the direct (Kronecker) product.²²⁾

The boundary values are incorporated in the grid-function

$$b(t) = b_1(t) + b_2(t) \in \mathbb{R}^M,$$

with $b_1(t)$ having non-zero components $\Delta x^{-2} u_\Gamma(x \pm \Delta x, y, t)$ on $(x, y) \in \Omega_h$ adjacent to the two vertical boundaries, and zeros elsewhere. Likewise, $b_2(t)$ has non-zero components $\Delta y^{-2} u_\Gamma(x, y \pm \Delta y, t)$ on grid points $(x, y) \in \Omega_h$ adjacent to the two horizontal boundaries, and zeros elsewhere. Taking the grid function $f(t)$ as the restriction of $s(x, y, t)$ to Ω_h , we have

$$g(t) = b(t) + f(t) \in \mathbb{R}^M. \quad (6.11)$$

All terms in the semi-discrete system $w'(t) = Aw(t) + g(t)$ are now defined. Note that A_1, A_2 and A are symmetrical. Stability of the semi-discrete system in the L_2 -norm is therefore determined by the spectrum of A .

Eigenvalues: Let $B = B_m \in \mathbb{R}^{m \times m}$ be the discretized 1D diffusion operator with Dirichlet conditions. For brevity of notation the subscript m will be dropped in the following. The matrix B has the eigenvalue-eigenvector decomposition

$$B = V \Lambda V^{-1}$$

with

$$V = [\phi_1, \phi_2, \dots, \phi_m], \quad \phi_j = \sqrt{2} (\sin(jh\pi), \sin(2jh\pi), \dots, \sin(mjh\pi))^T$$

and

$$\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m), \quad \lambda_j = -4h^{-2} \sin^2\left(\frac{1}{2}\pi jh\right).$$

²²⁾ For any two matrices $K = (\kappa_{ij}) \in \mathbb{R}^{k_1 \times k_2}$, $L \in \mathbb{R}^{l_1 \times l_2}$, the Kronecker product $K \otimes L$ is the $k_1 l_1 \times k_2 l_2$ block matrix with blocks $\kappa_{ij} L$. If $v \in \mathbb{R}^{k_2}$, $w \in \mathbb{R}^{l_2}$, then $(K \otimes L)(v \otimes w) = Kv \otimes Lw$. In particular, if $\lambda \in \sigma(K)$, $\mu \in \sigma(L)$ with square matrices, then it follows that $\lambda\mu \in \sigma(K \otimes L)$. See, for instance, Horn & Johnson (1991, Sect. 4.2) for further details.

Consider the two-dimensional problem where we assume for convenience of notation that $m_1 = m_2 = m$. Then the eigenvectors for the matrices A_1, A_2 and A in $\mathbb{R}^{M \times M}$, $M = m^2$ are the same; they are given by $\phi_i \otimes \phi_j \in \mathbb{R}^M$ with

$$A_1 \phi_i \otimes \phi_j = \lambda_j \phi_i \otimes \phi_j, \quad A_2 \phi_i \otimes \phi_j = \lambda_i \phi_i \otimes \phi_j,$$

and

$$A \phi_i \otimes \phi_j = (\lambda_i + \lambda_j) \phi_i \otimes \phi_j, \quad i, j = 1, \dots, m.$$

Hence the eigenvalues of A are given by the sum of the eigenvalues of the 1D operators in the x - and y -direction.

Remark 6.1 If we have, instead of the Dirichlet boundary conditions, the periodicity condition

$$u(x \pm 1, y \pm 1, t) = u(x, y, t), \quad (6.12)$$

then a similar derivation can be performed, giving the product Fourier modes $\phi_i \otimes \phi_j$ with complex exponentials as eigenvectors; see the corresponding 1D cases in Section I.3.3 and Section I.4.4. \diamond

Fractional powers: In the next chapter the 2D heat equation will be used occasionally as a model problem to analyze numerical schemes. We present here some technical results involving fractional powers of A_1, A_2 to be used later. We start again with the one-dimensional matrix $B \in \mathbb{R}^{m \times m}$ with $h = 1/(m+1)$. For completeness some derivations from Hundsdorfer & Verwer (1989) on this case are included here.

On \mathbb{R}^m we consider the inner product $\langle v, w \rangle = h v^T w$ and the discrete one-dimensional L_2 -norm $\|v\| = \langle v, v \rangle^{1/2}$. Note that we have orthonormality of the eigenvectors, $\langle \phi_i, \phi_j \rangle = \delta_{ij}$ (equals 1 if $i = j$, and 0 otherwise). Fractional powers of B are defined as $B^\alpha = V A^\alpha V^{-1}$, and thus we have for all $v \in \mathbb{R}^m$

$$B^\alpha v = \sum_{j=1}^m \langle \phi_j, v \rangle \lambda_j^\alpha \phi_j, \quad \|B^\alpha v\|^2 = \sum_{j=1}^m |\langle \phi_j, v \rangle \lambda_j^\alpha|^2.$$

For any two real functions f, g we use the notation $f \sim g$ ($x \downarrow 0$) if there are positive numbers $\gamma_0, \gamma_1, \kappa$ such that $\gamma_0 g(x) \leq f(x) \leq \gamma_1 g(x)$ for $0 < x \leq \kappa$.

Lemma 6.2 Suppose $f \in C(0, 1]$, $\beta \in \mathbb{R}$ and $f(x) \sim x^{-\beta}$ ($x \downarrow 0$). Then, for $h \downarrow 0$,

$$h \sum_{j=1}^m f(jh) \sim \begin{cases} 1 & \text{if } \beta < 1, \\ -\log(h) & \text{if } \beta = 1, \\ h^{1-\beta} & \text{if } \beta > 1. \end{cases}$$

Proof. The statement is trivial for $\beta \leq 0$, so assume that $\beta > 0$. Since we then can split f into a monotonically decreasing part ($\sim x^{-\beta}$ for $x \downarrow 0$) and a bounded remainder, it is clear that we may suppose without loss of generality that f itself is monotonically decreasing. Then

$$h \sum_{j=1}^m f(x_j) \geq \int_h^1 f(x) dx$$

with $x_j = jh$. On the other hand,

$$h \sum_{j=1}^m f(x_j) = hf(h) + h \sum_{j=2}^m f(x_j) \leq hf(h) + \int_h^1 f(x) dx.$$

We have $hf(h) \sim h^{1-\beta}$ ($x \downarrow 0$). The integral $\int_h^1 f(x) dx$ is $\sim -\log(h)$ ($h \downarrow 0$) if $\beta = 1$ and $\sim 1 + h^{1-\beta}$ ($h \downarrow 0$) if $\beta \neq 1$. \square

In the following we denote $e = (1, 1, \dots, 1)^T \in \mathbb{R}^m$ and e_j will stand for the vector in \mathbb{R}^m with j th component equal to 1 and the other components 0.

Lemma 6.3 *We have $\sup_{h>0} \|B^\alpha e\| < \infty$ iff $\alpha < \frac{1}{4}$.*

Proof. Since $Be = -h^{-2}(e_1 + e_m)$, it follows that

$$B^\alpha e = -h^{-2}B^{\alpha-1}(e_1 + e_m) = -h^{-2} \sum_{j=1}^m (h\phi_{j1} + h\phi_{jm})\lambda_j^{\alpha-1}\phi_j$$

with $\phi_{jk} = \sqrt{2}\sin(kjh\pi)$ the k th component of ϕ_j . We have $\phi_{j1} + \phi_{jm} = 0$ if j is even, while $\phi_{j1} + \phi_{jm} = 2\phi_{j1}$ for j odd. In the limit $h \downarrow 0$ we thus obtain

$$\begin{aligned} \|B^\alpha e\|^2 &= h^{-2} \sum_{j=1}^m |(\phi_{j1} + \phi_{jm})\lambda_j^{\alpha-1}|^2 \sim 2h^{-2} \sum_{j=1}^m |\phi_{j1}\lambda_j^{\alpha-1}|^2 \\ &= 4^{2\alpha-1}h^{2-4\alpha} \sum_{j=1}^m |\sin(jh\pi)| (\sin(\frac{1}{2}jh\pi))^{2\alpha-2}|^2 \\ &= 4^{2\alpha}h^{1-4\alpha}h \sum_{j=1}^m \cos^2(\frac{1}{2}jh\pi) (\sin(\frac{1}{2}jh\pi))^{4\alpha-2}. \end{aligned}$$

From Lemma 6.2 we see that $\|B^\alpha e\|^2 \sim h^{1-4\alpha}(1+h^{4\alpha-1})$ for $\alpha \neq \frac{1}{4}$, whereas $\|B^\alpha e\|^2 \sim -\log(h)$ ($h \downarrow 0$) for $\alpha = \frac{1}{4}$. \square

Lemma 6.4 *Let $\chi \in C^2[0, 1]$ and $v = (v_j) \in \mathbb{R}^m$ with $v_j = \chi(x_j)$. Suppose $\alpha < \frac{1}{4}$. Then $\sup_{h>0} \|B^\alpha v\| < \infty$.*

Proof. We have

$$Bv = (\xi_1 - h^{-2}\eta_1, \xi_2, \xi_3, \dots, \xi_{m-1}, \xi_m - h^{-2}\eta_m)^T$$

with $\xi_j = h^{-2}(\chi(x_{j-1}) - 2\chi(x_j) + \chi(x_{j+1})) \approx \chi''(x_j)$ and $\eta_1 = \chi(0)$, $\eta_m = \chi(1)$. Hence, with $\xi = (\xi_1, \xi_2, \dots, \xi_m)^T$, we obtain

$$B^\alpha v = B^{\alpha-1}\xi - h^{-2}B^{\alpha-1}(\eta_1 e_1 + \eta_m e_m).$$

Since B is negative definite with eigenvalues less than -1 , the vector $B^{\alpha-1}\xi$ is bounded uniformly for $h \downarrow 0$ whenever $\alpha < 1$. In the proof of the previous lemma we saw that $h^{-2}B^{\alpha-1}e_j$, with $j = 1$ or m , is also bounded uniformly in h , provided that $\alpha < \frac{1}{4}$. \square

In the application to the 2D heat equation we will work with grid functions on the uniform grid Ω_h with mesh width h in both the x - and y -direction. Grid functions $v = (v_{ij})$ can then be identified with vectors in \mathbb{R}^M , $M = m^2$, as noted above. The discrete L_2 -norm on Ω_h will be denoted by $\|\cdot\|$. The one-dimensional results are then extended by using the product Fourier decomposition,

$$v = \sum_{i,j=1}^m \hat{v}_{ij} \phi_i \otimes \phi_j, \quad \|v\|^2 = \sum_{i,j=1}^m |\hat{v}_{ij}|^2 = h^2 \sum_{i,j=1}^m |v_{ij}|^2.$$

The fractional powers of A_1, A_2 are then given by $A_1^\alpha = I \otimes B^\alpha$, $A_2 = B^\alpha \otimes I$. If $v = (v_{ij}) \in \mathbb{R}^M$ corresponds with a smooth grid function, we get again boundedness of $A_k^\alpha v$ for all $\alpha < \frac{1}{4}$.

Lemma 6.5 *Let $\chi \in C^2([0, 1]^2)$ and $v = (v_{ij}) \in \mathbb{R}^M$ with $v_{ij} = \chi(x_i, y_j)$. Then $\sup_{h>0} \|A_k^\alpha v\| < \infty$ for $\alpha < \frac{1}{4}$, and $\|A_k^{1/4}v\| = \mathcal{O}(|\log(h)|)$, $k = 1, 2$.*

Proof. Consider $k = 1$. Using the row-wise ordering of the grid function $v = (v_{ij})$ as in (6.9), we have

$$A_1^\alpha v = (I \otimes B^\alpha)v = ((B^\alpha v_1)^T, (B^\alpha v_2)^T, \dots, (B^\alpha v_m)^T)^T,$$

$$\|A_1^\alpha v\|^2 = h \sum_{j=1}^m \|B^\alpha v_j\|_{(1D)}^2,$$

with $\|B^\alpha v_j\|_{(1D)}$ the one-dimensional L_2 -norm. The result thus follows from Lemma 6.4. For $k = 2$ we can proceed in the same way by using a column-wise ordering for v . \square

Remark 6.6 It should be noted that if $v \in \mathbb{R}^M$ is the restriction to Ω_h of a smooth function that vanishes at the boundaries, then $\|A_k^\alpha v\| = \mathcal{O}(1)$ for $\alpha < \frac{5}{4}$, due to the fact that then $\tilde{v} = A_k v$ will be a smooth grid function that does not vanish (necessarily) at the boundaries. \diamond

Grid Orientations with the Laplace Operator

Employment of 1D discretization formulas for a multi-dimensional problem in Cartesian coordinates is the usual way to get spatial discretizations. As we saw, based on our 1D knowledge it is easy to apply it to problems with variable coefficients or nonlinearities, and for conservative problems we get schemes in conservation form. However, the systematic use of the underlying 1D discretizations in the coordinate directions sometimes shows up in the numerical results, in the sense that the errors exhibit a *grid orientation*. Working with genuine multi-dimensional discretizations may reduce such effects. In the following we consider discretizations of the Laplace operator on a uniform grid with equal spacing $\Delta x = \Delta y = h$. The common discretization is then given by (6.6) (with $d_1 = d_2 \equiv 1$).

The two-dimensional Laplace operator Δ reads $\partial_{xx} + \partial_{yy}$ in Cartesian coordinates, but it is in fact the divergence of the gradient. Both the divergence and the gradient operator can be viewed in geometrical terms without reference to the underlying coordinate system. In particular, the Laplace operator is invariant under rotations of the coordinate system. The common discretization is not invariant under such rotations; we get a 5-point stencil in which $w_{ij} \approx u(x_i, y_j)$ is coupled only with its direct neighbours $w_{i\pm 1,j}, w_{i,j\pm 1}$ in the coordinate directions. By using a 9-point stencil, giving also coupling in the diagonal directions, a discretization with better invariance properties can be obtained.

In stencil notation, these discretizations of the Laplace operator are

$$A^{[5]} = \frac{1}{h^2} \begin{bmatrix} 1 & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}, \quad A^{[9]} = \frac{1}{6h^2} \begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix}, \quad (6.13)$$

for the 5-point and 9-point stencil, respectively. The truncation errors are found to be

$$\sigma_h^{[5]} = \frac{1}{12}h^2(u_{xxxx} + u_{yyyy}) + \mathcal{O}(h^4)$$

for the 5-point discretization, whereas we get for the 9-point discretization

$$\sigma_h^{[9]} = \frac{1}{12}h^2\Delta^2 u + \mathcal{O}(h^4) = \frac{1}{12}h^2(u_{xxxx} + 2u_{xxyy} + u_{yyyy}) + \mathcal{O}(h^4).$$

Although the amplitude of the truncation errors will in general be of similar size, the leading term of the error with the 9-point stencil is invariant under rotations, whereas for the 5-point stencil it is not. As a result the 5-point stencil may reveal the underlying grid more clearly than with its 9-point counterpart.

Example 6.7 A simple diffusion-reaction model with spiraling solutions has been developed by Barkley (1991). The equations are

$$\begin{aligned} u_t &= \Delta u + \epsilon^{-1}u(1-u)(u - \alpha^{-1}(v + \beta)), \\ v_t &= \delta\Delta v + u - v. \end{aligned} \quad (6.14)$$

The region is taken here as $0 < x, y < 80$ and $t > 0$. The initial value is

$$u(x, y, 0) = \begin{cases} 0 & \text{if } x < 40, \\ 1 & \text{if } x \geq 40, \end{cases} \quad v(x, y, 0) = \begin{cases} 0 & \text{if } y < 40, \\ \frac{1}{2}\alpha & \text{if } y \geq 40. \end{cases}$$

At the boundaries a homogeneous Neumann condition is imposed for both u and v . The parameters are chosen as $\delta = 0$, $\epsilon = 0.002$, $\alpha = 0.25$ and $\beta = 0.001$. The solution develops into a spiral rotating around the center of the domain. Depending on the parameter choice the tip of the spiral will exhibit a meandering behaviour, see the review paper of Barkley (1995). Here we consider the solution at a given time on a fixed grid, to illustrate the grid orientation that may arise with the standard 5-point Laplace discretization.

The numerical solution for the u -component at time $t = 10$ on a uniform 400×400 grid is displayed in Figure 6.1, both with the 5-point and 9-point discretizations (6.13). With the 5-point discretization we see a clear grid orientation: instead of a round spiral the numerical solution assumes a somewhat square form, aligned to the grid.

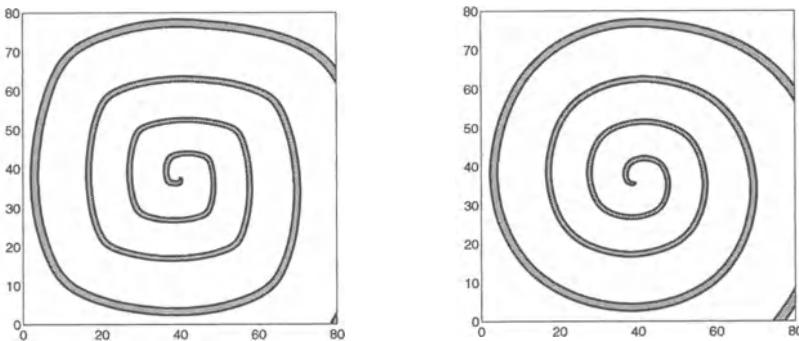


Fig. 6.1. Numerical solution at $t = 10$ of Barkley's model on a 400×400 grid with 5-point (left) and 9-point (right) stencils. Contours of the u -component (with grey = 1, white = 0).

It should be noted that with the present parameter choice the PDE is not very well resolved on this 400×400 grid. For that, finer grids are needed. Then the difference between the results for the 5-point and 9-point stencil becomes smaller; eventually both discretizations converge to the same (exact) solution, of course. \diamond

Remark 6.8 The above system on the 400×400 grid is not stiff, so suited for explicit time stepping. With respect to stability for explicit methods, the 9-point stencil has an additional advantage over the 5-point stencil. The eigenvalues for $A^{[9]}$ in a von Neumann analysis are in the interval $[-\frac{16}{3}h^{-2}, 0]$, compared to the interval $[-8h^{-2}, 0]$ for $A^{[5]}$. Consequently the time step restrictions for explicit methods are more relaxed with the 9-point stencil.

In 3D this difference between the second-order discretizations becomes more pronounced. The 5-point 2D stencil then becomes 7-point with eigenvalue range $[-12h^{-2}, 0]$. The 9-point 2D stencil can be generalized to give a 19-point 3D scheme with eigenvalue range $[-\frac{16}{3}h^{-2}, 0]$, the same as in 2D, see for instance Dowle et al. (1997). \diamond

6.3 Advection on Cartesian Grids

Within the method of lines (MOL) approach the 1D formulas for advection can be used in the various Cartesian directions, with or without limiting. This gives quite simple and effective schemes which are easily augmented with diffusion or reaction terms. On the other hand we then make little use of the fact that the solutions travel along the characteristics. With direct space-time (DST) discretizations this information can be better incorporated. However, these DST schemes are to be constructed anew for multi-dimensional problems; simple insertion of the 1D formulas does not lead to sensible schemes.

Example 6.9 We consider the 2D advection test problem

$$u_t + \underline{a} \cdot \nabla u = 0 \quad (6.15)$$

on the unit square with a given, constant velocity field \underline{a} , taken either as a diagonal flow ($\underline{a} = \underline{a}_D = (1, 1)^T$) or a horizontal flow ($\underline{a} = \underline{a}_H = (1, 0)^T$). The initial profile is a cone or cylinder with height 1, centered at $(0.2, 0.2)$ with radius 0.1, and at the inflow boundaries homogeneous Dirichlet conditions are imposed. The end time is taken as $T = 0.6$.

In Figure 6.2 the MOL solutions are shown for the cylinder on a 100×100 grid; the solutions for the cone are given in Figure 6.3 (together with an unsuccessful DST scheme). The fluxes in the semi-discrete formulation were

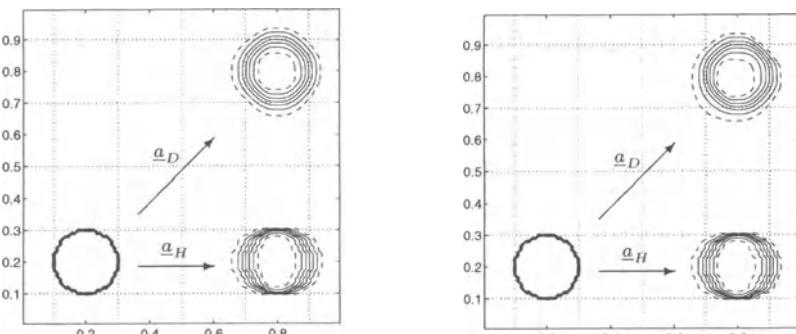


Fig. 6.2. Advection for the cylinder profile on a 100×100 grid. Left picture the semi-discrete solution; right picture for RK2 time stepping with 200 time steps, Courant numbers 0.3. Contour lines at levels $0.1, 0.3, \dots, 0.9$ (solid) and $0.01, 0.99$ (dashed).

calculated from one-dimensional third-order expressions (1.3) with limiting (1.7) to prevent spatial oscillations. The solutions have been obtained with the explicit trapezoidal rule (RK2) as time integration method, using 200 time steps; this gives one-dimensional Courant numbers of 0.3. For reference, also the solutions with very small time steps are displayed, corresponding with the exact solution of the semi-discrete system (6.4).

In the exact semi-discrete solution some grid orientation is visible. This is to be expected since the leading term in the truncation error will be

$$\sigma_h = \frac{1}{12}|a_1|\Delta x^3 u_{xxxx} + \frac{1}{12}|a_2|\Delta y^3 u_{yyyy}$$

for the non-limited scheme, with a_1, a_2 the components of the velocity field \underline{a} . Note from Figure 6.2 that the shape of the cylinder in the diagonal flow is not more deformed than with the horizontal flow. With the Runge-Kutta time stepping a slight additional shape deformation appears. In all, using the 1D discretizations gives good results here. Less dissipation could be obtained by taking higher-order 1D fluxes, or with smaller mesh widths, of course. \diamond

For DST advection schemes the situation is very different: naive use of the one-dimensional formulas will not lead to a sensible scheme. Suppose the one-dimensional scheme for $u_t + f(u)_x = 0$ reads

$$w_i^{n+1} = w_i^n + \frac{\tau}{\Delta x} \left(\bar{f}_{i-\frac{1}{2}}^n - \bar{f}_{i+\frac{1}{2}}^n \right),$$

where the fluxes are denoted with bars to distinguish them from those in the semi-discrete forms like (6.4). Then the naive extension for the 2D equation $u_t + f(u)_x + g(u)_y = 0$ is

$$w_{ij}^{n+1} = w_{ij}^n + \frac{\tau}{\Delta x} \left(\bar{f}_{i-\frac{1}{2},j}^n - \bar{f}_{i+\frac{1}{2},j}^n \right) + \frac{\tau}{\Delta y} \left(\bar{g}_{i,j-\frac{1}{2}}^n - \bar{g}_{i,j+\frac{1}{2}}^n \right). \quad (6.16)$$

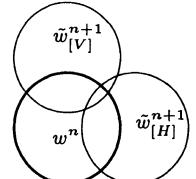
Here the horizontal fluxes $\bar{f}_{i\pm\frac{1}{2},j}$ will depend on values $w_{i\pm k,j}$ and horizontal Courant numbers $\tau a_1/\Delta x$, and similarly for $\bar{g}_{i,j\pm\frac{1}{2}}$ in the vertical direction. The basic problem with this scheme is that during the whole time step from t_n to t_{n+1} there is no diagonal exchange of information. In fact the scheme can be written as

$$w_{ij}^{n+1} = \tilde{w}_{ij[H]}^{n+1} + \tilde{w}_{ij[V]}^{n+1} - w_{ij}^n,$$

where

$$\tilde{w}_{ij[H]}^{n+1} = w_{ij}^n + \frac{\tau}{\Delta x} \left(\bar{f}_{i-\frac{1}{2},j}^n - \bar{f}_{i+\frac{1}{2},j}^n \right)$$

would be the result of one step in a purely horizontal flow, and, likewise, $\tilde{w}_{ij[V]}^{n+1}$ the result of one step in a purely vertical flow; note that for a constant velocity field these quantities approximate shifts in the horizontal and vertical direction. From this interpretation it is obvious that the scheme cannot



perform well. A little inspection also shows that the scheme will be consistent of order one only. Moreover it is clear that the standard 1D limiters will keep the values of $\tilde{w}_{ij[H]}^{n+1}$ and $\tilde{w}_{ij[V]}^{n+1}$ non-negative, rather than w_{ij}^{n+1} .

Example 6.10 Just how bad this naive scheme works is illustrated in Figure 6.3. The 1D fluxes are taken as (2.12) with limiter (2.15), (2.17). The test is as in Example 6.9, with a cone as initial profile and with horizontal and diagonal flow. The number of time steps is 200, giving one-dimensional Courant numbers 0.3. Along with the naive DST scheme also the solution with the MOL approach is given, using as before the explicit trapezoidal rule (RK2) in time.

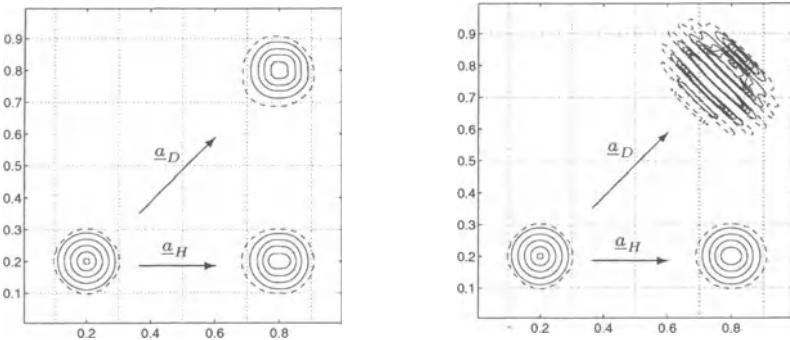


Fig. 6.3. Advection for the cone profile on a 100×100 grid with 200 time steps, Courant numbers 0.3. Contour levels as in Figure 6.2. Left picture RK2 time stepping, right picture with the naive DST implementation (negative values not shown).

With the horizontal flow the DST results are slightly better than the MOL results, due to the fact that the problem is then essentially one-dimensional. However, for the diagonal flow the naive DST scheme gives a result that is no longer recognizable as anything approximating a cone. With this scheme very small time steps would be needed to get reasonable approximations.

Finally we note that positivity for the solution in diagonal flow could be enforced by making the one-dimensional DST limiters more strict (replacing (2.16) by $\nu(1 + \mu) \leq 1/2$, for example). However in the experiment here this did not give an essential improvement of the numerical approximation. ◇

A proper way to use the one-dimensional DST formulas in multi-dimensional problems is by means of *dimensional (time) splitting*; this will be discussed in detail in the next chapter. An alternative is to use genuine two-dimensional DST schemes, which include diagonal transport. A well-known example is the second-order 2-stage MacCormack scheme

$$w_{ij}^* = w_{ij}^n + \frac{\tau}{h} (f(w_{ij}^n) - f(w_{i+1,j}^n)) + \frac{\tau}{h} (g(w_{ij}^n) - g(w_{i,j+1}^n)),$$

$$w_{ij}^{n+1} = \frac{1}{2} (w_{ij}^* + w_{ij}^n) + \frac{\tau}{2h} (f(w_{i-1,j}^*) - f(w_{ij}^*)) + \frac{\tau}{2h} (g(w_{i,j-1}^*) - g(w_{ij}^*)),$$

which is a straightforward extension of the 1D version (2.21); see also Mitchell & Griffiths (1980, p. 185) for related schemes and further references. This scheme reduces to the standard Lax-Wendroff scheme (1.6.7) for linear equations in one dimension. Hence the scheme will show quite strong oscillations if the solution is not very smooth. Schemes with better shape-preserving properties become quite complicated; see for example Rasch (1994) for a conservative scheme, Williamson & Rasch (1989) for positive semi-Lagrangian schemes (for equations in the advective form), and LeVeque (2002, Sect. 20). Moreover, such specialized DST schemes are not easily adapted to include diffusion or reaction terms. For these reasons we will consider in the following multi-dimensional advection only with time splitting or with the MOL approach. Although this MOL approach does not fully use the characteristic information, it is then easy to include additional diffusion or reaction terms.

Remark 6.11 Also for the simple donor-cell scheme, where first-order upwind in space is combined with explicit Euler in time, it holds that there is no diagonal transport during one time step. However, with this scheme the effect is overshadowed by the diffusive character of the scheme. ◇

Stability

Stability of the MOL schemes is usually studied by means of a von Neumann analysis, after linearization and freezing of coefficients. The situation is similar as for diffusion, except that instead of boundary conditions one should only consider spatial periodicity to avoid non-normal matrices.

In the two-dimensional case this means essentially that we consider the constant-coefficient equation $u_t + a_1 u_x + a_2 u_y = 0$ with periodicity (6.12). After spatial discretization this gives a semi-discrete system $w'(t) = Aw(t)$, $A = A_1 + A_2$, and L_2 -stability of the discretization holds if the eigenvalues $\lambda = \lambda_1 + \lambda_2$, with $\lambda_1 \in \sigma(A_1)$, $\lambda_2 \in \sigma(A_2)$, are in the left half complex plane. Using a time integration method with stability region \mathcal{S} then yields the stability requirement

$$\tau\lambda = \tau\lambda_1 + \tau\lambda_2 \in \mathcal{S}.$$

Again this is not fundamentally different from the 1D case. If the one-dimensional stability requirement reads $\tau|a_1|/\Delta x \leq C$, with $C > 0$ determined by the spatial discretization and the time integration method, then in 2D this will usually give the requirement²³⁾

$$\frac{\tau|a_1|}{\Delta x} + \frac{\tau|a_2|}{\Delta y} \leq C. \quad (6.17)$$

²³⁾ Usually, the eigenvalues λ_k ($k = 1, 2$) will be located on the boundary of $\nu_k \mathcal{G}$ with \mathcal{G} a convex set in \mathbb{C} , and then $\lambda \in (\nu_1 + \nu_2)\mathcal{G}$.

Hence the one-dimensional absolute Courant numbers should be added to get the corresponding stability restriction in 2D. With second-order central differences, C will be equal to the imaginary stability boundary β_I , which is the maximal number such that the portion $[iC, -iC]$ of the imaginary axis fits in the stability region.

The same considerations hold in 3D and also with inclusion of reaction and/or diffusion terms. Although this does not lead to conceptual difficulties, determination of precise stability bounds then quickly become very technical and complicated. Sufficient conditions for several interesting schemes can be found in Wesseling (2001, Sect. 5.8).

Remark 6.12 With the von Neumann stability analysis, we are in fact studying the evolution of Fourier modes

$$w_{jk}(t) = e^{\lambda t} e^{2i(\omega_1 j + \omega_2 k)}, \quad \frac{1}{2}\pi \leq \omega_1, \omega_2 \leq \frac{1}{2}\pi, \quad i = \sqrt{-1},$$

corresponding to wave numbers $\omega_1/\pi\Delta x$ and $\omega_2/\pi\Delta y$, which will give upon insertion into the semi-discrete system a dispersion relation of the form $\lambda = \lambda(\omega_1, \omega_2)$. For stability of the spatial discretization we require that $\operatorname{Re} \lambda \leq 0$ for all ω_1, ω_2 . Likewise, for fully discrete schemes we make the ansatz

$$w_{jk}^n = r^n e^{2i(\omega_1 j + \omega_2 k)}, \quad -\frac{1}{2}\pi \leq \omega_1, \omega_2 \leq \frac{1}{2}\pi, \quad i = \sqrt{-1},$$

with stability requirement $|r| \leq 1$ for all ω_1, ω_2 . In the MOL approach we will have $r = R(\tau\lambda)$ if R is the stability function of the ODE method. \diamond

Monotonicity

Spatial discretizations with one-dimensional limiters are also common for multi-dimensional advection problems. The TVD property is not that useful anymore as a condition on numerical schemes, since this is restricted to first-order schemes for multi-dimensional problems, see LeVeque (1992, Sect. 18.2). However, positivity can still be studied in a similar way as in 1D.

Consider the model problem $u_t + a_1 u_x + a_2 u_y = 0$ with constant $a_1, a_2 \geq 0$. Suppose in both spatial directions we use the 1D flux limiters (1.3), (1.5) with $\mu = 1$. Then the semi-discrete system can be written as

$$w'_{ij}(t) = \alpha_{ij}(w(t))(w_{i-1,j}(t) - w_{ij}(t)) + \beta_{ij}(w(t))(w_{i,j-1}(t) - w_{ij}(t))$$

with nonlinear functions α_{ij}, β_{ij} satisfying

$$0 \leq \alpha_{ij}(w) \leq \frac{2a_1}{\Delta x}, \quad 0 \leq \beta_{ij}(w) \leq \frac{2a_2}{\Delta y},$$

see formulas (1.19), (1.20). For a_1 or $a_2 < 0$, similar forms hold. Hence with forward Euler time stepping positivity is guaranteed provided that

$$\frac{\tau|a_1|}{\Delta x} + \frac{\tau|a_2|}{\Delta y} \leq \frac{1}{2}.$$

As in 1D, the forward Euler method itself is not a recommended time integration method, but positivity for higher-order Runge-Kutta or multistep methods now follows for those methods that can be written as a convex combination of forward Euler steps, possibly with some additional backward Euler steps for implicit methods. This will then lead to a restriction of the form

$$\frac{\tau|a_1|}{\Delta x} + \frac{\tau|a_2|}{\Delta y} \leq \frac{1}{2}K \quad (6.18)$$

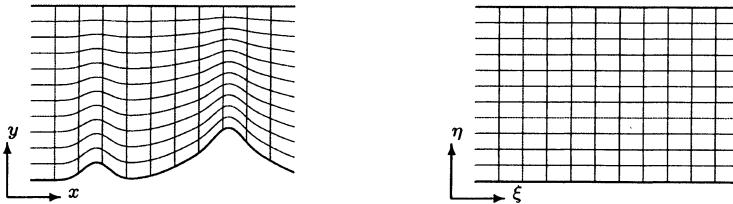
with K determined by the specific time integration method, as in Section II.4; for instance, with the explicit trapezoidal rule used in Example 6.9 we have $K = 1$. For the model problem with constant coefficients the positivity property will also imply the more general maximum principle, by which we then know that there will be no global undershoots or overshoots.

6.4 Transformed Cartesian Grids

To approximate PDEs on a non-Cartesian domain $\Omega \subset \mathbb{R}^d$ it is often possible to transform this *physical domain* into a rectangular *computational domain* $\mathcal{C} \subset \mathbb{R}^d$ on which the discretization takes place. Suppose we have a coordinate transformation that gives the physical coordinate $\underline{x} \in \Omega$ as a function of the computational coordinate $\underline{\xi} \in \mathcal{C}$,

$$\underline{x} = \underline{x}(\underline{\xi}), \quad \underline{x} \in \Omega, \quad \underline{\xi} \in \mathcal{C}.$$

If we impose a rectangular grid \mathcal{C}_h on the computational domain \mathcal{C} , the transformation results in a curvilinear physical grid Ω_h formed by curvilinear coordinate lines, where the structure of the grid is still logically rectangular.



By discretizing on \mathcal{C}_h , one still works with a rectangular Cartesian grid, on which known and well understood standard discretization formulas can be applied. The price one pays is that the PDE to be discretized on \mathcal{C}_h is a transformed version, which is usually more complicated than the original PDE in the Cartesian coordinates. For instance, due to the transformation metric coefficients are introduced. Moreover, cross derivatives will appear if the transformation is non-orthogonal. A detailed account of transformed grids with many examples and formulas for finite difference and finite volume methods is found in Thompson, Warsi & Mastin (1985). A more recent proceedings in this field is Thompson, Soni & Weatherill (1999).

Transformed grids can for instance be used to resolve boundary layers using locally refined grids. Transformed grids are often used in situations where the boundary of Ω is curvilinear and an accurate discretization of boundary conditions is desirable; the transformation is then chosen such that the grid is *boundary-fitted*. It is also possible that Ω_h has curvilinear grid lines that vary in time. This occurs, for example, in global air quality models where Ω is a shell around the earth. The grid Ω_h is then based on polar coordinates. The lower boundary will be fixed, curvilinear, following the orography (mountains). The upper boundary is determined by a zero pressure assumption. The grid cells in between depend on time-dependent air pressure conditions and thus their physical position may vary in time; see for instance Blom & Roemer (1997).

Remark 6.13 For complicated domains Ω the design of a curvilinear, boundary-fitted grid is difficult, certainly in 3D. In practice one often uses *grid generators*, which generate numerically a discrete coordinate mapping

$$\underline{x}_i = \underline{x}(\underline{\xi}_i), \quad \underline{x}_i \in \Omega_h, \quad \underline{\xi}_i \in \mathcal{C}_h.$$

This can be extended to a continuous transformation $\underline{x} = \underline{x}(\underline{\xi})$ by interpolation. A practical technique for this is piecewise bilinear interpolation in 2D. Bilinear interpolation maps quadrilaterals in Ω_h to rectangles in \mathcal{C}_h . Likewise, in 3D, we may use trilinear interpolation, which maps hexagons with straight edges in Ω to rectangular hexagons in \mathcal{C} . A detailed description of the use of such piecewise linear interpolation transformations in discretization formulas is found in Wesseling (2001, Chap. 11). \diamond

Transformed Advection-Diffusion-Reaction Equations

For the transformation of the advection-diffusion equation (6.1), from the Cartesian coordinates $\underline{x} = (x_1, \dots, x_d)^T$ to general curvilinear coordinates $\underline{\xi} = (\xi^1, \dots, \xi^d)^T$, it is convenient to use general forms based on the so-called covariant and contravariant basis vectors for the transformation $\underline{x} = \underline{x}(\underline{\xi}, t)$. Such forms have been developed in tensor analysis; here we merely present the transformed formula for equation (6.1).²⁴⁾ For this purpose we first write the equation in the form

$$u_t + \sum_{k=1}^d (a_k u)_{x_k} = \sum_{k=1}^d (\underline{e}_k \cdot D \sum_{j=1}^d (\underline{e}_j u_{x_j}))_{x_k} + s(\underline{x}, t, u), \quad (6.19)$$

with \underline{e}_k the k th Cartesian unit basis vector. Applying transformation relations from tensor analysis we then obtain a transformed equation with $v(\underline{\xi}, t) =$

²⁴⁾ We here follow Blom & Roemer (1997) where a summary compiled from Thompson et al. (1985) is given of the required transformation formulas in \mathbb{R}^3 for the differential equations and boundary conditions.

$u(\underline{x}, t)$, $\underline{x} = \underline{x}(\xi, t)$. In the curvilinear coordinate system (ξ, t) the advection-diffusion equation reads

$$\begin{aligned} & (\sqrt{g} v)_t + \sum_{k=1}^d (\sqrt{g} \alpha^k v)_{\xi^k} \\ &= \sum_{k=1}^d (\underline{c}^k \cdot D \sum_{j=1}^d (\sqrt{g} \underline{c}^j v)_{\xi^j})_{\xi^k} + \sqrt{g} s(\underline{x}(\xi, t), t, v). \end{aligned} \quad (6.20)$$

Here $\underline{c}^k = \nabla \xi^k$ is the k th contravariant basis vector and $\alpha^k = \underline{c}^k \cdot \underline{a}$ is the k th contravariant component of the velocity \underline{a} . Further $g = 1/\det|g^{kj}|$ with contravariant metric tensor $(g^{kj}) = (\underline{c}^k \cdot \underline{c}^j)$. One immediately sees the correspondence between (6.19) and (6.20); note in particular that in the transformed equation the spatially conserved quantity is now $\sqrt{g} v$. For actual applications the diffusion matrix D should be represented by its contravariant components.

Equation (6.20) can often be simplified. For instance, if the transformation is time-independent and we have a scalar diffusion coefficient D , then the transformed equation can be written as

$$v_t + \frac{1}{\sqrt{g}} \sum_{k=1}^d (\sqrt{g} \alpha^k v)_{\xi^k} = \frac{1}{\sqrt{g}} \sum_{k,j=1}^d (\sqrt{g} g^{kj} D v_{\xi^j})_{\xi^k} + s(\underline{x}(\xi, t), t, v),$$

from which we see the appearance of cross-derivatives for non-orthogonal transformations. This equation can now be discretized on C_h using standard finite difference or finite volume formulas. Even if the metric coefficients are given in analytical form, one may choose to use discrete representations if this would be more convenient. In setting up a discretization, one should however obey certain metric identities which show up for constant u . In this sense the numerical treatment of curvilinear coordinates is more involved. For details on this and other technicalities we refer to Thompson et al. (1985) and Wesseling (2001).

Example 6.14 As a simple illustration, consider the advection-diffusion-reaction equation

$$u_t + \nabla \cdot (\underline{a} u) = D \Delta u + s(u)$$

with constant D on a cylindrical region

$$\Omega = \{\underline{x} = (x, y, z) : x^2 + y^2 < 1, 0 < z < 1\} \subset \mathbb{R}^3.$$

Then it is appropriate, of course, to use cylindrical coordinates $\xi = (r, \theta, z)$ where $r^2 = x^2 + y^2$ and $\theta = \arctan(y/x)$. If we know in addition that the solution and coefficients are *radial symmetric*, that is, they do not depend on the horizontal angle θ , then we obtain the two-dimensional equation

$$v_t + \frac{1}{r} (r a^{(r)} v)_r + (a^{(z)} v)_z = \frac{D}{r} (r v_r)_r + D v_{zz} + s(v), \quad (6.21)$$

where $0 < r, z < 1$ and $a^{(r)}, a^{(z)}$ are the radial and vertical velocity components. Discretization with second-order central differences in conservative form on a uniform grid with grid points $r_i = (i - \frac{1}{2})\Delta r$, $i = 1, \dots, m_1$, $\Delta r = 1/m_1$ and $z_j = (j - \frac{1}{2})\Delta z$, $j = 1, \dots, m_2$, $\Delta z = 1/m_2$, will then give the semi-discrete system

$$\begin{aligned} v'_{ij} = & \frac{1}{2r_i\Delta r} \left(r_{i-\frac{1}{2}} a_{i-\frac{1}{2},j}^{(r)} (v_{i-1,j} + v_{ij}) - r_{i+\frac{1}{2}} a_{i+\frac{1}{2},j}^{(r)} (v_{ij} + v_{i+1,j}) \right) \\ & + \frac{1}{2\Delta z} \left(a_{i,j-\frac{1}{2}}^{(z)} (v_{i,j-1} + v_{ij}) - a_{i,j+\frac{1}{2}}^{(z)} (v_{ij} + v_{i,j+1}) \right) \\ & + \frac{D}{r_i \Delta r^2} \left(r_{i-\frac{1}{2}} (v_{i-1,j} - v_{ij}) - r_{i+\frac{1}{2}} (v_{ij} - v_{i+1,j}) \right) \\ & + \frac{D}{\Delta z^2} \left(v_{i,j-1} - 2v_{ij} + v_{i,j+1} \right) + s(v_{ij}). \end{aligned}$$

Higher-order discretization is obtained by replacement of the numerical flux values, such as $v_{i\pm 1/2,j} = \frac{1}{2}(v_{ij} + v_{i\pm 1,j})$, by the common higher-order 1D counterparts. Due to the radial symmetry we may take $v_{0j} = v_{1j}$ in the equations for $i = 1$. The other boundary conditions are of the usual form. \diamond

6.5 Unstructured Grids

In modern computational practice, multi-dimensional problems on complicated domains are more and more common, especially in industrial applications. Transformation to a simple computational domain is not always possible, or it may be too complex. For that reason unstructured grids are becoming increasingly popular. In this section we will briefly describe some simple schemes for two-dimensional problems on triangular grids. For theoretical results and higher-order schemes we will refer to more specialized literature.

We assume here that the physical region $\Omega \subset \mathbb{R}^2$ is covered by a set of non-overlapping triangles T_j such that all edges are fully shared by neighbouring triangles. On this grid we consider some basic finite element and finite volume methods for the linear advection-diffusion equation (6.1). The triangulation can be unstructured, in the sense that the number of edges meeting in a vertex may vary from vertex to vertex. Although this complicates programming, it greatly enhances the local flexibility.

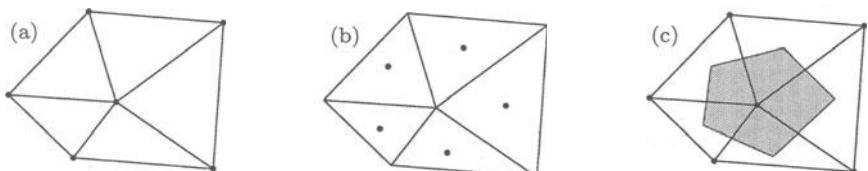


Fig. 6.4. Schematic illustration of the use of triangulations for (a) finite elements, (b) finite volumes on the principal grid and (c) finite volumes on a dual grid.

Finite Elements

Consider equation (6.1) for $t > 0$ with given initial function $u_0(\underline{x})$ and boundary conditions

$$u = \gamma_0(\underline{x}) \quad \text{on } \Gamma_0, \quad \underline{n} \cdot (\underline{a}u - D\nabla u) = \gamma_1(\underline{x}) \quad \text{on } \Gamma_1, \quad (6.22)$$

where $\Gamma_0 \cup \Gamma_1 = \partial\Omega$ is the boundary of Ω and \underline{n} is the outward normal vector on this boundary. First we derive the *weak form* of the equation, where u is required to be once differentiable only. This will allow us to take a piecewise linear approximation w^h for u .

Let for scalar functions v, w or vector valued functions $\underline{f}, \underline{g}$ the L_2 -inner product over Ω be denoted by

$$(w, v) = \int_{\Omega} w v \, d\Omega, \quad (\underline{g}, \underline{f}) = \int_{\Omega} \underline{g} \cdot \underline{f} \, d\Omega.$$

Using integration by parts, it follows that

$$(\nabla \cdot \underline{f}, v) = -(\underline{f}, \nabla v) + \int_{\partial\Omega} (\underline{f} \cdot \underline{n}) v \, d\Gamma.$$

We denote by \mathcal{H} the space of continuous functions on Ω such that ∇u lies in $L_2(\Omega)$, and consider

$$\mathcal{V} = \{v \in \mathcal{H} : v = \gamma_0 \text{ on } \Gamma_0\}, \quad \mathcal{V}_0 = \{v \in \mathcal{H} : v = 0 \text{ on } \Gamma_0\}.$$

Further we consider the bilinear form

$$[w, v] = (D\nabla w - \underline{a}w, \nabla v)$$

and the functional

$$G(v) = (s, v) - \int_{\Gamma_1} \gamma_1 v \, d\Gamma.$$

Then, multiplying (6.1) by a test function $v \in \mathcal{V}_0$ and integrating by parts gives us the following weak formulation for (6.1), (6.22): find $u(\cdot, t) \in \mathcal{V}$ such that

$$(u_t, v) + [u, v] = G(v) \quad \text{for all } v \in \mathcal{V}_0, t > 0. \quad (6.23)$$

Having a triangulation of Ω , a Galerkin finite element method is determined by choice of the basis functions. Let \underline{x}_j be the vertices of the triangulation. The simplest basis is formed by the piecewise linear (pyramid) functions ϕ_i that are linear on each triangle and such that

$$\phi_i(\underline{x}_j) = \delta_{ij}.$$

Suppose the vertices are numbered with index sets $\mathcal{J}_0, \mathcal{J}$ such that $j \in \mathcal{J}_0$ if $\underline{x}_j \in \Gamma_0$ and $j \in \mathcal{J}$ if $\underline{x}_j \in \Omega \cup \Gamma_1$. Formulation of the Galerkin finite element method is now similar as in 1D. We set

$$w^h(\underline{x}, t) = \sum_{j \in \mathcal{J} \cup \mathcal{J}_0} w_j(t) \phi_j(\underline{x}) \quad (6.24)$$

with

$$w_j(t) = \gamma_0(\underline{x}_j) \quad \text{if } j \in \mathcal{J}_0, \quad (6.25)$$

thus ensuring that $w^h \in \mathcal{V}$. The other weights are determined by the semi-discrete ODE system

$$\sum_{j \in \mathcal{J} \cup \mathcal{J}_0} (\phi_j, \phi_i) w'_j(t) = - \sum_{j \in \mathcal{J} \cup \mathcal{J}_0} [\phi_j, \phi_i] w_j(t) + G(\phi_i), \quad i \in \mathcal{J}, \quad (6.26)$$

with initial conditions $w_j(0) = u_0(\underline{x}_j)$, $j \in \mathcal{J}$. This formulation is also applicable for time-dependent boundary functions γ_0, γ_1 .

Thus we see that the basic finite element method is very easily generalized to two dimensions. Discretizations with higher order are obtained if we choose basis functions from higher-order piecewise polynomials. For convergence results and implementation issues we refer to Wait & Mitchell (1985) and Morton (1996).

Because of this ease to define the methods on arbitrary triangular grids, finite elements are increasingly important in numerical simulations for real-life problems. Upwinding can be achieved by tilting the test functions in the flow direction, with a Petrov-Galerkin formulation, leading for example to streamline-diffusion methods; see Morton (1996) and Roos, Stynes & Tobiska (1996) for error bounds in the stationary case. However, for time-dependent problems it is not clear to what extent oscillations will be avoided; see the discussion in Section 5.3. Moreover, with the finite element method there is no local mass conservation, in the sense that the method is not based on a mass balance for local control volumes.

Remark 6.15 The bilinear form $[\cdot, \cdot]$ used here is slightly different from the 1D version that was used in Section 5. This is due to the conservative form of the advection term in (6.1), for which we also applied integration by parts in the above. With an advective form $\underline{a} \cdot \nabla u$, integration by parts will not be applied to that term, and hence the bilinear form and source functional G are changed accordingly. Also other boundary conditions will lead to a modification of G . \diamond

Remark 6.16 The semi-discrete system (6.26) can be written in the familiar form $Bw'(t) = Aw(t) + g(t)$ with *mass matrix* B and *stiffness matrix* A . Time integration aspects are not fundamentally different from the 1D considerations in earlier sections. Although the structure of the matrices A and B will not be as regular as for finite difference schemes on Cartesian grids, these matrices are still *sparse* since the entries a_{ij}, b_{ij} will be zero if the support of ϕ_i and ϕ_j does not overlap.

The use of *mass lumping*, that is, replacement of B on the left-hand side of (6.26) by a diagonal matrix with the same row-sums, will lead to fully explicit schemes if we apply an explicit Runge-Kutta or multistep method. However, we already saw for the 1D case – see Remark 5.1 and the numerical illustration in Section 3.2 – that mass lumping may have an adverse effect on the accuracy of the scheme. \diamond

Finite Volumes (on Primary Grid)

With finite volume methods the discretization is derived from local mass balances on control volumes $\mathcal{C} \subset \Omega$. For the advection-diffusion problem (6.1) we have, according to the Gauss divergence theorem,

$$\int_{\mathcal{C}} u_t d\Omega = - \int_{\partial\mathcal{C}} \underline{n} \cdot (\underline{a}u - D\nabla u) d\Gamma + \int_{\mathcal{C}} s d\Omega. \quad (6.27)$$

Here \underline{n} is the outward normal vector on the boundary $\partial\mathcal{C}$. For this integral form of the equation, not even continuity in space of u is required, so we can approximate u by piecewise constant functions.

Given a triangulation of the domain Ω there are two basic choices for the control volumes. First we consider the case where the triangles themselves are the control volumes, schematically illustrated in Figure 6.4(b). This leads to a generalization of the cell centered schemes that were considered in Section 4.2 for the one-dimensional case.

Inclusion of diffusion terms is not straightforward in this approach. Therefore we first consider only

$$u_t + \nabla \cdot \underline{f}(u) = s.$$

Here $\underline{f}(u) = \underline{a}u$ for the linear advection problem, but more general nonlinear forms with $\underline{f}(u) = (f(u), g(u))^T$ can also be considered. On the boundary of Ω we assume inflow Dirichlet conditions. Consider a triangle T_i and let $|T_i|$ denote its area. Further let $\mathcal{E}_{i(k)}$, $k = 1, 2, 3$, stand for the edges, with length $|\mathcal{E}_{i(k)}|$ and corresponding neighbouring triangles $T_{i(k)}$. The mass balance (6.27) then leads to

$$\frac{d}{dt} \int_{T_i} u d\Omega = - \sum_{k=1,2,3} \int_{\mathcal{E}_{i(k)}} \underline{n} \cdot \underline{f}(u) d\Gamma + \int_{T_i} s d\Omega.$$

Let the average values of u and s over T_i at time t be denoted by $w_i(t)$, $s_i(t)$. To turn this local mass balance into a numerical scheme, we approximate the integral over $\mathcal{E}_{i(k)}$ by midpoint quadrature (1-point Gauss) and we replace $\underline{n} \cdot \underline{f}(u)$ by a numerical flux function $g_{i(k)}(w_i, w_{i(k)})$, defined in terms of the neighbouring values w_i and $w_{i(k)}$ by upwinding (and flux splitting for nonlinear problems, see Remark 1.6). This then gives the semi-discrete finite volume scheme

$$w'_i(t) = -\frac{1}{|T_i|} \sum_{k=1,2,3} |\mathcal{E}_{i(k)}| g_{i(k)}(w_i(t), w_{i(k)}(t)) + s_i(t). \quad (6.28)$$

If $\mathcal{E}_{i(k)}$ is part of the inflow boundary, the prescribed inflow condition will be used for the numerical flux. It is obvious that the restriction to triangles is not relevant here; we could just as well allow arbitrary polygons T_i . Although

the scheme (6.28) is basically just a generalization of the first-order upwind scheme, it is quite popular for nonlinear hyperbolic problems.

However, generalization to second order or higher, and inclusion of diffusion terms is not straightforward. The common technique for that is by means of *reconstructions* where the piecewise constant approximation is extended after each time step to a piecewise polynomial, from which diffusive fluxes and higher-order advective fluxes can be calculated. For results in this direction we refer to Kröner (1997) and Sonar (2002).

Finite Volumes (on Dual Grids)

Starting from a given triangulation of Ω we can also construct a dual grid on which the vertices \underline{x}_i of the triangles become centers of control volumes \mathcal{B}_i , see Figure 6.4(c) for a schematic illustration. A popular choice is to use so-called Voronoi boxes

$$\mathcal{B}_i = \{\underline{x} \in \Omega : |\underline{x} - \underline{x}_i| < |\underline{x} - \underline{x}_j| \text{ for all } j \neq i\},$$

with $|\cdot|$ being the Euclidean distance in R^2 . These are polygons that intersect the edges of the triangles half-way at straight angles. Using these polygons \mathcal{B}_i as control volumes leads to discretizations akin to the vertex centered 1D schemes discussed in Section 4.1.

Suppose for each vertex \underline{x}_i the neighbouring vertices are \underline{x}_j , $j \in \mathcal{J}(i)$. Let \mathcal{E}_j , $j \in \mathcal{J}(i)$ stand for the corresponding edges of the box \mathcal{B}_i with outward normal vectors $\underline{n}_j = (\underline{x}_j - \underline{x}_i)/\Delta_{ij}$, $\Delta_{ij} = |\underline{x}_i - \underline{x}_j|$. Consider the advection-diffusion equation (6.1) with scalar diffusion coefficient D . Applying (6.27) on the control volume \mathcal{B}_i then gives

$$\int_{\mathcal{B}_i} u_t d\Omega = - \sum_{j \in \mathcal{J}(i)} \int_{\mathcal{E}_j} (\underline{n}_j \cdot \underline{a} u - D \underline{n}_j \cdot \nabla u) d\Gamma + \int_{\mathcal{B}_i} s d\Omega.$$

Next we replace the integral over \mathcal{E}_j by 1-point quadrature in $\xi_{ij} = \frac{1}{2}(\underline{x}_i + \underline{x}_j)$. The average value w_i of u over \mathcal{B}_i can also be associated with the point value at \underline{x}_i . Hence for the fluxes on \mathcal{E}_j we can insert one-dimensional expressions. Using an upwind parameter ϑ_j we arrive at the semi-discrete scheme

$$w'_i = -\frac{1}{|\mathcal{B}_i|} \sum_{j \in \mathcal{J}(i)} |\mathcal{E}_j| \left(\alpha_j (\vartheta_j w_j + (1 - \vartheta_j) w_i) - D \frac{1}{\Delta_{ij}} (w_j - w_i) \right) + s_i. \quad (6.29)$$

Here $\alpha_j = \underline{n}_j \cdot \underline{a}$ and D are evaluated in the point ξ_{ij} and s_i is the average value of s over \mathcal{B}_i at time t . The upwind parameter ϑ_j will depend on α_j ; with pure upwinding we will take $\vartheta_j = 1$ if $\alpha_j > 0$ and $\vartheta_j = 0$ if $\alpha_j < 0$, whereas a central scheme is obtained with $\vartheta_j = \frac{1}{2}$. A popular choice is also exponential fitting, where

$$\vartheta_j = \frac{1}{\mu_j} - \frac{1}{e^{\mu_j} - 1}, \quad \mu_j = \frac{\alpha_j \Delta_{ij}}{D} \quad \text{for } j \in \mathcal{J}(i).$$

This gives a generalization of the Allen-Southwell-II'in scheme considered in Example 3.4 (with $\kappa = 1 - 2\vartheta$) for one-dimensional model problems.

For points \underline{x}_i on the boundary of Ω the control volume \mathcal{B}_i will contain two boundary edges of the primary triangulation. If boundary fluxes are given, this control volume is used for the mass balance. For Dirichlet conditions the values for w_i are simply inserted. Implementation of (6.22) is therefore straightforward.

Theoretical error bounds and numerical comparisons for steady-state problems with this exponentially fitted scheme can be found in Kröner (1997, Sect. 7.1). Instead of the Voronoi boxes, the control volumes around the vertices \underline{x}_i can also be chosen in so-called barycentric fashion; such schemes are discussed in Sonar (2002) for nonlinear problems.

7 Notes on Moving Grids and Grid Refinement

Non-uniform grids are useful if there are strong local variations in the solution that are hard to capture on a uniform grid. If the locations of strong local variations are fixed in time and known a priori, as for example with boundary layers or localized source terms, a temporally fixed, non-uniform grid can be selected. In general, however, regions of strong local variations need not be known in advance and also can move in time. In such cases one needs spatial grid adaptivity in time, in a manner that the refined parts of the grid capture emerging and moving regions of strong local variations.

Spatial grid adaptivity in time has been explored extensively, both for one- and higher-space dimensional problems. An obvious possibility is to use a Lagrangian grid such that grid points are moved continuously in time. This approach is often called *dynamic regridding*, in contrast to *static regridding* where grids are adjusted only at discrete time levels. In the following we will briefly describe some basic ideas behind these two approaches.

7.1 Dynamic Regridding

For describing the basic ideas behind dynamic regridding we will focus on the 1D case. Consider the PDE system

$$u_t = \mathcal{F}(x, t, u), \quad 0 < x < 1, \quad t > 0, \quad (7.1)$$

where $u = u(x, t)$ and \mathcal{F} is a spatial operator. The actual definition of \mathcal{F} with boundary conditions at $x = 0, 1$ and initial condition at $t = 0$ need not be specified for the discussion to follow.

As usual we adopt the method of lines approach, where on a chosen space grid the PDE problem is first converted into an ODE system by spatial discretization which then needs to be integrated in time by a suitable integration

method. Different from the standard MOL approach based on an Eulerian grid, it is here supposed that we have a moving grid, say composed of m time-dependent grid points $x_j(t)$,

$$0 = x_0 < x_1(t) < \cdots < x_j(t) < \cdots < x_m(t) < x_{m+1} = 1. \quad (7.2)$$

According to the Lagrangian grid approach, we then introduce the total derivative $u' = x' u_x + u_t$ along each trajectory $x(t) = x_j(t)$ and spatially approximate on the grid (7.2) the Lagrangian PDE system

$$u' = x' u_x + \mathcal{F}(x, t, u).$$

Using for example the standard second-order differences for u_x , this leads to

$$w'_j(t) = x'_j(t) \frac{w_{j+1}(t) - w_{j-1}(t)}{x_{j+1}(t) - x_{j-1}(t)} + F_j(t, w(t)), \quad 1 \leq j \leq m, \quad (7.3)$$

where $w_j(t)$ and $F_j(t, w(t))$ represent, respectively, the semi-discrete approximations to u and $\mathcal{F}(u, x, t)$ at $(x_j(t), t)$. The semi-discretization can be based on various finite difference, finite volume or finite element methods. System (7.3) is a semi-discrete system in a moving frame of reference. Required boundary conditions for (7.1) can be dealt with in the same way as with an Eulerian grid, by which the values $w_0(t)$ and $w_{m+1}(t)$ at the left and right boundary point are defined.

The issue here is how to define the grid trajectories $x_j(t)$, $t \geq 0$. Obviously, the first aim is to move the nodes such that in regions of high spatial activity there is enough spatial resolution. Further it would be advantageous to have also a smooth solution along the trajectories. For example, if we have a simple, steep running wave form $u(x, t) = \phi(x - at)$ as solution, the nodes should be moved with velocity a . In the ideal situation we then have a constant solution in the moving frame of reference.

An often used technique is *equidistribution* of a *monitor* function which is taken dependent on first or second spatial derivatives. Let $M(x, t, u) > 0$ denote a monitor function. The time-dependent grid (7.2) is then defined by the equidistribution relation

$$\int_{x_{j-1}(t)}^{x_j(t)} M(s, t, u) ds = \int_{x_j(t)}^{x_{j+1}(t)} M(s, t, u) ds, \quad 1 \leq j \leq m. \quad (7.4)$$

By this relation, the implicitly defined $x(t)$ -grid equidistributes integrals of monitor values over the spatial interval $[0, 1]$. This grid is thus determined by the solution profile, the chosen monitor, and the number of moving nodes $x_j(t)$. Trivially, where the monitor gets large the grid is refined and vice versa. Using the midpoint rule for approximating the integrals, we can discretize the equidistribution relation (7.4) to obtain, for $1 \leq j \leq m$,

$$(x_{j+1}(t) - x_j(t)) M_{j+\frac{1}{2}}(t) = (x_j(t) - x_{j-1}(t)) M_{j-\frac{1}{2}}(t), \quad (7.5)$$

where $M_{j+1/2}(t)$ approximates $M(x, t, u)$ at $x_{j+1/2} = \frac{1}{2}(x_j + x_{j+1})$.

For scalar problems a popular monitor function is

$$M(x, t, u) = (\alpha + (u_x)^2)^{1/2},$$

where the parameter $\alpha > 0$ serves to ensure strict positivity. Normally $\alpha = 1$, yielding the *arc-length* monitor and the nodes are then placed along uniform arc-length intervals. Again employing second-order central differences for u_x would give the discrete monitor values

$$M_{j+\frac{1}{2}}(t) = \left(\alpha + \left(\frac{w_{j+1}(t) - w_j(t)}{x_{j+1}(t) - x_j(t)} \right)^2 \right)^{1/2}. \quad (7.6)$$

For PDE systems with s components u_k , the arclength monitor could for example be taken as

$$M(x, t, u) = \left(\alpha + \sum_{k=1}^s \beta_k \left(\frac{\partial u_k(x, t)}{\partial x} \right)^2 \right)^{1/2} \quad (7.7)$$

with weights $\beta_k > 0$ for the different components.

We thus have obtained a system of *differential algebraic equations* (DAEs) with $2m$ unknowns, coupling the m differential equations (7.3) for the (scalar) unknowns $w_j(t)$ with the m algebraic equations (7.5), (7.6) for the unknowns $x_j(t)$. For integrating this DAE system in time, a suitable DAE solver should be used; some available codes have been listed in Section II.5.4.

However, as it stands, the moving-grid equation (7.5) can give rather irregular grid trajectories that oscillate in time. To avoid grid irregularities, some form of additional control over the grid movement is required by a *grid regularization*. For 1D problems an effective grid regularization technique has been proposed by Dorfi & Drury (1987) by which successive spatial grid size ratios can be bounded from below and above and temporal grid oscillations can be minimized. By their technique the moving-grid equation system (7.5) is extended to a nonlinear system of ODEs provided with two grid regularization parameters for steering the grid movement, one for spatial smoothing and one for temporal relaxation. The new system, replacing (7.5), is of the implicit form

$$\nu B(w, X)X' = G(w, X), \quad (7.8)$$

with $X = X(t) \in \mathbb{R}^m$ denoting the grid and ν representing the temporal relaxation parameter. Such systems are best solved with stiff solvers because the matrix B multiplying X' requires the solution of linear algebraic systems anyhow. Furthermore, with $\nu = 0$, as in (7.5), the full system reduces to a DAE system. For the technical derivation details and numerical results obtained with this specific regularization technique we refer to Dorfi & Drury (1987), Verwer et al. (1989) and Huang & Russell (1997).

Example 7.1 For a wide class of one-dimensional, time-dependent PDE systems, Blom & Zegeling (1994) have implemented the above moving-grid system (7.8) based on the arclength monitor (7.7) with the Dorfi-Drury regularization into a moving-grid interface, called MGI,²⁵⁾ which can be coupled with any existing DAE solver. MGI thus automizes the spatial discretization and the grid movement, and the DAE solver automizes the time integration. The spatial discretization is based on a central, second-order, lumped Galerkin finite element scheme due to Skeel & Berzins (1990).

We show results of MGI coupled to DASSL (Petzold, 1982) for a problem from Zegeling et al. (1992, Example III) describing the transport of brine in groundwater.²⁶⁾ This problem originates from models used to assess the risks of disposal of radioactive waste in salt formations. It has two components with a hydrodynamic pressure $p(x, t)$ and salt concentration $\omega(x, t)$ as independent variables,

$$(\psi\rho\beta)p_t + (\psi\rho\gamma)\omega_t = -(\rho q)_x ,$$

$$(\psi\rho)\omega_t = (-\rho q)\omega_x + (\rho\lambda|q|\omega_x)_x ,$$

were $0 < x < L = 1$ and $0 < t \leq T = 2$. Here $q = -k\mu^{-1}(p_x + \rho g)$ is the fluid velocity derived from Darcy's law and ρ is the fluid density for which the equation of state $\rho = \rho_0 \exp(\beta(p - p_0) + \gamma\omega)$ holds, with constant reference density ρ_0 , reference pressure p_0 , compressibility coefficient β and salt coefficient γ . Other constants are porosity ψ , permeability k , viscosity μ , gravity g and dispersion length λ . The initial functions are

$$\omega(x, 0) = 0 , \quad p(x, 0) = p_0((1 - x)p_l + xp_r) ,$$

with constants p_l, p_r , and the boundary conditions read

$$\omega(0, t) = \frac{1}{2}\left(1 - \tanh\left(10^3\left(t - \frac{3}{4}\right)\right)\right) , \quad \omega_x(1, t) = 0 ,$$

$$p(0, t) = p_0p_l , \quad p(1, t) = p_0p_r .$$

In this model, the compressibility coefficient β is very small, ruling out explicit time stepping.

From the equation for the salt transport we find, after freezing values, the Péclet number

$$|(L\rho q)/(\rho\lambda q)| = L/\lambda .$$

With our scaled variables we have $L = 1$. So with $\lambda \ll 1$ steep concentration gradients are expected and because MGI uses a central discretization, strong oscillations would arise on a fixed grid, as characterized by the cell Péclet condition (I.3.42). By using a moving grid, adapted to the steep gradients,

²⁵⁾ This FORTRAN program is available as Algorithm 731 from the *ACM Trans. Math. Softw.* and can be downloaded from <http://www.netlib.org/toms/731>

²⁶⁾ Paul Zegeling is acknowledged for carrying out the numerical tests.

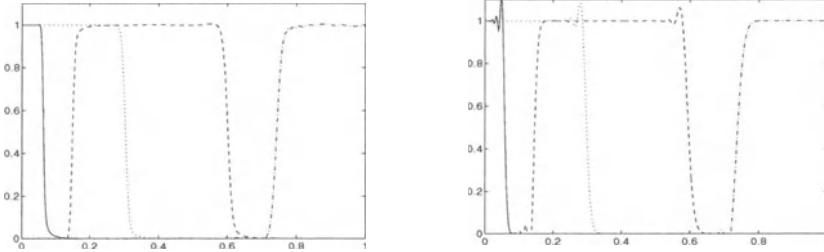


Fig. 7.1. Computed solutions ω for the brine transport problem at times $t = 0.1$ (solid), 0.5 (dotted), 1.0 (dashed) and 2.0 (dash-dotted). At the left the moving grid solution with 100 points, at the right a fixed grid solution with 500 points.

it is possible to solve the problem with relatively few grid points. Zegeling et al. (1992) used the value $\lambda = 10^{-3}$; here we have put $\lambda = 10^{-4}$ to create somewhat steeper profiles. The other parameter values are $p_0 = \omega_0 = \rho_0 = 1$, $g = 0.0981$, $\psi = 0.2$, $\gamma = 0.1794$, $\beta = 10^{-5}$ and $p_l = 1.7$, $p_r = 1.0$.

For these parameters, Figure 7.1 depicts the numerical solution for $\omega(\cdot, t)$ at times $t = 0.1, 0.5, 1.0, 2.0$, showing the following behaviour. The spatial step function at the initial time causes a front traveling to the right. At $t = 0.75$, the given boundary function for ω at $x = 0$ has a steep jump that generates a second front at $x = 0$, resulting in a block-shaped solution profile that travels to the right. The pressure p is not shown; it varies more slowly.

For a moving grid method this solution behaviour provides a challenging test. When starting from a uniform grid, which was chosen here with $m = 100$, the grid must rapidly refine near $x = 0$ to follow the front; then at $t = 0.75$ redistribute to capture the second emerging front, and follow the block profile; then redistribute again when the first front arrives at $x = 1$, and follow the second front. Figure 7.2 shows that the grid computed with MGI is in full accordance with the dynamics of the problem. The grid regularization parameters were chosen the same as in Zegeling et al. (1992).

The moving grid solution shown in Figure 7.1 is fairly accurate, but there are some small wiggles visible and the fronts are slightly smeared. In this

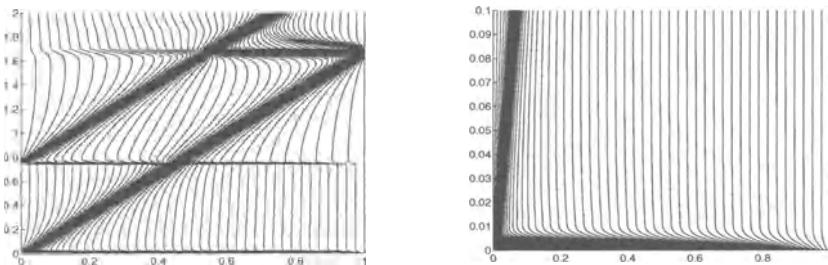


Fig. 7.2. The moving grid for the brine transport problem (horizontal axis x , vertical axis t). The plot at the right zooms in at the initial refinement near $x = 0$.

simulation the number of points was 100, which is about the minimum that could be used for this problem with the central MGI discretization scheme. As said before, using a fixed grid does require significantly more points with a central scheme and we see that even with 500 points quite large oscillations result. It is to be expected that a more sophisticated upwind-type discretization that takes into account the dominating advection nature of the problem will improve both the moving and fixed grid computation quite significantly, but schemes with limiters are difficult to combine with implicit time stepping. Finally we note that for obtaining optimal solutions with a moving grid code like MGI, some experience is needed to select proper regularization parameters. \diamond

There exist many more dynamic regridding methods. A well-known method is the moving finite element method, see Miller (1981) and Baines (1994). A recent proceedings on adaptive method of lines methods is Vande Wouwer, Saucez & Schiesser (2001) and another useful source on dynamic regridding literature is Zegeling (1999). Obviously, these methods are expected to be most useful for multi-dimensional problems because then the savings in computational effort would be largest. Whereas in higher space dimension equidistribution is not so easily applicable and grid regularization and tuning becomes even more essential, for special problem classes it can pay to design and apply dedicated moving-grid techniques, see for instance Cao, Huang & Russell (2002), Cao et al. (2003), and references therein.

7.2 Static Regridding

Spatial grid adaptivity in time based on static regridding seems to have a wider scope of applicability and generally seems more suitable and powerful for solving multi-dimensional problems. We speak of static regridding when space grids are adjusted only at discrete time levels. A static regridding method does not employ the Lagrangian form but solves the PDE on Eulerian grids. This approach offers the advantage of *locally refining* and *locally coarsening* the grid by adding and deleting points or cells, and, possibly, *local time stepping* with different step sizes over the grid. For multi-dimensional, complicated solution profiles this is an advantage compared to dynamic regridding and because points are not moved continuously one has easier control over the grid.

The idea of adaptivity based on static regridding is general. As a result it is found in many different versions, especially for solving hard technical engineering and CFD problems. Static regridding can for example be done by local uniform grid refinement (LUGR) as proposed by Gropp (1980), Berger & Oliger (1984) and Arney & Flaherty (1989). Figure 7.3 illustrates locally refined grids computed with the LUGR method found in Blom et al. (1996).

In this method the solution is forwarded in time on cascades of such nested, locally refined grids which are moved at discrete time levels. Basically

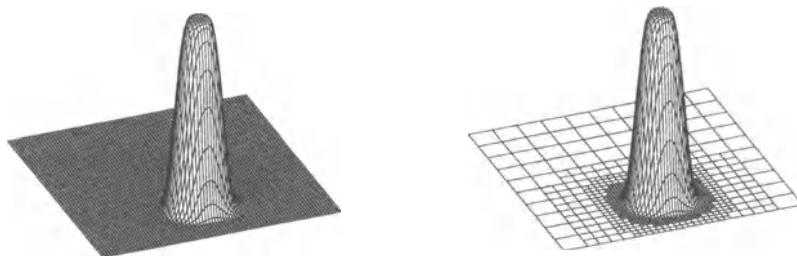


Fig. 7.3. A locally uniformly-refined Cartesian grid.

it works as follows: for a time step $t_n \mapsto t_{n+1}$, first a step on a coarse base grid is performed. In regions where the computed solution does not pass an heuristic spatial accuracy test, based on the curvature, the grid is refined by bisection. In the refined regions the time step $t_n \mapsto t_{n+1}$ is then redone. Required initial and boundary values for this second step may not be present, and then they are found by interpolation. For the time step $t_n \mapsto t_{n+1}$ this procedure is thus applied iteratively until the spatial accuracy test is satisfied everywhere, after which the whole procedure is repeated for the next time step $t_{n+1} \mapsto t_{n+2}$. For evolving time the refined grids with their solutions are saved, grid points are deleted and added, and the locally refined grids are thus moved at discrete time levels.

The method from Blom et al. (1996) has been designed primarily for 2D and 3D parabolic problems in Cartesian geometry and is implemented in the FORTRAN programs VLUGR2, VLUGR3 which were coded for general use by non-experts.²⁷⁾ The integration method is based on the two-step implicit BDF formula (II.3.11), provided with a variable step size strategy and advanced numerical algebra routines for solving large sparse linear systems. A powerful adaptive finite element method for parabolic problems using Rosenbrock formulas from class (II.1.25) for time integration is found in Lang (2000). This adaptive finite element method is implemented in 1D, 2D and 3D in a software package called KARDOS.²⁸⁾

Especially in the finite volume and finite element literature, highly sophisticated adaptive solution methods on structured and unstructured grids have been developed, see for instance Flaherty et al. (2000) and Sonar (2002). Obviously, the highly sophisticated level of these methods also requires a high level of expertise for their intended users. Unstructured grid methods are often advocated to deal with very complex geometries and for body fitted calculations. In contrast to this, Berger and co-workers have developed an

²⁷⁾ The codes are available as Algorithm 758 (VLUGR2), 759 (VLUGR3) from the *ACM Trans. Math. Softw.* and can be downloaded from <http://www.netlib.org/toms/758> and <http://www.netlib.org/toms/759>

²⁸⁾ Available from <http://www.zib.de/SciSoft/kardos/>

advanced adaptive technique for complex geometries with non-body fitted grids based on the more simple non-uniform Cartesian grids; see for instance Aftosmis, Berger & Melton (1999).

IV Splitting Methods

For many PDE problems in higher space dimension, such as the advection-diffusion-reaction system

$$u_t + \nabla \cdot (\underline{a} u) = \nabla \cdot (D \nabla u) + f(u),$$

it is in general inefficient or infeasible to apply one and the same integration formula to the different parts of the system. For example, the chemistry can be very stiff, which calls for an implicit ODE method. On the other hand, if the advection is discretized in space using a limiter, then explicit methods are often much more suitable for that part of the equation. Moreover, use of a single implicit integration formula for the whole problem readily leads to a nonlinear algebraic system too large to handle due to the simultaneous coupling over the species and over space. In such cases a more tuned approach based on an appropriate form of *splitting* is advocated. The general idea behind splitting is breaking down a complicated problem into smaller parts for the sake of time stepping, such that the different parts can be solved efficiently with suitable integration formulas.

1 Operator Splitting

This section is devoted to the technique called *operator splitting* or *time splitting*. We will discuss this form of splitting for ODE and PDE problems without invoking actual integration formulas. Hence in this section we mainly focus on concepts rather than on actual methods.

1.1 First-Order Splitting

Linear ODE Problems

Let us first illustrate the notion of splitting by considering a linear, homogeneous ODE system

$$w'(t) = Aw(t), \quad t > 0, \quad w(0) = w_0, \tag{1.1}$$

and assume for A a two-term splitting

$$A = A_1 + A_2.$$

System (1.1) may for example be seen as a semi-discretization of a linear PDE problem with homogeneous or periodic boundary conditions. The solution of (1.1) is given by

$$w(t_{n+1}) = e^{\tau A} w(t_n), \quad (1.2)$$

where $\tau = t_{n+1} - t_n$. If we wish to use only A_1 and A_2 separately, instead of the full A , then (1.2) can be approximated by

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n \quad (1.3)$$

with w_n approximating $w(t_n)$. This is the simplest splitting, in which we solve the two subproblems

$$\begin{aligned} \frac{d}{dt} w^*(t) &= A_1 w^*(t) \quad \text{for } t_n < t \leq t_{n+1} \text{ with } w^*(t_n) = w_n, \\ \frac{d}{dt} w^{**}(t) &= A_2 w^{**}(t) \quad \text{for } t_n < t \leq t_{n+1} \text{ with } w^{**}(t_n) = w^*(t_{n+1}), \end{aligned}$$

one after another, starting from w_n , and take $w_{n+1} = w^{**}(t_{n+1})$ to complete the splitting integration step.

Replacing (1.2) by (1.3) normally introduces an error, the so-called *splitting error*. Inserting the exact solution w of the original problem into (1.3) gives

$$w(t_{n+1}) = e^{\tau A_2} e^{\tau A_1} w(t_n) + \tau \rho_n,$$

with local truncation error ρ_n . Recall that $\tau \rho_n$ is the error introduced per step starting from the true solution, hence it is the *local* splitting error. We have

$$\begin{aligned} e^{\tau A} &= I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2(A_1 + A_2)^2 + \dots, \\ e^{\tau A_2} e^{\tau A_1} &= I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2(A_1^2 + 2A_2 A_1 + A_2^2) + \dots. \end{aligned}$$

The truncation error thus satisfies

$$\rho_n = \frac{1}{\tau} (e^{\tau A} - e^{\tau A_2} e^{\tau A_1}) w(t_n) = \frac{1}{2} \tau [A_1, A_2] w(t_n) + \mathcal{O}(\tau^2), \quad (1.4)$$

with

$$[A_1, A_2] = A_1 A_2 - A_2 A_1 \quad (1.5)$$

being the *commutator* of A_1 and A_2 . Obviously (1.3) is a first-order process unless A_1 and A_2 commute. In case A_1 and A_2 do commute we have

$$e^{\tau A_2} e^{\tau A_1} = e^{\tau A_2 + \tau A_1} = e^{\tau A},$$

which can be seen by using the power series expansion (I.2.19) for the exponential function. It follows that for commuting matrices the splitting (1.3) is exact, it leaves no splitting error.

In actual computation the subproblems are to be integrated by means of a suitable numerical ODE integration method. One has considerable freedom in selecting methods suitable for the various sub-steps. Whatever choice is made, however, integration errors are incurred in addition to the splitting error. In this section we focus on the splitting itself and assume that the subproblems are solved exactly (or sufficiently accurately).

The expansion in (1.4) assumes that the $\mathcal{O}(\tau)$ term is leading and dominates the higher-order terms. This requires that the commutator value is of moderate size, $[A_1, A_2]w(t_n) = \mathcal{O}(1)$. If the original PDE problem has homogeneous or periodic boundary conditions this is a reasonable assumption.

Stability

With regard to stability, if we have $\|e^{\tau A_k}\| \leq 1$, $k = 1, 2$, then it follows trivially that $\|w_{n+1}\| \leq \|w_n\|$ for the splitting (1.3). In the same way

$$\|e^{\tau A_k}\| \leq e^{\tau \omega_k}, \quad k = 1, 2 \quad \Rightarrow \quad \|w_{n+1}\| \leq e^{\tau \omega} \|w_n\|$$

with $\omega = \omega_1 + \omega_2$. If $\omega > 0$ this shows stability on finite time intervals $[0, T]$; if $\omega \leq 0$ the time interval may be arbitrarily large.

General stability results under the weaker assumption that $\|e^{t A_k}\| \leq K$ for $0 \leq t \leq T$ with a constant $K \geq 1$ seem unknown. However, in practice the splitting appears to be stable provided the sub-steps themselves are stable. If the performance of the splitting is disappointing, it is in general the local accuracy that needs improvement.

The Baker-Campbell-Hausdorff Formula

The Baker-Campbell-Hausdorff (BCH) formula expresses the product of two exponentials as one new exponential:

$$e^{\tau A_2} e^{\tau A_1} = e^{\tau \tilde{A}} \tag{1.6}$$

with

$$\begin{aligned} \tilde{A} = A + \frac{1}{2}\tau[A_2, A_1] + \frac{1}{12}\tau^2 &\left([A_2, [A_2, A_1]] + [A_1, [A_1, A_2]] \right) \\ &+ \frac{1}{24}\tau^3[A_2, [A_1, [A_1, A_2]]] + \mathcal{O}(\tau^4). \end{aligned} \tag{1.7}$$

Clearly, if A_1, A_2 commute all higher-order terms in the expansion vanish and $\tilde{A} = A$. This formula can be derived from the power series development for the three exponentials by comparing terms with the same power of τ . The calculation of the terms in \tilde{A} quickly becomes cumbersome if done in a straightforward fashion, but it can also be done in a recursive way, see Sanz-Serna & Calvo (1994) and the references given there. Using a Lie operator formalism, a similar formula can also be derived for nonlinear autonomous equations. This will be discussed in Section 1.4.

From formula (1.7) we can recover the truncation error (1.4), but we can also apply this formula in a global fashion,

$$(e^{\tau A_2} e^{\tau A_1})^n = e^{n\tau \tilde{A}} = e^{t_n \tilde{A}},$$

assuming a constant step size τ . Hence, the splitting process (1.3) exactly solves the *modified* equation $w'(t) = \tilde{A}w(t)$, rather than the original equation $w'(t) = Aw(t)$, and the global splitting error at $t = t_n$ reads

$$w(t_n) - w_n = (e^{t_n A} - e^{t_n \tilde{A}})w_0.$$

Multi-Component Splittings

The two-term splitting can be generalized to include more components. For example, if $A = A_1 + A_2 + A_3$, the first-order splitting (1.3) generalizes to

$$w_{n+1} = e^{\tau A_3} e^{\tau A_2} e^{\tau A_1} w_n.$$

This can be viewed as a repeated application of the two-term splitting: first write A as $A_1 + B$, $B = A_2 + A_3$, and then use splitting of B . Consequently the considerations on accuracy and stability carry over directly. In particular we now obtain the truncation error

$$\rho_n = \frac{1}{2}\tau \left([A_1, A_2] + [A_1, A_3] + [A_2, A_3] \right) w(t_n) + \mathcal{O}(\tau^2).$$

Nonlinear ODE Problems

For general nonlinear ODE systems

$$w'(t) = F(t, w(t)), \quad t > 0, \quad w(0) = w_0, \quad (1.8)$$

with the two-term splitting

$$F(t, v) = F_1(t, v) + F_2(t, v),$$

the above linear splitting (1.3) is reformulated as

$$\frac{d}{dt}w^*(t) = F_1(t, w^*(t)) \quad \text{for } t_n < t \leq t_{n+1} \text{ with } w^*(t_n) = w_n,$$

$$\frac{d}{dt}w^{**}(t) = F_2(t, w^{**}(t)) \quad \text{for } t_n < t \leq t_{n+1} \text{ with } w^{**}(t_n) = w^*(t_{n+1}),$$

giving $w_{n+1} = w^{**}(t_{n+1})$ as the next approximation. As in the linear case, in actual computation a suitable ODE method has to be used for approximately solving the subproblems. With $w_n = w(t_n)$ we now get the local truncation error

$$\rho_n = \frac{1}{2}\tau \left[\frac{\partial F_1}{\partial w} F_2 - \frac{\partial F_2}{\partial w} F_1 \right] (t_n, w(t_n)) + \mathcal{O}(\tau^2). \quad (1.9)$$

This truncation error is the nonlinear counterpart of (1.4) and again reveals a formal consistency order of one. This formula can be derived by Taylor expansions of $w^*(t_{n+1})$ and $w^{**}(t_{n+1})$ around $t = t_n$. If the bracketed term vanishes, the local truncation error becomes at least $\mathcal{O}(\tau^2)$.

1.2 Second-Order Symmetrical Splitting

Linear ODE problems

The splitting (1.3) starts in all steps with application of A_1 . Interchanging the order of A_1 and A_2 after each step will lead to symmetry and better accuracy. Carrying out two half steps with reversed sequence gives

$$w_{n+1} = \left(e^{\frac{1}{2}\tau A_1} e^{\frac{1}{2}\tau A_2} \right) \left(e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_1} \right) w_n = e^{\frac{1}{2}\tau A_1} e^{\tau A_2} e^{\frac{1}{2}\tau A_1} w_n. \quad (1.10)$$

This idea of symmetry in splitting has been proposed by Strang (1968) and Marchuk (1971). It is commonly called *Strang splitting*.

By a series expansion, and after some tedious calculations, the local truncation error is found to satisfy

$$\rho_n = \frac{1}{24}\tau^2 \left([A_1, [A_1, A_2]] + 2[A_2, [A_1, A_2]] \right) w(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^4), \quad (1.11)$$

revealing a formal consistency order of two. This result can also be found by repeated application of formula (1.6), (1.7). Due to symmetry around the point $t_{n+1/2}$, the expansion contains only even-order terms, see Remark II.5.2.

If we work with constant step sizes τ , method (1.10) will require almost the same amount of computational work as (1.3), which is clear by writing (1.10) as

$$w_n = e^{\frac{1}{2}\tau A_1} e^{\tau A_2} e^{\tau A_1} \cdots e^{\tau A_1} e^{\tau A_2} e^{\frac{1}{2}\tau A_1} w_0.$$

With variable step sizes the second-order process will be more expensive, but in general its order two pays off. This makes symmetrical splitting more popular than first-order splitting.

An earlier second-order splitting of Strang (1963) is

$$w_{n+1} = \frac{1}{2} \left(e^{\tau A_1} e^{\tau A_2} + e^{\tau A_2} e^{\tau A_1} \right) w_n. \quad (1.12)$$

Its truncation error reads

$$\rho_n = -\frac{1}{12}\tau^2 \left([A_1, [A_1, A_2]] + [A_2, [A_2, A_1]] \right) w(t_n) + \mathcal{O}(\tau^3). \quad (1.13)$$

This process is more expensive than (1.3). An advantage is that the factors $e^{\tau A_1} e^{\tau A_2}$ and $e^{\tau A_2} e^{\tau A_1}$ can be computed in parallel, but in general computation of individual exponentials $e^{\tau A_j}$ will already offer many possibilities for parallelism. For nonlinear problems the same considerations hold.

Multi-Component Splittings

If $A = A_1 + A_2 + A_3$, Strang's symmetrical splitting method (1.10) becomes

$$w_{n+1} = e^{\frac{1}{2}\tau A_1} e^{\frac{1}{2}\tau A_2} e^{\tau A_3} e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_1} w_n.$$

Note that this is just a repeated application of (1.10). First approximate $e^{\tau A}$ by $e^{\frac{1}{2}\tau A_1}e^{\tau(A_2+A_3)}e^{\frac{1}{2}\tau A_1}$ and then approximate $e^{\tau(A_2+A_3)}$ in the same fashion. The parallel splitting method (1.12) can be generalized as

$$w_{n+1} = \frac{1}{2} \left(e^{\tau A_1} e^{\tau A_2} e^{\tau A_3} + e^{\tau A_3} e^{\tau A_2} e^{\tau A_1} \right) w_n,$$

which also gives a second-order truncation error.

Nonlinear ODE Problems

Extending Strang's symmetrical splitting method (1.10) to nonlinear systems (1.8) is straightforward,

$$\frac{d}{dt}w^*(t) = F_1(t, w^*(t)), \quad t_n < t \leq t_{n+\frac{1}{2}}, \quad w^*(t_n) = w_n,$$

$$\frac{d}{dt}w^{**}(t) = F_2(t, w^{**}(t)), \quad t_n < t \leq t_{n+1}, \quad w^{**}(t_n) = w^*(t_{n+\frac{1}{2}}),$$

$$\frac{d}{dt}w^{***}(t) = F_1(t, w^{***}(t)), \quad t_{n+\frac{1}{2}} < t \leq t_{n+1}, \quad w^{***}(t_{n+\frac{1}{2}}) = w^{**}(t_{n+1}),$$

giving $w_{n+1} = w^{***}(t_{n+1})$ as the next approximation. We then also have a formal consistency order of two. The local truncation error now contains many terms. If we assume that the equation is autonomous, Taylor expansion gives after some calculations the expression

$$\begin{aligned} & \frac{1}{24}\tau^2 \left[\frac{\partial}{\partial w} \left(\frac{\partial F_1}{\partial w} F_1 \right) F_2 - 2 \frac{\partial}{\partial w} \left(\frac{\partial F_1}{\partial w} F_2 \right) F_1 + \frac{\partial}{\partial w} \left(\frac{\partial F_2}{\partial w} F_1 \right) F_1 \right. \\ & \left. - 2 \frac{\partial}{\partial w} \left(\frac{\partial F_2}{\partial w} F_2 \right) F_1 + 4 \frac{\partial}{\partial w} \left(\frac{\partial F_2}{\partial w} F_1 \right) F_2 - 2 \frac{\partial}{\partial w} \left(\frac{\partial F_1}{\partial w} F_2 \right) F_2 \right] + \mathcal{O}(\tau^4) \end{aligned} \quad (1.14)$$

with the F_j and derivative terms evaluated in $w(t_{n+1/2})$. For non-autonomous problems the truncation error becomes more complicated, but the method is still of order two. This Strang splitting method extends in a straightforward way to multiple splittings of F , in the same way as for linear problems.

1.3 Higher-Order Splittings

The above linear splitting methods fit in the more general form

$$w_{n+1} = \sum_{i=1}^s \alpha_i \left(\prod_{j=1}^r e^{\tau \beta_{ij} A_1} e^{\tau \gamma_{ij} A_2} \right) w_n \quad \text{with} \quad \sum_{i=1}^s \alpha_i = 1. \quad (1.15)$$

If we assume again $\|e^{tA_k}\| \leq 1$ for $t \geq 0$, $k = 1, 2$ and if all coefficients $\alpha_i, \beta_{ij}, \gamma_{ij} \geq 0$, we obtain as above the stability estimate $\|w_{n+1}\| \leq \|w_n\|$. One could try to find suitable parameter choices that give higher-order processes, but it was shown by Sheng (1989) that for having order $p > 2$ some of

the coefficients must be negative. The result of Sheng was refined by Goldman & Kaper (1996), who showed that

$$p > 2 \text{ and } \alpha_i > 0, 1 \leq i \leq s \implies \min \beta_{ij} < 0 \text{ and } \min \gamma_{ij} < 0,$$

and thus a step with negative time is necessary for both A_1 and A_2 . The proofs of these results are long and technical. Therefore we only discuss here briefly two examples of higher-order splittings.

Let $S_\tau = e^{\frac{1}{2}\tau A_1} e^{\tau A_2} e^{\frac{1}{2}\tau A_1}$ be the second-order symmetrical splitting operator. By using local Richardson extrapolation, one obtains the fourth-order splitting

$$w_{n+1} = \left(\frac{4}{3} (S_{\frac{1}{2}\tau})^2 - \frac{1}{3} S_\tau \right) w_n,$$

which has a negative weight $-\frac{1}{3}$. Because of this negative weight, the stability considerations considered above no longer hold. In fact, it is not precisely known for what kinds of problems this scheme will be stable or unstable. It was shown by Dia & Schatzman (1996) that the scheme will be stable on finite time intervals for a class of linear parabolic problems with dimensional splitting. If instabilities with this local extrapolation may occur, then *global* extrapolation at the end point $t = T$ might be more beneficial.

Another fourth-order splitting, derived by Yoshida (1990) and Suzuki (1990), reads

$$w_{n+1} = S_{\theta\tau} S_{(1-2\theta)\tau} S_{\theta\tau} w_n,$$

with $\theta = (2 - \sqrt[3]{2})^{-1} \approx 1.35$. Here we have $1 - 2\theta < 0$, and thus a step with negative time has to be taken.

For partial differential equations with boundary conditions such splittings with negative time steps seem of limited value. Diffusion or stiff reaction terms lead to ill-posedness when the time is reversed. For advection problems with Dirichlet conditions at the inflow boundary, taking a step backwards in time requires boundary values at the outflow boundary. Higher-order splittings are frequently used for conservative problems where boundary conditions are not relevant, such as with the Schrödinger equation (I.8.2), or for certain mechanical problems, see Sanz-Serna & Calvo (1994). Higher-order symplectic methods for Hamiltonian wave equations and ODEs are also found in McLachlan (1994, 1995) and McLachlan & Quispel (2002).

1.4 Abstract Initial Value Problems

Time splitting can also be discussed at the level of the PDE problem. The general ideas are independent of the particular spatial discretizations used. Considering directly the PDE problem will make it more clear in which cases splitting will be exact. In addition to the initial value problem for the ODE systems (1.1) and (1.8), we therefore now turn our attention to the initial value problem for abstract autonomous systems

$$u_t(\underline{x}, t) = \mathbf{f}(\underline{x}, u(\underline{x}, t)). \quad (1.16)$$

With this abstract problem one may associate any ODE or PDE initial value problem in autonomous form without boundary conditions. In the ODE case one may replace it by our more commonly used form $w'(t) = F(w(t))$ with $w(t)$ a vector or grid function instead of a space-continuous function $u(\underline{x}, t)$. In the PDE case \mathbf{f} is to be seen as a spatial partial differential operator, for example the operator

$$\mathbf{f}(u) = -\nabla \cdot (\underline{a}u) + \nabla \cdot (D \nabla u) + f(u) \quad (1.17)$$

of an advection-diffusion-reaction system where \underline{a} , D and f may depend on the spatial variable $\underline{x} \in \mathbb{R}^d$, and $u = u(\underline{x}, t)$ is a function of \underline{x} and the temporal variable $t \in [0, T]$.

For convenience of notation we suppress the dependence of u and \mathbf{f} on \underline{x} and write $u_t = \mathbf{f}(u)$. For most of what follows it is not needed to specify the spatial dimension d , the number of components of u and the function space u lives in. We only assume sufficient differentiability and denote the function space by \mathcal{U} . With (1.16) we associate the *solution operator* \mathbf{S}_τ acting on \mathcal{U} such that the exact solution u of (1.16) is given by

$$u(t + \tau) = \mathbf{S}_\tau(u(t)).$$

This generalizes the exponential operator $e^{\tau A}$ of the linear ODE system (1.1).

Now assume for \mathbf{f} the two-term splitting

$$\mathbf{f}(u) = \mathbf{f}_1(u) + \mathbf{f}_2(u),$$

and associate with $u_t = \mathbf{f}_k(u)$ the solution operator $\mathbf{S}_{k,\tau}$. For the abstract initial value problem (1.16), the basic splitting method (1.3) is then reformulated as

$$u_{n+1} = \mathbf{S}_{2,\tau}(\mathbf{S}_{1,\tau}(u_n)) \quad (1.18)$$

with $u_n \in \mathcal{U}$ approximating $u(t_n)$. As above, inserting the exact solution u of (1.16) gives the relation for the local truncation error,

$$u(t_{n+1}) = \mathbf{S}_{2,\tau}(\mathbf{S}_{1,\tau}(u(t_n))) + \tau \rho_n,$$

and by Taylor expansion one then finds the formal expression

$$\rho_n = \frac{1}{2}\tau \left[\frac{\partial \mathbf{f}_1}{\partial u} \mathbf{f}_2 - \frac{\partial \mathbf{f}_2}{\partial u} \mathbf{f}_1 \right] (u(t_n)) + \mathcal{O}(\tau^2), \quad (1.19)$$

similar to (1.9) for ODE problems. Likewise the counterpart of the local error (1.14) of the symmetrical splitting method (1.10) is recovered when this method is reformulated for the current problem. The derivations for multi-component splittings also follow the same lines, but become of course more lengthy.

An important observation is that similarly as for the local truncation error (1.4) in the linear ODE case, the complete truncation error (1.19) vanishes

with the bracketed term. In analogy with the linear ODE case, this bracketed term is called the *commutator* of the operators $\mathbf{f}_1, \mathbf{f}_2$. We denote this commutator by

$$[\mathbf{f}_1, \mathbf{f}_2](u) = \mathbf{f}'_1(u) \mathbf{f}_2(u) - \mathbf{f}'_2(u) \mathbf{f}_1(u), \quad (1.20)$$

where the primes denote differentiation with respect to u . Concrete examples are discussed in the next section.

The fact that the splitting process (1.18) leaves no splitting error if the commutator (1.20) is zero can be proven by means of a Lie operator formalism through which (1.18) is rewritten in a linear form analogous to the linear ODE case. For the presentation here we follow Sanz-Serna & Calvo (1994) and Sanz-Serna (1997). A classic introduction to Lie operators is Gröbner (1960), see also Hairer, Lubich & Wanner (2002).

The Lie Operator Formalism

To any operator \mathbf{f} acting on \mathcal{U} we associate a Lie operator, denoted by \mathcal{F} . This Lie operator \mathcal{F} is a *linear* operator on the space of operators acting on \mathcal{U} . By definition, \mathcal{F} maps any operator \mathbf{g} acting on \mathcal{U} into a new operator $\mathcal{F}\mathbf{g}$, such that for any element $v \in \mathcal{U}$,

$$\mathcal{F}\mathbf{g}(v) = \mathbf{g}'(v) \mathbf{f}(v). \quad (1.21)$$

For the solution $u(t)$ of $u_t = \mathbf{f}(u)$ it follows that

$$\mathcal{F}\mathbf{g}(u(t)) = \mathbf{g}'(u(t)) \mathbf{f}(u(t)) = \frac{\partial}{\partial t} \mathbf{g}(u(t)). \quad (1.22)$$

So $\mathcal{F}\mathbf{g}(u(t))$ measures the rate of change of \mathbf{g} along the solution of $u_t = \mathbf{f}(u)$. Likewise we get for $k \geq 1$

$$\mathcal{F}^k \mathbf{g}(u(t)) = \frac{\partial^k}{\partial t^k} \mathbf{g}(u(t)). \quad (1.23)$$

Using the exponentiated Lie operator form we then find ¹⁾

$$e^{\tau \mathcal{F}} \mathbf{g}(u(t)) = \sum_{k=0}^{\infty} \frac{\tau^k}{k!} \mathcal{F}^k \mathbf{g}(u(t)) = \sum_{k=0}^{\infty} \frac{\tau^k}{k!} \frac{\partial^k}{\partial t^k} \mathbf{g}(u(t)). \quad (1.24)$$

The right-hand side is the Taylor series of $\mathbf{g}(u(t + \tau)) = \mathbf{g}(\mathbf{S}_\tau(u(t)))$, so formally we have

$$e^{\tau \mathcal{F}} \mathbf{g}(\cdot) = \mathbf{g}(\mathbf{S}_\tau(\cdot)). \quad (1.25)$$

¹⁾ These series, which are called Lie-Taylor series, are to be interpreted as formal series whose convergence properties are set aside.

This equality holds for any \mathbf{g} defined on \mathcal{U} , in particular for the identity I . Hence the Lie-Taylor series operator $e^{\tau\mathcal{F}}I$ represents the solution operator \mathbf{S}_τ .

The same arguments apply to the subproblems $u_t = \mathbf{f}_k(u)$. If \mathcal{F}_k denotes the Lie operator associated to \mathbf{f}_k , then to each subproblem the operator equality (1.25) applies with \mathcal{F} replaced by \mathcal{F}_k and \mathbf{S}_τ by $\mathbf{S}_{k,\tau}$. We can now represent the splitting formula (1.18) by a composition of the Lie-Taylor series of \mathcal{F}_1 and \mathcal{F}_2 . This composition reads

$$u_{n+1} = e^{\tau\mathcal{F}_1} e^{\tau\mathcal{F}_2} I(u_n), \quad (1.26)$$

where one should note the reversed sequence compared to (1.18). This result is a consequence of the equality

$$e^{\tau\mathcal{F}_1} e^{\tau\mathcal{F}_2} \mathbf{g}(\cdot) = \mathbf{g}(\mathbf{S}_{2,\tau}(\mathbf{S}_{1,\tau}(\cdot))), \quad (1.27)$$

whose validity for any operator \mathbf{g} can be seen as follows. Introduce $\bar{\mathbf{g}} = e^{\tau\mathcal{F}_2} \mathbf{g}$. Applying (1.25) twice then gives

$$e^{\tau\mathcal{F}_1} e^{\tau\mathcal{F}_2} \mathbf{g} = e^{\tau\mathcal{F}_1} \bar{\mathbf{g}} = \bar{\mathbf{g}}(\mathbf{S}_{1,\tau}) = e^{\tau\mathcal{F}_2} \mathbf{g}(\mathbf{S}_{1,\tau}) = \mathbf{g}(\mathbf{S}_{2,\tau}(\mathbf{S}_{1,\tau})).$$

Note that the Lie version of the nonlinear symmetrical splitting method derived from (1.18) reads

$$u_{n+1} = e^{\frac{1}{2}\tau\mathcal{F}_1} e^{\tau\mathcal{F}_2} e^{\frac{1}{2}\tau\mathcal{F}_1} I(u_n), \quad (1.28)$$

with the same sequence as (1.10) due to the symmetry.

To sum up, by means of the Lie operator formalism, we have transformed a general, possibly nonlinear splitting to a linear one. This enables the use of the commutator concept and the BCH formulas (1.6)-(1.7) with A , A_k , \tilde{A} replaced by \mathcal{F} , \mathcal{F}_k , $\tilde{\mathcal{F}}$, respectively. In particular, if the commutator

$$[\mathcal{F}_2, \mathcal{F}_1] = \mathcal{F}_2 \mathcal{F}_1 - \mathcal{F}_1 \mathcal{F}_2 \quad (1.29)$$

is the zero operator we have a zero splitting error. The commutator (1.29) is also a Lie operator. From (1.21) it follows that

$$[\mathcal{F}_2, \mathcal{F}_1] \mathbf{g}(u) = (\mathbf{g}'(u) \mathbf{f}_1(u))' \mathbf{f}_2(u) - (\mathbf{g}'(u) \mathbf{f}_2(u))' \mathbf{f}_1(u),$$

and inserting I for \mathbf{g} then gives the *nonlinear commutator* (1.20). For a given operator $\mathbf{f}(u)$ and splitting $\mathbf{f}(u) = \mathbf{f}_1(u) + \mathbf{f}_2(u)$, this nonlinear commutator can be elaborated to verify if it vanishes, resulting in a zero splitting error for (1.18) and for the symmetrical method derived from it. As with the linear ODE case, also multi-component splittings can be studied this way.

1.5 Advection-Diffusion-Reaction Splittings

We proceed with some examples of splittings for advection-diffusion-reaction problems. These splittings can be used for all kinds of spatial discretizations, finite volumes or finite elements, on structured and unstructured grids. Dimensional splittings will be considered separately in the next section, since the use of such splittings is restricted to Cartesian grids.

In the following examples the error of first-order splitting will be examined using the abstract formulation of the previous section, allowing a discussion of the splitting errors on the PDE level along the lines of Lanser & Verwer (1999). If this splitting error does not vanish, then second-order symmetrical splitting is advocated.

Splitting Advection and Diffusion

The advection-diffusion equation

$$u_t = \mathbf{f}_A(u) + \mathbf{f}_D(u), \quad \mathbf{f}_A(u) = -\nabla \cdot (\underline{a}u), \quad \mathbf{f}_D(u) = \nabla \cdot (D \nabla u),$$

is considered for $\underline{x} = (x, y, z)^T \in \mathbb{R}^3$ with $\underline{a} = (a_1, a_2, a_3)^T$ and a 3×3 matrix D , both spatially dependent. Splitting of advection and diffusion has the advantage that diffusion can be treated implicitly and advection explicitly, possibly with a characteristic based method.

To study the commutator of $\mathbf{f}_A, \mathbf{f}_D$, note that

$$\mathbf{f}'_A(u)v = -\nabla \cdot (\underline{a}v), \quad \mathbf{f}'_D(u)v = \nabla \cdot (D \nabla v).$$

It follows that $[\mathbf{f}_A, \mathbf{f}_D](u) = 0$ iff

$$\nabla \cdot (D \nabla (\nabla \cdot (\underline{a}u))) = \nabla \cdot (\underline{a}(\nabla \cdot (D \nabla u))).$$

Elaborating these two expressions shows equality if both \underline{a} and D are independent of \underline{x} . Since this happens only in model situations, second-order symmetrical splitting is generally advocated here.

Splitting Advection and Reactions

For the advection-reaction system

$$u_t = \mathbf{f}_A(u) + \mathbf{f}_R(u), \quad \mathbf{f}_A(u) = -\nabla \cdot (\underline{a}u), \quad \mathbf{f}_R(u) = f(u),$$

splitting the advection from the reaction has obvious computational advantages. We can then use an explicit method for the advection terms and an implicit one for the reaction term if it is stiff. In the advection step there will be only coupling in space, whereas in the reaction step we will have only coupling between the chemical species in single grid cells. These decouplings simplify the numerical solution process, and they leave room for massively

parallel computing. For large-scale applications this is of practical importance, for example, for air pollution models where one has many species (up to 100) and the number of grid cells is in the order of millions, see Zlatev (1995), Verwer et al. (2002).

We have

$$\mathbf{f}'_A(u)\mathbf{f}_R(u) = -\nabla \cdot (\underline{a}f(u)), \quad \mathbf{f}'_R(u)\mathbf{f}_A(u) = -f'(u)\nabla \cdot (\underline{a}u).$$

Let $f_x(u)$ denote the partial derivative vector of $f(\underline{x}, u)$ with respect to x , and likewise for the y - and z -direction. An elementary calculation yields

$$\begin{aligned}\mathbf{f}'_A(u)\mathbf{f}_R(u) &= -(a_1 f(u))_x - (a_2 f(u))_y - (a_3 f(u))_z \\ &= -f'(u)(\underline{a} \cdot \nabla u) - (\nabla \cdot \underline{a})f(u) - (a_1 f_x(u) + a_2 f_y(u) + a_3 f_z(u))\end{aligned}$$

and

$$\mathbf{f}'_R(u)\mathbf{f}_A(u) = -f'(u)(\underline{a} \cdot \nabla u) - f'(u)(\nabla \cdot \underline{a})u.$$

It follows that

Advection commutes with reaction if: (i) f is independent of \underline{x} and $\nabla \cdot \underline{a} = 0$, or (ii) f is independent of \underline{x} and linear in u .

Concerning (i), note that nonlinearity of f is still allowed and that in practical applications divergence-free velocity fields often occur which makes this result of practical interest. The linearity requirement on f in (ii) is not commonly fulfilled in practical situations.

If we assume that f is independent of \underline{x} , then the first-order splitting process (1.18), with reaction followed by advection, has the local truncation error

$$\rho_n = \frac{1}{2}\tau \left[(\nabla \cdot \underline{a})(f(u) - f'(u)u) \right] (t_n) + \mathcal{O}(\tau^2),$$

see (1.19). If advection precedes reaction the sign is reversed. A similar $\mathcal{O}(\tau^2)$ truncation error, but with a more lengthy expression is found for the second-order symmetrical process derived from (1.18), see (1.14). Finally we note that the expression $f(u) - f'(u)u$ in the truncation error may become large if the reaction term is stiff. This situation will be discussed in more detail in Section 1.7.

Splitting Diffusion and Reactions

For the diffusion-reaction system

$$u_t = \mathbf{f}_D(u) + \mathbf{f}_R(u), \quad \mathbf{f}_D(u) = \nabla \cdot (D\nabla u), \quad \mathbf{f}_R(u) = f(u),$$

splitting diffusion from the reaction terms has again computational advantages since simultaneous coupling over space and the various chemical species is then avoided. Similar to the previous example for advection-reaction, this offers room for massively parallel computing.

For commutativity, $[\mathbf{f}_D, \mathbf{f}_R](u) = 0$, we need equality of

$$\mathbf{f}'_D(u)\mathbf{f}_R(u) = \nabla \cdot (D\nabla f(u)) \quad \text{and} \quad \mathbf{f}'_R(u)\mathbf{f}_D(u) = f'(u)(\nabla \cdot (D\nabla u)).$$

Hence we see that

Diffusion commutes with reaction if f is linear in u and independent of \underline{x} .

In practical situations this will not be satisfied often. Usually the reaction term will be nonlinear. However the above expressions show that if the reaction consists of a stiff linear part plus nonlinearities, say $f(u) = Lu + \tilde{f}(u)$ with L independent of \underline{x} , then the linear part drops out of the commutator, and the truncation error for the first-order splitting (1.18), with diffusion followed by reaction, will be

$$\rho_n = \frac{1}{2}\tau \left[\nabla \cdot (D\nabla \tilde{f}(u)) - \tilde{f}'(u)(\nabla \cdot (D\nabla u)) \right] (t_n) + \mathcal{O}(\tau^2).$$

The same applies to the truncation error of second-order symmetrical splitting.

1.6 Dimension Splitting

Dimension Splitting for Advection

Solving the d -dimensional advection equation

$$u_t + \nabla \cdot (\underline{\alpha}u) = 0 \tag{1.30}$$

by the standard MOL approach, with explicit time stepping and finite differences or finite volumes in space on a d -dimensional grid without splitting, will commonly lead to a CFL condition for stability of the type

$$\sum_{k=1}^d \frac{\tau}{h_k} |a_k| \leq C_0,$$

where $C_0 > 0$ is a constant determined by the method; see the considerations for (III.6.17). On the other hand, if we split the equation along the spatial dimensions, while using the same discretizations, we get the CFL condition

$$\max_k \left(\frac{\tau}{h_k} |a_k| \right) \leq C_0.$$

This condition obviously allows larger time steps. Furthermore, dimension splitting also allows the use of tailored 1D schemes, such as the third-order direct-space-time (DST) scheme discussed in Section III.2, for which good and simple multi-dimensional extensions are difficult to derive.

Splitting (1.30) along the spatial dimensions gives $d(d-1)/2$ commutators of similar form. For example, according to (1.20), putting $\mathbf{f}_1(u) = -(a_1 u)_x$ and $\mathbf{f}_2(u) = -(a_2 u)_y$ yields

$$[\mathbf{f}_1, \mathbf{f}_2](u) = (a_1(a_2 u)_y)_x - (a_2(a_1 u)_x)_y = -((a_1)_y a_2 u)_x + ((a_2)_x a_1 u)_y.$$

This commutator will vanish if

$$(a_1)_y = (a_2)_x = 0.$$

In 3D the corresponding condition reads

$$(a_1)_y = (a_1)_z = (a_2)_x = (a_2)_z = (a_3)_x = (a_3)_y = 0.$$

In general this will not be satisfied, resulting in splitting errors. Even for spatially constant solutions $\neq 0$, splitting thus may result in errors appearing as unwanted spatial oscillations. This is an awkward property of dimension splitting which will be illustrated in Example 1.2. Symmetrical dimension splitting therefore is advocated. In spite of this awkward property, in general good results can then be expected as symmetrical splitting gives an $\mathcal{O}(\tau^2)$ global error. Combining this with the CFL condition $\tau \sim h$, results in an $\mathcal{O}(h^2)$ splitting error and since with standard spatial discretizations we already have an $\mathcal{O}(h^2)$ spatial error, this splitting error contribution is acceptable in general.

Example 1.1 For a numerical illustration we solve the 2D test problem

$$u_t + (a_1 u)_x + (a_2 u)_y = 0, \quad t > 0, \quad 0 < x, y < 1,$$

with the divergence-free velocity field

$$a_1(x, y) = 2\pi(y - \frac{1}{2}), \quad a_2(x, y) = -2\pi(x - \frac{1}{2}).$$

The characteristics are circles with center $(\frac{1}{2}, \frac{1}{2})$ and the velocity field defines a clockwise rotation with period 1 around the center of the domain. This problem is often used as a test problem in the atmospheric and meteorological literature where it is known as the *Molenkamp-Crowley problem*. In this test we rotate a cylinder and cone of height 1, both with base radius 0.15 and centered initially at $(\frac{1}{2}, \frac{3}{4})$ and $(\frac{1}{2}, \frac{1}{4})$, respectively. As we have seen earlier, these profiles cannot be resolved with standard advection schemes without limiting if wiggles or negative solutions are to be avoided. The commutator

$$[\mathbf{f}_1, \mathbf{f}_2](u) = 4\pi^2 \left((x - \frac{1}{2}) u_x - (y - \frac{1}{2}) u_y \right),$$

is not identically zero, leading to splitting errors.

Tests were performed on a cell-centered 50×50 grid with three advection schemes. Two of them are based on the standard MOL approach without

splitting. These two use the third-order upwind-biased spatial discretization with limiter (III.1.7). They differ in the time stepping formula. The first scheme uses the second-order explicit trapezoidal rule (II.1.6), which we here refer to as RK2. The second uses the second-order extrapolated BDF method (II.3.13), here referred to as EBD2. The third scheme is based on the one-space dimensional, limited DST formula from Section III.2 and uses dimension splitting. As splitting methodology, the basic first-order splitting was used with the sequence

$$u_t^* + (a_1 u^*)_x = 0, \quad u_t^{**} + (a_2 u^{**})_y = 0.$$

Symmetrical splitting would be more natural of course, but for this particular test problem we found that basic first-order time splitting based on the DST formula is already more accurate than the two second-order (in time) MOL schemes. We owe this to the good DST performance and small splitting errors.

At the inflow boundaries fluxes are prescribed. For points adjacent to the outflow boundaries additional points outside the domain are needed because of the third-order upwind-biased discretization. Additional values were provided by constant extrapolation from the interior. The step size τ was chosen as $1/(100k)$, with integer k such that the numerical results for $\tau = 1/(100(k-1))$ were unstable. This resulted in

$$\tau = \frac{1}{400} \text{ for EBD2, } \tau = \frac{1}{300} \text{ for RK2, } \tau = \frac{1}{200} \text{ for DST.}$$

Note that EBD2 uses the maximal 2D Courant number

$$\frac{\tau}{h} \left(|a_1|_{max} + |a_2|_{max} \right) = \frac{1}{4}\pi,$$

which is larger than the previously given theoretical bound 0.46 for stability according to Table II.3.1. We still get a stable (and positive) solution due to the fact that the largest Courant numbers are found at the corners of the domain where boundary conditions are dominant and the solution is spatially constant. Likewise, RK2 uses the maximal 2D Courant number $\pi/3$ which

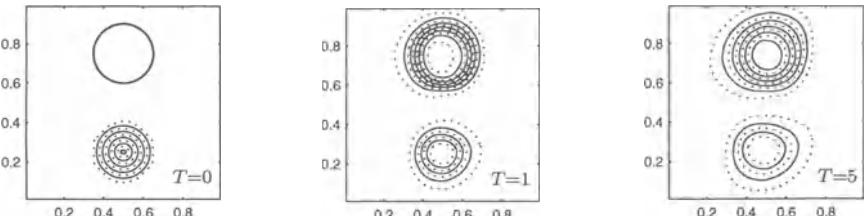


Fig. 1.1. Molenkamp-Crowley test on 50×50 grid. EBD2 contour lines at $T = 0, 1, 5$ for the cylinder and cone. Solid lines at contour values $0.1, 0.3, 0.5, 0.7, 0.9$, dotted lines at $0.2, 0.4, 0.6, 0.8$ and $0.01, 0.99$. Time step $\tau = 1/400$ (with $\tau = 1/300$ instability arises).

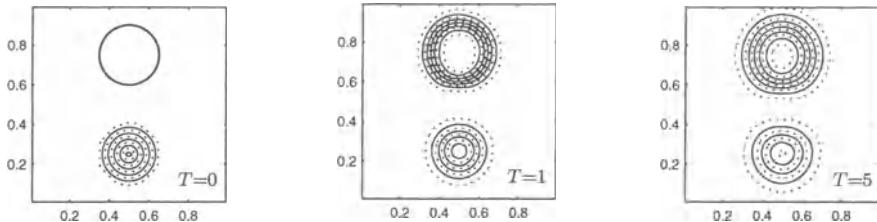


Fig. 1.2. Molenkamp-Crowley test on 50×50 grid. DST contour lines at $T = 0, 1, 5$ for the cylinder and cone. Solid lines at contour values $0.1, 0.3, 0.5, 0.7, 0.9$, dotted lines at $0.2, 0.4, 0.6, 0.8$ and $0.01, 0.99$. Time step $\tau = 1/200$ (with $\tau = 1/100$ instability arises).

is also larger than the theoretical stability bound 0.87 of Table II.1.2. The scheme DST is applied with Courant number

$$\frac{\tau}{h} \max(|a_1|_{\max}, |a_2|_{\max}) = \frac{1}{4}\pi,$$

which is within the range of the theoretical stability bound 1. To make comparison of the results easier, the unconditional stability option (with shifted stencils) for DST was not used.

The numerical solutions are shown as contour plots in Figure 1.1 and 1.2 for EBD2 and DST after 1, 2 and 5 full rotations. The pictures for RK2 were nearly identical to those of EBD2 and are therefore omitted here. Except for some small distortion for EBD2, the pictures for EBD2 and DST look similar, showing good shape preservation and phase speed. Of course, sharp profiles like cylinders and cones are diffused due to upwinding and flux-limiting. The results for DST are somewhat less diffusive than for EBD2 and RK2, which is due to the better properties of the underlying 1D scheme, see Section III.2.

Since the coefficients in the DST scheme depend on the Courant numbers, one step with this scheme requires slightly more work than with EBD2. However, this is more than balanced by the fact that DST allows larger step sizes. In conclusion, for this Molenkamp-Crowley test problem the results for the first-order dimension splitting scheme based on DST are better than for the second-order MOL schemes EBD2 and RK2. Here the main error source lies in the spatial discretization giving artificial diffusion. However, for all three schemes the artificial diffusion decreases upon grid refinement. The same conclusion with respect to DST and MOL schemes was reached with more challenging advection tests in Petersen et al. (1998). \diamond

Although dimension splitting works very well for this Molenkamp-Crowley advection problem, it is not always smooth sailing. An awkward property is that dimension splitting may introduce unwanted *spatial deformations* for advection in conservation form. As pointed out by Bott (1992), this can even happen for spatially constant solutions with a divergence free, variable

velocity field.²⁾ The explanation is simple. Consider $u_t + (a_1 u)_x + (a_2 u)_y = 0$ with $\operatorname{div}(a_1, a_2) = 0$, and suppose we solve the two subproblems exactly from $t = 0$ to $t = \tau$ with first-order splitting. Then the solution at $t = \tau$ of the first subproblem $u_t + (a_1 u)_x = 0$ will vary in space if $(a_1)_x \neq 0$ and so will the final result at $t = \tau$.³⁾ This variation is artificial and may look like a spatial oscillation. In particular, this deficiency may result in qualitatively bad results for problems with deformational flow fields and spatially constant solutions in part of the domain, which are background concentration values in applications.

Example 1.2 We will illustrate this deficiency numerically. For that purpose we consider the deformation test from Smolarkiewicz (1982) with background values; see also Bott (1992). This involves the solution of the 2D advection equation on the square $0 < x, y < 100$ with the flow field described by a stream-function ψ ,

$$a_1 = -\psi_y, \quad a_2 = \psi_x, \quad \psi(x, y) = 8 \sin\left(\frac{1}{25}\pi x\right) \cos\left(\frac{1}{25}\pi y\right),$$

in which a cone with center $(x, y) = (50, 50)$, base radius 15 and height 4 is placed on top of a background concentration equal to 1. The flow field consists of alternating cells of length 25 with clockwise or counter-clockwise rotation. In this flow, the cone is strongly deformed but the background concentration remains 1 since the flow field is divergence free.

In Figure 1.3 the numerical solutions are displayed for first- and second-order splitting after 38 time steps with step size $\tau = 0.7$, similar to Bott (1992). A uniform 100×100 mesh was used. The 1D subproblems were solved with the DST scheme from the previous example.

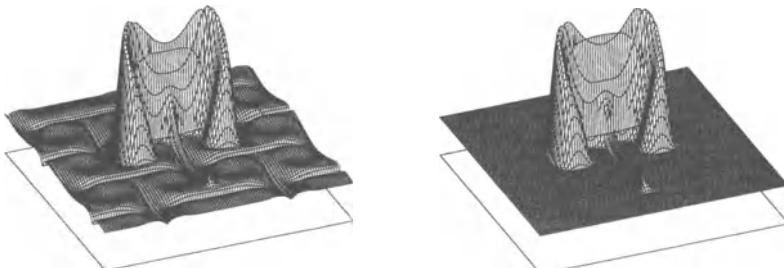


Fig. 1.3. Deformation test with first-order (left) and second-order splitting (right).

It is clear from Figure 1.3 that with first-order splitting not only the cone but also the background concentration is strongly deformed. The result

²⁾ If the velocity field is not divergence free, a spatially constant solution has to evolve in a spatially variable solution anyhow. For advection in the non-conservative form $u_t + \underline{a} \cdot \nabla u = 0$ the following discussion is not relevant.

³⁾ The velocity field of the Molenkamp-Crowley test problem is rather special since $(a_1)_x = (a_2)_y = 0$. Hence for this problem a spatially constant solution is preserved by splitting.

with second-order splitting is satisfactory. Although there still is a slight deformation of the background concentration, this is hardly visible anymore. However, if the background concentration were larger, relative to the height of the cone, some deformation would reappear. \diamond

A modification to diminish this deficiency has been proposed by Russell & Lerner (1981), which is described here for two spatial dimensions. Suppose the problem $u_t + (a_1 u)_x + (a_2 u)_y = 0$ is spatially discretized on a uniform Cartesian grid with grid sizes h_1, h_2 and that the given velocities are made divergence free in the discrete form

$$\frac{1}{h_1} (a_{1,i+\frac{1}{2},j} - a_{1,i-\frac{1}{2},j}) + \frac{1}{h_2} (a_{2,i,j+\frac{1}{2}} - a_{2,i,j-\frac{1}{2}}) = 0.$$

A constant solution field w_{ij}^n at time t_n then should still be constant at t_{n+1} . If we introduce the artificial densities $\rho_{ij}^n \equiv 1$ at time t_n and set

$$\rho_{ij}^* = \rho_{ij}^n + \frac{\tau}{h_1} (a_{1,i+\frac{1}{2},j} - a_{1,i-\frac{1}{2},j}), \quad \rho_{ij}^{n+1} = \rho_{ij}^* + \frac{\tau}{h_2} (a_{2,i,j+\frac{1}{2}} - a_{2,i,j-\frac{1}{2}}),$$

we have $\rho_{ij}^{n+1} \equiv 1$, in spite of the fact that the intermediate results ρ_{ij}^* may give spatial variations. The Russell-Lerner modification now consists of calculating the fluxes not from the approximate concentrations w_{ij}^n and w_{ij}^* , but from the artificial mixing ratios

$$v_{ij}^n = w_{ij}^n / \rho_{ij}^n, \quad v_{ij}^* = w_{ij}^* / \rho_{ij}^*.$$

The resulting scheme is

$$v_{ij}^n = w_{ij}^n / \rho_{ij}^n, \quad w_{ij}^* = w_{ij}^n + \frac{\tau}{h_1} (\bar{f}_{i-\frac{1}{2},j}^n - \bar{f}_{i+\frac{1}{2},j}^n), \\ v_{ij}^* = w_{ij}^* / \rho_{ij}^*, \quad w_{ij}^{n+1} = w_{ij}^* + \frac{\tau}{h_2} (\bar{g}_{i,j-\frac{1}{2}}^* - \bar{g}_{i,j+\frac{1}{2}}^*),$$

with mixing ratio fluxes $\bar{f}_{i+1/2,j}$ and $\bar{g}_{i,j+1/2}$. These fluxes are computed in the same way as the concentration fluxes, for example as in the above DST scheme, except that now the values v_{ij}^n, v_{ij}^* are used instead of w_{ij}^n, w_{ij}^* . With this modification it follows that⁴⁾

$$w_{ij}^n = C \quad \text{for all } i, j \quad \implies \quad w_{ij}^{n+1} = C \quad \text{for all } i, j,$$

a property not shared by the original splitting. Due to the fact that the intermediate quantities w_{ij}^* may be far from equilibrium, the flux computation in the second step may give large errors in the original splitting.

⁴⁾ If $w_{ij}^n \equiv C$, then $w_{ij}^* = C\rho_{ij}^*$ and likewise $w_{ij}^{n+1} = C\rho_{ij}^{n+1} \equiv C$. In meteorological applications genuine densities may be used, but then the mixing ratio fluxes are also multiplied by densities at the cell boundaries.

This modified splitting procedure produced favourable results in Petersen et al. (1998). In that paper also a convergence proof is presented for a scheme based on the donor cell algorithm. Similar splitting modifications with the same objective are found in Lin & Rood (1996) and Leonard, Lock & MacVean (1996).

Finally we note that splitting can also be used for nonlinear conservation laws $u_t + \nabla \cdot \underline{f}(u) = 0$. The standard first-order splitting in 2D, with $\underline{f} = (f, g)^T$, then amounts to solving subsequently

$$u_t^* + f(u^*)_x = 0, \quad u_t^{**} + g(u^{**})_y = 0$$

on each time interval $[t_n, t_{n+1}]$. For smooth solutions we can study the splitting error as in the linear advection case by Taylor expansions, showing second-order convergence for the symmetrical Strang splitting. Solutions of conservation laws are not smooth in general, see Section III.1.4, but theoretical justification for scalar problems with non-smooth solutions was given by Crandall & Majda (1980), who showed that if the sub-steps are solved by a monotone 1D scheme then the 2D splitting scheme converges to the correct weak solution.

Dimension Splitting for Diffusion

Spatial discretization of the d -dimensional diffusion equation

$$u_t = \nabla \cdot (D \nabla u) \quad (1.31)$$

results in a very large stiff ODE system. Application of a standard implicit Runge-Kutta or linear multistep method gives rise to huge systems of algebraic equations. Here dimension splitting can be used to break down the problem into easier 1D parts. This was the original motivation in many classical splitting schemes such as the ADI scheme of Peaceman & Rachford (1955), which will be discussed in following sections.

For example, in 3D with diagonal matrix D we have

$$u_t = (d_1 u_x)_x + (d_2 u_y)_y + (d_3 u_z)_z, \quad (1.32)$$

and with dimension splitting we sequentially solve 1D parabolic problems in a 3D domain. With standard second-order finite differences in space and an implicit integration method at each sub-step, we only have to solve tridiagonal systems of algebraic equations, one for each grid line. This is cheap since the solution costs for tridiagonal matrices are very low (linearly proportional to the dimension) and there are good possibilities for parallel computing.

Splitting diffusion terms normally gives a splitting error. With $\mathbf{f}_1(u) = (d_1 u_x)_x$ and $\mathbf{f}_2(u) = (d_2 u_y)_y$, the commutator (1.20) reads

$$[\mathbf{f}_1, \mathbf{f}_2](u) = (d_1(d_2 u_y)_{xy})_x - (d_2(d_1 u_x)_{xy})_y$$

which vanishes if $(d_1)_y = (d_2)_x = 0$ (conditions for a zero splitting error in three dimensions easily follow). In general these conditions are of course not satisfied, leaving splitting errors.

In spite of the splitting errors, dimension splitting for diffusion equations often leads to very economical schemes due to the 1D nature of the computations. With symmetrical splitting these errors are $\mathcal{O}(\tau^2)$, hence asymptotically of the same size as the spatial errors when second-order spatial differencing is used. It should be noted here that these asymptotic considerations are only justified if the commutator is of modest size. As we will see later the influence of boundary conditions may play a major role here.

Also note that dimension splitting is not always applicable. The spatial domain and the grid imposed on it should be such that the splitting into 1D problems is possible at the semi-discrete level. Complicated domains requiring complicated non-uniform grids prevent an easy use of dimension splitting. Furthermore, the operator itself complicates splitting if D is not a diagonal matrix giving cross diffusion terms (mixed space derivatives). For example, in two space dimensions with

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix},$$

the general form (1.31) reads

$$u_t = (d_{11}u_x)_x + (d_{12}u_y)_x + (d_{12}u_x)_y + (d_{22}u_y)_y.$$

Obviously, the common splitting into 1D problems is no longer possible. For this case special splitting schemes have been developed which treat the mixed terms explicitly in time; see Mitchell & Griffiths (1980, Chap. 2) for further references.

1.7 Boundary Values and Stiff Terms

For the analysis of splitting errors it was assumed thus far that the commutators are of moderate size. This however is not justified with inhomogeneous boundary conditions or with stiffness in reaction terms. Both problems are briefly discussed here. We note that inhomogeneous boundary conditions will give an inhomogeneous term in the semi-discrete system that is multiplied by a negative power of the mesh width, so this may act as a stiff forcing term.

Boundary Values

Major difficulties with splitting methods occur for problems where the boundary conditions are important. If we consider a PDE problem with boundary conditions, then these are physical conditions for the whole process and boundary conditions for the sub-steps (which may have little physical meaning) are missing. Therefore one may have to reconstruct boundary conditions for the specific splitting under consideration.

For example, consider a linear semi-discrete problem $w'(t) = Aw(t) + g(t)$, where $g(t)$ contains the given boundary conditions. Suppose that

$$Av + g(t) = (A_1 v + g_1(t)) + (A_2 v + g_2(t)) \quad (1.33)$$

with $g_k(t)$ containing the boundary conditions relevant to A_k . The exact solution satisfies

$$w(t_{n+1}) = e^{\tau A} w(t_n) + \int_0^\tau e^{(\tau-s)A} g(t_n + s) ds.$$

If we consider first-order splitting, with inhomogeneous terms \tilde{g}_1, \tilde{g}_2 , then

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n + e^{\tau A_2} \int_0^\tau e^{(\tau-s)A_1} \tilde{g}_1(t_n + s) ds + \int_0^\tau e^{(\tau-s)A_2} \tilde{g}_2(t_n + s) ds.$$

Even with commuting matrices, $A_1 A_2 = A_2 A_1$, and constant boundary terms, we will get a splitting error if we take $\tilde{g}_k = g_k$. An exact formula for commuting matrices is obtained by choosing

$$\tilde{g}_1(t_n + s) = e^{-sA_2} g_1(t_n + s), \quad \tilde{g}_2(t_n + s) = e^{(\tau-s)A_1} g_2(t_n + s). \quad (1.34)$$

Note that this correction for g_1 requires a *backward* time integration with A_2 , which will not always be feasible. With an implicit ODE method the implicit algebraic relations need no longer be well defined with negative step size.

A general analysis of boundary conditions for splitting methods is, at present, still lacking. As a rule of thumb it can be said that the treatment of the boundaries should coincide as much as possible with the scheme in the interior of the domain. This will be illustrated by the following example.

Example 1.3 Consider the model advection-reaction equation

$$u_t + u_x = u^2, \quad 0 < t \leq \frac{1}{2}, \quad 0 < x < 1,$$

from Section II.2.1, with given initial value at $t = 0$ and Dirichlet condition at $x = 0$ derived from the smooth exact solution

$$u(x, t) = \frac{\sin^2(\pi(x - t))}{1 - t \sin^2(\pi(x - t))}.$$

Since the equation is scalar and the nonlinear term is non-stiff, splitting is of course not needed from a practical point of view. However, this simple example is theoretically interesting as it clearly illustrates the unwanted influence of the boundary condition. Note that u_x and u^2 commute (see Section 1.5), so that without boundary conditions operator splitting would be exact.

As in Section II.2.1, spatial discretization is performed with fourth-order central differences in the interior and third-order one-sided approximations

at the boundaries, using a uniform grid with grid size h . This gives a semi-discrete system which we solve in time by three splittings, where the advection step is done with the classical fourth-order explicit Runge-Kutta method with Courant number $\tau/h = 2$ and the reaction step $u_t = u^2$ is solved exactly.

The three splitting processes are:

- (i) first-order (basic) splitting with reaction followed by advection, using in the advection step the given Dirichlet boundary value;
- (ii) second-order symmetrical splitting, again using in the advection step the given Dirichlet boundary value;
- (iii) first-order splitting as in (i), but now with a corrected boundary condition: before each advection sub-step the given boundary value $\gamma_0(t) = u(0, t)$ is corrected to

$$\tilde{\gamma}_0(t) = \frac{u(0, t)}{1 - (t_{n+1} - t) u(0, t)} \quad \text{for } t \in [t_n, t_{n+1}].$$

Hence we invoke the exact solution of the reaction $u_t = u^2$ as input at the boundary point for the advection step over $[t_n, t_{n+1}]$, similar to the linear case (1.34).⁵⁾ In doing so we treat the boundary point in the same way as the interior points of the grid. This is in fact just what we would do with the method of characteristics, which is exact here since the reaction step is solved exactly. So with this boundary correction we eliminate the splitting error incurred by the Dirichlet boundary condition.

τ	Basic splitting		Symmetrical splitting		Corrected boundary	
	L_2 -error	order	L_2 -error	order	L_2 -error	order
$\frac{1}{20}$	$0.26 \cdot 10^{-1}$		$0.14 \cdot 10^{-1}$		$0.88 \cdot 10^{-3}$	
$\frac{1}{40}$	$0.14 \cdot 10^{-1}$	0.94	$0.48 \cdot 10^{-2}$	1.58	$0.91 \cdot 10^{-4}$	3.27
$\frac{1}{80}$	$0.72 \cdot 10^{-2}$	0.96	$0.17 \cdot 10^{-2}$	1.54	$0.13 \cdot 10^{-4}$	2.80
$\frac{1}{160}$	$0.36 \cdot 10^{-2}$	0.98	$0.58 \cdot 10^{-3}$	1.52	$0.22 \cdot 10^{-5}$	2.57

Table 1.1. L_2 -errors and estimated orders for the advection-reaction model.

The relative errors at $t = \frac{1}{2}$ in the L_2 -norm, together with the estimated orders of convergence, are given in Table 1.1. Because the spatial and temporal advection errors are sufficiently small, basic splitting can be seen to converge with its order one. On the other hand, symmetrical splitting shows

⁵⁾ If we consider splitting with advection followed by reaction, then the splitting error can be avoided by a backward reaction solution of the boundary data as indicated by (1.34).

some order reduction and is less accurate than simple splitting with boundary correction (which has no splitting error). The error for the latter is dominated by the error of the Runge-Kutta method. For this method the actual convergence rate is less than four due to order reduction of the Runge-Kutta method, see the first column of Table II.2.2 for the same problem \diamond

The specific LOD and ADI methods discussed in the following sections experience the same difficulties with boundary values. For these methods certain correction techniques have been proposed; we will pay attention to such techniques in Section 2.4.

Stiff Terms

The order reduction for the Strang splitting in the above example, due to boundary conditions, shows that the accuracy is affected by large commutators.⁶⁾ Apart from diffusion, this may also happen by inclusion of stiff ODE terms, for instance originating from chemical reactions.

For a class of stiff linear ODE problems $w'(t) = (A_1 + A_2)w(t)$, with $\|A_2\|$ bounded but A_1 having some eigenvalues proportional to $-1/\epsilon$ with $\epsilon > 0$, $\tau/\epsilon \gg 1$, it was shown in Verwer & Sportisse (1998) and Sportisse (2000) that the simple first-order splitting will retain its first-order accuracy, but also the second-order Strang splitting will in general give only an order one accuracy. In these results the norm of the commutator $\|[A_1, A_2]\|$ was of the same size as $\|A_1\| \|A_2\|$. For many applications the norm of the commutator will be smaller.

Fairly general linear convergence results with order two for Strang splitting were obtained by Jahnke & Lubich (2000). In their analysis it was basically assumed that the commutator of A_1, A_2 can be bounded in a suitable manner by fractional powers of A_1 . The assumptions were shown to hold for a linear Schrödinger equation. Generalizations for the nonlinear Schrödinger equation (I.8.2) can be found in Besse, Bidégaray & Descombes (2002).

An analysis for the simple first-order splitting with nonlinear hyperbolic equations and stiff reactions has been given by Tang (1998), showing first-order convergence independent of the stiffness of the reactions. Modified procedures with applications to combustion problems are found for example in Helzel, LeVeque & Warnecke (2000).

⁶⁾ The commutators in this section have been defined only for autonomous problems, see (1.20). For linear problems with inhomogeneous, constant boundary conditions we get a semi-discrete system with $F_j(v) = A_j v + b_j$, where b_j contains the boundary data relevant to A_j , and then the commutator of F_1 and F_2 is given by

$$[F_1, F_2]v = [A_1, A_2]v + A_1 b_2 - A_2 b_1.$$

Even if $[A_1, A_2] = 0$, the commutator of F_1 and F_2 may be large. In general it will be zero only if certain compatibility conditions are satisfied which are not related to the smoothness of the solution.

Error bounds for splittings using backward Euler or trapezoidal rule approximations in the sub-steps will be examined more closely in the next sections. Particular attention will be paid to the order reduction caused by boundary conditions since that is often the main reason for a disappointing convergence behaviour with splitting methods.

2 LOD Methods

With operator splitting the original problem is decomposed into more manageable sub-steps. Following Yanenko (1971), these are also often called *fractional steps*. For the actual solution of these fractional steps some appropriate numerical integration method is needed. Two main approaches can be distinguished.

In the first approach, stable and accurate solvers based on one-step Runge-Kutta or Rosenbrock type methods are applied over the current split interval of length τ , possibly with a (variable) sub-step size smaller than τ to ensure that the integration errors are sufficiently small in comparison with the splitting errors. This is the approach which we advocate for general problems. It offers the possibility to choose a suitable implicit or explicit method for each of the fractional steps.

A more traditional approach is to select a fixed low-order one-step method and apply it with the same step size τ to obtain a specific splitting method. For multi-dimensional PDEs these methods are often based on dimension splitting, where the splitting is such that all computations become effectively one-dimensional. For this reason such methods are also known as *locally one-dimensional* (LOD) methods; in the following we will use the name LOD also with more general splittings for advection, diffusion and reactions.

This section is devoted to these classical LOD methods (not restricted, however, to dimension splitting). The first methods of this type were developed in the 1950's and 60's, mainly by numerical analysts from the Soviet Union. Some of the well-known contributors are D'Yakonov, Godunov, Marchuk, Samarskii and Yanenko, see for example Samarskii (1962), Yanenko (1971), Marchuk (1981, 1990) and references therein.⁷⁾ Their applications came from various fields from mathematical physics.

2.1 The LOD-Backward Euler Method

Again our starting point is the nonlinear semi-discrete ODE system $w'(t) = F(t, w(t))$ in \mathbb{R}^m for which we suppose the multiple splitting

$$F(t, v) = F_1(t, v) + F_2(t, v) + \cdots + F_s(t, v).$$

⁷⁾ The spelling of Russian names may vary in German, French or English translations.

The first-order operator splitting from Section 1.1, combined with the first-order backward Euler method for all the fractional steps, gives the first-order LOD-backward Euler (LOD-BE) method

$$\begin{aligned} v_0 &= w_n, \\ v_i &= v_{i-1} + \tau F_i(t_{n+1}, v_i), \quad i = 1, \dots, s, \\ w_{n+1} &= v_s, \end{aligned} \tag{2.1}$$

where v_1, \dots, v_{s-1} are internal vectors for the step from t_n to t_{n+1} . Order is here understood to be order of consistency in the classical sense with respect to the solution of the ODE problem on a fixed spatial grid, not with respect to the underlying PDE solution; such convergence will be studied more closely in following sections. Observe, in this connection, that none of the intermediate vectors v_i is a consistent approximation to the exact solution, which is typical for methods based on operator splitting. A consequence is that unlike standard ODE methods the method does not exactly return steady-state solutions \bar{w} ,

$$F(\bar{w}) = 0, \tag{2.2}$$

for autonomous problems. Putting $w_n = \bar{w}$ in (2.1) yields $w_{n+1} = \bar{w}$ only if \bar{w} is a steady state for each of the subsystems which is normally not true. Therefore the LOD method (2.1) is not a recommended method to march towards a steady state.

Example 2.1 If F represents a semi-discrete two-dimensional PDE problem and F_1 and F_2 are discretizations of operators acting in the x - and y -direction, we have an example of dimension splitting with $s = 2$. The classical example is the linear heat flow model (III.6.8), leading to a semi-discrete equation with $F_i(t, v) = A_i v + g_i(t)$, $i = 1, 2$, where the matrices A_1, A_2 are the finite difference approximations for ∂_{xx} and ∂_{yy} , respectively, and g_1, g_2 contain sources and boundary data. In terms of the usual one-dimensional difference matrices B_m we can write $A_1 = I_{m_2} \otimes B_{m_1}$, $A_2 = B_{m_2} \otimes I_{m_1}$ on an $m_1 \times m_2$ Cartesian grid, see (III.6.10). Method (2.1) then yields

$$\begin{aligned} w_{n+1}^* &= w_n + \tau A_1 w_{n+1}^* + \tau g_1(t_{n+1}), \\ w_{n+1} &= w_{n+1}^* + \tau A_2 w_{n+1}^* + \tau g_2(t_{n+1}). \end{aligned}$$

The computation of the first stage amounts to solving a linear system of algebraic equations with the matrix

$$I - \tau A_1 = I - \tau(I_{m_2} \otimes B_{m_1}) = I_{m_2} \otimes (I_{m_1} - \tau B_{m_1}),$$

which is composed of the m_2 decoupled tridiagonal matrices $I_{m_1} - \tau B_{m_1}$ of dimension m_1 , each of which is associated with a horizontal grid line. For the second stage computation we have exactly the same situation along the

y -direction. Hence by dimension splitting only one-space dimensional computations are performed, and this was the original motivation for considering splitting methods. Nowadays, solving simple linear algebra problems in two dimensions of the above type is no longer considered a challenging problem, but still in practice such splittings become interesting for more complicated problems where solution of the full multi-dimensional implicit systems in each time step can be time consuming. \diamond

Stability

In practice the LOD-BE method is stable for large classes of problems. It lends itself easily to the stability analysis of Section I.2. Error bounds given there can be used one after another for each of the stages of the splitting method.

The most simple situation from the analysis point of view is a linear problem with $F_i(t, v) = A_i v + g_i(t)$, where all A_i can be diagonalized with one and the same well conditioned eigensystem, $A_i = U \Lambda_i U^{-1}$, which implies commutativity of the matrices. This holds for constant coefficient linear PDE problems provided with periodic boundary conditions and spatially discretized on uniform grids with standard linear differences, where U will contain the multi-dimensional Fourier modes, see Section III.6. For stability considerations the inhomogeneous terms may be discarded. By a Fourier transformation it then suffices to consider the scalar test problem

$$w'(t) = \lambda_1 w(t) + \cdots + \lambda_s w(t), \quad (2.3)$$

where $\lambda_i \in \mathbb{C}$ represents an eigenvalue of A_i .

Let $z_i = \tau \lambda_i$, $i = 1, \dots, s$. Applying the s -stage LOD-BE method to (2.3) yields a recursion $w_{n+1} = R w_n$ with

$$R(z_1, \dots, z_s) = \prod_{i=1}^s (1 - z_i)^{-1}$$

as *stability function*. Obviously, stability for the test problem (2.3) requires $|R| \leq 1$ and because R is completely factorized we have stability if each of the factors has modulus at most 1.

More general results can be proved by using the nonlinear stability results available for the backward Euler method. For a nonlinear problem with splitting into terms $F_i(t, v)$, consider the Jacobian matrix $J_i(t, v) = (\partial F_i(t, v) / \partial v)$. Assume the logarithmic norm inequality (I.2.56),

$$\mu(J_i(t, v)) \leq \omega_i \quad \text{for all } t \geq 0, v \in \mathbb{R}^m,$$

is valid with constant ω_i for some suitable norm $\|\cdot\|$, and let \tilde{w}_n be a perturbation to w_n . The factorized nature of the s -stage LOD-BE method then

directly leads to the general stability inequality

$$\|w_{n+1} - \tilde{w}_{n+1}\| \leq \prod_{i=1}^s (1 - \tau\omega_i)^{-1} \|w_n - \tilde{w}_n\|, \quad (2.4)$$

provided $1 - \tau\omega_i > 0$ (Verwer, 1984).

As pointed out in Section I.4, with standard discretizations of order one and two, the logarithmic norm inequality is often easily verified for L_p -norms. For example, this is possible for second-order central spatial discretization of the scalar diffusion-reaction equation $u_t = \nabla \cdot (D\nabla u) + f(u)$ with D diagonal, positive definite and $f_u \leq \nu$ for some constant ν . It is this type of equation for which the LOD-BE method was originally developed. For advection-dominated problems the damping properties of the backward Euler steps are in general too strong.

2.2 LOD Crank-Nicolson Methods

By combining first-order operator splitting with the second-order implicit trapezoidal rule (Crank-Nicolson) for all the fractional steps, we obtain Yannenko's LOD Crank-Nicolson (LOD-CN) method, with $i = 1, 2, \dots, s$,

$$\begin{aligned} v_0 &= w_n, \\ v_i &= v_{i-1} + \frac{1}{2}\tau F_i(t_n + c_{i-1}\tau, v_{i-1}) + \frac{1}{2}\tau F_i(t_n + c_i\tau, v_i), \\ w_{n+1} &= v_s. \end{aligned} \quad (2.5)$$

We use the time levels $c_0 = 0, c_s = 1$. The other c_j are set to $\frac{1}{2}$, which is somewhat arbitrary since the vectors v_j are not consistent approximations to $w(t_n + c_j\tau)$. This method has also first-order consistency; it becomes second order if F_1, F_2 are homogeneous linear and commuting. Instead of the trapezoidal rule, integration of the fractional steps can also be based on the implicit midpoint rule, leading to a scheme with similar properties.

If we denote (2.5) as

$$w_{n+1} = S_{\tau;1,2,\dots,s}(t_n, w_n),$$

with the indices $1, 2, \dots, s$ denoting the sequence of F_1, F_2, \dots, F_s , then a symmetrical version, related to the Strang splitting (1.10), is obtained with

$$w_{n+\frac{1}{2}} = S_{\frac{1}{2}\tau;1,\dots,s}(t_n, w_n), \quad w_{n+1} = S_{\frac{1}{2}\tau;s,\dots,1}(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}). \quad (2.6)$$

Irrespective of the choice of the time levels c_j , this method is symmetrical and of order two in the classical ODE sense. This method therefore seems superior to (2.5). However, below it will be shown that this is not true, at least in general, due to order reduction in the presence of boundary conditions. In the same way one can construct a method using the implicit midpoint rule

in each of the fractional steps, which will lead again to a method with a very similar behaviour. The linear version of this symmetrical method has been proposed by Marchuk (1971).

Note that in terms of computational effort method (2.6) is twice as expensive as (2.5) for the step from t_n to t_{n+1} . However, if we double the step size for (2.6) and assume stepping from t_n to t_{n+2} , the two methods have an equal expense per time unit.

Second order in the classical ODE sense can also be obtained by using the parallel operator splitting method (1.12). The method

$$w_{n+1} = \frac{1}{2} S_{\tau;1,\dots,s}(t_n, w_n) + \frac{1}{2} S_{\tau;s,\dots,1}(t_n, w_n), \quad (2.7)$$

was introduced by Swayne (1987) for linear problems with $s = 2$. As we will see, given boundary data are more easily incorporated in this method than in (2.5) and (2.6).

Stability

The linear stability of the three methods (2.5), (2.6), (2.7) can be studied in the same way as for the LOD-BE method using again ideas and results from Section I.2. The only fundamental difference is that with these methods we have to restrict ourselves to inner product norms while for the LOD-BE method also the L_1 - and L_∞ -norm can be considered. The same holds for versions based on the implicit midpoint rule with regard to nonlinear stability using the known results for this method. Further we note that already the early literature mentioned in the beginning of this section contained many stability results, see for instance Samarskii (1962), Marchuk (1968, 1971).

In view of the following discussion on convergence, we elaborate the L_2 -stability here for linear systems with a 2-term splitting

$$F_i(t, v) = A_i v + g_i(t), \quad i = 1, 2, \quad (2.8)$$

and with internal perturbations on the stages. The basic LOD-CN method (2.5) can be written for $s = 2$ as

$$\begin{aligned} w_{n+\frac{1}{2}} &= w_n + \frac{1}{2}\tau F_1(t_n, w_n) + \frac{1}{2}\tau F_1(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}), \\ w_{n+1} &= w_{n+\frac{1}{2}} + \frac{1}{2}\tau F_2(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}) + \frac{1}{2}\tau F_2(t_{n+1}, w_{n+1}), \end{aligned} \quad (2.9)$$

where $w_{n+1/2}$ denotes the internal vector. We consider also a perturbed version

$$\begin{aligned} \tilde{w}_{n+\frac{1}{2}} &= \tilde{w}_n + \frac{1}{2}\tau F_1(t_n, \tilde{w}_n) + \frac{1}{2}\tau F_1(t_{n+\frac{1}{2}}, \tilde{w}_{n+\frac{1}{2}}) + r_{n+\frac{1}{2}}, \\ \tilde{w}_{n+1} &= \tilde{w}_{n+\frac{1}{2}} + \frac{1}{2}\tau F_2(t_{n+\frac{1}{2}}, \tilde{w}_{n+\frac{1}{2}}) + \frac{1}{2}\tau F_2(t_{n+1}, \tilde{w}_{n+1}) + r_{n+1}, \end{aligned} \quad (2.10)$$

with perturbations r_ℓ ($\ell = n + \frac{1}{2}, n + 1$).⁸⁾ Let also $\varepsilon_\ell = \tilde{w}_\ell - w_\ell$. Further we will use the notations

$$Z_i = \tau A_i, \quad P_i = I + \frac{1}{2}\tau A_i, \quad Q_i = I - \frac{1}{2}\tau A_i. \quad (2.11)$$

By subtracting (2.9) from (2.10), using linearity and eliminating the intermediate difference $\varepsilon_{n+1/2}$, it follows that

$$\varepsilon_{n+1} = R\varepsilon_n + \delta_n \quad (2.12)$$

with stability matrix

$$R = Q_2^{-1}P_2Q_1^{-1}P_1, \quad (2.13)$$

and with δ_n containing the perturbations,

$$\delta_n = Q_2^{-1}P_2Q_1^{-1}r_{n+\frac{1}{2}} + Q_2^{-1}r_{n+1}. \quad (2.14)$$

Stability is considered for the discrete L_2 -norm under the assumption

$$v^T A_i v \leq 0 \quad \text{for all } v \in \mathbb{R}^m. \quad (2.15)$$

According to Theorem I.2.11 for inner product norms, this implies

$$\|Q_i^{-1}\| \leq 1, \quad \|Q_i^{-1}P_i\| \leq 1 \quad (2.16)$$

for $i = 1, 2$ with arbitrary $\tau > 0$. Note that (2.15) imposes no restriction on the norm $\|A_i\|$. Hence the matrices A_i may contain negative powers of the mesh width $h > 0$. Likewise the A_i may stand for linearized reaction terms with arbitrary stiffness.

Collecting (2.12)–(2.16) thus gives the unconditional L_2 -stability result

$$\|R\| \leq 1$$

and

$$\|\varepsilon_{n+1}\| \leq \|\varepsilon_n\| + \|r_{n+\frac{1}{2}}\| + \|r_{n+1}\|.$$

It is obvious that similar results can be derived for the forms (2.6) and (2.7). Below these stability estimates will be used to obtain error bounds for the three LOD-CN methods.

Convergence and Order Reduction

The consistency orders p mentioned above are all orders in the classical ODE sense and not in the PDE sense. To study consistency and convergence with respect to $u_h(t)$, the PDE solution u restricted to the grid, we let τ and h tend to zero simultaneously and preferably independent of each other for

⁸⁾ In the convergence analysis we will sometimes use the stage perturbations in the form $r_\ell = \tau \rho_\ell$ where the ρ_ℓ can be viewed as residual truncation errors per stage.

unconditionally stable methods. The main objective of such an analysis is to establish the temporal order of convergence r in the error bound

$$\|u_h(t_n) - w_n\| \leq C_1 \tau^r + C_2 h^q \quad (\tau, h > 0, 0 \leq t_n \leq T) \quad (2.17)$$

with constants C_1, C_2 independent of τ, h , and with q denoting the spatial order of accuracy. Such an analysis is reminiscent of the analysis in Section II.2 for Runge-Kutta methods. In particular, we also encounter *order reduction* giving $r < p$ in the presence of boundary conditions.

For $s = 2$ an error analysis will be given for the three LOD Crank-Nicolson methods (2.5), (2.6), (2.7) aiming at the bound (2.17). Similar results are also valid for the methods based on the implicit midpoint rule. The error bounds will be elaborated in detail for the linear 2D heat flow problem with Dirichlet boundary conditions,

$$\begin{aligned} u_t &= u_{xx} + u_{yy} + f(x, y, t) && \text{on } \Omega = (0, 1)^2, \quad 0 < t \leq T, \\ u(x, y, t) &= u_\Gamma(x, y, t) && \text{on } \Gamma = \partial\Omega, \quad 0 \leq t \leq T, \\ u(x, y, 0) &= u_0(x, y) && \text{on } \Omega, \end{aligned} \quad (2.18)$$

using standard second-order finite differences on a uniform grid with mesh width h in both directions. In spite of its simplicity, the heat flow model problem reveals the essence of order reduction and the role of boundary conditions. For this model problem the effect of $h \rightarrow 0$ on the orders of consistency and convergence in time is summarized in Table 2.1. Recall that consistency of order r means that the local error, which is introduced in one step, is $\mathcal{O}(\tau^{1+r})$.

	Order of consistency			Order of convergence		
method	(2.5)	(2.6)	(2.7)	(2.5)	(2.6)	(2.7)
$h = h_0$	1	2	2	1	2	2
$h \rightarrow 0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{5}{4}$	1	$[\frac{1}{4}, \frac{1}{2}]$	2

Table 2.1. Temporal orders of consistency and convergence in the L_2 -norm for the LOD Crank-Nicolson methods with the linear 2D heat flow model problem (2.18).

These results are valid for the L_2 -norm. Only for fixed $h = h_0 > 0$ the classical ODE theory can be seen to apply. For $h \rightarrow 0$ a reduction in the local error is observed for all three methods. As we will see below, for methods (2.5), (2.7) there is no global reduction thanks to a favourable local error cancellation. For the global error and $h \rightarrow 0$, only method (2.6) suffers from reduction with an order between $\frac{1}{4}$ and $\frac{1}{2}$. This is remarkable in view of the

fact that the classical order p of this symmetrical method is two. It is caused by the fact that with this method there is no favourable cancellation of local errors.

To derive these error bounds we will first derive suitable expressions for the local discretization errors, and then consider the global errors.

Local Error Analysis (2D)

We proceed with the derivation of bounds for local discretization errors measured with respect to the PDE solution $u_h(t)$ restricted to the grid. The approach taken here is the same as with the Runge-Kutta methods in Section II.2. So we focus on the temporal order under simultaneous space-time refinement. For simplicity of presentation we assume a zero spatial error σ_h . Including this error in the analysis is always possible, as has been illustrated in Section II.2.4.

The error bounds will be based on derivatives of

$$u_h(t) \quad \text{and} \quad \psi(t) = F_2(t, u_h(t)) - F_1(t, u_h(t)).$$

If the PDE solution is smooth we may assume that these derivatives are bounded uniformly in the mesh width h . As in Section II.2.1 we will use the notation $\mathcal{O}(\tau^p)$ to denote a vector or matrix whose L_2 -norm is bounded by $C\tau^p$ with constant $C > 0$ independent of h . Notice that we do *not* have $A_i = \mathcal{O}(1)$ if A_i contains discretized spatial derivatives, such as for the 2D heat flow model.

First consider the basic LOD-CN method (2.5) in the form (2.9) for $s = 2$. Suitable expressions for the local discretization errors can be easily derived using the perturbed version (2.10) with internal perturbations. We take $\tilde{w}_\ell = u_h(t_\ell)$ ($\ell = n, n + \frac{1}{2}$) for all integers n ,⁹⁾ so that ε_n becomes the global discretization error at time $t = t_n$,

$$\varepsilon_n = u_h(t_n) - w_n,$$

and δ_n the local discretization error, that is, the error introduced in one single step of the method. With this choice of \tilde{w}_ℓ the corresponding residuals are easily found by Taylor expansions,

$$\begin{aligned} r_{n+\frac{1}{2}} &= \frac{1}{2}\tau\psi(t_{n+\frac{1}{2}}) - \frac{1}{8}\tau^2\psi'(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3), \\ r_{n+1} &= -\frac{1}{2}\tau\psi(t_{n+\frac{1}{2}}) - \frac{1}{8}\tau^2\psi'(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3). \end{aligned}$$

The local discretization error of the basic method (2.9) is given by (2.14). If the matrices A_1 and A_2 commute, this can be further elaborated to

⁹⁾ The choice $\tilde{w}_n = u_h(t_n)$ is natural: we want to find expressions for $u_h(t_n) - w_n$, $n \in \mathbb{N}$. On the other hand, taking $\tilde{w}_{n+1/2} = u_h(t_{n+1/2})$ may be considered as somewhat arbitrary, but the global recursion $\varepsilon_{n+1} = R\varepsilon_n + \delta_n$ will not be influenced by this choice.

$$\begin{aligned}\delta_n &= Q_2^{-1}Q_1^{-1}\left((I + \frac{1}{2}Z_2)r_{n+\frac{1}{2}} + (I - \frac{1}{2}Z_1)r_{n+1}\right) \\ &= \frac{1}{4}\tau Q_2^{-1}Q_1^{-1}\left(Z\psi(t_{n+\frac{1}{2}}) - (I - \frac{1}{4}(Z_1 - Z_2))\tau\psi'(t_{n+\frac{1}{2}})\right) + \mathcal{O}(\tau^3)\end{aligned}\quad (2.19)$$

with $Z = \tau A = Z_1 + Z_2$. In this subsection commutation of A_1 and A_2 is assumed; this is a valid assumption for the heat equation model (2.18). Due to boundary conditions this does not lead to a vanishing splitting error. In the non-stiff case, where $A_i = \mathcal{O}(1)$, we would get the expected estimate $\delta_n = \mathcal{O}(\tau^2)$, showing first-order consistency, because of the factor τ in $Z = \tau A$. However, if we merely assume boundedness of the rational expressions in (2.16) we get $\delta_n = \mathcal{O}(\tau)$ only. In general, this bound can be somewhat improved by a more detailed analysis.

For the heat equation with Dirichlet conditions we will have $A_j^\alpha\psi = \mathcal{O}(1)$, $j = 1, 2$ for $\alpha < \frac{1}{4}$, see Lemma III.6.5. Hence

$$\delta_n = \frac{1}{4}\tau^{1+\alpha} \sum_{j=1}^2 Q_2^{-1}Q_1^{-1}Z_j^{1-\alpha}[A_j^\alpha\psi(t_{n+\frac{1}{2}})] + \mathcal{O}(\tau^2).$$

The matrices $Q_2^{-1}Q_1^{-1}Z_j^{1-\alpha}$ are bounded, because they are normal with eigenvalues $(1 - \frac{1}{2}z_2)^{-1}(1 - \frac{1}{2}z_1)^{-1}z_j^{1-\alpha}$, $z_j < 0$, which leads to the estimate $\delta_n = \mathcal{O}(\tau^{1+\alpha})$. Hence we have consistency with order α . Since we can take any $\alpha < \frac{1}{4}$, this will be simply referred to as consistency with order $\frac{1}{4}$.¹⁰⁾

The local errors for the parallel LOD-CN method (2.7) are easily found as follows: let $\bar{\delta}_n$ be the same local error as (2.19) but with the indices 1,2 interchanged (this also implies that ψ is replaced by $-\psi$). Since we assume that A_1 and A_2 commute, we now get for (2.7) the global error recursion

$$\varepsilon_{n+1} = R\varepsilon_n + \delta_n^*, \quad \delta_n^* = \frac{1}{2}(\delta_n + \bar{\delta}_n) \quad (2.20)$$

with local error

$$\delta_n^* = \frac{1}{16}\tau^2 Q_2^{-1}Q_1^{-1}(Z_1 - Z_2)\psi'(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3). \quad (2.21)$$

Hence for the non-stiff case the standard estimate $\delta_n = \mathcal{O}(\tau^3)$ for the parallel method (2.7) is retrieved, whereas for the heat equation with Dirichlet conditions we get $\delta_n = \mathcal{O}(\tau^{2+\alpha})$ for $\alpha < \frac{1}{4}$, that is, consistency with order $\frac{5}{4}$.

For the symmetrical method (2.6) we can directly use the above derivations if we apply the scheme with step size 2τ , taking w_n to w_{n+2} (with w_{n+1} as intermediate result). This is merely for notational convenience. We then get

$$\varepsilon_{n+2} = R^2\varepsilon_n + \delta_n^{**}, \quad \delta_n^{**} = R\delta_n + \bar{\delta}_{n+1}. \quad (2.22)$$

¹⁰⁾ Recall from Lemma III.6.5 that for $\alpha = \frac{1}{4}$ precisely, an extra $\log(h)$ term should be included in the bound for $\|A_j^\alpha\psi\|$.

Since $\|R\| \leq 1$ we have $\|\delta_n^{**}\| \leq \|\delta_n\| + \|\bar{\delta}_{n+1}\|$. It follows that the order of consistency of (2.6) is at least equal to that of (2.5) for $s = 2$. However, the more interesting point is that for stiff situations it is in general not better either. For smooth solutions we will have $\bar{\delta}_{n+1} = -\delta_n(1 + \mathcal{O}(\tau))$ and hence¹¹⁾

$$\delta_n^{**} = (R - I)\delta_n(1 + \mathcal{O}(\tau)) = \frac{1}{4}\tau Q_2^{-2}Q_1^{-2}Z^2(\psi(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau)). \quad (2.23)$$

In the non-stiff ODE case this would yield $\delta_n^{**} = \mathcal{O}(\tau^3)$, but for our 2D heat equation model we merely get $\delta_n^{**} = \mathcal{O}(\tau^{1+\alpha})$ for $\alpha < \frac{1}{4}$, which gives the corresponding entry $\frac{1}{4}$ in Table 2.1 for the order of consistency.

Global Error Analysis (2D)

To derive global error bounds we will use again Lemma II.2.3 to examine favourable local error cancellations. Recall that with a stable recursion of the form $\varepsilon_{n+1} = S\varepsilon_n + \delta_n$, convergence with order r will hold if we have a local error decomposition

$$\begin{aligned} \delta_n &= (I - S)\xi_n + \eta_n \quad \text{with} \quad \xi_n = \mathcal{O}(\tau^r), \\ \eta_n &= \mathcal{O}(\tau^{r+1}), \quad \xi_{n+1} - \xi_n = \mathcal{O}(\tau^{r+1}). \end{aligned} \quad (2.24)$$

The above expressions for the local errors are used to obtain global results for the LOD-CN methods.

First consider the basic method (2.9) with recursion $\varepsilon_{n+1} = R\varepsilon_n + \delta_n$ and local error given by (2.19). It follows by some simple calculations that

$$\delta_n = \frac{1}{4}\tau(R - I)\psi(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^2).$$

Taking $\xi_n = -\frac{1}{4}\tau\psi(t_{n+1/2})$ shows that the method will be convergent with order one. Therefore the local order reduction disappears in the transition from local to global error.¹²⁾

In the same way, for the parallel method (2.7) the local error (2.21) can be written as

$$\delta_n^* = \frac{1}{16}\tau^2(R - I)(A^{-1}(A_2 - A_1))\psi(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3).$$

For the heat equation we will have $\|A^{-1}(A_2 - A_1)\| \leq 1$ since the matrix is normal with eigenvalues $(\lambda_1 + \lambda_2)^{-1}(\lambda_2 - \lambda_1)$, $\lambda_j < 0$. It follows again that the local order reduction is no longer present in the global errors, and we simply get $\varepsilon_n = \mathcal{O}(\tau^2)$ globally.

For the sequential symmetrical method (2.6) there is no favourable cancellation or damping of errors. The local error (2.23) can be written as

$$\delta_n^{**} = \frac{1}{4}\tau(R - I)^2\psi(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^2),$$

¹¹⁾ For commuting A_1, A_2 it easily follows that $R - I = Q_2^{-1}Q_1^{-1}Z$.

¹²⁾ Note that if we directly bound all local error contributions, $\|\varepsilon_{n+1}\| \leq \|\varepsilon_n\| + \|\delta_n\|$, then the local estimate $\|\delta_n\| = \mathcal{O}(\tau^{5/4})$ would only predict convergence with order 1/4.

but for first-order convergence we need, in view of (2.22), (2.24),

$$\delta_n^{**} = -(R^2 - I)\xi_n + \mathcal{O}(\tau^2), \quad \xi_n = \mathcal{O}(\tau).$$

This would be possible if the matrix

$$M = (R^2 - I)^{-1}(R - I)^2$$

were bounded. However, this matrix is normal and its eigenvalues are

$$\mu = \frac{(\mu_1\mu_2 - 1)^2}{(\mu_1^2\mu_2^2 - 1)}, \quad \mu_k = \frac{1 + \frac{1}{2}z_k}{1 - \frac{1}{2}z_k}$$

with $z_k = \tau\lambda_k$ the eigenvalues of τA_k . If we take $\lambda_1 = \mathcal{O}(1)$ and $\lambda_2 \sim -4h^{-2}$, then it is obvious that the matrix M will not be bounded uniformly in h . In fact, a more precise analysis (Hundsdorfer, 1992) shows that for the 2D heat equation model the order of convergence cannot be larger than $\frac{1}{2}$ if $\tau \sim h$. Since our local estimate already implied that the order of convergence is at least $\frac{1}{4}$, the entry $[\frac{1}{4}, \frac{1}{2}]$ of Table 2.1 is the obtained result.

Example 2.2 The order reduction predicted in the above analysis already shows up for the very simple stationary solution

$$u(x, y, t) = x(1-x)y(1-y)(16+y) \quad (2.25)$$

of (2.18). This solution is such that with second-order central differences no spatial errors are present. The source term f equals $-\Delta u$ and is split equally, that is, $g_1 = g_2 = \frac{1}{2}g$ with g being the restriction of f to the grid in the semi-discrete formulation. Note that the Dirichlet boundary values are zero in this example. Table 2.2 contains global errors at $T = 2$ for the two LOD-CN methods (2.5), (2.6) using (2.25) as initial condition. For this special problem with a steady-state solution and commuting matrices, method (2.7) resolves the solution exactly, as can be seen from the local error expression (2.21).

	L_2 -error, $h = 0.2$		L_2 -error, $h = 2\tau$		L_∞ -error, $h = 2\tau$	
τ	(2.5)	(2.6)	(2.5)	(2.6)	(2.5)	(2.6)
$\frac{1}{20}$	$0.23 \cdot 10^{-1}$	$0.10 \cdot 10^{-1}$	$0.34 \cdot 10^{-1}$	$0.45 \cdot 10^{-1}$	$0.73 \cdot 10^{-1}$	0.14
$\frac{1}{40}$	$0.12 \cdot 10^{-1}$	$0.27 \cdot 10^{-2}$	$0.20 \cdot 10^{-1}$	$0.31 \cdot 10^{-1}$	$0.46 \cdot 10^{-1}$	0.14
$\frac{1}{80}$	$0.59 \cdot 10^{-2}$	$0.68 \cdot 10^{-3}$	$0.10 \cdot 10^{-1}$	$0.22 \cdot 10^{-1}$	$0.25 \cdot 10^{-1}$	0.14

Table 2.2. Global errors of the LOD Crank-Nicolson methods (2.5) and (2.6) for the model problem (2.18) with steady-state solution (2.25).

The table nicely illustrates the L_2 -theory. On the fixed space grid, $h = 0.2$, where we are in the standard ODE situation, the symmetrical method (2.6) shows its order $p = 2$ and is more accurate than the basic first-order method (2.5). However, if we put $h = 2\tau$, then (2.6) is very inaccurate. Although the theory has been formulated for the L_2 -norm, it is illuminating to inspect the errors also in the max-norm, see again Table 2.2. In this norm the behaviour of the symmetrical method (2.6) is even worse: it appears that there is no convergence at all. Hence with this method there are $\mathcal{O}(1)$ errors locally, which turns out to be, not surprisingly, at the boundaries. \diamond

The above results are based on material from Hundsdorfer (1992). For the methods (2.5), (2.6) applied to the s -dimensional heat equation with s arbitrary a similar result was obtained by Čiegis & Kiškis (1994).

It is clear that in its present form Yanenko's symmetrical method (2.6) is not recommended. For this method we need *boundary correction* techniques to avoid the order reduction. Such corrections are presented in Section 2.4.

Finally we note that in view of the fact that (2.5) is of order one only, the LOD-BE method (2.1) is a better candidate for parabolic problems, since it mimics the damping properties of backward Euler. Rather general convergence results for the LOD-BE method were presented in Samarskii (1962). If this method does not provide sufficient accuracy, then the parallel version (2.7) seems a good candidate.

2.3 The Trapezoidal Splitting Method

The implicit trapezoidal rule can be viewed as a half step with forward Euler followed by a half step with backward Euler. With splitting this can be employed in the following way,

$$\begin{aligned} v_0 &= w_n, \\ v_i &= v_{i-1} + \frac{1}{2}\tau F_i(t_n, v_{i-1}), \quad i = 1, 2, \dots, s, \\ v_{s+i} &= v_{s+i-1} + \frac{1}{2}\tau F_{s+1-i}(t_{n+1}, v_{s+i}), \quad i = 1, 2, \dots, s, \\ w_{n+1} &= v_{2s}, \end{aligned} \tag{2.26}$$

which will be called Trapezoidal Splitting. Symmetry is incorporated by the reversal of the sequence with the backward Euler fractional steps. Again the vectors v_i ($1 \leq i \leq 2s - 1$) are internal quantities without physical relevance, except for v_s which is a consistent approximation to $w(t_{n+1/2})$.

Method (2.26) has order two in the classical ODE sense. If all F_i are linear and autonomous, $F_i(t, u) = A_i u$, then the method gives

$$w_{n+1} = (I - \frac{1}{2}\tau A_1)^{-1} \cdots (I - \frac{1}{2}\tau A_s)^{-1} (I + \frac{1}{2}\tau A_s) \cdots (I + \frac{1}{2}\tau A_1) w_n. \tag{2.27}$$

For commuting matrices this is also obtained with the LOD Crank-Nicolson methods (2.5), (2.7). Therefore all these methods have the same stability

function

$$R(z_1, \dots, z_s) = \prod_{i=1}^s (1 - \frac{1}{2}z_i)^{-1} (1 + \frac{1}{2}z_i),$$

and owing to the complete factorization we thus have stability if this is valid for each of the factors. For non-commuting matrices the method (2.26) is different from the LOD-CN methods of Section 2.2. In fact, for $s = 2$ it is closely related to the Peaceman-Rachford ADI method, which is discussed in the next section, also with respect to stability for non-commuting matrices. The linear version (2.27) of (2.26) was mentioned in Marchuk (1971) but not analyzed. Related linear and quasi-linear forms for $s = 2$ were examined by Beam & Warming (1976).

Stability with Internal Perturbations

In the following, a convergence analysis will be presented for linear problems, where it is assumed that the problem represents a semi-discrete PDE. This analysis is similar to the one for the LOD-CN schemes, but since the error structure will turn out to be simpler, also $s > 2$ can be considered. As before, we concentrate on linear problems with a multiple splitting

$$F_i(t, v) = A_i v + g_i(t), \quad 1 \leq i \leq s. \quad (2.28)$$

The dimension m depends on the mesh width in space, parameterized by h , and some or all of the matrices A_i will contain negative powers of h . The terms g_i will contain the source term and boundary values relevant to A_i , leading also to negative powers of h in g_i . The results will be obtained for the discrete L_2 -norm under the assumption (2.15), and we will use the notations (2.11) for Z_i , P_i , Q_i and $Z = \tau A = Z_1 + \dots + Z_s$.

Consider (2.26) with perturbations r_1, \dots, r_{2s} on the stages,

$$\begin{aligned} \tilde{v}_0 &= \tilde{w}_n, \\ \tilde{v}_i &= \tilde{v}_{i-1} + \frac{1}{2}\tau F_i(t_n, \tilde{v}_{i-1}) + r_i, \quad i = 1, \dots, s, \\ \tilde{v}_{s+i} &= \tilde{v}_{s+i-1} + \frac{1}{2}\tau F_{s+1-i}(t_{n+1}, \tilde{v}_{s+i}) + r_{s+i}, \quad i = 1, \dots, s, \\ \tilde{w}_{n+1} &= \tilde{v}_{2s}. \end{aligned} \quad (2.29)$$

Let $\varepsilon_n = \tilde{w}_n - w_n$. By subtracting (2.26) from (2.29) and eliminating the internal quantities $\tilde{v}_i - v_i$, it follows in a straightforward way that

$$\varepsilon_{n+1} = R \varepsilon_n + \delta_n, \quad (2.30)$$

with stability matrix

$$R = Q_1^{-1} Q_2^{-1} \cdots Q_s^{-1} P_s \cdots P_2 P_1, \quad (2.31)$$

and with δ_n containing the internal perturbations,

$$\begin{aligned}\delta_n = & Q_1^{-1} \cdots Q_s^{-1} \left(P_s \cdots P_2 r_1 + P_s \cdots P_3 r_2 + \cdots + P_s r_{s-1} + r_s \right) \\ & + Q_1^{-1} \cdots Q_s^{-1} r_{s+1} + Q_1^{-1} \cdots Q_{s-1}^{-1} r_{s+2} + \cdots + Q_1^{-1} r_{2s}.\end{aligned}\quad (2.32)$$

The matrix R determines how ε_n is propagated to ε_{n+1} , whereas δ_n stands for a local error introduced during the step due to the internal perturbations r_j . The usual step-by-step stability of the scheme is thus governed by R . Assuming that the matrices commute we have $\|R\| \leq 1$, and thus stability, due to (2.15). Under this assumption it also follows that

$$\|\delta_n\| \leq \|r_1\| + \|r_2\| + \cdots + \|r_{2s}\|, \quad (2.33)$$

since any explicit factor P_i occurring in (2.32) is balanced by its implicit counterpart Q_i^{-1} . This means that the internal perturbations r_i will not disrupt the result of a single step of the method. In this sense the method is *internally stable*.

Remark 2.3 Internal stability does not hold for all splitting methods. Consider for example the following method,

$$\begin{aligned}v_0 &= w_n, \\ v_i &= v_{i-1} + \frac{1}{2}\tau F_i(t_{n+\frac{1}{2}}, v_i), \quad i = 1, 2, \dots, s, \\ v_{s+i} &= v_{s+i-1} + \frac{1}{2}\tau F_{s+1-i}(t_{n+\frac{1}{2}}, v_{s+i-1}), \quad i = 1, 2, \dots, s, \\ w_{n+1} &= v_{2s}.\end{aligned}\quad (2.34)$$

The difference with (2.26) is that here the backward Euler steps precede the forward Euler steps, as in the implicit midpoint rule. Performing a similar analysis for internal perturbations now gives the recursion $\varepsilon_{n+1} = R\varepsilon_n + \delta_n$ with

$$R = P_1 P_2 \cdots P_s Q_s^{-1} \cdots Q_2^{-1} Q_1^{-1}$$

and

$$\begin{aligned}\delta_n = & P_1 \cdots P_s \left(Q_s^{-1} \cdots Q_1^{-1} r_1 + Q_s^{-1} \cdots Q_2^{-1} r_2 + \cdots + Q_s^{-1} r_s \right) \\ & + P_1 \cdots P_{s-1} r_{s+1} + P_1 \cdots P_{s-2} r_{s+2} + \cdots + P_1 r_{2s-1} + r_{2s}.\end{aligned}$$

The stability matrix R has a similar structure as with the trapezoidal splitting. Again, if the matrices A_j commute, then (2.15) implies $\|R\| \leq 1$. However, the propagation of the internal perturbations is now completely different. We only have a moderate propagation of r_1 and r_{2s} . For the other perturbations there are more explicit factors than implicit ones. With increasing stiffness, that is, if $h \rightarrow 0$, these explicit factors may introduce a blow-up of the local error δ_n . Therefore the midpoint splitting (2.34) is *not internally stable* for small h .

As a consequence the local discretization errors of this method are not bounded uniformly in h . A numerical illustration and detailed error analysis can be found in Hundsdorfer (1998b). \diamond

Local Error Analysis

We proceed with the derivation of bounds for local discretization errors of the trapezoidal splitting (2.26), similarly as for the LOD Crank–Nicolson methods in the previous section. So we focus on the temporal order of convergence with respect to the PDE solution $u_h(t)$ restricted to the grid under simultaneous space-time refinement, $h, \tau \rightarrow 0$, and for simplicity of presentation the spatial error $\sigma_h(t)$ is neglected.

The error bounds will be based on derivatives of

$$u_h(t) \quad \text{and} \quad \varphi_i(t) = F_i(t, u_h(t)), \quad 1 \leq i \leq s. \quad (2.35)$$

If the PDE solution is smooth, then with standard splittings, as discussed in Section 1.5 and 1.6, we may assume that the derivatives of $\varphi_i(t)$ are bounded uniformly in the mesh width h .

Suitable expressions for the local discretization errors can be easily derived by using the internal perturbations. First consider the perturbed formula (2.29) with $\tilde{w}_n = u_h(t_n)$, so that ε_n becomes the global discretization error, $\varepsilon_n = u_h(t_n) - w_n$, and δ_n the local discretization error. For the intermediate vectors \tilde{v}_i we take

$$\tilde{v}_i = u_h(t_n), \quad \tilde{v}_{s+i} = u_h(t_{n+1}), \quad 1 \leq i \leq s.$$

Note that the actual choice for these vectors is not important since we are only interested in the overall local error δ_n . But with the above choice we get simple expressions for the perturbations, namely $r_j = \tau \rho_j$ with residual truncation errors

$$\begin{aligned} \rho_i &= -\frac{1}{2}\tau\varphi_i(t_n), \quad i = 1, \dots, s, \\ \rho_{s+1} &= u_h(t_{n+1}) - u_h(t_n) - \frac{1}{2}\tau\varphi_s(t_{n+1}), \\ \rho_{s+i} &= -\frac{1}{2}\tau\varphi_{s+1-i}(t_{n+1}), \quad i = 2, \dots, s. \end{aligned}$$

Recall that for simplicity of presentation the spatial error $\sigma_h(t)$ is taken here as zero.

We will elaborate the local error δ_n for linear problems (2.28) with $s = 2$ and $s = 3$. Commutativity of the matrices A_i is not assumed. Inserting the above residuals in (2.32) gives for $s = 3$,

$$\begin{aligned} \delta_n &= Q_1^{-1}Q_2^{-1}Q_3^{-1} \left(-\frac{1}{2}\tau P_3 P_2 \varphi_1(t_n) - \frac{1}{2}\tau P_3 \varphi_2(t_n) - \frac{1}{2}\tau \varphi_3(t_n) + u_h(t_{n+1}) \right. \\ &\quad \left. - u_h(t_n) - \frac{1}{2}\tau \varphi_3(t_{n+1}) - \frac{1}{2}\tau Q_3 \varphi_2(t_{n+1}) - \frac{1}{2}\tau Q_3 Q_2 \varphi_1(t_{n+1}) \right). \end{aligned}$$

By expanding all terms around $t = t_{n+1/2}$ and using $u'_h(t) = \sum_{i=1}^s \varphi_i(t)$, it follows by some straightforward calculations that

$$\begin{aligned}\delta_n &= Q_1^{-1}Q_2^{-1}Q_3^{-1} \left(-\frac{1}{4}\tau Z_3 Z_2 \varphi_1(t_{n+\frac{1}{2}}) + \frac{1}{4}\tau^2 Z_3 \varphi'_2(t_{n+\frac{1}{2}}) \right. \\ &\quad \left. + \frac{1}{4}\tau^2 (Z_2 + Z_3) \varphi'_1(t_{n+\frac{1}{2}}) \right) + \mathcal{O}(\tau^3).\end{aligned}\quad (2.36)$$

The corresponding formula for $s = 2$ simply follows from this by setting $Z_3 = 0$, $\varphi_3 = 0$. So, for $s = 2$ the local discretization error is

$$\delta_n = \frac{1}{4}\tau^2 Q_1^{-1}Q_2^{-1} Z_2 \varphi'_1(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3). \quad (2.37)$$

Using (2.16), it follows directly that for $s = 2$ we have $\delta_n = \mathcal{O}(\tau^2)$, whereas for fixed h an $\mathcal{O}(\tau^3)$ bound holds due to the hidden τ in Z_2 . To obtain a similar bound uniformly in h , we need the compatibility condition

$$A_2 \varphi'_1(t) = \mathcal{O}(1). \quad (2.38)$$

This condition will only be satisfied in special cases, namely where $\varphi'_1(t)$ satisfies homogeneous boundary conditions relevant to A_2 . This will hold if there are no boundary conditions present, e.g. with spatial periodicity.

It should further be noted that also fractional order results are possible: if $A_2^\alpha \varphi_1(t) = \mathcal{O}(1)$ with $\alpha \in (0, 1)$, it can be shown that $\delta_n = \mathcal{O}(\tau^{2+\alpha})$. For the formula with $s = 3$ similar considerations hold. To guarantee that $\delta_n = \mathcal{O}(\tau^3)$ we now get several compatibility conditions. If we merely assume that A_2 and A_3 commute, it follows from (2.16), (2.36) only that $\delta_n = \mathcal{O}(\tau)$, which is a poor result of course since this is the error introduced in a single step. Below we will present some global error bounds where the compatibility conditions are replaced by less restrictive ones.

Global Error Analysis

We proceed with convergence results for the trapezoidal splitting method (2.26) for linear problem $w'(t) = Aw(t) + g(t)$, $0 < t \leq T$, with splitting (2.28) for $s = 2$ and $s = 3$. It is assumed in the remainder of this section that the method is stable, i.e.,

$$\|R^n\| \leq K \quad \text{for } n \geq 0, n\tau \leq T, \quad (2.39)$$

with a constant $K = \mathcal{O}(1)$. This certainly holds if the matrices A_i commute and satisfy assumption (2.15), see (2.27) and (2.16). To derive the global error bounds we will use again the decomposition (2.24) to exploit possible local error cancellation. The expressions, with $\varphi_i(t) = F_i(t, u_h(t))$, for the local errors then easily lead to global results.

Theorem 2.4 Consider the trapezoidal splitting method (2.26) with $s = 2$. Assume (2.39), $w_0 = u_h(0)$ and

$$A^{-1} A_2 \varphi_1^{(k)}(t) = \mathcal{O}(1) \quad (2.40)$$

for $t \in [0, T]$, $k = 1, 2$. Then $\|u_h(t_n) - w_n\| = \mathcal{O}(\tau^2)$ for $n\tau \leq T$.

Proof. We have

$$R - I = Q_1^{-1} Q_2^{-1} (Z_1 + Z_2).$$

Hence the local error (2.37) can be written as

$$\delta_n = \frac{1}{4} \tau^2 (R - I) A^{-1} A_2 \varphi_1'(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3).$$

Clearly this fits in the form (2.24) with $\xi_n = \frac{1}{4} \tau^2 A^{-1} A_2 \varphi_1'(t_{n+1/2})$, $r = 2$, $S = R$. The assumption (2.40) with $k = 2$ guarantees that $\xi_{n+1} - \xi_n = \mathcal{O}(\tau^3)$. Hence the second-order convergence result follows. \square

This result holds for noncommuting A_1 and A_2 . However, to verify condition (2.40) it is helpful to assume that the matrices do commute. It is easy to show that $A^{-1} A_2 = \mathcal{O}(1)$ if A_1 and A_2 are negative definite and commuting, since we then have a simultaneous diagonalization with orthogonal eigenvectors. Further it is obvious from the proof that the assumption in the theorem could be formulated a bit more general. What we need is only the existence of a function v with $v(t), v'(t) = \mathcal{O}(1)$ satisfying $Av(t) = A_2 \varphi_1'(t)$ for all t . This would allow A to be singular, which may happen for instance with Neumann conditions. The following results for $s = 3$ permit a similar generalization.

Theorem 2.5 Consider the trapezoidal splitting method (2.26) with $s = 3$, $w_0 = u_h(0)$. Assume (2.39) and let $M = A + \frac{1}{4} \tau^2 A_3 A_2 A_1$. If

$$M^{-1} A_i \varphi_j^{(k+1)}(t) = \mathcal{O}(1) \quad (i \geq 2, j < i), \quad M^{-1} A_3 A_2 \varphi_1^{(k)}(t) = \mathcal{O}(1) \quad (2.41)$$

for $t \in [0, T]$ and $k = 0, 1$, then we have $\|u_h(t_n) - w_n\| = \mathcal{O}(\tau^2)$ for $t_n \leq T$.

Proof. We have

$$R - I = Q_1^{-1} Q_2^{-1} Q_3^{-1} (Z_1 + Z_2 + Z_3 + \frac{1}{4} Z_3 Z_2 Z_1).$$

Hence the local error (2.36) can be written as

$$\delta_n = \frac{1}{4} \tau^2 (R - I) M^{-1} \left(-A_3 A_2 \varphi_1 + (A_2 + A_3) \varphi_1' + A_3 \varphi_2' \right) + \mathcal{O}(\tau^3)$$

with the terms $\varphi_j^{(k)}$ evaluated in $t_{n+1/2}$. The results thus follow in the same way as in the previous theorem. \square

Corollary 2.6 Let $s = 3$, $w_0 = u_h(0)$, and suppose the matrices A_i are negative definite and commuting. If either $A_2 = \mathcal{O}(1)$ or $A_3 = \mathcal{O}(1)$, then $\|u_h(t_n) - w_n\| = \mathcal{O}(\tau^2)$ for $t_n \leq T$.

Proof. If the A_i are commuting and negative definite, then

$$(A + \frac{1}{4}\tau^2 A_3 A_2 A_1)^{-1} A_i = \mathcal{O}(1),$$

and using $A_i = \mathcal{O}(1)$ for $i = 2$ or 3 , it follows that (2.41) is satisfied. \square

The above conditions for second-order convergence with $s = 3$ are quite strong and will often not be satisfied. Under weaker conditions first-order convergence can be proved. For instance, if (2.41) is replaced by the condition

$$\tau M^{-1} A_3 A_2 \varphi_1^{(k)}(t) = \mathcal{O}(1), \quad (2.42)$$

convergence with order one follows as in Theorem 2.5. As before also intermediate results with fractional powers are possible.

Applications of these theoretical results will be given in Section 2.5, including numerical comparisons with Yanenko's LOD-CN method (2.6).

2.4 Boundary Correction Techniques

The fact that boundary conditions can give rise to accuracy reduction has led to correction techniques by which the accuracy can be restored. The implementation of such techniques depends on the type of splitting method, on the type of boundary conditions, on the problem and also on the spatial geometry. Hence their description is rather technical and in general their use can become complicated. The general idea is always to treat boundary values as far as possible in the same way as the interior points. This will be briefly described here for the trapezoidal splitting (2.26) and Yanenko's LOD-CN method (2.6).

The possibility of accuracy reduction in splitting methods due to boundary conditions was mentioned already by D'Yakonov (1962). Boundary corrections were introduced by Fairweather & Mitchell (1967) for ADI methods. Examples of correction techniques are also found in Mitchell & Griffiths (1980, Chap. 2), Sommeijer et al. (1981) and in LeVeque (1985).

Boundary corrections can be easily derived for rectangular regions Ω . Assume for the moment that Dirichlet conditions are given on the whole boundary Γ . Let Γ_i be that part of the boundary on which the values are relevant to F_i , and let $\Gamma_{j,\dots,k} = \bigcup_{i=j}^k \Gamma_i$ for $j \leq k$. If F_i contains no discretized spatial derivatives, then Γ_i is empty. In case F_i does contain discretized derivatives we could apply these on Γ_j for $j \neq i$, but not on Γ_i itself. Application of F_i on the boundary will be denoted by $F_{i,\Gamma}$ and $v_{i,\Gamma}$ stands for the boundary grid function associated with v_i .

Due to its simple form it is easy to derive boundary corrections for the trapezoidal splitting. We note that $v_0 = w_n$ and $v_{2s} = w_{n+1}$ are consistent approximations to the exact solution. Further, in (2.26) we need the boundary values corresponding to v_{i-1} on Γ_i ($i = 1, \dots, s$), and for v_{s+i} on Γ_{s+1-i} ($i = 1, \dots, s$). For the corrected boundary conditions of the trapezoidal splitting we first take $v_{0,\Gamma} = u_{h,\Gamma}(t_n)$ on Γ , and subsequently

$$v_{i,\Gamma} = v_{i-1,\Gamma} + \frac{1}{2}\tau F_{i,\Gamma}(t_n, v_{i-1,\Gamma}) \quad \text{on } \Gamma_{i+1, \dots, s} \quad (2.43)$$

for $i = 1, 2, \dots, s-1$, and likewise $v_{2s,\Gamma} = u_{h,\Gamma}(t_{n+1})$ on Γ and

$$v_{2s-i,\Gamma} = v_{2s+1-i,\Gamma} - \frac{1}{2}\tau F_{i,\Gamma}(t_{n+1}, v_{2s+1-i,\Gamma}) \quad \text{on } \Gamma_{i+1, \dots, s} \quad (2.44)$$

for $i = 1, 2, \dots, s-1$. With Neumann boundary conditions the formulas (2.43) and (2.44) should be used to prescribe the outward normal derivatives for v_i and v_{s+i} .

With Yanenko's method (2.6) we can take $v_{0,\Gamma} = u_{h,\Gamma}(t_n)$, $v_{s,\Gamma} = u_{h,\Gamma}(t_{n+1/2})$ and $v_{2s,\Gamma} = u_{h,\Gamma}(t_{n+1})$, due to consistency. However, for the other stages the situation is more complicated, due to the fact that v_i can not be written explicitly in terms of either v_{i-1} or v_{i+1} , and the boundary values corresponding to v_i are now needed on both Γ_i and Γ_{i+1} ($i = 1, 2, \dots, s-1$) in the first s fractional steps; for the last stages this is similar. Consider, for example, the first stage, where v_1 is implicitly defined in terms of v_0 . Starting with $v_0 = w(t_n) = u_{h,\Gamma}(t_n)$ on Γ , we can approximate the implicit relation by

$$v_1 \approx w(t_n) + \frac{1}{2}\tau F_1(t_n, w(t_n)).$$

However, since F_1 cannot be applied on Γ_1 in general, we can use this formula only on Γ_2 in the second stage of the method. As we have

$$F_1(t, w(t)) = w'(t) - \sum_{j=2}^s F_j(t, w(t)),$$

we can also take the approximate formula

$$v_1 \approx w(t_{n+\frac{1}{2}}) - \frac{1}{2}\tau \sum_{j=2}^s F_j(t_{n+\frac{1}{2}}, w(t_{n+\frac{1}{2}})),$$

which now can be used on Γ_1 with boundary values $w(t_{n+1/2}) = u_{h,\Gamma}(t_{n+1/2})$ on Γ . For the other v_j we can proceed similarly. This gives for the v_i ($i = 1, 2, \dots, s-1$) the correction formulas

$$\begin{aligned} v_{i,\Gamma} &= u_{h,\Gamma}(t_n) + \frac{1}{2}\tau \sum_{j=1}^i F_{j,\Gamma}(t_n, u_{h,\Gamma}(t_n)) \quad \text{on } \Gamma_{i+1}, \\ v_{i,\Gamma} &= u_{h,\Gamma}(t_{n+\frac{1}{2}}) - \frac{1}{2}\tau \sum_{j=i+1}^s F_{j,\Gamma}(t_{n+\frac{1}{2}}, u_{h,\Gamma}(t_{n+\frac{1}{2}})) \quad \text{on } \Gamma_i. \end{aligned} \quad (2.45)$$

Likewise for the v_{s+i} ($i = 1, 2, \dots, s-1$) in the last stages we take

$$\begin{aligned} v_{s+i,\Gamma} &= u_{h,\Gamma}(t_{n+\frac{1}{2}}) + \frac{1}{2}\tau \sum_{j=s+1-i}^s F_{j,\Gamma}(t_{n+\frac{1}{2}}, u_{h,\Gamma}(t_{n+\frac{1}{2}})) \quad \text{on } \Gamma_{s-i}, \\ v_{s+i,\Gamma} &= u_{h,\Gamma}(t_{n+1}) - \frac{1}{2}\tau \sum_{j=1}^{s-i} F_{j,\Gamma}(t_{n+1}, u_{h,\Gamma}(t_{n+1})) \quad \text{on } \Gamma_{s+1-i}. \end{aligned} \quad (2.46)$$

Other choices for these correction formulas can be made. Numerical results in LeVeque (1985) indicate that more accuracy may be obtained if in (2.45), (2.46) higher-order terms of τ are included to give a better approximation of the implicit relations. However, if $s > 2$ or nonlinear terms are involved, this leads to rather complicated correction terms. A detailed error analysis for the above boundary corrections along the lines of the previous section is not available. Instead, we will present in the next section some numerical results, showing that with the current formulas second-order convergence is restored for Yanenko's method (2.6).

2.5 Numerical Comparisons

Below some test results are presented for Yanenko's LOD-CN method (2.6) and the trapezoidal splitting method (2.26) with and without boundary corrections. The LOD-CN method is twice as expensive per step with respect to computational work. For the comparisons we will therefore apply (2.6) with a step size twice as large as for (2.26). The errors below are the L_2 -errors $\|u_h(t_n) - w_n\|_2$. These include spatial errors, but it has been verified that the temporal errors are dominating.

A Diffusion-Reaction Problem with Traveling-Wave Solution

As a test example with time-dependent boundary conditions we solve the 2D diffusion-reaction equation

$$u_t = \epsilon(u_{xx} + u_{yy}) + \epsilon^{-1}u^2(1-u) \quad (2.47)$$

for $(x, y) \in \Omega = (0, 1)^2$, $t \in [0, 1]$, with $\epsilon = 10^{-1}$. The initial condition and Dirichlet boundary conditions are chosen according to the exact solution

$$u(x, y, t) = \left(1 + e^{\frac{1}{2}\epsilon^{-1}(x+y-t)}\right)^{-1}. \quad (2.48)$$

This solution is a smooth wave traveling diagonally across the domain. For this scalar 2D test example splitting is not really necessary, but the structure of the problem is similar to many real-life problems with large numbers of reactions, where splitting may be required in view of computer (memory) capacities.

The spatial derivatives are discretized with standard second-order differences on a uniform grid with mesh width h . Let $[\delta_x^2(t)]$ stand for the difference operator approximating ∂_{xx} with the associated time-dependent boundary conditions at $x = 0, 1$, and let $[\delta_y^2(t)]$ stand likewise for the approximation to ∂_{yy} with boundary conditions at $y = 0, 1$. For the semi-discrete system $w'(t) = F(t, w(t))$ a three-term splitting $F = F_1 + F_2 + F_3$ is used with

$$F_1 = \epsilon [\delta_x^2(t)] w, \quad F_2 = \epsilon [\delta_y^2(t)] w, \quad F_3 = \epsilon^{-1} w^2 (1 - w), \quad (2.49)$$

where multiplication of vectors is to be interpreted component-wise. Table 2.3 contains the L_2 -errors $\|u_h(t_n) - w_n\|_2$ at time $t_n = 1$ for different values of h , computed with and without the boundary corrections. The trapezoidal splitting scheme (2.26) is applied with $\tau = h$. For the LOD-CN method we take $\tau = 2h$, so that the computational costs are the same.

In this example trapezoidal splitting gives second-order accuracy without boundary corrections, which is related to Corollary 2.6. Although the assumptions of this corollary are not strictly fulfilled, the result seems to apply here since $A_3 = \mathcal{O}(1)$, with A_3 the Jacobian matrix of the reaction term F_3 . Yanenko's method gives a low-order of convergence without boundary corrections, which is in agreement with the theoretical results of Table 2.1. With boundary corrections the second-order convergence is restored and with respect to accuracy the method becomes competitive with the trapezoidal splitting method.

In the convergence results for the trapezoidal splitting method we have seen that the order in which the various splitting components are applied can be of importance. To illustrate this, problem (2.47) is considered once more, but now with

$$F_1 = \epsilon^{-1} w^2 (1 - w), \quad F_2 = \epsilon [\delta_x^2(t)] w, \quad F_3 = \epsilon [\delta_y^2(t)] w. \quad (2.50)$$

Here we cannot expect second-order convergence for the trapezoidal splitting since both F_2 and F_3 are not $\mathcal{O}(1)$. Hence, in contrast to the previous example,

h	Without Bnd.Corr.		With Bnd.Corr.	
	TS	LOD-CN	TS	LOD-CN
$\frac{1}{10}$	$3.8 \cdot 10^{-3}$	$1.5 \cdot 10^{-2}$	$3.2 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$
$\frac{1}{20}$	$9.9 \cdot 10^{-4}$	$6.9 \cdot 10^{-3}$	$8.2 \cdot 10^{-4}$	$2.9 \cdot 10^{-3}$
$\frac{1}{40}$	$2.5 \cdot 10^{-4}$	$4.1 \cdot 10^{-3}$	$2.0 \cdot 10^{-4}$	$7.4 \cdot 10^{-4}$
$\frac{1}{80}$	$6.3 \cdot 10^{-5}$	$2.7 \cdot 10^{-3}$	$5.1 \cdot 10^{-5}$	$1.8 \cdot 10^{-4}$

Table 2.3. Splitting (2.49) for (2.47). L_2 -errors for the trapezoidal splitting (TS) method (2.26) with $\tau = h$, and Yanenko's LOD-CN method (2.6) with $\tau = 2h$.

	Without Bnd.Corr.		With Bnd.Corr.	
h	TS	LOD-CN	TS	LOD-CN
$\frac{1}{10}$	$6.3 \cdot 10^{-3}$	$1.4 \cdot 10^{-2}$	$2.1 \cdot 10^{-3}$	$5.7 \cdot 10^{-3}$
$\frac{1}{20}$	$1.8 \cdot 10^{-3}$	$7.1 \cdot 10^{-3}$	$4.9 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$
$\frac{1}{40}$	$5.9 \cdot 10^{-4}$	$4.2 \cdot 10^{-3}$	$1.2 \cdot 10^{-4}$	$4.1 \cdot 10^{-4}$
$\frac{1}{80}$	$2.3 \cdot 10^{-4}$	$2.8 \cdot 10^{-3}$	$2.8 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$

Table 2.4. Splitting (2.50) for (2.47). L_2 -errors for the trapezoidal splitting (TS) method (2.26) with $\tau = h$, and Yanenko's LOD-CN method (2.6) with $\tau = 2h$.

Corollary 2.6 is not applicable here. Table 2.4 shows the errors with this splitting.

Indeed, we see that now boundary corrections are also needed for trapezoidal splitting to obtain second-order accuracy. Without these corrections a first-order convergence could be expected from condition (2.42). In the table the actual order of convergence seems slightly better, but tests with smaller τ and h did show an order of convergence close to one. As with the previous example, the results for trapezoidal splitting are more favourable than for Yanenko's method when taking the work load per unit of time into account.

3 ADI Methods

The second class of well-known splitting methods is formed by *alternating direction implicit* (ADI) methods. While the LOD methods were primarily developed by scientists from the Soviet Union during the 1950's and 60's, ADI methods were developed in that period by Douglas, Gunn, Peaceman and Rachford in the USA. An important application field at that time was formed by two- and three-dimensional parabolic problems from numerical oil reservoir models, see Peaceman (1977). As with the LOD methods, dimension splitting is the classical type of application, but we will also consider more general splittings of advection, diffusion and reaction terms. In contrast to the LOD methods, the intermediate stages of ADI methods yield approximations that are consistent with the full problem.

3.1 The Peaceman-Rachford Method

The ADI method of Peaceman & Rachford (1955) is one of the very first splitting methods proposed in the literature. For the nonlinear ODE system

$w'(t) = F(t, w(t))$ with the two-term splitting

$$F(t, v) = F_1(t, v) + F_2(t, v),$$

the Peaceman-Rachford method reads

$$\begin{aligned} w_{n+\frac{1}{2}} &= w_n + \frac{1}{2}\tau F_1(t_n, w_n) + \frac{1}{2}\tau F_2(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}), \\ w_{n+1} &= w_{n+\frac{1}{2}} + \frac{1}{2}\tau F_2(t_{n+\frac{1}{2}}, w_{n+\frac{1}{2}}) + \frac{1}{2}\tau F_1(t_{n+1}, w_{n+1}). \end{aligned} \quad (3.1)$$

This method could be viewed as being obtained by Strang-type operator splitting with alternate use of forward and backward Euler in a symmetrical fashion to get second-order (in the classical ODE sense). However, it is more natural to consider it as a method of its own. In contrast to the LOD methods of the previous section, both F_1 and F_2 are incorporated in each of the two stages, which makes the intermediate vector $w_{n+1/2}$ a consistent approximation at the time point halfway the step interval $[t_n, t_{n+1}]$. Because of this alternate implicit use of F_1 and F_2 , in the framework of dimension splitting the method is called alternating direction implicit. Although we will not restrict ourselves to dimension splitting, the name ADI will still be employed for schemes like (3.1) in which each stage is consistent.

A disadvantage compared to LOD methods is that (3.1) does not have a natural extension for more than two F -components. We therefore refrain from a detailed discussion of local and global errors and refer to Hundsdorfer & Verwer (1989) for comprehensive results on order reduction.¹³⁾ Notice that the method returns stationary solutions exactly: for autonomous problems we have

$$w_n = \bar{w}, \quad F(\bar{w}) = 0 \quad \implies \quad w_{n+\frac{1}{2}} = w_{n+1} = \bar{w}. \quad (3.2)$$

This is due to consistency of the internal stages.

Stability

If F_1, F_2 are linear, $F_i(t, v) = A_i v$, then the Peaceman-Rachford method gives $w_{n+1} = R w_n$ with

$$R = (I - \frac{1}{2}\tau A_1)^{-1} (I + \frac{1}{2}\tau A_2) (I - \frac{1}{2}\tau A_2)^{-1} (I + \frac{1}{2}\tau A_1).$$

Its stability analysis is more complicated than for LOD-CN methods, unless A_1, A_2 commute. For the commuting case the method is identical to the LOD method (2.5) with $s = 2$. In the more general linear case we can write

$$w_n = (I - \frac{1}{2}\tau A_1)^{-1} \hat{R}^n (I - \frac{1}{2}\tau A_1) w_0,$$

¹³⁾ The results are similar to those for the Douglas method (with $s = 2, \theta = \frac{1}{2}$) that are presented below: order reduction will in general only happen for the local errors. Still, boundary corrections, as given in Fairweather & Mitchell (1967), will improve accuracy since it will reduce the error constants.

$$\hat{R} = (I + \frac{1}{2}\tau A_2)(I - \frac{1}{2}\tau A_2)^{-1}(I + \frac{1}{2}\tau A_1)(I - \frac{1}{2}\tau A_1)^{-1}.$$

Hence, except for the start and completion of the integration, it is possible to rewrite (3.1) in the form of (2.5), and a similar transformation is also possible for nonlinear problems, see also Gourlay & Mitchell (1969). If we have

$$\text{cond}(I - \frac{1}{2}\tau A_1) \leq K_1, \quad \|\hat{R}^n\| \leq K_2, \quad n \geq 0,$$

then it is obvious that $\|R^n\| \leq K = K_1 K_2$ for all $n \geq 0$. However, boundedness of $\|I - \frac{1}{2}\tau A_1\|$ will in general only be valid under very strict conditions on the time step, such as $\tau/h^2 = \mathcal{O}(1)$ for parabolic problems. In practice such restrictions are not obeyed, but the Peaceman-Rachford method appears to be unconditionally stable for problems with smooth coefficients. Recent results on stability for certain classes of parabolic equations were obtained by Schatzman (1999).

Red-Black Splitting – the Hopscotch Method

Throughout this chapter we consider splittings between advection, diffusion and reaction processes, possibly augmented by splittings along dimensions. Other types of splittings exist, but these are not that widely applicable. An example is provided by the Hopscotch method for the one-dimensional advection-diffusion equation, which will be briefly discussed here, mainly for its historical interest.

Consider the 1D equation $u_t + au_x = du_{xx}$ with $t > 0$ and $x \in \mathbb{R}$, to avoid boundary conditions, and given initial condition. Discretization in space with standard second-order central differences gives $w'(t) = Aw(t)$ for which we consider a splitting in odd and even points,

$$A_i = I_i A \quad \text{for } i = 1, 2, \quad I_1 = I - I_2, \quad (I_2 v)_j = \begin{cases} v_j & \text{if } j \text{ is even,} \\ 0 & \text{if } j \text{ is odd.} \end{cases}$$

In two dimensions this would generalize to a red-black (checkerboard) splitting; here we only consider the 1D equation. If we apply the Peaceman-Rachford method with step size 2τ , we get

$$w_{n+1} = w_n + \tau A_1 w_n + \tau A_2 w_{n+1},$$

$$w_{n+2} = w_{n+1} + \tau A_2 w_{n+1} + \tau A_1 w_{n+2},$$

for $n = 0, 2, 4, \dots$. This defines the (odd-even) Hopscotch method. This method, with generalizations, is due to Gordon (1965), Gourlay (1970), Gourlay & McGuire (1971). By some simple manipulations it follows that

$$I_2 w_{n+1} = \frac{1}{2} I_2 (w_{n+2} + w_n),$$

$$I_2 w_{n+2} = I_2 w_n + 2\tau A_2 (I_1 w_{n+1} + I_2 w_{n+1}).$$

The scheme can now be written out in full; if $j + n$ is even we obtain

$$w_j^{n+2} = w_j^n + \frac{a\tau}{h} (w_{j-1}^{n+1} - w_{j+1}^{n+1}) + 2\frac{d\tau}{h^2} (w_{j-1}^{n+1} - w_j^n - w_j^{n+2} + w_{j+1}^{n+1}),$$

and

$$w_j^{n+1} = \frac{1}{2} (w_j^n + w_j^{n+2}).$$

Hence we have a complete decoupling of the points (x_j, t_n) with $j + n$ even from those with $j + n$ odd. On the odd points we simply have interpolation between adjacent time levels. In the following we restrict our attention to the points with $j + n$ even.

If $d = 0$ we recognize the resulting scheme as the explicit *leap-frog* scheme

$$w_j^{n+2} = w_j^n + \frac{a\tau}{h} (w_{j-1}^{n+1} - w_{j+1}^{n+1}) \quad (3.3)$$

for the advection equation $u_t + au_x = 0$. This is a second-order scheme which is stable (in the von Neumann sense) for $|a|\tau/h < 1$, see the Examples II.3.1 and II.3.5. Note that it was stated above that the Peaceman-Rachford method is usually unconditionally stable for problems with smooth coefficients. However, due to this odd-even splitting we have in fact introduced matrices A_1, A_2 which are as if they belong to a problem whose coefficients are very non-smooth.

If $a = 0$ the resulting scheme is effectively explicit, since the implicitness only arises in a scalar, linear way. It is the *Du Fort-Frankel* scheme

$$w_j^{n+2} = w_j^n + 2\frac{d\tau}{h^2} (w_{j-1}^{n+1} - w_j^n - w_j^{n+2} + w_{j+1}^{n+1}) \quad (3.4)$$

for the diffusion equation $u_t = du_{xx}$ (Du Fort & Frankel, 1953). Stability, in the sense of von Neumann, can be analyzed by writing the scalar 2-step recursion in 1-step system form. Then, by some calculations, see for instance Richtmyer & Morton (1967), it follows that the scheme is unconditionally stable. Of course this is very peculiar for an explicit scheme; it seems almost too good to be true, and indeed, there is a defect. That becomes clear if we consider the truncation error

$$\begin{aligned} \rho_n &= \frac{1}{\tau} \left(u(x_j, t_{n+2}) - u(x_j, t_n) \right) \\ &\quad - 2\frac{d}{h^2} \left(u(x_{j-1}, t_{n+1}) - u(x_j, t_n) - u(x_j, t_{n+2}) + u(x_{j+1}, t_{n+1}) \right) \\ &= 2d \left(\frac{\tau}{h} \right)^2 u_{tt}(x_j, t_{n+1}) + \mathcal{O}(\tau^2) + \mathcal{O}(h^2) + \mathcal{O}(\tau^4 h^{-2}). \end{aligned}$$

Hence the Du Fort-Frankel scheme is only consistent if $\tau/h \rightarrow 0$. Therefore the unconditional stability property is of very limited value. Occasionally, the scheme is still used for low-accuracy computations with problems having

very small diffusion coefficients, where the inconsistency does not matter too much, but in general it is not a scheme to be recommended. The same comment applies of course to the Hopscotch scheme.¹⁴⁾

It should be emphasized that the loss of unconditional stability or loss of consistency are due to the odd-even splitting, not to the Peaceman-Rachford time stepping. In the remainder of this chapter we will only consider standard type splittings, with separation of advection, diffusion, reactions and/or dimensions.

3.2 The Douglas Method

The Douglas ADI method shares the advantage of the LOD methods of applicability to multi-component splittings of F . Suppose we have a splitting

$$F(t, v) = F_0(t, v) + F_1(t, v) + \cdots + F_s(t, v), \quad (3.5)$$

where it is assumed that F_0 is non-stiff or mildly stiff so that this term can be treated explicitly. All the other terms may be stiff because they are treated implicitly in a sequential fashion as follows,

$$\begin{aligned} v_0 &= w_n + \tau F(t_n, w_n), \\ v_i &= v_{i-1} + \theta \tau (F_i(t_{n+1}, v_i) - F_i(t_n, w_n)), \quad i = 1, \dots, s, \\ w_{n+1} &= v_s. \end{aligned} \quad (3.6)$$

The method is of order one if $\theta = 1$ and of order two if $\theta = \frac{1}{2}$, $F_0 = 0$. We will call it the Douglas method since early ADI methods proposed by J. Douglas Jr and co-workers fit in if we put $F_0 = 0$, see Douglas (1955), Douglas & Rachford (1956), Douglas (1962) and Douglas & Gunn (1964). The method with $\theta = \frac{1}{2}$ was derived independently by Brian (1961) and Douglas (1962) for 3D parabolic problems with splitting along the dimensions.

Methods of the type (3.6) are also known as stabilizing correction methods, since the first forward Euler stage is followed by implicit stages which serve to stabilize this first explicit stage. As in LOD methods, these implicit stages use a single F_i . For references to early Soviet Union contributors on this type of ADI methods we refer to Marchuk (1990).

A nice property of (3.6) is that all internal vectors v_i are consistent approximations to $w(t_{n+1})$. Therefore, for problems where the boundary conditions are influential the accuracy is often better than for LOD methods. We will illustrate this later. Furthermore, the method returns stationary solutions exactly, similar to (3.2), which can be seen by considering the consecutive v_i , $i = 1, 2, \dots, s$.

¹⁴⁾ Nevertheless, an interesting computational property of the Hopscotch scheme is that it can be implemented in one array of storage. Hence in terms of memory it is attractive, which still can be an advantage in large-scale 3D computations.

Stability

The Douglas ADI method lends itself less easily to stability analysis than the LOD methods and the same observations as for the Peaceman-Rachford method apply. In fact, if $s \geq 3$ even the linear commuting case needs closer attention. Here we consider the most simple situation from the analysis point of view: linear problems with commuting, normal matrices A_i . Instead of test model (2.3) we consider

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t) + \cdots + \lambda_s w(t), \quad (3.7)$$

wherein the term $\lambda_0 w(t)$ is included to take the explicit term F_0 into account. Applying the ADI method to (3.7) yields a recursion $w_{n+1} = R w_n$ with

$$R(z_0, z_1, \dots, z_s) = 1 + \left(\prod_{i=1}^s (1 - \theta z_i) \right)^{-1} \sum_{i=0}^s z_i \quad (3.8)$$

as stability function.

For the stability analysis we will consider the wedge

$$\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : \zeta = 0 \text{ or } |\arg(-\zeta)| \leq \alpha\}$$

in the complex plane and examine stability for $z_i \in \mathcal{W}_\alpha$, $i \geq 1$. If F_i is a discretized linear advection-diffusion operator and λ_i an eigenvalue in the Fourier decomposition, then $\alpha < \frac{1}{2}\pi$ means that advection is not allowed to dominate too much, see Section I.3.4. For pure diffusion we have $z_i = \tau \lambda_i \in \mathcal{W}_0$, the line of non-positive real numbers. In the analysis it will be assumed that z_0, z_1, \dots, z_s are independent of each other.

Theorem 3.1 Suppose $z_0 = 0$ and $s \geq 2$, $1 \leq r \leq s-1$. For any $\theta \geq \frac{1}{2}$ we have

$$|R| \leq 1 \text{ for all } z_i \in \mathcal{W}_\alpha, 1 \leq i \leq s \iff \alpha \leq \frac{1}{s-1} \frac{\pi}{2}, \quad (3.9)$$

$$|R| \leq 1 \text{ for all } z_1, \dots, z_{s-r} \in \mathcal{W}_\alpha \text{ and } z_{s-r+1}, \dots, z_s \leq 0 \} \iff \alpha \leq \frac{1}{s-r} \frac{\pi}{2}. \quad (3.10)$$

Proof. For the complete proofs we refer to Hundsdorfer (1998a, 1999) as they are rather technical. However, necessity in (3.9) is easy to show. Take equal $z_k = -te^{i\alpha}$, $k \geq 1$ (with $i = \sqrt{-1}$). Then, for $t \rightarrow \infty$,

$$R = 1 - \frac{ste^{i\alpha}}{\theta^s t^s e^{is\alpha} + \mathcal{O}(t^{s+1})} = 1 - \frac{s}{\theta^s} t^{1-s} e^{i\alpha(1-s)} (1 + \mathcal{O}(t^{-1})),$$

and consequently $\operatorname{Re}(R) > 1$ if t is sufficiently large and $\alpha(1-s) > \frac{1}{2}\pi$.

To illustrate necessity in (3.10), consider $s = 3$ and $z_3 \leq 0$. Since R is fractional linear in z_3 , it follows that $|R| \leq 1$ for all $z_3 \leq 0$ iff this holds with z_3 equal to 0 or ∞ . This amounts to verification of the inequalities

$$\left| 1 + \frac{z_1 + z_2}{(1 - \theta z_1)(1 - \theta z_2)} \right| \leq 1, \quad \left| 1 - \frac{1}{\theta(1 - \theta z_1)(1 - \theta z_2)} \right| \leq 1.$$

For the first inequality we know from (3.9) that $\alpha \leq \frac{1}{2}\pi$ is necessary and sufficient, but for the second inequality it can be shown as above that we need $\alpha \leq \frac{1}{4}\pi$. \square

These two results show that unlike the LOD method the stability of the Douglas ADI method depends on the number of the split terms F_i . For $s = 2$, $z_0 = 0$ the method is *unconditionally stable* in the sense that $|R| \leq 1$ for all $z_1, z_2 \in \mathbb{C}^-$ with $\mathbb{C}^- = \mathcal{W}_{\pi/2}$ the left half-plane. However, this no longer holds for $s \geq 3$. For $z_1 \in \mathbb{C}^-$ arbitrary the essential condition for stability is

$$z_1 \in \mathbb{C}^- \quad \text{and} \quad z_2, \dots, z_s \leq 0.$$

This means that at most one of the implicitly treated terms F_i may have large, purely imaginary eigenvalues and all others are then required to have negative real eigenvalues if these are also large in modulus. With purely parabolic problems all eigenvalues will be negative real, and then we have unconditional stability.

Observe that in (3.10) with $r = 1$ we get the same angles α as in (3.9), which corresponds to $r = 0$. It is also unexpected that there is no difference between $\theta = \frac{1}{2}$ and $\theta = 1$. Similar stability results with an explicit term $z_0 \neq 0$ can be found in Hundsdorfer (1999); if we assume $|1 + z_0| \leq 1$, then having $\theta = \frac{1}{2}$ or $\theta = 1$ makes a difference. For $\theta = 1$ the above statements remain the same, but for $\theta = \frac{1}{2}$ we need $\alpha = 0$. This difference will be discussed in more detail in Section 4.1 for the simple case $s = 1$.

In Figure 3.1 the boundary of the stability region, where $|R| \leq 1$, is plotted for two special choices: $z_0 = 0$, $z_i = z$ ($1 \leq i \leq s$) (thick solid lines) and $z_s = \infty$ (dotted lines). The left picture shows the case $s = 2$ and the right picture the case $s = 3$. The dotted lines give contour levels $|R| = 0.1, \dots, 0.9$ with equal $z_i = z$.

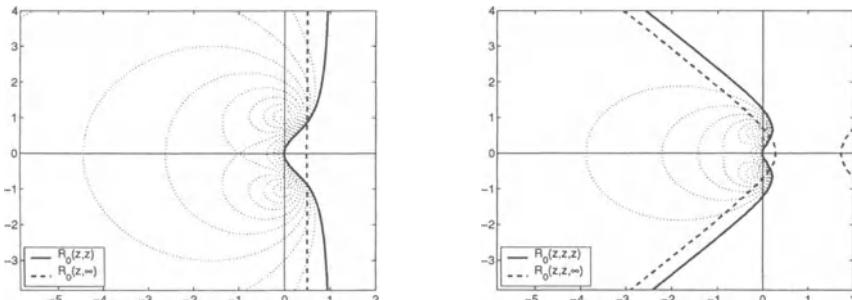


Fig. 3.1. Regions of stability $|R| \leq 1$ for $\theta = 1$, $z_0 = 0$, with equal $z_i = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$; the dotted lines give contour levels $|R| = 0.1, \dots, 0.9$ with equal $z_i = z$.

and $z_0 = 0$, $z_i = z$ ($1 \leq i \leq s - 1$), $z_s = \infty$ (dashed lines). Here $\theta = 1$; plots with $\theta = \frac{1}{2}$ look very similar. Also drawn, as dotted curved lines, are contour lines of $|R|$ at 0.1, 0.2, ..., 0.9 for the case that the z_i are equal. It is seen that we have little damping in general. If there are two z_i with large negative values, then $|R|$ will be close to 1.

Also far outside the stability region the value of $|R|$ may be very close to 1. This is an unpleasant property of the Douglas method because it may result in a very slow onset of instability which may be hard to detect in an actual computation.

Example 3.2 To illustrate this slow onset of instability, we solve the system of two advection-reaction equations

$$u_t = (a_1 u)_x + (a_2 u)_y + G u \quad \text{for } (x, y) \in (0, 1)^2, t > 0, \quad (3.11)$$

with components u_1 and u_2 , and G the 2×2 matrix

$$G = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

We take $k_1 = 1$. The second reaction constant k_2 can be used to vary the stiffness of the reaction term. Note that G has eigenvalues 0 and $-(k_1 + k_2)$ and that we have a chemical equilibrium if $u_1/u_2 = k_2/k_1$. The velocity field is the same as in the Molenkamp-Crowley problem from Example 1.1, i.e.,

$$a_1(x, y) = 2\pi(y - \frac{1}{2}), \quad a_2(x, y) = -2\pi(x - \frac{1}{2}).$$

The initial condition is chosen as

$$u_1(x, y, t) = c, \quad u_2(x, y, t) = (1 - c) + \tilde{c} e^{-80((x - \frac{1}{2})^2 + (y - \frac{3}{4})^2)},$$

with $c = k_2/(k_1 + k_2)$ and $\tilde{c} = 100/k_2$. So if k_2 increases we start closer to the chemical equilibrium to maintain some smoothness in the solution. After a short transient phase, where most of the Gaussian pulse is transferred from u_2 to u_1 , the problem becomes essentially identical to the Molenkamp-Crowley advection problem rotating the pulse around the center of the domain. Recall that at $t = 1$ one rotation is completed.

The spatial discretization is performed with simple second-order central differences. At the inflow boundaries Dirichlet conditions are prescribed and at the outflow boundaries the standard upwind discretization is used. Since the solution is not very steep, in the interior second-order central differences are possible using a uniform grid. We thus have created a linear semi-discrete system for which we consider a three-term splitting with F_1, F_2 the finite difference operators for advection in the x - and y -direction, respectively, and with F_3 defined by the linear reaction term. All three terms are treated implicitly by the Douglas method (3.6) with $F_0 = 0$. Recall that an implicit advection treatment is only justified for problems with smooth solutions, see

Section III.1.3, and that if steep gradients arise upwinding with flux limiting is to be preferred above central differences. However, the experiment here merely serves as an illustration of the theoretical results on the stability of the Douglas method with $s = 3$.

Although we are not in a model situation with commuting normal operators, a von Neumann analysis can be applied, by ignoring the boundary conditions and freezing the coefficients, to give eigenvalues λ_1, λ_2 on the imaginary axis whereas $\lambda_3 = 0$ or $-(k_1 + k_2)$. Our tests have been performed on a fixed 80×80 grid. Inserting the maximal absolute values for the velocities a_1, a_2 thus gives values z_1, z_2 between $-80\tau\pi i$ and $80\tau\pi i$ and $z_3 = 0$ or $z_3 = -\tau(k_1 + k_2)$. Two values of τ are considered, $\tau = 1/80$ and $\tau = 1/160$, giving the 1D Courant numbers π and $\pi/2$, respectively. On the basis of the stability result (3.10), where $s = 3$, $r = 1$, we now expect that the Douglas method will become unstable if we choose k_2 large. This indeed occurs, as is illustrated by Figure 3.2.

With $\theta = \frac{1}{2}$ in the Douglas method, the figure shows the numerical solution of component u_1 for $\tau = 1/160$, $k_2 = 2000$ at time $t = 1, 2, 3$ and $t = 4$ (with a different scale). Some smooth oscillations in the wake of the Gaussian pulse are seen, but these are caused by the central difference spatial discretization. The instabilities occur near the corners where the advection speeds are largest. Notable is that the build-up of the instabilities is very slow, and therefore it will be difficult to detect this with error estimators. To

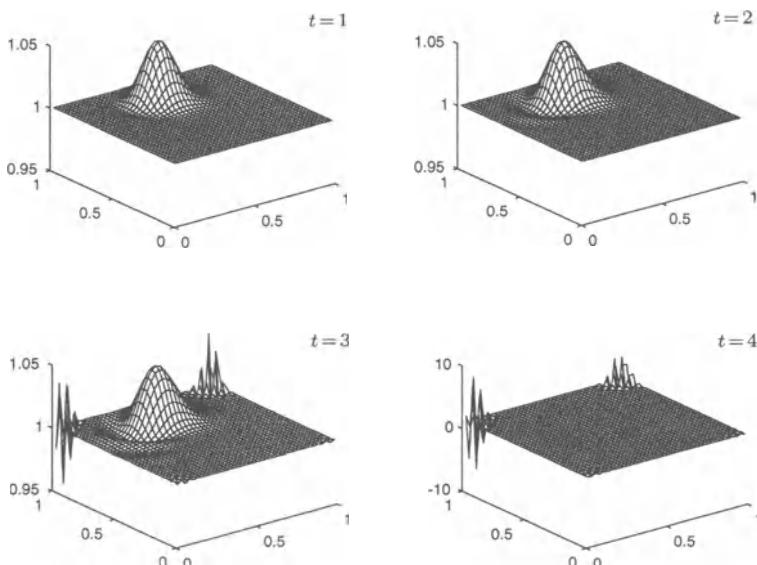


Fig. 3.2. The numerical solution for component u_1 of the advection-reaction problem (3.11) for $t = 1, 2, 3, 4$, computed by the ADI Douglas method (3.6) with $\theta = \frac{1}{2}$. Note the different scaling at $t = 4$.

k_2	τ^{-1}	Douglas, $t = 1$	Douglas, $t = 4$	LOD-TS, $t = 4$
500	80	$4.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-1}$	$1.0 \cdot 10^{-1}$
	160	$2.9 \cdot 10^{-2}$	$8.3 \cdot 10^{-2}$	$8.3 \cdot 10^{-2}$
1000	80	$2.2 \cdot 10^{-2}$	$8.1 \cdot 10^{+8}$	$5.0 \cdot 10^{-2}$
	160	$1.4 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$
2000	80	8.4	$7.6 \cdot 10^{+22}$	$2.5 \cdot 10^{-2}$
	160	$7.2 \cdot 10^{-3}$	7.5	$2.1 \cdot 10^{-2}$
4000	80	$1.9 \cdot 10^{+3}$	$5.6 \cdot 10^{+30}$	$1.2 \cdot 10^{-2}$
	160	$1.4 \cdot 10^{+3}$	overflow	$1.0 \cdot 10^{-2}$

Table 3.1. Maximum errors for the Douglas ADI method (3.6) with $\theta = \frac{1}{2}$ and the trapezoidal splitting method (2.26) for the advection-reaction problem (3.11).

some extent the slow onset of the instability can be attributed to the fact that it occurs near an outflow boundary, but related tests have confirmed that it is mainly caused by amplification factors only slightly larger than 1 in modulus. Also with $\theta = 1$ this type of instability has been observed.

Finally, in Table 3.1 the maximum errors $\|u_h(t_n) - w_n\|_\infty$ are given for different values of the reaction constant k_2 and again using $\theta = \frac{1}{2}$. These confirm that the instability increases with the magnitude of k_2 and that the transition from stable to unstable is very hesitant. Especially for smaller values of k_2 several rotations are needed to give a significant instability. For comparison this table also contains results for the trapezoidal splitting method (2.26). The latter does not suffer from instability. \diamond

In conclusion it can be said that the Douglas method (3.6) for advection-diffusion problems with dimension splitting seems only suited if either (i) advection dominates at most in one direction, or (ii) advection dominates in two directions and the other components F_j are non-stiff.

Error Analysis

In the remainder of this section an error analysis is presented for the Douglas method with $s \leq 3$. As usual, we restrict ourselves to linear problems, that is

$$F_i(t, v) = A_i v + g_i(t), \quad i = 0, 1, \dots, s,$$

where some of the matrices A_i will contain negative powers of the mesh width h . For inhomogeneous boundary conditions, the terms g_i will contain the boundary values relevant to A_i , which will also lead to negative powers of h . As before, the error bounds derived here will not be adversely affected by the mesh width h in the spatial discretization, and we will write $\mathcal{O}(\tau^p)$ to

denote vectors or matrices whose norm can be bounded by $C\tau^p$ with $C > 0$ independent of h . We use the notations

$$Z_i = \tau A_i, \quad Z = \tau A = Z_0 + Z_1 + \cdots + Z_s, \quad Q_i = I - \theta Z_i.$$

As the starting point, we consider along with (3.6) the perturbed scheme

$$\begin{aligned} \tilde{v}_0 &= \tilde{w}_n + \tau F(t_n, \tilde{w}_n) + r_0, \\ \tilde{v}_i &= \tilde{v}_{i-1} + \theta \tau (F_i(t_{n+1}, \tilde{v}_i) - F_i(t_n, \tilde{w}_n)) + r_i, \quad i = 1, \dots, s, \\ \tilde{w}_{n+1} &= \tilde{v}_s. \end{aligned} \quad (3.12)$$

As before, the perturbations r_i may stand for round-off or errors introduced in the solution of the implicit systems, for instance. We will use them here to derive an expression for the local discretization errors.

Let $\varepsilon_n = \tilde{w}_n - w_n$ and $\zeta_i = \tilde{v}_i - v_i - \varepsilon_n$. Subtraction of (3.6) from (3.12) gives the relations

$$\zeta_0 = Z\varepsilon_n + r_0, \quad \zeta_i = Q_i^{-1}(\zeta_{i-1} + r_i), \quad \varepsilon_{n+1} = \varepsilon_n + \zeta_s.$$

By elimination of the internal quantities ζ_i , it follows that

$$\varepsilon_{n+1} = R\varepsilon_n + \delta_n \quad (3.13)$$

with stability matrix

$$R = I + Q_s^{-1} \cdots Q_2^{-1} Q_1^{-1} Z \quad (3.14)$$

and with δ_n containing the internal perturbations,

$$\delta_n = Q_s^{-1} \cdots Q_1^{-1} (r_0 + r_1) + Q_s^{-1} \cdots Q_2^{-1} r_2 + \cdots + Q_s^{-1} r_s. \quad (3.15)$$

The matrix R determines how an error already present at time t_n will be propagated to t_{n+1} , whereas δ_n stands for a local error introduced during the step. It is assumed that the problem is such that the scheme is stable,

$$\|R^n\| \leq K \quad \text{for } n \geq 1, \quad \|Q_i^{-1}\| \leq C. \quad (3.16)$$

Under assumption (2.15) we have $C = 1$ in the L_2 -norm.

We will use (3.13)–(3.15) to obtain bounds for the discretization errors. As before, for ease of presentation, the spatial errors $\sigma_h(t)$ will be neglected here; these could be added to the perturbations. Let $\tilde{w}_n = u_h(t_n)$ so that $\varepsilon_n = u_h(t_n) - w_n$ is the global discretization error. To derive an expression for the local discretization error δ_n we are free to choose the \tilde{v}_i ; it is only the global relation (3.13) that matters. Simple expressions are obtained by taking $\tilde{v}_i = u_h(t_{n+1})$, $0 \leq i \leq s$. Then $r_j = \tau \rho_j$ with stage truncation errors ρ_j given by

$$\rho_0 = \frac{1}{2} \tau u''_h(t_n) + \frac{1}{6} \tau^2 u'''_h(t_n) + \cdots,$$

$$\rho_i = -\theta (\varphi_i(t_{n+1}) - \varphi_i(t_n)) = -\theta \tau \varphi'_i(t_n) - \frac{1}{2} \theta \tau^2 \varphi''_i(t_n) - \cdots$$

for $i = 1, \dots, s$, with $\varphi_i(t) = F_i(t, u_h(t))$. Recall that for standard splittings the derivatives of the φ_i will be bounded uniformly in the mesh width h . Inserting these residuals into (3.15) yields the local discretization error

$$\delta_n = \frac{1}{2}\tau^2 Q_s^{-1} \cdots Q_1^{-1} u''_h(t_n) - \theta\tau^2 \sum_{i=1}^s Q_s^{-1} \cdots Q_i^{-1} \varphi'_i(t_n) + \mathcal{O}(\tau^3). \quad (3.17)$$

Note that boundedness of the Q_i^{-1} factors implies that $\delta_n = \mathcal{O}(\tau^2)$ uniformly in the mesh width h , and by the stability assumption we thus obtain at least first-order convergence of the global errors ε_n independent of h .

If $F_0 = 0$ and $\theta = \frac{1}{2}$ this estimate can be improved, but then we need to take a closer look on the error propagation. We will elaborate this for $s \leq 3$. It follows by some calculations that we then have

$$\delta_n = \frac{1}{4}\tau^2 Q_3^{-1} Q_2^{-1} Q_1^{-1} \left(Z_1 \varphi'_2(t_n) + (Z_1 + Z_2 - \frac{1}{2}Z_1 Z_2) \varphi'_3(t_n) \right) + \mathcal{O}(\tau^3).$$

In case $s = 2$ this formula can be used with $Z_3 = 0$, $Q_3 = I$ and $\varphi_3 = 0$.

To prove second-order convergence the decomposition (2.24) will be used with $r = 2$. Using that framework, convergence results are now easily obtained. In the following it is assumed that the scheme is stable in the sense of (3.16), and that the solution is smooth, so that derivatives of u_h and φ_i are $\mathcal{O}(1)$.

Theorem 3.3 *Let $\theta = \frac{1}{2}$, $F_0 = 0$, $w_0 = u_h(0)$. Consider method (3.6) with $s = 2$, and assume (3.16) and*

$$A^{-1} A_1 \varphi_2^{(k)}(t) = \mathcal{O}(1)$$

for $t \in [0, T]$, $k = 1, 2$. Then $\|u_h(t_n) - w_n\| = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. If $s = 2$ we have

$$\delta_n = \frac{1}{4}\tau^2 Q_2^{-1} Q_1^{-1} Z_1 \varphi'_2(t_n) + \mathcal{O}(\tau^3) = \frac{1}{4}\tau^2 (R - I) Z^{-1} Z_1 \varphi'_2(t_n) + \mathcal{O}(\tau^3).$$

Thus in (2.24) we can take $\xi_n = -\frac{1}{4}\tau^2 A^{-1} A_1 \varphi'_2(t_n)$ and η_n containing the remaining $\mathcal{O}(\tau^3)$ terms. By the assumptions we have $\xi_n = \mathcal{O}(\tau^2)$ and $\xi_{n+1} - \xi_n = \mathcal{O}(\tau^3)$, which shows second-order convergence. \square

For many splittings with standard advection-diffusion problems we will have $\|A^{-1} A_1\| \leq 1$, and hence the assumption $A^{-1} A_1 \chi_2 = \mathcal{O}(1)$ for $\chi_2 = \varphi'_2, \varphi''_2$ in this theorem is natural. Furthermore we note that if A is singular, the above can be easily generalized: what we need to prove second-order convergence is the existence of a vector $v = \mathcal{O}(1)$ such that $Av = A_1 \chi_2$. In all of the following such generalizations apply.

Theorem 3.4 Let $\theta = \frac{1}{2}$, $F_0 = 0$, $w_0 = u_h(0)$. Consider method (3.6) with $s = 3$, and assume (3.16) and

$$A^{-1} A_i \varphi_j^{(k)}(t) = \mathcal{O}(1) \quad (j > i) \quad \text{and} \quad A^{-1} A_1 A_2 \varphi_3^{(k)}(t) = \mathcal{O}(\tau^{-1})$$

for $t \in [0, T]$, $k = 1, 2$. Then $\|u_h(t_n) - w_n\| = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. Since $R = I + Q_3^{-1} Q_2^{-1} Q_1^{-1} Z$, the local discretization error can be written as

$$\delta_n = \frac{1}{4} \tau^2 (R - I) Z^{-1} \left(Z_1 \varphi'_2(t_n) + (Z_1 + Z_2 - \frac{1}{2} Z_1 Z_2) \varphi'_3(t_n) \right) + \mathcal{O}(\tau^3).$$

Note that $A^{-1} A_1 A_2 \varphi_3^{(k)} = \mathcal{O}(\tau^{-1})$ implies $Z^{-1} Z_1 Z_2 \varphi_3^{(k)} = \mathcal{O}(1)$. Thus we can proceed in the same way as in the previous proof, with ξ_n containing the $\mathcal{O}(\tau^2)$ terms. \square

Compared to the case $s = 2$ in Theorem 3.3, the essential new condition here with $s = 3$ is $A^{-1} A_1 A_2 \chi_3 = \mathcal{O}(\tau^{-1})$, $\chi_3 = \varphi'_3, \varphi''_3$, that is,

$$Z^{-1} Z_1 Z_2 \chi_3 = \mathcal{O}(1).$$

This may hold also if $Z^{-1} Z_1 Z_2 \neq \mathcal{O}(1)$, as illustrated by the following example.

Example 3.5 As a classical example, consider the dimension splitting for the 3D heat equation $u_t = \Delta u + f$ with matrices A_i , $i = 1, 2, 3$, representing the one-dimensional second-order difference operators in x, y, z -directions, and with inhomogeneous Dirichlet conditions at the boundaries. Then the matrices A_i commute and $\|A^{-1} A_i\| \leq 1$ for all h . Further it holds that $A_1^\alpha A_2^\alpha \chi_3 = \mathcal{O}(1)$ for any $\alpha < \frac{1}{4}$, see Lemma III.6.5. Therefore we can write

$$Z^{-1} Z_1 Z_2 \chi_3 = \tau^{2\alpha} Z^{-1} Z_1^{1-\alpha} Z_2^{1-\alpha} [A_1^\alpha A_2^\alpha \chi_3].$$

The matrix $Z^{-1} Z_1^{1-\alpha} Z_2^{1-\alpha} = \mathcal{O}(1)$ since it is normal with bounded eigenvalues $z^{-1} z_1^{1-\alpha} z_2^{1-\alpha}$, $z = z_1 + z_2 + z_3$, $z_j < 0$. Taking $\tau \sim h^{1+\epsilon}$ with $\epsilon = 1 - 4\alpha > 0$, it follows that

$$\|Z^{-1} Z_1 Z_2 \chi_3\| \sim \tau^{2\alpha} \left(\frac{\tau}{h^2} \right)^{1-2\alpha} = \mathcal{O}(1).$$

Thus the conditions in Theorem 3.4 are fulfilled under a step size restriction $\tau \sim h^{1+\epsilon}$ with $\epsilon > 0$ arbitrarily small. Further we note that if φ_3 satisfies homogeneous boundary conditions on the boundaries relevant to A_1 and A_2 , then no condition on the ratio τ/h is necessary, since then $A_1 A_2 \chi_3 = \mathcal{O}(1)$.

In conclusion, Theorem 3.4 indicates that also with $s = 3$ we will often have second-order convergence, although a mild restriction on the step size might be necessary in this case. \diamond

For larger values of s a similar analysis could be performed, but verification of the accuracy conditions becomes increasingly technical. For example, if $s = 4$ we get, in addition to conditions as in Theorem 3.4, a requirement $A^{-1}A_1A_2A_3\chi_4 = \mathcal{O}(\tau^{-2})$. Although this may be fulfilled in many special cases, in general an order of convergence between one and two must now be expected.

Finally we note that the above convergence results with $\theta = \frac{1}{2}$, $s = 3$ are somewhat more favourable than those of Theorem 2.5 for the trapezoidal splitting method. For the latter method order two convergence for the 3D heat equation will only be valid under additional compatibility assumptions. Of course, convergence requires stability and there the trapezoidal splitting method (2.26) allows for larger classes of problems than the Douglas method.

Numerical Comparison

To compare the Douglas method with the LOD type Crank-Nicolson methods of the previous section, we consider again the 2D diffusion-reaction equation (2.47) with the traveling-wave solution (2.48). We take the same spatial discretizations and splittings with $s = 3$ and $F_0 = 0$ as in the Tables 2.3, 2.4. The results for the Douglas method (3.6) with $\theta = \frac{1}{2}$ are displayed in Table 3.2.

As expected the Douglas method gives second-order convergence in all cases. Due to the internal consistency of the method order reduction effects are not seen here. In this sense the Douglas method is to be preferred over the trapezoidal splitting (2.26) and certainly over the LOD-CN method (2.6).

Also for the Douglas method boundary corrections can be implemented, along the line of the formulas given in Mitchell & Griffiths (1980, p. 62). Although this is not necessary with respect to the order of convergence, such boundary corrections did produce somewhat smaller errors in the present test. However, since boundary corrections have to be derived anew for each individual problem, it is a favourable property of the Douglas method that it shows a second-order convergence behaviour without corrections. Of course, it should be emphasized once more that with the Douglas method it is stability that needs careful consideration for practical problems in case $s \geq 3$.

h	Splitting (2.49)	Splitting (2.50)
$\frac{1}{10}$	$6.0 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$
$\frac{1}{20}$	$1.4 \cdot 10^{-3}$	$3.0 \cdot 10^{-4}$
$\frac{1}{40}$	$3.3 \cdot 10^{-4}$	$7.3 \cdot 10^{-5}$
$\frac{1}{80}$	$8.1 \cdot 10^{-5}$	$1.8 \cdot 10^{-5}$

Table 3.2. Problem (2.47) with the splittings (2.49) and (2.50). L_2 -errors for the Douglas method (3.6), $\theta = \frac{1}{2}$, with $\tau = h$.

Finally we note that with the present diffusion-reaction problem the reaction term is not stiff, so this could also be treated explicitly. However, with the Douglas method (3.6) this would mean in a forward Euler fashion, which will lead therefore to first-order convergence only. Splitting type methods with improved treatment of explicit terms are studied in the next two sections.

4 IMEX Methods

For many problems there are natural splittings into two parts, one of which is non-stiff, or mildly stiff, and suited for explicit treatment. This can be handled in a straightforward way within operator splitting but then accuracy and boundary conditions are a major concern. Moreover, with operator splitting, multistep methods are not suited to solve the fractional steps, because if we solve a subproblem $v'(t) = F_i(t, v(t))$ on $[t_n, t_{n+1}]$, a multistep method will need past information for that particular subproblem, rather than values w_{n-j} for the whole problem.

In this section we consider IMEX methods, which consist of suitable mixtures of implicit and explicit methods. There exist IMEX methods of linear multistep and Runge-Kutta type. Here we will mainly focus on linear two-step methods. First we illustrate the ideas through the one-step IMEX- θ method.

4.1 The IMEX- θ Method

Suppose that the semi-discrete system is given by

$$w'(t) = F(t, w(t)) \equiv F_0(t, w(t)) + F_1(t, w(t)), \quad (4.1)$$

where F_0 is a non-stiff term suitable for explicit time integration, for instance discretized advection, and F_1 is a stiff term requiring an implicit treatment, say discretized diffusion or stiff reactions. Consider the following simple method

$$w_{n+1} = w_n + \tau F_0(t_n, w_n) + (1 - \theta)\tau F_1(t_n, w_n) + \theta\tau F_1(t_{n+1}, w_{n+1}), \quad (4.2)$$

with parameter $\theta \geq \frac{1}{2}$. Here one sees that the explicit Euler method is combined with the A -stable implicit θ -method. Such mixtures of implicit and explicit methods are called IMEX methods. Method (4.2) is called the IMEX- θ method.

Inserting the exact solution of (4.1) gives the temporal truncation error

$$\begin{aligned} \rho_n &= \tau^{-1} (w(t_{n+1}) - w(t_n)) - (1 - \theta)F(t_n, w(t_n)) - \theta F(t_{n+1}, w(t_{n+1})) \\ &\quad + \theta(\varphi(t_{n+1}) - \varphi(t_n)) = (\frac{1}{2} - \theta)\tau w''(t_n) + \theta\tau\varphi'(t_n) + \mathcal{O}(\tau^2), \end{aligned}$$

where $\varphi(t) = F_0(t, w(t))$. If F_0 represents discretized advection or other non-stiff terms, smoothness of w will usually also imply smoothness of φ , independent of boundary conditions or small mesh widths h . Therefore the structure of the truncation error is much more favourable than for methods based on operator splitting with fractional steps. For example, with a stationary solution we have a zero truncation error. On the other hand, with methods of this IMEX type it is stability that needs a careful examination.

The above method (4.2) is the most simple IMEX method. In fact this method can be viewed as a special case ($s = 1$) of the Douglas method (3.6), which has been studied already in Section 3.2 but mainly with $F_0 = 0$ and multiple implicit terms.

Stability

As before, let us consider the scalar complex test equation

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t), \quad (4.3)$$

and let $z_j = \tau \lambda_j$, $j = 0, 1$. Recall that in applications to PDEs these λ_j represent, after linearization, eigenvalues of the two components F_0 and F_1 found by inserting Fourier modes. One would hope that having $|1 + z_0| \leq 1$ (stability of the explicit method) and $\operatorname{Re}(z_1) \leq 0$ (stability of the implicit method) would be sufficient for linear stability of the IMEX method, but in general this is not true. Application of (4.2) to this test equation yields $w_{n+1} = R w_n$, $R = R(z_0, z_1)$, with

$$R(z_0, z_1) = \frac{1 + z_0 + (1 - \theta)z_1}{1 - \theta z_1} \quad (4.4)$$

and stability for the test equation thus requires $|R(z_0, z_1)| \leq 1$.

We first consider the set

$$\mathcal{D}_0 = \{z_0 \in \mathbb{C} : \text{the IMEX scheme is stable for any } z_1 \in \mathbb{C}^-\}. \quad (4.5)$$

So here we insist on A -stability with respect to the implicit part, and the question is whether \mathcal{D}_0 is smaller than the stability region of the explicit Euler method. Considering $z_1 = it$, $t \in \mathbb{R}$, and using the maximum modulus principle, it follows by some straightforward calculations that $z_0 = x_0 + iy_0 \in \mathcal{D}_0$ iff for all $t \in \mathbb{R}$

$$(2\theta - 1)t^2 + 2(\theta - 1)y_0 t - (2x_0 + x_0^2 + y_0^2) \geq 0,$$

which is for $\theta > \frac{1}{2}$ equivalent with

$$\theta^2 y_0^2 + (2\theta - 1)(1 + x_0)^2 \leq 2\theta - 1.$$

Plots of these ellipse-shaped regions are given in Figure 4.1. If $\theta = 1$ we recover the stability region of the explicit Euler method, but if we decrease θ

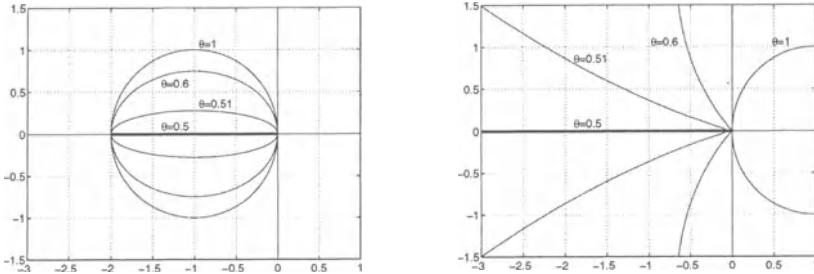


Fig. 4.1. Boundaries of regions \mathcal{D}_0 (left) and \mathcal{D}_1 (right) for the IMEX- θ method (4.2) with $\theta = 0.5, 0.51, 0.6$ and 1 .

the set starts to shrink, and for $\theta = \frac{1}{2}$ it even reduces to the negative line segment $[-2, 0]$.

Alternatively, one can insist on using the full stability region of the explicit method, $\mathcal{S}_0 = \{z_0 : |1 + z_0| \leq 1\}$, but then z_1 has to be restricted to the set

$$\mathcal{D}_1 = \{z_1 \in \mathbb{C} : \text{the IMEX scheme is stable for any } z_0 \in \mathcal{S}_0\}. \quad (4.6)$$

Then, for all $r_0 = 1 + z_0$ in the closed unit disk we should have $|r_0 + (1 - \theta)z_1| \leq |1 - \theta z_1|$. The left-hand side is largest if r_0 has the same argument as $(1 - \theta)z_1$ and lies on the unit circle. Thus we find that $z_1 \in \mathcal{D}_1$ iff

$$1 + |(1 - \theta)z_1| \leq |1 - \theta z_1|.$$

See again Figure 4.1 for an illustration. Note that it is only for $\theta = 1$ that we get the stability region of the implicit θ -method and that the set \mathcal{D}_1 equals the non-positive real line \mathbb{R}^- if $\theta = \frac{1}{2}$.

Method (4.2) with $\theta = 1$ could be viewed as a time splitting method where we first solve $v'(t) = F_0(t, v(t))$ on $[t_n, t_{n+1}]$ with forward Euler and then $v'(t) = F_1(t, v(t))$ with backward Euler. This explains the favourable stability results with $\theta = 1$. However, the structure of the truncation error is very different from the time splitting methods where intermediate results are inconsistent with the full equation. This is due to interference of the first-order splitting error with the first-order Euler errors.

Note that if $\theta > \frac{1}{2}$ the implicit θ -method is *strongly A-stable* (that is, A-stable with damping at ∞), whereas with $\theta = \frac{1}{2}$ the method is ‘just’ A-stable. Apparently, using a strongly A-stable implicit method gives better stability properties within the IMEX formula (4.2). On the other hand, the above criteria with the sets \mathcal{D}_0 and \mathcal{D}_1 are rather strict. If we confine z_1 to the negative real line, for example, then also $\theta = \frac{1}{2}$ gives stability for all $|1 + z_0| \leq 1$. Therefore, the IMEX- $\theta = \frac{1}{2}$ method should not be discarded, but extra care should be given to stability when applying it.

Finally we note that if F_0 is a genuinely non-stiff term, then we may assume that $|z_0| \leq L\tau$ with moderate constant L , in which case we can apply

a simple perturbation argument,

$$|R(z_0, z_1)| \leq |R(0, z_1)| + |1 - \theta z_1|^{-1} |z_0|,$$

$$|R(z_0, z_1)^n| \leq (1 + L\tau)^n \leq e^{Ltn} \quad \text{whenever } z_1 \in \mathbb{C}^-,$$

to show stability on finite intervals $[0, T]$ for any $\theta \geq \frac{1}{2}$.

Remark 4.1 In the above the values of λ_0 and λ_1 have been considered as independent, which is a reasonable assumption if F_0 and F_1 act in different directions, for instance with $F_0 \approx a\partial_x$ (horizontal coupling) and $F_1 \approx d\partial_{zz}$ (vertical coupling) or F_1 representing a reaction term (coupling over chemical species). Different results are obtained if there is a dependence between λ_0 and λ_1 . Then the implicit treatment of λ_1 can stabilize the process so that even $z_0 \in S_0$ may no longer be needed. Consider the standard example of the 1D advection-diffusion equation $u_t + au_x = du_{xx}$ with periodicity in space and with second-order spatial discretization. If advection is treated explicitly and diffusion implicitly, then for z_0, z_1 we may take

$$z_0 = -i\nu \sin(2\omega), \quad z_1 = -4\mu \sin^2(\omega)$$

with $\nu = a\tau/h$, $\mu = d\tau/h^2$ and $0 \leq \omega \leq \pi$, see Section I.3. A straightforward calculation shows that $|R| \leq 1$ iff

$$1 - 8(1 - \theta)\mu s + 16(1 - \theta)^2\mu^2 s^2 + 4\nu^2 s(1 - s) \leq 1 + 8\theta\mu s + 16\theta^2\mu^2 s^2$$

with $s = \sin^2(\omega)$. This holds for all $s \in [0, 1]$ iff

$$\nu^2 \leq 2\mu \quad \text{and} \quad 2(1 - 2\theta)\mu \leq 1.$$

So for any $\theta \geq \frac{1}{2}$ we now just have the condition $\nu^2 \leq 2\mu$ for stability, that is $\tau \leq 2d/a^2$. \diamond

In the following we will discuss several generalizations of the simple θ -method (4.2). Such generalizations are necessary for practical problems since the explicit Euler method is not well suited for advection and also first-order accuracy is often not sufficient.

4.2 IMEX Multistep Methods

With linear multistep methods (II.3.1) the above IMEX approach can be generalized as follows. Consider the fully implicit linear k -step method

$$\sum_{j=0}^k \alpha_j w_{n+j} = \tau \sum_{j=0}^k \beta_j (F_0(t_{n+j}, w_{n+j}) + F_1(t_{n+j}, w_{n+j})) \quad (4.7)$$

with separation of F_0 -terms and F_1 -terms. We can handle the F_0 -terms in an explicit manner by applying the extrapolation formula

$$\varphi(t_{n+k}) = \sum_{j=0}^{k-1} \gamma_j \varphi(t_{n+j}) + \mathcal{O}(\tau^q),$$

where $\varphi(t) = F_0(t, w(t))$. This leads to the k -step IMEX method

$$\sum_{j=0}^k \alpha_j w_{n+j} = \tau \sum_{j=0}^{k-1} \beta_j^* F_0(t_{n+j}, w_{n+j}) + \tau \sum_{j=0}^k \beta_j F_1(t_{n+j}, w_{n+j}), \quad (4.8)$$

with new coefficients $\beta_j^* = \beta_j + \beta_k \gamma_j$. Methods of this implicit-explicit multistep type were introduced by Crouzeix (1980) and Varah (1980). The order of consistency of (4.8) is easy to establish.

Theorem 4.2 *Assume the implicit linear multistep method (4.7) has order p and the extrapolation procedure has order q . Then the IMEX method (4.8) has order $r = \min(p, q)$.*

Proof. With $\varphi(t) = F_0(t, w(t))$, the local truncation error can be written as

$$\begin{aligned} & \frac{1}{\tau} \sum_{j=0}^k \left(\alpha_j w(t_{n+j}) - \tau \beta_j w'(t_{n+j}) \right) + \beta_k \left(\varphi(t_{n+k}) - \sum_{j=0}^{k-1} \gamma_j \varphi(t_{n+j}) \right) \\ &= C \tau^p w^{(p+1)}(t_n) + \mathcal{O}(\tau^{p+1}) + \beta_k C' \tau^q \varphi^{(q)}(t_n) + \mathcal{O}(\tau^{q+1}) \end{aligned}$$

with constants C, C' determined by the coefficients of the multistep method and the extrapolation procedure. \square

In this truncation error only total derivatives of w and φ arise, and therefore the error is not influenced by large Lipschitz constants (negative powers of the mesh width) in F_0 or F_1 . This is an important observation since it means absence of order reduction. On the other hand, stability results for the IMEX multistep methods are quite complicated, even for the simple test problem (4.3).

In the remainder of this section we focus on two-step methods and first give three examples of known, popular two-step schemes with $p = q = 2$. In these examples the most advanced time level is taken as t_{n+1} .

Example 4.3 Using the explicit midpoint (Leap-Frog) method for the explicit part and the trapezoidal rule (Crank-Nicolson) for the implicit part yields the popular IMEX-CNLF scheme

$$w_{n+1} - w_{n-1} = 2\tau F_0(t_n, w_n) + \tau F_1(t_{n+1}, w_{n+1}) + \tau F_1(t_{n-1}, w_{n-1}). \quad (4.9)$$

The stability region \mathcal{S}_0 of the explicit method is restricted to the imaginary axis between $-i$ and i , see Example II.3.5. The implicit method is the A -stable trapezoidal rule with step size 2τ . \diamond

Example 4.4 A second-order IMEX-BDF scheme can be derived from the implicit two-step backward differentiation formula (II.3.11) and its explicit counterpart (II.3.13). We consider the family of schemes

$$\begin{aligned} \frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} &= 2\tau F_0(t_n, w_n) - \tau F_0(t_{n-1}, w_{n-1}) \\ &+ \gamma\tau F_1(t_{n+1}, w_{n+1}) + 2(1-\gamma)\tau F_1(t_n, w_n) - (1-\gamma)\tau F_1(t_{n-1}, w_{n-1}) \end{aligned} \quad (4.10)$$

with parameter $\gamma \geq 0$. The order is two and the implicit method is A -stable for $\gamma \geq \frac{3}{4}$. With $\gamma = 1$, $F_0 = 0$ we regain the fully implicit BDF2 method.

In applications we will usually take $\gamma = 1$. Higher-order k -step IMEX-BDF type schemes are obtained starting with the fully implicit k -step BDF scheme together with k th order extrapolation for the explicit term, see Crouzeix (1980), Ascher, Ruuth & Wetton (1995). \diamond

Example 4.5 The third example is based on a class of second-order two-step Adams methods from Example II.3.2 with a parameter γ ,

$$\begin{aligned} w_{n+1} - w_n &= \frac{3}{2}\tau F_0(t_n, w_n) - \frac{1}{2}\tau F_0(t_{n-1}, w_{n-1}) + \gamma\tau F_1(t_{n+1}, w_{n+1}) \\ &+ \left(\frac{3}{2} - 2\gamma\right)\tau F_1(t_n, w_n) + \left(\gamma - \frac{1}{2}\right)\tau F_1(t_{n-1}, w_{n-1}). \end{aligned} \quad (4.11)$$

The explicit method is the two-step Adams-Basforth method. The implicit method is A -stable if $\gamma \geq \frac{1}{2}$. If $\gamma = \frac{1}{2}$ the implicit method reduces to the trapezoidal rule. The choice $\gamma = \frac{9}{16}$ was considered by Ascher et al. (1995); this choice yields maximal damping at $z_1 = \infty$. The implicit method with $\gamma = \frac{3}{4}$ was advocated by Nevanlinna & Liniger (1979) with regard to contractivity for scalar problems. \diamond

In Figure 4.2 the stability regions \mathcal{S}_0 of the explicit methods in (4.10) and (4.11) are plotted together with the regions \mathcal{D}_0 , defined as in (4.5), where we allow for arbitrary $z_1 \in \mathbb{C}^-$. We see from the figure that \mathcal{D}_0 is really smaller than \mathcal{S}_0 and if the implicit method is just A -stable, the region \mathcal{D}_0 reduces to a line. Formulas for the boundary of \mathcal{D}_0 can be found in Frank et al. (1997), where it was also shown that $\mathcal{D}_0 = \mathcal{S}_0$ for the IMEX-CNLF scheme (4.9).

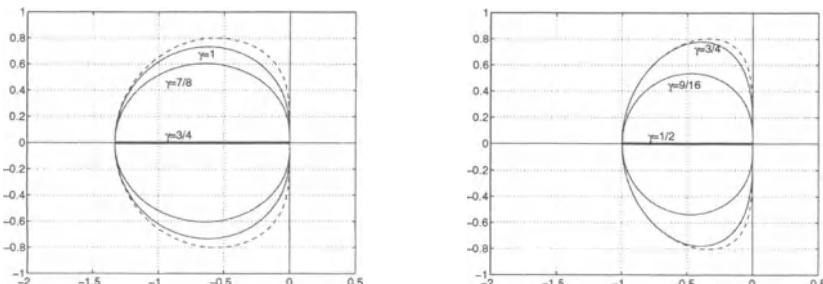


Fig. 4.2. Explicit stability regions \mathcal{S}_0 (dashed) and regions \mathcal{D}_0 for the IMEX-BDF2 methods (4.10) (left) and two-step IMEX-Adams methods (4.11) (right).

Remark 4.6 Stability in actual applications is determined by specific spatial operators and selected spatial discretizations. Often, the non-stiff part F_0 emanates from advection and the stiff part F_1 from reaction and diffusion terms. We here consider the specific case that λ_0 is associated to advection discretized by the first-order upwind or third-order upwind-biased scheme, while λ_1 may still take on arbitrary values in \mathbb{C}^- . For a given IMEX method this leads to a CFL restriction. In a manner similar to Table II.3.1, such restrictions are given in Table 4.1 for the IMEX-BDF2 scheme (4.10) with $\gamma = 1$ and the IMEX-Adams scheme (4.11) with $\gamma = \frac{3}{4}, \frac{9}{16}$ for z_0 in \mathcal{S}_0 or \mathcal{D}_0 . The numbers have been determined experimentally by comparisons of the eigenvalues with the stability sets of Figure 4.2.

	BDF2, $\gamma = 1$		Adams2, $\gamma = \frac{3}{4}$		Adams2, $\gamma = \frac{9}{16}$	
	\mathcal{S}_0	\mathcal{D}_0	\mathcal{S}_0	\mathcal{D}_0	\mathcal{S}_0	\mathcal{D}_0
$z_0 = z_{a,1}$	0.66	0.66	0.50	0.50	0.50	0.50
$z_0 = z_{a,3}$	0.46	0.23	0.58	0.43	0.58	0.16

Table 4.1. CFL restrictions for the IMEX methods (4.10) and (4.11) with first-order upwind $z_{a,1}$ and third-order upwind-biased advection $z_{a,3}$ (as in Table II.3.1).

With regard to applications, the results for the third-order upwind-biased discretization are more important than those for first-order upwind. For the latter discretization the CFL restrictions are the same for \mathcal{S}_0 and \mathcal{D}_0 . With the third-order discretization the requirement of A -stability for the implicit term has a large effect on the allowable Courant number of the IMEX-BDF2 method; even though the region \mathcal{D}_0 seems only slightly smaller than \mathcal{S}_0 in Figure 4.2, this results in a reduction of the maximal Courant number by approximately one half. The reason for this is that eigenvalues of the third-order scheme are very close to the imaginary axis in the vicinity of the origin. In this respect, among the two-step IMEX schemes considered here, the Adams method (4.11) with $\gamma = \frac{3}{4}$ gives the best results.

Central advection discretization of even order leads to purely imaginary eigenvalues. Therefore, among the two-step methods considered here only the Crank-Nicolson Leap-Frog method (4.9) will be stable. For the other two-step methods some upwinding or diffusion is necessary. The Leap-Frog method cannot be used with upwinding, of course. \diamond

Although A -stability is a valuable property, in many practical situations one can settle for less, such as $A(\alpha)$ -stability. Some sufficient analytical results for stability with arbitrary $z_0 \in \mathcal{S}_0$ and with z_1 in a wedge \mathcal{W}_α , i.e., $|\arg(-z_1)| \leq \alpha$, were obtained by Frank et al. (1997). Some pictures of the sets \mathcal{D}_1 , defined as in (4.6), are displayed in Figure 4.3; these pictures were

obtained numerically. By zooming in on the origin one can establish an experimental bound of the admissible angles α for stability with arbitrary $z_0 \in \mathcal{S}_0$ and $z_1 \in \mathcal{W}_\alpha$. For the IMEX-BDF2 method (4.10) with $\gamma = 1$ it was found that $\alpha \approx 0.32\pi$. For the IMEX-Adams method (4.11) with $\gamma = \frac{3}{4}$ and $\gamma = \frac{9}{16}$ the experimental bound was found to be $\alpha \approx 0.30\pi$ and $\alpha \approx 0.14\pi$, respectively. In Frank et al. (1997) it was also shown that for the IMEX-CNLF scheme (4.9) the A -stability property is retained.

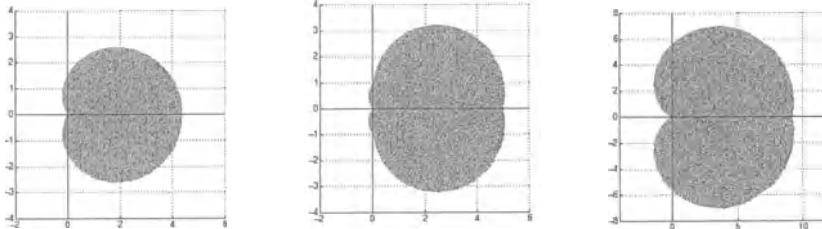


Fig. 4.3. Exterior of shaded region, \mathcal{D}_1 : stability for z_1 with arbitrary $z_0 \in \mathcal{S}_0$ for the IMEX-BDF2 method with $\gamma = 1$ (left) and the two-step IMEX-Adams method with $\gamma = \frac{3}{4}$ (middle) and $\gamma = \frac{9}{16}$ (right).

Above we followed a standard ODE stability analysis in the sense that the eigenvalues λ_0 and λ_1 were allowed to take on arbitrary complex values in certain regions in the complex plane. In actual applications these regions are determined by specific spatial discretizations. Often, the non-stiff part F_0 emanates from advection and the stiff part F_1 from reaction-diffusion terms. For example, for atmospheric transport-chemistry models a useful test model is the system $u_t + a_1 u_x + a_2 u_y = \epsilon u_{zz} + f(u)$ where u is a vector of chemical concentrations, $a_1 u_x + a_2 u_y$ models advection in a horizontal wind field, ϵu_{zz} models a vertical diffusion process, which includes parameterized turbulence, and $f(u)$ is a set of atmospheric stiff chemical reactions. Application of IMEX schemes to atmospheric problems has been investigated in Verwer et al. (1996).

Results for 1D linear advection-diffusion equations $u_t + a u_x = d u_{xx}$ with constant coefficients, based on Fourier decompositions as in Remark 4.1, can be found in Varah (1980), Ascher et al. (1995). Sufficient conditions for multi-dimensional problems with constant coefficients are found in Wesseling (2001). More general stability results, valid for noncommuting operators, are given in Crouzeix (1980). Generalizations to problems that are nonlinear in F_0 can be found in Akrivis et al. (1999).

Remark 4.7 Multistep methods for splittings (3.5) with s implicit terms can be derived by using the stabilizing corrections idea, where one starts with an explicit scheme and then adds implicitness as corrections. For $s = 1$ this leads to IMEX methods. For $s \geq 2$, however, this seems to give quite

poor stability properties, essentially restricting such schemes to problems like the heat equation with dimension splitting for which all implicit terms are negative definite; see Hundsdorfer (2002) for details.

Different splitting schemes based on multistep methods have been derived by Warming & Beam (1979), again valid with multiple implicit terms. These methods appear more stable, see van der Houwen & Sommeijer (2001), but they contain internal vectors that are not consistent with the full problem, similar to the LOD methods of Section 2. For good convergence properties this again requires construction of special boundary conditions, which is unattractive in comparison with the consistency properties of Theorem 4.2 for the IMEX schemes. \diamond

4.3 Notes on IMEX Runge-Kutta Methods

In this section we have thus far discussed only multistep schemes, in particular two-step ones. Combinations of explicit and implicit methods can also be considered for Runge-Kutta methods. The IMEX- θ method (4.2) provides the most simple example. Here we consider some slightly more sophisticated schemes based on a mixture of the implicit trapezoidal rule and its explicit counterpart (II.1.6), the explicit trapezoidal rule.

Example 4.8 Combining the implicit and explicit trapezoidal rule can be done in various ways. First we consider the method

$$\begin{aligned} w_{n+1}^* &= w_n + \tau F_0(t_n, w_n) + \frac{1}{2}\tau F_1(t_n, w_n) + \frac{1}{2}\tau F_1(t_{n+1}, w_{n+1}^*), \\ w_{n+1} &= w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_{n+1}^*). \end{aligned} \quad (4.12)$$

It can be seen by some calculations that the method has order two in the classical ODE sense. If we apply (4.12) to the scalar test equation $w'(t) = (\lambda_0 + \lambda_1)w(t)$ we get $w_{n+1} = R w_n$ with $R = R(z_0, z_1)$, $z_j = \tau \lambda_j$,

$$R(z_0, z_1) = 1 + \frac{1 + \frac{1}{2}z_0}{1 - \frac{1}{2}z_1} (z_0 + z_1). \quad (4.13)$$

In the limit $z_1 \rightarrow -\infty$ we have $|R| = |1 + z_0|$, and thus the explicit Euler stability restriction $|1 + z_0| \leq 1$ should then be obeyed. This already shows that the IMEX scheme (4.12) will not be suited for advection combined with stiff reactions if the advection term is treated explicitly and discretized with higher-order differences. On the other hand, there are problems where the scheme does perform well; an example will be provided below by a reaction-diffusion problem with non-stiff reaction.

A detailed error analysis for linear problems with $F_j(t, v) = A_j v + g_j(t)$, $j = 0, 1$, can be performed along the same lines as in the previous sections. Let $Z_j = \tau A_j$, $Z = Z_0 + Z_1$ and assume that we have an inner product norm with $\langle v, Z_1 v \rangle \leq 0$ for all v and $Z_0 = \mathcal{O}(1)$, where, as before, the order symbol

$\mathcal{O}(\tau^q)$ refers to a norm estimate uniformly in the mesh width h . If we insert the exact solution into (4.12), then by using residual errors on the stages we obtain after some manipulation the recursion $\varepsilon_{n+1} = R\varepsilon_n + \delta_n$ for the global errors $\varepsilon_n = w(t_n) - w_n$ with amplification matrix

$$R = I + Z(I - \frac{1}{2}Z_1)^{-1}(I + \frac{1}{2}Z_0)$$

and local discretization errors given by

$$\delta_n = \frac{1}{4}\tau^2 Z(I - \frac{1}{2}Z_1)^{-1}\varphi'(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3),$$

where $\varphi(t) = F_0(t, w(t))$. Hence we can write $\delta_n = (R - I)\xi_n + \mathcal{O}(\tau^3)$ provided

$$(I + \frac{1}{2}Z_0)\xi_n = \frac{1}{4}\tau^2\varphi'(t_{n+\frac{1}{2}}).$$

Assuming stability, the decomposition (2.24) shows that the essential condition for convergence with order two is $\xi_n = \mathcal{O}(\tau^2)$. This condition is not that different from the condition for the explicit trapezoidal rule itself (obtained if $F_1 \equiv 0$, in which case $\varphi(t) = w'(t)$). Stability for non-commuting matrices seems difficult to establish however. \diamond

Example 4.9 Another way to combine the implicit and explicit trapezoidal rule is by

$$\begin{aligned} w_{n+1}^* &= w_n + \tau F(t_n, w_n), \\ w_{n+1} &= w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F_0(t_{n+1}, w_{n+1}^*) + \frac{1}{2}\tau F_1(t_{n+1}, w_{n+1}). \end{aligned} \quad (4.14)$$

Here the second, final stage is implicit in F_1 . For the scalar test equation we find again the stability function (4.13).

For systems of linear equations, using the notation of the previous example, we find the global error recursion $\varepsilon_{n+1} = R\varepsilon_n + \delta_n$ with

$$R = I + (I - \frac{1}{2}Z_1)^{-1}(I + \frac{1}{2}Z_0)Z,$$

$$\delta_n = \frac{1}{4}\tau^2(I - \frac{1}{2}Z_1)^{-1}Z_0w''(t_{n+\frac{1}{2}}) + \mathcal{O}(\tau^3).$$

Although the local errors are slightly different from the previous example, the conditions for second-order convergence are comparable.

In a numerical test below, for a reaction-diffusion problem, only results for method (4.12) will be presented; the results for (4.14) were very much the same. Examples where either (4.12) or (4.14) outperforms the other one can be constructed. Preference for (4.12) or (4.14) may therefore depend on the problem class at hand. \diamond

Higher-order combinations of explicit and diagonally implicit Runge-Kutta methods have been derived by Zhong (1996), Ascher, Ruuth & Spiteri

(1997), Calvo, de Frutos & Novo (2001) and Kennedy & Carpenter (2003), for example. Related methods were also obtained by Knoth & Wolke (1998). The stability derivations in these papers mostly deal with the model advection-diffusion equation $u_t + au_x = du_{xx}$ through Fourier analysis, as in Remark 4.1. An analysis for scalar, independent z_0, z_1 for several IMEX Runge-Kutta schemes can be found in Pareschi & Russo (2000). Linear stability and convergence results for non-commuting matrices, formulated in terms of Rosenbrock methods, have been obtained by Ostermann (2002).

At present no clear device exists to construct such IMEX Runge-Kutta pairs. For this reason we will not discuss these IMEX methods in detail. Instead, in the next section several Rosenbrock methods with special Jacobian approximation will be introduced, which have the additional advantage that also multiple splittings of the implicit terms are allowed in a natural way. In case only one implicit term is present ($s = 1$) these formulas can be viewed as linearized IMEX Runge-Kutta methods.

4.4 Concluding Remarks and Tests

Modified Newton Iterations

The implementation of an IMEX scheme (4.8) will in general involve a Newton-type iteration with Jacobian matrix $I - \theta\tau A_1$, where $\theta = \beta_k/\alpha_k$, $A_1 = \partial_w F_1(t_{n+1}, \bar{w}_{n+1})$ and \bar{w}_{n+1} is a prediction for w_{n+1} . We can of course also apply a Newton type iteration with this same Jacobian matrix to the fully implicit multistep scheme (4.7). If this iteration converges, then the favourable stability properties of the implicit schemes are fully employed. For atmospheric dispersion problems this approach has been examined in Ahmad & Berzins (1997). Convergence of such iterations will be discussed in a more general context in Section 5.4 (see Remark 5.3).

Assuming convergence, the difference with the IMEX approach will be that more evaluations of F are required, but if such evaluations are computationally cheap then this approach can be attractive. On the other hand, if evaluations of F are expensive, as with advection discretizations with limiters, then the IMEX approach seems more promising.

Numerical Illustration

As a test model for the IMEX schemes, we consider the following system of reaction-diffusion equations on the unit square $\Omega = (0, 1)^2$,

$$\begin{aligned} u_t &= D_1 \Delta u + \kappa(a - u + u^2 v), \\ v_t &= D_2 \Delta v + \kappa(b - u^2 v). \end{aligned} \tag{4.15}$$

The initial values are given by

$$u(0, x, y) = a + b + 10^{-3} e^{-100((x-\frac{1}{3})^2 + (y-\frac{1}{2})^2)}, \quad v(0, x, y) = b/(a+b)^2,$$

and the boundary conditions are taken as homogeneous Neumann. The parameter values are $\kappa = 100$, $a = 0.1305$, $b = 0.7695$, $D_1 = 0.05$, $D_2 = 1$. The initial value consists of a small Gaussian perturbation on top of a chemical steady state ($u \equiv 0.90$, $v \equiv 0.95$). Due to reaction and diffusion this initial perturbation is amplified and spread, leading to a formation of spots, which then slowly move and interact. The time evolution of the u -component is illustrated in Figure 4.4 with contour lines at levels $0.5, 1, \dots, 2.5$ in the (x, y) -plane. The peak values for u are 2.8, approximately; the values for v range between 0.4 and 1.1, with smallest values in the spots where u is maximal.

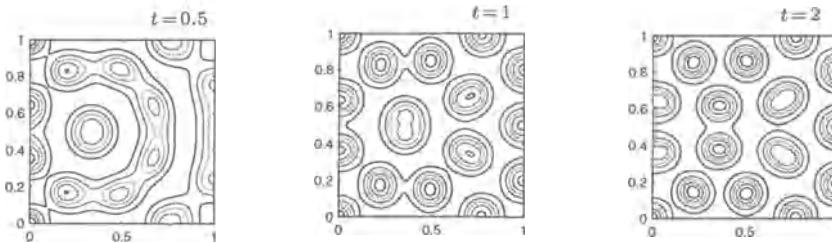


Fig. 4.4. Time evolution u -component for the reaction-diffusion model (4.15).

The present model is due to Schnakenberg (1979). It is somewhat related to the Gray-Scott model considered in Section I.1.4. Numerically, however, there is a clear difference: the present model is stiff, already on rather coarse grids, due to the relatively large diffusion coefficients. Therefore we consider this model here to test some of the IMEX schemes, treating the diffusion terms implicitly and the reactions explicitly. In the tests the L_2 -errors in the solution at time $t = 1$ will be examined. Related tests and an analysis of qualitative behaviour of IMEX multistep schemes for pattern formation can be found in Ruuth (1995).

The equation is discretized on a cell-centered 200×200 grid using standard second-order differences in space for the diffusion terms. Let A_D stand for the discretized Laplace operator with the homogeneous Neumann conditions incorporated. With the IMEX schemes we have to solve each time step linear systems of the form $(I - \theta\tau A_D)w = b$ with given b , and for this the fast Poisson solver FISHPACK was used; see Remark 4.10 below for some details. To make the comparison between the various methods easier we used constant step sizes, but it should be noted that for an efficient implementation variable step sizes would be advisable here since most of the activity in the model occurs around time $t = 0.5$; before and afterwards the evolution is slower.

The temporal L_2 -errors in the solution at time $t = 1$ are plotted in Figure 4.5 as function of the CPU times in seconds. For the various time stepping schemes we used constant step sizes $\tau = 1/N$ with $N = 2^k 100$, $k \geq 0$; for some of the schemes additional numbers of steps were considered if insta-

bilities did arise within this range of step sizes. In the figure the horizontal dashed line indicates the spatial error, which is $2.0 \cdot 10^{-3}$ approximately.

In Figure 4.5(a) we have the results for the IMEX- θ scheme (4.2) with θ equal to $\frac{1}{2}, 1$. Due to the explicit Euler method for the reaction term these results are rather inaccurate and very small time steps are needed to achieve a temporal error close to the spatial error $2 \cdot 10^{-3}$. Much better results are obtained with the two-stage IMEX trapezoidal rule (4.12) by which the reaction term is treated as with the explicit trapezoidal rule.

The results for the IMEX-BDF2 scheme (4.10), $\gamma = 1$, are given in Figure 4.5(b). Although somewhat less accurate in this test than the IMEX trapezoidal rule (4.12), these results are satisfactory. The two-step IMEX Adams schemes (4.9) turned out to be slightly more accurate in this test, in particular for $\gamma = \frac{1}{2}$, but the overall behaviour was quite close to (4.10) and therefore these results have been omitted in the figure. Much better results were obtained in this test with the IMEX-BDF3 scheme which combines (II.3.12) and (II.3.14). Apparently the higher order pays off here.

With these multistep schemes the results were found to be sensitive to the starting procedure. For the IMEX-BDF2 scheme the first approximation was computed by the IMEX- θ scheme with $\theta = 1$, that is, the IMEX-BDF1 scheme. In Figure 4.5(b) there are additional results for IMEX-BDF2, indicated with dashed lines, where (4.12) was used to compute w_1 ; for the larger step sizes this gave somewhat smaller errors. For the IMEX-BDF3 scheme we used IMEX-BDF1 for the first step, but in an extrapolated fashion (first a w_1^* was computed with step size τ , then a w_1^{**} by two steps with $\frac{1}{2}\tau$, and finally $w_1 = 2w_1^{**} - w_1^*$). For the second step we used IMEX-BDF2. The sensitivity for starting procedures is to some extent due to the fact that we are dealing here with constant step sizes. Usually, a multistep scheme is started with a small time step, which is then gradually increased.

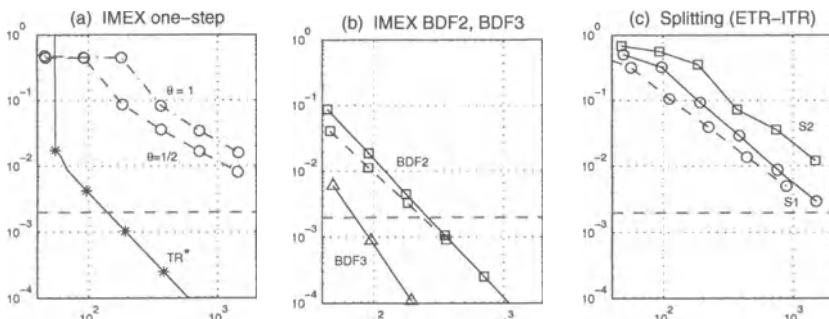


Fig. 4.5. Errors versus CPU (seconds) for the reaction-diffusion model (4.15). Results: (a) IMEX- θ with $\theta = 1, \frac{1}{2}$ and method (4.12) (indicated as TR*); (b) IMEX-BDF2 and IMEX-BDF3; (c) First-order splitting (S1) and Strang splitting (S2) with explicit and implicit trapezoidal rule. See text for details.

In Figure 4.5(c) the errors are shown for time splitting where the fractional steps were solved by one step with the explicit trapezoidal rule for reaction and with the implicit trapezoidal rule for diffusion. These results are very inaccurate, in particular for the second-order Strang splitting. Apparently there is some form of order reduction here, see also the remarks in Section 1.7. The sequence in which reaction and diffusion steps were taken was found to be not very relevant (for this particular problem). For the first-order splitting some improvement was obtained when the fractional reaction steps were solved by the explicit trapezoidal rule with smaller (sub-)time steps. In the figure, the dashed line corresponds with 5 sub-steps by the explicit trapezoidal rule for the reaction. With Strang splitting this did not lead to a significant improvement.

In conclusion, it can be said that for the present reaction-diffusion model the IMEX- θ scheme and the time splitting schemes were not sufficiently accurate (in particular the Strang splitting); to achieve a temporal error comparable to the spatial error quite small step sizes were needed. The results with the other IMEX schemes seem satisfactory. As said before, an efficient implementation would require variable step sizes for this problem. Constant steps were taken here to facilitate the comparison between the various schemes, otherwise the different error estimators also become important in the comparison.

With the IMEX schemes, the local errors depend not only on derivatives of $w(t)$, but also on derivatives of $\varphi(t) = F_0(t, w(t))$. As noted above, the temporal evolution is rather slow towards $t = 1$. On the other hand, the solution is far from chemical steady state, and this indicates that the derivatives of $\varphi(t)$ may be much larger than the derivatives of $w(t)$. Therefore, a fully implicit method might be more efficient here. To some extent this is true. We implemented the trapezoidal rule with a modified Newton iteration, using only the diffusion terms in the Jacobians and an explicit Euler initiation. The iteration was terminated if the L_2 -residual was less than $10^{-2}\tau^2$. This stopping criterion was based on the observation that with large tolerances a wrong solution was sometimes returned, whereas a very small tolerance leads to unnecessary iterations. Still, the stopping criterion is somewhat arbitrary. With a variable time step this matter does not show up in this way; usually one then takes the tolerance to be a certain fraction, say 10%, of the estimated local error.

The results for the implicit trapezoidal rule are given in Table 4.2. Indeed for a given time step, the errors with the implicit scheme are much smaller than for the corresponding IMEX scheme (4.12), but on the other hand more work is needed per step. For low accuracies the IMEX scheme is more efficient here, but for higher accuracies the implicit scheme quickly becomes more efficient. For smaller step sizes fewer Newton iterations per step are needed, even though the tolerance here is proportional to τ^2 . To achieve a temporal

N	100	200	400	800
TR impl.	$4.8 \cdot 10^{-3}$ (253)	$6.5 \cdot 10^{-4}$ (240)	$1.8 \cdot 10^{-4}$ (328)	$3.8 \cdot 10^{-5}$ (531)
TR-IMEX	$4.2 \cdot 10^{-3}$ (96)	$1.0 \cdot 10^{-3}$ (191)	$2.5 \cdot 10^{-4}$ (379)	$5.9 \cdot 10^{-5}$ (724)

Table 4.2. Results for the implicit trapezoidal rule (step size $\tau = \frac{1}{N}$) and the IMEX scheme (4.8) ($\tau = \frac{1}{4N}$). The entries are the temporal L_2 -errors together with the CPU times (seconds) in brackets.

error comparable to the spatial error ($2.0 \cdot 10^{-3}$ on the 200×200 grid) the two schemes are quite similar in efficiency.

A similar comparison was also found for the implicit BDF2 scheme and its IMEX counterpart. Comparisons between very different types of methods, like fully implicit and IMEX, will depend very much on implementations and also on the given class of problems. The fact that the IMEX schemes are less accurate here than the implicit methods is not a generic situation. If the derivatives of $\varphi(t)$ are more comparable in size to those of $w(t)$, then the accuracy of the IMEX schemes will also be closer to the accuracy of the corresponding implicit methods.

Finally we note that for the present problem, a fully explicit scheme like the explicit trapezoidal rule would need a step size $\tau \leq \frac{1}{4} h^2 = 6.25 \cdot 10^{-6}$, leading to large CPU times. On finer grids this becomes even worse, of course.

Remark 4.10 In these tests, the arising linear systems were solved by the code FISHPACK, written by Adams, Swarztrauber and Sweet. It is available at

<http://www.netlib.org/fishpack/index.html>

This code is based on the Buneman cyclic reduction algorithm, see Golub & van Loan (1996) or Stoer & Bulirsch (1980). It solves Helmholtz equations $\Delta u - cu = f$ on a square region with second-order differences, and it does it very fast. On an $m \times m$ grid the code turned out to be numerically stable (w.r.t. round-off) up to $m = 500$ in single precision and $m = 2000$ in double precision. The Helmholtz equation is of course more general than the Poisson equation $\Delta u = f$, but codes like FISHPACK are commonly known as *fast Poisson solvers*.

The use of this fast solver, or alternative schemes based on FFT, is confined to constant coefficient Helmholtz equations. For more general systems, preconditioned iterative solvers are commonly used, see Golub & van Loan (1996) and the ‘templates’ in Barrett et al. (1994). We decided to use FISHPACK based on the comparisons in Botta et al. (1997), which show that for special problems like the Poisson equation, tailored fast solvers like FISHPACK are indeed very much faster than more general solvers. ◇

5 Rosenbrock AMF Methods

In this section we discuss a number of splitting methods derived from the Rosenbrock methods with inexact Jacobian matrices that have been introduced in Section II.1.5.¹⁵⁾ Splitting is realized here by choosing special approximations to the Jacobian matrix A used in these formulas. The basic idea is to simplify and economize the linear system solutions with the involved matrix $I - \gamma\tau A$. This is achieved by omitting entries in A and in particular by factorizing the matrix $I - \gamma\tau A$. The common name for this technique is *approximate matrix factorization* (AMF).

The use of approximate matrix factorization is classic and is related to some of the previously discussed LOD and ADI methods. In particular for the most simple Rosenbrock method there is a close relation with the Douglas method (3.6). Two early papers where AMF is applied are due to D'Yakonov (1964) and Beam & Warming (1976). A recent survey has been given by van der Houwen & Sommeijer (2001).

Because the Rosenbrock methods have been formulated for autonomous systems, in this section we primarily focus on autonomous, semi-discrete systems

$$w'(t) = F(w(t)) \equiv F_0(w(t)) + F_1(w(t)) + \cdots + F_s(w(t)), \quad (5.1)$$

where F_0 is a non-stiff term. Time-dependent terms will be treated by a transformation to an augmented autonomous form. In case $s = 1$ the methods in this section can be viewed as IMEX type schemes.

5.1 One-Stage Methods of Order One and Two

We first consider the one-stage Rosenbrock method from Example II.1.9,

$$w_{n+1} = w_n + \tau B^{-1} F(w_n), \quad B = I - \gamma\tau A, \quad (5.2)$$

where A is an approximation to $F'(w_n)$. Let us assume

$$A_j = F'_j(w_n) + \mathcal{O}(\tau), \quad j = 1, \dots, s, \quad (5.3)$$

and replace $B = I - \gamma\tau A$ in (5.2) by the approximate factorization

$$B = (I - \gamma\tau A_1) (I - \gamma\tau A_2) \cdots (I - \gamma\tau A_s). \quad (5.4)$$

The resulting scheme thus reads

$$w_{n+1} = w_n + (I - \gamma\tau A_s)^{-1} \cdots (I - \gamma\tau A_1)^{-1} \tau F(w_n). \quad (5.5)$$

¹⁵⁾ Rosenbrock methods employing arbitrary Jacobian approximations are sometimes called W -methods, see Steihaug & Wolfbrandt (1979) and Hairer & Wanner (1996, Sect. IV.7).

Note that the non-stiff Jacobian $F'_0(w_n)$ is not present here. So, loosely speaking, the F_0 term is treated explicitly (as in explicit Euler) and the other terms linearly implicitly one after another. Because F itself is not split, stationary solutions \bar{w} satisfying $F(\bar{w}) = 0$ are maintained. Also note that in a concrete case a change in sequence of the factors $I - \gamma\tau A_j$ gives a different algorithm, unless the matrices A_j commute.

With this approximate factorization the order of consistency (in the classical ODE sense) is one in general for any γ . We do get order two if $\gamma = \frac{1}{2}$ and $F_0 = 0$. Formally (5.4) can be identified with a special choice of $A = (\gamma\tau)^{-1}(I - B)$, which equals $F'(w_n) + \mathcal{O}(\tau)$ in case $F_0 = 0$.

If the problem is linear this one-stage Rosenbrock AMF method is identical to the Douglas ADI method (3.6) and for $s = 1$ to the IMEX- θ method (4.2) (in both cases with $\theta = \gamma$). Hence the linear stability and convergence properties relevant to this Rosenbrock AMF method can be found in Section 3.2 for $s > 1$ (mainly with $F_0 = 0$) and in Section 4.1 for $s = 1$.

To apply the method to a non-autonomous problem $w'(t) = F(t, w(t))$ we first rewrite this in the augmented autonomous form

$$v'(t) = G(v(t)) \quad \text{with} \quad v = \begin{pmatrix} t \\ w \end{pmatrix}, \quad G(v) = \begin{pmatrix} 1 \\ F(t, w) \end{pmatrix},$$

to which the method can be applied. Then t is formally also considered as an unknown, but it is easily seen that the approximations t_n found with this method still equal $n\tau$. When reformulated in terms of w_n , the methods will now also involve approximations to the time derivatives $F_t(t, w)$. Taking $A_j \approx \partial_w F_j(t_{n+\gamma}, w_n) \in \mathbb{R}^{m \times m}$, $b_j \approx \partial_t F_j(t_{n+\gamma}, w_n) \in \mathbb{R}^m$ and

$$B_j^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ (I - \gamma\tau A_j)^{-1} \gamma\tau b_j & (I - \gamma\tau A_j)^{-1} \end{pmatrix} \in \mathbb{R}^{(m+1) \times (m+1)},$$

the factorized Rosenbrock scheme (5.2) now reads

$$\begin{pmatrix} t_{n+1} \\ w_{n+1} \end{pmatrix} = \begin{pmatrix} t_n \\ w_n \end{pmatrix} + B_s^{-1} \cdots B_2^{-1} B_1^{-1} \begin{pmatrix} \tau \\ \tau F(t_n, w_n) \end{pmatrix}.$$

We will have $t_{n+1} = t_n + \tau$, as it should be, and the computation of w_{n+1} can be written in the more transparent recursive form with increments dv_j ,

$$\begin{aligned} dv_0 &= \tau F(t_n, w_n), \\ dv_j &= (I - \gamma\tau A_j)^{-1} (dv_{j-1} + \gamma\tau^2 b_j), \quad j = 1, \dots, s, \\ w_{n+1} &= w_n + dv_s. \end{aligned} \tag{5.6}$$

Setting $v_j = w_n + dv_j$, we see once more the close relation with the Douglas scheme (3.6) for linear problems.

A disadvantage of this one-stage Rosenbrock method is that when the non-stiff term F_0 represents semi-discrete advection, the method is of limited

use because of the connection with the explicit Euler method, which is known to be unstable with second-order central and higher-order upwind-biased spatial advection discretizations. This instability can easily manifest itself if the one-stage Rosenbrock AMF method is applied with F_1, \dots, F_s representing diffusion and reaction terms. We therefore proceed with a two-stage method connected with the explicit trapezoidal rule, which is stable when combined with the third-order upwind-biased advection scheme under the appropriate CFL restriction, see Table II.1.2.

5.2 Two-Stage Methods of Order Two and Three

Consider the 2-stage method (II.1.27). Before we introduce the approximate factorization we remove the matrix vector multiplication Ak_1 in the second stage. To that end we substitute $k_1 = \tilde{k}_1, k_2 = \tilde{k}_2 - \gamma_{21}k_1/\gamma$. Omitting the tilde and imposing the relations for order two then gives the method

$$\begin{aligned} w_{n+1} &= w_n + (2 - b_2)k_1 + b_2k_2, \\ B k_1 &= \tau F(w_n), \\ B k_2 &= \tau F(w_n + \frac{1}{2b_2}k_1) - \frac{1}{b_2}k_1, \end{aligned} \tag{5.7}$$

where $B = I - \gamma\tau A$ and γ, b_2 are free parameters. With the approximate factorization (5.4) this method remains of order two (in the ODE sense), even with $F_0 \neq 0$, because the method is of second order for any A and (5.4) can be interpreted as a special choice of A . For the second-order methods we take $b_2 = \frac{1}{2}$ so that the method includes the explicit trapezoidal rule for $A = 0$. Further we will take $\gamma \geq \frac{1}{4}$ and use the approximate factorization (5.4). As will be seen below, the choice of γ is important for stability.

Trivially, method (5.7) provided with matrix factorization returns stationary solutions exactly because splitting of F itself is not used. The AMF methods can also be applied to non-autonomous problems $w'(t) = F(t, w(t))$ in the same way as for the one-stage method. For both stages this leads to a recursion of the type (5.6).

An AMF counterpart of the third-order method (II.1.29) could be obtained with $b_2 = \frac{3}{4}, \gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. However, for order three it is necessary to require $F_0 = 0$, since in the order conditions for (II.1.29) it was supposed that $A = F'(w_n) + \mathcal{O}(\tau)$. With the factorization (5.4) this will only be satisfied if $F_0 = 0$.

Stability

The non-factorized method (5.7) is A -stable for any $\gamma \geq \frac{1}{4}$. With matrix factorization the stability properties change. As for the Douglas ADI method,

consider the scalar test model (3.7). Applying (5.7) to this test model gives the stability function

$$R(z_0, z_1, \dots, z_s) = 1 + 2\frac{z}{\varpi} - \frac{z}{\varpi^2} + \frac{z^2}{2\varpi^2}, \quad (5.8)$$

where

$$z = \sum_{i=0}^s z_i, \quad \varpi = \prod_{i=1}^s (1 - \gamma z_i),$$

and γ is still free (the parameter b_2 has cancelled in this expression). If $s = 0$ we get $R = P(z_0) = 1 + z_0 + \frac{1}{2}z_0^2$, the stability function of the explicit trapezoidal rule. Putting $\varpi = 1 - \gamma z$ returns the stability function (II.1.18) of the unfactorized methods.

As for the Douglas method, we consider the wedge \mathcal{W}_α in the left half-plane, consisting of $\zeta \in \mathbb{C}$ with $|\arg(-\zeta)| \leq \alpha$, and examine stability under the condition

$$z_i \in \mathcal{W}_\alpha, \quad 1 \leq i \leq s \quad \text{with either} \quad z_0 = 0 \quad \text{or} \quad |1 + z_0 + \frac{1}{2}z_0^2| \leq 1.$$

Since only in a few cases analytical results on lower bounds for the admissible angles α can be obtained, we present computed results in the Figures 5.1, 5.2. In these figures the fraction a corresponding to the angle $\alpha = a\pi/2$ is plotted as function of the method parameter γ . These admissible angles have been determined in an experimental way by computing the maximal value of $|R|$ for sufficiently many points $|\arg(-z_i)| = \alpha$.

First consider Figure 5.1. This figure is of interest for the second-order Rosenbrock AMF method with $F_0 \neq 0$. For F_0 one can think of a semi-discrete advection term or another mildly stiff term. Note that the complex number z_0 associated with F_0 runs freely in the stability region of the explicit trapezoidal rule and the remaining z_i run freely in the wedge \mathcal{W}_α .

A few remarks are in order. If $s = 1$, that is, if we have one explicitly and one implicitly treated term, then A -stability for the implicit term is preserved with $\gamma = \frac{1}{2}$. This is surprising in view of the fact that without factorization the stability function is just A -stable and does not damp at infinity. This

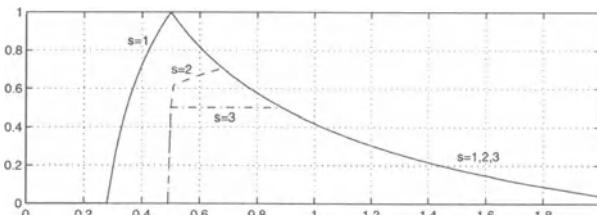


Fig. 5.1. Fractions $a = 2\alpha/\pi$ versus γ for $s = 1, 2, 3$ and $|1 + z_0 + \frac{1}{2}z_0^2| \leq 1$.

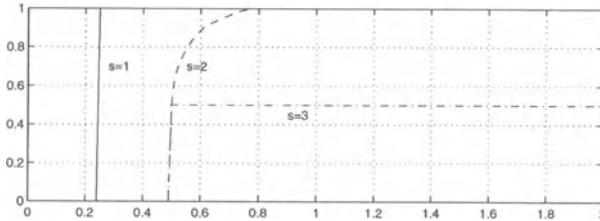


Fig. 5.2. Fractions $a = 2\alpha/\pi$ versus γ for $s = 1, 2, 3$ and $z_0 = 0$.

result can be seen as follows: for $s = 1$ we can write $R = R(z_0, z_1)$ as

$$R = \frac{(1 + z_0 + \frac{1}{2}z_0^2) + (1 - 2\gamma)(1 + z_0)z_1 + (\frac{1}{2} - 2\gamma + \gamma^2)z_1^2}{(1 - \gamma z_1)^2}, \quad (5.9)$$

and with $\gamma = \frac{1}{2}$, $|1 + z_0 + \frac{1}{2}z_0^2| \leq 1$ we therefore have

$$|R| \leq \frac{1 + \frac{1}{4}|z_1|^2}{|1 - \frac{1}{2}z_1|^2} \leq 1 \quad \text{whenever } \operatorname{Re}(z_1) \leq 0.$$

Also noteworthy is that if we have two or three implicitly treated terms, we get $\alpha = 0$ for $\frac{1}{4} \leq \gamma \leq \frac{1}{2}$. This happens with $z_1 \approx (2\gamma - 1)/(2\gamma^2)$ and $z_2 \rightarrow -\infty$.

Next consider Figure 5.2 which is of interest for both the second-order and third-order Rosenbrock AMF method with $F_0 = 0$. Now $s = 1$ means that we are looking at the original unfactorized methods, and thus $\alpha = \frac{1}{2}\pi$ for $\gamma \geq \frac{1}{4}$. For $s = 2$ we have $\alpha = \frac{1}{2}\pi$ if γ exceeds a threshold, which turns out to be $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$, see Lanser et al. (2001), the value of γ defining the third-order method. So, with $F_0 = 0$ and $s = 2$, both the second- and third-order Rosenbrock method can preserve the A -stability property with AMF. If $s = 3$ the situation is completely different: the A -stability property is lost and $\alpha = \frac{1}{4}\pi$ for $\gamma \geq \frac{1}{2}$.

Remark 5.1 Angle barriers $(s - 1)\alpha \leq \frac{1}{2}\pi$ for stability were encountered already in the inequality (3.9) for the Douglas scheme and consequently also for the one-stage method (5.2) with approximate factorization. This barrier is universal for one-step AMF schemes.

To demonstrate this, let $z = z_1 + z_2 + \dots + z_s$ and consider the function $R = R(z_1, \dots, z_s)$ given by

$$R = 1 + \frac{\phi(z_1, z_2, \dots, z_s)}{\psi_1(z_1)\psi_2(z_2)\dots\psi_s(z_s)} z,$$

where ϕ is a polynomial in each z_j and the ψ_j are all polynomials without zeros in the left half-plane. This is the general form of a stability function of a one-step method with approximate factorization of implicit terms.

Now consider equal $z_k = -t e^{i\beta}$, $k = 1, \dots, s$, with $0 \leq \beta \leq \alpha$ and assume that

$$\frac{\phi(-t e^{i\beta}, \dots, -t e^{i\beta})}{\psi_1(-t e^{i\beta}) \cdots \psi_s(-t e^{i\beta})} = C(t e^{i\beta})^{-r} + \mathcal{O}(t^{-r-1}), \quad t \rightarrow \infty,$$

with integer r and constant $C \neq 0$. Then, for $t \rightarrow \infty$,

$$R = 1 - s C t^{1-r} e^{i(1-r)\beta} + \mathcal{O}(t^{-r}).$$

Hence stability for all $\beta \leq \alpha$ requires $C > 0$ and $|r-1|\alpha \leq \frac{1}{2}\pi$. Stability for fixed $z_k < 0$ ($k \neq j$) and $z_j \rightarrow -\infty$ implies that the degree of ϕ in z_j is less than the degree of ψ_j . Consequently $r \geq s$ and thus we get the condition

$$\alpha \leq \frac{1}{s-1} \frac{\pi}{2},$$

the same upper bound as in (3.9). \diamond

Remark 5.2 The second-order Rosenbrock AMF method has been proposed in Verwer et al. (1999) where it was applied to some 3D atmospheric transport-chemistry problems. There the method was used with $s = 2$ and with F_0 representing advection, F_1 diffusion, and F_2 reaction. The free parameter γ was taken as $\gamma = 1 + \frac{1}{2}\sqrt{2}$ to have optimal damping (L -stability). The eigenvalues of F_1 and F_2 were negative real and therefore stability problems were not expected, and indeed did not occur as long as the CFL condition for the explicit trapezoidal rule was satisfied. It is for these kinds of problems, where the structure of the eigenvalue spectra can be well predicted in advance, that these Rosenbrock AMF methods are primarily suited. In general, values γ in the range $[\frac{1}{2}, 1]$ seem preferable over $\gamma = 1 + \frac{1}{2}\sqrt{2}$, because this value gives relatively large error constants.

The A -stable third-order Rosenbrock AMF method has been proposed in Lanser et al. (2001) where it was successfully applied to the spherical shallow water equations, which is a prototype model for atmospheric circulation. For this two-dimensional hyperbolic problem fast gravity waves impede explicit time stepping and the AMF property combined with A -stability allows large time steps with computational efficiency.

For $s = 1$ rather general linear stability results have been obtained by Ostermann (2002). The essential assumption in that analysis is a sectorial bound for A_1 together with a bound $\|A_1^{-1/2} A_0\| \leq C$, which will hold for example with advection-diffusion in the same direction. If A_0, A_1 act in different directions or if $s \geq 2$, scalar stability results as presented in this section give necessary conditions for stability in more general cases. \diamond

5.3 A Three-Stage Method of Order Two

Using for A the zero matrix in the second-order method (5.7) yields the explicit trapezoidal rule. With regard to stability and positivity improved

second-order explicit Runge-Kutta methods are found in the class (II.4.15). Taking the 3-stage member of this class, Gerisch & Verwer (2002) have constructed a 3-stage Rosenbrock AMF method which has order two for arbitrary Jacobian approximations A , similar to method (5.7).

In standard form, using the notation from (II.1.25), this Rosenbrock method reads

$$\begin{aligned} w_{n+1} &= w_n + \frac{1}{3}(k_1 + k_2 + k_3), \\ (I - \gamma\tau A) k_i &= \tau F\left(w_n + \frac{1}{2} \sum_{j=1}^{i-1} k_j\right) + \tau A \sum_{j=1}^{i-1} \gamma_{ij} k_j, \quad i = 1, 2, 3, \end{aligned} \quad (5.10)$$

where

$$\begin{aligned} \gamma_{21} &= -(3\gamma + \gamma_{31} + \gamma_{32}), \quad \gamma_{32} = \frac{1}{2} - 3\gamma, \\ \gamma_{31} &= \frac{-1}{1 + 2\gamma_{32}} \left(6\gamma^3 - 12\gamma^2 + 6(1 + \gamma_{32})\gamma + 2\gamma_{32}^2 - \frac{1}{2} \right), \\ \gamma &= 1 - \frac{1}{2}\sqrt{2}\cos\theta + \frac{1}{2}\sqrt{6}\sin\theta, \quad \theta = \frac{1}{3}\arctan\left(\frac{1}{4}\sqrt{2}\right). \end{aligned}$$

The value for γ is approximately 0.43586652. The parameters are chosen such that the method is L -stable and of order 3 for homogeneous linear problems with constant coefficients. For its implementation the scheme is to be rearranged such that A appears only on the left-hand side of the equations to save matrix-vector multiplications. In this 3-stage case this is achieved through redefining the k_i to \tilde{k}_i by

$$k_1 = \tilde{k}_1, \quad k_2 = \tilde{k}_2 - \frac{1}{\gamma}\gamma_{21}k_1, \quad k_3 = \tilde{k}_3 - \frac{1}{\gamma}\gamma_{31}k_1 - \frac{1}{\gamma}\gamma_{32}k_2. \quad (5.11)$$

After this, one can apply the method with the approximate matrix factorization in the same manner as above, replacing the matrix $I - \gamma\tau A$ by its factorized counterpart B defined by (5.4).

The factorization does not affect the order of accuracy (in the ODE sense), since the method is of order two for any choice of A . In comparison with (5.7), a better performance is expected for this method if positivity, TVD or stability properties of the underlying explicit method are of importance. Some numerical results with this Rosenbrock AMF method are given in Section 6.

Stability

Stability can be experimentally studied in the same way as for the previous method (5.7) on the scalar test problem (3.7). Let $P(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{12}z^3$ be the stability polynomial of the explicit method, see (II.4.14). Further let $R = R(z_0, z_1, \dots, z_s)$ be the stability function of method (5.10) with the

transformation (5.11) and with the approximate matrix factorization (5.4). We are interested in the maximal angle α giving $|R| \leq 1$ for arbitrary $z_i \in \mathcal{W}_\alpha$, $i = 1, \dots, s$, with either $z_0 = 0$ or $|P(z_0)| \leq 1$. The experimentally found maximum fractions $a = 2\alpha/\pi$ for $s = 1, 2, 3$ are given in Table 5.1;¹⁶⁾ the accuracy of these entries is within 1%. Note that the fraction 0.50 found for $s = 3$ is optimal in view of Remark 5.1.

	$s = 1$	$s = 2$	$s = 3$
$z_0 = 0$	1	1	0.50
$ P(z_0) \leq 1$	0.73	0.73	0.50

Table 5.1. Fractions $a = 2\alpha/\pi$, $s = 1, 2, 3$, for stability with $z_i \in \mathcal{W}_\alpha$, $i = 1, \dots, s$.

5.4 Concluding Remarks and Tests

Modified Newton Iteration

Instead of the Rosenbrock AMF methods, one can also use a well-known fully implicit method and then try to economize on the Newton process by approximate matrix factorization. In this way AMF is used iteratively within the Newton process. The advantage is that if the iteration converges the theoretical properties of the fully implicit method are valid. One then has also the option to choose high-order methods, see for instance the survey paper by van der Houwen & Sommeijer (2001).

Consider a generic implicit relation

$$w_{n+1} = W_n + \gamma\tau F(w_{n+1}), \quad (5.12)$$

where W_n contains the information up to t_n . This may come from the backward Euler method ($\gamma = 1, W_n = w_n$), trapezoidal rule ($\gamma = \frac{1}{2}, W_n = w_n + \frac{1}{2}\tau F(w_n)$) or the implicit BDF2 method ($\gamma = \frac{2}{3}, W_n = \frac{4}{3}w_n - \frac{1}{3}w_{n-1}$), for example. Then the Newton iteration to solve the implicit relation will be of the form

$$v_{i+1} = v_i - B^{-1}(v_i - \gamma\tau F(v_i) - W_n), \quad i = 0, 1, 2, \dots, \quad (5.13)$$

with initial guess v_0 . Standard modified Newton iteration is obtained with $B = I - \gamma\tau A$, $A \approx F'(v_0)$. For systems of multi-dimensional PDEs this leads to a very big system of linear algebraic equations that could be solved for example by a preconditioned conjugate gradient or multigrid method.

As an alternative one can consider approximate factorization inside the Newton process. Assuming the splitting (5.1), this means that B is redefined

¹⁶⁾ The admissible angles in this table were communicated to us by Alf Gerisch.

by the factorization (5.4). This clearly affects the convergence of the iteration. We will examine this for the scalar test equation (3.7). For this test equation the AMF iteration process has a convergence factor¹⁷⁾

$$S = 1 - \left(\prod_{j=1}^s (1 - \gamma z_j) \right)^{-1} \left(1 - \gamma \sum_{j=0}^s z_j \right) \quad (5.14)$$

and for the iteration to converge we need $|S| < 1$. This convergence factor looks very similar to the stability factor for the Douglas ADI and the Rosenbrock AMF methods. Indeed, the statements given in Section 3.2 for the stability factor $|R| \leq 1$ with the z_j in wedges are also valid for this convergence factor S . For details, see Hundsdorfer (1999). In particular, it should be noted that convergence can be very slow.

To illustrate this, in Figure 5.3 the boundaries of the convergence region are plotted for $z_0 = 0$ and the special choices $z_i = z$ ($1 \leq i \leq s$) and $z_i = z$ ($1 \leq i \leq s-1$), $z_s = \infty$, similar to Figure 3.1. The dotted curved lines are the contour lines for $|S|$ at levels $0.1, 0.2, \dots, 0.9$ for the case that all z_j are equal. If the z_j assume large negative values, then $|S|$ is very close to 1 and thus the convergence will be very slow. Moreover, divergence may occur for $s \geq 3$ if two or more of the z_j are near the imaginary axis.

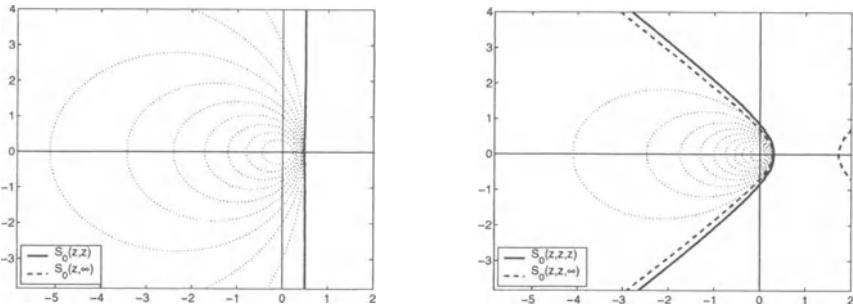


Fig. 5.3. Regions of convergence $|S| < 1$ for $\gamma = 1$, $z_0 = 0$ with equal $z_j = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$; the dotted lines give contour levels $|S| = 0.1, \dots, 0.9$ with equal $z_j = z$.

In conclusion, the convergence of modified Newton AMF iteration can be rather slow, especially for solutions rich in high frequencies. Hence it is not an approach that is recommended for general equations. Of course, there are special problems, especially with smooth solutions, where this approach may work well, see e.g. van der Houwen, Sommeijer & Kok (1997). However, the class of problems for which the iteration does not diverge seems to be close

¹⁷⁾ If v_* is the limit solution of the iteration, then

$v_{i+1} - v_* = v_i - v_* - B^{-1}(v_i - v_* - \gamma \tau(F(v_i) - F(v_*)))$
and for the scalar linear test problem this reduces to $v_{i+1} - v_* = S(v_i - v_*)$.

to the class where the Rosenbrock AMF methods are stable, see Figures 3.1 and 5.3. In those cases the simpler Rosenbrock schemes will be more efficient and with these schemes smoothness of the solution is not required.

Remark 5.3 If $s = 1$, the convergence factor (5.14) equals

$$S = (1 - \gamma z_1)^{-1} \gamma z_0.$$

Hence if we allow z_1 to range over the left half-plane or the negative real axis, the convergence condition is $|\gamma z_0| < 1$. This may be more favourable than the stability condition for the corresponding IMEX schemes, but of course with the IMEX schemes fewer F evaluations are performed per step.

Some results for such iteration were presented in Table 4.2 for a reaction-diffusion model, which will be used again below. For the implicit trapezoidal rule and the BDF2 scheme on the same problem we also tried modified Newton iteration with approximate factorization, involving both the Jacobian matrices for diffusion and reaction, but this generally gave slower convergence of the iteration, in particular for the trapezoidal rule. \diamond

Numerical Illustration

For a numerical comparison of the Rosenbrock methods with approximate Jacobian matrices we consider again the reaction-diffusion model (4.15) with a 200×200 grid, using standard second-order differences for the diffusion terms. In this model the reaction terms are mildly stiff, and thus it is not a priori clear whether these terms should be treated implicitly or explicitly. With the Rosenbrock methods we can consider both choices. In case of explicit treatment of the reactions we then have schemes comparable to the IMEX schemes considered before.

In this test we again used constant step sizes; error control and variable steps will be used in the next section for more complicated examples. As in Section 4.4, the arising linear systems were solved by FISHPACK. The results for the reaction-diffusion model (4.15) are displayed in Figure 5.4. Recall that the spatial L_2 -error is here $2.0 \cdot 10^{-3}$. In the figure the errors are the temporal L_2 -errors at time $t = 1$, found by comparison with a reference solution on the same grid with a very small step size.

In Figure 5.4(a) we have the results for some of the Rosenbrock methods with a Jacobian where only the diffusion terms are taken along. Method (5.2) with $\gamma = \frac{1}{2}$ (indicated here as R1) then becomes the same as the IMEX- θ method (4.2) with $\theta = \frac{1}{2}$ for which the results were given in Figure 4.5. More accurate results are obtained with the second-order scheme (5.7). In the figure the results are given for $b_2 = \frac{1}{2}$ with $\gamma = \frac{1}{2}$ (R2) and with $\gamma = 1 - \frac{1}{2}\sqrt{2}$ (R2*). Both choices lead to second-order convergence but the error constants with $\gamma = 1 - \frac{1}{2}\sqrt{2}$ are smaller. The choice $b_2 = \frac{3}{4}$, $\gamma = \frac{1}{2} - \frac{1}{3}\sqrt{6}$ did give somewhat larger errors than the other two choices.

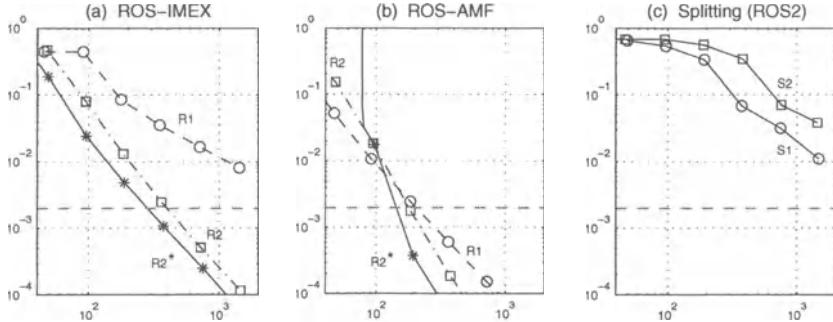


Fig. 5.4. Errors versus CPU (seconds) for the reaction-diffusion model (4.15). Results: (a) ROS-IMEX: method (5.2), $\gamma = \frac{1}{2}$ (indicated as R1), method (5.7) with $b_2 = \frac{1}{2}$, $\gamma = \frac{1}{2}$ (R2) and $\gamma = 1 - \frac{1}{2}\sqrt{2}$ (R2*); (b) ROS2-AMF: same indications as in (a); (c) First-order splitting (S1) and Strang splitting (S2). See text for details.

In Figure 5.4(b) the same Rosenbrock methods are considered but now with approximate factorization, using also a Jacobian factor for the reaction term. Method (5.2) then becomes of order two. Quite surprisingly, the results for method (5.7) with $b_2 = \frac{1}{2}$ and $\gamma = \frac{1}{2}$ or $\gamma = 1 - \frac{1}{2}\sqrt{2}$ are closer to order three. The errors here however have only been displayed up to the level 10^{-4} (in view of the spatial errors). For higher accuracies the temporal errors for (5.7) did show an order two behaviour. This indicates that for this specific problem the error constant in front of the τ^2 terms in the global errors is very small. Results were also obtained for (5.7) with $b_2 = \frac{3}{4}$, $\gamma = \frac{1}{2} - \frac{1}{3}\sqrt{6}$, which is of order three, but in the displayed range the errors for that scheme were somewhat larger than with the other two choices.

For the present problem it is mainly accuracy which determines the success or failure of the schemes. Nevertheless, it is interesting to note that the AMF scheme (5.7), $b_2 = \frac{1}{2}$, did become unstable for increasing step sizes much sooner for $\gamma = 1 - \frac{1}{2}\sqrt{2}$ than for $\gamma = \frac{1}{2}$. This is in agreement with the results in Figure 5.1 for the scalar test equations.

Finally, in Figure 5.4(c) the errors are given for operator splitting. Here we used method (5.7) with $b_2 = \frac{1}{2}$, $\gamma = 1 - \frac{1}{2}\sqrt{2}$ for the fractional steps. Within the reaction sub-step, the Jacobian matrix was set equal to 0, resulting in the explicit trapezoidal rule. As in Figure 4.5 these results with splitting are disappointing, due to the fact that for this reaction-diffusion problem, the patterns are formed by a subtle interplay between reaction and diffusion, which is not sufficiently represented with operator splitting.

6 Numerical Examples

In Section I.1.4 a class of chemo-taxis problems from mathematical biology was presented. The simultaneous occurrence of advection by chemo-taxis, diffusion and reaction calls for a tuned time stepping approach. By way of illustration, we will solve in this section two 2D problems from this class by means of four related schemes, two of which are based on operator splitting and two on the Rosenbrock approximate matrix factorization approach. The emphasis lies in this section on time integration. As a reference method the general purpose BDF-Krylov solver VODPK from Byrne (1992) will be applied. The results presented here have been taken from Gerisch & Verwer (2002). Interested readers are referred to this paper and references therein for the full information. For the sake of readability some of the technical details are omitted here.

6.1 Two Chemo-taxis Problems from Biology

The first problem is the 2D counterpart of the tumour angiogenesis model (I.1.32) which was also discussed briefly in Section III.1.2 with regard to spatial discretization. The 2D set-up resembles that of (I.1.32); in fact the 1D model was obtained from the current 2D version. The second example is a model for a tumour invasion.

The Tumour Angiogenesis Model

The model equations for tumour angiogenesis are

$$\begin{aligned}\frac{\partial \rho}{\partial t} &= \epsilon \Delta \rho - \nabla \cdot (\kappa \rho \nabla c) + \mu \rho (1 - \rho) \max(0, c - c^*) - \beta \rho, \\ \frac{\partial c}{\partial t} &= \Delta c - \lambda c - \frac{\alpha \rho c}{\gamma + c},\end{aligned}$$

where ρ is a cell density and c a chemical concentration (TAF) secreted by the tumour, which induces growth of cells in the direction of the source through the ∇c term, see (I.1.32) for a more detailed description. The spatial (x, y) -domain is taken as the unit square and the time interval is $[0, T]$. Two cases are considered, the case with diffusion for the cell density ρ with $\epsilon = 10^{-3}$ and the case without diffusion, $\epsilon = 0$. The other parameter values are

$$\alpha = 10, \quad \beta = 4, \quad \gamma = 1, \quad \kappa = 0.7, \quad \lambda = 1, \quad \mu = 100, \quad c^* = 0.2.$$

A tumour is located on the left boundary of the spatial domain ($x = 0$), which acts as the source of the TAF concentration c . The initial condition for c is given in Figure 6.1 (left). This figure also depicts the initial cell concentration ρ corresponding to a parent blood vessel on the right boundary

of the domain ($x = 1$) with some already developed capillary sprouts. The boundary conditions for $c(x, y, t)$ are

$$c(0, y, t) = c(0, y, 0), \quad c(1, y, t) = c(1, y, 0), \quad c_y(x, 0, t) = c_y(x, 1, t) = 0.$$

An inflow boundary condition for ρ on the right boundary is used,

$$\rho(1, y, t) = \rho(1, y, 0).$$

If $\epsilon = 0$ no other boundary conditions for ρ are imposed whereas for $\epsilon = 10^{-3}$ Neumann boundary conditions are added for ρ on the remaining part of the boundary. The final simulation time for the described setups are $T = 1.3$ for $\epsilon = 0$ and $T = 1.1$ for $\epsilon = 10^{-3}$. Thereafter the assumptions underlying the model do not hold anymore because the blood vessels have reached the tumour and other processes take over.

The Tumour Invasion Model

This model has three components: ρ is a tumour cell density and c_1 and c_2 are the density of an extracellular matrix (ECM) and the concentration of the matrix degradative enzymes (MDE), respectively. The system of equations reads

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= \epsilon \Delta \rho - \nabla \cdot (\gamma \rho \nabla c_1), \\ \frac{\partial c_1}{\partial t} &= -\eta c_2 c_1, \quad \frac{\partial c_2}{\partial t} = d_2 \Delta c_2 + \alpha \rho - \beta c_2. \end{aligned}$$

For the spatial (x, y) -domain we again take the unit square and the time interval is $[0, T]$. The problem is provided with Neumann boundary conditions for ρ and c_2 ,

$$\underline{n} \cdot (\nabla \rho) = 0, \quad \underline{n} \cdot (\nabla c_2) = 0,$$

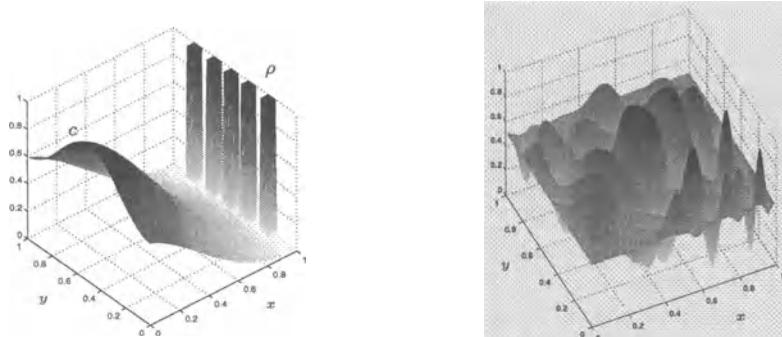


Fig. 6.1. Left picture: Initial conditions for the angiogenesis problem – the smooth function is the initial TAF concentration c and the function consisting of the blocks at the right boundary is the initial cell distribution for ρ . Right picture: Initial condition for the heterogeneous ECM concentration c_1 for the tumour invasion problem.

with \underline{n} the outward unit normal vector. We choose the parameters

$$\epsilon = 0.001, \quad \gamma = 0.005, \quad \eta = 10, \quad d_2 = 0.001, \quad \alpha = 0.1, \quad \beta = 0.$$

In addition we will also consider zero diffusion for ρ (i.e. $\epsilon = 0$). Boundary conditions have no notable influence on the result of the simulations because the cell density near the boundary is virtually zero during simulation time. The initial conditions for tumour cells and MDE (produced by the tumour) are concentrated in the center of the domain and given by

$$\rho(x, y, 0) = c_2(x, y, 0) = e^{-400((x-\frac{1}{2})^2 + (y-\frac{1}{2})^2)}.$$

A hypothetical heterogeneous ECM initial distribution is assumed as depicted in Figure 6.1 (right). The final simulation time for the described setup is $T = 15$. Questions concerning this model are, for instance, how far tumour cells can invade into the surrounding tissue and under which circumstances the initial cell mass does break up into pieces.

6.2 The Numerical Methods

The numerical solutions are obtained on equidistant grids. The diffusion terms are approximated by second-order central differences. For the taxis terms central differencing is not used as this would lead to wiggles and negative concentrations in the solution. Similar as with the more common advection dominated advection-diffusion problems, these arise in the vicinity of steep gradients if the transport of ρ induced by the chemo-taxis term is much stronger than the diffusion transport. Therefore the taxis terms are approximated by the limited third-order upwind-biased scheme in the same way as in Section III.1.2 for the 1D angiogenesis problem, except that here van Leer's limiter (III.2.19) is used. The discretization at the boundaries is similar to the description in Section III.1.1, using virtual values with quadratic extrapolation. For more details we refer to Gerisch & Verwer (2002).

The result of the spatial discretization is the autonomous semi-discrete system $w'(t) = F(w(t))$ assembling at all grid cells the approximations to the population density and the chemicals. The function F is split as

$$F(v) = F_A(v) + F_{D_x}(v) + F_{D_y}(v) + F_R(v),$$

where F_A contains the chemo-taxis terms, which are of the form $\text{div}(\underline{a}\rho)$ with $\underline{a} = \text{grad } c$. Further F_R contains the reaction terms, and F_{D_x} and F_{D_y} stand for the x - and y -diffusion discretizations. In the time integration the taxis terms will be treated explicitly and the remaining (at most one-space dimensional) terms linearly implicitly by approximate matrix factorization and operator splitting. Four second-order integration procedures are applied.

The first procedure is here abbreviated as AMF-ETR. It is based on the Rosenbrock method (5.7) with $b_2 = \frac{1}{2}$ and $\gamma = 1 - \frac{1}{2}\sqrt{2}$ and provided with the approximate matrix factorization (AMF)

$$I - \gamma\tau A = (I - \gamma\tau F'_R(w_n)) (I - \gamma\tau F'_{D_y}(w_n)) (I - \gamma\tau F'_{D_x}(w_n)). \quad (6.1)$$

The abbreviation ETR refers to the explicit trapezoidal rule, also known as the modified Euler method, which is the explicit method that results from the Rosenbrock method if we set $A = 0$. Closely related is AMF-RK32. This procedure is based on the Rosenbrock method (5.10), (5.11), for which the corresponding explicit method (RK32) has a larger stability and positivity domain than the explicit trapezoidal rule.

The third procedure is called OPS-ETR and is based on Strang-type operator splitting (OPS)

$$w_{n+1} = S_{0, \frac{1}{2}\tau}(w^{**}), \quad w^{**} = S_{1,\tau}(w^*), \quad w^* = S_{0, \frac{1}{2}\tau}(w_n), \quad (6.2)$$

where $S_{0,\tau'}$ and $S_{1,\tau'}$ are approximate evolution operators for the functions $F_0 = F_A$ and $F_1 = F_{D_x} + F_{D_y} + F_R$, respectively, over a step size τ' . We will use the explicit trapezoidal rule applied to F_A to approximate $S_{0,\tau/2}$. To approximate the fractional step $S_{1,\tau}$ we use again the implicit Rosenbrock method (5.7) with $b_2 = \frac{1}{2}$ and $\gamma = 1 - \frac{1}{2}\sqrt{2}$ in a similar way as in AMF-ETR, but now without the explicit term F_A . Operator splitting is applied in the order given in (6.2) because then we use only half the step size of the splitting step for the explicit method. This doubles the stability and positivity domain of the explicit method and hence is expected to lead overall to fewer time steps. The fourth procedure, OPS-RK32, differs only from OPS-ETR in the choice of the explicit method, which is now taken as RK32 instead of the explicit trapezoidal rule.

6.3 Numerical Experiments

Following standard practice the four methods have been applied with variable step sizes. The embedded first-order approximations $w_n + \tau k_1$ in AMF-ETR and $w_n + \frac{1}{2}(k_1 + k_2)$ in AMF-RK32 were used to obtain an estimate of the local error in the two AMF methods. The time step was selected on the basis of an error per step (EPS) control which aims to keep this estimate below a mixed (relative and absolute) threshold depending on the tolerance Tol ($= Tol_R = Tol_A$). The second-order solution is used to advance an accepted step (local extrapolation, Shampine (1994, p. 342)). The two OPS methods use Richardson extrapolation to obtain a local error estimate and then the same EPS control to select the step size. They step forward with the solution obtained after two half-steps (doubling, no local extrapolation to third order, in the terminology of Shampine (1994, p. 342)). Jacobian matrices were evaluated at the beginning of a time step (AMF) or at the

beginning of a Richardson step (OPS). Finite difference approximations to the true Jacobian matrices of the functions were used.

On the selected space grid the computed solutions w_N at the final time $t_N = T$ were compared against accurate reference solutions $w(t_N)$ of the semi-discrete system. Hence spatial errors play no role in the error measurements but splitting errors do. The errors and local error estimates are all measured in the discrete L_2 -norm. This norm is also used in the accuracy–CPU time plots below. Note that due to the embedding the local error estimate for the two AMF schemes is $\mathcal{O}(\tau^2)$, whereas with the Richardson extrapolation for the two OPS schemes it is $\mathcal{O}(\tau^3)$. This will cause the AMF schemes to take more steps than the OPS schemes for a given tolerance Tol and this should be accounted for when comparing the two approaches (the accuracy–CPU time plots below reveal the larger number of steps through larger CPU time and more accuracy). In all test cases the four integration procedures were used for tolerance values $Tol = 10^{-3}, 10^{-3.5}, \dots, 10^{-6}$.

For comparison also the code VODPK was applied using the same range of tolerance values. VODPK is a variable-coefficient solver with high-order implicit BDF methods and the Krylov method GMRES for the solution of linear systems, see Byrne (1992). It is based on the VODE and LSODPK packages. This code was used with default parameters, without preconditioning, and with method parameter MF= 21, giving BDF formulas up to order five.

Results for the Tumour Angiogenesis Model

The concentration ρ for the endothelial cell density has initially peaks near the right boundary of the domain. The cells there are migrating to the left, forming a stream which moves up the present gradient of the TAF concentration c as time proceeds. No cell proliferation takes place in the beginning of the simulation because the c concentration at the cells is below the threshold c^* . Later proliferation leads to a strong, local increase of the cell density. The cells also take up TAF. This results in changes in the TAF gradients and causes lateral cell movement and hence a widening of the cell streams. The cell streams turn towards the center of the TAF source (the tumour) once they are close enough to the left boundary. Figure 6.2 gives cell density plots. Note that the process proceeds faster if cell diffusion is present and that in this case also the lateral cell movement is more pronounced, leading to a closed wave front towards final time. Apart from this the solutions look very similar.

Figures 6.3 and 6.4 give accuracy–CPU time plots for the problems with and without cell diffusion, respectively, and on two grids, with mesh widths $h = 1/100, 1/200$. On the vertical axes the \log_{10} of the L_2 -errors are displayed. In all four test situations and up to moderate accuracy, the approximate matrix factorization AMF and operator splitting OPS schemes are clearly much more efficient than the standard code VODPK. This is especially true

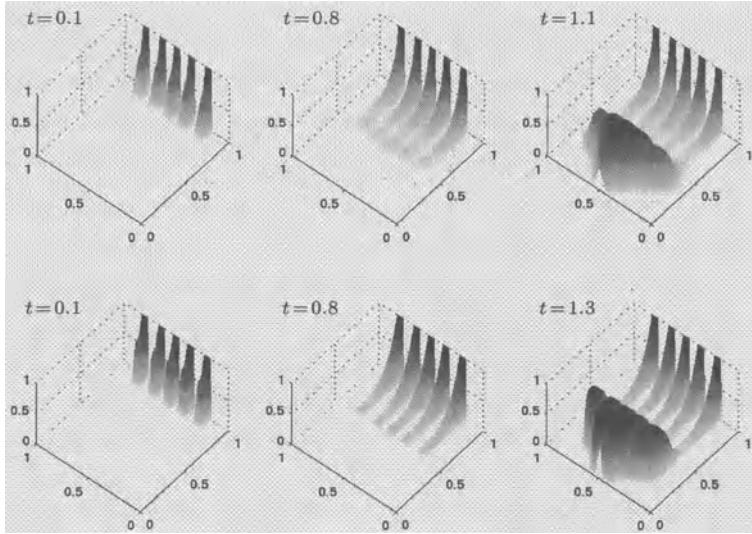


Fig. 6.2. Simulation results for the cell density ρ of the tumour angiogenesis model with cell diffusion (top row) and without diffusion (bottom row). The output times are given in the title of each plot.

for the finer grid resolution. VODPK is more efficient for higher accuracy requirements because of its higher order. However, we note that the point of intersection between the VODPK curve and the OPS-RK32 curve is at a higher achieved accuracy on the finer grid. Thus, if the spatial accuracy is increased then the splitting schemes are also more efficient for higher temporal accuracy. Further we observe that for this problem the OPS procedures are more efficient than the AMF procedures and this observation is independent of the choice of the cell diffusion coefficient ϵ .

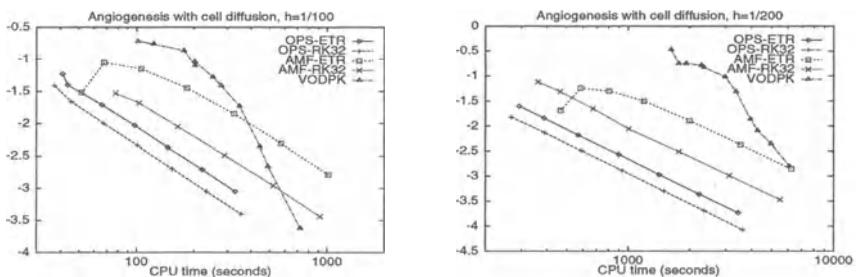


Fig. 6.3. Results ($\log_{10}(\text{err})$ v.s. CPU) for the angiogenesis model with cell diffusion on spatial grids with $h = 1/100$ (left) and $h = 1/200$ (right). AMF-RK32 failed for $Tol = 10^{-3}$ in the left plot; additional points for Tol up to 10^{-8} were included for VODPK.

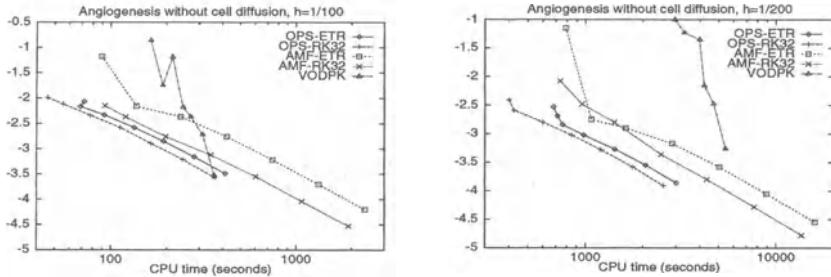


Fig. 6.4. Results ($\log_{10}(\text{err})$ v.s. CPU) for the angiogenesis model without cell diffusion on spatial grids with $h = 1/100$ (left) and $h = 1/200$ (right). VODPK failed for $Tol = 10^{-3}$ in the case of $h = 1/200$.

Results for the Tumour Invasion Model

The solution ρ of the cell density equation of this problem has an initial peak in the center of the domain. This peak spreads outward moving up gradients of the ECM density c_1 which is heterogeneous initially. This leads to a heterogeneous pattern in the cell density solution. These patterns are sharper if there is no cell diffusion (a break up of the initially compact cell mass can be observed) and more smeared with cell diffusion (the break up of cell mass is not so pronounced in this case). The total cell mass in the domain is a conserved quantity of the model. The tumour cells release MDE, with concentration c_2 , which (slowly) diffuses within the spatial domain. This

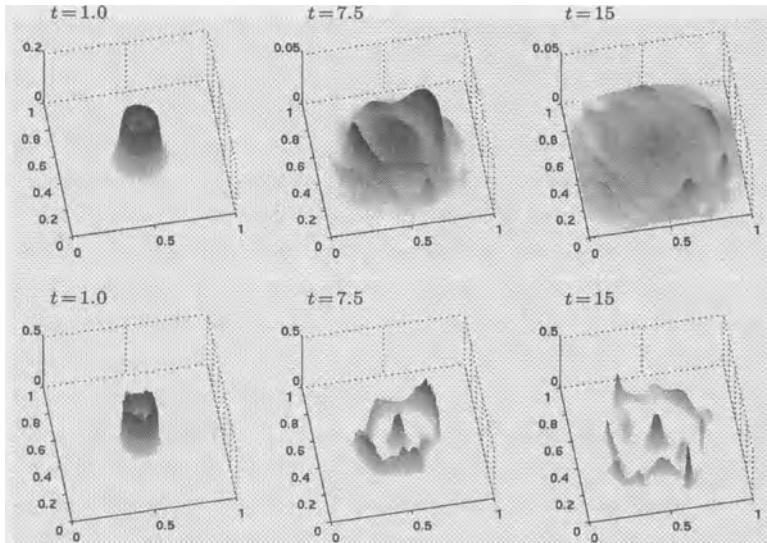


Fig. 6.5. Simulation results for the cell density ρ of the tumour invasion model with cell diffusion (top row) and without diffusion (bottom row). The output times are given in the title of each plot.

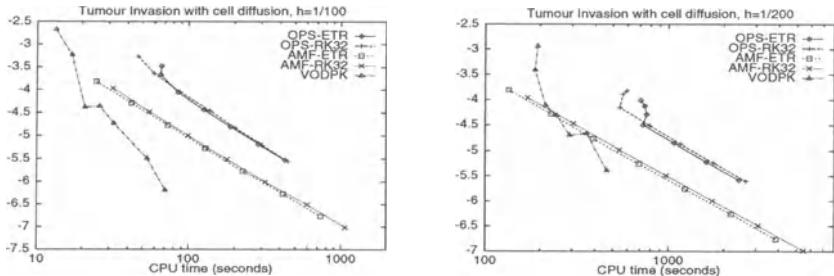


Fig. 6.6. Results ($\log_{10}(\text{err})$) v.s. CPU) for the tumour invasion model with cell diffusion on spatial grids with $h = 1/100$ (left) and $h = 1/200$ (right).

MDE in turn degrades ECM and hence leads to new gradients in the ECM density which give rise to further migration of the cells. The most interesting component of this model is the cell density. Solution plots are given in Figure 6.5.

Figures 6.6 and 6.7 give accuracy–CPU time plots. In contrast to the angiogenesis problem, the BDF code VODPK turns out to be very efficient for this invasion model. This advantage decreases for the finer grid resolution and more significantly if the small diffusion coefficients ϵ or d_2 are enlarged; hence VODPK, as used here, with default parameters and no preconditioning, seems somewhat sensitive to increasing stiffness in this problem. We clearly see that AMF is more suitable than OPS for the test case with cell diffusion; without cell diffusion the situation is the opposite and OPS generally demonstrates a better performance. Needless to say that the problem without cell diffusion is the more difficult one. It can also be observed that for this problem the differences between the RK32 and ETR based schemes are rather minor. In all cases the initial cell mass was preserved up to a difference of the size of round-off.

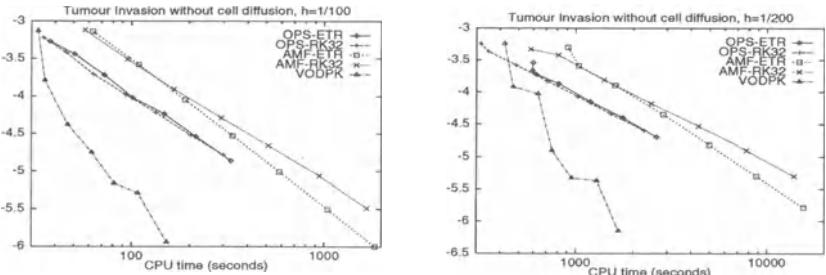


Fig. 6.7. Results ($\log_{10}(\text{err})$) v.s. CPU) for the tumour invasion model without cell diffusion on spatial grids with $h = 1/100$ (left) and $h = 1/200$ (right).

Remarks on the Test Results

The above test results illustrate some time integrations for two taxis-diffusion-reaction models from mathematical biology by splitting methods based on two approaches. These differ mainly in the way how the taxis discretization is treated in the time stepping process: OPS separates the taxis and the diffusion-reaction parts completely, whereas with AMF the different parts are treated more or less simultaneously, but also in an appropriate explicit or implicit way.

For readability and to save space we have not discussed the test results in great detail, see Gerisch & Verwer (2002) for a full account. The results show the potential of these splitting schemes, but they also show that an overall ‘best’ method does not exist. The relative performances are problem dependent. However, the tests do point to the RK32 based methods as the better ones compared to the ETR based methods. We owe this to the larger stability and positivity domain of RK32. Further, with cell diffusion the AMF approach seems better than OPS, whereas without cell diffusion it is just the other way around. Within OPS the taxis computation stands on its own and this seems to be an advantage for the steeper gradient computations encountered without cell diffusion.

For really time-consuming applications (3D) it will be worthwhile to use a more sophisticated fine-tuning of the AMF and OPS codes. For example, the local error and step size control that has been used here is simple and standard, and at every time step the required Jacobian matrices have been simply recalculated and decomposed. In the VODE code Jacobian updates are only performed when the convergence in the Newton process for the implicit relations is deemed to be insufficient. On the other hand it should also be mentioned that the VODPK code was used with default parameters and without preconditioning. Hence with appropriate fine-tuning its performance might also be improved for these particular chemo-taxis problems.

V Stabilized Explicit Runge-Kutta Methods

In this chapter we discuss special purpose explicit Runge-Kutta methods for systems of ODEs in \mathbb{R}^m

$$w'(t) = F(t, w(t)), \quad t > 0, \quad w(0) = w_0,$$

representing semi-discrete, multi-space dimensional parabolic problems. Often, parabolic problems give rise to stiff systems having a symmetric Jacobian matrix $\partial F(t, w)/\partial w$ with a spectral radius proportional to h^{-2} , h representing a spatial mesh width. Standard explicit methods are then highly inefficient due to their severe stability constraint, see Section II.1.4. On the other hand, unconditionally stable implicit methods, like backward Euler or the implicit trapezoidal rule, do require one or more linear or nonlinear algebraic system solutions at each integration step, which can become costly in higher space dimension.

The stabilized Runge-Kutta methods discussed here are in between: they are explicit, and thus avoid algebraic system solutions, and possess *extended real stability intervals* with a length proportional to s^2 , with s the number of stages. This quadratic dependence is derived from properties of first kind Chebyshev polynomials. Hence these stabilized methods are also named *Runge-Kutta-Chebyshev* methods. The quadratic dependence is very attractive, since it means that the scaled stability interval length, taking into account the work load per step (the number of stages), increases linearly with s . These stabilized methods are therefore very useful for modestly stiff, semi-discrete parabolic problems for which the implicit system solution is really costly in terms of CPU time or difficult due to memory constraints. In case of severe stiffness these stabilized methods can of course become inefficient since then a very large number of stages will be needed to achieve stability with reasonable step sizes. In such situations the use of an implicit, unconditionally stable method is advocated.

We will focus on two families of methods: the family of second-order Runge-Kutta-Chebyshev (RKC) methods proposed by van der Houwen & Sommeijer (1980) and the more recent family of second- and fourth-order Orthogonal-Runge-Kutta-Chebyshev (ROCK) methods proposed by Abdulle (2001, 2002) and Abdulle & Medovikov (2001). A related third family of methods is the DUMKA family developed by Lebedev and Medovikov, see

Lebedev (1994, 2000), Medovikov (1998) and references therein. We here refrain from discussing DUMKA as this offers no new insight after having discussed RKC and ROCK and the latter is believed to improve DUMKA, see Abdulle (2001). Other references of interest are van der Houwen (1996), Verwer (1996) and Hairer & Wanner (1996, Sect. IV.2).

1 The RKC Family

1.1 Stability Polynomials

For any explicit Runge-Kutta method applied to the test model $w'(t) = \lambda w(t)$ with $\lambda < 0$, stability from step to step is determined by its *real stability boundary* β_R . By definition, $[-\beta_R, 0]$ is the largest segment of the negative real axis contained in the stability region $\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$ as determined by the polynomial stability function

$$R(z) = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \cdots + \gamma_s z^s, \quad \gamma_0 = \gamma_1 = 1. \quad (1.1)$$

Here $\gamma_0 = \gamma_1 = 1$ for first-order consistency. In this section we discuss Chebyshev-type polynomials which yield boundaries proportional to s^2 , and we begin with a theorem which proves that $\beta_R = 2s^2$ is optimal.

First-Order Stability Polynomials

Theorem 1.1 *For any explicit, consistent Runge-Kutta method we have $\beta_R \leq 2s^2$. The optimal stability polynomial is the shifted Chebyshev polynomial of the first kind*

$$P_s(z) = T_s\left(1 + \frac{z}{s^2}\right), \quad \beta_R = 2s^2. \quad (1.2)$$

Proof.¹⁾ The Chebyshev polynomials T_s can be defined by the relation $T_s(x) = \cos(s \arccos(x))$, $x \in [-1, 1]$, or recursively for $z \in \mathbb{C}$ by

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_j(z) = 2zT_{j-1}(z) - T_{j-2}(z), \quad 2 \leq j \leq s. \quad (1.3)$$

Hence $P_s(z) = 1 + z + \mathcal{O}(z^2)$ and therefore these shifted polynomials belong to the class of stability polynomials (1.1) that can be generated by explicit consistent Runge-Kutta methods. By the definition of $T_s(x)$ it follows that $|P_s(x)| \leq 1$ for $-2s^2 \leq x \leq 0$. On this interval P_s alternates between -1 and $+1$ and has $s - 1$ points of tangency with these horizontal lines (illustrated in Figure 1.1 for $2 \leq s \leq 5$).

¹⁾ The proof goes back to Markoff (1892). Quoting van der Houwen (1996), the use of the shifted Chebyshev polynomials P_s for solving parabolic equations was first mentioned by Yuan' Chzao-Din (1958), Franklin (1959) and Guillou & Lago (1961). These authors were not aware of each others work.

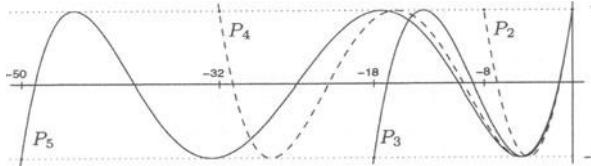


Fig. 1.1. The shifted Chebyshev polynomials $P_s(x)$ for $x \leq 0$, $2 \leq s \leq 5$.

This property determines it as the unique polynomial with the largest real stability boundary. Suppose a second polynomial (1.1) exists with $\beta_R \geq 2s^2$. Since $P_s(x)$ has $s - 1$ points of tangency with the lines $y = \pm 1$, this second polynomial then must intersect $P_s(x)$ at least $s - 1$ times for $x < 0$, where intersecting points with common tangent are counted double. Hence the difference polynomial has at least $s - 1$ negative roots, again with roots of multiplicity 2 counted double. However, the difference is of the form $x^2(\tilde{\gamma}_2 + \cdots + \tilde{\gamma}_s x^{s-2})$ and thus can have at most $s - 2$ negative roots, which is a contradiction. \square

According to Abramowitz & Stegun (1968, Eqs. 15.4.1 and 15.4.3), $T_s(x)$ can also be written as

$$T_s(x) = \sum_{i=0}^s \frac{(-s)_i (s)_i}{(\frac{1}{2})_i i!} \left(\frac{1-x}{2} \right)^i,$$

where $(a)_i$ is defined as $(a)_0 = 1$ and $(a)_i = a(a+1)\cdots(a+i-1)$ for $a \in \mathbb{R}$, $i \geq 1$. A simple calculation then yields for the coefficients γ_i of (1.1) the expression

$$\gamma_0 = \gamma_1 = 1, \quad \gamma_i = \frac{1 - (i-1)^2/s^2}{i(2i-1)} \gamma_{i-1} \quad \text{for } i = 2, \dots, s.$$

By way of illustration we list

$$\begin{aligned} P_2(z) &= 1 + z + \frac{1}{8}z^2, \\ P_3(z) &= 1 + z + \frac{4}{27}z^2 + \frac{4}{729}z^3, \\ P_4(z) &= 1 + z + \frac{5}{32}z^2 + \frac{1}{128}z^3 + \frac{1}{8192}z^4, \\ P_5(z) &= 1 + z + \frac{4}{25}z^2 + \frac{28}{3125}z^3 + \frac{16}{78125}z^4 + \frac{16}{9765625}z^5, \end{aligned}$$

and it is seen that for a given value of s the coefficients γ_i rapidly decrease with increasing i , which is a prerequisite for a large stability interval. The top picture in Figure 1.2 illustrates the stability region \mathcal{S} of P_5 .

For any fixed i , the coefficient γ_i tends with $s \rightarrow \infty$ to the limit value

$$\gamma_i^* = 2^i / (2i)! . \tag{1.4}$$

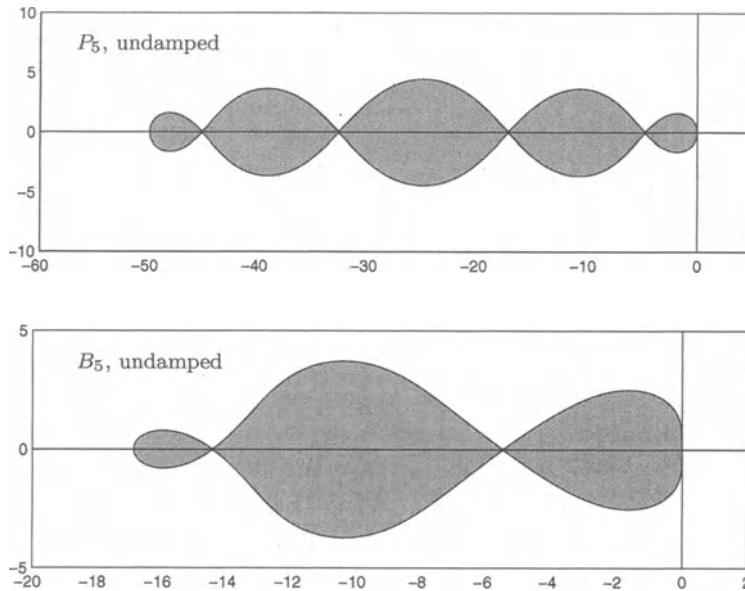


Fig. 1.2. Stability regions for the first- and second-order shifted Chebyshev polynomials P_5 and B_5 , undamped case.

This follows from induction on the limiting recurrence relation

$$\gamma_0^* = 1, \quad \gamma_i^* = \frac{1}{i(2i-1)} \gamma_{i-1}^* \quad \text{for } i \geq 1.$$

Consequently, for $P_s(z)$ we find the pointwise limit

$$\lim_{s \rightarrow \infty} P_s(z) = \cos(\sqrt{-2z}) = \sum_{i=0}^{\infty} \frac{(2z)^i}{(2i)!},$$

which shows that for s large,

$$P_s(z) \approx e^z - \frac{1}{3}z^2 + \mathcal{O}(z^3), \quad z \rightarrow 0,$$

revealing a leading error coefficient of approximately 1/3.

Second-Order Stability Polynomials

For actual computational practice first-order consistency is often too low. So let us look for stability polynomials

$$R(z) = 1 + z + \cdots + \frac{1}{p!}z^p + \gamma_{p+1}z^{p+1} + \cdots + \gamma_sz^s \quad (1.5)$$

of consistency order $p \geq 2$, where the free coefficients γ_i should be chosen to get β_R as large as possible. Riha (1972) proved the existence of such optimal polynomials for any order $p \geq 1$ and degree $s \geq p$. Using a similar reasoning as in the proof of Theorem 1.1, it can be shown that among all polynomials of order p and degree s , the optimal one satisfies the so-called *equal ripple property*: there exist $s-p+1$ points $z_0 < z_1 < \dots < z_{s-p} < 0$, with $z_0 = -\beta_R$, such that

$$\begin{aligned} R(z_i) &= -R(z_{i+1}), \quad i = 0, 1, \dots, s-p-1, \\ |R(z_i)| &= 1, \quad i = 0, 1, \dots, s-p. \end{aligned} \tag{1.6}$$

This equal ripple property has been used by various authors to construct approximations to the optimal polynomials, since no explicitly given analytical expressions for the optimal coefficients are known for $p > 1$.

Numerical approximations for $p = 2, 3, 4$ given in van der Houwen (1977, Tab. 2.6.6) indicate that the optimal bounds β_R always depend quadratically on s and satisfy $\beta_R = c_p s^2$ for $s \rightarrow \infty$ with

$$c_2 \approx 0.82, \quad c_3 \approx 0.49, \quad c_4 \approx 0.34. \tag{1.7}$$

Estimates for c_p up to order $p = 11$ can be found in Abdulle (2001).

For $p = 2$ a suitable approximate polynomial in *analytical* form was given by Bakker (1971),

$$B_s(z) = \frac{2}{3} + \frac{1}{3s^2} + \left(\frac{1}{3} - \frac{1}{3s^2} \right) T_s \left(1 + \frac{3z}{s^2 - 1} \right), \quad \beta_R \approx \frac{2}{3}(s^2 - 1). \tag{1.8}$$

This polynomial generates about 80% of the optimal interval. For an even degree, β_R equals $\frac{2}{3}(s^2 - 1)$ exactly, while for an odd degree β_R is slightly larger. This follows from the observation that $B_s(z)$ alternates between the values $\frac{1}{3} + \frac{2}{3}s^{-2}$ and 1 for $z \in [-\frac{2}{3}(s^2 - 1), 0]$, while for an odd degree the exact boundary is determined by the point z where $B_s(z)$ intersects the line -1 . This point is slightly smaller than $-\frac{2}{3}(s^2 - 1)$. The bottom picture in Figure 1.2 illustrates the stability domain \mathcal{S} of B_5 .

Remark 1.2 Although not optimal with respect to stability, this polynomial B_s is attractive as it has a relatively small error constant, see (1.9) below. In the construction of their RKC methods, van der Houwen and Sommeijer therefore preferred this Bakker polynomial above their own analytically given second-order version

$$R(z) = \frac{2}{2-z} - \frac{z}{2-z} T_s \left(\cos \frac{\pi}{s} + \frac{z}{2} (1 - \cos \frac{\pi}{s}) \right), \quad s \geq 2,$$

with the nearly optimal stability boundary

$$\beta_R = \frac{2}{\tan^2(\pi/2s)} \approx 8 \frac{s^2}{\pi^2} \approx 0.81 s^2$$

for large s , see van der Houwen (1996). Hence we also proceed with the Bakker polynomial (1.8). \diamond

Following the derivation in the first-order case, the coefficients γ_i of B_s are found to satisfy $\gamma_2 = \frac{1}{2}$ and

$$\gamma_i = 3 \frac{1 - (i-1)^2/s^2}{i(2i-1)(1-1/s^2)} \gamma_{i-1} \quad \text{for } i = 3, \dots, s.$$

In particular,

$$\begin{aligned} B_3(z) &= 1 + z + \frac{1}{2}z^2 + \frac{1}{16}z^3, \\ B_4(z) &= 1 + z + \frac{1}{2}z^2 + \frac{2}{25}z^3 + \frac{1}{250}z^4, \\ B_5(z) &= 1 + z + \frac{1}{2}z^2 + \frac{7}{80}z^3 + \frac{1}{160}z^4 + \frac{1}{6400}z^5, \end{aligned}$$

and the limiting polynomial is given by

$$\lim_{s \rightarrow \infty} B_s(z) = \frac{2}{3} + \frac{1}{3} \cos(\sqrt{-6z}) = \frac{2}{3} + \frac{1}{3} \sum_{i=0}^{\infty} \frac{(6z)^i}{(2i)!}.$$

Notice that the coefficients are a factor 3^{i-1} larger than in the first-order case. For large s we have

$$B_s(z) \approx e^z - \frac{1}{15}z^3 + \mathcal{O}(z^4), \quad z \rightarrow 0, \quad (1.9)$$

which shows that the Bakker-Chebyshev polynomials possess a rather small error constant for a second-order approximation to the exponential. The error constant of the optimal second-order polynomial is slightly larger, viz. 0.074 approximately, see Abdulle (2001, Tab. 4.1).

Damped Stability Polynomials

For the polynomials (1.2) and (1.8) the stability interval contains interior points $z \in (-\beta_R, 0)$ where $|R(z)| = 1$. This means that a small imaginary perturbation on z might yield instability. For this reason the polynomials are slightly modified so as to introduce a little damping.

Adopting the choice made by Guillou & Lago (1961), the damped form of (1.2) reads

$$P_s(z) = \frac{T_s(\omega_0 + \omega_1 z)}{T_s(\omega_0)}, \quad \omega_1 = \frac{T_s(\omega_0)}{T'_s(\omega_0)}, \quad (1.10)$$

where $\omega_0 > 1$ is a parameter and ω_1 is chosen such that $P'_s(0) = 1$, implying first-order consistency. The stability interval is determined by the relation $-\omega_0 \leq \omega_0 + \omega_1 z \leq \omega_0$, giving $\beta_R = 2\omega_0/\omega_1$. In the interior of the stability interval, $P_s(z)$ now alternates between $T_s(\omega_0)^{-1}$ and $-T_s(\omega_0)^{-1}$. A convenient choice for ω_0 is $\omega_0 = 1 + \epsilon/s^2$ with ϵ a small positive number. Expanding at

$\omega_0 = 1$ and using $T'_s(1) = s^2$, $T''_s(1) = \frac{1}{3}s^2(s^2 - 1)$ then shows $T_s(\omega_0) \approx 1 + \epsilon$ and

$$\beta_R = \frac{2\omega_0 T'_s(\omega_0)}{T_s(\omega_0)} \approx (2 - \frac{4}{3}\epsilon)s^2.$$

A suitable value for ϵ is 0.05. For practical problems this gives sufficient damping (approximately 5%) and it gives only a minor decrease of the stability boundary to approximately $1.93 s^2$.

In a similar manner the second-order polynomial (1.8) is damped (van der Houwen & Sommeijer, 1980):

$$B_s(z) = 1 + \frac{T''_s(\omega_0)}{(T'_s(\omega_0))^2} \left(T_s(\omega_0 + \omega_1 z) - T_s(\omega_0) \right), \quad \omega_1 = \frac{T'_s(\omega_0)}{T''_s(\omega_0)}. \quad (1.11)$$

Using $T'_s(1) = s^2$, $T''_s(1) = \frac{1}{3}s^2(s^2 - 1)$ and $T'''_s(1) = \frac{1}{15}s^2(s^2 - 1)(s^2 - 4)$, the boundary β_R can be seen to satisfy

$$\beta_R \approx \frac{(\omega_0 + 1)T''_s(\omega_0)}{T'_s(\omega_0)} \approx \frac{2}{3}(s^2 - 1)\left(1 - \frac{2}{15}\epsilon\right). \quad (1.12)$$

Taking $\epsilon = 2/13$ we get approximately 5% damping in the interior of the stability interval and a reduction in the stability boundary of about 2% compared to the undamped case.

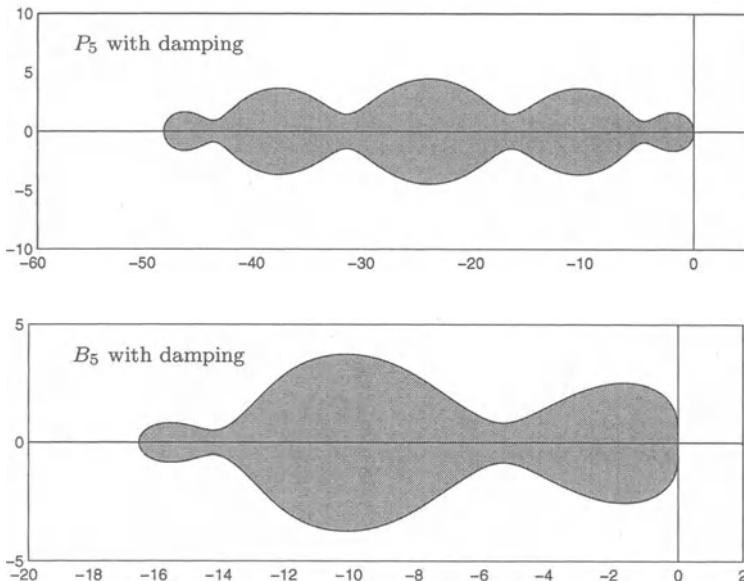


Fig. 1.3. Stability regions for the first- and second-order shifted Chebyshev polynomials P_5 and B_5 with damping.

The effect of damping is that the stability regions \mathcal{S} contain a narrow strip along the negative real line segment, which increases the robustness in actual calculations. Figure 1.3 illustrates this for the damped versions of P_5 and B_5 .

1.2 Integration Formulas

Having selected appropriate stability polynomials, one can construct explicit Runge-Kutta formulas generating these polynomials. However, for stabilized methods it turns out that the choice of the formulas is crucial with regard to another form of stability, viz. *internal stability*. Consider the explicit Runge-Kutta formula written in the form

$$\begin{aligned} w_{n0} &= w_n, \\ w_{nj} &= w_n + \tau \sum_{k=0}^{j-1} \alpha_{jk} F(t_n + c_k \tau, w_{nk}), \quad j = 1, \dots, s, \\ w_{n+1} &= w_{ns}, \end{aligned} \quad (1.13)$$

and its perturbed version

$$\begin{aligned} \tilde{w}_{n0} &= \tilde{w}_n, \\ \tilde{w}_{nj} &= \tilde{w}_n + \tau \sum_{k=0}^{j-1} \alpha_{jk} F(t_n + c_k \tau, \tilde{w}_{nk}) + r_j, \quad j = 1, \dots, s, \\ \tilde{w}_{n+1} &= \tilde{w}_{ns}, \end{aligned} \quad (1.14)$$

with local perturbations r_j , representing for example round-off errors.

We study the propagation of the perturbations for linear systems

$$w'(t) = Aw + g(t), \quad 0 < t \leq T, \quad w(0) = w_0. \quad (1.15)$$

Denoting $e_n = \tilde{w}_n - w_n$ and $e_{nj} = \tilde{w}_{nj} - w_{nj}$, we have

$$e_{nj} = R_j(\tau A) e_n + \sum_{k=1}^j Q_{jk}(\tau A) r_k, \quad 1 \leq j \leq s, \quad (1.16)$$

where $R_j(\tau A)$ and $Q_{jk}(\tau A)$ are matrix polynomials of degree j and $j - k$, respectively. In particular, $e_{n+1} = e_{ns}$ and

$$e_{n+1} = R(\tau A) e_n + \sum_{j=1}^s Q_{sj}(\tau A) r_j, \quad (1.17)$$

see also formula (II.1.21).

Remark 1.3 With regard to consistency, recall that $R(z)$ approximates e^z for $z \rightarrow 0$. In particular, $R(z) = e^z + \mathcal{O}(z^{p+1})$ for $p \leq 2$ implies consistency of order p also for the general nonlinear problem $w'(t) = F(t, w(t))$, because the consistency conditions of Runge-Kutta methods for order $p \leq 2$ are the same for linear and nonlinear problems. The intermediate polynomials $R_j(z)$ ($1 \leq j \leq s-1$) play a similar role with respect to $e^{c_j z}$ and w_{nj} . \diamond

The error scheme (1.17) gives a complete description of the linear stability. The polynomials Q_{sj} are called *internal stability polynomials* as they determine the accumulation of the stage perturbations r_j within a single integration step. If the matrix A is normal, then

$$\|e_{n+1}\|_2 \leq \max_{z=\tau\lambda} |R(z)| \|e_n\|_2 + \sum_{j=1}^s \max_{z=\tau\lambda} |Q_{sj}(z)| \|r_j\|_2, \quad (1.18)$$

where λ runs through the spectrum of A . Hence if A is normal and the common stability condition $|R(z)| \leq 1$ is satisfied, then we have the usual stability from step to step for the propagation of e_n . This, however, does not guarantee internal stability as Example 1.4 will illustrate.

Example 1.4 Consider the following special method written in the form (1.13),

$$\begin{aligned} w_{n0} &= w_n, \\ w_{nj} &= w_n + \tau \alpha_{j,j-1} F(t_n + c_{j-1}\tau, w_{n,j-1}), \quad 1 \leq j \leq s, \\ w_{n+1} &= w_{ns}. \end{aligned}$$

By defining $\alpha_{j,j-1} = \gamma_{s+1-j}/\gamma_{s-j}$ it generates the stability polynomial (1.1). This method is simple and requires a minimum amount of storage and has therefore been considered quite extensively in the past, see van der Houwen (1977). If the stability polynomial is of order $p \leq 2$, this method is also of order p for general nonlinear problems (see Remark 1.3 below). However, the method is of limited practical use as it is severely internally unstable.

To see this, let us consider the internal stability polynomials Q_{sj} introduced in (1.17). We find

$$Q_{sj}(z) = \gamma_{s-j} z^{s-j}, \quad j = 1, \dots, s.$$

The rapid growth of z^{s-j} for z in the real stability interval $[-\beta_R, 0]$ renders the method useless for large values of s . If we substitute the limiting values $\gamma_{s-j}^* = 2^{s-j}/(2s-2j)!$ from (1.4) belonging to the first-order stability polynomial (1.2) and put $z = -\beta_R = -2s^2$, we get

$$|Q_{sj}(-\beta_R)| \approx \frac{(2s)^{2s-2j}}{(2s-2j)!}.$$

Hence, for j small, that is for the early stages, we get growth factors approximately equal to $(2s)^{2s}/(2s)!$. If the r_j stand for rounding errors, these factors are to be multiplied by the machine precision of the computer used. However, they increase so rapidly with s that in actual application only a limited number of stages can be used. For example, already for $s = 12$ we have $\approx 10^9$ so that with a machine precision of 16 digits at most 7 digits remain. The internal accumulation of rounding errors observed in actual computations is in line with this estimate.

Although we have not given a precise mathematical definition of internal stability, from a practical point of view it is clear that we may call the current method internally unstable. The error growth is caused by the strong growth of internal stability polynomials $Q_{sj}(z)$ when z approaches the stability boundary $-\beta_R$. \diamond

To avoid internal instability, van der Houwen & Sommeijer (1980) developed a family of methods from the three-term Chebyshev recursion (1.3). Their Runge-Kutta-Chebyshev (RKC) family contains first- and second-order formulas which can be used for any (practical) value of s . Their main idea was to determine the formulas in such a way that all polynomials R_j belonging to the intermediate stages of the error scheme (1.16) are defined by this three-term Chebyshev recursion and all share the stability interval $[-\beta_R, 0]$ of the target polynomial $R = R_s$. We now first construct the RKC formulas. The resulting internal stability polynomials Q_{sj} will be discussed in Section 1.3.

Thus the ansatz is made that all R_j ($1 \leq j \leq s$) are of the form

$$R_j(z) = a_j + b_j T_j(\omega_0 + \omega_1 z), \quad a_j = 1 - b_j T_j(\omega_0),$$

with $R_s(z) = R(z)$ being the earlier derived damped stability polynomial (1.10) or (1.11). This means that ω_0, ω_1 and b_s are defined and that b_j is yet undetermined for $1 \leq j < s$. Define $R_0(z) = a_0 + b_0 \equiv 1$. Imposing the recursion (1.3) and using $R_j(0) = 1$ then shows that the R_j satisfy

$$R_0(z) = 1, \quad R_1(z) = 1 + \tilde{\mu}_1 z,$$

$$R_j(z) = (1 - \mu_j - \nu_j) + \mu_j R_{j-1}(z) + \nu_j R_{j-2}(z) + \tilde{\mu}_j R_{j-1}(z)z + \tilde{\gamma}_j z,$$

where $j = 2, \dots, s$ and

$$\begin{aligned} \tilde{\mu}_1 &= b_1 \omega_1, \quad \mu_j = \frac{2b_j \omega_0}{b_{j-1}}, \quad \nu_j = \frac{-b_j}{b_{j-2}}, \\ \tilde{\mu}_j &= \frac{2b_j \omega_1}{b_{j-1}}, \quad \tilde{\gamma}_j = -a_{j-1} \tilde{\mu}_j. \end{aligned} \tag{1.19}$$

From these relations we can deduce the RKC integration formulas for the nonlinear problem $w'(t) = F(t, w(t))$ by associating R_j with the intermediate approximation w_{nj} and the occurrence of z with a function evaluation:

$$\begin{aligned}
w_{n0} &= w_n, \\
w_{n1} &= w_n + \tilde{\mu}_1 \tau F_{n0}, \\
w_{nj} &= (1 - \mu_j - \nu_j) w_n + \mu_j w_{n,j-1} + \nu_j w_{n,j-2} + \tilde{\mu}_j \tau F_{n,j-1} + \tilde{\gamma}_j \tau F_{n0}, \\
w_{n+1} &= w_{ns}.
\end{aligned} \tag{1.20}$$

Here $j = 2, \dots, s$ and F_{nk} denotes $F(t_n + c_k \tau, w_{nk})$. This formula obviously belongs to class (1.13). What remains is to define b_j ($1 \leq j < s$) dependent on what target stability polynomial R we have in mind.

First-Order Formulas

Choose for $R(z)$ the first-order damped polynomial (1.10). With the damping parameters

$$\omega_0 = 1 + \epsilon / s^2, \quad \omega_1 = T_s(\omega_0) / T'_s(\omega_0),$$

we select b_j such that

$$R_j(z) = T_j(\omega_0 + \omega_1 z) / T_j(\omega_0), \quad j = 1, \dots, s.$$

All R_j then share the real stability interval $[-\beta_R, 0]$ of the target stability polynomial $R = R_s$. A little inspection then reveals that the parameters b_j should be taken as

$$b_j = 1 / T_j(\omega_0), \quad j = 0, \dots, s, \tag{1.21}$$

defining the first-order method (1.20). Observe that $R_j(z) = e^{c_j z} + \mathcal{O}(z^2)$ with

$$c_j = \frac{T_s(\omega_0)}{T'_s(\omega_0)} \frac{T'_j(\omega_0)}{T_j(\omega_0)} \approx \frac{j^2}{s^2} \quad (1 \leq j \leq s-1), \quad c_s = 1.$$

Second-Order Formulas

Choose for $R(z)$ the second-order damped polynomial (1.11). The free parameters b_j can now be chosen in slightly different ways. Here we adopt the choice of Sommeijer & Verwer (1980) by which all intermediate approximations w_{nj} ($2 \leq j \leq s$) are second-order consistent.²⁾ This means that the expansion

$$R_j(z) = 1 + b_j \omega_1 T'_j(\omega_0) z + \frac{1}{2} b_j \omega_1^2 T''_j(\omega_0) z^2 + \mathcal{O}(z^3)$$

must be matched with $R_j(z) = 1 + c_j z + \frac{1}{2} (c_j z)^2 + \mathcal{O}(z^3)$. An elementary calculation yields

$$b_j = T''_j(\omega_0) / (T'_j(\omega_0))^2, \quad j = 2, \dots, s. \tag{1.22}$$

²⁾ In the original method of van der Houwen & Sommeijer (1980) all intermediate approximations are of order one. This seems less attractive, but with respect to accuracy of the step approximations w_n no essential difference exists for practical values of s .

With the damping parameters chosen as

$$\omega_0 = 1 + \epsilon / s^2, \quad \omega_1 = T'_s(\omega_0) / T''_s(\omega_0),$$

we now have

$$R_j(z) = 1 + \frac{T''_j(\omega_0)}{(T'_j(\omega_0))^2} \left(T_j(\omega_0 + \omega_1 z) - T_j(\omega_0) \right), \quad j = 2, \dots, s.$$

Then $R_j(z) = e^{c_j z} + \mathcal{O}(z^3)$ with

$$c_j = \frac{T'_s(\omega_0)}{T''_s(\omega_0)} \frac{T''_j(\omega_0)}{T'_j(\omega_0)} \approx \frac{j^2 - 1}{s^2 - 1} \quad (2 \leq j \leq s-1), \quad c_s = 1.$$

For $j = 1$ only first order is possible, which yields some freedom. We put

$$b_0 = b_2, \quad b_1 = b_2, \quad (1.23)$$

so that $c_1 = c_2/T'_2(\omega_0) \approx c_2/4$. All values $t_n + c_j \tau$ now lie in the span of the integration step. The choices (1.22), (1.23) define the second-order method (1.20). \diamond

Remark 1.5 Sommeijer, Shampine & Verwer (1997) have implemented the second-order method into a FORTRAN code RKC that uses variable step sizes based on common local error control. It also works with a variable amount of stages to minimize computational costs. For that purpose it has been equipped with a spectral radius estimator. In Section 3 the behaviour of this code will be illustrated for some nonlinear problems.³⁾ \diamond

1.3 Internal Stability and Full Convergence Properties

Internal Stability

In deciding to exploit recursion (1.3), van der Houwen & Sommeijer (1980) were inspired by the stability of this two-step Chebyshev recursion for the Chebyshev semi-iterative method for solving elliptic problems, see e.g. Varga (1962, Sect. 5.1). In this truly iterative application, accumulation of round-off is bounded and merely proportional to the condition number of the matrix under consideration; see also Woźniakowski (1977). In the current, non-iterative RKC application of (1.3) the situation is different. Here we do have round-off accumulation per step which behaves quadratic with the stage number s . For practical purposes the quadratic growth with the stage number s is harmless.

Consider the error bound (1.18) with the internal stability polynomials Q_{sj} derived for linear systems (1.15). For these internal stability functions a

³⁾ The source code can be obtained from <ftp://ftp.cwi.nl/pub/bsom/rkc> or from <http://www.netlib.org/ode/>

closed expression can be derived. Using (1.16) and (1.19), (1.20) it follows by some calculations that the polynomials Q_{jk} satisfy

$$\begin{aligned} Q_{kk}(z) &= 1, \quad Q_{k+1,k} = 2 \frac{b_{k+1}}{b_k} (\omega_0 + \omega_1 z), \\ Q_{jk} &= 2 \frac{b_j}{b_{j-1}} (\omega_0 + \omega_1 z) Q_{j-1,k}(z) - \frac{b_j}{b_{j-2}} Q_{j-2,k}(z), \quad j = k+2, \dots, s, \end{aligned}$$

for $k = 1, \dots, s-2$, while

$$Q_{s-1,s-1}(z) = Q_{ss}(z) = 1 \quad \text{and} \quad Q_{s,s-1} = 2 \frac{b_s}{b_{s-1}} (\omega_0 + \omega_1 z),$$

see Verwer, Hundsdorfer & Sommeijer (1990) for details. It follows that the $b_j^{-1} Q_{jk}(z)$ satisfy the recursion for the shifted *Chebyshev polynomials of the second kind*. Hence

$$Q_{sj}(z) = \frac{b_s}{b_j} U_{s-j}(\omega_0 + \omega_1 z), \quad j = 1, \dots, s, \quad (1.24)$$

with $U_i(x)$ being the i th degree Chebyshev polynomial of the second kind, see Abramowitz & Stegun (1968).

In contrast to the first kind Chebyshev polynomials $T_i(x)$, these $U_i(x)$ are not bounded by ± 1 in the interval $[-1, 1]$. There holds $U_i(\pm 1) = (\pm 1)^i (i+1)$ and $i+1$ is also the maximal value on $[-1, 1]$. On the greater part of this interval, $U_i(x)$ does however alternate between (approximately) $+1$ and -1 . The slope of $U_i(x)$ near $x = 1$ is also larger than that of $T_i(x)$. There holds $U'_i(1) = i(i+1)(i+2)/3$ while $T'_i(1) = i^2$.

Using these properties, for any of the choices for b_j made above, that is, (1.21) and (1.22)–(1.23), it can be shown that as long as $z \in [-\beta_R, 0]$,

$$|Q_{sj}(z)| \leq \frac{b_s}{b_j} (s-j+1) (1+C\epsilon), \quad j = 1, \dots, s,$$

where ϵ is the previously introduced damping parameter and C is a constant of moderate size independent of s . Consequently, if A is negative definite and the common stability condition $\tau\lambda \in [-\beta_R, 0]$, $\lambda \in \sigma(A)$, is satisfied, we obtain from (1.18) the error bound

$$\|e_{n+1}\|_2 \leq \|e_n\|_2 + \sum_{j=1}^s \frac{b_s}{b_j} (s-j+1) (1+C\epsilon) \|r_j\|_2.$$

For both the first- and second-order method, with and without damping, examination of the parameters b_j then leads to

$$\begin{aligned} \|e_{n+1}\|_2 &\leq \|e_n\|_2 + K \sum_{j=1}^s (s-j+1) \|r_j\|_2 \\ &\leq \|e_n\|_2 + \frac{1}{2}s(s+1) K \max_j \|r_j\|_2, \end{aligned} \quad (1.25)$$

where K is again a constant of moderate size independent of A , τ and s .

Thus, stepwise accumulation of internal perturbations, such as round-off errors, is independent of the spectrum of A as long as $\tau\rho(A) \leq \beta_R$, with $\rho(A)$ denoting the spectral radius. This obviously is a tremendous improvement over the method from Example 1.4. The estimate says that perturbations grow at most *quadratically* with s . A numerical experiment in Verwer et al. (1990) shows that in actual computation this quadratic growth indeed takes place. For rounding errors this renders no problem at all. For example, if $s = 1000$, which for a serious application of course is a hypothetical value, the local perturbation is at most $\approx 10^6 \max \|r_j\|_2$. With a machine precision of 16 digits, this local perturbation still leaves 10 digits for accuracy which for PDEs is more than enough.

Full Convergence Properties

For linear semi-discrete systems (1.15), the error bound (1.25) can also be exploited to examine the fully discrete error

$$\varepsilon_n = u_h(t_n) - w_n$$

with respect to exact PDE solutions $u_h(t_n)$ restricted to a space grid. We assume again that the matrix A is negative definite and the number of stages is large enough for stability. The following convergence results can be found in Verwer et al. (1990); the analysis is akin to that of Section II.2 for standard Runge-Kutta methods.

Let $\sigma_h(t)$ be the local spatial error (II.2.6). For the first-order method (1.20) defined by the coefficient set (1.21), there holds

$$\|\varepsilon_n\|_2 \leq C \left(\tau \max_{0 \leq t \leq T} \|u_h^{(2)}(t)\|_2 + \max_{0 \leq t \leq T} \|\sigma_h(t)\|_2 \right),$$

for all $n = 1, 2, \dots$ with $n\tau \leq T$. Here C is a moderately sized constant independent of A, τ, s and the spatial mesh width h . This proves first-order temporal convergence and as C is constant, this convergence does not involve a restriction on τ in terms of h (unconditional convergence). Recall, however, that the error bound (1.25) assumes the stability condition $\tau\rho(A) \leq \beta_R$ which can be satisfied by increasing the number of stages s .

For the second-order method (1.20) defined by the coefficients (1.22)-(1.23), there holds

$$\|\varepsilon_n\|_2 \leq C \left(\frac{\tau}{s^3} \max_{0 \leq t \leq T} \|u_h^{(2)}(t)\|_2 + \tau^2 \max_{0 \leq t \leq T} \|u_h^{(3)}(t)\|_2 + \max_{0 \leq t \leq T} \|\sigma_h(t)\|_2 \right).$$

The first-order consistency of the first stage introduces the $\mathcal{O}(\tau/s^3)$ -term in this estimate. In Verwer et al. (1990) also a genuine second-order result was demonstrated for non-stiff problems. For the original second-order method of van der Houwen & Sommeijer (1980), where all intermediate stages are

only first-order consistent, the $\mathcal{O}(\tau)$ -term is a factor s larger. For practice these $\mathcal{O}(\tau)$ -terms do not matter much since s is commonly large enough to render these terms negligibly small, which means second-order convergence in practice.

In conclusion we can thus state that for $s \geq s_0$ with s_0 modestly large, the above error bound for method (1.20) shows convergence with practically order two, independent of s . This has been supported by ample practical numerical evidence, both for linear and nonlinear problems.

2 The ROCK Family

We proceed with the ROCK (Orthogonal-Runge-Kutta-Chebyshev) methods proposed by Abdulle (2001, 2002) and Abdulle & Medovikov (2001). These methods cover nearly 100% of the optimal real stability interval and there exist methods of order two and four.⁴⁾ A property shared with the RKC methods is the use of stable, two-step recurrence formulas for internal stability. However, in contrast to RKC, the ROCK methods are not given in a closed analytic form because they have been constructed partly numerically.⁵⁾ This construction is ingenious but rather lengthy; hence for a full description we refer to the original papers and here only the main results are reviewed.

2.1 Stability Polynomials

The first result to be mentioned is the following theorem due to Abdulle (2000), who has proven this result with the help of the order star theory. We consider here polynomials of degree s which give an approximation in the origin to the exponential function of order $p \geq 1$.

Theorem 2.1 *The optimal polynomials $R(z)$ as defined by the equal ripple property (1.6) possess exactly p complex zeros if p is even and exactly $p - 1$ complex zeros if p is odd. The remaining real zeros are distinct and are all in $[-\beta_R, 0]$. Furthermore, the first point z_{s-p} left of the origin with $|R(z)| = 1$ satisfies $R(z_{s-p}) = (-1)^p$.*

This theorem shows that on $[-\beta_R, 0]$ the second- and fourth-order optimal polynomials can be decomposed as

$$R(z) = W_p(z) P_{s-p}(z), \quad (2.1)$$

⁴⁾ Also the DUMKA stability polynomials (Zolotarev polynomials) cover nearly 100% of the optimal stability interval and there exist DUMKA methods of order 2, 3, 4; see Medovikov (1998), Lebedev (1994, 2000) and references therein. The DUMKA methods use a similar decomposition of the stability function as described here for ROCK.

⁵⁾ Earlier attempts to construct the optimal stability polynomials with least squares techniques (van der Houwen, 1977) and the Remes algorithm (Wanner, personal communication, 1996) failed for degrees above about 10 due to severely ill-conditioned matrix systems.

where $W_p(z)$ is a polynomial of degree p without real zeros and $P_{s-p}(z)$ is a polynomial of degree $s-p$ with $s-p$ negative distinct zeros. As $R(0) = 1$, it can be assumed that $W_p(0) = P_{s-p}(0) = 1$.

The optimal first-order stability polynomial (1.2) is an orthogonal polynomial with respect to the weight function $1/\sqrt{1-x^2}$ when shifted back to the interval $-1 \leq x \leq 1$ by setting $x = 1 + 2z/\beta_R$. General orthogonal polynomials often have excellent approximation properties and satisfy three-term recurrence relations. This led Abdulle & Medovikov (2001) to the idea of seeking approximations to the second- and fourth-order optimal stability polynomials (2.1) through suitably constructed orthogonal polynomials related to (1.2). For this particular construction, a classic theorem of Bernstein (1930) proved fundamental; this theorem generalizes the property of orthogonality and minimality of Chebyshev polynomials to more general weight functions. For (2.1) the generalized weight function is $W_p^2(x)/\sqrt{1-x^2}$ and theoretical approximations to $P_{s-p}(x)$ can be identified.

The actual construction of the p th-order ($p = 2, 4$) ROCK stability polynomials of degree $s \geq p+1$,

$$R(x) = \tilde{W}_p(x) \tilde{P}_{s-p}(x), \quad (2.2)$$

with approximate \tilde{W}_p and \tilde{P}_{s-p} , is rather laborious. It amounts to finding: (i) near-optimal positive, p th degree polynomials $\tilde{W}_p(x)$ with coefficients depending on s , and (ii) orthogonal polynomials $\tilde{P}_{s-p}(x)$ of degree $s-p$ associated with the weight function $\tilde{W}_p^2(x)/\sqrt{1-x^2}$, such that $R(x)$ results in a p th-order stability polynomial which remains bounded as long as possible on the negative real axis and at the same time provides damping; similar as for the RKC methods a damping of 5% is imposed, and (iii) three-term recurrence formulas

$$\tilde{P}_j(z) = (\mu_j z - \nu_j) \tilde{P}_{j-1}(z) - \kappa_j \tilde{P}_{j-2}(z), \quad j = 2, \dots, s-p, \quad (2.3)$$

where $\tilde{P}_0(z) = 1$ and $\tilde{P}_1(z) = 1 + \mu_1 z$.

For order $p = 2$, Abdulle & Medovikov (2001) constructed polynomials (2.2) of degree s equal to 3 up to more than 1000, with 5% damping and

$$\beta_R \approx 0.809 s^2.$$

For practical purposes this can be considered as optimal, see (1.7). Likewise, for $p = 4$ Abdulle (2002) achieved

$$\beta_R \approx 0.354 s^2.$$

Figure 2.1 shows the stability regions for the second-order polynomial of degree five (to be compared with B_5 with damping in Figure 1.3) and the fourth-order polynomial of degree nine; see also Abdulle (2002).

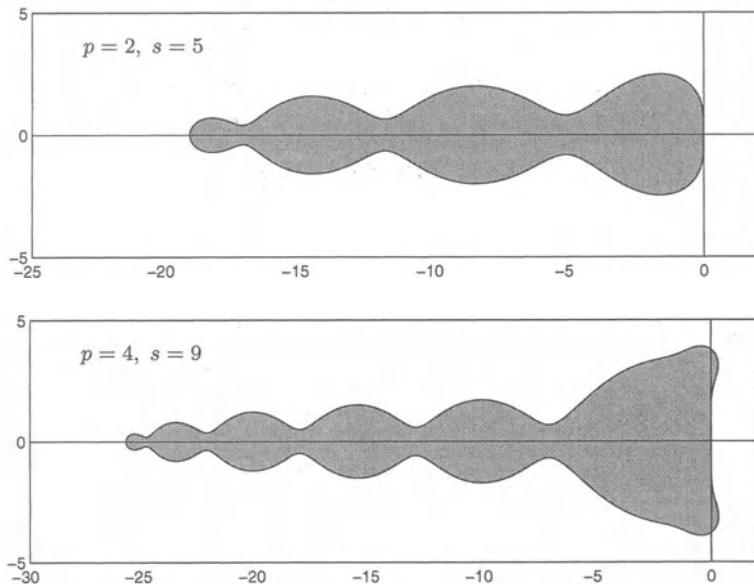


Fig. 2.1. Stability regions for the damped ROCK polynomials (2.2) of order $p = 2$, degree $s = 5$ and order $p = 4$, degree $s = 9$.

2.2 Integration Formulas

Following the idea of van der Houwen & Sommeijer (1980) for their RKC methods, the ROCK methods with $p = 2, 4$ exploit the recurrence relation (2.3) for the first $s - p$ stages. This gives

$$\begin{aligned} w_{n0} &= w_n, \\ w_{n1} &= w_{n0} + \tau \mu_1 F_{n0}, \\ w_{nj} &= \tau \mu_j F_{n,j-1} - \nu_j w_{n,j-1} - \kappa_j w_{n,j-2}, \end{aligned} \tag{2.4}$$

where $j = 2, \dots, s - p$ and F_{nk} has the same meaning as in (1.20). When applied to the test equation $w'(t) = \lambda w(t)$, these stages build up the polynomial $\tilde{P}_{s-p}(z)$ given in (2.2). The remaining factor $\tilde{W}_p(z)$ is built up by a p -stage finishing explicit Runge-Kutta procedure, so that when applied to the test equation we get $w_{n+1} = R(z)w_n$ with $R(z)$ the stability polynomial from (2.2).

For second-order methods the consistency conditions are the same for linear and nonlinear problems. This does not hold for $p > 2$, and therefore the construction of the fourth-order methods for general nonlinear problems is much more complicated. This construction is discussed in detail in Abdulle (2001, 2002) and is similar to that for the higher-order DUMKA methods of Medovikov (1998). Use is made of composition of methods by the B-series

theory, cf. Hairer et al. (1993), to set up the consistency conditions for order four.

Remark 2.2 Abdulle (2001) has implemented the second- and fourth-order methods into FORTRAN codes ROCK2 and ROCK4.⁶⁾ These codes work the same as the RKC code referred to in Remark 1.5. We will illustrate ROCK4 in Section 3. \diamond

2.3 Internal Stability and Convergence

The internal stability properties depend on the polynomials Q_{sj} occurring in (1.17) where the r_j represent, for example, round-off. Precise internal stability analysis results, as outlined in Section 1.3 for the RKC methods, are not available for the ROCK methods. Ideally, for a good internal stability behaviour, all polynomials Q_{sj} should take on modest values for $z \in [-\beta_R, 0]$ and not grow too much as z approaches its maximal value $-\beta_R$. Although such a growth seems unavoidable for the ROCK methods due to the p finishing stages based on common explicit Runge-Kutta formulas, the first $s - p$ three-term recurrence stages seem to provide sufficient damping to ensure that round-off accumulation stays within practical limits, see Abdulle (2001, Sect. 4.4). For example, if we wish to preserve approximately 7 digits for accuracy with double precision arithmetic (16 digits), the second-order method can take up to 200 stages and the fourth-order method up to 50. The stronger accumulation for the fourth-order method is due to its larger number of finishing stages. The maximal stage number in the code ROCK2 is limited to 200 and in ROCK4 to 152. In spite of this limitation put on s , we will see below that the behaviour of these codes is occasionally somewhat erratic, probably due to unfavourable propagations of errors over the stages.

As for RKC, the ROCK methods applied to linear semi-discrete systems $w'(t) = Aw(t) + g(t)$ take the form of the error scheme (1.17), that is,

$$w_{n+1} = R(\tau A) w_n + \sum_{j=1}^s Q_{sj}(\tau A) \tau g(t_n + c_j \tau).$$

The global error bounds outlined in Section 1.3 for RKC tell us that for this linear case the PDE convergence order practically equals the ODE convergence order. For the ROCK methods such detailed analytical results are not known, but it is anticipated that due to the p finishing stages order reduction can occur. In particular this can be expected with time-dependent boundary conditions because these enter the inhomogeneous term $g(t)$. When it occurs, this order reduction will manifest itself by a decrease in temporal accuracy when the spatial grid is refined, or, similarly, by a disappointing accuracy on a single fine grid due to very large error constants, see also the discussion in Section II.2.

⁶⁾ See <http://www.unige.ch/math/folks/hairer/software.html> for the source code.

Example 2.3 We illustrate this order reduction for the test problem

$$u_t = u_{xx} + f(x, t), \quad t > 0, \quad 0 < x < 1, \quad (2.5)$$

with solution

$$u(x, t) = e^{-t}x(x+1), \quad (2.6)$$

where the source term $f(x, t)$ and Dirichlet boundary conditions are derived from the prescribed, exact solution. For the spatial discretization a uniform grid with mesh width h was chosen with second-order central finite differences. Since the solution is a quadratic polynomial in x there is no spatial error and we see only temporal errors. With an adjustment for a constant step size τ , the codes ROCK2 and ROCK4 mentioned in Remark 2.2 were used to produce the Tables 2.1, 2.2.⁷⁾ Given τ and h , the stage number s was chosen minimal such that $\tau\rho(A) \leq \beta_R$.

	$\tau = \frac{1}{20}$	$\tau = \frac{1}{40}$	$\tau = \frac{1}{80}$	$\tau = \frac{1}{160}$
$h = \frac{1}{20}$	$5.0 \cdot 10^{-3}$	$6.5 \cdot 10^{-4}$	$8.9 \cdot 10^{-5}$	$1.8 \cdot 10^{-5}$
$h = \frac{1}{40}$	$2.6 \cdot 10^{-2}$	$3.1 \cdot 10^{-3}$	$3.1 \cdot 10^{-4}$	$4.0 \cdot 10^{-5}$
$h = \frac{1}{80}$	$7.2 \cdot 10^{-2}$	$9.3 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$1.4 \cdot 10^{-4}$
$h = \frac{1}{160}$	$2.6 \cdot 10^{-1}$	$4.0 \cdot 10^{-2}$	$4.8 \cdot 10^{-3}$	$5.9 \cdot 10^{-4}$

Table 2.1. Convergence test for ROCK2 applied to (2.5), (2.6). Maximum-norm global errors at $t = 1$ shown as a function of h and τ .

	$\tau = \frac{1}{20}$	$\tau = \frac{1}{40}$	$\tau = \frac{1}{80}$	$\tau = \frac{1}{160}$
$h = \frac{1}{20}$	$2.6 \cdot 10^{-1}$	$7.9 \cdot 10^{-3}$	$4.8 \cdot 10^{-4}$	$1.5 \cdot 10^{-5}$
$h = \frac{1}{40}$	$1.1 \cdot 10^{+1}$	$4.7 \cdot 10^{-1}$	$1.7 \cdot 10^{-2}$	$6.2 \cdot 10^{-4}$
$h = \frac{1}{80}$	$9.8 \cdot 10^{+2}$	$3.6 \cdot 10^{+1}$	$1.0 \cdot 10^{+0}$	$3.1 \cdot 10^{-2}$
$h = \frac{1}{160}$	$6.0 \cdot 10^{+4}$	$2.6 \cdot 10^{+3}$	$7.1 \cdot 10^{+1}$	$2.0 \cdot 10^{+0}$

Table 2.2. Convergence test for ROCK4 applied to (2.5), (2.6). Maximum-norm global errors at $t = 1$ shown as a function of h and τ .

The tables speak for themselves. On a fixed space grid the methods converge with the right order, and even beyond that, for the present range of step sizes. However, when fixing τ and halving h , in both tables the temporal error increases and most strongly for ROCK4. This means that leading error

⁷⁾ We thank Ben Sommeijer for his assistance with the numerical tests in the current and the next section.

constants C in classical ODE error bounds depend on negative powers of h , say $C = C(h^{-q})$. The current test results indicate $q \approx 2$ for the second-order methods and even $q \approx 6$ in case of the fourth-order method.

In sharp contrast to this behaviour, RKC yields nearly the same error for all h in this test, which is in accordance with the error bounds given in Section 1.3. For example, for $h = 1/160$ the corresponding errors are $1.5 \cdot 10^{-5}$, $3.0 \cdot 10^{-6}$, $5.5 \cdot 10^{-7}$, $1.2 \cdot 10^{-7}$ for $\tau = 1/20, 1/40, 1/80, 1/160$. The accuracy is high since we have prescribed a very smooth solution proportional to e^{-t} . \diamond

The order reduction in the above example could be mitigated by transforming the problem to one with homogeneous boundary conditions. Order reduction is problem dependent and the extent to which it will be felt will depend on how influential the boundary conditions and source terms are. As the above example shows, it can completely annihilate the advantage of a high, classical ODE order and it makes a code less reliable in actual applications.

It is already clear from this simple test that the present version of the ROCK codes may not be as robust as RKC. Note however that the development of the near-optimal polynomials leading to these ROCK codes is quite recent, and therefore also the implementation is not as mature as for the RKC code, for which also less robust implementations were tried before.

3 Numerical Examples

We present here some numerical results obtained with the codes RKC and ROCK4 that have been discussed in the previous sections, see in particular the Remarks 1.5 and 2.2. We omit ROCK2 here since this second-order code does not offer a real improvement over RKC. Two nonlinear 2D test problems are treated. The first comes from combustion and the second from radiation-diffusion. More comparisons are found in Abdulle (2001, 2002) and Abdulle & Medovikov (2001); these comparisons were done on fixed grids with high temporal accuracy and there ROCK4 did give in general the best results. In the present tests we will also refine the spatial grids to keep the spatial and temporal errors in balance.

Recall that the Runge-Kutta-Chebyshev codes are fully explicit and hence share the ease of use of standard, explicit Runge-Kutta integrators. Compared to implicit integrators they also have a low storage demand which makes them attractive for modestly stiff large-scale parabolic problems. The codes use a variable step size strategy based on a local L_2 -error estimation procedure governed by a tolerance parameter Tol . The step size strategies differ somewhat; see Sommeijer et al. (1997) for details on RKC and Abdulle (2001, 2002) for ROCK4. In addition, both codes work with a variable number of stages s . Having selected a step size τ , the codes minimize s under the stability criterion $\tau\rho \leq \beta_R$ where ρ is an estimate of the spectral radius of the

Jacobian matrix. The codes usually can compute ρ internally; otherwise the user has to provide ρ .

3.1 A Combustion Model

Our first problem is a scalar two-dimensional nonlinear (hotspot) problem from combustion theory,

$$u_t = d \Delta u + f(u), \quad f(u) = \frac{R}{\alpha \delta} (1 + \alpha - u) e^{\delta(1-u)},$$

defined on the unit square for $t > 0$. The problem is subjected to the initial condition $u(x, y, 0) = 1$ and for $t > 0$ to the zero Neumann boundary condition at $x = 0, y = 0$ and the Dirichlet boundary condition $u = 1$ at $x = 1, y = 1$. The parameter values in this problem are $d = 1$, $\alpha = 1$, $\delta = 20$ and $R = 5$.

This problem models a reaction of a mixture of two chemicals with u representing the temperature of the mixture. For small times u gradually increases in a small circular region around the origin. Then ignition occurs, slightly beyond $t = 0.25$, causing u to jump from near unity to $1 + \alpha$, while simultaneously a reaction front is formed which circularly propagates towards the outer Dirichlet boundaries. When the front reaches the boundary, at $t \approx 0.34$, a steady state results. Figure 3.1 shows an accurate reference solution of the traveling front along the diagonal line $x = y$ at several output times, with $r = \sqrt{x^2 + y^2}$ on the horizontal axis.

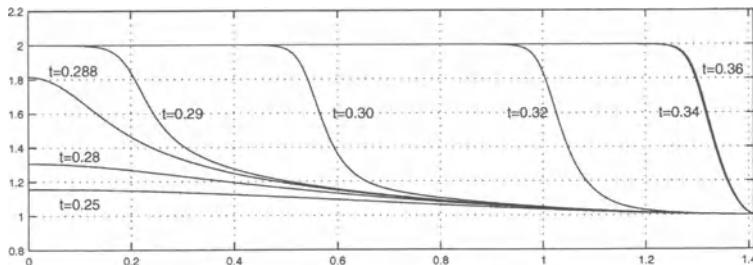


Fig. 3.1. The traveling front solution of the combustion problem along the diagonal $x = y$. At $t = 0.34, 0.36$ the plots almost coincide due to steady-state arrival.

Because of the traveling front, this problem has often been used as a test model for adaptive-grid methods for time-dependent PDE problems, in particular with smaller values of d which give steeper fronts; it has also been considered in Verwer (1996) with a previous version of RKC. A similar 3D problem with two components was treated with the current version of RKC in Sommeijer et al. (1997).

Stabilized explicit methods are natural candidates for the numerical integration of this problem. First, the traveling reaction front limits the step size of any integration scheme, be it implicit or explicit. Secondly, the problem becomes locally unstable in the course of time because $f'(u)$ varies approximately between +1000 (for $u \approx 1.6$) and -5500 (for $u \approx 2.0$). Consequently, irrespective the integrator used, rather small step sizes are required to maintain sufficient accuracy in the transient phase, especially during the ignition. Only during the starting phase and in the near approach to steady state the step size can be increased to a level that would justify the use of an implicit method. For the difficult transient phase a high order is advocated. In this regard the fourth-order code ROCK4 should perform better than RKC.

Table 3.1 shows full global errors in the max-norm at time $t = 0.32$ for the uniform 40×40 and 80×80 space grids and time tolerances $Tol = 10^{-3}, 10^{-5}, 10^{-7}$. The max-norm global space error for the two grids is also given. The spatial discretization of the Laplacian and Neumann boundary conditions was based on second-order central differencing. The codes computed the spectral radius estimate ρ internally.

The table also gives numbers of function evaluations, numbers of accepted integration steps and rejected ones. Recall that the number of function evaluations per step (the stage number s) varies with the step size. During the starting phase and in the approach to steady state larger steps with a higher s are taken than during the transient phase.

In all test cases the temporal errors are seen to be quite large compared to the imposed tolerances. This is mainly due to the local instability in the problem and the use of the maximum norm in this table (whereas the codes use L_2 -error estimators). The numerically computed fronts do have the cor-

		$h = \frac{1}{40}, \text{ err}_{\infty,s} = 5.1 \cdot 10^{-2}$		$h = \frac{1}{80}, \text{ err}_{\infty,s} = 1.1 \cdot 10^{-2}$	
Tol		err_{∞}	Costs	err_{∞}	Costs
RKC	10^{-3}	$7.2 \cdot 10^{-1}$	630 (66+5)	$7.2 \cdot 10^{-1}$	1178 (59+4)
	10^{-5}	$1.1 \cdot 10^{-1}$	1110 (350+0)	$7.8 \cdot 10^{-2}$	1945 (287+0)
	10^{-7}	$5.4 \cdot 10^{-2}$	3759 (1523+0)	$1.4 \cdot 10^{-2}$	4730 (1487+0)
ROCK4	10^{-3}	$1.8 \cdot 10^{-1}$	1104 (96+16)	$6.8 \cdot 10^{-2}$	3158 (72+14)
	10^{-5}	$8.5 \cdot 10^{-2}$	1195 (134+5)	$2.0 \cdot 10^{-2}$	1929 (112+5)
	10^{-7}	$5.1 \cdot 10^{-2}$	2223 (304+0)	$1.1 \cdot 10^{-2}$	3099 (296+0)

Table 3.1. Results for the combustion problem with L_{∞} -errors and Costs; err_{∞} is the full global error and $\text{err}_{\infty,s}$ is the global space error, both with respect to the reference solution. Costs is given as $N_f(N + N_{rej})$ where N_f is the number of function evaluations, including those for the spectral radius estimates, N is the number of successful steps and N_{rej} the number of rejected steps.

rect shape but travel somewhat too slow, causing a notable error in the front region. By further decreasing Tol and h this error will vanish. In the table the total error for ROCK4 is close to the spatial error for the smallest tolerance value; for RKC there still is a non-negligible temporal error. ROCK4 behaves slightly erratic on the 80×80 grid for $Tol = 10^{-3}$, using a large number of function evaluations; this may be due to the spectral radius estimator. Apart from this, both codes behave reliable. For the more stringent tolerance values the higher order of the ROCK4 scheme clearly starts to pay off.

3.2 A Radiation-Diffusion Model

The second test problem is a system of two strongly nonlinear diffusion equations with a highly stiff reaction term from Mousseau, Knoll & Rider (2000). The problem is known as a radiation-diffusion model and is an idealization of non-equilibrium radiation diffusion in a material. The dependent variables are E and T , representing, respectively, radiation energy and material temperature. Models of this type are found for example in laser fusion applications.

The equations are defined on the unit square for $t > 0$,

$$E_t = \nabla \cdot (D_1 \nabla E) + \sigma(T^4 - E),$$

$$T_t = \nabla \cdot (D_2 \nabla T) - \sigma(T^4 - E),$$

with

$$\sigma = \frac{Z^3}{T^3}, \quad D_1 = \frac{1}{3\sigma + |\nabla E|/E}, \quad D_2 = k T^{5/2}.$$

Here $|\nabla E|$ is the Euclidean norm and $Z = Z(x, y)$ represents the atomic mass number which may vary in the spatial domain to reflect inhomogeneities in the material. We take

$$Z(x, y) = \begin{cases} Z_0 & \text{if } |x - \frac{1}{2}| \leq \frac{1}{6} \text{ and } |y - \frac{1}{2}| \leq \frac{1}{6}, \\ 1 & \text{otherwise,} \end{cases}$$

with $Z_0 = 1$ or $Z_0 = 5$. Further, the temperature diffusion coefficient k is taken as $k = 5 \cdot 10^{-3}$. The initial values are constant in space,

$$E(x, y, 0) = 10^{-5}, \quad T(x, y, 0) = E(x, y, 0)^{1/4} \approx 5.62 \cdot 10^{-2},$$

and the boundary conditions are

$$\frac{1}{4}E - \frac{1}{6\sigma}E_x = 1 \quad \text{at } x = 0,$$

$$\frac{1}{4}E + \frac{1}{6\sigma}E_x = 0 \quad \text{at } x = 1,$$

$$T_x = 0 \quad \text{at } x = 0, 1,$$

together with homogeneous Neumann conditions for E and T at $y = 0, 1$.

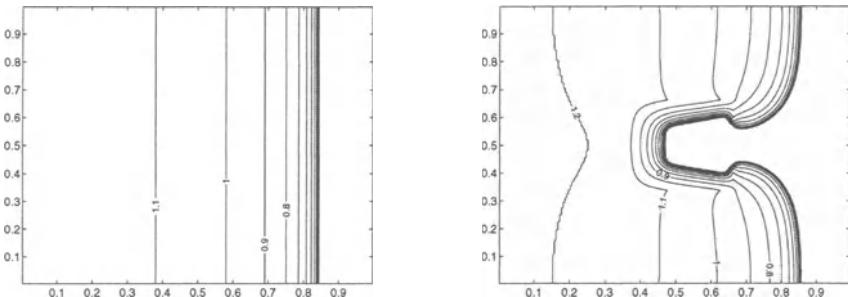


Fig. 3.2. Material temperature T at time $t = 3$ for $Z_0 = 1$ (left) and $Z_0 = 5$ (right). Contour levels: $0.1, 0.2, \dots, 1.2$.

An accurate reference solution for T at the end time $t = 3$ is displayed in Figure 3.2. The solution consists of a steep temperature front moving to the right. For $Z_0 = 1$ the solution is essentially one-dimensional. For $Z_0 = 5$ the movement is hampered by the interior region with large atomic mass number (and corresponding small diffusion). The solution of the radiation energy E is for the most part almost equal to T^4 , except near the front where it is slightly larger with a steeper profile; see also the illustrations in Mousseau et al. (2002) for different values $k = 0, 0.1$ of the temperature diffusion coefficient.

Spatial discretization: The spatial discretization has been performed on a uniform cell centered mesh with cells

$$\Omega_{ij} = [(i-1)h, ih] \times [(j-1)h, jh], \quad i, j = 1, \dots, m, \quad h = 1/m,$$

and conservative second-order central discretization. This gives approximations $E_{ij} = E_{ij}(t)$ and $T_{ij} = T_{ij}(t)$ at the cell centers of the mesh $(x_i, y_j) = ((i - \frac{1}{2})h, (j - \frac{1}{2})h)$. The diffusion coefficients D at the boundaries of cell Ω_{ij} were taken as the algebraic average of the cell values, that is,

$$D_{i+\frac{1}{2},j} = \frac{1}{2}(D_{ij} + D_{i+1,j}),$$

and for the cell values D_{ij} in the energy equation we used again central differences to approximate $|\nabla E|$.⁸⁾

Implementation of the boundary conditions has been done with virtual values. For example on the left boundary, the condition $\frac{1}{4}E - \frac{1}{6\sigma}E_x = 1$ is

⁸⁾ Note that for $Z_0 \neq 1$ the diffusion coefficients will have (large) jumps. Often it is then better to take geometric averages like

$$D_{i+\frac{1}{2},j} = 2D_{ij}D_{i+1,j}/(D_{ij} + D_{i+1,j}),$$

but for the present problem the algebraic average turned out to be preferable, probably due to the nonlinearities in the diffusion coefficients.

discretized as

$$\frac{1}{8}(E_{1j} + E_{0j}) - \frac{1}{6\sigma_{1j}h}(E_{1j} - E_{0j}) = 1,$$

defining the virtual value E_{0j} corresponding to an approximation of the energy E at $x = -\frac{1}{2}h$, $y = y_j = (j - \frac{1}{2})h$ outside the physical region. Note that due to the homogeneous Neumann condition for T we have $T_{0j} = T_{1j}$, and therefore we can take the boundary value $\sigma_{1/2,j}$ equal to $\sigma_{1j} = Z_{1j}^3/T_{1j}^3$.

Spectral radius estimation: The estimate for the spectral radius of the Jacobian was prescribed; the code's internal estimator failed on this. For that purpose we used the following heuristic considerations. The solution range for T is approximately $[0.056, 1.2]$. Hence

$$D_1 \leq \frac{1}{3\sigma} = \frac{T^3}{3Z^3} \lesssim Z^{-3}, \quad D_2 = kT^{5/2} \ll 1.$$

The contribution of diffusion to the spectral radius is therefore estimated by

$$\rho_D = 8h^{-2}Z^{-3}$$

for moderate Z -values. For the reaction term we have

$$F_R = \begin{pmatrix} Z^3 T^{-3} (T^4 - E) \\ -Z^3 T^{-3} (T^4 - E) \end{pmatrix}, \quad F'_R = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix},$$

with

$$\alpha = \frac{Z^3}{T^3}, \quad \beta = Z^3 \left(1 + \frac{3E}{T^4}\right).$$

The eigenvalues of this reaction part are 0 and $-(\alpha + \beta)$. In the expression for $\alpha + \beta$ the term Z^3/T^3 will be the dominating one. Since we know that $1/T^3 \lesssim 5.6 \cdot 10^3$ we can estimate the contribution of reaction to the spectral radius as

$$\rho_R = \alpha + \beta \leq 6000 Z^3.$$

In total we get the estimate $\rho = \rho_D + \rho_R$ for the spectral radius, which is then maximized over the spatial region. With increasing Z this spectral radius quickly becomes large.

Discussion of the results: For both choices of the atomic mass number Z , the code ROCK4 did not achieve an efficient integration. With reasonable tolerances, the code started to produce negative values after short times, leading to instability. This behaviour is probably related to the lack of internal stability at the four finishing stages of the algorithm.

The code RKC succeeded to integrate this difficult nonlinear problem. In the Tables 3.2 and 3.3, the temporal L_2 -errors are listed for $t = 3$ with various tolerances on 50×50 and 100×100 grids. The temporal errors were obtained by comparison with a reference solution, also computed with RKC but with

$Z_0 = 1$	$h = \frac{1}{50}$, $err_{2,s} = 3.0 \cdot 10^{-2}$	$h = \frac{1}{100}$, $err_{2,s} = 8.5 \cdot 10^{-3}$
Tol	$err_{2,t}$	Costs
10^{-1}	$2.3 \cdot 10^{-2}$	2175 (36+2 , 82)
10^{-2}	$3.6 \cdot 10^{-3}$	3020 (68+3 , 101)
10^{-3}	$1.3 \cdot 10^{-3}$	5779 (180+33 , 49)
Tol	$err_{2,t}$	Costs
10^{-1}	$7.4 \cdot 10^{-3}$	5207 (52+7 , 122)
10^{-2}	$3.0 \cdot 10^{-3}$	6393 (101+2 , 78)
10^{-3}	$4.4 \cdot 10^{-4}$	12484 (266+47 , 54)

Table 3.2. Results for RKC for the radiation-diffusion problem for $Z_0 = 1$ with L_2 -errors and Costs; $err_{2,t}$ is the temporal error and $err_{2,s}$ is the spatial error. Costs is given as $N_F(N + N_{rej}, s_{max})$ with N_F number of function evaluations, N number of accepted steps, N_{rej} number of rejected steps, and s_{max} the maximal number of stages.

$Z_0 = 5$	$h = \frac{1}{50}$, $err_{2,s} = 7.8 \cdot 10^{-2}$	$h = \frac{1}{100}$, $err_{2,s} = 2.7 \cdot 10^{-2}$
Tol	$err_{2,t}$	Costs
10^{-1}	$2.3 \cdot 10^{-2}$	11598 (33+3 , 459)
10^{-2}	$4.1 \cdot 10^{-3}$	15678 (67+2 , 513)
10^{-3}	$1.5 \cdot 10^{-3}$	28980 (173+27 , 249)
Tol	$err_{2,t}$	Costs
10^{-1}	$9.1 \cdot 10^{-3}$	15496 (52+7 , 395)
10^{-2}	$3.6 \cdot 10^{-3}$	18624 (99+2 , 213)
10^{-3}	$4.3 \cdot 10^{-4}$	31868 (254+20 , 142)

Table 3.3. Results for RKC for the radiation-diffusion problem for $Z_0 = 5$. Entries are as in Table 3.2.

a very small tolerance. Also given are estimated spatial L_2 -errors; these were obtained by comparison with reference solutions on finer grids (with twice as many grid points in both spatial directions). From the tables we see that with decreasing tolerance Tol the temporal errors quickly become insignificant in comparison to the spatial errors.

The results with RKC are satisfactory for $Z_0 = 1$. For $Z_0 = 5$ the system becomes very stiff leading to very high stage numbers s . Because the stiffness originates mostly from the reaction part in the equations, the performance could probably be much improved by using operator splitting, thereby using RKC only for the diffusion part and treating the reaction terms in an implicit manner. It should be noted that a full implicit treatment for the whole problem does become complicated, see Mousseau et al. (2000) and Brown & Woodward (2001), for instance. Because RKC works fully explicit, the use of this integrator is (relatively) straightforward for such nonlinear problems.

Conclusions

Among the many time stepping techniques discussed in this text, such as operator splitting, ADI, AMF, IMEX and stabilized explicit methods, there

is no overall ‘best’ method. Different classes of PDE problems require different approaches. The stabilized explicit methods are able to handle nonlinear parabolic problems with moderate stiffness in an efficient way. Moreover, these explicit methods are easy to implement and they have a low storage demand. The higher order approach of ROCK4 certainly has promise, but at this stage of development its robustness is still not satisfactory.

For problems with significant advection the code RKC is not very suited; ROCK4 then has the advantage of having a portion of the imaginary axis in its stability region, see Figure 2.1. As an alternative recent development we finally mention the approach of Bermejo & El Amrani (2001), where a combined Lagrangian-Eulerian approach has been advocated, in which advection is treated in a Lagrangian fashion and RKC is used for the remaining parabolic part.

Bibliography

- S. Abarbanel, D. Gottlieb, M.H. Carpenter (1996), *On the removal of boundary errors caused by Runge-Kutta integration of nonlinear partial differential equations*. SIAM J. Sci. Comput. 17, pp. 777–782.
- A. Abdulle (2000), *On roots and error constants of optimal stability polynomials*. BIT 40, pp. 177–182.
- A. Abdulle (2001), *Chebyshev methods based on orthogonal polynomials*. Thesis No. 3266, Dept. Math., Univ. of Geneva.
- A. Abdulle (2002), *Fourth order Chebyshev methods with recurrence relation*. SIAM J. Sci. Comput. 23, pp. 2042–2055.
- A. Abdulle, A.A. Medovikov (2001), *Second order Chebyshev methods based on orthogonal polynomials*. Numer. Math. 90, pp. 1–18.
- M. Abramowitz, I.A. Stegun (1968), *Handbook of Mathematical Functions*. Fifth edition, Dover Publications, New York.
- M.J. Aftosmis, M.J. Berger, J.E. Melton (1999), *Adaptive Cartesian mesh generation*. In: *Handbook of Grid Generation*. Eds. J.F. Thompson, B.K. Soni, N.P. Weatherill, CRC Press, Chapter 22.
- I. Ahmad, M. Berzins (1997), *An algorithm for ODEs from atmospheric dispersion problems*. Appl. Numer. Math. 25, pp. 137–149.
- G. Akrivis, M. Crouzeix, C. Makridakis (1999), *Implicit-explicit multistep methods for quasilinear parabolic equations*. Numer. Math. 82, pp. 521–541.
- D. Allen, R. Southwell (1955), *Relaxation methods applied to determining the motion, in two dimensions, of a viscous fluid past a fixed cylinder*. Quart. J. Mech. Appl. Math. 8, pp. 129–145.
- T.M. Apostol (1964), *Calculus*. Blaisdell, New York.
- A.R.A. Anderson, M.A.J. Chaplain, E.L. Newman, R.J.C. Steele, A.M. Thompson (2000), *Mathematical modelling of tumour invasion and metastasis*. J. Theor. Med. 2, pp. 129–154.
- R. Aris (1965), *Introduction to the Analysis of Chemical Reactors*. Prentice-Hall, Englewood Cliffs.
- D.C. Arney, J.E. Flaherty (1989), *An adaptive local mesh refinement method for time-dependent partial differential equations*. Appl. Numer. Math. 5, pp. 257–274.
- U.M. Ascher, L.R. Petzold (1998), *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, Philadelphia.

- U.M. Ascher, S.J. Ruuth, R.J. Spiteri (1997), *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*. Appl. Numer. Math. 25, pp. 151–167.
- U.M. Ascher, S.J. Ruuth, B. Wetton (1995), *Implicit-explicit methods for time-dependent PDE's*. SIAM J. Numer. Anal. 32, pp. 797–823.
- M.J. Baines (1994), *Moving Finite Elements*. Clarendon Press, Oxford.
- M. Bakker (1971), *Analytical aspects of a minimax problem* (in Dutch). Technical Note TN 62, Mathematical Centre, Amsterdam.
- D. Barkley (1991), *A model for fast computer simulation of waves in excitable media*. Physica D 49, pp. 61–70.
- D. Barkley (1995), *Spiral meandering*, In: *Chemical Waves and Patterns*. Eds. R. Kapral, K. Showalter, Kluwer, Dordrecht, pp. 163–189.
- R. Barrett, M. Berry, T.F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. van der Vorst (1994), *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Second edition, SIAM, Philadelphia.
- R.M. Beam, R.F. Warming (1976), *An implicit finite-difference algorithm for hyperbolic systems in conservation-law form*. J. Comput. Phys. 22, pp. 87–110.
- A.E. Berger, J.M. Solomon, M. Ciment, S.H. Leventhal, B.C. Weinberg (1980), *Generalized OCI schemes for boundary layer problems*. Math. Comp. 35, pp. 695–731.
- M.J. Berger, J. Oliger (1984), *Adaptive mesh refinement for hyperbolic partial differential equations*. J. Comput. Phys. 53, pp. 484–512.
- R. Bermejo, M. El Amrani (2001), *A finite element semi-Lagrangian explicit Runge-Kutta-Chebyshev method for reaction-diffusion problems*. Report MA-UCM 2001-20, Dept. Math., Universidad Complutense de Madrid.
- S.N. Bernstein (1930), *Sur les polynômes orthogonaux relatifs à un segment fini*. Journal de Mathématiques 9, pp. 127–177.
- M. Berzins (2001), *Modified mass matrices and positivity preservation for hyperbolic and parabolic PDEs*. Comm. Numer. Meth. in Eng. 17, pp. 659–666.
- M. Berzins, R.M. Furzeland (1986), *A user's manual for SPRINT - A versatile software package for solving systems of algebraic, ordinary and partial differential equations: Part 2 - Solving partial differential equations*. Report No. 202, Dept. of Computer Studies, Univ. of Leeds.
- C. Besse, B. Bidégaray, S. Descombes (2002), *Order estimates in time of splitting methods for the nonlinear Schrödinger equation*. SIAM J. Numer. Anal. 40, pp. 26–40.
- J.G. Blom, M.G.M. Roemer (1997), *Description of the 3D LOTOS model. Part I: Dynamics*. Report MAS-N9701, CWI, Amsterdam.
- J.G. Blom, R.A. Trompert, J.G. Verwer (1996), *Algorithm 758: VLUGR2: A vectorizable adaptive grid solver for PDEs in 2D*. ACM Trans. Math. Softw. 22, pp. 302–328.
- J.G. Blom, J.G. Verwer (1996), *Algorithm 759: VLUGR3: A vectorizable adaptive grid solver for PDEs in 3D – II. Code description*. ACM Trans. Math. Softw. 22, pp. 329–347.

- J.G. Blom, P.A. Zegeling (1994), *A moving-grid interface for systems of one-dimensional time-dependent partial differential equations*. ACM Trans. Math. Softw. 20, pp. 194–214.
- C. Bolley, M. Crouzeix (1978), *Conservation de la positivité lors de la discréétisation des problèmes d'évolution paraboliques*. RAIRO Anal. Numer. 12, pp. 237–245.
- J.P. Boris, D.L. Book (1973), *Flux corrected transport I: SHASTA, a fluid transport algorithm that works*. J. Comput. Phys. 11, pp. 38–69.
- N. Borovych, D. Drissi, M.N. Spijker (2000), *A note about Ritt's condition, related resolvent conditions and power bounded operators*. Numer. Funct. Anal. Optim. 21, pp. 425–438.
- A. Bott (1992), *Monotone flux limitation in the area-preserving flux-form advection algorithm*. Monthly Weather Review 120, pp. 2595–2602.
- E.F.F. Botta, K. Dekker, Y. Notay, A. van der Ploeg, C. Vuik, F.W. Wubs, P.M. de Zeeuw (1997), *How fast the Laplace equation was solved in 1995*. Appl. Numer. Math. 24, pp. 439–455.
- J.P. Boyd (2001), *Chebyshev and Fourier Spectral Methods*. Dover Publications, Mineola.
- R.W. Brankin, I. Gladwell, L.F. Shampine (1992), *RKSUITE: a suite of Runge-Kutta codes for the initial value problem for ODEs*. Softreport 92-S1, Dept. of Math., Southern Methodist University, Dallas.
- K. Brenan, S. Campbell, L.R. Petzold (1989), *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Second edition, SIAM, Philadelphia.
- Y. Brenier (1984), *Averaged multivalued solutions for scalar conservation laws*. SIAM J. Numer. Anal. 21, pp. 1013–1037.
- P. Brenner, M. Crouzeix, V. Thomée (1982), *Single step methods for inhomogeneous linear differential equations*. RAIRO Anal. Numer. 16, pp. 5–26.
- S.C. Brenner, L.R. Scott (1994), *The Mathematical Theory of Finite Elements*. Texts in Applied Mathematics, Vol. 15, Springer, New York.
- P.L.T. Brian (1961), *A finite-difference method of high order accuracy for the solution of three-dimensional transient heat conduction problems*. AIChE Journal 7, pp. 367–370.
- P.N. Brown, G.D. Byrne, A.C. Hindmarsh (1989), *VODE, a variable coefficient ODE solver*. SIAM J. Sci. Stat. Comput. 10, pp. 1038–1051.
- P.N. Brown, C.S. Woodward (2001), *Preconditioning strategies for fully implicit radiation diffusion with material-energy transfer*. SIAM J. Sci. Comput. 23, pp. 499–516.
- K. Burrage, J.C. Butcher (1979), *Stability criteria for implicit Runge-Kutta methods*. SIAM J. Numer. Anal. 16, pp. 46–57.
- J.C. Butcher (1975), *A stability property of implicit Runge-Kutta methods*. BIT 15, pp. 358–361.
- J.C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*. John Wiley & Sons, Chichester.

- G.D. Byrne (1992), *Pragmatic experiments with Krylov methods in the stiff ODE setting*. In: *Computational Ordinary Differential Equations*. Eds. C.J. Cash, I. Gladwell, Oxford Univ. Press, pp. 323–356.
- M.P. Calvo, J. de Frutos, J. Novo (2001), *Linearly implicit Runge-Kutta methods for advection-diffusion-reaction problems*. Appl. Numer. Math. 37, pp. 535–549.
- M.P. Calvo, C. Palencia (2002), *Avoiding the order reduction of Runge-Kutta methods for linear initial-boundary value problems*. Math. Comp. 71, pp. 1529–1543.
- C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang (1988), *Spectral Methods in Fluid Dynamics*. Springer Series in Computational Physics, Springer, New York.
- W. Cao, W. Huang, R.D. Russell (2002), *A moving mesh method based on the geometric conservation law*. SIAM J. Sci. Comput. 24, pp. 118–142.
- W. Cao, R. Carretero-Gonzalez, W. Huang, R.D. Russell (2003), *Variational mesh adaptation for axisymmetrical problems*. SIAM J. Numer. Anal. 41, pp. 235–257.
- M.H. Carpenter, D. Gottlieb, S. Abarbanel, W.S. Don (1995), *The theoretical accuracy of Runge-Kutta time discretizations for the initial-boundary value problem: a study of the boundary error*. SIAM J. Sci. Comput. 16, pp. 1241–1252.
- M.A.J. Chaplain, A.M. Stuart (1993), *A model mechanism for the chemotactic response of endothelial cells to tumour angiogenesis factor*. IMA J. Math. Appl. Med. Biol. 10, pp. 149–168.
- R. Čiegis, K. Kiškis (1994), *On the stability of LOD difference schemes with respect to boundary conditions*. Lithuanian Acad. of Sc., Informatica 5, pp. 297–323.
- E.A. Coddington, N. Levinson (1955), *Theory of Ordinary Differential Equations*. McGraw-Hill, New York.
- P. Colella, P. Woodward (1984), *The piecewise parabolic method (PPM) for gas-dynamical simulations*. J. Comput. Phys. 54, pp. 174–201.
- J.W. Cooley, J.W. Tukey (1965), *An algorithm for the machine calculation of complex Fourier series*. Math. Comp. 19, pp. 297–301.
- W.A. Coppel (1965), *Stability and Asymptotic Behaviour of Differential Equations*. Heath Mathematical Monographs, D.C. Heath & Co., Boston.
- R. Courant, K.O. Friedrichs, H. Lewy (1928), *Über die partiellen Differenzen-gleichungen der mathematischen Physik*. Math. Anal. 100, pp. 32–74.
- R. Courant, E. Isaacson, M. Rees (1952), *On the solution of nonlinear hyperbolic differential equations by finite differences*. Comm. Pure Appl. Math. 2, pp. 243–255.
- M.G. Crandall, A. Majda (1980), *Monotone difference approximations for scalar conservation laws*. Math. Comp. 34, pp. 1–21.
- J. Crank, P. Nicolson (1947), *A practical method for numerical integration of solutions of partial differential equations of heat-conduction type*. Proc. Cambridge Philos. Soc. 43, pp. 50–67.
- M. Crouzeix (1975), *Sur l'approximation des équations différentielles opérationnelles par des méthodes de Runge-Kutta*. Thesis, Univ. Paris VI.

- M. Crouzeix (1979), *Sur la B-stabilité des méthodes de Runge-Kutta*. Numer. Math. 32, pp. 75–82.
- M. Crouzeix (1980), *Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques*. Numer. Math. 35, pp. 257–276.
- M. Crouzeix, P.A. Raviart (1980), *Approximation des Problèmes d'Evolution*. Lecture Notes, Université de Rennes.
- C.F. Curtiss, J.O. Hirschfelder (1952), *Integration of stiff equations*. Proc. Nat. Acad. Sci. U.S.A. 38, pp. 235–243.
- G. Dahlquist (1956), *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand. 4, pp. 33–53.
- G. Dahlquist (1963), *A special stability problem for linear multistep methods*. BIT 3, pp. 27–43.
- G. Dahlquist (1975), *Error analysis for a class of methods of stiff nonlinear initial value problems*. Procs. Dundee Conference 1975, Lectures Notes in Math., Vol. 506, Ed. G.A. Watson, Springer, Berlin, pp. 60–74.
- R. Dautray, J.-L. Lions (1993), *Mathematical Analysis and Numerical Methods for Science and Technology 6 - Evolution Problems II*. Springer, Berlin.
- K. Dekker, J.G. Verwer (1984), *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*. North-Holland, Amsterdam.
- C. Desoer, H. Haneda (1972), *The measure of a matrix as a tool to analyze computer algorithms for circuit analysis*. IEEE Trans. Circuit Theory 19, pp. 480–486.
- B.O. Dia, M. Schatzman (1996), *Commutateurs des certain semi-groupes holomorphes et applications aux directions alternées*. RAIRO Modél. Math. Anal. Numér. 30, pp. 343–383.
- E.A. Dorfi, L. O'C. Drury (1987), *Simple adaptive grids for 1D initial value problems*. J. Comput. Phys. 69, pp. 175–195.
- J.R. Dormand, P.J. Prince (1980), *A family of embedded Runge-Kutta formulae*. J. Comp. Appl. Math. 6, pp. 19–26.
- J.L.M. van Dorsselaer, J.F.B.M. Kraaijevanger, M.N. Spijker (1993), *Linear stability analysis in the numerical solution of initial value problems*. Acta Numerica 1993, pp. 199–237.
- J. Douglas Jr. (1955), *On the numerical integration of $u_{xx} + u_{yy} = u_t$ by implicit methods*. J. Soc. Ind. Appl. Math. 3, pp. 42–65.
- J. Douglas Jr. (1962), *Alternating direction methods for three space variables*. Numer. Math. 4, pp. 41–63.
- J. Douglas Jr., J.E. Gunn (1964), *A general formulation of alternating direction methods*. Numer. Math. 6, pp. 428–453.
- J. Douglas Jr., H.H. Rachford Jr. (1956), *On the numerical solution of heat conduction problems in two and three space variables*. Trans. Amer. Math. Soc. 82, pp. 421–439.
- M. Dowle, R.M. Mantel, D. Barkley (1997), *Fast simulations of waves in three-dimensional excitable media*. Int. J. Bifurcation and Chaos 7, pp. 2529–2545.

- E.C. Du Fort, S.P. Frankel (1953), *Stability conditions in the numerical treatment of parabolic differential equations*. Math. Tables and other Aids to Computation 7, pp. 135–152.
- D.R. Durran (1999), *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, Texts in Applied Mathematics, Vol. 32, Springer, New York.
- E.G. D'Yakonov (1962), *Difference schemes with splitting operator for multi-dimensional non-stationary problems*. Zh. Vychisl. Mat. i Mat. Fiz. 2, pp. 549–568.
- E.G. D'Yakonov (1964), *Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives*. USSR Comput. Math. Math. Phys. 4(5), pp. 206–216.
- H. Dym, H.P. McKean (1972), *Fourier Series and Integrals*. Academic Press, New York.
- B.L. Ehle (1969), *A-stable methods and Padé approximations to the exponential*. SIAM J. Math. Anal. 4, pp. 671–680.
- T.M. El-Mistikawy, M.J. Werle (1978), *Numerical methods for boundary layers with blowing – the exponential box scheme*. AIAA J. 16, pp. 749–751.
- B. Engquist, S. Osher (1980), *Stable and entropy satisfying approximations for transonic flow calculations*. Math. Comp. 34, pp. 45–75.
- G. Fairweather, A.R. Mitchell (1967), *A new computational procedure for A.D.I. methods*. SIAM J. Numer. Anal. 4, pp. 163–170.
- L. Ferracina, M.N. Spijker (2003), *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*. Report Univ. Leiden; to appear.
- J.E. Flaherty, R.M. Loy, M.S. Shephard, J.D. Teresco (2000), *Software for the parallel adaptive solution of conservation laws by discontinuous Galerkin methods*. In: *Discontinuous Galerkin Methods. Theory, Computation, and Applications*. Eds. B. Cockburn, G. Karniadakis, S.-W. Shu, Lecture Notes in Computational Science and Engineering, Vol. 11, Springer, Berlin, pp. 113–124.
- J. Frank, W. Hundsdorfer, J.G. Verwer (1997), *On the stability of implicit-explicit linear multistep methods*. Appl. Numer. Math. 25, pp. 193–205.
- R. Frank, J. Schneid, C.W. Ueberhuber (1985), *Order results for implicit Runge-Kutta methods applied to stiff systems*. SIAM J. Numer. Anal. 22, pp. 515–534.
- J.N. Franklin (1959), *Numerical stability in digital and analogue computation for diffusion problems*. J. Math. Phys. 37, pp. 305–315.
- J.E. Fromm (1968), *A method for reducing dispersion in convective difference schemes*. J. Comput. Phys. 3, pp. 176–189.
- G.R. Gavalas (1968), *Nonlinear Differential Equations of Chemically Reacting Systems*. Springer Tracts in Natural Philosophy, Vol. 17, Springer, Berlin.
- C.W. Gear (1971), *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs.
- A. Gerisch, J.G. Verwer (2002), *Operator splitting and approximate factorization for taxis-diffusion-reaction models*. Appl. Numer. Math. 42, pp. 159–176.
- A. Gerisch, R. Weiner (2003), *On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods*. Comp. and Math. with Appl. 45, pp. 53–67.

- E. Godlewski, P.-A. Raviart (1996), *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Applied Mathematical Sciences, Vol. 118, Springer, New York.
- S.K. Godunov (1959), *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*. Mat. Sb. 47, pp. 271–306.
- D. Goldman, T.J. Kaper (1996), *N-th order split operator schemes and non-reversible systems*. SIAM J. Numer. Anal. 33, pp. 349–367.
- G.H. Golub, C.F. van Loan (1996), *Matrix Computations*. Third edition. John Hopkins Univ. Press, Baltimore.
- C. González, C. Palencia (1998), *Stability of time-stepping methods for abstract time-dependent parabolic equations*. SIAM J. Numer. Anal. 35, pp. 973–989.
- P. Gordon (1965), *Nonsymmetric difference equations*. J. Soc. Ind. Appl. Math. 13, pp. 667–673.
- S. Gottlieb, C.-W. Shu (1998), *Total variation diminishing Runge-Kutta schemes*. Math. Comp. 67, pp. 73–85.
- S. Gottlieb, C.-W. Shu, E. Tadmor (2001), *Strong stability preserving high-order time discretization methods*. SIAM Review 42, pp. 89–112.
- A.R. Gourlay (1970), *Hopscotch, a fast second order partial differential equation solver*. J. Inst. Maths. Applics. 6, pp. 375–390.
- A.R. Gourlay, G.R. McGuire (1971), *General hopscotch methods for partial differential equations*. J. Inst. Maths. Applics. 7, pp. 216–227.
- A.R. Gourlay, A.R. Mitchell (1969), *The equivalence of certain alternating direction and locally one-dimensional difference methods*. SIAM J. Numer. Anal. 6, pp. 37–46.
- T.E. Graedel, P.J. Crutzen (1995), *Atmosphere, Climate and Change*. Scientific American Library, Freeman and Company, New York.
- W.B. Gragg (1965), *On extrapolation algorithms for ordinary initial value problems*. SIAM J. Numer. Anal. 2, pp. 384–403.
- J.A. van de Griend, J.F.B.M. Kraaijevanger (1986), *Absolute monotonicity of rational functions occurring in the numerical study of initial value problems*. Numer. Math. 49, pp. 413–424.
- D.F. Griffiths, J.M. Sanz-Serna (1986), *On the scope of the method of modified equations*. SIAM J. Sci. Comput. 7, pp. 994–1008.
- W. Gröbner (1960), *Die Lie-Reihen und ihre Anwendungen*. VEB Deutscher Verlag der Wissenschaften, Berlin, Second edition 1967.
- W.D. Gropp (1980), *A test of moving mesh refinement for 2D scalar hyperbolic problems*. SIAM J. Sci. Statist. Comput. 1, pp. 191–197.
- A. Guillou, B. Lago (1961), *Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles, a pas séparés et a pas liés*. Recherche de formules a grand rayon de stabilité. Ier Congr. Assoc. Fran. Calcul, AFCAL, Grenoble, Sept. 1960, pp. 43–56.
- B. Gustafsson (1975), *The convergence rate for difference approximations to mixed initial boundary value problems*. Math. Comp. 29, pp. 396–406.

- B. Gustafsson, H.-O. Kreiss, J. Oliger (1995), *Time Dependent Problems and Difference Methods*. John Wiley & Sons, New York.
- B. Gustafsson, H.-O. Kreiss, A. Sundström (1972), *Stability theory of difference approximations for mixed initial boundary value problems. II*. Math. Comp. 26, pp. 649–686.
- E. Hairer, C. Lubich, G. Wanner (2002), *Geometric Numerical Integration*. Springer Series in Computational Mathematics, Vol. 31, Springer, Berlin.
- E. Hairer, S.P. Nørsett, G. Wanner (1993), *Solving Ordinary Differential Equations I – Nonstiff Problems*. Second edition, Springer Series in Computational Mathematics, Vol. 8, Springer, Berlin.
- E. Hairer, G. Wanner (1996), *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*. Second edition, Springer Series in Computational Mathematics, Vol. 14, Springer, Berlin.
- A. Harten (1983), *High resolution schemes for hyperbolic conservation laws*. J. Comput. Phys. 49, pp. 357–393.
- A. Harten, J.M. Hyman, P.D. Lax (1976), *On finite-difference approximations and entropy conditions for shocks*, with appendix by B. Keyfitz. Comm. Pure Appl. Math. 29, pp. 297–322.
- A. Harten, P.D. Lax, B. van Leer (1983), *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*. SIAM Review 25, pp. 35–61.
- C. Helzel, R.J. LeVeque, G. Warnecke (2000), *A modified fractional step method for the accurate approximation of detonation waves*. SIAM J. Sci. Comput. 22, pp. 1489–1510.
- P.W. Hemker (1977), *A Numerical Study of Stiff Two-Point Boundary Problems*. Thesis, University of Amsterdam.
- P. Henrici (1962), *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons, New York.
- D. Hilhorst, R. van der Hout, L.A. Peletier (1996), *The fast reaction limit for a reaction-diffusion system*. J. of Math. Anal. and Applic. 199, pp. 349–373.
- C. Hirsch (1988), *Numerical Computation of Internal and External Flows 1: Fundamentals and Numerical Discretization*. John Wiley & Sons, Chichester.
- R.A. Horn, C.R. Johnson (1985), *Matrix Analysis*. Cambridge University Press.
- R.A. Horn, C.R. Johnson (1991), *Topics in Matrix Analysis*. Cambridge University Press.
- Z. Horváth (1998), *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*. Appl. Numer. Math. 28, pp. 309–326.
- P.J. van der Houwen (1977), *Construction of Integration Formulas for Initial Value Problems*. North-Holland, Amsterdam.
- P.J. van der Houwen (1996), *The development of Runge-Kutta methods for partial differential equations*. Appl. Numer. Math. 20, pp. 261–273.
- P.J. van der Houwen, B.P. Sommeijer (1980), *On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values*. Z. Angew. Math. Mech. 60, pp. 479–485.

- P.J. van der Houwen, B.P. Sommeijer (2001), *Approximate factorization for time-dependent partial differential equations*. J. Comp. Appl. Math. 128, pp. 447–466.
- P.J. van der Houwen, B.P. Sommeijer, J. Kok (1997), *The iterative solution of fully implicit discretizations of three-dimensional transport models*. Appl. Numer. Math. 25, pp. 243–256.
- W. Huang, R.D. Russell (1997), *Analysis of moving mesh partial differential equations with spatial smoothing*. SIAM J. Numer. Anal. 34, pp. 1106–1126.
- W. Hundsdorfer (1992), *Unconditional convergence of some Crank-Nicolson LOD methods for initial-boundary value problems*. Math. Comp. 53, pp. 81–101.
- W. Hundsdorfer (1998a), *A note on stability of the Douglas splitting method*. Math. Comp. 67, pp. 183–190.
- W. Hundsdorfer (1998b), *Trapezoidal and midpoint splittings for initial-boundary value problems*. Math. Comp. 67, pp. 1047–1062.
- W. Hundsdorfer (1999), *Stability of approximate factorizations with θ -methods*. BIT 39, pp. 473–483.
- W. Hundsdorfer (2001), *Partially implicit BDF2 blends for convection dominated flows*. SIAM J. Numer. Anal. 38, pp. 1763–1783.
- W. Hundsdorfer (2002), *Accuracy and stability of splitting with stabilizing corrections*. Appl. Numer. Math. 42, pp. 213–233.
- W. Hundsdorfer, B. Koren, M. van Loon, J.G. Verwer (1995), *A positive finite-difference advection scheme*. J. Comput. Phys. 117, pp. 35–46.
- W. Hundsdorfer, S.J. Ruuth, R.J. Spiteri (2003), *Monotonicity-preserving linear multistep methods*. SIAM J. Numer. Anal. 41, pp. 605–623.
- W. Hundsdorfer, R.A. Trompert (1994), *Method of lines and direct discretization: a comparison for linear advection*. Appl. Numer. Math. 13, pp. 469–490.
- W. Hundsdorfer, J.G. Verwer (1989), *Stability and convergence of the Peaceman-Rachford ADI method for initial-boundary value problems*. Math. Comp. 53, pp. 81–101.
- A.M. Il'in (1969), *Differencing scheme for a differential equation with a small parameter affecting the highest derivative* (in Russian). Mat. Zametki 6, pp. 237–248.
- A. Iserles (1996), *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics, Cambridge University Press.
- A. Iserles, S.P. Nørsett (1991), *Order Stars*. Applied Mathematics and Mathematical Computation, Vol. 2, Chapman & Hall, London.
- A. Iserles, G. Strang (1983), *The optimal accuracy of difference schemes*. Trans. Amer. Math. Soc. 277, pp. 779–803.
- T. Jahnke, Ch. Lubich (2000), *Error bounds for exponential operator splitting*. BIT 40, pp. 735–744.
- R. Jeltsch (1988), *Order barriers for difference schemes for linear and nonlinear hyperbolic problems*. In: Procs. Num. Anal. Conf. Dundee 1987, Eds. D.F. Griffiths, G.A. Watson, Pitman Research Notes in Mathematics, Vol. 170, Longman.

- R. Jeltsch, O. Nevanlinna (1978), *Largest disk of stability of explicit Runge-Kutta methods*. BIT 18, pp. 500–502.
- R. Jeltsch, O. Nevanlinna (1981), *Stability of explicit time discretizations for solving initial value problems*. Numer. Math. 40, pp. 245–296.
- C. Johnson (1987), *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press.
- C.A. Kennedy, M.H. Carpenter (2003), *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*. Appl. Numer. Math. 44, pp. 139–181.
- I.P.E. Kinnmark, W.G. Gray (1984), *One-step integration methods of third-fourth order accuracy with large hyperbolic stability limits*. Math. Comput. Simulation 18, pp. 181–184.
- M. Kline (1972), *Mathematical Thought from Ancient to Modern Times*. Oxford Univ. Press, New York.
- O. Knoth, R. Wolke (1998), *Implicit-explicit Runge-Kutta methods for computing atmospheric reactive flows*. Appl. Numer. Math. 28, pp. 327–341.
- B. Koren (1993), *A robust upwind discretization for advection, diffusion and source terms*. In: *Numerical Methods for Advection-Diffusion Problems*. Eds. C.B. Vreugdenhil, B. Koren, Notes on Numerical Fluid Mechanics, Vol. 45, Vieweg, Braunschweig, pp. 117–138.
- J.F.B.M. Kraaijevanger (1985), *B-convergence of the implicit midpoint rule and the trapezoidal rule*. BIT 25, pp. 652–666.
- J.F.B.M. Kraaijevanger (1986), *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*. Numer. Math. 48, pp. 303–322.
- J.F.B.M. Kraaijevanger (1991), *Contractivity of Runge-Kutta methods*. BIT 31, pp. 482–528.
- D. Kröner (1997), *Numerical Schemes for Conservation Laws*. Wiley and Teubner, Chichester, Stuttgart.
- P. Laasonen (1949), *Über eine Methode zur Lösung der Wärmeleitungsgleichung*. Acta Math. 81, pp. 309–317.
- J.D. Lambert (1991), *Numerical Methods for Ordinary Differential Equations, the Initial Value Problem*. John Wiley & Sons, Chichester.
- P. Lancaster, M. Tismenetsky (1985), *The Theory of Matrices*. Second edition, Academic Press, New York.
- J. Lang (2000), *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications*. Lecture Notes in Computational Science and Engineering, Vol. 16, Springer, Berlin.
- J. Lang, J.G. Verwer (2001), *ROS3P – An accurate third order Rosenbrock solver designed for parabolic problems*. BIT 41, pp. 731–738.
- D. Lanser, J.G. Blom, J.G. Verwer (2001), *Time integration of the shallow water equations in spherical geometry*. J. Comput. Phys. 171, pp. 373–393.
- D. Lanser, J.G. Verwer (1999), *Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling*. J. Comp. Appl. Math. 111, pp. 201–216.

- P.D. Lax, B. Wendroff (1960), *Systems of conservation laws*. Comm. Pure Appl. Math. 13, pp. 217–237.
- V.I. Lebedev (1994), *How to solve stiff systems of differential equations by explicit methods*. In: *Numerical Methods and Applications*. Ed. G.I. Marchuk, CRC Press, pp. 45–80.
- V.I. Lebedev (2000), *Explicit difference schemes for solving stiff problems with a complex or separable spectrum*. Comp. Math. and Math. Phys. 40, pp. 1801–1812.
- B. van Leer (1974), *Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme*. J. Comput. Phys. 14, pp. 361–370.
- B. van Leer (1977), *Towards the ultimate conservative difference scheme III. Upstream-centered finite-difference schemes for ideal compressible flow*. J. Comput. Phys. 23, pp. 263–275.
- B. van Leer (1979), *Towards the ultimate conservative difference scheme V. A second order sequel to Godunov’s method*. J. Comput. Phys. 32, pp. 101–136.
- B. van Leer (1985), *Upwind-difference methods for aerodynamic problems governed by the Euler equations*. In: *Large Scale Computations in Fluid Mechanics*. Eds. B.E. Engquist, S. Osher, R.C.J. Somerville, AMS Series, American Mathematical Society, pp. 327–336.
- H.W.J. Lenferink (1989), *Contractivity preserving explicit linear multistep methods*. Numer. Math. 55, pp. 213–223.
- H.W.J. Lenferink (1991), *Contractivity preserving implicit linear multistep methods*. Math. Comp. 56, pp. 177–199.
- B.P. Leonard (1988), *Simple high accuracy resolution program for convective modeling of discontinuities*. Int. J. Numer. Meth. Fluids 8, pp. 1291–1318.
- B.P. Leonard, A.P. Lock, M.K. MacVean (1996), *Conservative explicit unrestricted time step multidimensional constancy preserving advection schemes*. Monthly Weather Review 124, pp. 2588–2606.
- R.J. LeVeque (1982), *Large time step shock-capturing techniques for scalar conservation laws*. SIAM J. Numer. Anal. 19, pp. 1091–1109.
- R.J. LeVeque (1985), *Intermediate boundary conditions for LOD, ADI and approximate factorization methods*. ICASE Report 85-21, Langley Research Center.
- R.J. LeVeque (1992), *Numerical Methods for Conservation Laws*. Lecture Notes in Mathematics, ETH Zürich, Birkhäuser Verlag, Basel.
- R.J. LeVeque (1996), *High-resolution conservative algorithms for advection in incompressible flow*. SIAM J. Numer. Anal. 33, pp. 627–665.
- R.J. LeVeque (2002), *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics, Cambridge University Press.
- S.J. Lin, R.B. Rood (1996), *Multidimensional flux-form and semi-Lagrangian transport scheme*. Monthly Weather Review 124, pp. 2046–2070.
- W.M. Lioen, J.J.B. de Swart, W.A. van der Veen (1996), *Test set for IVP solvers*. Report NM-R9615, CWI, Amsterdam.

- Ch. Lubich, A. Ostermann (1993), *Runge-Kutta methods for parabolic equations and convolution quadrature*. Math. Comp. 60, pp. 105–131.
- Ch. Lubich, A. Ostermann (1995a), *Runge-Kutta approximation of quasilinear parabolic equations*. Math. Comp. 64, pp. 601–627.
- Ch. Lubich, A. Ostermann (1995b), *Interior estimates for time discretization of parabolic equations*. Appl. Numer. Math. 18, pp. 241–251.
- Ch. Lubich, A. Ostermann (1995c), *Linearly implicit time discretization of nonlinear parabolic equations*. IMA J. Numer. Anal. 15, pp. 555–583.
- R.W. MacCormack (1969), *The effect of viscosity in hypervelocity impact cratering*. AIAA Hypervelocity Impact Paper No. 69–354.
- T.A. Manteuffel, A.B. White (1986), *The numerical solution of second-order boundary-value problems on nonuniform meshes*. Math. Comp. 47, pp. 511–535.
- G.I. Marchuk (1968), *Some applications of splitting-up methods to the solution of mathematical physics problems*. Aplikace Matematiky 13, pp. 103–132.
- G.I. Marchuk (1971), *On the theory of the splitting-up method*. In: *SYNSPADE 1970*. Ed. B. Hubbard, Procs. of the Second Symposium on the Numerical Solution of Partial Differential Equations, Academic Press, New York.
- G.I. Marchuk (1981), *Methods of Numerical Mathematics*. Second edition, Springer, Berlin.
- G.I. Marchuk (1990), *Splitting and alternating direction methods*. In: *Handbook of Numerical Analysis I*. Eds. P.G. Ciarlet, J.L. Lions, North-Holland, Amsterdam, pp. 197–462.
- V.A. Markoff (1892), *Über Polynome, die in einem gegebenen Intervall möglichst wenig von Null abweichen*. Math. Ann. 77 (1916), pp. 213–258 (translation by J. Grossman of original Russian article in Acad. Sc. St. Petersburg, 1892).
- J.E. Marsden (1970), *Basic Complex Analysis*. W.H. Freeman and Company, San Francisco.
- R.I. McLachlan (1994), *Symplectic integration of Hamiltonian wave equations*. Numer. Math. 66, pp. 465–492.
- R.I. McLachlan (1995), *On the numerical integration of ordinary differential equations by symmetric composition methods*. SIAM J. Sci. Comput. 16, pp. 151–168.
- R.I. McLachlan, G.R.W. Quispel (2002), *Splitting methods*. Acta Numerica 2002, pp. 341–434.
- A.A. Medovikov (1998), *High order explicit methods for stiff ordinary differential equations*. BIT 38, pp. 372–390.
- K. Miller (1981), *Moving finite elements II*. SIAM J. Numer. Anal. 18, pp. 1033–1057.
- A.R. Mitchell, D.F. Griffiths (1980), *The Finite Difference Method in Partial Differential Equations*. John Wiley & Sons, Chichester.
- K.W. Morton (1980), *Stability of finite difference approximations to a diffusion-convection equation*. Internat. J. Numer. Methods Engrg. 15, pp. 677–683.
- K.W. Morton (1996), *Numerical Solution of Convection-Diffusion Problems*. Applied Mathematics and Mathematical Computation, Vol. 12, Chapman & Hall, London.

- V.A. Mousseau, D.A. Knoll, W.J. Rider (2000), *Physics-based preconditioning and the Newton-Krylov method for non-equilibrium radiation diffusion*. J. Comput. Phys. 160, pp. 743–765.
- J.D. Murray (1989), *Mathematical Biology*, Biomathematics Texts, Vol. 19, Springer, Berlin.
- O. Nevanlinna (1985), *Matrix valued versions of a result of von Neumann with an application to time discretization*. J. Comp. Appl. Math. 12 & 13, pp. 475–489.
- O. Nevanlinna, W. Liniger (1979), *Contractive methods for stiff differential equations, II*. BIT 19, pp. 53–72.
- S.P. Nørsett (1974), *Semi-explicit Runge-Kutta methods*. Rep. Math. and Comp. No. 6/74, Dept. of Math., Univ. of Trondheim.
- G.G. O'Brien, M.A. Hyman, S. Kaplan (1951), *A study of the numerical solution of partial differential equations*. J. Math. and Phys. 29, pp. 223–251.
- J.M. Ortega, W.C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*. Computer Science and Applied Mathematics, Academic Press, New York.
- A. Ostermann (2002), *Stability of W-methods with applications to operator splitting and to geometric theory*. Appl. Numer. Math. 42, pp. 353–366.
- A. Ostermann, M. Roche (1993), *Rosenbrock methods for partial differential equations and fractional orders of convergence*. SIAM J. Numer. Anal. 30, pp. 1084–1098.
- L. Pareschi, G. Russo (2000), *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*. In: Advances on Computation: Theory and Praxis, Vol. 3. Ed. D. Trigiante. Nova Science Publishers, pp. 269–289.
- D. Pathria (1997), *The correct formulation of intermediate boundary conditions for Runge-Kutta time integration of initial boundary value problems*. SIAM J. Sci. Comput. 18, pp. 1255–1266.
- D.W. Peaceman (1977), *Fundamentals of Numerical Reservoir Simulation*. Elsevier, Amsterdam.
- D.W. Peaceman, H.H. Rachford, Jr. (1955), *The numerical solution of parabolic and elliptic differential equations*. J. Soc. Indust. Appl. Math. 3, pp. 28–41.
- J.E. Pearson (1993), *Complex patterns in a simple system*. Science 261, pp. 189–192.
- A.C. Petersen, E.J. Spee, H. van Dop, W. Hundsdorfer (1998), *An evaluation and intercomparison of four new advection schemes for use in global chemistry models*. J. Geophys. Res. 103, pp. 19253–19269.
- L.R. Petzold (1982), *A description of DASSL: a differential-algebraic system solver*. Procs. IMACS World Congress, Montreal, Canada.
- A. Pinkus, S. Zafrany (1997), *Fourier Series and Integral Transforms*. Cambridge University Press.
- A. Prothero, A. Robinson (1974), *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*. Math. Comp. 28, pp. 145–162.

- A. Quarteroni, A. Valli (1994), *Numerical Approximation of Partial Differential Equations*. Springer Series in Computational Mathematics, Vol. 23, Springer, Berlin.
- P.J. Rasch (1994), *Conservative shape-preserving two-dimensional transport on a spherical reduced grid*. Monthly Weather Review 122, pp. 1337–1350.
- S.C. Reddy, L.N. Trefethen (1992), *Stability of the method of lines*. Numer. Math. 62, pp. 235–267.
- R.D. Richtmyer, K.W. Morton (1967), *Difference Methods for Initial-Value Problems*. Second edition, John Wiley & Sons, Interscience Publishers, New York.
- W. Riha (1972), *Optimal stability polynomials*. Computing 9, pp. 37–43.
- P.L. Roe (1985), *Some contributions to the modeling of discontinuous flows*. Lect. Notes. Appl. Math. 22, Amer. Math. Soc., Providence, pp. 163–193.
- H.-G. Roos, M. Stynes, L. Tobiska (1996), *Numerical Methods for Singularly Perturbed Differential Equations*. Springer Series in Computational Mathematics, Vol. 24, Springer, Berlin.
- H.H. Rosenbrock (1963), *Some general implicit processes for the numerical solution of differential equations*. Comput. J. 5, pp. 329–330.
- G.L. Russell, J.A. Lerner (1981), *A new finite-differencing scheme for the tracer transport equation*. J. Appl. Meteorol. 20, pp. 1483–1498.
- S.J. Ruuth (1995), *Implicit-explicit methods for reaction-diffusion problems in pattern formation*. J. Math. Biol. 34, pp. 148–176.
- A.A. Samarskii (1962), *On an economical difference method for the solution of a multi-dimensional parabolic equation in an arbitrary region*. Zh. Vychisl. Mat. i Mat. Fiz. 2, pp. 894–926.
- A.A. Samarskij (1984), *Theorie der Differenzenverfahren*. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig (published in Russian in 1977).
- A. Sandu (1999), *Positive numerical integration methods for chemical kinetic systems*. J. Comput. Phys. 170, pp. 589–602.
- J.M. Sanz-Serna (1984), *Methods for the numerical solution of the nonlinear Schrödinger equation*. Math. Comp. 43, pp. 21–27.
- J.M. Sanz-Serna (1997), *Geometric integration*. In: *The State of the Art in Numerical Analysis*. Eds. I.S. Duff, G.A. Watson, Clarendon Press, pp. 121–143.
- J.M. Sanz-Serna, M.P. Calvo (1994), *Numerical Hamiltonian Problems*. Applied Mathematics and Mathematical Computation, Vol. 7, Chapman & Hall, London.
- J.M. Sanz-Serna, J.G. Verwer (1986), *Conservative and nonconservative schemes for the solution of the nonlinear Schrödinger equation*. IMA J. Numer. Anal. 6, pp. 25–42.
- J.M. Sanz-Serna, J.G. Verwer (1989), *Stability and convergence at the PDE/stiff ODE interface*. Appl. Numer. Math. 5, pp. 117–132.
- J.M. Sanz-Serna, J.G. Verwer, W.H. Hundsdorfer (1987), *Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations*. Numer. Math. 50, pp. 405–418.

- M. Schatzman (1999), *Stability of the Peaceman-Rachford approximation*. J. Funct. Anal. 162, pp. 219–255.
- J. Schnakenberg (1979), *Simple chemical reaction systems with limit cycle behaviour*. J. Theor. Biol. 81, pp. 389–400.
- L.F. Shampine (1994), *Numerical Solution of Ordinary Differential Equations*. Chapman & Hall, New York.
- Q. Sheng (1989), *Solving partial differential equations by exponential splittings*. IMA J. Numer. Anal. 9, pp. 199–212.
- G.I. Shishkin (1990), *Grid Approximation of Singularly Perturbed Elliptic and Parabolic Equations* (in Russian). Second doctoral thesis, Keldysh Institute, Moscow.
- C.-W. Shu (1988), *Total-variation-diminishing time discretizations*. SIAM J. Sci. Stat. Comput. 9, pp. 1073–1084.
- C.-W. Shu (1999), *High order ENO and WENO schemes for computational fluid dynamics*. In: *High-Order Methods for Computational Physics*, Eds. T.J. Barth, H. Deconinck, Lecture Notes in Computational Science and Engineering, Vol. 9, Springer, Berlin, pp. 439–582.
- C.-W. Shu, S. Osher (1988), *Efficient implementation of essentially non-oscillatory shock-capturing schemes*. J. Comput. Phys. 77, pp. 439–471.
- R.D. Skeel, M. Berzins (1990), *A method for the spatial discretization of parabolic equations in one space variable*. SIAM J. Sci. Stat. Comput. 11, pp. 1–32.
- P.K. Smolarkiewicz (1982), *The multi-dimensional Crowley advection scheme*. Monthly Weather Review 110, pp. 1968–1983.
- B.P. Sommeijer, P.J. van der Houwen, J.G. Verwer (1981), *On the treatment of time-dependent boundary conditions in splitting methods for parabolic differential equations*. Internat. J. Numer. Methods Engrg. 17, pp. 335–346.
- B.P. Sommeijer, L.F. Shampine, J.G. Verwer (1997), *RKC: An explicit solver for parabolic PDEs*. J. Comp. Appl. Math. 88, pp. 315–326.
- B.P. Sommeijer, J.G. Verwer (1980), *A performance evaluation of a class of Runge-Kutta-Chebyshev methods for solving semi-discrete parabolic differential equations*. Report NW91/80, Mathematical Center, Amsterdam.
- T. Sonar (2002), *Methods on unstructured grids, WENO and ENO recovery techniques*. In: *Hyperbolic Partial Differential Equations: Theory, Numerics and its Applications*. Eds. A. Meister, J. Struckmeier, Vieweg, Wiesbaden, pp. 115–268.
- P. Sonneveld, B. van Leer (1985), *A minimax problem along the imaginary axis*. Nieuw Archief voor Wiskunde 3, pp. 19–22.
- M.N. Spijker (1983), *Contractivity in the numerical solution of initial value problems*. Numer. Math. 42, pp. 271–290.
- R.J. Spiteri, S.J. Ruuth (2002), *A new class of optimal high-order strong-stability-preserving time-stepping schemes*. SIAM J. Numer. Anal. 40, pp. 469–491.
- B. Sportisse (2000), *An analysis of operator splitting in the stiff case*. J. Comput. Phys. 161, pp. 140–168.

- T. Steihaug, A. Wolfbrandt (1979), *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations*. Math. Comp. 33, pp. 521–534.
- G. Steinebach (1995), *Order reduction of ROW methods for DAEs and method of lines applications*. Preprint 1741, Technische Hochschule Darmstadt, Germany.
- J. Stoer, R. Bulirsch (1980), *Introduction to Numerical Analysis*, Second edition, Texts in Applied Mathematics, Vol. 12, Springer, New York.
- G. Strang (1962), *Trigonometric polynomials and difference methods of maximal accuracy*. J. Math. and Phys. 41, pp. 147–154.
- G. Strang (1963), *Accurate partial difference methods I: linear Cauchy problems*. Arch. Rat. Mech. Anal. 12, pp. 392–402.
- G. Strang (1968), *On the construction and comparison of difference schemes*. SIAM J. Numer. Anal. 5, pp. 506–517.
- G. Strang (1986), *Introduction to Applied Mathematics*. Wellesley-Cambridge Press.
- G. Strang, J. Fix (1973), *An Analysis of the Finite Element Method*. Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs.
- C.J. Strikwerda (1989), *Finite Difference Schemes and Partial Differential Equations*. Chapman & Hall, New York.
- M. Suzuki (1990), *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*. Phys. Lett. A 146, pp. 319–323.
- D.A. Swayne (1987), *Time dependent boundary and interior forcing in locally one-dimensional schemes*. SIAM J. Sci. Statist. Comput. 8, pp. 755–767.
- P.K. Sweby (1984), *High resolution schemes using flux-limiters for hyperbolic conservation laws*. SIAM J. Numer. Anal. 21, pp. 995–1011.
- T. Tang (1998), *Convergence analysis of operator-splitting methods applied to conservation laws with stiff source terms*. SIAM J. Numer. Anal. 35, pp. 1939–1968.
- V. Thomée (1984), *Galerkin Finite Element Methods for Parabolic Problems*. Lecture Notes in Mathematics, Vol. 1054, Springer, Berlin.
- V. Thomée (1990), *Finite difference methods for linear parabolic equations*. In: *Handbook of Numerical Analysis I*. Eds. P.G. Ciarlet, J.L. Lions, North-Holland, pp. 5–196.
- J.F. Thompson, B.K. Soni, N.P. Weatherill (1999), *Handbook of Grid Generation*. CRC Press, Boca Raton.
- J.F. Thompson, Z.U.A. Warsi, C.W. Mastin (1985), *Numerical Grid Generation*. North-Holland, New York.
- L.N. Trefethen (1983), *Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundström*. J. Comput. Phys. 49, pp. 199–217.
- L.N. Trefethen (1984), *Instability of difference models for hyperbolic initial boundary value problems*. Comm. Pure Appl. Math. 37, pp. 329–367.
- L.N. Trefethen (1997), *Pseudospectra of linear operators*, SIAM Review 39, pp. 383–406.

- R. Tyson, S.R. Lubkin, J.D. Murray (1999), *Model and analysis of chemotactic bacterial patterns in a liquid medium*. J. Math. Biol. 38, pp. 359–375.
- A. Vande Wouwer, Ph. Saucez, W.E. Schiesser (Eds.) (2001), *Adaptive Method of Lines*. Chapman & Hall / CRC.
- J.M. Varah (1980), *Stability restrictions on second order, three-level finite-difference schemes for parabolic equations*. SIAM J. Numer. Anal. 17, pp. 300–309.
- R.S. Varga (1962), *Matrix Iterative Analysis*. Prentice Hall, Englewood Cliffs.
- A.E.P. Veldman, K. Rinzema (1992), *Playing with nonuniform grids*. J. Eng. Math. 26, pp. 119–130.
- J.G. Verwer (1984), *Contractivity of locally one-dimensional splitting methods*. Numer. Math. 44, pp. 247–259.
- J.G. Verwer (1986), *Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines*. In: *Numerical Analysis*. Eds. D.F. Griffiths, G.A. Watson, Pitman Research Notes in Mathematics, Vol. 140, pp. 220–237.
- J.G. Verwer (1996), *Explicit Runge-Kutta methods for parabolic partial differential equations*. Appl. Numer. Math. 22, pp. 359–379.
- J.G. Verwer, J.G. Blom, R.M. Furzeland, P.A. Zegeling (1989), *A moving grid method for one-dimensional PDEs based on the method of lines*. In: *Adaptive Methods for Partial Differential Equations*. Eds. J.E. Flaherty, P.J. Paslow, M.S. Shephard, J.D. Vasilakis, SIAM Proceedings Series, pp. 160–175.
- J.G. Verwer, J.G. Blom, W. Hundsdorfer (1996), *An implicit-explicit approach for atmospheric transport-chemistry problems*. Appl. Numer. Math. 20, pp. 191–209.
- J.G. Verwer, W.H. Hundsdorfer, J.G. Blom (2002), *Numerical time integration for air pollution models*. Surveys on Mathematics for Industry 10, pp. 107–174.
- J.G. Verwer, W.H. Hundsdorfer, B.P. Sommeijer (1990), *Convergence properties of the Runge-Kutta-Chebyshev method*. Numer. Math. 57, pp. 157–178.
- J.G. Verwer, E.J. Spee, J.G. Blom, W. Hundsdorfer (1999), *A second order Rosenbrock method applied to photochemical dispersion problems*. SIAM J. Sci. Comput. 20, pp. 1456–1480.
- J.G. Verwer, B. Sportisse (1998), *A note on operator splitting in a stiff linear case*. Report MAS-R9830, CWI, Amsterdam.
- R. Wait, A.R. Mitchell (1985), *Finite Element Analysis and Applications*. John Wiley & Sons, Chichester.
- G. Wanner, E. Hairer, S.P. Nørsett (1978), *Order stars and stability theorems*. BIT 18, pp. 475–489.
- R.F. Warming, B.J. Hyett (1974), *The modified equation approach to the stability and accuracy analysis of finite difference methods*. J. Comput. Phys. 14, pp. 159–179.
- R.F. Warming, R.M. Beam (1979), *An extension of A-stability to alternating direction methods*. BIT 19, pp. 395–417.

- A. Weiser, M.F. Wheeler (1988), *On convergence of block-centered finite differences for elliptic problems*. SIAM J. Numer. Anal. 25, pp. 351–375.
- P. Wesseling (2001), *Principles of Computational Fluid Dynamics*, Springer Series in Computational Mathematics, Vol. 29, Springer, Berlin.
- G.B. Whitham (1974), *Linear and Nonlinear Waves*. Wiley-Interscience, New York.
- D.L. Williamson, P.J. Rasch (1989), *Two-dimensional semi-Lagrangian transport with shape-preserving interpolation*. Monthly Weather Review 117, pp. 102–129.
- H. Woźniakowski (1977), *Numerical stability of the Chebyshev method for the solution of large linear systems*. Numer. Math. 28, pp. 191–209.
- N.N. Yanenko (1971), *The Method of Fractional Steps*. Springer, Berlin.
- H. Yoshida (1990), *Construction of higher order symplectic integrators*. Phys. Lett. A 150, pp. 262–268.
- Yuan' Chzao-Din (1958), *Some difference schemes for the solution of the first boundary value problem for linear differential equations with partial derivatives*, Thesis (in Russian), Moscow State University.
- H.C. Yuen, W.E. Ferguson (1978), *Relationships between Benjamin-Feir instability and recurrence in the nonlinear Schrödinger equation*. Phys. Fluids 21, pp. 1275–1278.
- V.E. Zakharov, A.B. Shabat (1972), *Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media*. Sov. Phys. JETP 34, pp. 62–69.
- S.T. Zalesak (1979), *Fully multidimensional flux-corrected transport algorithms for fluids*. J. Comput. Phys. 31, pp. 335–362.
- S.T. Zalesak (1987), *A preliminary comparison of modern shock-capturing schemes: linear advection*. In: *Advances in Computer Methods for Partial Differential Equations*. Eds. R. Vichnevetsky, R.S. Stepleman, IMACS Proceedings 1987, pp. 15–22.
- P.A. Zegeling (1999), *Moving grid techniques*. In: *Handbook of Grid Generation*. Eds. J.F. Thompson, B.K. Soni, N.P. Weatherill, CRC Press, Chapter 37.
- P.A. Zegeling, J.G. Verwer, J.C.H. van Eijkeren (1992), *Application of a moving grid method to a class of 1D brine transport problems in porous media*. Int. J. Num. Methods in Fluids 15, pp. 175–191.
- X. Zhong (1996), *Additive semi-implicit Runge-Kutta methods for computing high-speed nonequilibrium reactive flows*. J. Comp. Phys. 128, pp. 19–31.
- Z. Zlatev (1995), *Computer Treatment of Large Air Pollution Models*. Kluwer, Dordrecht.

Index

- absolute monotonicity, 186
- ADI (Alternating Direction Implicit) method, *see* splitting method
- adsorption model, 222
- advection discretization
 - κ -scheme, 219
 - box scheme, 252
 - conservative form
 - first-order upwind, 79, 215
 - limiting, *see* limiting
 - second-order central advection, 80
 - third-order upwind biased, 216
 - third-order upwind-biased, 83
 - direct space-time, *see* direct space-time discretization
 - finite elements, *see* finite element method
 - first-order upwind, 52, 81
 - fourth-order central, 60
 - high-resolution scheme, 221
 - optimal-order schemes, 59
 - second-order central, 53, 81
 - second-order upwind, 70
 - second-order upwind-biased, 70
 - spatially implicit, 250
 - third-order upwind-biased, 60
- advection-diffusion discretization, 66
 - adaptive upwind, 254
 - advection-dominated, 66
 - cell Péclet number, 67
 - El-Mistikawy-Werle, 255
 - exponential fitting, 254, 316
 - fourth-order central, 66
 - operator compact implicit, 253
 - second-order central, 66
 - spatially implicit, 250
- Allen-Southwell-II'in scheme, *see* advection-diffusion discretization, exponential fitting
- AMF (Approximate Matrix Factorization), *see* splitting method, Rosenbrock AMF
- anti-diffusion, 216
- artificial diffusion (dissipation), dispersion, 55, 58, 114
- atmospheric chemistry, 6
- Barkley model, 301
- BCH (Baker-Campbell-Hausdorff) formula, 327
- BDF (Backward Differentiation Formula), *see* linear multistep method
- boundary condition
 - Dirichlet, Neumann, Robin, 17
 - inflow, 17
 - outflow, 84
 - periodic, 9
- boundary layer, 84, 279
- boundary treatment
 - cell centered grid, 90, 222
 - higher-order schemes, 92
 - Neumann diffusion, 88
 - outflow central advection, 87
 - vertex centered grid, 90
- brine transport model, 319
- Buckley-Leverett equation, 237
- Butcher array, 140
- Cartesian grid, 293, *see also* transformed Cartesian grid
- Cauchy-Schwarz inequality, 29
- cell average, 1, 78, 83, 274, 293
- cell centered grid, 90, 272

- cell centered scheme, 272
- cell Péclet number, 67, 120, 253
- cell-vertex scheme, 258
- CFL condition, 102
 - explicit multistep method, 182
 - explicit Runge-Kutta method, 150
- CFL number, *see* Courant number
- characteristic polynomial, 174
- characteristics, 10, 16, 99
- chemical kinetics
 - mass action law, 3, 4, 7
 - mass conservation law, 3, 6
 - positivity, 3, 6, 121
 - production-loss form, 5
 - rate function, 4
 - stoichiometric coefficient, 4
 - stoichiometric matrix, 5
- chemo-taxis, 19
- clipping, 224, 225
- combustion model, 439
- comparison principle, 117
- compression, 229
- conservation law (nonlinear), 233
 - advective form, 233
 - conservative space differencing, 235
 - shock speed, 225, 235
- consistency
 - direct space-time discretization, 101
 - spatial discretization, 71, 72, 86
 - time integration
 - θ -method, 43
 - ADI method, 379
 - linear multistep method, 172
 - LOD method, 355, 362
 - Rosenbrock method, 152
 - Runge-Kutta method, 141
- contractivity, 196
- control volume, 264, 314
- convection (advection), 2
- convergence
 - direct space-time discretization, 101
 - spatial discretization, 71, 72, 85
 - first-order upwind, 75
 - second-order central advection, 75
 - second-order central diffusion, 75
 - third-order upwind-biased, 75
 - time integration
 - B -convergence, 170
- θ -method, 43
- ADI method, 380
- linear multistep method, 182
- LOD method, 353, 363
- Runge-Kutta method, 141
- Runge-Kutta MOL, 155
- Courant number, 102, 149, 307
- Courant-Friedrichs-Lowy condition
 - see CFL condition, 102
- Courant-Isaacson-Rees scheme, 96, 109
- Crank-Nicolson scheme, 125
- cylindrical coordinates, 310
- DAE (Differential Algebraic Equation) system, 205
- Dahlquist barrier
 - first barrier, 175
 - second barrier, 179
- diffusion discretization
 - 5-point, 9-point stencil, 301
 - fourth-order central, 65
 - higher-order schemes, 65
 - second-order central, 62, 75, 82
- direct (Kronecker) product, 297
- direct space-time discretization, 95
 - 2D advection, 303
 - conservative form, 248
 - Courant-Isaacson-Rees scheme, 109
 - Lax-Wendroff scheme, 98
 - limiting, 245
 - MacCormack scheme, 247, 305
 - non-uniform grid, 282
 - optimal-order scheme, 239
 - QUICKEST scheme, 247
- DIRK, *see* Runge-Kutta method
- discretization error (global)
 - direct space-time discretization, 100
 - operator splitting, 328
 - spatial discretization, 72, 73, 157
 - time integration
 - θ -method, 43
 - ADI method, 380
 - linear multistep method, 171
 - LOD method, 357, 363
 - Runge-Kutta method, 140
 - Runge-Kutta MOL, 157, 161
- discretization error (local)
 - direct space-time discretization, 100
 - operator splitting, 326

- spatial discretization, 73
- local error decomposition, 86
- time integration
- θ -method, 43
- ADI method, 379
- linear multistep method, 172
- LOD method, 355, 362
- Runge-Kutta method, 141
- Runge-Kutta MOL, 158
- divergence
 - free, 16
 - operator, 15
- domain of dependence, 102
- donor-cell scheme, 306
- Douglas method, *see* splitting method, ADI
- DST scheme, *see* direct space-time discretization
- Du Fort-Frankel scheme, 372
- DUMKA method, 419
- eigenvalue criterion, *see* stability, time integration
- ENO (essentially non-oscillatory) schemes, 220
- Euler's method, 24, 142
- fast Poisson solver, 397
 - FISHPACK, 397
- FCT (Flux Corrected Transport), 220
- finite element method, 283, 312
 - coercivity property, 289
 - conforming, 289
 - energy norm, 288
 - essential boundary condition, 285
 - fourth-order advection scheme, 287
 - Galerkin, 285
 - mass matrix, 287
 - moving finite element, 321
 - natural boundary condition, 285
 - Petrov-Galerkin, 291, 313
 - stiffness matrix, 287
 - streamline-diffusion, 313
 - Taylor-Galerkin, 292
 - test space, 285
 - trial space, 285
 - unstructured grid, 312
 - weak form, 284, 312
- finite volume method, 79, 215, 294
 - non-uniform grid, 264
 - unstructured grid, 314, 315
- flux limiting, *see* limiting
- flux splitting, 237
- Fourier analysis, 10, 49
 - discrete transform, 52
 - dispersion relation, 14
 - FFT (Fast Fourier Transform), 52
 - Fourier coefficient, 11, 50
 - Fourier mode, 10, 49, 56, 111, 298
 - Fourier series, 11
 - frequency, 14
 - inversion formula, 50
 - phase error, 58
 - phase velocity, 58
 - spatial discretization
 - first-order upwind advection, 56
 - fourth-order central advection advection, 61
 - second-order central advection, 56
 - second-order central diffusion, 63
 - third-order upwind-biased advection, 61
 - time integration
 - von Neumann analysis, 111, 295, 306, 307
 - wave number, 11, 14, 52
- fractional order, 363
- fractional step method, *see* splitting method, operator splitting
- Fromm scheme, *see* advection discretization, second-order upwind-biased
- fully discrete approximation, 94
- Gibbs phenomenon, 14
- GKS theory, 94, 115
- global error, *see* discretization error (global)
- Godunov barrier, 119
- gradient operator, 15
- Gray-Scott model, 21
- grid orientation, 302, 304
- grid refinement, *see also* moving grid
 - non-body fitted grid, 323
 - static regridding, 321
- Hölder inequality, 28
- Helmholtz equation, 397

- hopscotch method, *see* splitting method, ADI
- Il'in scheme, *see* advection-diffusion discretization, exponential fitting
- IMEX method, *see* splitting method
- Laasonen scheme, 125
- Lagrange interpolation, 240
- Lagrangian scheme, 99
- Laplace operator, 15
 - 5-point, 9-point stencil (2D), 301
 - grid orientation (2D), 301
- Lax equivalence theorem, 101
- Lax-Wendroff scheme, 98, 113
- leap-frog scheme, 173, 372
- Lie operator, 333
- limiting
 - direct space-time discretization, 245
 - higher-order methods, 281
 - limiter, 216
 - κ -scheme, 219
 - Koren limiter, 217
 - MUSCL limiter, 247
 - superbee limiter, 247
 - target limiter, 219
 - third-order upwind-biased, 217
 - van Leer limiter, 219
 - non-uniform cell centered grid, 281
 - positivity, 216, 221
- linear invariant, *see* linear ODE conservation law
- linear multistep method, 170, *see also* splitting method, IMEX
 - Adams-Basforth, 173
 - Adams-Moulton, 173
 - backward differentiation (BDF), 173
 - codes, 205
 - explicit midpoint rule (2-step), 173, 176
 - predictor-corrector, 173
- linear ODE conservation law, 140
- linearization, 47
- Lipschitz condition
 - classical, two-sided, 23
 - one-sided, 46, 148
- local error, *see* discretization error (local)
- local error decomposition
- Runge-Kutta MOL, 162
- spatial truncation error, 86
- LOD (Locally One-Dimensional) method, *see* splitting method
- logarithmic norm, 32, 40, 41, 44, 83, 350
- MacCormack scheme, 247, 305
- mass conservation law, 1, 78
- mass flux, 1
- mass lumping, 287
- matrix
 - M -matrix, 262
 - circulant, 50
 - companion, 180
 - condition number, 28
 - exponential, 30
 - Hermitian, 29
 - non-negative definite, 29
 - normal, 29, 104
 - orthogonal, 29
 - positive definite, 29
 - skew-Hermitian, 29
 - skew-symmetric, 29
 - symmetric, 29
 - unitary, 29
- maximum modulus theorem, 37
- maximum principle, 9, 118
- mean-value theorem, 25, 44, 47
- method of lines, 94
- mixing ratio, 16, 342
- modified equation
 - θ -method, 108
 - first-order upwind, 54, 75
 - operator splitting, 328
 - second-order central advection, 75
 - second-order central diffusion, 63, 75
 - third-order upwind-biased, 75
 - third-order upwind-biased advection, 60
- MOL, *see* method of lines
- Molenkamp-Crowley problem, 338, 376
- moving grid, 316
 - arc-length monitor, 318
 - dynamic regridding, 316
 - equidistribution, 317
 - interface MGI, 319
 - monitor function, 317
 - moving finite element, 321

- Newton iteration (modified), 127, 405
- non-uniform grid, 264, 308, 311, 317, 321
 - smooth grid, 266
- norm
 - absolute, 29, 39
 - discrete L_p , 28
 - matrix, 28
 - spectral, 28
- norm invariance, 51
- numerical diffusion (dissipation), dispersion, *see* artificial diffusion (dissipation), dispersion
- OCI (Operator Compact Implicit) scheme, 253
- ODE codes, 205
- one-leg formulation, 184
- one-way wave equation, 12
- operator splitting method, *see* splitting method, operator splitting
- order conditions
 - linear multistep method, 172
 - Rosenbrock method, 152
 - Runge-Kutta method, 141
- order reduction
 - spatial discretization, 86
 - time integration
 - DIRK method, 165
 - fourth-order Runge-Kutta method, 159, 163
 - implicit midpoint rule, 163, 165
 - implicit trapezoidal rule, 164
 - LOD method, 353
 - ROCK method, 437
 - Rosenbrock method, 170
 - Runge-Kutta MOL, 155
 - Strang splitting, 347
- ozone model, 7
- Péclet number, 84, 288, *see also* cell Péclet number
- Padé approximation, 146
- parasitic root, 176
- Parseval's identity, 11, 50
- pattern formation, 21, 301, 394
- Peaceman-Rachford method, *see* splitting method, ADI
- Poisson equation, 397
- positivity
 - chemical reaction system, *see* chemical kinetics
 - implicit spatial discretization, 261
 - ODE system
 - linear case, 117
 - nonlinear case, 116
 - spatial discretization, *see also* TVD property
 - advection, linear case, 118, 215
 - diffusion, linear case, 119
 - limited advection, 216, 221
 - linear systems, 227
 - time integration
 - θ -method, linear case, 122
 - θ -method, nonlinear case, 124
 - 2-step Adams method, 195
 - 2-step BDF method, 194, 195
 - absolute monotonicity, 186
 - Bolley-Crouzeix theorem, 186
 - explicit trapezoidal rule, nonlinear case, 124
 - linear multistep method (linear case), 192
 - one-step method (linear case), 185
 - one-step method (nonlinear case), 190
 - optimal 3-step, 4-step method, 193
 - order-one barrier, 187
 - Padé polynomials, 188
 - Shu linear multistep form, 193
 - Shu-Osher DIRK form, 190
 - threshold factor γ_R , 186, 192
 - power boundedness, 104, 148, 180
 - predictor-corrector method, *see* linear multistep method
- principal root, 176
- pseudo-spectra, 115
- quasi-uniform grid, 271
- radiation-diffusion model, 441
- residual error, *see* truncation error (local)
- resolvent condition, 115
- Richardson extrapolation, 331
- RKC (Runge-Kutta-Chebyshev) method, 420
 - Bakker polynomial, 423

- Chebyshev polynomial of the first kind, 420
- Chebyshev polynomial of the second kind, 431
- Chebyshev recursion, 420
- damped stability polynomial, 424
- equal ripple property, 423
- first-order stability polynomial, 420
- full convergence property, 432
- internal stability, 426, 430
- internal stability polynomial, 427, 431
- RKC code, 430
- RKC formula, 428
- second-order stability polynomial, 422
- shifted Chebyshev polynomial, 420
- stability boundary estimates, 423
- ROCK (Orthogonal-Runge-Kutta-Chebyshev) method, 433
- equal ripple property, 433
- internal stability, 436
- optimal stability polynomial, 433
- order reduction, 437
- ROCK formula, 435
- ROCK2, ROCK4 code, 436
- stability boundary estimates, 434
- root condition, 175
- root locus curve, 177
- Rosenbrock method, 151, *see also* splitting method, Rosenbrock AMF
 - linearized θ -method, 154
- round-off errors, 48, 426
- Runge-Kutta method, 139
 - θ -method, 35, 94, 103
 - classical fourth-order method, 143
 - codes, 205
 - collocation (Gauss, Radau, Lobatto), 143
 - DIRK (diagonally implicit), 144
 - DUMKA, *see* DUMKA method
 - explicit (forward) Euler, 24, 142
 - explicit midpoint rule (one-step), 142
 - explicit trapezoidal rule, 103, 142
 - Heun's third-order method, 143
 - implicit (backward) Euler, 35, 143
 - implicit midpoint rule, 143
 - implicit trapezoidal rule, 35, 143
- Kraaijevanger's positive second-order method, 191
- RKC, *see* RKC (Runge-Kutta-Chebyshev) method
- ROCK, *see* ROCK (Orthogonal-Runge-Kutta-Chebyshev) method
- Shu-Osher form, 190
- Scharfetter-Gummel scheme, *see* advection-diffusion discretization, exponential fitting
- Schnakenberg model, 394
- Schrödinger equation, 128, 209
- semi-discrete system, 48, 71, 94
- semi-Lagrangian scheme, 99, 240
- Shishkin grid, 280
- shock speed, 225, 235
 - Rankine-Hugoniot relation, 235
- Simpson quadrature, 288
- Sobolev space, 284
- soliton solutions, 129
- spectral method, 288
- spectral radius, 28
- spectrum, 28
- splitting method
 - ADI (Alternating Direction Implicit), 369
 - Douglas (stabilizing correction method), 373
 - hopscotch, 371
 - Peaceman-Rachford, 370
 - boundary correction, 346, 365
 - dimension splitting, 337, 343
 - IMEX (Implicit Explicit), 383
 - θ method, 383
 - 2-step Adams, 388
 - 2-step BDF, 388
 - Crank-Nicolson Leap-Frog, 387
 - LOD (Locally One-Dimensional), 348
 - backward Euler, 349
 - Crank-Nicolson, 351
 - trapezoidal splitting, 359
 - operator splitting, 325
 - BCH (Baker-Campbell-Hausdorff) formula, 327
 - commutator, 326, 333, 334
 - Lie operator, 333
 - Strang (parallel) splitting, 329
 - Strang (symmetrical) splitting, 329

- Rosenbrock AMF (Approximate Matrix Factorization), 398
- spurious root*, *see* parasitic root
- stability*
 - direct space-time discretization, 100, 241
 - linear ODE system, 31
 - spatial discretization, 71, 72, 87
 - L_2 -stability advection, 59
 - time integration
 - A -stability, 37, 146, 179
 - $A(\alpha)$ -stability, 179
 - B -stability, 148
 - G -stability, 185
 - L -stability, 38, 146
 - θ -method (advection-diffusion), 105
 - ADI method, 374
 - CFL condition, 103
 - eigenvalue criterion, 104, 109, 149
 - imaginary boundary, 150
 - linear multistep method, 174, 180
 - LOD method, 350, 352, 360
 - nonlinear results, 44
 - one-step method, 39, 43
 - power boundedness, 104, 148
 - real boundary, 151, 420, 433
 - Rosenbrock AMF method, 400
 - Rosenbrock method, 153
 - stability function, 37, 103, 145, 153
 - stability region, 37, 103, 145, 176
 - stiffness, *see* stiffness
 - strong A -stability, 38, 146
 - von Neumann analysis, 111, 295, 306, 307
 - zero-stability, 175
- stabilized explicit Runge-Kutta method*, *see* DUMKA, ROCK, RKC
- stabilizing correction method*, *see* splitting method, ADI, Douglas
- stage order*, 142, 144, 159
- stencil*, 59
- stiffness*, 9, 36, 37, 64, 144, 148, 347, 419
- symmetric ODE method*, 200
- taxis model*, *see* chemo-taxis
- threshold factor* γ_R , 186, 192
- time splitting*, *see* splitting method, operator splitting
- total variation*, 227
- transformed Cartesian grid, 308
- advection-diffusion-reaction model, 310
- contravariant basis, 310
- covariant basis, 310
- curvilinear coordinate lines, 308
- grid generator, 309
- transient phase*, 27, 35
- truncation error (local)*
 - direct space-time discretization, 99
 - implicit spatial discretization, 251
 - operator splitting, 326, 328
 - spatial discretization, 72
 - time integration
 - θ -method, 42
 - linear multistep method, 172
- tumorous tissue model*, 206
- tumour angiogenesis model*, 20, 134, 409
- tumour invasion model*, 410
- TVB (Total Variation Boundedness*, 231)
- TVD (Total Variation Diminishing)*
 - backward Euler, 230
 - explicit Euler, 228
 - linear systems, 226
 - nonlinear systems, 228
 - ODE methods, 230
 - Shu-Osher form, 230
- unstructured grid*, 311, *see also* finite volume method, finite element method
- variable step size control*, 197
 - asymptotic expansion, 198
 - error per step, 203
 - step size selection, 197
- variation of constants formula*, 30
- vertex centered grid*, 90, 265
- vertex centered scheme*, 265
- von Neumann*
 - analysis, 111, 295, 306, 307
 - condition, 113
 - theorem, 41, 148
- Voronoi box*, 315
- wave equation*, 12
- wiggles (numerical oscillations)*, 53