

38

Springer Series in
Computational
Mathematics

High Order Difference Methods for Time Dependent PDE

Bertil Gustafsson



Springer

Editorial Board

R. Bank
R.L. Graham
J. Stoer
R. Varga
H. Yserentant

Bertil Gustafsson

High Order Difference Methods for Time Dependent PDE

With 94 Figures and 12 Tables



Bertil Gustafsson
Ledungsvägen 28
75440 Uppsala
Sweden
bertil@stanford.edu

ISBN 978-3-540-74992-9

e-ISBN 978-3-540-74993-6

DOI [10.1007/978-3-540-74993-6](https://doi.org/10.1007/978-3-540-74993-6)

Springer Series in Computational Mathematics ISSN 0179-3632

Library of Congress Control Number: 2007940500

Mathematics Subject Classification (2000): 65M06

© 2008 Springer-Verlag Berlin Heidelberg

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. Violations are liable to prosecution under the German Copyright Law.

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: WMX Design GmbH, Heidelberg

Typesetting: by the author using a Springer L^AT_EX macro package

Production: LE-T_EX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Many books have been written on finite difference methods (FDM), but there are good reasons to write still another one. The main reason is that even if higher order methods have been known for a long time, the analysis of stability, accuracy and effectiveness is missing to a large extent. For example, the definition of the formal high order accuracy is based on the assumption that the true solution is smooth, or expressed differently, that the grid is fine enough such that all variations in the solution are well resolved. In many applications, this assumption is not fulfilled, and then it is interesting to know if a high order method is still effective. Another problem that needs thorough analysis is the construction of boundary conditions such that both accuracy and stability is upheld. And finally, there has been quite a strong development during the last years, in particular when it comes to very general and stable difference operators for application on initial-boundary value problems.

The content of the book is not purely theoretical, neither is it a set of recipes for various types of applications. The idea is to give an overview of the basic theory and construction principles for difference methods without going into all details. For example, certain theorems are presented, but the proofs are in most cases left out. The explanation and application of the theory is illustrated by using simple model examples. Among engineers, one is often talking about “toy problems” with a certain scepticism, under the assumption that the results have no significance for real world problems. When looking at the scientific production over the years, there is a certain truth to this claim. A method may be working very well, and better than other well known methods, for a certain model problem in one space dimension, but it is not even clear how the method can be generalized to several space dimensions. In this book we try to avoid falling into this trap. The generalization should of course always be possible in the sense that an algorithm based on the same principle can be constructed for the full problem, and hopefully the essential properties of the numerical method have been caught by the analysis of the simpler problem. Sometimes, the theoretical considerations carry over without difficulty, but sometimes we have to rely upon intuition and numerical experiments to be confident about the performance on the large problem. On the other hand, there are many cases where the model problem tells it all. For example, for a hyperbolic system of PDE in one

space dimension, one can transform the system to a diagonal one and limit the analysis to a scalar PDE with constant coefficients. By applying the transformation back, and generalizing to variable coefficients, the results hold for the original problem as well.

When discussing stability and accuracy, the model examples in this book are often of low order accuracy, since they have a simpler structure. The goal is simply to illustrate how the theory works, and it is easier to see the basic mechanisms for simpler schemes. Once the application of the theory is well understood, it should be clear how to apply it for more complicated and powerful high order methods. However, there are two main application areas where we choose to go into more detail about the full implementation of the suggested methods. One application is wave propagation with relevance for acoustics and electromagnetics. In this case the most interesting problems are so large, that there is no realistic low order computational alternative. The other one is incompressible flow governed by the Navier-Stokes equations. Even if low order methods have been used for many applications, the real challenge is turbulent flow, where there is no realistic alternative to high order methods if the smallest scales are to be represented well. In both cases, we present a fairly detailed description of the methods. One good reason for this is to illustrate how the analysis and construction described in earlier chapters is carried out for a more technically complicated problem.

There is a pervading theme in the book, and that is compactness of the computational stencils. A comparison between two approximations of the same order always comes out in favor of the more compact one, i.e., as few grid points as possible should be involved, both in space and time. The smaller error constants often make quite a dramatic change. Padé approximations and staggered grids are examples of this, as well as the box scheme described in 8.

The outline of the book is as follows.

In Chapter 1 there is an analysis of the effectiveness of higher order methods. It is based on Fourier analysis, and the necessary number of points per wave length is estimated for different types of PDE and different orders of accuracy.

Chapter 2 contains a survey of the theory for well-posedness and stability, and the different tools for analysis are described. These tools are based on Fourier analysis for problems with periodic solutions and the energy method and Laplace transform method for initial-boundary value problems. Different kinds of stability definitions are necessary in this case, and we discuss the implications of each one.

In Chapter 3 we discuss how the order of accuracy is connected to the convergence rate, i.e., how fast the numerical solution approaches the true solution when the grid is refined. This is a straightforward analysis for periodic problems, but less obvious when boundaries are involved.

Chapters 2 and 3 can be read independently of the rest of the book as an introduction and a survey of the stability theory for difference approximations in general.

The next three chapters contain a systematic presentation of the different ways of constructing high order approximations. There are the standard centered difference operators, but also Padé type operators as well as schemes based on staggered grids. When constructing the difference schemes, one principle is to approximate the dif-

ferential equation in space first, and then approximate the resulting ODE system in time. But there are other ways to obtain effective difference schemes where the space and time discretizations are not separated, and this is described in Chapter 6.

In Chapter 7 we bring up the special problem of constructing proper approximations of the boundary conditions, and of the modifications of the scheme near the boundaries. A major part of this chapter is devoted to the so called SBP operators, which are based on summation by parts leading to stability in the sense of the energy method.

Chapter 8 is a little different from the other ones, since the box scheme discussed there is only second order accurate. However, because of the compact nature, it is sometimes more accurate than higher order ones, and some recent results regarding this property are presented. Another advantage is that it can easily be generalized to nonuniform grids.

The next two chapters are devoted to applications. The purpose is not to give a survey of problems and numerical methods, but rather to illustrate the application of certain high order methods in some detail for a few problems of high interest in the engineering and scientific communities. In this way, we hope to give an idea of how to handle the technical details. In Chapter 9 we discuss wave propagation problems described by first order hyperbolic PDE with application in acoustics and electromagnetics. Here we concentrate on a class of high order methods based on staggered grids, and demonstrate the effectiveness, even if the solutions are not smooth. We also present a new method of embedding the boundary with a certain definition of the coefficients in the PDE such that the true boundary conditions are well represented. In Chapter 10 we discuss incompressible fluid dynamics. The flow is governed by the Navier–Stokes equations, but we give a special presentation of the Stokes equations, since they play a special role in the Navier–Stokes solver. The method is semi-implicit and fourth order accurate in space. Large linear systems of equations must be solved for each time step, and we present the iterative solver in some detail as well.

A big challenge, particularly in gas dynamics, is computation of solutions to nonlinear problems containing discontinuities or shocks. This requires some special theory and specialized methods. Techniques that work well for nonlinear problems with smooth solutions, like the incompressible Navier–Stokes equations in Chapter 10, are not well suited. We discuss the theory and methods for these problems in Chapter 11. This is an area where low order methods dominate even more than for linear problems, but we give some emphasis on those methods that can be generalized to high order.

Each chapter has a summary section at the end. These sections contain a brief summary of the theory and results of the chapter, and also some historical remarks and comments on available literature.

When electronic computers came in use in the forties, the field of Numerical Analysis expanded very quickly. Already from the beginning, the solution of ordinary and partial differential equations was in focus, and the development of FDM set full speed. Hardly no other methods were considered, and it was not until the late sixties, that finite element methods (FEM) started emerging for solving PDE

problems. In the beginning, FEM were used mostly for steady state problems, and for approximation of the space part of time dependent problems. For approximation in time, finite difference methods were used, and this is still the case for many applications. The great advantage with FEM is their flexibility when it comes to approximation of irregular domains. This flexibility is shared by finite volume methods (FVM), but construction of high order FVM is not that easy.

The geometric flexibility of FEM is not shared by spectral and pseudo-spectral methods, which was the next class of numerical methods that emerged for PDE. Their strength is the very high accuracy relative to the required work. However, these methods have at least as many restrictions as FDM on the type of computational domain, in particular those that are based on approximation by trigonometric polynomials (Fourier methods). Later this difficulty was partly overcome by the use of spectral element methods, where the domain is partitioned into many subdomains with orthogonal polynomials used for approximation on each one of them.

During the last decade, discontinuous Galerkin methods have arisen as a new interesting class of methods. They can be seen as a generalization of FEM, and have the potential of leading to faster algorithms.

In the final chapter, we give a brief introduction to all of the methods mentioned here.

The available commercial and public software is today dominated by algorithms based on finite element methods (or finite volume methods for problems in fluid problems). One reason for this is that the development of unstructured grid generators has made enormous progress during the last decades, and this is important for many applications. On the other hand, for problems where structured grids provide an acceptable representation of the computational domain, it is hard to beat a good high order difference method when it comes to implementation, speed and effectiveness.

We have limited the presentation in this book almost exclusively to uniform grids. The use of Cartesian grids is for convenience, we could of course use any other of the classic coordinate systems like for example cylindrical coordinates. In general, the Cartesian uniform grid can be seen as the model for all cases where a smooth mapping takes the physical domain to a rectangle in 2-D or hyper-rectangle in higher dimensions.

If a smooth mapping cannot be used for transformation of the whole domain, one can use Cartesian coordinates together with some sort of interpolation procedure near the boundary, and construct the finite differences locally on unstructured grids. Another way is to construct one local structured grid that fits the irregular boundary, and another grid that is used in the main part of the domain. These two (or more) grids are then connected via interpolation. A third way is to couple a finite difference method with a finite element or finite volume method near the boundary, and in this way arriving at a hybrid method. Still another way is to embed the irregular boundary in a larger domain with regular boundaries, and to enforce the boundary conditions by some modification of the PDE.

If the need for a nonstructured grid arises from the fact that the solution varies on very different scales, the most general finite difference technique is based on piecewise uniform grids, that are coupled by an interpolation procedure.

There is also the possibility to construct FDM directly on unstructured grids in the whole computational domain. However, the stability is then a harder issue, and furthermore, the effectiveness will not be much better than with FEM.

Acknowledgment

The main part of this book was written after my retirement from the chair at the Division of Scientific Computing at Uppsala University. However, the university has provided all the necessary infrastructure, such as libraries and computers during the whole project. Furthermore, the staff at the Department of Information Technology have been very helpful when needed. During the writing period I have also spent some time as a visitor at Stanford University, and I want to express my gratitude to the Center for Turbulence Research (CTR) and the Institute for Computational Mathematics in Engineering (ICME) for providing very good working conditions.

Finally I would like to thank my wife Margareta for enduring still another major undertaking from my side after formal retirement.

Uppsala, Sweden, November 2007

Bertil Gustafsson

Contents

1	When are High Order Methods Effective	1
1.1	Preliminaries	1
1.2	Wave Propagation Problems	2
1.3	Parabolic Equations	8
1.4	Schrödinger Type Equations	11
1.5	Summary	12
2	Well-posedness and Stability	13
2.1	Well Posed Problems	13
2.2	Periodic Problems and Fourier Analysis	16
2.2.1	The PDE Problem	17
2.2.2	Difference Approximations	21
2.3	Initial–Boundary Value Problems and the Energy Method	29
2.3.1	The PDE Problem	29
2.3.2	Semidiscrete Approximations	33
2.3.3	Fully Discrete Approximations	38
2.4	Initial–Boundary Value Problems and Normal Mode Analysis for Hyperbolic Systems	41
2.4.1	Semidiscrete Approximations	41
2.4.2	Fully Discrete Approximations	59
2.5	Summary	66
3	Order of Accuracy and the Convergence Rate	69
3.1	Periodic Solutions	69
3.2	Initial–Boundary Value Problems	72
3.3	Summary	79
4	Approximation in Space	81
4.1	High Order Formulas on Standard Grids	81
4.2	High Order Formulas on Staggered Grids	85
4.3	Compact Padé Type Difference Operators	87

4.4	Optimized Difference Operators	91
4.5	Summary	93
5	Approximation in Time	95
5.1	Stability and the Test Equation	95
5.2	Runge–Kutta Methods	97
5.3	Linear Multistep Methods	102
5.4	Deferred Correction	108
5.5	Richardson Extrapolation	111
5.6	Summary	113
6	Coupled Space-Time Approximations	115
6.1	Taylor Expansions and the Lax–Wendroff Principle	115
6.2	Implicit Fourth Order Methods	117
6.3	Summary	124
7	Boundary Treatment	127
7.1	Numerical Boundary Conditions	127
7.2	Summation by Parts (SBP) Difference Operators	130
7.3	SBP Operators and Projection Methods	140
7.4	SBP Operators and Simultaneous Approximation Term (SAT) Methods	147
7.5	Summary	155
8	The Box Scheme	157
8.1	The Original Box Scheme	157
8.2	The Shifted Box Scheme	161
8.3	Two Space Dimensions	165
8.4	Nonuniform Grids	169
8.5	Summary	176
9	Wave Propagation	177
9.1	The Wave Equation	177
9.1.1	One Space Dimension	178
9.1.2	Two Space Dimensions	185
9.2	Discontinuous Coefficients	192
9.2.1	The Original One Step Scheme	193
9.2.2	Modified Coefficients	201
9.2.3	An Example with Discontinuous Solution	206
9.3	Boundary Treatment	209
9.3.1	High Order Boundary Conditions	209
9.3.2	Embedded Boundaries	210
9.4	Summary	216

10 A Problem in Fluid Dynamics	219
10.1 Large Scale Fluid Problems and Turbulent Flow	219
10.2 Stokes Equations for Incompressible Flow	220
10.3 A Fourth Order Method for Stokes Equations	223
10.4 Navier–Stokes Equations for Incompressible Flow	228
10.5 A Fourth Order Method for Navier–Stokes Equations	231
10.6 Summary	242
11 Nonlinear Problems with Shocks	245
11.1 Difference Methods and Nonlinear Equations	245
11.2 Conservation Laws	246
11.3 Shock Fitting	251
11.4 Artificial Viscosity	252
11.5 Upwind Methods	257
11.6 ENO and WENO Schemes	261
11.7 Summary	265
12 Introduction to Other Numerical Methods	267
12.1 Finite Element Methods	267
12.1.1 Galerkin FEM	267
12.1.2 Petrov–Galerkin FEM	281
12.2 Discontinuous Galerkin Methods	283
12.3 Spectral Methods	289
12.3.1 Fourier Methods	290
12.3.2 Polynomial Methods	295
12.4 Finite Volume Methods	300
A Solution of Difference Equations	307
B The Form of SBP Operators	311
B.1 Diagonal H -norm	311
B.2 Full H_0 -norm	317
B.3 A Padé Type Operator	323
References	325
Index	331

Acronyms

BDF	Backward Differentiation Formulas
CFL	Courant–Friedrichs–Levy
CG	Conjugate Gradient (methods)
CGSTAB	Conjugate Gradient Stabilized (method)
DIRK	Diagonal Implicit Runge–Kutta (methods)
DFT	Discrete Fourier Transform
ENO	Essentially Nonoscillating
ERK	Explicit Runge–Kutta (methods)
FDM	Finite Difference Method(s)
FEM	Finite Element Method(s)
FFT	Fast discrete Fourier Transform
FVM	Finite Volume Method(s)
GKS	Gustafsson–Kreiss–Sundström
IBVP	Initial–Boundary Value Problems
IRK	Implicit Runge–Kutta (methods)
n-D	n space Dimensions
ODE	Ordinary Differential Equations
PDE	Partial Differential Equations
RK	Runge–Kutta
SAT	Simultaneous Approximation Term
SBP	Summation By Parts
SSP	Strong Stability Preserving
TV	Total Variation
TVD	Total Variation Diminishing
WENO	Weighted Essentially Nonoscillating

Chapter 1

When are High Order Methods Effective?

In the modern era of computational mathematics beginning in the forties, most methods in practical use were first or second order accurate. Actually, that is the case even today, and the reason for this low accuracy is probably the simpler implementation. However, from an efficiency point of view, it is most likely that they should be substituted by higher order methods. These require more programming work, and the computer has to carry out more arithmetic operations per grid point. However, for a given error tolerance, the number of grid points can be reduced substantially, and in several dimensions, one may well reduce the computing time and memory requirement by orders of magnitude.

In this chapter, we shall investigate how the order of accuracy affects the performance of the method. We shall use simple model problems to get an idea of what we can expect for different types of PDE.

1.1 Preliminaries

Every numerical method for solution of differential equations is based on some sort of discretization, such that the computer can handle it in finite time. The most common discretization parameter is the step size h , which denotes the typical distance between points in a grid where the solution can be computed. If the true solution can formally be expressed as an infinite sum, the discretization parameter is N , which denotes the finite number of terms in the approximating sum. For difference methods, the approximation is related to the differential equation by the *truncation error* τ , and the *order of accuracy* is defined as p if $\tau \sim h^p$. Under certain conditions that will be described in Chapter 3, this leads to error estimates of the same order, i.e., the *error in the solution* is also proportional to h^p . Those methods that have $p > 2$ are usually called higher order methods.

The difference approximations throughout this book will be built by the basic centered, forward and backward difference operators on a uniform grid with step size h .

$$x_j = jh, \quad j = 0, \pm 1, \pm 2, \dots$$

Grid functions in space are defined by $u(x_j) \rightarrow u_j$, and the difference operators are

$$\begin{aligned} D_0 u_j &= (u_{j+1} - u_{j-1})/(2h), \\ D_+ u_j &= (u_{j+1} - u_j)/h, \\ D_- u_j &= (u_j - u_{j-1})/h. \end{aligned}$$

We also define the shift operator by

$$E u_j = u_{j+1}.$$

All the difference operators commute, such that for example

$$D_0 D_+ D_- = D_+ D_- D_0 = D_0 D_- D_+.$$

For any difference operator Q we use the simplified notation Qu_j , which is to be interpreted as $(Qu)_j$. This notation is used even when j is fixed, i.e., Qu_0 means $(Qu)_{j=0}$.

The time discretization is done on a uniform grid

$$t_n = nk, \quad n = 0, 1, \dots,$$

where k is the time step. The approximation of a function $u(x_j, t_n)$ is denoted by u_j^n .

1.2 Wave Propagation Problems

In this section we shall consider wave propagation problems represented by the simple model equation

$$u_t = u_x$$

satisfied by the simple wave $e^{i\omega(x+t)}$, where ω is the wave number. It may seem as a complication to consider complex solutions in the analysis, also when we know that the solutions are real, but actually it is a simplification. The reason is that the algebraic operations become easier in this way when Fourier analysis is used.

The most straightforward way of finding the order of accuracy of a certain difference approximation is Taylor expansion. It is easily shown that for any sufficiently smooth function $u(x)$, we have

$$D_0 u(x) = u_x + \frac{h^2}{6} u_{xxx} + \mathcal{O}(h^4),$$

i.e., D_0 is a second order approximation of $\partial/\partial x$. The leading part of the truncation error can now be eliminated by including a difference approximation of it. Again, it is easily shown by Taylor expansion that

$$D_0 D_+ D_- u(x) = u_{xxx} + \mathcal{O}(h^2),$$

which gives us the fourth order approximation

$$Q_4 u(x) = u_x + \mathcal{O}(h^4),$$

where

$$Q_4 = D_0 \left(I - \frac{h^2}{6} D_+ D_- \right).$$

After a few more Taylor expansions, one obtains the sixth order approximation

$$Q_6 = D_0 \left(I - \frac{h^2}{6} D_+ D_- + \frac{h^4}{30} D_+^2 D_-^2 \right).$$

Since the solution of our model problem is periodic, we consider the computational domain $0 \leq x \leq 2\pi$, $0 \leq t$, and the grid

$$x_j = jh, \quad j = 0, 1, \dots, N, \quad (N+1)h = 2\pi.$$

With the notation $Q_2 = D_0$, we have the three alternative approximations

$$\frac{du_j}{dt} = Q_p u_j, \quad p = 2, 4, 6. \quad (1.1)$$

With the ansatz

$$u_j(t) = \hat{u}(t) e^{i\omega x_j},$$

we get the *Fourier transform* of the equation (1.1)

$$\frac{d\hat{u}}{dt} = \hat{Q}_p \hat{u},$$

where \hat{Q}_p is the Fourier transform of Q_p . Since

$$\begin{aligned} D_0 e^{i\omega x} &= \frac{i}{h} \sin \xi e^{i\omega x}, \\ D_+ D_- e^{i\omega x} &= -\frac{4}{h^2} \sin^2 \frac{\xi}{2} e^{i\omega x}, \end{aligned}$$

where $\xi = \omega h$, we get

$$\begin{aligned} \hat{Q}_2 &= \frac{i}{h} \sin \xi, \\ \hat{Q}_4 &= \frac{i}{h} \sin \xi \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} \right), \\ \hat{Q}_6 &= \frac{i}{h} \sin \xi \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} + \frac{8}{15} \sin^4 \frac{\xi}{2} \right). \end{aligned} \quad (1.2)$$

The solution in Fourier space is $\hat{u}(t) = \exp(\hat{Q}_p t)$, which gives the solution in physical space

$$u_j(t) = e^{i\omega x_j + \hat{Q}_p t}.$$

The approximation changes $i\omega$ in the exponent to $\hat{Q}_p = \hat{Q}_p(\xi)$, and it makes sense to compare these quantities. After normalization by the factor h/i , we compare ξ with $h\hat{Q}_p(\xi)/i$. Assuming for convenience that N is an even number, the highest wave number that can be represented on the grid is $\omega = N/2 = (\pi - h/2)/h$. Since h is arbitrarily small, the range of ξ is $0 \leq |\xi| \leq \pi$. Figure 1.1 shows how $h\hat{Q}_p(\xi)/i$ approaches ξ for increasing p . The true wave speed for our problem is -1, and for the approximation it is $-\hat{Q}_p/(i\omega)$, which is always less than 1 in magnitude. The waves will be slowed down by the approximation, more for higher frequencies, and this error is called the *dispersion error*.

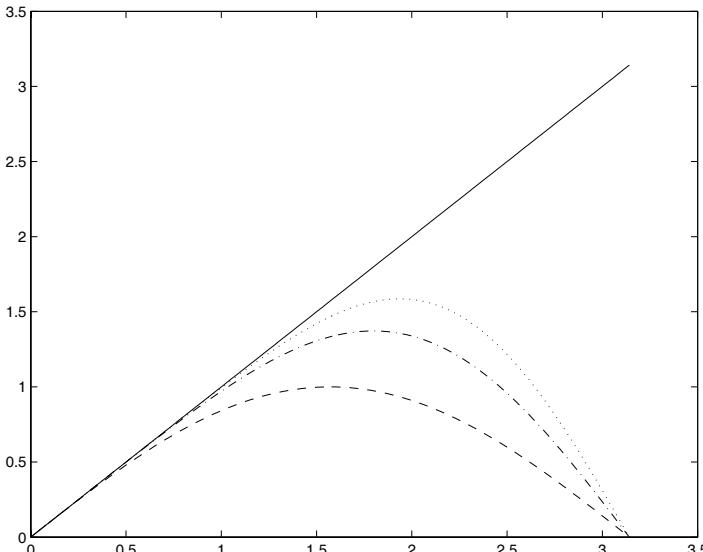


Fig. 1.1 $\xi(-), \hat{Q}_p(\xi)h/i$, $p = 2(--)$, $4(-\cdot)$, $6(\cdots)$

Next we will estimate the number of grid points $N_p = N + 1$ that are necessary for achieving a certain accuracy, and for convenience, we assume that ω is positive. The solution is periodic in both time and space, and the length of one period in time is $2\pi/\omega$. If we want to compute q periods, the time interval is $0 \leq t \leq 2\pi q/\omega$. We also introduce the number of grid points per wavelength

$$M_p = N_p/\omega = 2\pi/\xi, \quad p = 2, 4, 6.$$

We assume that $\xi \ll 1$, and investigate the error defined as

$$v^{(p)}(t) = \max_j |e^{i\omega(x_j+t)} - e^{i\omega x_j + \hat{Q}_p t}| = |1 - e^{(-i\omega + \hat{Q}_p)t}|.$$

By Taylor expansion in terms of ξ , we obtain

$$v^{(2)}(t) \approx \frac{\omega t \xi^2}{6} \leq \frac{\pi q \xi^2}{3} = \frac{4\pi^3 q}{3M_2^2}. \quad (1.3)$$

By prescribing the maximum error ε for $v^{(2)}$, we obtain an expression for M_2 in terms of q and ε . By applying the same procedure for $p = 4$ and $p = 6$, we get the complete list

$$\begin{aligned} M_2 &\approx 2\pi \left(\frac{\pi q}{3\varepsilon}\right)^{1/2}, \\ M_4 &\approx 2\pi \left(\frac{\pi q}{15\varepsilon}\right)^{1/4}, \\ M_6 &\approx 2\pi \left(\frac{\pi q}{70\varepsilon}\right)^{1/6}. \end{aligned} \quad (1.4)$$

Since the work per grid point increases by a constant factor (independent of q and ε) for each level of increased accuracy, we note that a higher order method always wins if the accepted error level is low enough, and/or the time interval is large enough. Furthermore, the gain is more pronounced in several space dimensions. If an explicit time integrator is used, there is a limit on the time step for stability reasons. If the number of grid points in each space direction can be reduced by a factor $\alpha > 1$ by using a higher order method, the total reduction of the number of grid points for a problem with three space dimensions, is a factor α^4 .

Indeed, there is a substantial gain already in one space dimension, and quite modest error levels. Table 1.1 shows M_p for a 1 % error level and $q = 20$ and $q = 200$ respectively.

Table 1.1 M_p for 1 % error level

q	M_2	M_4	M_6
20	287	28	13
200	909	51	20

For several space dimensions, the total number of grid points for a second order method becomes totally unrealistic for long time integration. Clearly it pays to use a fourth order method in these cases, even if the computing time per grid point is longer. Going to sixth order is more doubtful in one or two space dimensions.

The formal order of accuracy, as well as the estimates above, are derived under the assumption that the solution $u(x, t)$ is smooth. In practice this is seldom the case, and one might wonder how the higher order methods behave for less smooth solutions. Consider for example the problem

$$\begin{aligned} u_t &= u_x, \quad -1 \leq x \leq 1, \quad 0 \leq t, \\ u(x, 0) &= |\sin(\pi x/2)|^r, \end{aligned} \tag{1.5}$$

where r is an odd number. The solution $|\sin(\pi(x+t)/2)|^r$ is 2-periodic in both time and space. The derivative of order r is discontinuous, i.e., the solution becomes smoother for higher r . Figures 1.2, 1.3, 1.4 show the solutions for $r = 1, 3, 5$ and its approximations $u^{(p)}$ obtained by a formally p th order accurate method. The figures show the solution at $t = 6$ when the true solution is back at its initial state for the third time. Even for $r = 1$, the higher order methods give better solutions. The dramatic change between the 2nd and 4th order methods is clearly visible.

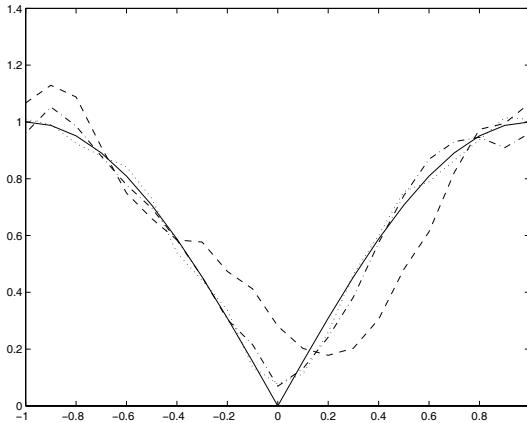


Fig. 1.2 $u(x, 6)$, $r = 1$ ($-$), $u^{(p)}$, $p = 2$ ($--$), $p = 4$ ($-\cdot$), $p = 6$ ($\cdot\cdot$), $N = 20$

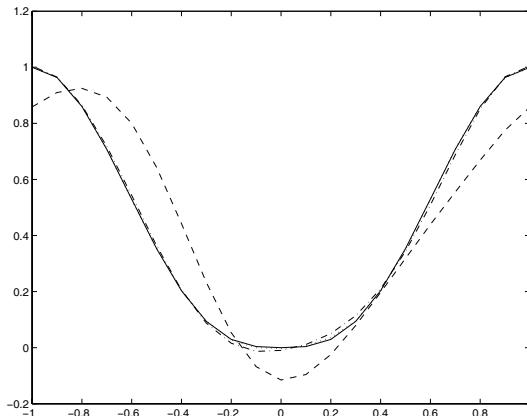


Fig. 1.3 $u(x, 6)$, $r = 3$ ($-$), $u^{(p)}$, $p = 2$ ($--$), $p = 4$ ($-\cdot$), $p = 6$ ($\cdot\cdot$), $N = 20$

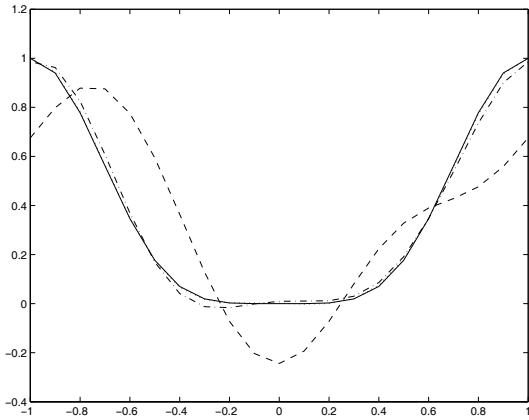


Fig. 1.4 $u(x, 6)$, $r = 5$ ($-$), $u^{(p)}$, $p = 2$ ($--$), $p = 4$ ($-\cdot$), $p = 6$ ($\cdot\cdot$), $N = 20$

The next three figures show the l_2 -error $\sum_j |v_j(t)|^2 h$ for $r = 1, 3, 5$ as a function of time for $N = 20$ and $N = 40$. For the case $r = 1$, the convergence rate is roughly linear ($\sim h$) for all three methods, but the error is significantly smaller for the higher order ones. For the smoother cases $r = 3$ and $r = 5$, the convergence rate goes up considerably as expected.

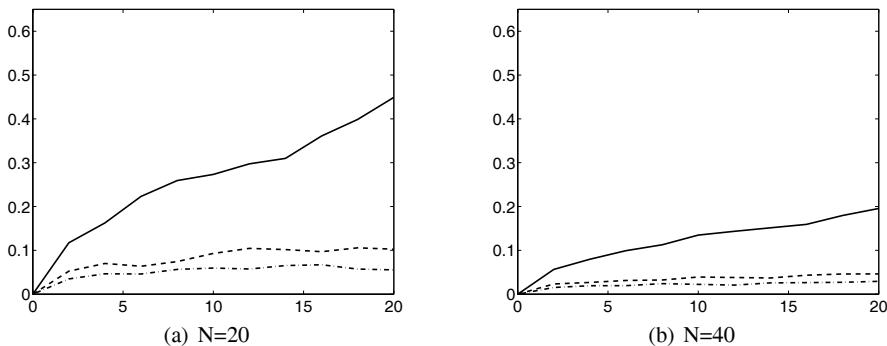


Fig. 1.5 l_2 -error, $r = 1$, $p = 2$ ($--$), $p = 4$ ($-\cdot$), $p = 6$ ($\cdot\cdot$)

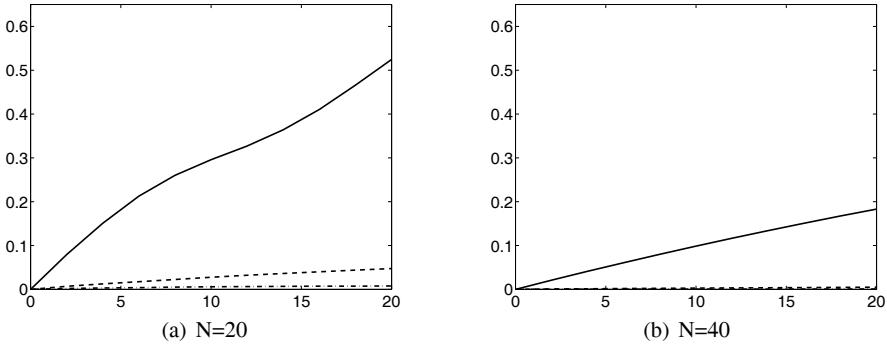


Fig. 1.6 l_2 -error, $r = 3$, $p = 2(--)$, $p = 4(-\cdot)$, $p = 6(\cdots)$

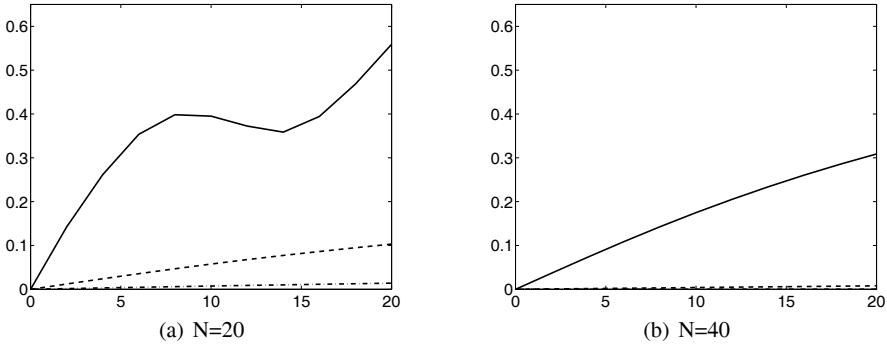


Fig. 1.7 l_2 -error, $r = 5$, $p = 2(--)$, $p = 4(-\cdot)$, $p = 6(\cdots)$

1.3 Parabolic Equations

For *parabolic* equations, the situation is a little different. Consider the heat equation in its simplest form and the model problem

$$\begin{aligned} u_t &= u_{xx}, \\ u(x, 0) &= e^{i\omega x} \end{aligned}$$

with the solution $u(x, t) = e^{-\omega^2 t + i\omega x}$. The standard second order approximation is

$$\begin{aligned} \frac{du_j}{dt} &= D_+ D_- u_j, \\ u_j(0) &= e^{i\omega x_j} \end{aligned}$$

with the solution $u_j(t) = e^{\hat{P}_2(\xi)t+i\omega x_j}$, where

$$\hat{P}_2(\xi) = -\frac{4}{h^2} \sin^2 \frac{\xi}{2}$$

is the Fourier transform of D_+D_- . Apparently, the accuracy is determined by the ability of $\hat{P}_2(\xi)$ to approximate $-\omega^2$.

The fourth and sixth order approximations are

$$\begin{aligned}\frac{du_j}{dt} &= D_+D_-(I - \frac{h^2}{12}D_+D_-)u_j, \\ \frac{du_j}{dt} &= D_+D_-(I - \frac{h^2}{12}D_+D_- + \frac{h^4}{90}(D_+D_-)^2)u_j\end{aligned}$$

with the solution $u_j(t) = e^{\hat{P}_p(\xi)t+i\omega x_j}$, $p = 4, 6$, where

$$\begin{aligned}\hat{P}_4(\xi) &= -\frac{4}{h^2} \sin^2 \frac{\xi}{2} \left(1 + \frac{1}{3} \sin^2 \frac{\xi}{2}\right). \\ \hat{P}_6(\xi) &= -\frac{4}{h^2} \sin^2 \frac{\xi}{2} \left(1 + \frac{1}{3} \sin^2 \frac{\xi}{2} + \frac{8}{45} \sin^4 \frac{\xi}{2}\right).\end{aligned}$$

Figure 1.8 shows a comparison between $\xi^2 = \omega^2 h^2$ and $-\hat{P}_p(\xi)h^2$, $p = 2, 4, 6$. All three approximations give a stronger damping with time than the true solution has.

In order to estimate the necessary number of grid points, we consider the error

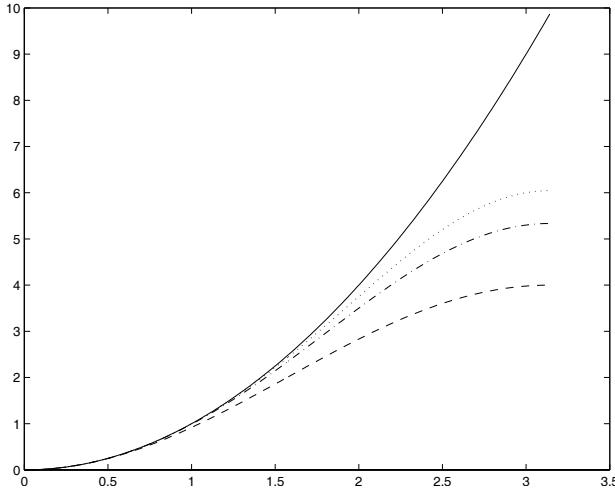


Fig. 1.8 $\xi^2(-)$, $-\hat{P}_p(\xi)h^2$, $p = 2(--)$, $4(-\cdot)$, $6(..)$

$$v^{(p)}(t) = \max_j |e^{i\omega x_j - \omega^2 t} - e^{i\omega x_j + \hat{P}_p(\xi)t}| = e^{-\omega^2 t} |1 - e^{(\omega^2 + \hat{P}_p(\xi))t}|, \quad p = 2, 4, 6.$$

For small $|\xi|$, we have by Taylor expansion

$$\hat{P}_2(\xi) \approx -\omega^2 \left(1 - \frac{\xi^2}{12} + \mathcal{O}(\xi^4)\right),$$

giving the approximative error

$$v^{(2)}(t) \approx \frac{1}{12} \omega^2 \xi^2 t e^{-\omega^2 t}. \quad (1.6)$$

In contrast to the hyperbolic case, the error is maximal at a point in time which is independent of the number of computed periods:

$$\max_t (v^{(2)}(t)) \approx \frac{\xi^2}{12e} \quad \text{for } t = \frac{1}{\omega^2}.$$

For the prescribed error level ε , the necessary number of points per wave length is

$$M_2 \approx \frac{2\pi}{(12e\varepsilon)^{1/2}}.$$

A similar calculation for the fourth and sixth order cases gives the number of points per wave length

$$M_4 \approx \frac{2\pi}{(90e\varepsilon)^{1/4}},$$

$$M_6 \approx \frac{2\pi}{(560e\varepsilon)^{1/6}}.$$

Table 1.2 shows M_p for the error levels $\varepsilon = 0.01, 0.001, 0.0001$.

Table 1.2 M_p for different error levels, parabolic equations

ε	M_2	M_4	M_6
0.01	11	5	4
0.001	35	9	6
0.0001	110	16	9

Since M_p is independent of the length of the time integration (except for the first short part), it takes stronger accuracy requirements to get a real advantage of higher order accurate methods for second order parabolic problems. It can be shown that this conclusion holds also for higher order parabolic equations, since they all have strong damping with time, which causes the error to peak after a short time.

1.4 Schrödinger Type Equations

The *Schrödinger equation* in its simplest form is

$$u_t = iu_{xx},$$

with complex solutions u . With the usual Fourier component as initial function, the solution is $u(x, t) = e^{i(\omega x - \omega^2 t)}$. The only difference from the parabolic case, is the extra constant i multiplying the space derivative. However, this difference is quite significant when it comes to choosing the best order of approximation, since we don't have any damping of the amplitudes any longer. The behavior is more like the hyperbolic case, with increasing error all the time until it reaches a level $\mathcal{O}(1)$.

The approximations of the Schrödinger equation are obtained precisely as for the parabolic model problem in the previous section, except for an extra factor i multiplying them. Therefore, we don't have to go through the whole analysis again, but rather substitute $e^{-\omega^2 t}$ by $e^{-i\omega^2 t}$. The error derived in (1.6) is now obtained as

$$v^{(2)}(t) \approx \frac{1}{12} \omega^2 \xi^2 t,$$

and similarly for the $v^{(4)}$ and $v^{(6)}$. Therefore, we have the same situation as for the hyperbolic case. The only difference is that the length of a period in time is now $2\pi/\omega^2$, but for q periods it leads to the same type of inequality as (1.3):

$$v^{(2)} \approx \frac{\omega^2 t \xi^2}{12} \leq \frac{\pi q \xi^2}{6} = \frac{2\pi^3 q}{3M_2^2}.$$

With a prescribed error level ε , we get in analogy with (1.4)

$$\begin{aligned} M_2 &\approx 2\pi \left(\frac{\pi q}{6\varepsilon} \right)^{1/2}, \\ M_4 &\approx 2\pi \left(\frac{\pi q}{45\varepsilon} \right)^{1/4}, \\ M_6 &\approx 2\pi \left(\frac{\pi q}{270\varepsilon} \right)^{1/6}. \end{aligned} \tag{1.7}$$

Compared to the parabolic case, there is now an extra factor q involved. The error will grow with increasing time intervals for integration, but the influence becomes weaker with increasing order of accuracy. For the 1 % error level, we get Table 1.3. The results are very similar to the hyperbolic case, and again the most dramatic advantage is obtained by going from second to fourth order. Note however, that one period in time is now $2\pi/\omega^2$, which is shorter than in the hyperbolic case. So if the total time interval for integration is $[0, T]$ for both cases, the comparison between lower and higher order methods comes out even more favorable for the higher order ones in the case of Schrödinger type equations.

Table 1.3 M_p for 1% error level

q	M_2	M_4	M_6
20	203	22	11
200	642	38	16

1.5 Summary

The first theoretical analysis regarding optimal order of accuracy for first order hyperbolic equations was done by Kreiss and Oliger 1972 [Kreiss and Oliger, 1972]. This type of analysis has been presented in this chapter, also for higher order differential equations. The first rule of thumb is that the advantage of high order methods is more pronounced for problems where small errors in the solution are required. Secondly, for real equations $\partial u / \partial t = a \partial^q u / \partial t^q$, a real and q odd, there is an extra advantage with high order methods for long time integrations. This extra advantage is there also for complex equations with $a = i$ and even q , giving the solutions a wave propagation character. For problems in several space dimensions, the advantage with high order methods is even more pronounced.

For parabolic problems with real a and even q , the advantage with higher order methods is less. The reason for this is that there is an inherent damping in the equation, which means that the errors are not allowed to accumulate with time as for hyperbolic problems. Even if the integration is carried out over a long time interval, the error behaves more like short time integration.

The analysis presented here is based on the behavior of the approximation when applied to a single wave with a fixed wave number ω . If, for a whole wave package, the highest wave number ω_0 that is of interest to us is determined a priori, then the guidelines derived in this chapter tell us what method should be used to obtain a certain accuracy for the whole solution. One could of course have more involved criteria, where for example less accuracy is required for higher wave numbers, and then the conclusions would be modified. One can also discuss in terms of group velocity as Trefethen did in [Trefethen, 1983], see also [Strikwerda, 1989].

The time discretization can of course also be included in the analysis, as was done in [Swartz and Wendroff, 1974] for hyperbolic problems. The results are in line with the ones summarized above. A more detailed comparison between different schemes for wave propagation problems was carried out by Zingg in [Zingg, 2000]. Further investigations are carried out in Chapters 6 and 9.

Chapter 2

Well-posedness and Stability

Stability is a fundamental concept for any type of PDE approximation. A stable approximation is such that small perturbations in the given data cause only small perturbations in the solutions. Furthermore, the solutions converge to the true solution of the PDE as the step size h tends to zero. The extra condition required for this property is that the PDE problem is well posed. In this chapter we shall present a survey of the basic theory for the well-posedness and stability. The theory can be divided into three different techniques: Fourier analysis for Cauchy and periodic problems, the energy method and Laplace analysis (also called normal mode analysis) for initial-boundary value problems. In order to emphasize the similarities between the continuous and discrete case, we treat the application of each technique to both the PDE and the finite difference approximations in the same section (the Laplace technique for PDE is omitted).

2.1 Well Posed Problems

We consider a general initial-boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= Pu + F, \quad 0 \leq t, \\ Bu &= g, \\ u &= f, \quad t = 0. \end{aligned} \tag{2.1}$$

Here P is a differential operator in space, and B is a boundary operator acting on the solution at the spacial boundary. (Throughout this book, we will refer to t as the time coordinate, and to the remaining independent variables as the space variables, even if the physical meaning may be different.) There are three types of data that are fed into the problem: F is a given forcing function, g is a boundary function and f is an initial function. (By “function” we mean here the more general concept “vector function”, i.e., we are considering systems of PDE.) A well posed problem

has a unique solution u , and there is an estimate

$$\|u\|_I \leq K(\|f\|_H + \|F\|_{III} + \|g\|_{IV}), \quad (2.2)$$

where K is a constant independent of the data. In general, there are four different norms involved, but $\|\cdot\|_I$ and $\|\cdot\|_H$ are often identical.

Let v be the solution of the perturbed problem

$$\begin{aligned} \frac{\partial v}{\partial t} &= Pv + F + \delta F, \quad 0 \leq t, \\ Bv &= g + \delta g, \\ v &= f + \delta f, \quad t = 0. \end{aligned} \quad (2.3)$$

Assuming that P and B are linear operators, we subtract (2.1) from (2.3), and obtain for the perturbation $w = v - u$ of the solution

$$\begin{aligned} \frac{\partial w}{\partial t} &= Pw + \delta F, \quad 0 \leq t, \\ Bw &= \delta g, \\ w &= \delta f, \quad t = 0. \end{aligned}$$

The estimate (2.2) can now be applied to w :

$$\|w\|_I \leq K(\|\delta f\|_H + \|\delta F\|_{III} + \|\delta g\|_{IV}).$$

Hence, if K has a moderate size, small perturbations δf , δF , δg in the data cause a small perturbation w in the solution.

As an example, consider a scalar problem in one space dimension and $0 \leq x \leq 1$. Then $u = u(x, t)$, $F = F(x, t)$, $g = g(t)$, $f = f(x)$, and we choose

$$\|u\|_I^2 = \|u(\cdot, t)\|^2 = \int_0^1 |u(x, t)|^2 dx.$$

If the boundary conditions are

$$\begin{aligned} u(0, t) &= g_0(t), \\ u(1, t) &= g_1(t), \end{aligned}$$

then a typical estimate has the form

$$\|u(\cdot, t)\|^2 \leq K \left(\|f(\cdot)\|^2 + \int_0^t \|F(\cdot, \tau)\|^2 d\tau + \int_0^t (|g_0(\tau)|^2 + |g_1(\tau)|^2) d\tau \right),$$

where K may depend on t , but not on f , F , g_0 , g_1 . We could of course reformulate this estimate as in (2.2), but it is more convenient to keep the squared norms.

If the domain in space doesn't have any boundaries, there is of course no boundary condition. However, for a difference approximation, there has to be boundaries

for computational reasons. A special, but frequent case, is that the solutions are periodic in space. In that case the computation is done in a bounded domain, with the requirement that the solution and all its derivatives are equal at the both ends of the interval. This is often used as a model problem, since Fourier analysis can be used for investigating stability.

For periodic problems, the solution can be written as a Fourier series, and the behavior of the coefficients is the key issue. Let $v(x) = [v^{(1)} \ v^{(2)} \ \dots \ v^{(m)}]^T$ be a 2π -periodic vector function. The following lemma connects the size of these coefficients with the L_2 -norm of $v(x)$, which is defined by

$$\|v(\cdot)\|^2 = \int_0^{2\pi} |v(x)|^2 dx, \quad \|v\|^2 = \sum_{\nu=1}^m |v^{(\nu)}|^2.$$

Lemma 2.1. (Parseval's relation) *Let $v(x)$ be represented by its Fourier series*

$$v(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{v}(\omega) e^{i\omega x}.$$

Then

$$\|v(\cdot)\|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{v}(\omega)|^2.$$

□

As an example, consider the heat equation in its simplest form

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x \leq 2\pi, \quad 0 \leq t, \\ u(x, 0) &= f(x). \end{aligned} \tag{2.4}$$

The solution can be written as a Fourier series with time dependent coefficients

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{u}(\omega, t) e^{i\omega x}, \tag{2.5}$$

and the coefficients satisfy

$$\begin{aligned} \hat{u}_t &= -\omega^2 \hat{u}, \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned}$$

where $\hat{f}(\omega)$ are the Fourier coefficients of the initial data. The solution is

$$\hat{u}(\omega, t) = e^{-\omega^2 t} \hat{f}(\omega),$$

and by Parseval's relation

$$\|u(\cdot, t)\|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{u}(\omega, t)|^2 \leq \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = \|f(\cdot)\|^2.$$

Here we have proven well-posedness by simply finding the explicit form of the solution.

Assume next, that we want to solve the heat equation backward

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x \leq 2\pi, \quad 0 \leq t \leq T, \\ u(x, T) &= \phi(x). \end{aligned}$$

By the variable transformation $\tau = T - t$, $v(x, \tau) = u(x, T - \tau) = u(x, \tau)$, we get

$$\begin{aligned} v_\tau &= -v_{xx}, \quad 0 \leq x \leq 2\pi, \quad 0 \leq \tau \leq T, \\ v(x, 0) &= \phi(x). \end{aligned}$$

By the same procedure as above, we obtain

$$\hat{v}(\omega, \tau) = e^{\omega^2 \tau} \hat{\phi}(\omega).$$

It is now impossible to obtain an estimate of the type

$$\|v(\cdot, \tau)\| \leq K \|\phi(\cdot)\|,$$

where K is a constant, since

$$\|v(\cdot, \tau)\|^2 = \sum_{\omega=-\infty}^{\infty} e^{2\omega^2 \tau} |\hat{\phi}(\omega)|^2.$$

This shows that, given a measured heat distribution at a given time T , it is in theory impossible to find the true heat distribution at an earlier time, except for very smooth ϕ with fast decaying Fourier coefficients. The problem is ill posed . In practice it means, that it is extremely difficult to get any reasonable accuracy at $t = 0$, since small errors in the measurements give rise to large errors in the solution. (It should be said that there are numerical methods for ill posed problems, but a discussion of those is outside the scope of this book.)

2.2 Periodic Problems and Fourier Analysis

In this section we shall discuss the so called *Cauchy problem* , i.e., the domain in space is the whole real line. However, for convenience we will assume that the solutions are 2π -periodic in space, i.e., $u(x, t) = u(x + 2\pi, t)$, such that we can deal with a finite interval $[0, 2\pi]$. When the concept Cauchy problem is used a few times in the text, it refers either to the periodic case or to the case where $u(x, t) = 0$ outside some finite interval in space.

We begin by considering the PDE problem before discretization.

2.2.1 The PDE Problem

Consider the general problem in one space dimension

$$\begin{aligned}\frac{\partial u}{\partial t} &= P(\partial/\partial x)u + F(x, t), \quad 0 \leq t, \\ u(x, 0) &= f(x),\end{aligned}\tag{2.6}$$

where $P(\partial/\partial x)$ is a linear differential operator, i.e.,

$$P(\partial/\partial x)(\alpha u + v) = \alpha P(\partial/\partial x)u + P(\partial/\partial x)v,$$

if α is a constant. Before defining well-posedness, we consider the example

$$u_t = Au_x, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix}.$$

The matrix A can be diagonalized, i.e., there is a *similarity transformation* such that

$$T^{-1}AT = \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix},$$

where

$$T = \begin{bmatrix} 1 & 1 \\ 1/2 & -1/2 \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}.$$

By the substitution $v = T^{-1}u$, $g = T^{-1}f$, we get the new system

$$\begin{aligned}v_t &= \Lambda v_x, \\ v(x, 0) &= g(x),\end{aligned}$$

with the solution

$$\begin{aligned}v^{(1)}(x, t) &= g^{(1)}(x + 2t), \\ v^{(2)}(x, t) &= g^{(2)}(x - 2t).\end{aligned}$$

The norm is defined by

$$\|v(\cdot, t)\|^2 = \int_0^{2\pi} |v(x, t)|^2 dx, \quad |v(x, t)|^2 = |v^{(1)}(x, t)|^2 + |v^{(2)}(x, t)|^2,$$

and we have the inequality

$$\|Tv\| = \left(\int_0^{2\pi} |Tv|^2 dx \right)^{1/2} \leq \left(\int_0^{2\pi} |T|^2 |v|^2 dx \right)^{1/2} = |T| \cdot \|v\|^2,$$

where $|T|$ is the matrix norm defined by

$$|T| = \max_{|\nu|=1} |Tv|.$$

By periodicity it follows that

$$\int_0^{2\pi} |g^{(\nu)}(x \pm 2t)|^2 dx = \int_0^{2\pi} |g^{(\nu)}(x)|^2 dx,$$

and we get

$$\|u(\cdot, t)\| \leq |T| \cdot \|v(\cdot, t)\| = |T| \cdot \|g(\cdot)\| = |T| \cdot \|T^{-1}f(\cdot)\| \leq |T| \cdot |T^{-1}| \cdot \|f(\cdot)\|.$$

Since the matrix A is not symmetric, the *condition number* $K = |T| \cdot |T^{-1}|$ is greater than 1, but the estimate

$$\|u(\cdot, t)\| \leq K \|f(\cdot)\|$$

is the best one we can get.

Next, consider the trivial example $u_t = \alpha u$, where α is a positive constant. Obviously the solution satisfies

$$\|u(\cdot, t)\| = e^{\alpha t} \|f(\cdot)\|.$$

These two examples indicate that the following definition of well-posedness is appropriate:

Definition 2.1. The problem (2.6) is well posed if for $F(x, t) = 0$ there is a unique solution satisfying

$$\|u(\cdot, t)\| \leq K e^{\alpha t} \|f(\cdot)\|, \quad (2.7)$$

where K and α are constants independent of $f(x)$. \square

If the forcing function F is nonzero, one can show that for a well posed problem, the estimate

$$\|u(\cdot, t)\| \leq K e^{\alpha t} \left(\|f(\cdot)\| + \int_0^t \|F(\cdot, \tau)\| d\tau \right) \quad (2.8)$$

holds. This is a useful estimate, and it means that the forcing function can be disregarded in the analysis. The norm $\|\cdot\|_{III}$ in (2.2) is defined by

$$\|F\|_{III} = \int_0^t \|F(\cdot, \tau)\| d\tau.$$

For the simple examples treated so far, the existence of solutions is a trivial matter, in fact we have constructed them. However, questions concerning existence of solutions in the general case is beyond the scope of this book. Uniqueness, on the other hand, follows immediately from the condition (2.7). Assume that there is another solution v of the problem (2.6). Then by linearity, the difference $w = u - v$ satisfies the initial value problem

$$\begin{aligned}\frac{\partial w}{\partial t} &= P(\partial/\partial x)w, \quad 0 \leq t, \\ w(x, 0) &= 0.\end{aligned}$$

The condition (2.7) then implies that $w = 0$, i.e., $v = u$.

Next we shall discuss how to verify that the estimate (2.7) holds. In the previous section we saw how the problem (2.4) was converted into a simple set of ordinary differential equations by the Fourier transform, i.e., after writing the solution as a Fourier series. The operator $\partial/\partial x^2$ becomes $-\omega^2$ acting on the Fourier coefficients \hat{u} . Let us now apply this technique to general problems in one space dimension. Consider the problem (2.6), where $P(\partial/\partial x)$ is a differential operator with *constant coefficients*. This means that it has the form

$$P(\partial/\partial x) = \sum_{\nu=0}^q A_\nu \frac{\partial^\nu}{\partial x^\nu},$$

where the matrices A_ν are independent of x and t . By writing the solution as a Fourier series of the form (2.5), the vector coefficients are obtained as

$$\hat{u}(\omega, t) = e^{\hat{P}(i\omega)t} \hat{f}(\omega),$$

where

$$\hat{P}(i\omega) = \sum_{\nu=0}^q A_\nu (i\omega)^\nu.$$

Note that if u is a vector with m components, i.e., there are m differential equations in (2.6), then $\hat{P}(i\omega)$ is an $m \times m$ matrix, and it is called the *symbol* or *Fourier transform* of $P(\partial/\partial x)$.

By Parseval's relation, we get

Theorem 2.1. *The problem (2.6) is well posed if and only if there are constants K and α such that for all ω*

$$|e^{\hat{P}(i\omega)t}| \leq Ke^{\alpha t}. \quad (2.9)$$

□

It is often easier to study the eigenvalues of a matrix rather than the norm. We have

Definition 2.2. The *Petrovski condition* is satisfied if the eigenvalues $\lambda(\omega)$ of $\hat{P}(i\omega)$ satisfy the inequality

$$\operatorname{Re}(\lambda(\omega)) \leq \alpha, \quad (2.10)$$

where α is a constant independent of ω .

□

Clearly this condition is necessary for stability. There are many ways of prescribing extra conditions such that it is also sufficient for well-posedness. One such condition is given by

Theorem 2.2. *The Petrovski condition is necessary for well-posedness. It is sufficient if there is a constant K and a matrix $T(\omega)$ such that $T^{-1}(\omega)\hat{P}(i\omega)T(\omega)$ is diagonal and $|T^{-1}(\omega)| \cdot |T(\omega)| \leq K$ for all ω .* □

Problems in several space dimensions are treated in the same way. By defining the vectors

$$\mathbf{x} = [x^{(1)} \ x^{(2)} \ \dots \ x^{(d)}]^T, \\ \boldsymbol{\omega} = [\omega^{(1)} \ \omega^{(2)} \ \dots \ \omega^{(d)}]^T,$$

the symbol $\hat{P}(i\boldsymbol{\omega})$ is well defined by the formal transition $\partial/\partial x^{(\nu)} \rightarrow i\omega^{(\nu)}$. For example, when using the more common notation $x = x^{(1)}$, $y = x^{(2)}$, the differential operator

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial}{\partial x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial}{\partial y}$$

has the symbol

$$\hat{P}(i\boldsymbol{\omega}) = i \begin{bmatrix} \omega^{(1)} & \omega^{(2)} \\ \omega^{(2)} & \omega^{(1)} \end{bmatrix}.$$

The two theorems above now hold exactly as stated with the generalized definition of $\omega \rightarrow \boldsymbol{\omega}$.

The symbol in our example has purely imaginary eigenvalues, which implies that the Petrovski condition is satisfied with $\alpha = 0$. Furthermore, since $\hat{P}(i\boldsymbol{\omega})$ is skew-Hermitian, i.e., $(\hat{P}^*(i\boldsymbol{\omega}) = -\hat{P}(i\boldsymbol{\omega}))$, it can be diagonalized by a unitary matrix. This implies that the conditions of Theorem 2.2 are satisfied (with $K = 1$ and $\alpha = 0$), which makes the Petrovski condition sufficient for well-posedness.

General first order systems have the form

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^d A_\nu \frac{\partial u}{\partial x^{(\nu)}}, \quad (2.11)$$

and they are quite common in applications. They are called *hyperbolic* if the symbol

$$\hat{P}(i\boldsymbol{\omega}) = i \sum_{\nu=1}^d A_\nu \omega^{(\nu)}$$

has real eigenvalues and can be diagonalized by a matrix $T(\boldsymbol{\omega})$ with bounded condition number. Obviously, the Petrovski condition is satisfied for such systems.

If the PDE system doesn't have constant coefficients A_ν , then the Fourier analysis cannot be applied in a straightforward way as was done above. If $A_\nu = A_\nu(x)$, and the Fourier series is formally inserted, we get equations where the coefficients $\hat{u}(\boldsymbol{\omega})$ depend also on x , and it doesn't lead anywhere. The analysis can still be based on Fourier technique, but the theory becomes much more involved (see [Hörmander, 1985]), and we don't discuss it further here. (For difference approximations we shall briefly indicate what can be done.)

2.2.2 Difference Approximations

The discretization in time is done on a uniform grid $t_n = nk$, $n = 0, 1, \dots$, where k is the step size. Consider first the classic simple approximation of (2.4)

$$\begin{aligned} u_j^{n+1} &= Qu_j^n, \quad j = 0, 1, \dots, N, \\ u_j^0 &= f_j, \quad j = 0, 1, \dots, N, \end{aligned} \tag{2.12}$$

where $Q = I + kD_+ D_-$. The solution can be expanded in a finite Fourier series

$$u_j^n = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{u}_\omega^n e^{i\omega x_j},$$

where, for convenience, it is assumed that N is even. The coefficients are obtained by

$$\hat{u}_\omega^n = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^N u_j^n e^{-i\omega x_j} h,$$

which is called the *Discrete Fourier Transform* (DFT), often called the *Fast Fourier Transform* (FFT), which refers to the fast algorithm for computing it. The Fourier series is plugged into (2.12), and since the grid functions $\{e^{i\omega x_j}\}_{\omega=-N/2}^{N/2}$ are linearly independent, we obtain

$$\hat{u}_\omega^{n+1} e^{i\omega x_j} = Q \hat{u}_\omega^n e^{i\omega x_j} = (I + kD_+ D_-) \hat{u}_\omega^n e^{i\omega x_j} = (1 - 4\lambda \sin^2 \frac{\xi}{2}) \hat{u}_\omega^n e^{i\omega x_j},$$

where $\xi = \omega h$, $\lambda = k/h^2$. The function $\hat{Q}(\xi) = 1 - 4\lambda \sin^2 \frac{\xi}{2}$ is called the (discrete) Fourier transform of the difference operator Q , and we have

$$\hat{u}_\omega^{n+1} = \hat{Q} \hat{u}_\omega^n. \tag{2.13}$$

Instead of having a difference operator acting on the whole grid function u_j^n , we have obtained a very simple scalar equation for each Fourier component. The obvious condition for nongrowing solutions is

$$|\hat{Q}(\xi)| \leq 1, \quad |\xi| \leq \pi, \tag{2.14}$$

and it is satisfied if and only if $\lambda \leq \frac{1}{2}$.

Going back to the physical space, we introduce the discrete norm

$$\|u^n\|_h^2 = \sum_{j=0}^N |u_j^n|^2 h. \tag{2.15}$$

In analogy with Lemma 2.1 we have

Lemma 2.2. (The discrete Parseval's relation) Let v_j be represented by its Fourier series

$$v_j = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{v}_\omega e^{i\omega x_j}.$$

Then

$$\|v\|_h^2 = \sum_{\omega=-N/2}^{N/2} |\hat{v}_\omega|^2.$$

□

By using the discrete Parseval's relation it follows from (2.14) that

$$\|u^{n+1}\|_h^2 = \sum_{\omega=-N/2}^{N/2} |\hat{u}_\omega^{n+1}|^2 = \sum_{\omega=-N/2}^{N/2} |\hat{Q}\hat{u}_\omega^n|^2 \leq \sum_{\omega=-N/2}^{N/2} |\hat{u}_\omega^n|^2 = \|u^n\|_h^2,$$

and by repeating this inequality for decreasing n , we obtain the final estimate

$$\|u^n\|_h \leq \|f\|_h$$

in analogy with the continuous case.

Let us next consider the same problem, but with a lower order term added:

$$u_t = u_{xx} + \alpha u, \quad \alpha > 0.$$

The difference scheme is (2.12), but now with $Q = I + kD_+D_- + \alpha kI$. By doing the same analysis as above, we arrive at

$$\hat{u}^{n+1} = (1 - 4\lambda \sin^2 \frac{\xi}{2} + \alpha k)\hat{u}^n, \quad |\xi| \leq \pi,$$

and the best estimate we can obtain for all ξ is

$$|\hat{u}^{n+1}| \leq (1 + \alpha k)|\hat{u}^n|.$$

This leads to

$$\|u^n\|_h^2 \leq (1 + \alpha k)^{2n} \|f\|_h^2 \leq e^{2\alpha nk} \|f\|_h^2 = e^{2\alpha t_n} \|f\|_h^2.$$

Referring back to the discussion of well-posedness above, this growth corresponds to the growth of the solution to the differential equation itself for the special case $f = \text{const.}$, i.e., $\hat{f}(\omega) = 0$ for $\omega \neq 0$. The solution to the differential equation is $u = e^{\alpha t}$ if $f \equiv 1$. Hence, we cannot expect any better estimate.

In order to include such lower order terms in the class of problems we want to solve, it is reasonable to generalize the stability definition to

$$\|u^n\|_h \leq e^{\alpha t_n} \|f\|_h.$$

Note the important difference between $e^{\alpha t_n}$ and $e^{\alpha n}$. In the first case the solution is bounded by $e^{\alpha T} \|f\|_h$ for $0 \leq t_n \leq T$, while in the second case there is no bound for any fixed $\bar{t} = t_n$ as $k \rightarrow 0$.

Next we will generalize to systems of PDE, and consider

$$\begin{aligned} u_t &= Au_{xx}, \\ u(x, 0) &= f(x), \end{aligned}$$

where u is a vector function with m components $u^{(\nu)}$, and A is a constant $m \times m$ matrix with real and positive eigenvalues a_j . The difference scheme becomes

$$u_j^{n+1} = (I + kAD_+D_-)u_j^n,$$

with the Fourier transform

$$\hat{u}_\omega^{n+1} = \hat{Q}\hat{u}_\omega^n,$$

where $\hat{Q} = I - 4\lambda A \sin^2(\xi/2)$. The Fourier coefficients and Parseval's relation are well defined as above for vectors u_j^n and \hat{u}_ω^n . Assuming that A can be diagonalized such that

$$T^{-1}AT = \tilde{A} = \text{diag}(a_1 \ a_2 \ \dots \ a_m),$$

we introduce the new vector $\hat{v}_\omega = T^{-1}\hat{u}_\omega$, and get

$$\hat{v}_\omega^{n+1} = (I - 4\lambda \tilde{A} \sin^2 \frac{\xi}{2})\hat{v}_\omega^n.$$

This is an uncoupled system, and by requiring nongrowing solutions for each component, the condition becomes

$$\lambda a \leq \frac{1}{2},$$

where $a = \max_j \{a_j\}$. We obtain $|\hat{v}_\omega^n| \leq |\hat{v}_\omega^0|$, and furthermore

$$|\hat{u}_\omega^n| = |T\hat{v}_\omega^n| \leq |T| \cdot |\hat{v}_\omega^n| \leq |T| \cdot |\hat{v}_\omega^0| = |T| \cdot |T^{-1}\hat{u}_\omega^0| \leq K|f_\omega|,$$

where K is the condition number of T . The final estimate becomes

$$\|u^n\|_h^2 = \sum_{\omega=-N/2}^{N/2} |\hat{u}_\omega^n|^2 = \sum_{\omega=-N/2}^{N/2} |\hat{Q}\hat{u}_\omega^{n-1}|^2 \leq K^2 \sum_{\omega=-N/2}^{N/2} |\hat{f}_\omega|^2 = K^2 \|f\|_h^2.$$

When taking this example into account, as well as the example above with a lower order term in the differential equation, it is reasonable to make the following definition of stability for general one step schemes:

Definition 2.3. A difference approximation (2.12) with a general difference operator Q is *stable* if the solution satisfies

$$\|u^n\|_h \leq K e^{\alpha t_n} \|f\|_h, \quad (2.16)$$

where the constants K and α do not depend on the initial data f . \square

In the last example, the Fourier transform is a matrix. That may be the case even if the underlying PDE is a scalar equation. Consider the first order wave equation

$$u_t = u_x$$

and the *leap-frog scheme*

$$u_j^{n+1} = u_j^{n-1} + 2kD_0 u_j^n.$$

After inserting the Fourier series for u_j^n , we get

$$\hat{u}_\omega^{n+1} = \hat{u}_\omega^{n-1} + 2\lambda i(\sin \xi) \hat{u}_\omega^n, \quad \lambda = k/h.$$

By introducing the vector $\hat{\mathbf{u}}_\omega^n = [\hat{u}_\omega^{n+1} \ \hat{u}_\omega^n]^T$, the difference equation can be written in the form

$$\hat{\mathbf{u}}_\omega^{n+1} = \hat{Q} \hat{\mathbf{u}}_\omega^n, \quad (2.17)$$

where

$$\hat{Q} = \begin{bmatrix} 2\lambda i \sin \xi & 1 \\ 1 & 0 \end{bmatrix}.$$

If $\lambda < 1$, this matrix can be diagonalized, and it has the eigenvalues

$$z_{1,2} = \lambda i \sin \xi \pm \sqrt{1 - \lambda^2 \sin^2 \xi}$$

with $|z_1| = |z_2| = 1$. Therefore we have, just as for the parabolic case above,

$$|\hat{\mathbf{u}}_\omega^n| \leq K |\hat{\mathbf{u}}_\omega^0|,$$

where K is the condition number of the matrix diagonalizing \hat{Q} . The vector corresponding to \hat{u}_ω^n is in physical space is $\mathbf{u}^n = [u^{n+1} \ u^n]^T$, and we have

$$\|\mathbf{u}^n\|_h \leq K \|\mathbf{f}\|_h,$$

where $\mathbf{f} = [u^1 \ u^0]^T$. The final estimate is

$$\|u^{n+1}\|_h^2 + \|u^n\|_h^2 \leq K^2 (\|u^1\|_h^2 + \|u^0\|_h^2).$$

Note that the reformulation to one step form is done only for the purpose of analysis. The eigenvalues z_1, z_2 could as well be obtained as the roots of the *characteristic equation*

$$z^2 - 2\lambda i(\sin \xi)z - 1 = 0.$$

This equation is formally obtained by setting $\hat{u}_\omega^n = z^n$ and dividing by z^{n-1} . Note also, that we could as well have derived the matrix form of \hat{Q} by first reformulating the original leap-frog scheme as a one step scheme

$$\mathbf{u}_j^{n+1} = \begin{bmatrix} 2kD_0 & I \\ I & 0 \end{bmatrix} \mathbf{u}_j^n,$$

which has the Fourier transform (2.17).

Consider next a general multistep scheme

$$\begin{aligned} Q_{-1} u_j^{n+1} &= \sum_{\sigma=0}^q Q_\sigma u_j^{n-\sigma} + kF_j^n, \\ u_j^\sigma &= f_j^\sigma, \quad \sigma = 0, 1, \dots, q, \end{aligned}$$

where

$$Q_\sigma = \sum_{\nu=-r}^p A_\nu^{(\sigma)} E^\nu.$$

We define the vector $\mathbf{u}^n = [u^{n+q} \ u^{n+q-1} \ \dots \ u^n]^T$ and similarly for \mathbf{F}^n and \mathbf{f} . Then we get the one step scheme

$$\begin{aligned} \mathbf{u}_j^{n+1} &= Q\mathbf{u}_j^n + \mathbf{F}_j^n, \\ \mathbf{u}_j^0 &= \mathbf{f}_j, \end{aligned} \tag{2.18}$$

where

$$Q = \begin{bmatrix} Q_{-1}^{-1} Q_0 & Q_{-1}^{-1} Q_1 & \dots & \dots & Q_{-1}^{-1} Q_q \\ I & 0 & \dots & \dots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & I & 0 \end{bmatrix}.$$

The vector \mathbf{u}_j has $m(q+1)$ components, and Q is an $m(q+1) \times m(q+1)$ matrix as well as the Fourier transform $\hat{Q}(\xi)$ (also called the *symbol* or the *amplification matrix*). For convenience we use from now on the notation u_j^n even if it is a vector.

The stability condition derived above is a condition on the eigenvalues of the symbol. In analogy with the Petrovski condition we have

Definition 2.4. The *von Neumann condition* is satisfied if the eigenvalues $z(\xi)$ of the symbol $\hat{Q}(\xi)$ satisfy

$$|z(\xi)| \leq e^{\alpha k}, \quad |\xi| \leq \pi. \tag{2.19}$$

□

If the symbol can be expressed in terms of $\xi = \omega h$ without any explicit k -dependence, the von Neumann condition becomes

$$|z(\xi)| \leq 1, \quad |\xi| \leq \pi. \tag{2.20}$$

This condition was used in the examples above, since $z(\xi)$ was expressed in terms of $\lambda = k/h^2$ in the parabolic case, and in terms of $\lambda = k/h$ in the hyperbolic case.

We have shown for our examples, that the von Neumann condition is sufficient for stability. In general this is not the case. For example, if \hat{Q} cannot be diagonalized,

the matrix T does not exist, and we don't have a well defined condition number K . The leap-frog scheme with $\lambda = 1$ is such a case. On the other hand, it is easy to see that the condition is necessary. If it is not satisfied, we just pick an eigenvector of \hat{Q} as an initial vector, and get an unbounded solution. We formulate the result as a theorem.

Theorem 2.3. *Let $\hat{Q}(\xi)$ be the symbol of the difference approximation*

$$u_j^{n+1} = Qu_j^n.$$

Then the von Neumann condition is necessary for stability. If $\hat{Q}(\xi)$ can be diagonalized by a similarity transformation $T^{-1}\hat{Q}T$ with $|T| \cdot |T^{-1}| \leq K$, where K is a constant, then the condition is also sufficient for stability. \square

There is an abundance of stability theory developed with the von Neumann condition as a basis. One important extension is due to Kreiss, who introduced the concept of dissipativity (of order $2r$), which means that the eigenvalues satisfy an inequality

$$|z(\xi)| \leq 1 - \delta|\xi|^{2r}, \quad \delta > 0, \quad |\xi| \leq \pi. \quad (2.21)$$

For parabolic equations, this is a natural property, since the differential equation itself is damping all amplitudes corresponding to nonzero frequencies. For hyperbolic equations it has to be introduced artificially either by explicitly adding extra terms, or as an inherent consequence of the structure of the approximation.

Real life problems do not in general have constant coefficients. If the coefficients depend on x , such that

$$u_j^{n+1} = Q(x_j)u_j^n,$$

then, just as for the continuous case, the Fourier analysis cannot be applied directly. However, we can freeze the coefficients by letting $x_j \rightarrow \bar{x}$, and consider the Fourier transform of the constant coefficient equation

$$u_j^{n+1} = Q(\bar{x})u_j^n$$

for each possible \bar{x} . The symbol is $\hat{Q} = \hat{Q}(\bar{x}, \xi)$, with eigenvalues $z(\bar{x}, \xi)$, and the von Neumann condition is well defined. The stability analysis now requires deeper theory, but in [Kreiss, 1964] it is proven, that under certain extra conditions, the scheme is stable for the variable coefficient case if the dissipativity property (2.21) holds pointwise for every \bar{x} in the computational domain. Therefore, the Fourier analysis is a significant tool also for realistic application problems.

We shall next consider two types of additional terms in the difference scheme, and show that they can be disregarded in the stability analysis. A PDE of the form $u_t = Pu + F$ introduces an extra forcing function also in the difference approximation. For convenience we drop the space index j and let

$$u^{n+1} = Qu^n \quad (2.22)$$

be an approximation of the homogeneous equation, with Q possibly obtained via a transformation of a multistep scheme as demonstrated above. The scheme approximating the PDE with a forcing function has the form

$$v^{n+1} = Qv^n + kF^n. \quad (2.23)$$

If the original approximation is a one step scheme, the function F^n is usually the same as F in the PDE. In the multistep case, F^n is a vector with zero elements in all positions except the first one. Note that the factor k is naturally occurring, since the scheme is normalized such that it approximates $ku_t = kPu + kF$. For any number of time steps, the solution of (2.22) satisfies

$$u^n = Q^{n-\nu} u^\nu,$$

where it is assumed that Q is independent of t_n . Obviously, stability implies

$$\|Q^{n-\nu}\|_h \leq Ke^{\alpha(t_n-t_\nu)},$$

where the norm of the difference operator is defined by

$$\|Q\|_h = \max_{\|v\|_h=1} \|Qv\|_h.$$

Consider now the solutions of (2.23). It is easily shown that

$$v^n = Q^n v^0 + k \sum_{\nu=0}^{n-1} Q^{n-\nu-1} F^\nu,$$

and we get

$$\begin{aligned} \|v^n\|_h &\leq \|Q^n\|_h \|v^0\|_h + k \sum_{\nu=0}^{n-1} \|Q^{n-\nu-1}\|_h \|F^\nu\|_h \\ &\leq Ke^{\alpha t_n} \|v^0\|_h + \left(\max_{0 \leq \nu \leq n-1} \|F^\nu\|_h \right) k \sum_{\nu=0}^{n-1} Ke^{\alpha(t_{n-1}-t_\nu)}. \end{aligned}$$

But

$$k \sum_{\nu=0}^{n-1} e^{\alpha(t_{n-1}-t_\nu)} = k \sum_{\nu=0}^{n-1} e^{\alpha(n-\nu-1)k} = k \frac{e^{\alpha nk} - 1}{e^{\alpha k} - 1} \leq \frac{e^{\alpha t_n} - 1}{\alpha}$$

if $\alpha \neq 0$. (In the last inequality we have used a Taylor expansion of $e^{\alpha k} - 1$.) If $\alpha = 0$, then the last expression is substituted by t_n (limit of $(e^{\alpha t_n} - 1)/\alpha$ as $\alpha \rightarrow 0$). The final estimate therefore becomes

$$\|v^n\|_h \leq Ke^{\alpha t_n} \|v^0\|_h + K_1(t_n) \max_{0 \leq \nu \leq n-1} \|F^\nu\|_h, \quad (2.24)$$

where

$$K_1(t) = \begin{cases} \frac{e^{\alpha t} - 1}{\alpha} & \text{if } \alpha \neq 0, \\ t & \text{if } \alpha = 0. \end{cases}$$

The conclusion is that if there is a forcing function in the equation, it can be disregarded in the stability analysis.

It was shown earlier for an example, that terms which are proportional to ku can be added without affecting the stability. We formulate this as a theorem for the general case:

Theorem 2.4. *Assume that (2.22) is stable. Then the perturbed problem*

$$v^{n+1} = Qv^n + kRv^n$$

is also stable if R is a bounded operator, i.e., $\|R\|_h \leq K$, where K is independent of h, k . \square

The proof uses the same technique as was applied above, by considering Rv^n as a forcing function. We omit the details here.

It should be said that even if the scheme is stable, an exponential growth with time may be bad in practice. It is particularly bad if the lower order term is damping the true solution. For example, if the PDE is

$$u_t = Pu - \alpha u, \quad \alpha > 0,$$

with solutions $\sim e^{-\alpha t}$, then a scheme with solutions behaving like $\sim e^{\alpha t}$ is of little practical use.

We finish this section by stating the von Neumann condition for semidiscrete approximations like the ones analyzed for accuracy in Chapter 1. They have the general form

$$\begin{aligned} \frac{du_j}{dt} &= Qu_j + F_j(t), \\ u_j(0) &= f_j, \end{aligned} \tag{2.25}$$

where Q is a difference operator in space with constant coefficients. The largest elements are of order h^{-q} if the differential operator in space of the PDE is of order q . Then we have corresponding to Definition 2.4 (and corresponding to the Petrovski condition for PDE)

Definition 2.5. Consider the semidiscrete approximation (2.25). The *von Neumann condition* is satisfied if the eigenvalues $z(\xi, h)$ of the symbol $\hat{Q}(\xi, h)$ satisfy

$$\operatorname{Re}(z(\xi, h)) \leq \alpha, \quad |\xi| \leq \pi, \tag{2.26}$$

where α is a constant independent of ξ and h . \square

If there are no terms of order one in Q , then the von Neumann condition has the form

$$\operatorname{Re}(z(\xi, h)) \leq 0 \quad |\xi| \leq \pi.$$

Obviously the von Neumann condition is necessary for stability also in the semidiscrete case.

2.3 Initial–Boundary Value Problems and the Energy Method

In this section we consider the general initial–boundary value problem (2.1). For general boundary conditions, the Fourier technique can no longer be applied, since the solutions obey boundary conditions that are not satisfied by the Fourier components. We saw in the previous section, that even for periodic problems, the Fourier technique cannot be directly implied if the coefficients are x -dependent. Both of these difficulties are overcome by the so called *energy method*, which we shall describe here.

2.3.1 The PDE Problem

Consider first the model problem

$$\begin{aligned} u_t &= (au_x)_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= 0, \\ u_x(1, t) &= 0, \\ u(x, 0) &= f(x), \end{aligned}$$

where $a = a(x, t) \geq \delta > 0$. Define the scalar product and norm for real functions v and w by

$$(v(\cdot), w(\cdot)) = \int_0^1 v(x)w(x)dx, \quad \|v(\cdot)\|^2 = (v(\cdot), v(\cdot)).$$

We get, when using the boundary conditions,

$$\begin{aligned} \frac{d}{dt}\|u\|^2 &= 2(u, u_t) = 2(u, (au_x)_x) = 2u(1, t)a(1, t)u_x(1, t) - 2u(0, t)a(0, t)u_x(0, t) \\ &\quad - 2(u_x, au_x) = -2(u_x, au_x) \leq -2\delta\|u_x\|^2 \leq 0. \end{aligned}$$

When integrating this differential inequality, we obtain

$$\|u(\cdot, t)\|^2 \leq \|f(\cdot)\|^2,$$

i.e., the problem is well posed. Actually, we have the stronger estimate

$$\|u(\cdot, t)\|^2 + 2\delta \int_0^t \|u_x(\cdot, \tau)\|^2 d\tau \leq \|f(\cdot)\|^2,$$

showing that as long as the solution is not constant, there is a decrease of the norm $\|u\|$. This is typical for parabolic equations, describing for example heat conduction or diffusion.

This problem is an example of a *semibounded* operator. The differential operator $P = (\partial/\partial x) a\partial/\partial x$ applied on differentiable functions v satisfying the boundary conditions, satisfies $(v, Pv) \leq 0$.

Consider now the general initial-boundary value problem (2.1). The scalar product and norm are generalized to

$$(v(\cdot), w(\cdot)) = \int_0^1 q(x) \langle v(x), Hw(x) \rangle dx, \quad ||v(\cdot)||^2 = (v(\cdot), v(\cdot)),$$

where

$$\langle v, w \rangle = \sum_{\nu=1}^m \overline{v^{(\nu)}} w^{(\nu)}, \quad |v|^2 = \langle v, v \rangle.$$

Here $q(x)$ is a positive function and H is a positive definite Hermitian matrix. If not stated otherwise, we assume that $q(x) = 1$ and $H = I$ in the examples below. Semiboundedness is defined by

Definition 2.6. Let \mathcal{V} be the space of differentiable functions satisfying the boundary conditions $Bv = 0$. The differential operator P is *semibounded* if for all $v \in \mathcal{V}$ the inequality

$$(v, Pv) \leq \alpha ||v||^2$$

holds, where α is a constant independent of v . \square

(In the complex case, (v, Pv) is substituted by $Re(v, Pv)$.) If a solution exists, semibounded operators guarantee well-posedness, since

$$\frac{d}{dt} ||u||^2 = 2(u, u_t) = 2(u, Pu) \leq 2\alpha ||u||^2,$$

which after integration leads to

$$||u(\cdot, t)|| \leq e^{\alpha t} ||f(\cdot)||.$$

Before proceeding, let us first make a general remark. The integration by parts procedure, that is the core of the energy method, produces certain boundary terms. If they have the right sign, we are in good shape, and we get a semibounded operator with $\alpha = 0$. If they are not, there is no chance to estimate them in terms of $||u||^2$, and in that way obtain a semibounded operator, but with an $\alpha > 0$. As a counterexample, consider the function $f(x) = x^{-1/4}$ on the interval $[-1, 1]$. The L_2 -norm is

$$\left(\int_{-1}^1 |f(x)|^2 dx \right)^{1/2} = \left(\int_{-1}^1 \frac{1}{\sqrt{|x|}} dx \right)^{1/2} = 2,$$

while $f(x) \rightarrow \infty$ as $x \rightarrow 0$. This shows that $f(x)$ at $x = 0$ cannot be estimated in terms of its L_2 -norm.

Existence of solutions for initial-boundary value problems is in general a more difficult issue compared to Cauchy or periodic problems. Consider the example

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= 0, \\ u(1, t) &= 0, \\ u(x, 0) &= f(x). \end{aligned}$$

Obviously we have a semibounded operator, since

$$(v, Pv) = \frac{1}{2}(|v(1)|^2 - |v(0)|^2) = 0$$

for all v satisfying the boundary conditions. However, the solution u is uniquely defined without the boundary condition at $x = 0$ as

$$u(x, t) = \begin{cases} f(x+t) & \text{if } x+t \leq 1, \\ 0 & \text{if } x+t > 1. \end{cases} \quad (2.27)$$

In particular, $u(0, t) = f(t)$ for $t \leq 1$, and if f is not identically zero, the boundary condition $u(0, t) = 0$ is contradicted.

This example shows that there is a need for restricting the semiboundedness concept in order to guarantee the existence of a solution:

Definition 2.7. The differential operator P is *maximally semibounded* if it is semibounded in the function space \mathcal{V} , but not semibounded in any space with fewer boundary conditions. \square

In the example above, the function space \mathcal{V} is too small for allowing the existence of a solution, and it must be made bigger by dropping the boundary condition at the left such that

$$\mathcal{V} = \{v(x), v(1) = 0\}.$$

Then P is still semibounded, since

$$(v, Pv) = -\frac{|v(0)|^2}{2} \leq 0.$$

If also the boundary condition at the right is dropped, we get

$$(v, Pv) = \frac{1}{2}(|v(1)|^2 - |v(0)|^2),$$

which cannot be bounded in terms of $\|v\|^2$.

When discussing existence of solutions, there is still another difficulty arising from the presence of boundaries. If in the example above, $f(1) \neq 0$, then there is an incompatibility at the point $\{x = 1, t = 0\}$, since $g(0) = 0$. Still it makes sense to define the solution as given by (2.27), since it is well defined everywhere. However, there is a discontinuity at $x + t = 1$, and there is trouble giving the derivative a meaningful definition there. Therefore $u(x, t)$ is called a *generalized solution*, and it can be defined as a limit of smooth solutions $u_\nu(x, t)$ resulting from smooth initial

functions $f_\nu(x)$ satisfying $f_\nu(1) = 0$ for every fixed ν . (This concept of generalized solutions is introduced already for the pure initial value problem, if the initial function $f(x)$ is not smooth.) In our definitions of well-posedness we shall always assume that the solution is smooth, relying on the fact that it can be extended to the case where one or more of the functions F, g, f in (2.1) are nonsmooth, and/or the boundary and initial data are incompatible.

Maximal semiboundedness implies well-posedness, and we make the formal definition

Definition 2.8. The initial–boundary value problem (2.1) is *well posed* if for $F = 0, g = 0$ there is a unique solution satisfying

$$\|u(\cdot, t)\| \leq K e^{\alpha t} \|f(\cdot)\|,$$

where K and α are constants independent of $f(x)$. \square

If the forcing function F is nonzero, one can show that an estimate involving $\|F(\cdot, t)\|$ can be obtained, just as for the pure initial value problem, see (2.8).

Let us next consider the example

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(1, t) &= g(t), \\ u(x, 0) &= f(x), \end{aligned}$$

where $g(t)$ is nonzero. We have

$$\frac{d}{dt} \|u\|^2 = 2(u, u_x) = |g(t)|^2 - |u(0, t)|^2,$$

which leads to

$$\|u(\cdot, t)\|^2 + \int_0^t |u(0, \tau)|^2 d\tau = \|f(\cdot)\|^2 + \int_0^t |g(\tau)|^2 d\tau.$$

Here is a case where we get an estimate of the solution in terms of the initial function *and* the boundary function. However, there are problems for which this is not possible, and therefore we make a special definition for the stronger estimate:

Definition 2.9. The initial–boundary value problem (2.1) is *strongly well posed* if there is a unique solution satisfying

$$\|u(\cdot, t)\|^2 \leq K e^{\alpha t} \left(\|f(\cdot)\|^2 + \int_0^t (\|F(\cdot, \tau)\|^2 + |g(\tau)|^2) d\tau \right),$$

\square

where K and α are constants independent of $f(x)$, $F(x, t)$ and $g(t)$. (Note that the estimate for the example above is even stronger, since the integral of the solution at the boundary is included on the left hand side of the inequality as well.) The number

of components g_ν in the vector g depends on the number of boundary conditions. For example, if $P(\partial/\partial x) = A\partial/\partial x$, where A is an $m \times m$ matrix with r negative and $m - r$ positive eigenvalues, then the boundary conditions can be written as

$$B_0 u(0, t) = \begin{bmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_r(t) \end{bmatrix}, \quad B_1 u(1, t) = \begin{bmatrix} g_{r+1}(t) \\ g_{r+2}(t) \\ \vdots \\ g_m(t) \end{bmatrix},$$

where B_0 and B_1 are $r \times m$ and $(m - r) \times m$ matrices, respectively. The norm of g is then defined as

$$\|g\| = \left(\sum_{\nu=1}^m |g_\nu|^2 \right)^{1/2}.$$

Let us summarize this section so far. The central concept for getting an energy estimate is semiboundedness. A maximally semibounded operator leads to a well posed problem for homogeneous boundary conditions as defined in Definition 2.8. Furthermore, from this follows an estimate of type (2.8) for nonzero forcing functions. This property is usually sufficient for most applications in practice. However, if one really wants to make sure that the problem is strongly well posed for nonzero boundary data, further analysis is required. There are general theorems for certain classes of problems, where strong well-posedness follows from semiboundedness. For example, first order hyperbolic systems with the correct number of boundary conditions on each side, are strongly well posed, see [Gustafsson et al., 1995]. As was demonstrated above, it is sometimes possible to apply the energy method on the problem in a direct way, to derive this property.

2.3.2 Semidiscrete Approximations

When dealing with initial–boundary value problems, it is useful to first deal with the semidiscrete problem, also called *the method of lines*. Let us first discuss the formulation, by studying the problem

$$\begin{aligned} \frac{du_j}{dt} &= D_0 u_j, \quad j = 1, 2, \dots, N-1, \\ u_0 &= 2u_1 - u_2, \\ u_N(t) &= g(t), \\ u_j(0) &= f_j, \quad j = 0, 1, \dots, N, \end{aligned} \tag{2.28}$$

where the step size h is defined by $Nh = 1$. Since there is no physical boundary condition at the left side, we use linear extrapolation to define u_0 .

If we want to use a standard ODE solver for this problem, there is a technical difficulty here. The ODE solver expects a closed system of ODE complemented by

an initial condition, which requires a reformulation. In order to achieve this, the end components u_0 and u_N can be eliminated, and the first and last differential equation become

$$\begin{aligned}\frac{du_1}{dt} &= D_+ u_1, \\ \frac{du_{N-1}}{dt} &= -\frac{u_{N-2}}{2h} + \frac{g(t)}{2h}.\end{aligned}$$

Denoting by \mathbf{u} the vector with components u_1, u_2, \dots, u_{N-1} , and similarly for \mathbf{f} , we can use vector/matrix notation, and we get

$$\begin{aligned}\frac{d\mathbf{u}}{dt} &= Q\mathbf{u} + \mathbf{F}, \\ \mathbf{u}(0) &= \mathbf{f},\end{aligned}\tag{2.29}$$

where

$$Q = \frac{1}{h} \begin{bmatrix} -1 & 1 & 0 & \dots & \dots & 0 \\ -1/2 & 0 & 1/2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ \vdots & & & -1/2 & 0 & 1/2 \\ 0 & \dots & \dots & \dots & -1/2 & 0 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ g/(2h) \end{bmatrix}.$$

This is a proper formulation for applying an ODE solver. However, when it comes to analysis, we have to remember that the vector F is unbounded as a function of h .

The original formulation (2.28) has certain advantages when it comes to analysis of the system. The reason is that the boundary conditions are singled out, which allows for analysis analogous to the one carried out for the PDE problem above.

There is a third possible formulation that is as natural as the other two. We have seen that u_0 can be eliminated, which results in the modified difference operator D_+ at the left end. Let the difference operator Q be defined by

$$Qu_j = \begin{cases} D_+ u_j, & j = 1, \\ D_0 u_j, & j = 2, 3, \dots, N-1. \end{cases}$$

Then we can write the approximation as

$$\begin{aligned}\frac{du_j}{dt} &= Qu_j, \quad j = 1, 2, \dots, N-1, \\ u_N(t) &= g(t), \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, N.\end{aligned}\tag{2.30}$$

In this way, the somewhat disturbing introduction of F as an unbounded vector function is avoided.

There is actually still another possibility. We can interpret the one-sided approximation at the boundary as an extra boundary condition, and use the form

$$\begin{aligned}\frac{du_j}{dt} &= D_0 u_j, \quad j = 2, 3, \dots, N-1, \\ \frac{du_1}{dt} &= D_+ u_1, \\ u_N(t) &= g(t), \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, N.\end{aligned}\tag{2.31}$$

This last one is similar to (2.28), the only difference is that u_0 has been eliminated, and the extra boundary condition to the left is formulated differently.

Whatever form is used, the general problem is formulated as

$$\begin{aligned}\frac{du_j}{dt} &= Qu_j + F_j, \quad j = 1, 2, \dots, N-1, \\ B_h u &= g(t), \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, N.\end{aligned}\tag{2.32}$$

(The numbering is here normalized, such that x_1, x_2, \dots, x_{N-1} are considered as inner points.) Here $B_h u = g(t)$ denotes the complete set of boundary conditions, which means that it has as many components as are required for the ODE system to have a unique solution. In contrast to the PDE problem, the number of boundary conditions is trivially obtained for the ODE system. (The number of boundary conditions should always be understood as the number of linearly independent conditions.) If Qu_j involves r points to the left of x_j and p points to the right, and $u(x, t)$ is a vector with m components, then the vector $B_h u$ (and g) has $m(r+p)$ scalar components. In our example above with $m = 1$, we have demonstrated three alternatives: two, one or zero boundary conditions. We can use either one of them, but prefer (2.28) or (2.31), where the unbounded F is not introduced. However, we repeat that the form (2.29) is the most convenient one for the application of an ODE solver.

We can now define semibounded difference operators in complete analogy with differential operators. The discrete scalar product and norm for real grid vector functions are defined by

$$(u, v)_h = \sum_{j=1}^{N-1} q_j \langle u_j, \tilde{H} v_j \rangle h, \quad \|u\|_h^2 = (u, u)_h.\tag{2.33}$$

Here $q_j > 0$ and \tilde{H} is a positive definite symmetric matrix. This definition is analogous to the continuous case, but later we shall generalize it further. Furthermore, the numbering of the grid may be different in the formulation of the problem, such that for example, the scalar product starts the summation at some other point $x_{j_0} \neq x_1$.

Definition 2.10. Let \mathcal{V}_h be the space of grid vector functions v satisfying the boundary conditions $B_h v = 0$. The difference operator Q is *semibounded* if for all $v \in \mathcal{V}_h$ the inequality

$$(v, Qv)_h \leq \alpha \|v\|_h^2$$

holds, where α is a constant independent of v and h . \square

As an example, we consider the problem (2.30), but shift the grid point numbering one step in order to comply with the numbering that will be used later in Chapter 7:

$$\begin{aligned} \frac{du_j}{dt} &= Du_j, \quad j = 0, 1, \dots, N-1, \\ u_N(t) &= 0, \\ u_j(0) &= f_j, \end{aligned} \tag{2.34}$$

where

$$Du_j = \begin{cases} D_+ u_j & j = 0, \\ D_0 u_j & j = 1, 2, \dots, N-1. \end{cases}$$

The scalar product is defined by

$$(u, v)_h = \frac{h}{2} u_0 v_0 + \sum_{j=1}^{N-1} u_j v_j h.$$

Then

$$(u, Du)_h = \frac{1}{2} u_0 (u_1 - u_0) + \frac{1}{2} \sum_{j=1}^{N-1} u_j (u_{j+1} - u_{j-1}) = -\frac{1}{2} |u_0|^2 \leq 0,$$

showing that we have a semibounded operator.

For the continuous problem, the fundamental tool for verifying semiboundedness is integration by parts. In the discrete case, we use instead *summation by parts*. Unfortunately, the rules become a little more involved for difference operators compared to differential operators. With the notation

$$(u, v)_{r,s} = \sum_{j=r}^s \bar{u}_j v_j h,$$

we have

$$\begin{aligned} (u, D_+ v)_{r,s} &= -(D_- u, v)_{r+1,s+1} + \bar{u}_{s+1} v_{s+1} - \bar{u}_r v_r, \\ (u, D_0 v)_{r,s} &= -(D_0 u, v)_{r+1,s+1} + \frac{1}{2} (\bar{u}_s v_{s+1} + \bar{u}_{s+1} v_s) - \frac{1}{2} (\bar{u}_{r-1} v_r + \bar{u}_r v_{r-1}). \end{aligned} \tag{2.35}$$

For periodic grid functions with $u_j = u_{j+N+1}$, we get the simple relations

$$\begin{aligned} (u, D_+ v)_{1,N} &= -(D_- u, v)_{1,N}, \\ (u, D_0 v)_{1,N} &= -(D_0 u, v)_{1,N}. \end{aligned} \tag{2.36}$$

Semiboundedness is the basis for stability, which we first formally define:

Definition 2.11. The problem (2.32) is *stable* if for $F = 0, g = 0$ the solution satisfies

$$\|u(t)\|_h \leq Ke^{\alpha t} \|f\|_h,$$

where K and α are constants independent of f and h . \square

(K and α are generic constants, and do not have to be the same as for the underlying PDE problem.) Note the important extra condition, that the constants K and α must be independent of h . The estimate must be uniform independent of the grid, otherwise we get into trouble when studying the convergence of the numerical solution to the true solution as $h \rightarrow 0$. Or in more practical terms: if the grid is refined, then a more accurate solution is expected, and that could of course not be guaranteed if the constants depend on h .

In analogy with well-posedness for the continuous case we have

Theorem 2.5. If the difference operator Q is semibounded, then the problem (2.32) is stable. \square

If the forcing function F is nonzero, an estimate is obtained just as for the PDE problem. However, here is where a formulation of the problem, which allows for an unbounded F , causes trouble. An estimate in terms of $\|F(t)\|$ is not good enough in such a case.

For completeness, we also make

Definition 2.12. The problem (2.32) is *strongly stable* if there is a unique solution satisfying

$$\|u(t)\|_h^2 \leq Ke^{\alpha t} \left(\|f\|_h^2 + \int_0^t (\|F(\cdot, \tau)\|_h^2 + |g(\tau)|^2) d\tau \right),$$

\square

where K and α are constants independent of f, F, g and h .

As we shall see in the next section, semiboundedness for the method of lines, sometimes leads to stability also for the fully discrete approximation. In such a case, there is of course a considerable simplification if the analysis for a given problem can be limited to the discrete operator in space.

Until now we have discussed various stability definitions which all have the common property that the solutions are bounded in some sense, but always independent of h when $h \rightarrow 0$. The formal stability definitions allow for unlimited growth in time for *fixed* step sizes h and k , and in practice this cannot be allowed except for short time integrations. Stability is certainly necessary, but it is natural to introduce an extra condition requiring boundedness of the norm as a function of time (for zero forcing function and boundary data). There are many different notations for this property in the literature, here we will use the quite common name *time stable*. We make the formal definition

Definition 2.13. The problem (2.32) is *time stable* if for $F = 0$ and $g = 0$ there is a unique solution satisfying

$$\|u(t)\|_h \leq K\|f\|_h,$$

where K is independent of f , h and t . \square

Obviously, we have the following theorem:

Theorem 2.6. *If the difference operator Q is semibounded and satisfies*

$$(v, Qv)_h \leq 0$$

for all v satisfying the homogeneous boundary conditions, then the approximation (2.32) is time stable. \square

Clearly, we must require that the underlying PDE problem has this property. Already for periodic solutions, the differential equation

$$u_t = a(x)u_x$$

may have growing solutions, while

$$u_t = (a(x)u)_x + a(x)u_x$$

has not. In the latter case, the differential operator $P = (\partial/\partial x)a(x) + a(x)\partial/\partial x$ is skewsymmetric, i.e., $(u, Pv) = -(Pu, v)$. This means that $(u, Pu) = 0$, i.e., P is semibounded.

2.3.3 Fully Discrete Approximations

We consider the general difference scheme

$$\begin{aligned} u_j^{n+1} &= Qu_j^n + kF_j^n, \quad j = 1, 2, \dots, N-1 \\ B_h u^n &= g^n, \\ u_j^0 &= f_j, \quad j = 1, 2, \dots, N-1, \end{aligned} \tag{2.37}$$

Here B_h is a boundary operator, that connects points locally in the neighborhood of the boundary point (x_0, t_n) and (x_N, t_n) respectively. In the statement of the problems, we will define the initial function f_j starting at $j = 1$, and then assume that it satisfies the boundary conditions such that Qu_j^0 is always well defined.

The energy method principle is very simple just as in the semidiscrete case, where one tries to find a norm such that

$$\frac{d}{dt} \|u\|_h^2 \leq 0,$$

for the case $F_j(t) = 0$, $g(t) = 0$. In the fully discrete case, we try to find a norm such that

$$\|u^{n+1}\|_h^2 \leq \|u^n\|_h^2, \tag{2.38}$$

for $F_j^n = 0$, $g^n = 0$. As an example, consider the semidiscrete approximation (2.32), where Q is semibounded. The *trapezoidal rule* approximation in time leads to the *Crank–Nicholson scheme*

$$\begin{aligned} u_j^{n+1} - u_j^n &= \frac{k}{2} Q(u_j^{n+1} + u_j^n), \quad j = 1, 2, \dots, N-1, \\ B_h u^n &= 0, \\ u_j^0 &= f_j, \quad j = 1, 2, \dots, N-1. \end{aligned} \tag{2.39}$$

The difference operator Q is semibounded for a certain scalar product $(u, v)_h$, where u and v satisfy the homogeneous boundary conditions. When taking the scalar product of the first equation in (2.39) with the grid function $u_j^{n+1} + u_j^n$, we obtain

$$\|u^{n+1}\|_h^2 - \|u^n\|_h^2 = (u_j^{n+1} + u_j^n, \frac{k}{2} Q(u_j^{n+1} + u_j^n))_h.$$

By linearity of B_h , we have $B_h(u_j^{n+1} + u_j^n) = 0$, and by linearity and semiboundedness of Q , the right hand side is nonpositive, i.e., (2.38) is satisfied.

The stability definitions are analogous to the semidiscrete case:

Definition 2.14. The problem (2.37) is *stable* if for $F_j^n = 0$, $g^n = 0$ the solution satisfies

$$\|u^n\|_h \leq K e^{\alpha t_n} \|f\|_h,$$

where K and α are constants independent of f , h and k . \square

Definition 2.15. The problem (2.37) is *strongly stable* if the solution satisfies

$$\|u^n\|_h^2 \leq K e^{\alpha t_n} \left(\|f\|_h^2 + \sum_{\nu=0}^{n-1} (\|F^\nu\|_h^2 + |g^\nu|^2) k \right),$$

where K and α are constants independent of F^n , g^n , f , h and k . \square

The definitions are here given for one step schemes. However, as was demonstrated for the Cauchy problem, any multistep scheme can be rewritten as a one step scheme by introducing the vector

$$\mathbf{u}^n = [u^{n+q} \ u^{n+q-1} \ \dots \ u^n]^T$$

and similarly for F , g and f . The stability definitions then become the same as above but with u , F , g , f substituted by \mathbf{u} , \mathbf{F} , \mathbf{g} , \mathbf{f} . For a PDE system with m differential equations and a $(q+1)$ -step scheme, the standard l_2 -norm of \mathbf{u} is defined by

$$\|\mathbf{u}^n\|_h^2 = \sum_{j=1}^{N-1} \sum_{\mu=1}^m \sum_{\nu=0}^q |(u_j^{(\mu)})^{n+\nu}|^2 h,$$

and similarly for \mathbf{F} , \mathbf{g} , \mathbf{f} .

The meaning of the concept *method of lines* is that some standard ODE solver is applied to a stable semidiscrete approximation. With u denoting a vector containing the discrete solution at the grid points in space, the ODE system is

$$\begin{aligned}\frac{du}{dt} &= Qu, \\ u(0) &= f,\end{aligned}\tag{2.40}$$

where Q is a matrix representation of the difference operator. The ideal situation would be that the fully discrete approximation corresponding to the ODE solver is stable as well, at least for some reasonable bound on the time step. Unfortunately, the theoretical basis for such conclusions is not very complete. We saw one example above, where the trapezoidal rule leads to unconditional stability, another example is the *Euler backward* approximation

$$(I - kQ)u^{n+1} = u^n,$$

which is also stable for semibounded difference operators Q . However, in general the situation is not that favorable, in particular for high order time discretizations.

The stability theory for ODE solvers is well developed for systems with a fixed size N . However, for discretizations of PDE systems, N increases without bounds as $h \rightarrow 0$. If Q can be diagonalized, then a new system of scalar equations is obtained after transformation. If the matrix Q in (2.40) is constant we define the new vector $v = T^{-1}u$, where $T^{-1}QT = \Lambda$ is diagonal. The system becomes

$$\begin{aligned}\frac{dv}{dt} &= \Lambda v, \\ v(0) &= T^{-1}f.\end{aligned}$$

We assume that a certain time discretization is stable for each scalar equation, and consequently we have for the whole system

$$\|v^n\|_h \leq K \|T^{-1}f\|_h,$$

where K is a constant. However, in order to get an estimate of u^n in terms of f , we need a bounded condition number $\text{cond}(T) = \|T\|_h \|T^{-1}\|_h$, where

$$\|T\|_h = \max_{\|v\|_h=1} \|Tv\|_h.$$

In such a case we have

$$\|u^n\|_h \leq \|T\|_h \|v^n\|_h \leq K \|T\|_h \|T^{-1}f\|_h \leq K \|T\|_h \|T^{-1}\|_h \|f\|_h.$$

However, for initial-boundary value problems, it may be hard to verify the boundedness of the condition number.

For real world problems, one is often choosing the time step based on the von Neumann condition, i.e., the diagonal matrix Λ represents the eigenvalues of the Fourier transform of the difference operator in space without boundary conditions. A further step may be to compute the eigenvalues of the operator when the boundary conditions are included as well. Neither method is of course sufficient for stability. In practice, if the choice of time step based on this type of analysis turns out to produce instabilities, an experimental trial and error approach often forms a somewhat uncertain basis for the choice of a restricted time step.

Analysis of scalar ODE may serve as a first tool when comparing different methods, and furthermore, it can be used to rule out a certain method. If a method is no good for the scalar equation, it is no good for the real problem. In Chapter 5 we shall further discuss different time discretization methods.

2.4 Initial–Boundary Value Problems and Normal Mode Analysis for Hyperbolic Systems

As we have seen in the previous section, the energy method has a simple structure, and is convenient to use when it works. However, there are two different drawbacks: it is limited to PDE with Hermitian coefficient matrices (otherwise integration by parts doesn't work), and it gives only sufficient conditions for stability. If one fails to prove well-posedness or stability, it may be because one has not found the right form of the scalar product and norm. Therefore there is a need for a more powerful tool for analysis, and it is provided by the so called *normal mode analysis*, which is based on the Laplace transform. It can be applied to PDE of arbitrary order, but in this book we shall limit ourselves to first order hyperbolic systems (2.11) in one space dimension with application for example to wave propagation and fluid dynamics. The continuous problems in 1-D are easily understood by analysis of the characteristics, which provides necessary and sufficient conditions for well-posedness. On the contrary, the discretized problems require further tools already in one space dimension as demonstrated above.

The normal mode analysis is useful for the continuous problem in several space dimensions, where well-posedness is a nontrivial matter. However, in this book we turn directly to the semidiscrete problem.

2.4.1 Semidiscrete Approximations

As an introduction we consider the model problem

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(x, 0) &= f(x), \\ \|u(\cdot, t)\| &< \infty, \end{aligned}$$

where $\|u(\cdot, t)\|^2 = \int_0^\infty |u(x, t)|^2 dx$. The right boundary has been removed here, and the boundary condition is substituted by the requirement that the L_2 -norm is finite, i.e., we are considering the *quarter space problem*. As we shall see, this simplifies the analysis considerably, and the more practical situation with a boundary also on the right hand side can be handled by applying a general theorem to be presented later.

We consider again the centered second order difference approximation

$$\begin{aligned} \frac{du_j}{dt} &= D_0 u_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ u_0 &= 2u_1 - u_2, \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, \\ \|u(t)\|_h &< \infty, \end{aligned} \tag{2.41}$$

where $\|u(t)\|_h^2 = \sum_{j=1}^\infty |u_j(t)|^2 h$. First, we are looking for solutions of the type $\phi_j e^{st}$, where s is a complex number. The system of equations for ϕ becomes

$$\begin{aligned} \tilde{s}\phi_j &= \frac{1}{2}(\phi_{j+1} - \phi_{j-1}), \quad j = 1, 2, \dots, \quad 0 \leq t, \\ \phi_0 &= 2\phi_1 - \phi_2, \\ \|\phi(t)\|_h &< \infty, \end{aligned} \tag{2.42}$$

where $\tilde{s} = sh$. This is actually an eigenvalue problem. We are looking for the eigenvalues \tilde{s} and eigenfunctions ϕ of the difference operator hD_0 in the space of grid functions with a finite l_2 -norm and satisfying the boundary condition. Assume that we find such an eigenvalue \tilde{s}_0 with $\operatorname{Re} \tilde{s}_0 > 0$. This means that for each h there is a solution $\phi_j e^{\tilde{s}_0 t/h}$ of (2.41) with ϕ_j as initial data. For decreasing h , we have a sequence of solutions with increasing exponential growth, and the method is obviously unstable. The conclusion is that an eigenvalue of (2.42) with $\operatorname{Re} \tilde{s} > 0$ cannot be allowed.

Let us next find out how to solve the eigenvalue problem. We can of course consider the difference operator as a matrix, and apply all the well known techniques from linear algebra for finding the eigenvalues. However, difference operators with constant coefficients have a very special structure, allowing for a special technique of computation.

The first equation of (2.42) is a difference equation with constant coefficients, and the solution is defined in terms of the roots of its *characteristic equation* (see Appendix A)

$$\kappa^2 - 2\tilde{s}\kappa - 1 = 0. \tag{2.43}$$

These roots are

$$\begin{aligned} \kappa_1 &= \tilde{s} - \sqrt{1 + \tilde{s}^2}, \\ \kappa_2 &= \tilde{s} + \sqrt{1 + \tilde{s}^2}. \end{aligned} \tag{2.44}$$

Since \tilde{s} and κ are complex, it is necessary to define the square root uniquely, and we use the standard definition, i.e., $\sqrt{-1} = i$ (not $-i$). We shall use this convention throughout the remainder of the book.

If the roots are distinct, the solution ϕ is

$$\phi_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j, \quad (2.45)$$

where σ_1 and σ_2 are constants to be determined by the boundary conditions. The problem at hand is to determine whether or not there is an eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} > 0$, i.e., if there is any nontrivial solution ϕ for $\operatorname{Re} \tilde{s} > 0$. In order to do that, we first note that there is no root κ on the unit circle. Assume that there is one, i.e., $\kappa = e^{i\xi}$, where ξ is real. Then the characteristic equation implies

$$\tilde{s} = \frac{1}{2}(\kappa - \kappa^{-1}) = \frac{1}{2}(e^{i\xi} - e^{-i\xi}) = i \sin \xi,$$

which is a contradiction. Since the roots are continuous functions of its coefficients, it follows that the roots are either strictly inside or outside the unit circle as long as \tilde{s} is strictly in the right halfplane. Therefore one can pick any value of \tilde{s} to the right of the imaginary axis to determine the location of the roots κ . For convenience, we pick $\tilde{s} = \delta$, where $\delta > 0$ is arbitrary small. A Taylor expansion gives

$$\begin{aligned} \kappa_1 &= \delta - \sqrt{1 + \delta^2} = -1 + \delta + \mathcal{O}(\delta^2), \\ \kappa_2 &= \delta + \sqrt{1 + \delta^2} = 1 + \delta + \mathcal{O}(\delta^2), \end{aligned} \quad (2.46)$$

i.e., $|\kappa_1| < 1$ and $|\kappa_2| > 1$. Obviously, the roots are distinct, and (2.45) is the proper form of the solution. The condition $\|\phi(t)\|_h < \infty$ implies $\sigma_2 = 0$, which means that the solution has the simple form

$$\phi_j = \sigma_1 \kappa_1^j.$$

Substitution into the boundary condition gives the final condition

$$\sigma_1 (\kappa_1 - 1)^2 = 0.$$

Since $\kappa_1 \neq 1$, the only possibility is $\sigma_1 = 0$, which means that $\phi_j \equiv 0$. The conclusion is that there is no eigenvalue in the right halfplane.

Next we consider the general semidiscrete problem

$$\begin{aligned} \frac{du_j}{dt} &= Qu_j + F_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ B_h u &= g(t), \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, \\ \|u(t)\|_h &< \infty. \end{aligned} \quad (2.47)$$

Here Q is a difference operator of the form

$$Q = \frac{1}{h} \sum_{\nu=-r}^p A_\nu E^\nu, \quad (2.48)$$

where A_ν are uniformly bounded matrices as $h \rightarrow 0$. The boundary operator B_h is acting on u in a neighborhood of $j = 0$. For convenience it is assumed that B_h is such that the grid function outside the computational domain can be explicitly expressed as

$$u_\nu = \sum_{j=1}^q B_{\nu j} u_j + g_\nu, \quad \nu = 0, -1, \dots, -r+1,$$

where $B_{\nu j}$ are uniformly bounded matrices. Later we shall relax this condition and allow for more general boundary conditions, even including differential operators d/dt . The eigenvalue problem is

$$\begin{aligned} \tilde{s}\phi_j &= hQ\phi_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ B_h\phi &= 0, \\ \|\phi\|_h &< \infty. \end{aligned} \quad (2.49)$$

We have

Definition 2.16. The *Godunov–Ryabenkii condition* is satisfied if there is no eigenvalue \tilde{s} of the problem (2.49) with $\operatorname{Re} \tilde{s} > 0$. \square

This condition corresponds to the von Neumann condition for the periodic case. The difference here is that the space of eigenfunctions is different from the Fourier modes, which span a particularly simple space.

An analysis completely analogous to the one for the model example above, shows that the Godunov–Ryabenkii condition is necessary for stability of the general approximation (2.47).

Let us next consider how to verify the Godunov–Ryabenkii condition. Following the same principles as for the example above, we first define the general form of the solution to the first equation of (2.49), which is

$$(\tilde{s}I - \sum_{\nu=-r}^p A_\nu E^\nu)\phi_j = 0, \quad j = 1, 2, \dots$$

Referring to Appendix A, we look for nontrivial vector solutions ψ of

$$(\tilde{s}I - \sum_{\nu=-r}^p A_\nu \kappa^\nu)\psi = 0, \quad \psi = \begin{bmatrix} \psi^{(1)} \\ \psi^{(2)} \\ \vdots \\ \psi^{(m)} \end{bmatrix}. \quad (2.50)$$

These vectors ψ are a kind of generalized eigenvectors corresponding to the roots κ , but should be distinguished from the eigensolutions ϕ corresponding to the eigenvalues \tilde{s} of (2.49). The $m \times m$ coefficient matrix must be singular, and we have the characteristic equation

$$\operatorname{Det}(\tilde{s}I - \sum_{\nu=-r}^p A_\nu \kappa^\nu) = 0. \quad (2.51)$$

There are $m(r+p)$ roots κ_μ to this equation. We showed for the example above, that there were no roots κ_μ on the unit circle for $\operatorname{Re} \tilde{s} > 0$. This property holds also for the general case:

Lemma 2.3. *Assume that the von Neumann condition is satisfied for the periodic problem. If $\operatorname{Re} \tilde{s} > 0$, then*

i) *There are no roots κ_μ of (2.51) with $|\kappa_\mu| = 1$.*

ii) *If A_{-r} is nonsingular, then there are exactly rm roots κ_μ with $|\kappa_\mu| < 1$.* \square

The proof of the first part follows just as for the example above by assuming that $\operatorname{Re} \tilde{s} > 0$ and $\kappa_\mu = e^{i\xi}$, where ξ is real. But then \tilde{s} is an eigenvalue of the matrix $\sum_{\nu=-r}^p A_\nu e^{i\nu \xi}$, which is the Fourier transform of the difference operator in space, and this contradicts the von Neumann condition.

The proof of the second part follows by considering large values of $\operatorname{Re} \tilde{s}$.

The lemma shows that the set of complex roots κ_μ can be split into two subsets, which are uniquely defined for all \tilde{s} in the right halfplane.

If there are $m(r+p)$ linearly independent eigenvectors ψ_μ of (2.50), then the grid function ϕ has the form

$$\phi_j = \sum_{\mu=1}^{m(r+p)} \sigma_\mu \psi_\mu \kappa_\mu^j,$$

where σ_μ are scalar constants. Furthermore, Lemma 2.3 and the condition $\|\phi\|_h < \infty$ imply that there are mp constants σ_μ that must be zero. Therefore, with the proper numbering of the roots κ_μ , the solution has the form

$$\phi_j = \sum_{\mu=1}^{mr} \sigma_\mu \psi_\mu \kappa_\mu^j, \quad |\kappa_\mu| < 1,$$

for $\operatorname{Re} \tilde{s} > 0$.

As shown in Appendix A, if there is not a full set of eigenvectors ψ_μ , the solution ϕ has the more general form

$$\phi_j = \sum_{|\kappa_\mu| < 1} P_\mu(j) \kappa_\mu^j. \quad (2.52)$$

Here, $P_\mu(j)$ are polynomials in j with vector coefficients, containing mr constants σ_μ .

Whatever form the solution ϕ has, it is substituted into the boundary conditions $B_h \phi = 0$. The condition for a nontrivial solution results in a linear system of equations for the σ -coefficients:

$$C(\tilde{s})\boldsymbol{\sigma} = 0, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{mr} \end{bmatrix} \neq \mathbf{0}. \quad (2.53)$$

Since we don't allow such a solution, we require

$$\text{Det } C(\tilde{s}) \neq 0, \quad \text{Re } \tilde{s} > 0, \quad (2.54)$$

which is equivalent to the Godunov–Ryabenkii condition.

We summarize the procedure for verifying this condition:

1. Assume that $\text{Re } \tilde{s} > 0$.
2. Find the roots κ_μ of the characteristic equation (2.51) with $|\kappa_\mu| < 1$.
3. Find the corresponding eigenvectors ψ_μ of the equation (2.50).
4. Find the general form (2.52) of the eigensolution ϕ .
5. Substitute this solution into the boundary condition $B_h\phi = 0$, and derive the matrix $C(\tilde{s})$ in (2.53).
6. Investigate whether or not $\text{Det } C(\tilde{s}) = 0$ for some value \tilde{s} with $\text{Re } \tilde{s} > 0$. If such an eigenvalue \tilde{s} exists, then the Godunov–Ryabenkii condition is violated.

Indeed, this procedure can often be simplified, since the explicit expression of κ_μ in terms of \tilde{s} is not always required. In our example above, we need only to know that the solution has the form $\phi_j = \sigma_1 \kappa_1^j$, where $|\kappa_1| < 1$. The boundary condition then implies that the condition for an eigenvalue is $\kappa_1 = 1$, which is a contradiction. For our example, this is the complete procedure for verifying the Godunov–Ryabenkii condition.

The next step in the stability theory is to find sufficient conditions for stability, and this theory is based on the Laplace transform. Assume that $u(t)$ is a function of t . Then the Laplace transformed function \hat{u} is defined by

$$\hat{u}(s) = \mathcal{L}u = \int_0^\infty e^{-st} u(t) dt,$$

where $s = i\xi + \eta$ is a complex number with η large enough such that the integral exists. The inverse Laplace transform is

$$u(t) = \mathcal{L}^{-1}\hat{u} = \frac{1}{2\pi i} \int_{\text{Re } s=\eta} e^{st} \hat{u}(s) ds,$$

which can also be written as

$$u(t) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} e^{(i\xi+\eta)t} \hat{u}(i\xi + \eta) d\xi.$$

Parseval's relation is

$$\int_0^\infty e^{-2\eta t} |u(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{u}(i\xi + \eta)|^2 d\xi.$$

Just as for the periodic case, this relation makes it convenient to derive stability estimates in Laplace space, and then transform back to physical space.

We also need the transform of the differential operator in time. It is convenient to assume that $u(0) = 0$, and we get

$$\mathcal{L} \left(\frac{du}{dt} \right) = \int_0^\infty e^{-st} \frac{du}{dt} dt = [e^{-st} u(t)]_{t=0}^\infty + s \int_0^\infty e^{-st} u(t) dt.$$

Here it is assumed that s is such that all integrals exist, and furthermore that $e^{-st} u(t) \rightarrow 0$ as $t \rightarrow \infty$. Then we get the relation

$$\mathcal{L} \left(\frac{du}{dt} \right) = s \mathcal{L} u.$$

We begin by analyzing the familiar example once more. This time we assume that the initial data are zero, but there is instead nonzero boundary data:

$$\begin{aligned} \frac{du_j}{dt} &= D_0 u_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ u_0 - 2u_1 + u_2 &= g(t), \\ u_j(0) &= 0, \quad j = 1, 2, \dots, \\ \|u(t)\|_h &< \infty. \end{aligned} \tag{2.55}$$

The assumption of zero initial data is no more strange than assuming zero boundary data in our previous analysis. In both cases it is just a step in the analysis; when solving the real problem numerically, there is of course no such restriction.

We now take the Laplace transform of the whole approximation, and we get

$$\begin{aligned} \tilde{s}\hat{u}_j &= \frac{1}{2}(\hat{u}_{j+1} - \hat{u}_{j-1}), \quad j = 1, 2, \dots, \quad 0 \leq t, \\ \hat{u}_0 - 2\hat{u}_1 + \hat{u}_2 &= \hat{g}, \\ \|\hat{u}(s)\|_h &< \infty, \end{aligned}$$

where \hat{g} is the Laplace transform of g . Note the similarity with the eigenvalue problem (2.42). The only difference is that there is now a nonzero function \hat{g} in the boundary condition. This means that the tools for analyzing the eigenvalue problem can be used here as well. The general form of the solution of the first equation is the same, and by applying the boundedness condition on the norm, one of the two components of the solution is eliminated. The solution is

$$\hat{u}_j = \sigma_1 \kappa_1^j,$$

and the boundary condition implies

$$\sigma_1 (\kappa_1 - 1)^2 = \hat{g},$$

i.e.,

$$\hat{u}_j = \frac{\hat{g}}{(\kappa_1 - 1)^2} \kappa_1^j.$$

Obviously we need an estimate from below of $|\kappa_1 - 1|^2$. Recalling Parseval's relation above, we need this estimate to hold for all s with $\operatorname{Re}s > \eta$ for some constant η . But the root κ_1 is a function of $\tilde{s} = sh$, and since h is arbitrarily small, the estimate must hold uniformly for \tilde{s} arbitrarily close to the imaginary axis. In other words, we need

$$|\kappa_1 - 1| \geq \delta > 0, \quad \operatorname{Re}\tilde{s} \geq 0.$$

Clearly, $\kappa_1 \neq 1$ for $\operatorname{Re}\tilde{s} > 0$. The question now is whether there is any \tilde{s} on the imaginary axis such that $\kappa_1(\tilde{s}) = 1$. The characteristic equation (2.43) shows that the corresponding critical point for \tilde{s} is $\tilde{s} = 0$. But according to the explicit expression for the roots κ_μ in (2.44), we have $\kappa_1(0) = -1$. The other root κ_2 is one, but that component is already eliminated, and has no influence on the estimate.

The conclusion is that for any fixed j the estimate

$$|\hat{u}_j| \leq K |\hat{g}(s)|$$

holds for all s with $\operatorname{Re}s \geq 0$, where K is independent of s . When using Parseval's relation, we must choose an integration line $\operatorname{Re}s = \eta$ such that the integral of $|\hat{g}|^2$ exists. However, $\hat{g}(s)$ is the Laplace transform of $g(t)$. Formally we are integrating to infinity in time, but for the practical computation, one is interested in a finite time interval $0 \leq t \leq T$. Therefore, we can without restriction assume that $g(t) = 0$ for $t > T$, which allows for integration along the line $\eta = 0$. The estimate in physical space is

$$\int_0^\infty |u_j(t)|^2 dt \leq \frac{K^2}{2\pi} \int_{-\infty}^\infty |\hat{g}(i\xi)|^2 d\xi = K^2 \int_0^\infty |g(t)|^2 dt$$

for any fixed j . As explained above, the upper limit in the last integral can be substituted by T , and the final estimate becomes (by trivially doing the same substitution in the first integral)

$$\int_0^T |u_j(t)|^2 dt \leq K_1 \int_0^T |g(t)|^2 dt$$

for some constant K_1 .

We now turn to the general problem, and consider

$$\begin{aligned} \frac{du_j}{dt} &= Qu_j, \quad j = 1, 2, \dots, \quad 0 \leq t \\ B_h u &= g(t), \\ u_j(0) &= 0, \quad j = -r+1, -r+2, \dots, \\ \|u(t)\|_h &< \infty. \end{aligned} \tag{2.56}$$

The Laplace transformed system is

$$\begin{aligned}\tilde{s}\hat{u}_j &= hQ\hat{u}_j, \quad j = 1, 2, \dots, \\ \hat{B}_h\hat{u} &= \hat{g}, \\ \|\hat{u}\|_h &< \infty.\end{aligned}\tag{2.57}$$

The same procedure as in the example is carried out for the general problem, giving the solution

$$\hat{u}_j = \sum_{|\kappa_\mu|<1} P_\mu(j) \kappa_\mu^j,$$

where the right hand side contains mr parameters σ_μ . These parameters are determined by the system

$$C(\tilde{s})\boldsymbol{\sigma} = \hat{g}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{mr} \end{bmatrix} \neq \mathbf{0}.$$

We need also here an estimate of $|\hat{u}|$ for all \tilde{s} in the right halfplane including the imaginary axis, and it has the form

$$|\hat{u}_j(\tilde{s})| \leq K|\hat{g}|, \quad Re \tilde{s} > 0,$$

where K is a constant independent of \tilde{s} . This is the Kreiss condition, but before making the formal definition, we shall discuss two alternative forms.

One alternative way is to make the definition in terms of $\text{Det } C(\tilde{s})$. The difference compared to the Godunov–Ryabenkii condition is that the condition (2.54) is sharpened such that the imaginary axis is included in the domain for \tilde{s} .

When talking about eigenvalues \tilde{s} , we run into a problem. If $\text{Det } C(\tilde{s}_0) = 0$, then there is a nontrivial solution to the eigenvalue problem. If $Re \tilde{s}_0 > 0$, then \tilde{s}_0 is an eigenvalue. However, if $Re \tilde{s}_0 = 0$, we recall that the roots κ_μ do not necessarily separate into two sets located on opposite sides of the unit circle. Accordingly, there is the possibility that there is at least one κ_μ with $|\kappa_\mu(\tilde{s})| = 1$. If this is the case, then the condition $\|\phi\|_h < \infty$ is violated, and the eigenfunction is not in the proper function space. Therefore we make

Definition 2.17. If there is a $\tilde{s}_0 = i\xi_0$ such that $\text{Det } C(i\xi_0) = 0$ and at least one root $\kappa_\mu(i\xi_0)$ of the characteristic equation is located on the unit circle, then \tilde{s}_0 is called a *generalized eigenvalue*. \square

We can now make the formal definition:

Definition 2.18. The Kreiss condition is satisfied if either one of the following equivalent conditions is satisfied.

- i) The solutions of (2.57) satisfy for every fixed j

$$|\hat{u}_j(\tilde{s})| \leq K|\hat{g}|, \quad Re \tilde{s} > 0,$$

where K is a constant independent of \tilde{s} .

ii)

$$\text{Det } C(\tilde{s}) \neq 0, \quad \text{Re } \tilde{s} \geq 0.$$

iii) The problem (2.49) has no eigenvalue or generalized eigenvalue in the right half-plane $\text{Re } \tilde{s} \geq 0$. \square

In the first version i), it would be possible to substitute $\text{Re } \tilde{s} > 0$ by $\text{Re } \tilde{s} \geq 0$. However, by keeping the strict inequality, we automatically separate out the part of the solution corresponding to the roots κ_μ inside the unit circle.

The second version ii) is often called the *determinant condition*. Note that the matrix $C(\tilde{s})$ is well defined, also at the imaginary axis as the limit as $\text{Re } \tilde{s} \rightarrow 0_+$.

As for the example above, the Kreiss condition leads to the estimate

$$\int_0^T |u_j(t)|^2 dt \leq K_1 \int_0^T |g(t)|^2 dt \quad (2.58)$$

for any fixed j . If a difference approximation (2.56) has the property (2.58), it is sometimes called *boundary stable*.

For the Godunov–Ryabenkii condition, a step by step scheme was given for the verification procedure, and it was also indicated how it can be simplified. The procedure is very much the same for the Kreiss condition. Indeed it is identical until the last step, where the condition $\text{Det } C(\tilde{s}) \neq 0$ must hold in the whole right halfplane *including* the imaginary axis. For our example, we don't have to know the explicit expression (2.44) for κ_1 in the solution $\phi_j = \sigma_1 \kappa_1^j$. The boundary condition and the characteristic equation imply that the critical point is $\tilde{s} = 0$ and $\kappa = 1$. Recall that the matrix $C(\tilde{s})$ is unique, and its form on the imaginary axis is defined as the limit when $\text{Re } \tilde{s} \rightarrow 0_+$. Therefore, the only remaining question is whether $\kappa_1(0) = 1$ or $\kappa_2(0) = 1$, where by definition $|\kappa_1(\tilde{s})| < 1$ for $\text{Re } \tilde{s} > 0$. In order to find out, we do a simple perturbation calculation of the characteristic equation (2.43). Since the roots κ_μ cannot cross the unit circle as long as \tilde{s} is strictly to the right of the imaginary axis, we choose the simplest perturbation of \tilde{s} , i.e., we let $\tilde{s} = \delta$, where δ is real, positive and small. The characteristic equation then leads to

$$\begin{aligned} \kappa_1 &= -1 + \delta + \mathcal{O}(\delta^2), \\ \kappa_2 &= 1 + \delta + \mathcal{O}(\delta^2), \end{aligned}$$

i.e., $\kappa_1(0) = -1$ and $\kappa_2(0) = 1$. Consequently, $\tilde{s} = 0$ is *not* a generalized eigenvalue, and the Kreiss condition is satisfied.

For our simple example, the explicit expressions (2.44) for κ_1 and κ_2 were easily found, but for more realistic problems, it is not always the case. Then the procedure described here may be a significant simplification of the analysis.

With the Kreiss condition complete, we consider next the problem

$$\begin{aligned} \frac{du_j}{dt} &= Qu_j + F_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ B_h u &= g(t), \\ u_j(0) &= f_j, \quad j = -r+1, -r+2, \dots, \\ \|u(t)\|_h &< \infty, \end{aligned} \tag{2.59}$$

where nonzero forcing and initial functions have been introduced. For symmetric hyperbolic systems, the following theorem holds:

Theorem 2.7. *Assume that the difference operator Q in (2.48) is semibounded for the Cauchy problem, and that $r \geq p$. If the Kreiss condition is satisfied, then the approximation (2.59) is strongly stable.* \square

(Strong stability is defined in Definition 2.12 with obvious modification of the range of j .) The proof of this theorem is more complicated, and is found in [Gustafsson et al., 1995].

In a practical situation, there is a boundary also at the right as defined in the model problem (2.28). When using normal mode analysis, we investigate each boundary by itself, and we have already treated the right quarter space problem (2.56). The corresponding left quarter space problem is

$$\begin{aligned} \frac{du_j}{dt} &= D_0 u_j, \quad j = N-1, N-2, \dots, \quad 0 \leq t, \\ u_N(t) &= g(t), \\ u_j(0) &= f_j, \quad j = N-1, N-2, \dots, \\ \sum_{-\infty}^{N-1} |u_j|^2 h &< \infty. \end{aligned} \tag{2.60}$$

The analysis of this problem can be carried out after a transformation corresponding to $\tilde{x} = 1 - x$ for the PDE, i.e., j is substituted by $N - j$ in (2.60). This puts the problem in standard form

$$\begin{aligned} \frac{du_j}{dt} &= D_0 u_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ u_0(t) &= g(t), \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, \\ \|u(t)\|_h &< \infty. \end{aligned}$$

The analysis is now identical to the one carried out for the problem (2.55) all the way until the final step, where we get the equation

$$\sigma_1 \kappa_1^0 = \hat{g},$$

i.e., $\sigma_1 = \hat{g}$. This means that the Kreiss condition is trivially satisfied, and after a transformation back, it is clear that the left quarter space problem (2.60) is boundary stable.

The analysis is simplified considerably by treating each quarter space problem separately. The reason for this is that all the components in the solution \hat{u} corresponding to roots κ_μ with $|\kappa_\mu| > 1$ are eliminated. Assuming that both quarter space problems satisfy the Kreiss condition, then it is possible to show that the problem with two boundaries is boundary stable, i.e., there is an estimate of $|\hat{u}_j|$ in terms of the boundary data on both sides if the forcing function F and initial function f are zero.

To prove strong stability for the strip problem with two boundaries, we need a semibounded difference operator Q , and also the condition $r \geq p$. This last condition must be substituted by $p \geq r$ for the left quarter space problem, i.e., we must have $p = r$. Under this condition the strip problem (2.32) is strongly stable.

The theory based on normal mode analysis actually leads more naturally to another type of stability definition for the full problem (2.59). The boundary stability concept is stated in terms of an integral in time of the norm, which is a natural consequence of Parseval's relation. For problems where we don't have an energy estimate for the Cauchy problem, strong stability in the original sense does not follow. Instead we introduce a weaker stability definition:

Definition 2.19. The problem (2.59) is *stable in the generalized sense* if for $f = 0$ and $g = 0$ there is a constant η_0 such that the solution satisfies the estimate

$$\int_0^\infty e^{-\eta t} \|u(t)\|_h^2 dt \leq K(\eta) \int_0^\infty e^{-\eta t} \|F(t)\|_h^2 dt \quad (2.61)$$

for all $\eta > \eta_0$, where $K(\eta) \rightarrow 0$ as $\eta \rightarrow \infty$. \square

A stronger version is given by

Definition 2.20. The problem (2.59) is *strongly stable in the generalized sense* if for $f = 0$ there is a constant η_0 such that the solution satisfies the estimate

$$\int_0^\infty e^{-\eta t} \|u(t)\|_h^2 dt \leq K(\eta) \int_0^\infty e^{-\eta t} (\|F(t)\|_h^2 + |g(t)|^2) dt \quad (2.62)$$

for all $\eta > \eta_0$, where $K(\eta) \rightarrow 0$ as $\eta \rightarrow \infty$. \square

In both definitions, the initial function f is zero. In the actual computation, the initial data are of course in general nonzero. The question is whether we should expect any extra trouble for this reason. Actually we should not. Let $v_j(t)$ be a grid vector function which satisfies the conditions

$$\begin{aligned} v_j(0) &= f_j, \quad j = 1, 2, \dots, \\ \|v(t)\|_h &< \infty, \\ \int_0^\infty e^{-\eta t} \|v(t)\|_h^2 dt &< K_1 \|f\|_h, \end{aligned}$$

where K_1 is a constant. Then the difference $w = u - v$ satisfies the system

$$\begin{aligned}\frac{dw_j}{dt} &= Qw_j + \tilde{F}_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ B_h w(t) &= \tilde{g}(t), \\ w_j(0) &= 0, \quad j = 1, 2, \dots, \\ \|w(t)\|_h < \infty,\end{aligned}$$

where

$$\begin{aligned}\tilde{F}_j &= F_j + Qv_j - \frac{dv_j}{dt}, \\ \tilde{g}(t) &= g(t) - B_h v.\end{aligned}$$

If the approximation is strongly stable in the generalized sense, then there is an estimate of w in terms of \tilde{F} and $\tilde{g}(t)$, and by the assumptions made on v , it follows that there is an estimate also for $u = v + w$:

$$\int_0^\infty e^{-\eta t} \|u(t)\|_h^2 dt \leq K(\eta) \int_0^\infty e^{-\eta t} (\|\tilde{F}(t)\|_h^2 + |\tilde{g}(t)|^2) dt + K_1 \|f\|_h.$$

If Qv_j and dv_j/dt are bounded such that the integral on the right hand side exists, then this estimate is good enough. There is no problem with dv_j/dt , since we can always make v smooth in time. The only complication is if the initial function f is very rough such that $\|Qf\|_h$ is not bounded for small h . Then there may be a growth of the order $1/h$, but this is the worst case that can occur.

If the approximation is only stable in the generalized sense, then the grid vector function v must be constructed such that it satisfies also the boundary condition $B_h v = g(t)$. Then the boundary condition for w is homogeneous, and the estimate (2.61) can be applied with a modified forcing function. In addition to the restriction on f , we now have an extra restriction on g , since dv_j/dt involves dg/dt .

The conclusion is that the zero initial data restriction in the definition of stability in the generalized sense is in general not severe. It is only in the case of rough initial data, and possibly also rough boundary data, that we may expect a weak instability.

Next we shall discuss an example which has a semibounded difference operator, but still violates the Kreiss condition.

Let the matrix A be defined by

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

and consider the problem

$$\begin{aligned}u_t + Au_x &= F, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u^{(1)}(0, t) &= g(t), \\ u(x, 0) &= f(x), \\ \|u\| &< \infty.\end{aligned}$$

The semidiscrete approximation is

$$\begin{aligned} \frac{du_j}{dt} + AD_0 u_j &= F_j, \quad j = 1, 2, \dots, \quad 0 \leq t, \\ u_0^{(1)}(t) &= g(t), \\ \frac{du_0^{(2)}}{dt} + D_+ u_0^{(1)} &= F_0^{(2)}, \\ u_j(0) &= f_j, \quad j = 1, 2, \dots, \\ \|u\|_h < \infty. \end{aligned} \tag{2.63}$$

Here we have introduced an extra numerical boundary condition, which is of non-standard type, since it includes a differential operator in time.

We shall first show that there is an energy estimate for this problem. In order to do that, we shall first reformulate the approximation in accordance with the form (2.30) of the scalar example above. Let the difference operator Q be defined by

$$\begin{aligned} (Qu_j)^{(2)} &= -D_+ u_0^{(1)}, \quad j = 0, \\ Qu_j &= -AD_0 u_j, \quad j = 1, 2, \dots, \end{aligned}$$

and define $\tilde{u} = [u_0^{(2)}, u_1^{(1)}, u_1^{(2)}, \dots]^T$. Then we can write the approximation for $F = 0$ and $g = 0$ as

$$\begin{aligned} \frac{d\tilde{u}_j}{dt} &= Qu_j, \quad j = 0, 1, \dots, \\ u_0^{(1)}(t) &= 0, \\ u_j(0) &= f_j, \quad j = 0, 1, \dots, \\ \|u\|_h < \infty. \end{aligned}$$

With the scalar product defined by

$$(u, v)_h = u_0^{(2)} v_0^{(2)} \frac{h}{2} + \sum_{j=1}^{\infty} \langle u_j, v_j \rangle h,$$

we have

$$\frac{d}{dt} \|u\|_h^2 = 2(u, \frac{d\tilde{u}}{dt})_h = 2(u, Qu)_h,$$

i.e., we are back to the standard case, where Q must be shown to be semibounded. We have

$$(u, Qu)_h = -\frac{1}{2} u_0^{(2)} (u_1^{(1)} - u_0^{(1)}) - \frac{1}{2} \sum_{j=1}^{\infty} \left(u_j^{(1)} (u_{j+1}^{(2)} - u_{j-1}^{(2)}) + u_j^{(2)} (u_{j+1}^{(1)} - u_{j-1}^{(1)}) \right) = 0.$$

Hence, the difference operator Q is semibounded in the function space where $v_0^{(1)} = 0$ and $\|v\|_h < \infty$, and it follows that the approximation is stable.

We shall next investigate the Kreiss condition. The Laplace transform technique allows for differential operators in the boundary condition, since d/dt is substituted by the complex scalar s , and we can therefore use the original formulation (2.63). The eigenvalue problem corresponding to the Laplace transformed system is

$$\begin{aligned}\tilde{s}\phi_j + hAD_0\phi_j &= 0, \quad j = 1, 2, \dots, \\ \phi_0^{(1)} &= 0, \\ \tilde{s}\phi_0^{(2)} + \phi_1^{(1)} - \phi_0^{(1)} &= 0, \\ \|\phi\|_h < \infty.\end{aligned}$$

The characteristic equation corresponding to the first equation is

$$\text{Det} \begin{bmatrix} 2\tilde{s}\kappa & \kappa^2 - 1 \\ \kappa^2 - 1 & 2\tilde{s}\kappa \end{bmatrix} = 0,$$

and the roots

$$\begin{aligned}\kappa_1 &= -\tilde{s} + \sqrt{1 + \tilde{s}^2}, \\ \kappa_2 &= \tilde{s} - \sqrt{1 + \tilde{s}^2}\end{aligned}$$

are the ones inside the unit circle for $\text{Re } \tilde{s} > 0$. The solution is

$$\phi_j = \sigma_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \kappa_1^j + \sigma_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \kappa_2^j, \quad (2.64)$$

which is introduced in the boundary conditions. A nontrivial solution ϕ is obtained if the system

$$\begin{bmatrix} 1 & 1 \\ \tilde{s} + \kappa_1 - 1 & -\tilde{s} + \kappa_2 - 1 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = 0$$

has a nonzero solution. The coefficient matrix is singular if and only if either $\tilde{s} = i$ corresponding to $\kappa_1 = -\kappa_2 = -i$, or $\tilde{s} = -i$ corresponding to $\kappa_1 = -\kappa_2 = i$. Apparently, there are two generalized eigenvalues, and the Kreiss condition is not satisfied.

Recall that the Kreiss condition implies strong stability, and therefore the fact that it is violated here, is not a contradiction. Semibounded operators lead to stability, but not necessarily to strong stability. On the other hand, this problem is an example of a very special case, and we shall further investigate this.

The Laplace transformed variable \hat{u}_j satisfies the boundary conditions

$$\begin{aligned}\hat{u}_0^{(1)} &= \hat{g}^{(1)}, \\ \tilde{s}\hat{u}_0^{(2)} + \hat{u}_1^{(1)} - \hat{u}_0^{(1)} &= \hat{g}^{(2)},\end{aligned}$$

and we want an estimate of $|\hat{u}_j|$ in terms of $|\hat{g}| = |[\hat{g}^{(1)} \hat{g}^{(2)}]^T|$, where $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ are the Laplace transforms of g and $hF^{(2)}$ respectively. The form of the solution is

(2.64) with ϕ substituted by \hat{u} , and the coefficients σ_μ are determined by

$$\begin{bmatrix} 1 & 1 \\ \tilde{s} + \kappa_1 - 1 & -\tilde{s} + \kappa_2 - 1 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \hat{g},$$

where the coefficient matrix C has the determinant

$$\text{Det } C(\tilde{s}) = -2\sqrt{1 + \tilde{s}^2}.$$

We know already from the discussion above that $\text{Det } C(\pm i) = 0$, and we shall take a closer look at the behavior near these two singular points. For convenience we perturb \tilde{s} by a real and positive small number such that $\tilde{s} = i + \delta$. Then

$$|\text{Det } C(\tilde{s})| = 2\sqrt{2\delta} + \mathcal{O}(\delta),$$

which leads to the estimate

$$|\hat{u}_j| \leq \frac{K}{\sqrt{\delta}} |\hat{g}| = \frac{K}{\sqrt{Re \tilde{s}}} |\hat{g}|$$

for some constant K . The other singular point $-i$ leads to the same type of estimate.

If the Kreiss condition is satisfied, then $|\text{Det } C(\tilde{s})|$ is bounded from below also at the imaginary axis. If this condition is violated, the normal behavior is that $|\text{Det } C(\tilde{s})| \approx Re \tilde{s}$, and the approximation is unstable in any reasonable sense. In our example we have the exceptional case that the determinant tends to zero as $\sqrt{Re \tilde{s}}$ for small $Re \tilde{s}$. It can be shown that even in the more general case, the slower growth of the coefficients σ_μ as $Re \tilde{s} \rightarrow 0$, leaves the possibility of a stable approximation, but not a strongly stable one.

The fact that we get a determinant of the size $\sqrt{Re \tilde{s}}$ is due to the fact that κ_1 and κ_2 are both double roots of the characteristic equation at the singular points. The root κ_1 coincides with $\kappa_3 = -\tilde{s} - \sqrt{1 + \tilde{s}^2}$, and κ_2 coincides with $\kappa_4 = \tilde{s} + \sqrt{1 + \tilde{s}^2}$. Note that κ_3 and κ_4 are the ones outside the unit circle for $Re \tilde{s} > 0$, and they don't interfere with the general form (2.64) of the solution, which is based on distinct roots κ_μ .

Another case, where we can expect a well behaving solution, even when the Kreiss condition is not satisfied, is where there is a true eigenvalue \tilde{s}_0 on the imaginary axis, i.e., all the corresponding $\kappa_\mu(\tilde{s}_0)$ are strictly inside the unit circle. In this case the better behavior is an effect of the form of the solution. Even if the coefficients σ_μ are unbounded for \tilde{s} near \tilde{s}_0 , the component $\kappa_\mu^j(\tilde{s}_0)$ decreases exponentially as j increases.

The whole theory presented in this section has been developed for constant coefficients in the PDE and in the difference approximation. Real problems have in most cases variable coefficients, and fortunately the theory can be generalized to this case. The coefficients are frozen just as demonstrated for the periodic case in Section 2.2. To get the boundary conditions right, the coefficients are frozen at the values they have at the boundaries x_0 and x_N respectively.

The normal mode analysis is a more powerful method for stability analysis than the energy method. In order to demonstrate this, we shall analyze another example. In applications it is quite common that one uses different approximations in different subdomains. For example, if the solution is discontinuous, one might want to add a dissipative term to an otherwise norm conserving method in a local subdomain around the discontinuity. One question then is if the sudden change of method introduces some kind of instability.

As a model problem, we consider the equation $u_t = u_x$ with the standard second order method applied for $x \geq 0$, and with added dissipation for $x < 0$:

$$\begin{aligned}\frac{du_j}{dt} &= D_0 u_j, \quad j = 0, 1, \dots, \\ \frac{du_j}{dt} &= (D_0 + \alpha h D_+ D_-) u_j, \quad j = -1, -2, \dots,\end{aligned}$$

Here $\alpha > 0$, and we note that the special value $\alpha = 1/2$ changes the method to a pure upwind method

$$\frac{du_j}{dt} = D_+ u_j.$$

When working with the energy method, it seems natural to use the standard scalar product and norm. We use the notation

$$(u, v)_{r,s} = \sum_{j=r}^s u_j v_j h, \quad \|u\|_{r,s}^2 = \sum_{j=r}^s |u_j|^2 h,$$

and get

$$\frac{1}{2} \frac{d}{dt} \|u\|_{-\infty, \infty}^2 = (u, D_0 u)_{-\infty, \infty} + (u, \alpha h D_+ D_- u)_{-\infty, -1} = (u, \alpha h D_+ D_- u)_{-\infty, -1}.$$

The question now is whether or not this expression can be positive for any grid function u_j , and the answer is yes. Let $u_j = 0$, $j = -2, -3, \dots$. Then

$$(u, h D_+ D_- u)_{-\infty, -1} = u_{-1} (u_0 - 2u_{-1}),$$

which is positive if $u_{-1} > 0$ and $u_0 > 2u_{-1}$. Since point values u_0, u_{-1} cannot be estimated in terms of $\|u\|_{-\infty, \infty}$, the energy method doesn't work with this norm. This does not imply that the scheme is unstable, it just says that the norm we have used here may have a local growth in time. Instead of looking for another scalar product that gives a semibounded operator, we shall now prove that the method is stable by using the normal mode analysis.

For convenience, we reformulate the problem as an initial–boundary value problem by so called folding and define a new variable $v(x, t) = u(-x, t)$ with boundary condition $v(0, t) = u(0, t)$. On the discrete side we get

$$\begin{aligned}\frac{du_j}{dt} &= D_0 u_j, \quad j = 0, 1, \dots, \\ \frac{dv_j}{dt} &= -(D_0 - \alpha h D_+ D_-) v_j, \quad j = 1, 2, \dots, \\ u_0 &= v_0, \\ u_{-1} &= v_{-1}.\end{aligned}$$

In order to check the Kreiss condition, we look for eigenvalues \tilde{s} of the problem

$$\begin{aligned}2\tilde{s}\phi_j &= \phi_{j+1} - \phi_{j-1}, \\ 2\tilde{s}\psi_j &= -(\psi_{j+1} - \psi_{j-1} - 2\alpha(\psi_{j+1} - 2\psi_j + \psi_{j-1})), \\ \phi_0 &= \psi_0, \\ \phi_{-1} &= \psi_{-1}.\end{aligned}$$

Since the coupling between the variables is limited to the boundary, we get the general form of the solution as

$$\begin{aligned}\phi_j &= \sigma \kappa^j, \\ \psi_j &= \tau \mu^j,\end{aligned}$$

where κ and μ are the roots of the characteristic equations

$$\begin{aligned}2\tilde{s}\kappa &= \kappa^2 - 1, \\ 2\tilde{s}\mu &= -(\mu^2 - 1 - 2\alpha(\mu - 1)^2).\end{aligned}$$

(For the pure upwind case $\alpha = 1/2$ there is only one root of the second equation.) The boundary conditions imply

$$\begin{aligned}\sigma - \tau &= 0, \\ \kappa^{-1}\sigma - \mu^{-1}\tau &= 0,\end{aligned}$$

and the requirement of a nontrivial solution enforces the condition $\kappa = \mu$. When plugging that into the characteristic equations, and subtracting one from the other, we end up with the equation

$$(\kappa - 1)(\kappa + 1) - \alpha(\kappa - 1)^2 = 0,$$

which has the two roots $\kappa = 1$ and $\kappa = -(1 + \alpha)/(1 - \alpha)$. But $\kappa = 1$ implies $\tilde{s} = 0$, and we know from the form of κ above, that this is a contradiction. ($\kappa = 1$ corresponds to the other root of the characteristic equation, which is outside the unit circle for $\operatorname{Re} \tilde{s} > 0$.) The other possible solution $\kappa = -(1 + \alpha)/(1 - \alpha)$ does not give rise to any instability, since $|\kappa| > 1$ contradicts our assumption. Since the extra conditions of Theorem 2.7 are also satisfied, we have proven strong stability.

The case where the dissipation term is applied for $x > 0$ instead of $x < 0$, is treated in exactly the same manner, and it follows that the Kreiss condition is satisfied also in this case.

2.4.2 Fully Discrete Approximations

We begin by considering one step schemes

$$\begin{aligned} u_j^{n+1} &= Qu_j^n + kF_j^n, \quad j = 1, 2, \dots, \\ B_h u_0^n &= g^n, \\ u_j^0 &= f_j, \quad j = 1, 2, \dots, \\ ||u^n||_h &< \infty. \end{aligned} \tag{2.65}$$

Also here Q has the form

$$Q = \sum_{\nu=-r}^p A_\nu E^\nu, \tag{2.66}$$

but the matrices A_ν are of course not the same as in the semidiscrete case (2.48). The boundary operator B_h connects a fixed number of points in the neighborhood of (x_j, t_n) . It is assumed that $k/h = \lambda = \text{const}$, such that $Q = Q(\lambda)$.

The Laplace transform requires that the function is defined for all t , and therefore we have to extend the definition of u_j^n . The function

$$u_j(t) = u_j^n, \quad t_n \leq t < t_{n+1}$$

is piecewise constant. It is defined everywhere, and the Laplace transform is

$$\hat{u}_j(s) = \int_0^\infty e^{-st} u_j(t) dt.$$

The properties that are derived for $u_j(t)$, can immediately be transferred to u_j^n via the identity

$$\int_0^\infty |u_j(t)|^2 dt = \sum_{n=0}^\infty |u_j^n|^2 h.$$

We need to know the effect of a shift in time on the Laplace transform. If $f = 0$, we have

$$\int_0^\infty e^{-st} u_j(t+k) dt = \int_k^\infty e^{-s(t-k)} u_j(t) dt = e^{sk} \int_0^\infty e^{-st} u_j(t) dt,$$

i.e.,

$$\mathcal{L}u_j(t+k) = e^{sk} \mathcal{L}u_j(t).$$

The grid functions F_j^n and g^n are extended in the same way as u_j^n . The transformed system (2.65) is for $f = 0$

$$\begin{aligned} e^{sk}\hat{u}_j &= Q\hat{u}_j + k\hat{F}_j, \quad j = 1, 2, \dots, \\ \hat{B}_h\hat{u}_0 &= \hat{g}, \\ \|\hat{u}\|_h &< \infty. \end{aligned} \tag{2.67}$$

Note that the Laplace transform of $B_h u$ is in general different from $\hat{B}_h \hat{u}$, but in our case there are no time dependent coefficients. On the other hand, we must substitute \hat{B}_h for B_h , since the boundary conditions may include u_j^n at different time levels.

The complex number s occurs as the combination e^{sk} everywhere, and we introduce $z = e^{sk}$. As before, we need the solution to be well defined for all s with $\operatorname{Re} s \geq \eta_0 > 0$, and since k is arbitrarily small, the unit circle $|z| = 1$ will play the same role as the imaginary axis $\operatorname{Re} \tilde{s} = 0$ did for the semidiscrete case. With

$$\tilde{u}_j(z) = \hat{u}_j(s), \quad \tilde{F}_j(z) = \hat{F}_j(s), \quad \tilde{g}(z) = \hat{g}(s), \quad \tilde{B}_h(z) = \hat{B}_h(s),$$

the system (2.67) takes the form

$$\begin{aligned} z\tilde{u}_j &= Q\tilde{u}_j + k\tilde{F}_j, \quad j = 1, 2, \dots, \\ \tilde{B}_h\tilde{u}_0 &= \tilde{g}, \\ \|\tilde{u}\|_h &< \infty. \end{aligned} \tag{2.68}$$

This is actually the *z-transform* of (2.65), and we could as well have used the theory for this transform when developing the stability theory for the fully discrete problem. However, the Laplace transform is used here since the proper tools for this technique have already been introduced.

The eigenvalue problem corresponding to (2.68) is

$$\begin{aligned} z\phi_j &= Q\phi_j, \quad j = 1, 2, \dots, \\ \tilde{B}_h\phi &= 0, \\ \|\phi\|_h &< \infty, \end{aligned} \tag{2.69}$$

and we have

Definition 2.21. The *Godunov–Ryabenkii condition* is satisfied if the eigenvalues z of the problem (2.69) satisfy $|z| > 1$. \square

This condition is necessary for stability of the difference approximation (2.64).

The technique for verifying the Godunov–Ryabenkii condition is exactly the same as for the semidiscrete case, except that the domain $\operatorname{Re} \tilde{s} > 0$ is substituted by $|z| > 1$. Here follows a short summary. We look for nontrivial vector solutions ψ of

$$(zI - \sum_{\nu=-r}^p A_\nu E^\nu)\psi = 0, \quad \psi = \begin{bmatrix} \psi^{(1)} \\ \psi^{(2)} \\ \vdots \\ \psi^{(m)} \end{bmatrix}$$

for $|z| > 1$. This requires solution of the characteristic equation

$$\operatorname{Det}(zI - \sum_{\nu=-r}^p A_\nu \kappa^\nu) = 0. \quad (2.70)$$

We have in complete analogy with Lemma 2.3

Lemma 2.4. *Assume that the von Neumann condition is satisfied for the periodic problem. If $|z| > 1$, then*

- i) *There are no roots κ_μ of (2.70) with $|\kappa_\mu| = 1$.*
- ii) *If A_{-r} is nonsingular, then there are exactly rm roots κ_μ with $|\kappa_\mu| < 1$.*

□

The solution ϕ_j has the form

$$\phi_j = \sum_{|\kappa_\mu| < 1} P_\mu(j) \kappa_\mu^j, \quad (2.71)$$

where $P_\mu(j)$ are polynomials in j with vector coefficients, containing mr constants σ_μ . The boundary conditions imply

$$C(z)\boldsymbol{\sigma} = \mathbf{0}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{mr} \end{bmatrix} \neq \mathbf{0}, \quad (2.72)$$

and the Godunov–Ryabenkii is equivalent to

$$\operatorname{Det} C(z) \neq 0, \quad |z| > 1.$$

The z -transformed system for $F = 0$ and $f = 0$ is

$$\begin{aligned} z\tilde{u}_j &= Q\tilde{u}_j, \quad j = 1, 2, \dots, \\ \tilde{B}_h &= \tilde{g}, \\ \|\tilde{u}\|_h &< \infty. \end{aligned} \quad (2.73)$$

The Kreiss condition is introduced in analogy with the semidiscrete case. The first version requires an estimate of \tilde{u} in terms of $|\tilde{g}|$. When considering the eigenvalue problem (2.69), we have also here the possibility of generalized eigenvalues:

Definition 2.22. Consider the matrix $C(z)$ in (2.72). If there is a complex number z_0 on the unit circle such that $\operatorname{Det} C(z_0) = 0$, and if there is at least one root $\kappa_\mu(z_0)$ of the characteristic equation with $|\kappa_\mu(z_0)| = 1$, then z_0 is called a *generalized eigenvalue*. □

We have

Definition 2.23. The Kreiss condition is satisfied if either one of the following equivalent conditions is satisfied.

- i) The solutions of (2.73) satisfy for every fixed j

$$|\tilde{u}_j(z)| \leq K|\tilde{g}|, \quad |z| > 1,$$

where K is a constant independent of z .

- ii)

$$\text{Det } C(z) \neq 0, \quad |z| \geq 1.$$

(The *determinant condition*)

- iii) The problem (2.69) has no eigenvalue or generalized eigenvalue outside or on the unit circle $|z| \geq 1$. \square

By using Parseval's relation and going back to the grid function u_j^n one can prove

Theorem 2.8. *If the Kreiss condition is satisfied, then there is a constant K_1 such that the solutions of (2.68) with $\tilde{F} = 0$ satisfy the estimate*

$$\sum_{\nu=1}^n |u_j^\nu|^2 k \leq K_1 \sum_{\nu=1}^n |g^\nu|^2 k$$

for any fixed j . \square

Let us next briefly discuss general multistep schemes

$$\begin{aligned} Q_{-1} u_j^{n+1} &= \sum_{\sigma=0}^q Q_\sigma u_j^{n-\sigma} + k F_j^n, \quad j = 1, 2, \dots, \\ B_h u_0^n &= g^n, \\ u_j^\sigma &= f_j^\sigma, \quad j = 1, 2, \dots, \quad \sigma = 0, 1, \dots, q, \\ \|u^n\|_h &< \infty, \end{aligned}$$

where

$$Q_\sigma = \sum_{\nu=-r}^p A_\nu^{(\sigma)} E^\nu.$$

The z -transformed system for $F = 0$, $f = 0$ is also in this case obtained by formally putting $u_j^n = z^n \tilde{u}_j$ and then dividing by z^n :

$$\begin{aligned} Q_{-1} z \tilde{u}_j &= \sum_{\sigma=0}^q Q_\sigma z^{-\sigma} \tilde{u}_j, \quad j = 1, 2, \dots, \\ \tilde{B}_h \tilde{u}_0 &= \tilde{g}, \\ \|\tilde{u}\|_h &< \infty. \end{aligned}$$

At this point we have a system of difference equations in space, and therefore everything that follows is in principle exactly the same as for one-step schemes. The characteristic equation is

$$\text{Det} \left(\sum_{\nu=-r}^p (z A_\nu^{(-1)} - \sum_{\sigma=0}^q z^{-\sigma} A_\nu^{(\sigma)}) \kappa^\nu \right) = 0,$$

and we arrive at the Godunov–Ryabenkii and the Kreiss conditions in the same way as above.

As an illustration, we consider the *leap-frog scheme* for the equation $u_t = u_x$:

$$\begin{aligned} u_j^{n+1} - u_j^{n-1} &= \lambda (u_{j+1}^n - u_{j-1}^n), \quad j = 1, 2, \dots, \\ u_0^n &= 2u_1^n - u_2^n, \\ \|u^n\|_h &< \infty, \end{aligned}$$

where it is assumed that $\lambda = k/h < 1$ such that the von Neumann condition is satisfied. Assuming that $|z| > 1$, the eigensolution has the form

$$\phi_j = \sigma_1 \kappa_1^j$$

where κ_1 is the root of the characteristic equation

$$(z^2 - 1)\kappa = \lambda z(\kappa^2 - 1), \quad (2.74)$$

which satisfies $|\kappa_1| < 1$. The boundary condition implies

$$(\kappa_1 - 1)^2 \sigma_1 = 0,$$

i.e., $C(z) = (\kappa_1 - 1)^2$. We know without solving the characteristic equation that $\kappa_1 \neq 1$ for $|z| > 1$. The remaining question is whether or not $\kappa_1 = 1$ for some z on the unit circle. The characteristic equation shows that there are two possibilities: $z = 1$ or $z = -1$. We know that one of the roots κ_μ is one, the question is whether it is κ_1 or κ_2 . This can be determined by perturbing z in the characteristic equation. We begin by investigating the point $z = -1$, and let $z = -(1 + \delta)$, where δ is real, positive and small. We get

$$2\delta\kappa = -\lambda(1 + \delta)(\kappa^2 - 1) + \mathcal{O}(\delta^2),$$

which has the roots

$$\kappa = \pm 1 - \frac{1}{\lambda}\delta + \mathcal{O}(\delta^2).$$

By definition, $|\kappa_1| < 1$ for $\delta > 0$. Consequently, $\kappa_1(-1) = 1$, and we have found a generalized eigenvalue $z = -1$. (For $z = 1$, it is shown in the same way that $\kappa_1(1) = -1$, i.e., $z = 1$ is not a generalized eigenvalue.) Recall that the semidiscrete approximation with the same boundary condition satisfies the Kreiss condition. The instability found here is an effect of the time discretization.

The extrapolation used here can be modified to the more general extrapolation

$$(hD_+)^p u_0^n = 0.$$

The only difference in the analysis is that we get $C(z) = (\kappa_1 - 1)^p$. For the z -transformed system we have the boundary condition

$$(\kappa_1 - 1)^p \sigma_1 = \tilde{g},$$

i.e.,

$$\tilde{u}_j = \frac{\tilde{g}}{(\kappa_1 - 1)^p} \kappa_1^j.$$

This shows that with extrapolation of higher order p , the singularity at $z = -1$ becomes more severe, and we should expect a stronger instability. We have the somewhat unusual situation here, that higher order accuracy leads to a worse solution. This is of course no contradiction, stability is a concept by itself, and is in principle not connected to the order of accuracy.

We ran the leap-frog scheme for the two cases $p = 1$ and $p = 2$ for the solution $u(x, t) = \sin |\pi(x+t)/2|$ in the interval $-1 \leq x \leq 1$. Figure 2.1 shows the solution at the boundary $x_0 = -1$ for $N = 80$. The first order extrapolation $u_0^n = u_1^n$ is unstable, but it is still possible to recognize the true shape of the solution $u(x, t) = \sin |\pi(t-1)/2|$. For the second order extrapolation $u_0^n = 2u_1^n - u_2^n$, the instability is much more severe, as expected.

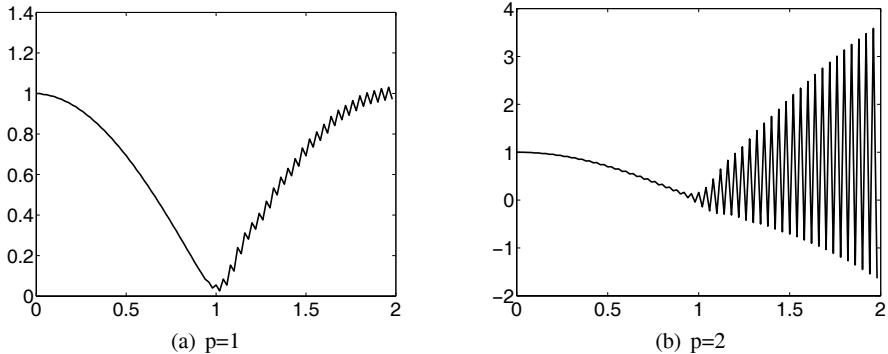


Fig. 2.1 u_0^n for the leap-frog scheme, $0 \leq t_n \leq 2$, $N = 80$

The boundary condition results in a nonlinear system of equations for z and κ , which can be called the characteristic equation for the boundary conditions. This equation, together with the characteristic equation for the difference scheme at inner points, constitutes the complete system that is the basis for the analysis. The strength with the normal mode analysis is that one can use this system not only for checking a certain given boundary condition, but also for constructing new stable boundary conditions. In our example we saw that the combination $z = -1$, $\kappa = 1$ gave a generalized eigenvalue. If we could force the condition $z = \kappa$ to hold, then we would be in a better situation. With the boundary condition

$$u_0^{n+1} = 2u_1^n - u_2^n \quad (2.75)$$

corresponding to extrapolation in the x/t direction, we get the equation

$$z^2 - 2z\kappa + \kappa^2 = 0,$$

or equivalently

$$(z - \kappa)^2 = 0.$$

With $z = \kappa$ in the characteristic equation (2.74) we get the only possibilities $z = \pm 1$ for $\lambda < 1$. But this case is already treated above, and we know that the condition $z = \kappa_1$ cannot hold. The conclusion is that the Kreiss condition is satisfied.

In the same way it is easily shown that the boundary condition

$$u_0^{n+1} = u_0^n + \lambda(u_1^n - u_0^n) \quad (2.76)$$

satisfies the Kreiss condition.

Figure 2.2 shows the solution for these two stable boundary conditions.

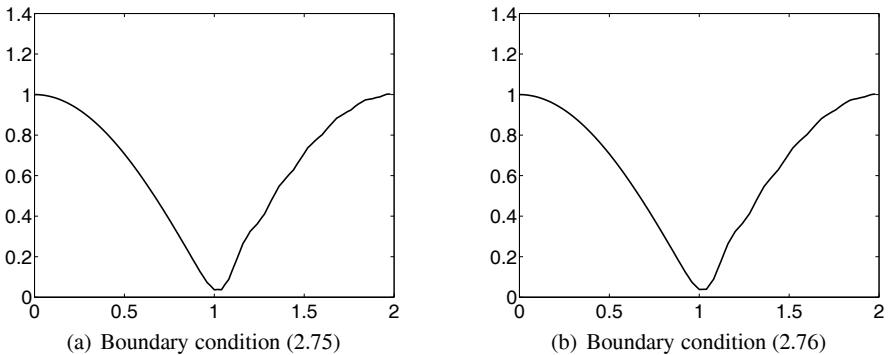


Fig. 2.2 u_0^n for the leap-frog scheme, $0 \leq t_n \leq 2$, $N = 80$

The stability definition connected to the Kreiss condition is a natural extension of the semidiscrete version:

Definition 2.24. Assume zero initial data f (or $f^{(\sigma)}$ in the multistep case). Then the approximation is *stable in the generalized sense* if there are constants K and η_0 such that the solution satisfies

$$\sum_{n=1}^{\infty} e^{-\eta t_n} \|u^n\|_h^2 k \leq K(\eta) \sum_{n=1}^{\infty} e^{-\eta t_n} \|F^{n-1}\|_h^2 k,$$

for all $\eta \geq \eta_0$, where $K(\eta) \rightarrow 0$ as $\eta \rightarrow \infty$.

It is *strongly stable in the generalized sense* if

$$\sum_{n=1}^{\infty} e^{-\eta t_n} \|u^n\|_h^2 k \leq K(\eta) \sum_{n=1}^{\infty} e^{-\eta t_n} \left(\|F^{n-1}\|_h^2 + |g^{n-1}|^2 \right) k,$$

□

Under certain technical assumptions on the difference scheme, it is possible to show that the Kreiss condition is equivalent to strong stability in the generalized sense, see [Gustafsson et al., 1972].

For symmetric hyperbolic systems, there is a theorem analogous to the semidiscrete version (2.7). We need an energy estimate for the Cauchy problem

$$u_j^{n+1} = Qu_j^n, \quad j = 0, \pm 1, \pm 2, \dots,$$

where we require that $u_j^n \rightarrow 0$ sufficiently fast as $j \rightarrow \pm\infty$ such that the norm

$$\|u^n\|_{-\infty, \infty} = \left(\sum_{j=-\infty}^{\infty} |u^n|^2 h \right)^{1/2}$$

exists. We have

Theorem 2.9. *Assume that there is an energy estimate*

$$\|u^{n+1}\|_{-\infty, \infty}^2 \leq \|u^n\|_{-\infty, \infty}^2$$

for the Cauchy problem and that the Kreiss condition is satisfied. If the coefficient matrices in (2.66) can be simultaneously diagonalized and $r \geq p$, then (2.65) is strongly stable. □

(Strong stability is defined in Definition 2.15 with obvious modification of the range of j .)

2.5 Summary

In this chapter we have given a fairly comprehensive survey of the stability theory.

The classical von Neumann condition based on Fourier analysis is necessary for stability in any reasonable sense. This is often the only condition (if any) that is checked for real world problems. Already for periodic problems it is necessary to add some other conditions to be sure about stability. Such a condition is the dissipativity concept, which was introduced by Kreiss [Kreiss, 1964].

The real challenge is to construct numerical boundary conditions such that the initial-boundary value problem is stable. For these problems, there are two fundamentally different methods of analyzing stability, namely the energy method and the normal mode analysis. The principle with the first one, is to construct a norm which does not grow from one time level to the next, if the forcing function F and

the boundary data g are zero. The drawback is that if one fails to construct such a norm, it is not clear whether or not the scheme is stable. Perhaps one has not been clever enough in the search for a proper norm.

The normal mode analysis is based on a Laplace transformation in time of the approximation. The resulting difference scheme is a boundary value problem, and the crucial question is if it satisfies the Kreiss condition, which is a condition on a certain matrix of small size. This condition can often be investigated analytically, and if it is satisfied, then the scheme is stable in a norm defined over space and time that naturally occurs with the Laplace transform. Under certain conditions, an estimate is obtained in the same type of norm over space as for the energy method. If the Kreiss condition is not satisfied, then the scheme is usually unstable, but there are examples where a weaker form of stability still holds.

In this book we have presented the normal mode analysis only for the 1-D case. However, it can be applied to problems in d space dimensions as well by Fourier transforming in the x_2, \dots, x_d directions. This again gives rise to a 1-D difference equation, but now containing $d - 1$ wave numbers $\omega_2, \dots, \omega_d$ as parameters. The Kreiss condition is now defined as for 1-D, but the estimates have to be independent of these parameters. The theoretical basis for multidimensional problems was given by Michelson in [Michelson, 1983].

As we have seen, there are a number of different stability definitions for initial-boundary value problems, and this may seem a little confusing to somebody with a real problem at hand. However, the Kreiss condition is the fundamental concept, and our experience is, that if this condition is satisfied for the linearized problem with constant coefficients, then the approximation is behaving very well even for nonlinear problems with solutions that are well resolved by the grid.

Many papers have been written on how to verify the Kreiss condition. Goldberg and Tadmor derived a series of easy to use criteria in [Goldberg and Tadmor, 1981], [Goldberg and Tadmor, 1985], [Goldberg and Tadmor, 1987], [Goldberg and Tadmor, 1989] and [Goldberg, 1991]. One of the results found there is the following: Assume that the scalar approximation

$$u_j^{n+1} = \sum_{\nu=-r}^p u_{j+\nu}^n, \quad j = 1, 2, \dots, \quad (2.77)$$

is stable for the Cauchy problem. Then the Kreiss condition is satisfied with the boundary conditions

$$u_\nu^{n+1} = g_\nu^{n+1}, \quad j = -r + 1, -r + 2, \dots, 0.$$

This means that one can always prescribe all the necessary data pointwise, regardless of the direction of the characteristic. Of course it may be difficult to find these data for an outflow problem (outgoing characteristic), but the trouble is the accuracy, not the stability.

Another general result is that for a dissipative scalar one-step scheme (2.77), one can always use extrapolation

$$(hD_+)^{\mu} u_{\nu}^{n+1} = 0, \quad j = -r+1, -r+2, \dots, 0$$

of any order μ . This result was first proven in a different way by Kreiss [Kreiss, 1968].

From a historical point of view, there are a few outstanding names in the development of stability theories. One of them was John von Neumann, and in his paper [von Neumann, 1944] the von Neumann condition was introduced in analogy with the Petrovski condition for PDE. Another one is Peter Lax, who wrote a number of important papers on stability in the fifties and sixties. Perhaps the most well known is the paper [Lax and Richtmyer, 1956] with Richtmyer, where the famous *Lax equivalence theorem* (sometimes called the Lax–Richtmyer equivalence theorem) was presented, see also [Richtmyer and Morton, 1967]. H.-O. Kreiss proved the *Matrix Theorem* in [Kreiss, 1962], which provides necessary and sufficient conditions on the matrix \hat{Q} , such that all powers \hat{Q}^n are bounded. Another important paper by Kreiss is [Kreiss, 1964] mentioned in Section 2.2.2, where sufficient conditions for stability of dissipative approximations are given.

The initial–boundary value problem was taken up early by the Russian school, where Godunov was one of the central individuals, and the normal mode analysis was introduced 1963 in [Godunov and Ryabenkii, 1963] leading to the Godunov–Ryabenkii condition for the fully discrete case. The same name has later been given to the analogous condition for semidiscrete approximations. In the sixties, Kreiss began working on this type of problems, and he presented the first complete theory for dissipative approximations in [Kreiss, 1968], with a follow up by Osher [Osher, 1969]. The general stability theory based on the Laplace transform technique (in the form of the z -transform) was given in [Gustafsson et al., 1972] leading to the Kreiss condition. This paper gave rise to the concept *GKS theory*, which is sometimes used as a synonym for normal mode analysis. The paper is an extension to the discrete case of the theory in the famous paper [Kreiss, 1970] by Kreiss, where he solved the problem of well-posedness for hyperbolic initial–boundary value PDE problems. The theorems in these papers give estimates of the solution in a norm integrated over time. Theorem 2.7 was proven in [Gustafsson et al., 1995], and it gives an estimate in the more traditional l_2 -norm in space. Another view was given by Trefethen in [Trefethen, 1983], where he related the Kreiss condition to the group velocity. The same author took still another approach, when he developed a new theory based on pseudo-eigenvalues, see [Reddy and Trefethen, 1990] and [Reddy and Trefethen, 1992]. These results can be seen as a generalization of the Kreiss matrix theorem.

In this book we have presented the normal mode analysis by treating the semidiscretized problem before the fully discretized one, but historically, it was developed in the opposite order. The semidiscrete case was actually treated first by Strikwerda in [Strikwerda, 1980] as a follow up on the Kreiss theory for the continuous and fully discrete case. However, in [Gustafsson et al., 1995] a different technique was used such that the estimate of the solution is obtained without involving the time integrals, leading to strong stability.

Chapter 3

Order of Accuracy and the Convergence Rate

In Chapter 1 comparisons were made between methods with different orders of accuracy, where the order is defined in terms of the truncation error. This error is a measure of how well the difference scheme is approximating the differential equation, but by itself it doesn't tell how well the *solutions* of the difference scheme approximates the *solutions* of the differential equation. This requires stability, and we have seen in the previous chapter, that there are several different stability concepts. Furthermore, the truncation error and the order of accuracy has to be defined very precisely when it comes to boundary conditions.

When discussing the truncation error, it is usually assumed that the solution of the differential equation is sufficiently smooth. By this we mean that the highest order derivatives of the solution $u(x, t)$ occurring in the truncation error are bounded. This is a quite strong condition, in particular for initial-boundary value problems in several space dimensions, and in practice it is often violated. We have already seen in Chapter 1 that higher order methods do a better job also for problems with non-smooth solutions. In Chapter 9 we shall further demonstrate this by presenting an example from acoustics. However, if not stating otherwise, smoothness is assumed in this chapter.

3.1 Periodic Solutions

In this section we consider problems with 2π -periodic solutions

$$\begin{aligned}\frac{\partial u}{\partial t} &= P(\partial/\partial x)u + F(x, t), \\ u(x, 0) &= f(x),\end{aligned}\tag{3.1}$$

and difference approximations

$$\begin{aligned} u_j^{n+1} &= Qu_j^n + kF_j^n, \quad j = 0, 1, \dots, N, \\ u_j^0 &= f_j, \quad j = 0, 1, \dots, N, \end{aligned} \tag{3.2}$$

where $(N+1)h = 2\pi$. Here it is assumed that if the difference scheme is a multi-step method, it has been rewritten as a one step scheme. For convenience, we use the original notation u_j^n even in that case, but it should be interpreted as the vector representing several time levels. For a two step scheme, it is the approximation of $[u(x_j, t_{n+1}) \ u(x_j, t_n)]^T$, and similarly for F and f . Assuming that the solution $u(x, t)$ is smooth, we plug it into the difference approximation (3.2), and obtain

$$\begin{aligned} u(x_j, t_{n+1}) &= Qu(x_j, t_n) + kF_j^n + k\tau(x_j, t_n), \quad j = 0, 1, \dots, N, \\ u(x_j, 0) &= f_j + \phi_j, \quad j = 0, 1, \dots, N, \end{aligned} \tag{3.3}$$

where again $u(x, t)$ represents the solution at several time levels in the multi-step case. Even if there may be an error ϕ in the initial data, the order of accuracy is defined in terms of the main difference approximation:

Definition 3.1. The function $\tau(x_j, t_n)$ is called the *truncation error*. The *order of accuracy* is (p, q) if

$$|\tau(x_j, t_n)| \leq K(h^p + k^q),$$

where K is a constant independent of h and k . The approximation is *consistent* if $p > 0, q > 0$. \square

Note that the difference scheme is written in a normalized form such that it approximates

$$k \frac{\partial u}{\partial t} = kP(\partial/\partial x) + kF(x, t),$$

i.e., $Q = I + kQ_1$ in the one step case. Hence there is an extra factor k multiplying the truncation error τ . Sometimes $k\tau$ is called the *local truncation error*.

By subtracting (3.2) from (3.3) we obtain the *error equation*

$$\begin{aligned} w_j^{n+1} &= Qw_j^n + k\tau_j^n, \quad j = 0, 1, \dots, N, \\ w_j^0 &= \phi_j, \quad j = 0, 1, \dots, N \end{aligned}$$

for the error $w_j^n = u(x_j, t_n) - u_j^n$. An error estimate is now easily obtained from the stability definition:

Theorem 3.1. Consider the difference approximation (3.2) on a finite time interval $0 \leq t_n \leq T$, and assume that it is stable and accurate of order (p, q) . If the truncation error in the initial data satisfies $|\phi_j| = \mathcal{O}(h^p + k^q)$, then the error $w_j^n = u(x_j, t_n) - u_j^n$ satisfies the estimate

$$||w^n||_h \leq K(T)(h^p + k^q),$$

where $K(T)$ is a constant depending only on T . \square

The approximation of the initial function $f(x)$ is usually of infinite accuracy and causes no trouble. However, the condition on the accuracy of the initial data f_j is

important for multi-step schemes, since f_j represents the numerical solution at all the time levels required for getting the multi-step scheme started. As an example, consider the leap-frog scheme $u^{n+1} = 2kQ_2u^n + u^{n-1}$, which has order of accuracy $(2, 2)$ if Q_2 is a second order approximation of the differential operator $P(\partial/\partial x)$. Assume that the first time level is computed by the Euler forward method

$$u_j^1 = (I + kQ_2)u_j^0.$$

This approximation is first order accurate in time, but for the first time step we have

$$u(x_j, k) = u(x_j, 0) + kQ_2u(x_j, 0) + \mathcal{O}(kh^2 + k^2).$$

This means that if u_j^0 is exact, then the initial error satisfies

$$|u(x_j, k) - u_j^1| = \mathcal{O}(kh^2 + k^2).$$

Obviously the condition in the theorem on the initial data is satisfied for $p = q = 2$, and if the PDE is a first order hyperbolic system, then there is an error estimate $\mathcal{O}(h^2 + k^2)$ if k/h is small enough such that the leap-frog scheme scheme is stable. Note that the forward Euler scheme is unstable for hyperbolic problems, but for a single time step, that has no significance.

This is an example of a general observation: for multistep schemes, the necessary initial data can be generated by a method with one order lower accuracy in time. Actually, if $k/h = \text{const}$ in our example, we could have lowered the order of accuracy one step also in space, since there is an extra factor k multiplying the error in space. However, the relaxation of accuracy in time is more important, since it allows for generation of the first time level by a simple one-step scheme.

If the solution $u(x, t)$ is not smooth, we may still have convergence as the step sizes tend to zero, but the convergence rate deteriorates. For technical reasons, we must assume that the space and time steps are defined as sequences h_ν and k_ν tending to zero as the integer ν tends to infinity. We write $h \rightarrow 0$, $k \rightarrow 0$ for this procedure, and we make the formal definition

Definition 3.2. The method is *convergent* if the numerical solution u_j^n converges to the true solution $u(x, t)$, i.e., if for $0 \leq t \leq T$

$$\|u(\cdot, t) - u^{t/k}\|_h \rightarrow 0 \text{ as } h \rightarrow 0, k \rightarrow 0.$$

□

This definition doesn't contain any quantitative information about the accuracy of the solution. The practical interpretation is, that if a certain computation with a convergent difference scheme doesn't give sufficient accuracy, then one should expect a more accurate solution if the grid is refined.

There is a classic theorem connecting consistency, stability and convergence:

Theorem 3.2. (The Lax equivalence theorem)

Assume that the PDE problem is well posed and that the difference approximation is consistent. Then the approximation is convergent if and only if it is stable. \square

The proof can be found in [Richtmyer and Morton, 1967], and in a different form in [Gustafsson et al., 1995].

3.2 Initial–Boundary Value Problems

Here we consider a general initial–boundary value problem and the approximation

$$\begin{aligned} u_j^{n+1} &= Qu_j^n + kF_j^n, \quad j = 1, 2, \dots, N-1, \\ B_h u^n &= g^n, \\ u_j^0 &= f_j. \end{aligned} \tag{3.4}$$

The normalization of the approximation at inner points was discussed in the previous section. The definition of accuracy order of the boundary conditions is a little more complicated, and we begin by looking at the familiar example

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x \leq 1, \quad 0 \leq t \leq T, \\ u(1, t) &= g(t), \\ u(x, 0) &= f(x), \end{aligned}$$

and the Crank–Nicholson approximation

$$\begin{aligned} (I - \frac{k}{2}D_0)u_j^{n+1} &= (I + \frac{k}{2}D_0)u_j^n, \quad j = 1, 2, \dots, N-1, \\ u_0^n &= 2u_1^n - u_2^n, \\ \frac{1}{2}(u_N^n + u_{N-1}^n) &= g^n, \\ u_j^0 &= f_j. \end{aligned}$$

The boundary condition at the right is a strange one, but we choose it in order to illustrate the basic principles of accuracy at the boundary. Assuming that $k/h = \text{const}$, the truncation error at inner points is $\mathcal{O}(h^2)$, and we assume that the initial data f_j are exact. For the boundary condition at the left, we have for any smooth function

$$u(0, t_n) = 2u(h, t_n) - u(2h, t_n) + \mathcal{O}(h^2).$$

If the grid is located such that $(x_N + x_{N-1})/2 = 1$, then the true solution $u(x, t)$ satisfies

$$\frac{1}{2}(u(x_N, t_n) + u(x_{N-1}, t_n)) = g(t_n) + \mathcal{O}(h^2).$$

In Section 2.3.3 we showed that the scheme is stable, and it is easily shown that it is strongly stable as well. The error $w_j^n = u(x_j, t_n) - u_j^n$ satisfies

$$\begin{aligned} (I - \frac{k}{2}D_0)w_j^{n+1} &= (I + \frac{k}{2}D_0)w_j^n + \mathcal{O}(kh^2), \quad j = 1, 2, \dots, N-1, \\ w_0^n - 2w_1^n + w_2^n &= \mathcal{O}(h^2), \\ \frac{1}{2}(w_N^n + w_{N-1}^n) &= \mathcal{O}(h^2), \\ w_j^0 &= 0, \end{aligned}$$

and by the definition of strong stability it follows immediately that

$$\|w^n\|_h \leq K_1(T)h^2.$$

Let us next assume that the grid is located such that $x_N = 1$. The stability properties are not affected, since from a stability point of view, the scheme doesn't know the connection to the PDE and its boundary conditions. However, there is now a larger truncation error, since the true solution satisfies

$$\frac{1}{2}(u(x_N, t_n) + u(x_{N-1}, t_n)) = g(t_n) + \mathcal{O}(h).$$

The strong stability estimate for $\|u^n\|_h^2$ now contains a term $Ke^{\alpha T} \sum_{\nu=0}^{n-1} |g^\nu|^2 k$, where $|g^\nu| = \mathcal{O}(h)$, which leads to the final estimate

$$\|w^n\|_h \leq K_1(T)h.$$

This estimate is sharp, and it can be generalized to PDE of any order. From this we learn, that in order to get full accuracy corresponding to the approximation at the inner points, the boundary conditions of the PDE must be approximated to the same order.

Another question is what kind of truncation error order we associate with the extra numerical boundary condition at the left. It is written as a second order extrapolation of $u(x, t)$, but is not related to the PDE or to any boundary condition for the PDE. On the other hand, we recall from Section 2.3 that the approximation can be written in the following way:

$$\begin{aligned} (I - \frac{k}{2}D_0)u_j^{n+1} &= (I + \frac{k}{2}D_0)u_j^n, \quad j = 2, 3, \dots, N-1, \\ (I - \frac{k}{2}D_+)u_1^{n+1} &= (I + \frac{k}{2}D_+)u_1^n, \\ \frac{1}{2}(u_N^n + u_{N-1}^n) &= g^n, \\ u_j^0 &= f_j. \end{aligned}$$

The centered difference operator has been substituted by a one-sided operator at the boundary, and this can be considered as another kind of numerical boundary condition. For the true solution we have

$$(I - \frac{k}{2}D_+)u(h, t_{n+1}) = (I + \frac{k}{2}D_+)u(h, t_n) + k\mathcal{O}(k^2 + h),$$

i.e., the approximation of the *differential equation* is first order accurate. Consequently, if the order of accuracy is related to the differential equation, then the approximation is one order lower at the boundary. This is the background to the often used statement “one order lower accuracy at the boundary is allowed”. Recall that we had exactly the same situation for the initial conditions as described in the previous section. The initial and boundary conditions for the PDE must be approximated to the same order as the approximation at inner points. However, the accuracy for the extra initial and boundary conditions can be relaxed. In Section 7.1 we shall discuss the accuracy of boundary condition approximations also for higher order PDE.

The true normalization of the boundary conditions is of course determined by the condition of strong stability. With the boundary condition $B_h u^n = g^n$, there must be an estimate for $\|u^n\|_h^2$ containing the term $\sum_{\nu=0}^{n-1} |g^\nu|^2 k$. If the boundary condition is rewritten in the form $\tilde{B}_h u^n = \tilde{g}^n$, where $\tilde{B}_h = hB_h$ and $\tilde{g}^n = hg^n$ such that it looks like a one step higher order of accuracy, then the cheating is revealed when we try to prove strong stability. We would get an estimate in terms of $\sum_{\nu=0}^{n-1} |\tilde{g}^\nu / h|^2 k$, violating strong stability.

For the general difference scheme (3.4), we assume for convenience that there is a relation $k = k(h)$ such that the truncation error can be expressed in terms of h :

$$\begin{aligned} u(x_j, t_{n+1}) &= Qu(x_j, t_n) + kF_j^n + kh^{p_1}\tilde{F}_j^n, \quad j = 1, 2, \dots, N-1, \\ B_h u(x_j, t_n) &= g^n + h^{p_2}\tilde{g}^n, \\ u(x_j, 0) &= f_j + h^{p_3}\tilde{f}_j, \end{aligned} \tag{3.5}$$

where the grid functions $\tilde{F}, \tilde{g}, \tilde{f}$ are bounded independent of h . The error $w_j^n = u(x_j, t_n) - u_j^n$ satisfies

$$\begin{aligned} w_j^{n+1} &= Qw_j^n + kh^{p_1}\tilde{F}_j^n, \quad j = 1, 2, \dots, N-1, \\ B_h w^n &= h^{p_2}\tilde{g}^n, \\ w_j^0 &= h^{p_3}\tilde{f}_j. \end{aligned} \tag{3.6}$$

If the difference scheme is strongly stable as defined in Definition 2.15, then we get

$$\|w^n\|_h \leq K_1(T)h^{p_0}, \tag{3.7}$$

where $p_0 = \min(p_1, p_2, p_3)$.

In Sections 2.3 and 2.4 the concept of strong stability in the generalized sense was discussed. It applies to first order hyperbolic systems, and requires zero initial data. For one-step schemes it is applied without complications, since it is very natural that

$\tilde{f}_j = 0$. The error estimate is

$$\sum_{n=1}^{\infty} e^{-\eta t_n} \|w^n\|_h^2 k \leq K \sum_{n=1}^{\infty} e^{-\eta t_n} (\|h^{p_1} \tilde{F}^{n-1}\|_h^2 + |h^{p_2} \tilde{g}^{n-1}|^2) k \leq K_1 h^{2p_0},$$

where $p_0 = \min(p_1, p_2)$ and $\eta \geq \eta_0$.

As we have seen, strong stability leads directly to optimal error estimates. All other stability concepts for initial–boundary value problems requires some restriction on the data. This doesn't necessarily mean that the error estimates deteriorate. In Section 2.4.1 we discussed how to subtract a certain grid function $v_j(t)$, satisfying the initial condition, from the numerical solution $u_j(t)$ such that the problem is transformed to another one for $u - v$ with zero initial data. The same procedure can be applied to fully discrete approximations with obvious modifications. In order to derive error estimates, we use this procedure for the system (3.6) for the error w , and subtract a grid function v_j^n satisfying

$$v_j^0 = h^{p_3} \tilde{f}_j, \\ \sum_{n=1}^{\infty} e^{-\eta t_n} \|v^n\|_h^2 k \leq \mathcal{O}(h^{2p_0})$$

with p_0 defined in (3.7). Then the resulting system for $\phi = w - v$ is

$$\begin{aligned} \phi_j^{n+1} &= Q\phi_j^n + kh^{p_0} \tilde{G}_j^n, \quad j = 1, 2, \dots, N-1, \\ B_h \phi^n &= h^{p_0} \tilde{d}^n, \\ \phi_j^0 &= 0, \end{aligned}$$

where \tilde{G} and \tilde{d} are bounded independent of h . Strong stability in the generalized sense gives the proper estimate for ϕ , and the final estimate

$$\sum_{n=1}^{\infty} e^{-\eta t_n} \|w^n\|_h^2 k \leq \mathcal{O}(h^{2p_0})$$

follows from $w = v + \phi$.

Recall that if exact data are used for $t = 0$, then this technique for deriving the convergence rate is required only for multi-step schemes. Furthermore, the only extra restrictions imposed by this construction, is the requirement of a certain smoothness of the truncation error in the initial condition. We need $Q\tilde{f} = \mathcal{O}(h)$, and since $Q = I + kQ_1$, where Q_1 is consistent with a first order differential operator in space, the smoothness condition is fulfilled if for example \tilde{f} is Lipschitz continuous.

For an approximation which is stable in the generalized sense, the construction is more complicated, since v must satisfy also the boundary conditions, and still have a certain smoothness property. The complete construction for this case is carried out in [Gustafsson, 1975].

Error estimates can sometimes be derived by direct application of the energy method. For illustration, consider the differential equation $u_t = u_x$ with a boundary condition to the right, and the standard second order semidiscrete approximation (2.30) (with a renumbering such that $x_j = 0$ is the first point):

$$\begin{aligned}\frac{du_j}{dt} &= Qu_j, \quad j = 0, 1, \dots, N-1, \\ u_N(t) &= g(t), \\ u_j(0) &= f_j, \quad j = 0, 1, \dots, N-1, \\ Qu_j &= \begin{cases} D_+ u_j, & j = 0, \\ D_0 u_j, & j = 1, 2, \dots, N-1. \end{cases}\end{aligned}$$

The error $w_j(t) = u(x_j, t) - u_j(t)$ satisfies the system

$$\begin{aligned}\frac{dw_j}{dt} &= Qw_j + F_j, \quad j = 0, 1, \dots, N-1, \\ w_N(t) &= 0, \\ w_j(0) &= 0, \quad j = 0, 1, \dots, N-1.\end{aligned}$$

where

$$F_j = \begin{cases} \mathcal{O}(h), & j = 0 \\ \mathcal{O}(h^2), & j = 1, 2, \dots, N-1, \end{cases}$$

With the scalar product defined by

$$(u, v)_h = \frac{h}{2} u_0 v_0 + \sum_{j=1}^{N-1} u_j v_j h,$$

we get

$$\frac{d}{dt} \|w\|_h^2 = 2(w, Qw + F)_h = -|w_0|^2 + w_0 F_0 h + 2 \sum_{j=1}^{N-1} w_j F_j h.$$

By using the inequalities

$$2|w_j F_j| \leq |w_j|^2 + |F_j|^2, \quad j = 1, 2, \dots, N-1,$$

and

$$|w_0 F_0 h| \leq \delta |w_0|^2 + \frac{1}{4\delta} |F_0|^2 h^2$$

with $\delta = 1$, we get

$$\frac{d}{dt} \|w\|_h^2 \leq \frac{1}{4} |F_0|^2 h^2 + \sum_{j=1}^{N-1} (|w_j|^2 + |F_j|^2) h \leq \|w\|_h^2 + \mathcal{O}(h^4).$$

After integration of this inequality, we obtain

$$\|w(t)\|_h \leq K(t)h^2, \quad (3.8)$$

where the constant $K(t)$ depends on t but not on h .

The result obtained here shows that the global accuracy is one order higher than the local approximation at the boundary. The technique used here works because the summation by parts provides a term $-|u_0|^2$ which can be used to cancel the term $\delta|u_0|^2$ that is obtained when bounding $u_0 F_0 h$. The factor h can be used in full to raise the order of the local truncation error at $x = x_0$.

The term $-|u_0|^2$ represents an energy leakage through the boundary. For other types of problems, where this kind of term is not present, a factor $\mathcal{O}(h^{1/2})$ is lost in the error estimate by the straightforward application of the energy method. However, by using a deeper theory, but still based on the energy method, it is possible to obtain optimal error estimates for such problems as well. This was done for the fully discrete case in [Gustafsson, 1981], and the same technique can be used also for the semidiscrete case.

The normal mode analysis is a more general and powerful tool, both for investigation of stability and of the convergence rate. The presentation of this type of analysis has been limited to hyperbolic systems in this book. The Kreiss condition is fundamental in this theory, and if it is satisfied, one can in most cases allow for one order lower approximation for the extra boundary conditions.

Sometimes the order of accuracy near the boundary is lowered more than one step for stability reasons. Assume that this local accuracy has order $p - s$ for a scheme of order p , where $s \geq 0$. Then one can show that the global error satisfies

$$\|u(t)\|_h \leq K(t)h^{\min(p,p-s+1)} \quad (3.9)$$

for hyperbolic problems. For strongly stable schemes, this result is included in (3.7), but for weaker types of stability, some extra conditions might be needed as discussed for the case $s = 1$.

Let us now include also parabolic problems, and consider the quarter space problem

$$u_t = \frac{\partial^r u}{\partial x^r}, \quad 0 \leq x < \infty, \quad 0 \leq t,$$

$$\begin{aligned} B_r u(0, t) &= 0, \\ u(x, 0) &= f(x), \\ \|u(\cdot, t)\| &< \infty. \end{aligned}$$

Here $r = 1$ or $r = 2$, and B_r represents the physical boundary condition. We use the fourth order approximation

$$\frac{du_j}{dt} = Q_4^{(r)} u_j, \quad j = 1, 2, \dots,$$

at inner points and the extrapolation

$$u_{-1} = 4u_0 - 6u_1 + 4u_2 - u_3 \quad (3.10)$$

as an extra boundary condition. This is a fourth order approximation of the function itself at $x = -h$. The fourth order difference approximation of u_x at $x = h$ is

$$Q_4^{(1)} u_1 = D_0(I - \frac{h^2}{6} D_+ D_-) u_1 = \frac{1}{h} (\frac{1}{12} u_{-1} - \frac{2}{3} u_0 + \frac{2}{3} u_2 - \frac{1}{12} u_3).$$

When substituting (3.10) into this formula, we get

$$Q_3^{(1)} u_1 = \frac{1}{h} (-\frac{1}{3} u_0 - \frac{1}{2} u_1 + u_2 - \frac{1}{6} u_3) = D_0 u_1 - \frac{h^2}{6} D_+^2 D_- u_1.$$

The last term approximates the leading truncation error $h^2 u_{xxx}/6$ of $D_0 u$. Since the operator $D_+^2 D_-$ is not centered properly, we get a first order approximation of u_{xxx} resulting in a third order approximation of u_x at $x = h$.

Consider next the fourth order approximation of $u_{xx}(h)$:

$$Q_4^{(2)} u_1 = D_+ D_-(I - \frac{h^2}{12} D_+ D_-) u_1.$$

The extrapolation (3.10) can be written as $(D_+ D_-)^2 u_1 = 0$, and changes the approximation to

$$Q_2^{(2)} u_1 = D_+ D_- u_1,$$

which is the standard second order approximation of u_{xx} . Hence, one and the same fourth order extrapolation corresponds to an approximation of the differential operator of order 3 if $r = 1$ and of order 2 if $r = 2$, i.e., the accuracy is lowered r steps.

For the analysis, the Laplace transform technique can be used also for higher order PDE. The Laplace transform acts in the time direction, and it does not change the extrapolation formula (3.10). The transformed error equation for $w_j(t) = u(x_j, t) - u_j(t)$ is

$$\tilde{w}_{-1} - 4\tilde{w}_0 + 6\tilde{w}_1 - 4\tilde{w}_2 + \tilde{w}_3 = h^4 \tilde{g},$$

where \tilde{g} is bounded. Another boundary condition is required, and it is an approximation of the boundary condition for the PDE. At this point we refer back to the analysis in Section 2.4, and assume that the two boundary conditions together are such that we can solve uniquely for $\tilde{w}_{-1}, \tilde{w}_0$ in terms of inner points. Then these two relations have an error term of order h^4 . The conclusion is that the error is of fourth order in both cases $r = 1$ and $r = 2$.

This discussion indicates that in general we can expect that the accuracy of the extra boundary conditions can be relaxed as many steps as the order of the PDE in space. Some results in this direction are obtained in [Svärd and Nordström, 2006]. The concept *pointwise strongly stable* is introduced there, which is essentially the same as strongly stable, but in the l_∞ -norm. For such approximations the result indicated above is proven.

There is also another aspect on error estimates. The constant $K(t)$ in (3.9) depends on t , and it is interesting to know how it behaves for a fixed step size and increasing t . This problem was treated in [Abarbanel et al., 2000] for a large class of approximations. For hyperbolic problems and $s = 0, 1$, one can show that $K(t) \sim \sqrt{t}$ for $t \ll 1$, and then changes smoothly to $K(t) \sim t$. For $s \geq 2$, the increase is weaker, and we have $K(t) \sim \sqrt{t}$ for all time.

3.3 Summary

The convergence rate as the step size tends to zero has been treated in this chapter. For periodic problems, the question of global accuracy has easy answers. The initial steps that are necessary to start a multistep scheme, can have one order lower accuracy $p - 1$ without destroying the overall accuracy p . Initial-boundary value problems are more complicated. The physical boundary conditions that hold for the differential equation must always be approximated to at least the same order as the main difference scheme has. The question regarding the extra numerical boundary conditions is more involved. For strongly stable schemes with truncation errors defined in (3.5), the error estimate (3.7) is easily obtained directly from the stability definition. However, one must be careful with the definition of the boundary operator B_h . It has to be normalized properly such that it really is strongly stable, i.e., the solution can be estimated in terms of the right hand side of the boundary condition.

When considering the extra boundary conditions as approximations of the differential operator of degree r in space, then the general rule is that the approximation order can be lowered r steps locally. We have indicated that this rule holds for large classes of approximations. Our experience from a large number of applications is that if the Kreiss condition is satisfied, or if there is an energy estimate, then full accuracy is obtained. If the strong stability requirement is relaxed, there are counterexamples to this result. However, if stability holds in any of the other senses defined in Chapter 2, the known counterexamples show at most the weak deterioration $1/\sqrt{h}$, and this is by no means any disaster. Actually, as we shall see in Chapters 7 and 9, there is often good reason for going down more than r steps in order to keep some other desirable property, and these schemes may very well be very effective despite the lower formal accuracy.

One should be aware that all the results in this chapter hold only if the solution is sufficiently smooth. For initial-boundary value problems, this is a quite strong assumption. For nonparabolic problems (with no damping), the initial and boundary data must be prescribed carefully with sufficient compatibility at the corners, such that no discontinuities are created in the solution. On the other hand, if the solutions are nonsmooth such that the formal accuracy goes down, the high order approximations may well be more effective than lower order ones. This was demonstrated in Chapter 1, and will be further discussed in Chapter 6 and 9.

Some of the material in this chapter is found in the two papers [Gustafsson, 1975] and [Gustafsson, 1981]. The first one is about approximations of first order

hyperbolic PDE, and is based on normal mode analysis. The second one treats approximations of more general PDE, and it is based on the energy method.

A reminder that the boundary accuracy has to be handled very carefully is given by Carpenter et.al. in [Carpenter et al., 1995]. They show that the seemingly very natural inflow boundary condition $u(0,t) = g(t)$ for a scalar hyperbolic problem, does not give full accuracy when used with a Runge-Kutta method. The RK methods are one step schemes going from t_n to t_{n+1} , but they have several stages, and the final truncation error at $t = t_{n+1}$ has to be carefully investigated at the boundary. One cure for the problem is to substitute the boundary condition by the differentiated version $\partial^r u(0,t)/\partial t^r = d^r g/dt^r$, where r is the number of stages.

The general results concerning truncation errors and convergence rate hold without any restrictions on the grid. However, for nonregular domains, it is almost always a good idea to do a transformation to a computational space with a uniform grid (if it is possible), and then apply the numerical method there. The truncation error is easily found, and the stability theory is more easily applied. However, for first or second order methods, one can stay in physical space if the grid has been generated by a smooth transformation. This is illustrated for finite volume methods in Section 12.4, where the connection to difference methods is discussed.

Chapter 4

Approximation in Space

Given the PDE

$$\frac{\partial u}{\partial t} = Pu,$$

where P is a differential operator in space, one can first discretize in space and obtain a system of ODE

$$\frac{dv}{dt} = Qv. \quad (4.1)$$

Here v is a grid function, and Q is a difference operator. This is called a *semidiscrete approximation*, or the *method of lines*. The latter name originates from the graph in the x/t plane, where the discretization is shown by vertical, but continuous lines. It is assumed that the system (4.1) is solved by some standard method in time. In the following we shall discuss the derivation of different types of operators Q .

4.1 High Order Formulas on Standard Grids

In this section we assume that the difference operator $Q = Q_p$ is centered at the point x_j on a uniform grid, and begin by the approximation of $\partial/\partial x$. Because of symmetry, we make the ansatz

$$Q_p = D_0 \sum_{\nu=0}^{p/2-1} (-1)^\nu \alpha_\nu (h^2 D_+ D_-)^\nu,$$

where α_ν are the coefficients to be determined. When disregarding possible boundaries, the solution can be expanded in terms of Fourier components, and referring back to Chapter 1 we get for each component

$$Q_p e^{i\omega x} = \frac{i}{h} \sin \xi \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} (\sin \frac{\xi}{2})^{2\nu} e^{i\omega x}, \quad \xi = \omega h.$$

A Taylor expansion results in the expression

$$Q_p u = \frac{du}{dx} + \mathcal{O}(h^p \frac{d^{p+1}u}{dx^{p+1}}),$$

i.e., in Fourier space

$$Q_p e^{i\omega x} = (i\omega + \mathcal{O}(h^p \omega^{p+1})) e^{i\omega x},$$

and this gives the equation

$$\xi + \mathcal{O}(\xi^{p+1}) = \sin \xi \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} (\sin \frac{\xi}{2})^{2\nu}.$$

When substituting $\theta = \sin(\xi/2)$, we get

$$\frac{\arcsin \theta}{\sqrt{1 - \theta^2}} = \theta \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} \theta^{2\nu} + \mathcal{O}(\theta^{p+1}).$$

The left hand side is expanded in a Taylor series, and by identifying the coefficients of the powers of θ , the coefficients α_ν are uniquely determined. The result can be expressed by the recursion

$$\begin{aligned} \alpha_0 &= 1, \\ \alpha_\nu &= \frac{\nu}{4\nu + 2} \alpha_{\nu-1}, \quad \nu = 1, 2, \dots, p/2 - 1. \end{aligned} \tag{4.2}$$

The same principle can be used for higher order derivatives. For $\partial^2/\partial x^2$, the approximation is based on the expansion

$$D_+ D_- \sum_{\nu=0}^{p/2-1} (-1)^\nu \beta_\nu (h^2 D_+ D_-)^\nu,$$

etc.

We rewrite the difference approximation of $\partial^r/\partial u^r$ in the form

$$Q_p^{(r)} u_j = \frac{1}{h^r} \sum_{\nu=-p/2}^{p/2} a_\nu^{(r)} u_{j+\nu}. \tag{4.3}$$

Table 4.1 shows the coefficients $a_\nu^{(r)}$ for $1 \leq r \leq 4$.

Table 4.1 Coefficients $a_\nu^{(r)}$ in (4.3)

$h^r \frac{\partial^r}{\partial x^r}$	p	x_{j-4}	x_{j-3}	x_{j-2}	x_{j-1}	x_j	x_{j+1}	x_{j+2}	x_{j+3}	x_{j+4}
$h \frac{\partial}{\partial x}$	2				$-\frac{1}{2}$	0	$\frac{1}{2}$			
	4			$\frac{1}{12}$	$-\frac{2}{3}$	0	$\frac{2}{3}$	$-\frac{1}{12}$		
	6		$-\frac{1}{60}$	$\frac{3}{20}$	$-\frac{3}{4}$	0	$\frac{3}{4}$	$-\frac{3}{20}$	$\frac{1}{60}$	
	8	$\frac{1}{280}$	$-\frac{4}{105}$	$\frac{1}{5}$	$-\frac{4}{5}$	0	$\frac{4}{5}$	$-\frac{1}{5}$	$\frac{4}{105}$	$-\frac{1}{280}$
$h^2 \frac{\partial^2}{\partial x^2}$	2				1	-2	1			
	4			$-\frac{1}{12}$	$\frac{4}{3}$	$-\frac{5}{2}$	$\frac{4}{3}$	$-\frac{1}{12}$		
	6		$\frac{1}{90}$	$-\frac{3}{20}$	$\frac{3}{2}$	$-\frac{49}{18}$	$\frac{3}{2}$	$-\frac{3}{20}$	$\frac{1}{90}$	
	8	$-\frac{1}{560}$	$\frac{8}{315}$	$-\frac{1}{5}$	$\frac{8}{5}$	$-\frac{205}{72}$	$\frac{8}{5}$	$-\frac{1}{5}$	$\frac{8}{315}$	$-\frac{1}{560}$
$h^3 \frac{\partial^3}{\partial x^3}$	2			$-\frac{1}{2}$	1	0	-1	$\frac{1}{2}$		
	4			$\frac{1}{8}$	-1	$\frac{13}{8}$	0	$-\frac{13}{8}$	1	$-\frac{1}{8}$
	6	$-\frac{7}{240}$	$\frac{3}{10}$	$-\frac{169}{120}$	$\frac{61}{30}$	0	$-\frac{61}{30}$	$\frac{169}{120}$	$-\frac{3}{10}$	$\frac{7}{240}$
	8									
$h^4 \frac{\partial^4}{\partial x^4}$	2			1	-4	6	-4	1		
	4			$-\frac{1}{6}$	2	$-\frac{13}{2}$	$\frac{28}{3}$	$-\frac{13}{2}$	2	$-\frac{1}{6}$
	6	$\frac{7}{240}$	$-\frac{2}{5}$	$\frac{169}{60}$	$-\frac{122}{15}$	$\frac{91}{8}$	$-\frac{122}{15}$	$\frac{169}{60}$	$-\frac{2}{5}$	$\frac{7}{240}$
	8									

For periodic problems, it is natural to work with centered approximations as given here. When boundaries are involved, we need extra numerical boundary conditions. As we have seen in Chapter 2, one possibility is to use noncentered approximations. In general, at some points, the grid function values may be available on both sides of the approximation point, but here we limit ourselves to one sided formulas

$$\left(\frac{\partial^r u}{\partial x^r} \right)_{x=x_0} \approx \frac{1}{h^r} \sum_{\nu=0}^{r+p-1} b_\nu^{(r)} u_\nu, \quad (4.4)$$

where p is the order of accuracy. The coefficients $b_\nu^{(r)}$ are given in Table 4.2.

Table 4.2 Coefficients $b_{\nu}^{(r)}$ for one-sided approximations (4.4)

$h^r \frac{\partial^r}{\partial x^r}$	p	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$h \frac{\partial}{\partial x}$	1	-1	1							
	2	$-\frac{3}{2}$	2	$-\frac{1}{2}$						
	3	$-\frac{11}{6}$	3	$-\frac{3}{2}$	$\frac{1}{3}$					
	4	$-\frac{25}{12}$	4	-3	$\frac{4}{3}$	$-\frac{1}{4}$				
	5	$-\frac{137}{60}$	5	-5	$\frac{10}{3}$	$-\frac{5}{4}$	$\frac{1}{5}$			
	6	$-\frac{49}{20}$	6	$-\frac{15}{2}$	$\frac{20}{3}$	$-\frac{15}{4}$	$\frac{6}{5}$	$-\frac{1}{6}$		
	7	$-\frac{363}{140}$	7	$-\frac{21}{2}$	$\frac{35}{3}$	$-\frac{35}{4}$	$\frac{21}{5}$	$-\frac{7}{6}$	$\frac{1}{7}$	
	8	$-\frac{761}{280}$	8	-14	$\frac{56}{3}$	$-\frac{35}{2}$	$\frac{56}{5}$	$-\frac{14}{3}$	$\frac{8}{7}$	$-\frac{1}{8}$
$\frac{\partial^2}{\partial x^2}$	1	1	-2	1						
	2	2	-5	4	-1					
	3	$\frac{35}{12}$	$-\frac{26}{3}$	$\frac{19}{2}$	$-\frac{14}{3}$	$\frac{11}{12}$				
	4	$\frac{15}{4}$	$-\frac{77}{6}$	$\frac{107}{6}$	-13	$\frac{61}{12}$	$-\frac{5}{6}$			
	5	$\frac{203}{45}$	$-\frac{87}{5}$	$\frac{117}{4}$	$-\frac{254}{9}$	$\frac{33}{2}$	$-\frac{27}{5}$	$\frac{137}{180}$		
	6	$\frac{469}{90}$	$-\frac{223}{10}$	$\frac{879}{20}$	$-\frac{949}{18}$	41	$-\frac{201}{10}$	$\frac{1019}{180}$	$-\frac{7}{10}$	
	7	$\frac{29531}{5040}$	$-\frac{962}{35}$	$\frac{621}{10}$	$-\frac{4006}{45}$	$\frac{691}{8}$	$-\frac{282}{5}$	$\frac{2143}{90}$	$-\frac{206}{35}$	$\frac{363}{560}$
$\frac{\partial^3}{\partial x^3}$	1	-1	3	-3	1					
	2	$-\frac{5}{2}$	9	-12	7	$-\frac{3}{2}$				
	3	$-\frac{17}{4}$	$\frac{71}{4}$	$-\frac{59}{2}$	$\frac{49}{2}$	$-\frac{41}{4}$	$\frac{7}{4}$			
	4	$-\frac{49}{8}$	29	$-\frac{461}{8}$	62	$-\frac{307}{8}$	13	$-\frac{15}{8}$		
	5	$-\frac{967}{120}$	$\frac{638}{15}$	$-\frac{3929}{40}$	$\frac{389}{3}$	$-\frac{2545}{24}$	$\frac{268}{5}$	$-\frac{1849}{120}$	$\frac{29}{15}$	
	6	$-\frac{801}{80}$	$\frac{349}{6}$	$-\frac{18353}{120}$	$\frac{2391}{10}$	$-\frac{1457}{6}$	$\frac{4891}{30}$	$-\frac{561}{8}$	$\frac{527}{30}$	$-\frac{469}{240}$
$\frac{\partial^4}{\partial x^4}$	1	1	-4	6	-4	1				
	2	3	-14	26	-24	11	-2			
	3	$\frac{35}{6}$	-31	$\frac{137}{2}$	$-\frac{242}{3}$	$\frac{107}{2}$	-19	$\frac{17}{6}$		
	4	$\frac{28}{3}$	$-\frac{111}{2}$	142	$-\frac{1219}{6}$	176	$-\frac{185}{2}$	$\frac{82}{3}$	$-\frac{7}{2}$	
	5	$\frac{1069}{80}$	$-\frac{1316}{15}$	$\frac{15289}{60}$	$-\frac{2144}{5}$	$\frac{10993}{24}$	$-\frac{4772}{15}$	$\frac{2803}{20}$	$-\frac{536}{15}$	$\frac{967}{240}$

The approximations in Table 4.2 are of course not unique. By including more points to the right, other formulas with the same order of accuracy can be obtained. Indeed, the so called summation by parts (SBP) operators that will be discussed in Section 7.2, do not use the one sided approximations in the table (except for $r = 1$ and $p = 1$). The reason is that the SBP operators are constructed in such a way that an energy estimate is guaranteed, and this requires more than the minimal number of points for obtaining a certain order of accuracy.

We have limited the discussion to the approximation of derivatives. On a uniform grid the inclusion of variable coefficients in the PDE does not introduce any difficulties. A term $a(x)u_x$ is simply approximated by $a(x_j)Qu_j$, where Q is one of the operators above. On a staggered grid, an approximation centered at $x_{j+1/2}$ has the form $(Ia)_j\tilde{Q}u_j$, where $(Ia)_j$ is an interpolation of $a(x)$ that has the same order of accuracy as the difference operator \tilde{Q} has.

The formulas here should be applied in the computational domain, which is possibly obtained by a transformation from some other domain in physical space. There is of course a possibility to compute directly in physical space without transformation. If the grid there is obtained by a smooth transformation from a uniform grid, one can actually apply the difference approximation without modification if the accuracy is limited to second order. This will be demonstrated in the final section for finite volume methods, and the result can be transferred to difference methods without modification. However, in order to keep full accuracy, higher order computations should be carried out in the uniform grid space. Indeed, this is a good rule for all methods.

4.2 High Order Formulas on Staggered Grids

Consider the wave propagation system

$$\begin{aligned} p_t &= c^2 u_x, \\ u_t &= p_x, \end{aligned}$$

which will be discussed later in Section 9.1. The structure is such that one can use a staggered grid, where u is stored at $x_j = jh$, and p at the intermediate half-points $x_{j+1/2} = (j + 1/2)h$. The semidiscrete second order system is

$$\begin{aligned} \frac{dp_{j+1/2}}{dt} &= c^2 \frac{u_{j+1} - u_j}{h}, \\ \frac{du_j}{dt} &= \frac{p_{j+1/2} - p_{j-1/2}}{h}. \end{aligned}$$

Since the truncation error is proportional to the square of the step size, the error becomes four times smaller compared to a standard grid. For example, as demonstrated above, we have

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

which leads to

$$\frac{u(x+h/2) - u(x-h/2)}{h} = u_x(x) + \frac{h^2}{24} u_{xxx} + \mathcal{O}(h^4).$$

We can now expand the right hand side further just as in the previous section, and obtain a difference operator of arbitrarily high order:

$$\tilde{Q}_p^{(r)} u_j = \frac{1}{h^r} \sum_{\nu=-p/2}^{p/2-1} \tilde{a}_\nu^{(r)} u_{j+\nu+1/2}. \quad (4.5)$$

Note that this formula holds also for half-points, i.e., for all j in the set $\{0, \pm 1/2, \pm 1, \pm 3/2, \pm 2, \dots\}$.

For more general systems, it may be necessary to interpolate the function itself, and in that case we may need the expansion (4.5) also for $r = 0$. The following table shows the coefficients $\tilde{a}_\nu^{(r)}$.

Table 4.3 Coefficients $\tilde{a}_\nu^{(r)}$ in (4.5) for staggered grids

$h^r \frac{\partial^r}{\partial x^r}$	p	$x_{j-7/2}$	$x_{j-5/2}$	$x_{j-3/2}$	$x_{j-1/2}$	$x_{j+1/2}$	$x_{j+3/2}$	$x_{j+5/2}$	$x_{j+7/2}$
I	2				$\frac{1}{2}$	$\frac{1}{2}$			
	4			$-\frac{1}{16}$	$\frac{9}{16}$	$\frac{9}{16}$	$-\frac{1}{16}$		
	6		$\frac{3}{256}$	$-\frac{25}{256}$	$\frac{75}{128}$	$\frac{75}{128}$	$-\frac{25}{256}$	$\frac{3}{256}$	
	8	$-\frac{5}{2048}$	$\frac{49}{2048}$	$-\frac{245}{2048}$	$\frac{1225}{2048}$	$\frac{1225}{2048}$	$-\frac{245}{2048}$	$\frac{49}{2048}$	$-\frac{5}{2048}$
$h \frac{\partial}{\partial x}$	2				-1	1			
	4			$\frac{1}{24}$	$-\frac{9}{8}$	$\frac{9}{8}$	$-\frac{1}{24}$		
	6		$-\frac{3}{640}$	$\frac{25}{384}$	$-\frac{75}{64}$	$\frac{75}{64}$	$-\frac{25}{384}$	$\frac{3}{640}$	
	8	$\frac{5}{7168}$	$-\frac{49}{5120}$	$\frac{245}{3072}$	$-\frac{1225}{1024}$	$\frac{1225}{1024}$	$-\frac{245}{3072}$	$\frac{49}{5120}$	$-\frac{5}{7168}$
$h^2 \frac{\partial^2}{\partial x^2}$	2			$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$		
	4			$-\frac{5}{48}$	$\frac{13}{16}$	$-\frac{17}{24}$	$-\frac{17}{24}$	$\frac{13}{16}$	$-\frac{5}{48}$
	6	$\frac{259}{11520}$	$-\frac{499}{2304}$	$\frac{1299}{1280}$	$-\frac{1891}{2304}$	$-\frac{1891}{2304}$	$\frac{1299}{1280}$	$-\frac{499}{2304}$	$\frac{259}{11520}$
$h^3 \frac{\partial^3}{\partial x^3}$	2			-1	3	-3	1		
	4			$\frac{1}{8}$	$-\frac{13}{8}$	$\frac{17}{4}$	$-\frac{17}{4}$	$\frac{13}{8}$	$-\frac{1}{8}$
	6	$-\frac{37}{1920}$	$\frac{499}{1920}$	$-\frac{1299}{640}$	$\frac{1891}{384}$	$-\frac{1891}{384}$	$\frac{1299}{640}$	$-\frac{499}{1920}$	$\frac{37}{1920}$
$h^4 \frac{\partial^4}{\partial x^4}$	2			$\frac{1}{2}$	$-\frac{3}{2}$	1	1	$-\frac{3}{2}$	$\frac{1}{2}$
	4	$-\frac{7}{48}$	$\frac{59}{48}$	$-\frac{45}{16}$	$\frac{83}{48}$	$\frac{83}{48}$	$-\frac{45}{16}$	$\frac{59}{48}$	$-\frac{7}{48}$

As for standard grids, noncentered formulas may be needed at the boundaries. Assuming that we need the approximation at $x = x_0$, and that there is one point $x_{-1/2}$ available to the left, the approximation has the form

$$\left(\frac{\partial^r u}{\partial x^r} \right)_{x=x_0} \approx \frac{1}{h^r} \sum_{\nu=-1}^{r+p-2} \tilde{b}_{\nu+1/2}^{(r)} u_\nu, \quad (4.6)$$

where p is the order of accuracy. Also here we may need to interpolate the function itself. The coefficients $\tilde{b}_{\nu+1/2}^{(r)}$ are given in Table 4.4.

Table 4.4 Coefficients $\tilde{b}_\nu^{(r)}$ in (4.6) for staggered grids

p	$x_{-1/2}$	$x_{1/2}$	$x_{3/2}$	$x_{5/2}$	$x_{7/2}$	$x_{9/2}$	$x_{11/2}$	$x_{13/2}$	$x_{15/2}$
1	1								
2	$\frac{1}{2}$	$\frac{1}{2}$							
3	$\frac{3}{8}$	$\frac{3}{4}$	$-\frac{1}{8}$						
I	4	$\frac{5}{16}$	$\frac{15}{16}$	$-\frac{5}{16}$	$\frac{1}{16}$				
	5	$\frac{35}{128}$	$\frac{35}{32}$	$-\frac{35}{64}$	$\frac{7}{32}$	$-\frac{5}{128}$			
	6	$\frac{63}{256}$	$\frac{315}{256}$	$-\frac{105}{128}$	$\frac{63}{128}$	$-\frac{45}{256}$	$\frac{7}{256}$		
	7	$\frac{231}{1024}$	$\frac{693}{512}$	$-\frac{1155}{1024}$	$\frac{231}{256}$	$-\frac{495}{1024}$	$\frac{77}{512}$	$-\frac{21}{1024}$	
	8	$\frac{429}{2048}$	$\frac{3003}{2048}$	$-\frac{3003}{2048}$	$\frac{2145}{2048}$	$-\frac{1001}{2048}$	$-\frac{273}{2048}$	$-\frac{33}{2048}$	
	2	-1	1						
	3	$-\frac{23}{24}$	$\frac{7}{8}$	$\frac{1}{8}$	$-\frac{1}{24}$				
	4	$-\frac{11}{12}$	$\frac{17}{24}$	$\frac{3}{8}$	$-\frac{5}{24}$	$\frac{1}{24}$			
$h \frac{\partial}{\partial x}$	5	$-\frac{563}{640}$	$\frac{67}{128}$	$\frac{143}{192}$	$-\frac{37}{64}$	$\frac{29}{128}$	$-\frac{71}{1920}$		
	6	$-\frac{1627}{1920}$	$\frac{211}{640}$	$\frac{59}{48}$	$-\frac{235}{192}$	$\frac{91}{128}$	$-\frac{443}{1920}$	$\frac{31}{960}$	
	7	$-\frac{88069}{107520}$	$\frac{2021}{15360}$	$\frac{28009}{15360}$	$-\frac{6803}{3072}$	$\frac{5227}{3072}$	$-\frac{12673}{15360}$	$\frac{3539}{15360}$	$-\frac{3043}{107520}$
	8	$-\frac{1423}{1792}$	$-\frac{491}{7168}$	$\frac{7753}{3072}$	$-\frac{18509}{5120}$	$\frac{3535}{1024}$	$-\frac{2279}{1024}$	$\frac{953}{1024}$	$-\frac{1637}{7168}$
									$\frac{2689}{107520}$

(Tables 4.1–4.4 are reproduced with permission from [Fornberg, 1988].)

4.3 Compact Padé Type Difference Operators

We have seen in Section 4.1 that for smooth functions u

$$u_x = D_0 u - \frac{h^2}{6} D_0 D_+ D_- u + \mathcal{O}(h^4). \quad (4.7)$$

Since $D_0 u = u_x + \mathcal{O}(h^2)$ we have

$$D_0 D_+ D_- u = D_+ D_- D_0 u = D_+ D_- u_x + \mathcal{O}(h^2),$$

and it follows that

$$(I + \frac{h^2}{6} D_+ D_-) u_x = D_0 u + \mathcal{O}(h^4).$$

Assuming that the inverse of the operator on the left hand side exists and is bounded independently of h , we have

$$u_x = (I + \frac{h^2}{6} D_+ D_-)^{-1} D_0 u + \mathcal{O}(h^4). \quad (4.8)$$

In this way, we have obtained a new difference operator $(I + \frac{h^2}{6} D_+ D_-)^{-1} D_0$, that has fourth order accuracy.

Another way to derive (4.8) is as follows. The approximation (4.7) can be written

$$u_x = (I - \frac{h^2}{6} D_+ D_-) D_0 u + \mathcal{O}(h^4).$$

When applied to smooth functions, the term $\frac{h^2}{6} D_+ D_-$ is in a certain sense a small perturbation of the identity operator I . Just as the algebraic equality

$$1 - \varepsilon = \frac{1}{1 + \varepsilon} + \mathcal{O}(\varepsilon^2)$$

holds for small ε , we have

$$I - \frac{h^2}{6} D_+ D_- = (I + \frac{h^2}{6} D_+ D_-)^{-1} + \mathcal{O}(h^4),$$

which leads to (4.8).

At a first glance it seems like we have just complicated things by introducing the inverse of a nondiagonal operator in the approximation. However, the advantage is that the error coefficient multiplying h^4 becomes smaller than for the standard operator Q_4 . Furthermore, the solution procedure requires only the solution of tridiagonal systems, which is a fast procedure.

In Chapter 1 we demonstrated the dispersion error as given in Fourier space, see Figure 1.1. The Fourier transform of the Padé operator $Q_4^{[P]}$ above is

$$\hat{Q}_4^{[P]} = \frac{i \sin \xi}{h \left(1 - \frac{2}{3} \sin^2(\xi/2) \right)}.$$

Figure 4.1 shows a comparison between the explicit operators \hat{Q}_4 , \hat{Q}_6 (shown also in Figure 1.1) and $\hat{Q}_4^{[P]}$. The 4th order Padé approximation is better than the 6th order explicit one almost everywhere. However, a closeup near $\xi = 0$ would show that \hat{Q}_6 is better there.

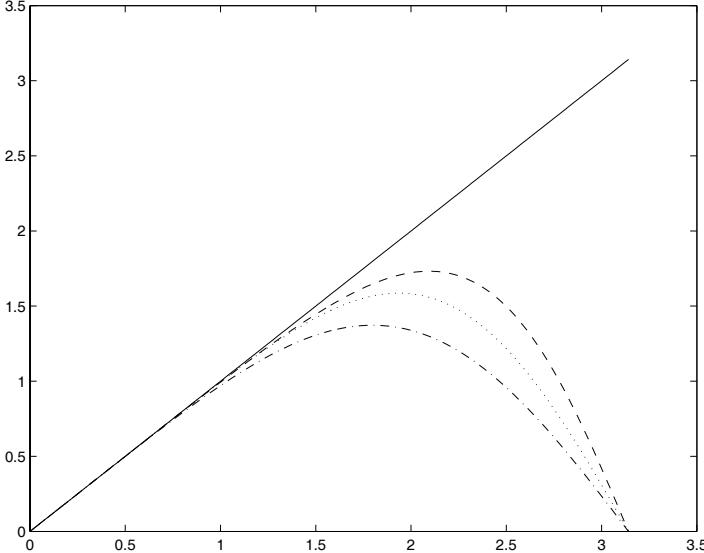


Fig. 4.1 Dispersion for $Q_4(-\cdot)$, $Q_6(\cdot\cdot)$, $Q_4^{[P]}(--)$

For general derivatives $v^{(r)} = \partial^r u / \partial x^r$, we rewrite the difference approximation in the form

$$P_p^{(r)} v_j^{(r)} = Q_p^{(r)} u_j,$$

where

$$\begin{aligned} P_p^{(r)} v_j &= \sum_{\nu=-\nu_0}^{\nu_0} b_\nu^{(r)} v_{j+\nu}, \\ Q_p^{(r)} u_j &= \frac{1}{h^r} \sum_{\nu=-\nu_1}^{\nu_1} c_\nu^{(r)} u_{j+\nu}. \end{aligned} \quad (4.9)$$

From an implementation point of view, it is convenient to keep the bandwidth on the left hand side as small as possible. In Table 4.5 we list the coefficients $b_\nu^{(r)}$ and $c_\nu^{(r)}$ for $r = 1, 2$ for tridiagonal left hand sides, i.e., $\nu_0 = 1$.

Padé type approximations can be developed for staggered grids as well. In Chapter 10 we shall describe some fourth order operators, and demonstrate how they can be used for the incompressible Navier-Stokes equations.

Let us next discuss the implementation of this type of difference approximations. Consider the PDE

$$u_t + a(x, t)u_x = 0,$$

and the leap-frog approximation

$$u_j^{n+1} = u_j^{n-1} - 2ka_j^n \left(I + \frac{h^2}{6} D_+ D_- \right)^{-1} D_0 u_j^n,$$

Table 4.5 Coefficients $b_\nu^{(r)}$ and $c_\nu^{(r)}$ in (4.9)

$\frac{\partial^r}{\partial x^r}$	P	$x_{j-2} \ x_{j-1} \ x_j \ x_{j+1} \ x_{j+2}$
$\frac{\partial}{\partial x}$	4	$b_\nu^{(1)}$ $c_\nu^{(1)}$
		1 4 1 -3 0 3
$\frac{\partial}{\partial x}$	6	$b_\nu^{(1)}$ $c_\nu^{(1)}$
		1 3 1 $-\frac{1}{12}$ $-\frac{7}{3}$ 0 $\frac{7}{3}$ $\frac{1}{12}$
$\frac{\partial^2}{\partial x^2}$	4	$b_\nu^{(1)}$ $c_\nu^{(1)}$
		1 10 1 12 -24 12
$\frac{\partial^2}{\partial x^2}$	6	$b_\nu^{(1)}$ $c_\nu^{(1)}$
		1 $\frac{11}{2}$ 1 $\frac{9}{24}$ 6 $-\frac{51}{4}$ 6 $\frac{9}{24}$

where n denotes the time level, and k the time step. Let $v_j^n = (I + \frac{h^2}{6}D_+D_-)^{-1}D_0u_j^n$ be the approximation of u_x on the grid $j = 1, 2, \dots, N$ with periodic boundary conditions $x_j = x_{j+N}$. We want to advance the scheme from time level n to time level $n+1$, i.e., u_j^{n+1} is known. The system

$$(I + \frac{h^2}{6}D_+D_-)v_j^n = D_0u_j^n, \quad j = 1, 2, \dots, N, \quad (4.10)$$

is solved for v_j^n , and u_j^{n+1} is then computed from

$$u_j^{n+1} = u_j^{n-1} - 2ka_j^n v_j^n.$$

The system is tridiagonal except for the nonzero corners of the coefficient matrix. The easiest way to solve it is by Fourier transform. With

$$u_j^n = \frac{1}{\sqrt{2\pi}} \sum \hat{u}_\omega^n e^{i\omega x_j},$$

$$v_j^n = \frac{1}{\sqrt{2\pi}} \sum \hat{v}_\omega^n e^{i\omega x_j},$$

we get for each coefficient

$$k\hat{v}_\omega^n = \frac{\lambda i \sin(\omega h)}{1 - \frac{2}{3} \sin^2(\omega h/2)} \hat{u}_\omega^n, \quad \lambda = \frac{k}{h}. \quad (4.11)$$

Hence, by using the FFT on u_j^n , the solution kv^n is obtained by computing $k\hat{v}_\omega^n$ from (4.11), and then using the inverse FFT. We also note that by Parseval's relation it follows from (4.11) that the inverse of the operator $I + \frac{h^2}{6}D_+D_-$ exists and is bounded.

For nonperiodic solutions, boundary conditions must be provided on each side, and we have a true tridiagonal system (or possibly a wider band matrix, depending on the boundary conditions). This system is solved by a fast direct method.

The structure of the system to be solved is independent of the differential equation. For example, for the nonlinear equation

$$u_t + a(x, t, u)u_x = 0,$$

we solve for $v = u_x$ just as above. Similarly, for the conservation law

$$u_t + f_x(u) = 0,$$

we put $g = f_x(u)$ and $g_j = (f_x(u))_{x=x_j}$. Then with u_j^n and $f_j^n = f(u_j^n)$ being known, g_j^n is computed from

$$(I + \frac{h^2}{6}D_+D_-)g_j^n = D_0f_j^n, \quad j = 1, 2, \dots, N.$$

For a system

$$\mathbf{u}_t + A(x, t, \mathbf{u})\mathbf{u}_x = 0,$$

where \mathbf{u} is a vector with m components, and A is an $m \times m$ matrix, the system (4.10) is solved for each component of \mathbf{u} .

Finally, for a PDE in several space dimensions

$$u_t + a(x, y, z, t, u)u_x + b(x, y, z, t, u)u_y + c(x, y, z, t, u)u_z = 0,$$

we solve a number of one-dimensional systems of the type (4.10) to obtain $v = u_x$, $w = u_y$, $s = u_z$, which are then inserted into the main difference approximation for each time step.

4.4 Optimized Difference Operators

The difference approximations we have described above are constructed such that they are very accurate near $\xi = 0$ in Fourier space. This means that the convergence rate as $h \rightarrow 0$ is high, i.e., $|u(x_j, t) - u_j(t)|$ is very small when h is small. In applications one is often faced with the problem of getting accurate solutions when the solutions are not smooth, i.e., when the solution contains components corresponding to large wave numbers ω . To get into the region with small $\xi = \omega h$, one must choose h so small that the computational problem becomes too large. In other

words, to get sufficient accuracy for the wave numbers at the high end, the solution is unnecessarily accurate at the low end.

An alternative is then to give up the formal high order, but still get good accuracy over a larger interval in the wave number domain. The dispersion error is then minimized over an interval $0 \leq \xi \leq \xi_0$.

As an example, we consider first order PDE, and the fourth order approximation

$$Q_4 = D_0(I - \frac{h^2}{6}D_+D_-).$$

Define instead

$$Q_4(\alpha) = D_0(I - \alpha \frac{h^2}{6}D_+D_-),$$

where α is a constant. In Fourier space we have

$$\hat{Q}_4(\alpha) = \frac{i}{h} \sin \xi \left(1 + \frac{2\alpha}{3} \sin^2 \frac{\xi}{2}\right),$$

which should be compared to $i\omega$. After dividing by $i\omega$, the relative error is

$$d(\xi) = 1 - \frac{\sin \xi}{\xi} \left(1 + \frac{2\alpha}{3} \sin^2 \frac{\xi}{2}\right).$$

We now define

$$G(\alpha) = \int_0^{\xi_0} |d(\xi)|^2 d\xi,$$

and solve the least square problem:

Find α_0 such that

$$\min_{\alpha} G(\alpha) = G(\alpha_0). \quad (4.12)$$

Figure 4.2 shows the difference between the standard 4th order approximation $\alpha = 1$ and the case $\alpha = 1.4$, which is obtained with the choice $\xi_0 = \pi/2$. For larger wave numbers, $\alpha = 1.4$ gives considerably better accuracy. The right figure shows the result in a different scale, and one can see that even if $\alpha = 1.4$ gives a slight “overshoot” for lower wave numbers, the error is still very small.

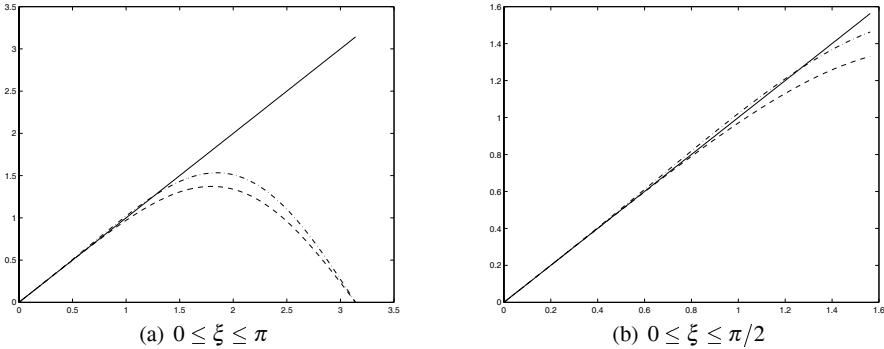


Fig. 4.2 Dispersion for $Q_4(\alpha)$, $\alpha = 1$ (—), $\alpha = 1.4$ (—·)

4.5 Summary

The presentation of difference approximations in this chapter is to be used for the method of lines, i.e., only the space part of the PDE is approximated. The formulas are well known, but the algorithms for generating the coefficients are nontrivial. The coefficients in Tables 4.1, 4.2, 4.3, 4.4 were computed by Fornberg in [Fornberg, 1988]. He used a clever formula that is quite simple to implement, see also [Fornberg, 1990] and [Fornberg, 1998].

The centered formulas are convenient from the point of view that the Fourier transforms are either purely real or purely imaginary. This means that the stability is easily verified for large classes of time discretizations. However, it should be emphasized that the one-sided formulas in Tables 4.2 and 4.4 are constructed only from an accuracy point of view. The stability is a much more complicated question. When they are used locally near the boundary for initial-boundary value problems, the stability must be investigated carefully as demonstrated in Chapter 2. In Section 7.2 there is a more systematic construction of boundary closures on standard grids, and they do not use the formulas in Table 4.2.

The Padé type approximations of differential operators were introduced by Collatz, see [Collatz, 1966]. The name refers to the French mathematician Henri Padé, who developed rational function approximations of given functions in his work in the late 19th century. Kreiss introduced them for time dependent PDE in unpublished work in the seventies, and Lele gave a more extensive derivation in [Lele, 1992]. The great advantage is the compact nature leading to a much smaller error coefficient for a given order of accuracy.

The last section is another demonstration of the fact that the formal order of accuracy does not tell the whole story about the accuracy on a given grid. By optimizing the coefficients such that they take into account a larger range of wave numbers, a more effective scheme can be obtained even if the formal order of accuracy is lower. The example given there is due to Efraimsson [Efraimsson, 1998]. Many others have

been using the similar ideas, see for example [Jakobsson, 2006], [Nehrbass et al., 1998], [Tam and Webb, 1993], [Zingg, 2000], [Zingg et al., 1993], [Zingg et al., 1996].

Another method for obtaining high order approximations is Richardson extrapolation. Low order computations on grids of different size are combined in a certain way to obtain higher accuracy. We shall describe this method as applied for approximation in time in Section 5.5.

Chapter 5

Approximation in Time

In Chapter 4, high order discretizations in space were discussed. With u denoting a vector containing the discrete solution at the grid points in space, we consider the ODE system

$$\frac{du}{dt} = Qu, \quad (5.1)$$

where Q is a matrix representation of the difference operator. For example, if Q corresponds to the standard 2nd order difference operator D_0 in one space dimension, and $u(x, t)$ is periodic in space, the system is

$$\frac{du}{dt} = \frac{1}{2h} \begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots \\ & & & & & -1 & 0 & 1 \\ 1 & & & & & & -1 & 0 \end{bmatrix} u$$

We shall first discuss the stability concept for ODE, and then introduce some of the classic numerical methods and its properties.

5.1 Stability and the Test Equation

The stability theory for high order ODE solvers is well developed for systems with a fixed size N . However, for discretizations of PDE systems, N increases without bounds as $h \rightarrow 0$. If Q can be diagonalized, i.e., there is a bounded matrix T with bounded inverse T^{-1} such that

$$T^{-1}QT = \Lambda,$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$ is a diagonal matrix, then it is sufficient to study the so called test equation

$$\frac{du}{dt} = qu, \quad (5.2)$$

where q is a complex number. If the method is stable for all $q = \lambda_j$, $j = 1, 2, \dots, N$, then it is stable also for the original system. Problems with periodic solutions often have the required properties. The matrix T above is then simply the discrete Fourier transform, which is a unitary transformation, and it has the condition number $|T| \cdot |T^{-1}| = 1$. In the nonperiodic case, it is a much harder problem to prove stability, as we saw in Chapter 2. However, the test equation still plays an important role. The von Neumann condition for periodic problems is a necessary condition for stability, and it gives an indication of the choice of method and time step. Sometimes it is possible to get some idea about the eigenvalues λ_j also for the initial-boundary value problem, and the stability domain to be discussed below, must contain these eigenvalues. However, it should be emphasized that a guarantee of stability can be obtained only by applying the more powerful tools derived in Chapter 2.

In the next sections we shall discuss various ODE methods, and we shall investigate their stability domains, i.e., the q -domain in the complex plane where they are stable. The centered approximations in space presented above lead to purely imaginary q for odd order PDE, and to purely real q for even order PDE. Therefore it is of primary interest to consider the imaginary and real axis of the stability domains. Furthermore, the ODE systems quickly become very *stiff*, i.e., the eigenvalues of Q have a very wide span on the axis where they reside, wider for increasing order of the PDE. For example, the heat equation $u_t = u_{xx}$ with standard 2nd order approximation in space, produces a matrix Q with eigenvalues distributed between 0 and $-4/h^2$. As we shall see, many methods require that $-\alpha_0 \leq \operatorname{Re} \mu \leq 0$ for some positive number α_0 , where $\mu = kq$, and k is the time step. Then the stability condition for our example is

$$k \leq \frac{\alpha_0 h^2}{4},$$

which is very restrictive if $|\alpha_0|$ is not very large. Consequently, such a method is not well suited for parabolic problems. The ideal situation is that the method is stable for all μ in the left halfplane: $\{\mu; \operatorname{Re} \mu \leq 0\}$. If this is the case, the method is called *A-stable*.

Hyperbolic first order systems of PDE is a special case. The resulting ODE system is stiff, but explicit methods can still be very useful. The equation $u_t + u_x = 0$ has propagation speed 1, which means that any particular feature in the solution should have roughly the same resolution in space as in time. The eigenvalues of the centered approximations that were derived in the previous section, are such that the eigenvalues of Q are distributed on the imaginary axis in an interval $[-ia_0/h, ia_0/h]$, where a_0 is a positive constant independent of h . Many explicit ODE solvers require that $|\operatorname{Im} \mu| \leq \beta_0$, where β_0 is a constant independent of h . The stability condition then becomes

$$\frac{ka_0}{h} \leq \beta_0.$$

The left hand side is the *CFL-number* (also called the *Courant number*), where the acronym refers to the paper [Courant et al., 1928]. The restriction is not severe in this case, since the resulting time step will be of the same order as the space step, and this is natural for a hyperbolic problem. We shall therefore consider both explicit and implicit methods in the following.

It should also be noted that the differential equations in applications often contain derivatives of different order. This means that the spectrum of the corresponding Q will not be located on the real or imaginary axis. However, the highest derivative of order r has more strength, since the order of the largest eigenvalues will be proportional to $1/h^r$. For example, the standard second order accurate approximation of $\partial^2/\partial x^2 + \partial/\partial x$ has eigenvalues

$$-\frac{4}{h^2} \sin^2 \frac{\xi}{2} + \frac{i}{h} \sin \xi.$$

For small h these are distributed in a long band along the negative real axis, which means that the time-step will be restricted more severely by the parabolic part than by the imaginary part.

When taking also nonperiodic boundary conditions into account, the situation becomes more complicated as we have seen in Chapter 2.

5.2 Runge–Kutta Methods

Perhaps the class of Runge–Kutta methods is the most well known among ODE solvers. Consider the general system

$$\frac{du}{dt} = F(t, u).$$

An RK method is a one step method, i.e., it is enough to know the solution u^n at one time level, in order to compute u^{n+1} . However, the solution is obtained by s separate stages, where $F(t, u)$ must be evaluated at each stage. We consider first explicit methods (ERK), which can be defined in the following way. Given a set of coefficients a_{ij} , b_j , c_j , usually arranged in a table

$$\begin{array}{c|ccccccccc} 0 & & & & & & & & & \\ \hline c_2 & a_{21} & & & & & & & & \\ c_3 & a_{31} & a_{32} & & & & & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & & & \\ \vdots & \vdots & \vdots & \vdots & & \ddots & & & & \\ \vdots & \vdots & \vdots & \vdots & & & \ddots & & & \\ \hline c_s & a_{s1} & a_{s2} & \dots & \dots & \dots & a_{s,s-1} & & & \\ \hline b_1 & b_2 & \dots & \dots & b_{s-1} & b_s, & & & & \end{array}$$

the method is defined by

$$\begin{aligned}k_1 &= F(t_n, u^n) \\k_2 &= F(t_n + c_2 k, u^n + k a_{21} k_1) \\k_3 &= F(t_n + c_3 k, u^n + k(a_{31} k_1 + a_{32} k_2)) \\\vdots \\k_s &= F(t_n + c_s k, u^n + k(a_{s1} k_1 + \dots + a_{s,s-1} k_{s-1})) \\u^{n+1} &= u^n + k(b_1 k_1 + \dots + b_s k_s)\end{aligned}$$

The function F is evaluated s times, but no system of equations has to be solved. For $s = 1$ we recognize the Euler forward method, for $s = 2$ it is the Heun's method. We are looking for the high order ones, and $s = 4$ gives the classic 4th order method defined by the table

0				
$1/2$	$1/2$			
$1/2$	0	$1/2$		
1	0	0		1
	$1/6$	$1/3$	$1/3$	$1/6$

For the test equation obtained by putting $F(t, u) = qu$, $\mu = kq$, we make the ansatz $u^n = z^n u_0$, where z is a complex number. This results in a Taylor expansion of the exponential e^μ for the first four methods, and for $s = 3$ and $s = 4$ we have

$$z = 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6},$$

$$z = 1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} + \frac{\mu^4}{24}.$$

The *stability domain* is defined as the set $S(\mu)$ in the complex plane which satisfies $|z(\mu)| < 1$. Figure 5.1 shows $S(\mu)$ for these two cases.

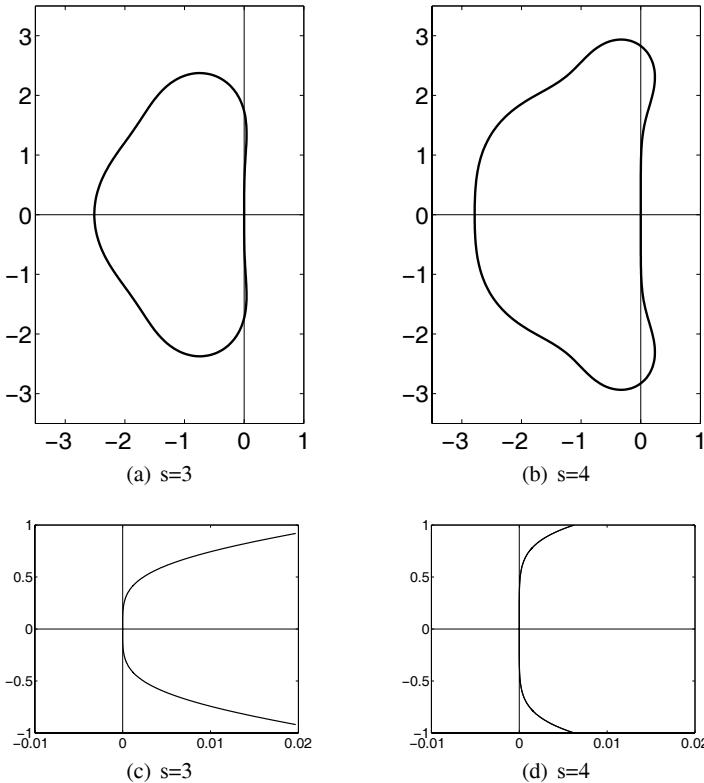


Fig. 5.1 Stability domain for 3rd and 4th order ERK, two different scales

For $s = 3$ the cutoff at the real axis is $\mu = -2.51$, and at the imaginary axis it is $\mu = \pm i\sqrt{3} = \pm 1.73i$. For $s = 4$ the corresponding values are $\mu = -2.78$ and $\mu = \pm 2i\sqrt{2} = \pm 2.82i$. The lower two figures show the boundary of the stability domain near $\mu = 0$ with a rescaled x -axis. The more pronounced tangential approach to the imaginary axis for $s = 4$ is an effect of the higher order accuracy.

For Runge–Kutta methods there is a general stability theorem which is applicable for the initial–boundary value problem, and it is based on the concept of stability in the generalized sense. It is defined for the semidiscrete case in Definition 2.19, and for the fully discrete case in Definition 2.24. Assume that the semidiscrete approximation has the form

$$\begin{aligned} \frac{du}{dt} &= Qu + F, \\ u(0) &= f, \end{aligned} \tag{5.3}$$

where the boundary conditions have been eliminated. Here $u = u(t)$ is the vector with u_j as components, and Q is the matrix corresponding to the linear difference approximation of a first order hyperbolic system. We consider Runge–Kutta methods of the form

$$\begin{aligned} u^{n+1} &= P(kQ)u^n + kP_1(kQ)F^n, \\ u^0 &= f, \end{aligned} \tag{5.4}$$

where $P(z)$ and $P_1(z)$ are polynomials in z . It is assumed that $\lambda = k/h$ is a constant, such that kQ is a bounded matrix for all h and k .

For the stability theorem we need the assumption that the stability domain contains a full half disc in the left half of the complex domain, and there are two other technical assumptions as well:

Theorem 5.1. Assume that

i) There is an open half disc

$$|\mu| < R, \quad \operatorname{Re} \mu < 0$$

which belongs to S .

ii) If the boundary of S touches the imaginary axis such that

$$P(\mu) = e^{i\alpha}, \quad \operatorname{Re} \mu = 0, \quad |\mu| \leq R,$$

then μ is a simple root of this equation, and it is the only one on the imaginary axis with $|\mu| \leq R$ for this α .

iii) The function $K(\eta)$ in Definition 2.19 satisfies $K(\eta) \leq K_1/\eta$, where K_1 is a constant.

Then the Runge-Kutta scheme is stable in the generalized sense if

$$\|kQ\|_h \leq R_1 \tag{5.5}$$

for any R_1 with $R_1 < R$. □

The estimate obtained in this way is given in terms of $\|P_1(kQ)F^n\|_h$. However, since $P_1(kQ)$ is a bounded matrix, we obtain an equivalent estimate in terms of $\|F^n\|_h$.

The technical assumptions are satisfied for most reasonable approximations, and they don't introduce any difficulties. For the third and fourth order methods the assumptions in the theorem are easily verified. A half disc can clearly be contained in the stability domain, for $s = 3$ the radius R_1 can be chosen as $R_1 = 1.73$. For $-1.73 \leq \operatorname{Im} \mu \leq 1.73$ there is only one point where the boundary of S touches the imaginary axis, and that is $\mu = 0$. But this is a simple root of

$$1 + \mu + \frac{\mu^2}{2} + \frac{\mu^3}{6} = 1,$$

and therefore all the assumptions are satisfied. The stability condition is $\|kQ\|_h < \sqrt{3}$. This may be a severe restriction on the time step. We shall see later that the construction of boundary conditions may introduce quite large coefficients in the matrix Q near the boundary, and for those approximations, $\|kQ\|_h$ may be very large. Practical computations show that k can be chosen well above the limit given by the

sufficient condition (5.5). However, no necessary and sufficient stability condition has been derived so far.

The class of Runge–Kutta methods has also been treated by Tadmor and co-authors, and their work is based on the energy method. Most results are for approximations using so called *coercive* operators Q with a certain dissipative property, see for example [Levy and Tadmor, 1998]. For general semibounded operators, there is essentially one result, and it holds for the third order Runge–Kutta method. If Q has time independent coefficients, then the method is

$$u^{n+1} = (I + kQ + \frac{k^2}{2}Q^2 + \frac{k^3}{6}Q^3)u^n, \quad (5.6)$$

and the following theorem holds:

Theorem 5.2. *Assume that the difference operator Q is semibounded with the scalar product $(\cdot, \cdot)_h$. Then the third order Runge–Kutta method (5.6) is stable if*

$$k\|Q\|_h \leq 1. \quad (5.7)$$

□

The time step limit is also here given in terms of the norm of Q , and sometimes it may be quite restrictive. It is conjectured, but not proven, that the condition is sharp when including all semibounded approximations, i.e., there exists a semibounded operator Q , where the condition is necessary. For a given approximation, it is sufficient, but not necessarily sharp. We note that the condition given by Theorem 5.2 is more restrictive than the one given by Theorem 5.1.

Figure 5.2 shows the stability domain for the 3rd and 4th order Runge–Kutta methods and the inscribed semicircle with radius $R = 1.73$ and $R = 2.58$ respectively, where R is the limit in the stability condition (5.5). In the third order case, R is determined by the cutoff at the imaginary axis, while in the 4th order case, it is determined by the dent in the kidney boundary to the left.

Explicit Runge–Kutta methods are classic, but there are also *implicit Runge–Kutta methods* (IRK). A set of coefficients

$$\begin{aligned} a_{ij}, \quad & 1 \leq i, j \leq s, \\ b_i, \quad & 1 \leq i \leq s, \\ c_i = \sum_{j=1}^{i-1} a_{ij} \end{aligned}$$

are given, and we assume that u^n is known. The IRK methods are defined by

$$\begin{aligned} k_i &= F(t_n + c_i k, u^n + k \sum_{j=1}^s a_{ij} k_j), \quad i = 1, 2, \dots, s, \\ u^{n+1} &= u^n + k \sum_{i=1}^s b_i k_i. \end{aligned}$$

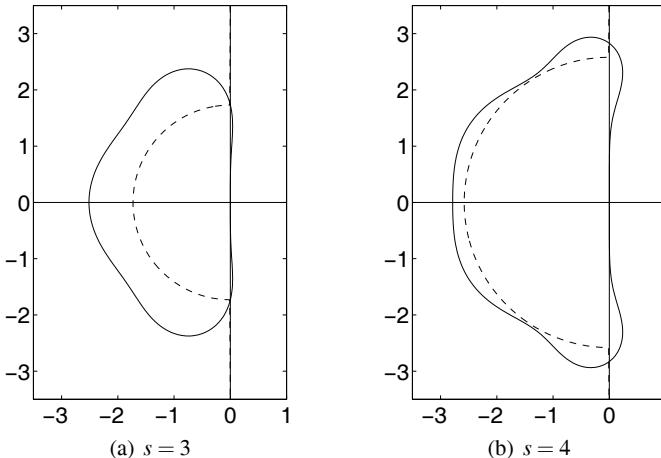


Fig. 5.2 Semicircle with radius R inside the stability domain for the 3rd and 4th order Runge-Kutta method

If $a_{ij} = 0$ for $i \leq j$, then we have an explicit method.

If $a_{ij} = 0$ for $i < j$, and at least one $a_{ii} \neq 0$, then we have a diagonal implicit Runge-Kutta method (DIRK).

It is easily checked that the parameter values $s = 1$, $a_{11} = 1$, $b_1 = 1$ gives the Euler backward method, and $s = 1$, $a_{11} = 1/2$, $b_1 = 1$ gives the trapezoidal rule, and both of them are A-stable.

The drawback with IRK is of course the need for solving the large systems of equations in each time step. For general IRK, all the stages are coupled to each other, and cannot be factored into s separate systems.

5.3 Linear Multistep Methods

In this section we consider the system (5.1) and discuss general m -step multistep methods of the form

$$\alpha_m u^{n+m} + \alpha_{m-1} u^{n+m-1} + \dots + \alpha_0 u^n = kQ(\beta_m u^{n+m} + \beta_{m-1} u^{n+m-1} + \dots + \beta_0 u^n). \quad (5.8)$$

Note that the concept “linear multistep methods” does not mean that the ODE system (5.1) has to be linear, but refers to the fact that Qu occurs in a linear fashion in (5.8) even if Q is a nonlinear operator. For nonlinear equations, there is also the possibility to use the different form

$$k(\beta_m Qu^{n+m} + \beta_{m-1} Qu^{n+m-1} + \dots + \beta_0 Qu^n)$$

for the right hand side.

For the test equation (5.2) obtained by setting $Q = q$, we define the *characteristic equation*

$$\rho(z) - \mu\sigma(z) = 0, \quad \mu = kq, \quad (5.9)$$

where k is the time step and

$$\begin{aligned} \rho(z) &= \alpha_m z^m + \alpha_{m-1} z^{m-1} + \dots + \alpha_0, \\ \sigma(z) &= \beta_m z^m + \beta_{m-1} z^{m-1} + \dots + \alpha_0. \end{aligned}$$

The characteristic equation is formally obtained by setting $u^n = z^n$ in (5.8) and then divide by z^n . (We accept here the ambiguity of the notation; u^n means time level t_n , while z^n means the n th power of z .) Since the characteristic equation may have multiple roots, the stability domain S must be defined with a little more care than for one step methods:

$$\begin{aligned} S = [\mu; \text{all roots } z_\nu(\mu) \text{ of (5.9) satisfy } |z_\nu(\mu)| \leq 1, \\ \text{multiple roots satisfy } |z_\nu(\mu)| < 1]. \end{aligned}$$

As an example, we study the leap-frog scheme for $du/dt = qu$

$$u^{n+2} - 2\mu u^{n+1} - u^n = 0,$$

and its characteristic equation

$$z^2 - 2\mu z - 1 = 0$$

with roots

$$z = \mu \pm \sqrt{1 + \mu^2}.$$

The only possibility to avoid one root being outside the unit circle, is to keep μ on the imaginary axis, and furthermore satisfying $|\mu| \leq 1$. In order to avoid multiple roots on the unit circle, the restriction is further strengthened to $|\mu| < 1$. Accordingly, we have a very small stability domain $(-i, i)$ on the imaginary axis.

Still the method may be useful for first order systems. Consider the model equation $u_t = u_x$ with the approximation $du/dt = D_0 u$. For periodic solutions, the diagonalization is obtained by a Fourier transform, and we get

$$\frac{d\hat{u}}{dt} = \frac{i \sin \xi}{h} \hat{u}.$$

This means that $\mu = i\lambda \sin \xi$, $\lambda = k/h$ in the analysis above, and the stability restriction is $\lambda < 1$.

The leap-frog scheme has a very simple structure with a centered 2nd order difference operator in both space and time. To get higher order in space, we just expand the stencil symmetrically around the central point x_j as demonstrated in Section 4.1. Unfortunately, this procedure cannot be used in time. The fourth order scheme would be (see Table 4.1)

$$-\frac{1}{12}u_j^{n+2} + \frac{2}{3}u_j^{n+1} - \frac{2}{3}u_j^{n-1} + \frac{1}{12}u_j^{n-2} = kD_0(I - \frac{h^2}{6}D_+D_-)u_j^n.$$

A necessary requirement for stability is that the characteristic equation for $k = 0$ does not have any roots outside the unit circle (this is called *zero-stability*). In our case, one root of

$$z^4 - 8z^3 + 8z - 1 = 0$$

is $z_1 = 7.873$, showing that the scheme is strongly unstable.

We next turn to some well known ODE methods, and for convenience, we formulate them for the test equation (5.2).

Explicit m -step Adams methods

It is convenient to introduce the backwards undivided difference operator in time defined by

$$\Delta_{-t} u^n = u^n - u^{n-1}.$$

The explicit m -step Adams method (also called the *Adams–Basforth method*) is

$$u^{n+1} = u^n + \mu \sum_{\nu=0}^{m-1} \gamma_\nu \Delta_{-t}^\nu u^n. \quad (5.10)$$

The order of accuracy is $p = m$ if the coefficients are chosen properly. The first coefficients are

$$\gamma_0 = 1, \gamma_1 = 1/2, \gamma_2 = 5/12, \gamma_3 = 3/8.$$

Figure 5.3 shows the stability domain for $m = 2, 3, 4$, and there is a significant difference between them. Indeed one can show for all m that the stability domain in the left halfplane $\operatorname{Re} \mu \leq 0$ diminishes with increasing m , and accordingly, these methods are not very suitable for PDE problems, except possibly for first order hyperbolic systems.

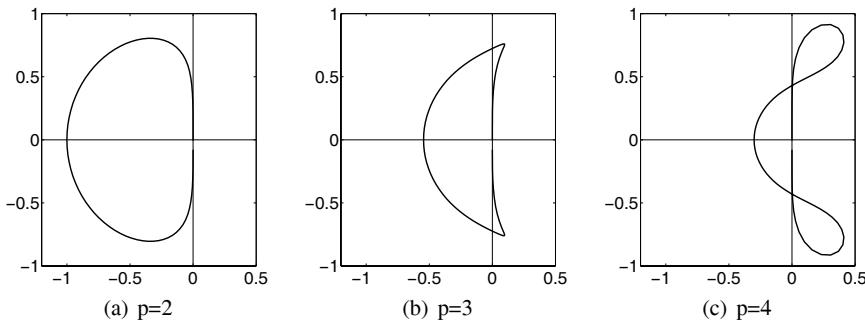


Fig. 5.3 Stability domain for explicit Adams methods of order p

There is another disadvantage with the second order method, besides its low accuracy. A closer look shows that except for the origin, the imaginary axis is not included at all in the stability domain. For the third and fourth order method it is, and the cutoff is at $\pm 0.72i$ and $\pm 0.43i$ respectively. (Note the strange extra fingers in the right halfplane for $p = 4$.)

Implicit m -step Adams methods

Larger stability domains are obtained with the implicit m -step Adams method (also called the *Adams–Moulton method*)

$$u^{n+1} = u^n + \mu \sum_{\nu=0}^m \tilde{\gamma}_\nu \Delta_{-t}^\nu u^{n+1}, \quad (5.11)$$

where the first coefficients are

$$\tilde{\gamma}_0 = 1, \tilde{\gamma}_1 = -1/2, \tilde{\gamma}_2 = -1/12, \tilde{\gamma}_3 = -1/24, \tilde{\gamma}_4 = -19/720, \tilde{\gamma}_5 = -3/160.$$

The order of accuracy is $p = m + 1$. The special choice $m = 1$ is the well known trapezoidal rule, which is A-stable. Higher order linear multistep methods cannot be A-stable according to Dahlquist's theorem:

Theorem 5.3. *Any A-stable linear multistep method has order of accuracy $p \leq 2$.*

□

Figure 5.4 shows the stability domain for $p = 3, 4, 5, 6$. Again, the size goes down for increasing p , but it is larger than for the explicit version. A closer look at the third and fourth order method reveals that the imaginary axis is not included at all in the stability domain, so it should not be used for PDE of odd order. The fifth and sixth order method on the other hand, do not only include part of the imaginary axis, but also a small part of the right half-plane. The cutoff on the imaginary axis is $\pm 1.21i$ for $p = 5$ and $\pm 1.37i$ for $p = 6$.

BDF methods

BDF stands for Backwards Differentiation Formulas, and is a class of implicit methods defined by

$$\sum_{\nu=1}^m \frac{1}{\nu} \Delta_{-t}^\nu u^{n+1} = \mu u^{n+1}. \quad (5.12)$$

The order of accuracy is $p = m$, and it is A-stable for $p \leq 2$. The first order method is the classic Euler backward method. For increasing p , the stability domain becomes smaller, and for $p \geq 7$ they go unstable even for $\mu = 0$. Unfortunately, for $p = 3, 4$, the imaginary axis is not included near the origin, which means that they are useless for first order hyperbolic systems. For $p = 5, 6$, there is a large gap on the imaginary axis making also these methods less useful for hyperbolic problems. However, for parabolic problems, they are all good, since the stability domain includes the whole negative real axis, see Figure 5.5.

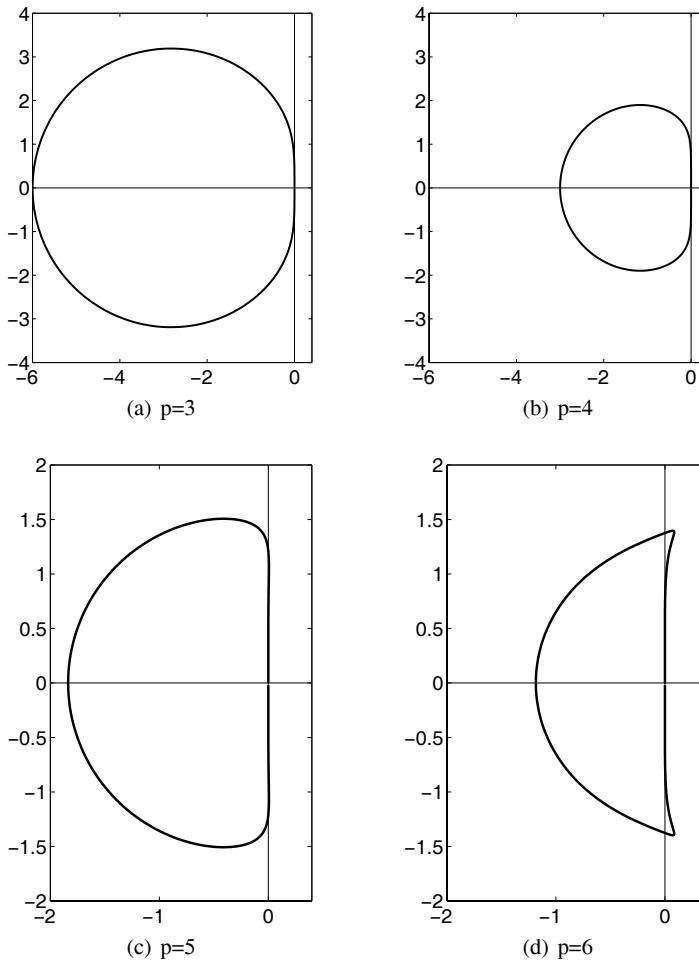


Fig. 5.4 Stability domain for implicit Adams methods of order p

The Milne method

By using the notation D_{0t} , D_{+t} , D_{-t} for the centered, forward and backward difference operator in time, we can define the Padé type difference approximation in time as

$$\frac{du}{dt} \approx \left(I + \frac{k^2}{6} D_{+t} D_{-t}\right)^{-1} D_{0t} u,$$

which has an $\mathcal{O}(k^4)$ truncation error. This leads to the implicit linear multistep method

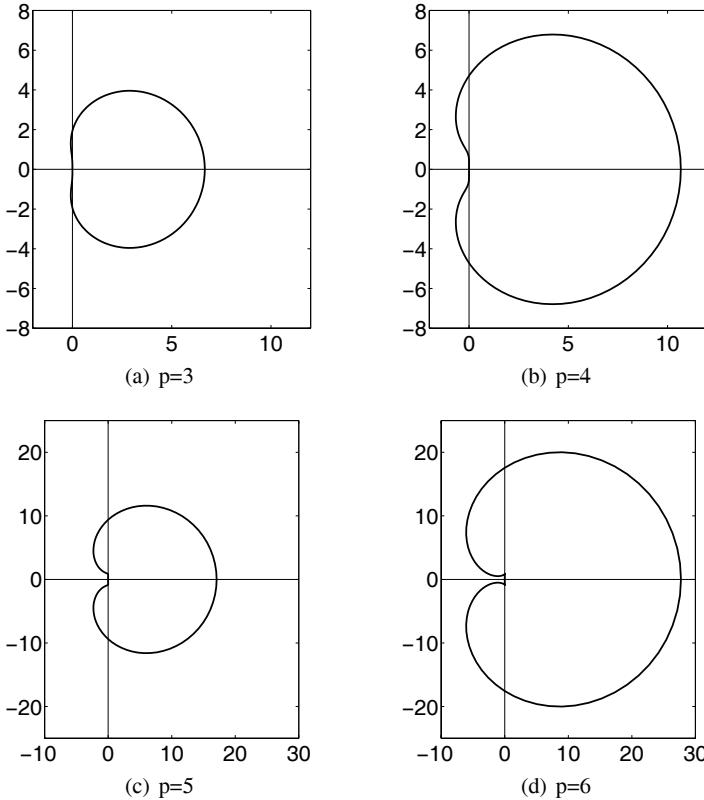


Fig. 5.5 Stability domain (outside the curve) for BDF methods of order p

$$u^{n+1} - u^{n-1} = \frac{k}{3} Q(u^{n+1} + 4u^n + u^{n-1}), \quad (5.13)$$

which is the classic Milne method. The stability domain is limited to part of the imaginary axis: $(-i\sqrt{3}, i\sqrt{3})$.

For PDE approximations, as we discuss here, implicit methods have the disadvantage that a large system of algebraic equations must be solved in each time step. This is almost always done by an iterative method. If there is a stability limit on the time step, it doesn't seem to be much to gain from an implicit method. However, if a small number of iterations is enough, it may be worthwhile. First of all, a good guess $u_{[0]}^{n+1}$ for u^{n+1} is essential, and a natural choice is $u_{[0]}^{n+1} = u^n$. A simple iteration formula is then obtained as

$$u_{[\nu+1]}^{n+1} = u^{n-1} + \frac{k}{3} (Qu_{[\nu]}^{n+1} + 4Qu_{[\nu]}^n + Qu_{[\nu]}^{n-1}), \quad \nu = 0, 1, \dots$$

It is essential to keep the number of iterations low. This can be achieved by using a more accurate method for obtaining $u_{[0]}^{n+1}$, for example by using an explicit ODE solver. This solver is then called a *predictor*, and the whole procedure is called a *predictor–corrector* method. The same principle can of course be used for any implicit method.

One should be aware, that with a finite number of iterations the method becomes explicit, and the stability domain changes accordingly.

5.4 Deferred Correction

The principle of deferred correction for improving the accuracy is well known for boundary value problems, but has until recently not been used much for initial value problems. The idea is to start out with a computed low order solution, and then raise the accuracy one or two levels by using this one for estimating the leading truncation errors. This procedure can then be continued to any order of accuracy. The advantage is that the structure and properties of the basic low order scheme to a large extent characterizes the whole algorithm.

The trapezoidal rule is A-stable, and it is a compact one step scheme, which leads to a small error constant. These are two attractive properties, in particular for PDE discretizations, and we choose it as the basic scheme.

We consider the problem

$$\begin{aligned} \frac{du}{dt} &= Qu, \\ u(0) &= f, \end{aligned} \tag{5.14}$$

and for convenience we assume here that the matrix Q is independent of t . In this section we leave out the subscript t on the difference operators, and use the notation

$$\begin{aligned} D_+ u^n &= (u^{n+1} - u^n)/k, \\ D_- u^n &= (u^n - u^{n-1})/k, \\ Du^{n+1/2} &= (u^{n+1} - u^n)/k, \\ Au^{n+1/2} &= (u^{n+1} + u^n)/2. \end{aligned}$$

The trapezoidal rule is the first step in the algorithm, and it can be written in the form

$$\begin{aligned} Du^{2,n+1/2} &= QAu^{2,n+1/2}, \\ u^{2,0} &= f, \end{aligned}$$

where the superscript 2 indicates the order of accuracy. The truncation errors are found from

$$\begin{aligned}\frac{u(t_{n+1}) - u(t_n)}{k} &= u_t(t_{n+1/2}) + \frac{k^2}{24} u_{ttt}(t_{n+1/2}) + \mathcal{O}(k^4), \\ \frac{u(t_{n+1}) + u(t_n)}{2} &= u(t_{n+1/2}) + \frac{k^2}{8} u_{tt}(t_{n+1/2}) + \mathcal{O}(k^4),\end{aligned}$$

and this is the basis for the next step. Since the solution $u^{2,n}$ is now available, it can be used for approximation of the truncation error. It is enough to approximate the derivatives u_{tt} and u_{ttt} to second order accuracy, since there is an extra factor k^2 multiplying them. The second step is

$$\begin{aligned}Du^{4,n+1/2} - \frac{k^2}{24} DD_+ D_- u^{2,n+1/2} &= QAu^{4,n+1/2} - \frac{k^2}{8} QAD_+ D_- u^{2,n+1/2}, \\ u^{4,0} &= f,\end{aligned}$$

which is a fourth order approximation. The scheme requires the vector $u^{2,-1}$, which has to be computed in some way. Using for example an explicit Runge-Kutta method for integration backwards does not work. The reason is that the resulting change of method at $t = 0$ introduces a nonsmooth truncation error, which severely affects the performance of the deferred correction method. Therefore, we use extrapolation. It can be shown that the error must be at least of order $\mathcal{O}(k^5)$, which is obtained by

$$D_+^5 u^{2,-1} = 0.$$

The trapezoidal rule is compact and requires only the storage of one vector. However, when used with deferred correction, we need also to store the lower order solutions. In the fourth order case we need $u^{2,\nu}$, $\nu = n-1, n, n+1, n+2$ in addition to $u^{4,n}$, when computing $u^{4,n+1}$.

The procedure of eliminating the truncation errors by using the lower order solution can now be continued. We separate between the approximation of the derivative and of the function itself. The coefficients c_j and d_j are determined recursively from the expansions

$$c_1 = 1,$$

$$\frac{du}{dt} = \sum_{\nu=1}^{j-1} c_\nu k^{2\nu-2} D(D_+ D_-)^{\nu-1} u + c_j k^{2j-2} \frac{d^{2j-1} u}{dt^{2j-1}} + \mathcal{O}(k^{2j}), \quad j = 2, 3, \dots, p/2,$$

$$d_1 = 1,$$

$$u = \sum_{\nu=1}^{j-1} d_\nu k^{2\nu-2} A(D_+ D_-)^{\nu-1} u + d_j k^{2j-2} \frac{d^{2j-2} u}{dt^{2j-2}} + \mathcal{O}(k^{2j}), \quad j = 2, 3, \dots, p/2.$$

With the coefficients known, the general deferred correction method of order p is defined by

$$\begin{aligned}
& Du^{2j,n+1/2} + \sum_{\nu=2}^j c_\nu k^{2\nu-2} D(D_+D_-)^{\nu-1} u^{2j-2,n+1/2} \\
& = QAu^{2j,n+1/2} + \sum_{\nu=2}^j d_\nu k^{2\nu-2} QA(D_+D_-)^{\nu-1} u^{2j-2,n+1/2}, \quad j = 1, 2, \dots, p/2, \\
& u^{2j,0} = f, \quad j = 1, 2, \dots, p/2.
\end{aligned}$$

The storage requirement is $p^2/4 + p/2 - 1$ vectors u of full length. Figure 5.6 shows the structure of the method for $p = 6$, when computing $u^{6,n+1}$. The storage is represented by the points inside the dashed polygon.

Let us now assume that the matrix Q corresponds to a semibounded operator, i.e., $(u, Qu)_h \leq 0$ for all vectors u . (We use the same notation $(\cdot, \cdot)_h$ for vectors here as we used earlier for grid functions.) The trapezoidal rule is A-stable, and furthermore, there is an estimate

$$\|u^n\|_h \leq \|f\|_h.$$

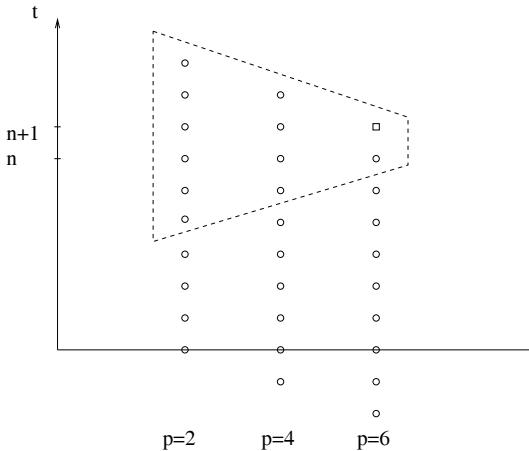


Fig. 5.6 Deferred correction, 6th order accuracy

This allows for an estimate also for the deferred correction method, and one can prove

Theorem 5.4. *Assume that the matrix Q is semibounded. Then the solutions of the deferred correction method of order p satisfies*

$$\|u^{p,n}\|_h \leq K \max_{1 \leq j \leq p/2} \|Q^{j-1} f\|_h,$$

where the constant K depends on p and t_n , but not on k or Q . □

If Q corresponds to a discretization of a differential operator in space, the estimate is independent of the step size h in space. However, one cannot call the method unconditionally stable in the sense we have discussed before, since the estimate is in terms of $\|Q^{j-1}f\|_h$ rather than in terms of $\|f\|_h$. This means that there is a condition on the smoothness of the initial data.

The final error estimate

$$\|u^{p,n} - u(t_n)\|_h = \mathcal{O}(k^p)$$

follows in the usual way from Theorem 5.4.

5.5 Richardson Extrapolation

We consider again the problem (5.14), and apply the trapezoidal rule

$$\frac{u^{n+1} - u^n}{k} = Q \frac{u^{n+1} + u^n}{2}. \quad (5.15)$$

Let $\bar{t} = nk$ be a fixed point in time, with numerical solution $u_{[k]}^n$. The method is second order accurate, and we assume that there is an expansion of the error of the type

$$u_{[k]}^n - u(\bar{t}) = ck^2 + \mathcal{O}(k^4) \quad (5.16)$$

for all n and k with $nk = \bar{t}$, where c is independent of k . This means that we also have

$$u_{[2k]}^{n/2} - u(\bar{t}) = 4ck^2 + \mathcal{O}(k^4).$$

Therefore,

$$3u(\bar{t}) = 4u_{[k]}^n - u_{[2k]}^{n/2} + \mathcal{O}(k^4),$$

which leads to

$$u(\bar{t}) = \frac{4u_{[k]}^n - u_{[2k]}^{n/2}}{3} + \mathcal{O}(k^4).$$

This is an example of *Richardson extrapolation*, which is a well known technique for a large class of computational problems, where a step size k is involved. The obvious advantage is that a 4th order solution is obtained with only one computation on the finest grid with step size k . If Q is a discretization of a differential operator in space, it is reasonable to assume that the required work for advancing the scheme to $t = \bar{t}$ is αn^2 , where α is a constant. The corresponding computation of $u_{[2k]}^{n/2}$ requires only $\alpha n^2/4$ arithmetic operations.

We know that there is an error estimate $|u^n - u(t_n)| = \mathcal{O}(k^2)$ for the trapezoidal rule and smooth solutions, but the existence of the form (5.16) is not obvious, and we shall prove it.

We make the ansatz

$$u^n - u(t_n) = k^2 v(t_n) + k^4 w^n, \quad (5.17)$$

and we want to prove that $v(t)$ is a function that is independent of k and well defined for all t . Furthermore, we shall prove that w^n is a grid function that is bounded independent of k . When using (5.15) we get

$$\begin{aligned} \frac{v(t_{n+1}) - v(t_n)}{k} - Q \frac{v(t_{n+1}) + v(t_n)}{2} &= -\frac{1}{k^2} \left(\frac{u(t_{n+1}) - u(t_n)}{k} - Q \frac{u(t_{n+1}) + u(t_n)}{2} \right) \\ &\quad - k^2 \left(\frac{w^{n+1} - w^n}{k} - Q \frac{w^{n+1} + w^n}{2} \right). \end{aligned} \quad (5.18)$$

At $t = t_{n+1/2}$ the left hand side is

$$v_t - Qv + \frac{k^2}{24} v_{ttt} - \frac{k^2}{8} Qv_{tt} + \mathcal{O}(k^4),$$

and we have the same expansion for u on the right hand side. By collecting the terms of order zero, we solve the initial value problem

$$\begin{aligned} v_t - Qv &= \frac{1}{8} Qu_{tt} - \frac{1}{24} u_{ttt}, \\ v(0) &= 0, \end{aligned} \quad (5.19)$$

where the equality $u_t = Qu$ has been used. By well known theory for ODE, we know that a smooth solution $v(t)$ exists, and obviously there is no k -dependence.

It remains to show that the grid function w^n is bounded. From (5.18) we have

$$\begin{aligned} \frac{w^{n+1} - w^n}{k} &= Q \frac{w^{n+1} + w^n}{2} + r^n, \\ w^0 &= 0, \end{aligned}$$

and since $u(t)$ and $v(t)$ are smooth, r^n is a bounded grid function. Since the trapezoidal rule is unconditionally stable, there is an estimate of w^n in terms of r^n , and it follows that w^n is bounded. This completes the proof that the expansion (5.17) exists, which allows for the Richardson extrapolation.

The process can of course be continued, such that even higher order approximations can be obtained. If there is an expansion

$$u_{[k]}^n - u(t_n) = \sum_{\nu=1}^{p/2} k^{2\nu} v^{(\nu)}(t_n) + k^{p+2} w^n,$$

then we can combine the computed solutions $u_{[k]}^n, u_{[2k]}^{n/2}, u_{[3k]}^{n/3}, \dots$ to obtain approximations of any even order.

There are of course practical limits for how far we can go. One limit is given by the smoothness of the solution. The Taylor expansion of the truncation error is necessary, and the higher order derivatives may not exist, or if they do, they may not be well resolved by the coarser grids. In other words, if the grid is coarse enough, then terms of formally high order in k may be as large as the lower order ones, and the asymptotic expansion is of no use.

One should note the importance of having a zero initial condition for the initial value problem (5.19), or the corresponding problems for higher accuracy. This allows for the necessary k -independence of the solution $v(t)$. For a multistep scheme, a special numerical method must be used for starting up the scheme, and then we don't have the simple form of the initial condition any longer. This is one reason for using the trapezoidal rule as the basis for Richardson extrapolation. Another reason is that we gain two orders of accuracy for each step.

5.6 Summary

In this chapter we have given a survey of the most common difference methods, which are of relevance for time discretizations of PDE. Runge-Kutta and linear multistep methods are the two dominating classes, and our presentation of these is concentrated on stability domains and accuracy. Discretization in space by centered approximations for periodic problems leads to quite simple forms of the Fourier transforms, i.e., of the eigenvalues of the space operator. A first choice of method can be made on the basis of the test equation, but when boundary conditions enter the problem, a more thorough analysis has to be made. Indeed, the very common use of the method of lines, is very seldom based on a strict stability analysis. If an energy estimate is not available, an eigenvalue analysis of the space operator with constant coefficients is sometimes performed, most often numerically. The time step is then chosen such that these eigenvalues (after multiplication with the time step) are inside the stability domain for the test equation. However, to be sure about stability, the theorems 5.1 proven in [Kreiss and Wu, 1993], or 5.2 proven in [Tadmor, 2002], should be applied. There are similar theorems also for linear multistep methods. These sufficient conditions are more restrictive than those based on the eigenvalue analysis, and from practical experience we know that they are sometimes not necessary. General necessary and sufficient conditions for the method of lines were actually given by Reddy and Trefethen in [Reddy and Trefethen, 1992]. Their theory is based on the concept of *pseudo-eigenvalues*, but in order to find these for a given approximation, one has in most cases to rely upon numerical estimates.

Discretizations in space of PDE give rise to very stiff systems of ODE, i.e., the eigenvalues of the coefficient matrix have a large span, and they are proportional to $1/h^r$, where r is the order of the differential operator. This means that explicit methods can be used for $r = 1$, but they usually become too expensive for $r \geq 2$. The ideal situation would be if the method is A-stable, i.e., there is no stability limit on the time step. Unfortunately, linear multistep methods suffer from the Dahlquist

barrier, which says that the order of accuracy cannot be higher than 2. In this book we are advocating higher order methods, and if we insist upon unconditional stability, we must give up the linear multistep methods. One choice is then implicit Runge-Kutta methods, another one is deferred correction methods. In the latter case the implementation is quite easy, but the amount of storage limits the order to 4 or possibly 6 for large scale practical applications.

The crucial question of barriers on accuracy and stability has been thoroughly investigated by R. Jeltsch and O. Nevanlinna, see [Jeltsch and Nevanlinna, 1981], [Jeltsch and Nevanlinna, 1982] and [Jeltsch and Nevanlinna, 1983]. In [Jeltsch and Smit, 1987] and [Jeltsch and Smit, 1998], there is a special treatment of barriers for explicit approximations of hyperbolic equations. For explicit one step “upwind” schemes of the form

$$u_j^{n+1} = \sum_{\nu=0}^r \alpha_\nu u_{j+\nu}^n$$

with order of accuracy p , the barrier is

$$p \leq \min(r, 2).$$

This means that explicit upwind schemes with order higher than two are unstable.

Difference methods for ordinary differential equations has been an area of research for a long time, and there is an abundance of methods and theory. Some of the methods are associated with mathematicians who worked long before computers existed, for example J.C. Adams, F. Bashforth, M.V. Kutta and C. Runge. The first introduction of Runge-Kutta methods was made by Runge [Runge, 1895], and then Kutta [Kutta, 1901] formulated the general class. The Milne method was introduced 1926 by Milne [Milne, 1926]. The Richardson extrapolation is due to L. F. Richardson, who is well known for introducing numerical weather prediction long before computers existed, see [Richardson, 1922] (more recent editions are available). Later, in the second half of the previous century, G. Dahlquist and J. Butcher did some of the most important work on ODE. Dahlquist published a number of important papers on linear multistep methods, and his paper [Dahlquist, 1963] on the stability barrier for A-stable schemes was published 1963. Butcher has been the outstanding researcher on Runge-Kutta methods for many decades, and has written several books. His latest book Numerical Methods for Ordinary Differential Equations [Butcher, 2003] came out as late as 2003. If the books [Hairer et al., 1993] and [Hairer and Wanner, 1996] by Hairer et.al. are added to this one, there is not much more that is needed for getting a complete picture of ODE approximations. However, the results on deferred correction methods are quite new, and are found in [Gustafsson and Kress, 2001]. A special investigation for application of these methods to initial-boundary value problems is presented in [Kress and Gustafsson, 2002].

Chapter 6

Coupled Space-Time Approximations

In the previous chapters we have separated the discretization in space and time. This does not give all possible discretizations of a time dependent PDE, and in this chapter we shall discuss other possibilities. We shall assume throughout this chapter that the solutions are periodic.

6.1 Taylor Expansions and the Lax–Wendroff Principle

Assume that we have a linear PDE

$$u_t = P(\partial/\partial x)u$$

and a discretization in space

$$\frac{du}{dt} = Qu, \quad (6.1)$$

where u is a vector containing the values at the grid points, and Q is the matrix representation of a difference operator, which is independent of t . Looking for a one step method, we start from a Taylor expansion in time

$$u(t_{n+1}) = u(t_n) + k \frac{du}{dt}(t_n) + \frac{k^2}{2} \frac{d^2u}{dt^2}(t_n) + \mathcal{O}(k^3). \quad (6.2)$$

It is natural to substitute du/dt by Qu . For d^2u/dt^2 , we differentiate (6.1), and obtain

$$\frac{d^2u}{dt^2} = Q \frac{du}{dt} = Q^2 u.$$

The resulting approximation is

$$u^{n+1} = (I + kQ + \frac{k^2}{2}Q^2)u^n, \quad (6.3)$$

which has an $\mathcal{O}(k^2)$ truncation error plus an error from the space discretization Q . Clearly, this procedure can be generalized to any order in time; the scheme is

$$u^{n+1} = \sum_{\nu=0}^m \frac{k^\nu}{\nu!} Q^\nu u^n. \quad (6.4)$$

For this linear case, this type of approximation is very similar to explicit Runge–Kutta methods. In fact, for $1 \leq m \leq 4$, the method (6.4) is equivalent with the standard m stage Runge–Kutta methods described above.

If $Q = Q(t)$, the same principle can be applied, but the scheme becomes more complicated, since extra terms containing time derivatives of Q must be included.

One disadvantage with (6.4) is that the computational stencil becomes very wide for large m , and this complicates the implementation when nonperiodic boundaries are involved. For example, if Q is the centered fourth order approximation of $\partial/\partial x$, the width of the stencil is $4m + 1$ points.

A more compact scheme is obtained by using a different principle, while still keeping the Taylor expansion as a basis. We use the simple model problem

$$u_t = a(x)u_x,$$

and start from the second order Taylor expansion (6.2). Instead of using the semidiscrete approximation, we use the original differential equation, and get

$$u(x, t_{n+1}) = u(x, t_n) + kau_x(x, t_n) + \frac{k^2}{2}a(au_x(x, t_n))_x + \mathcal{O}(k^3).$$

To obtain a second order approximation also in space, we use difference approximations that are as compact as possible and centered at $x = x_j$:

$$u_j^{n+1} = u_j^n + ka_j D_0 u_j^n + \frac{k^2}{2} a_j D_+(a_{j-1/2} D_- u_j^n).$$

This is the famous Lax–Wendroff scheme. The computational stencil is only 3 points wide, compared to 5 points for the corresponding straightforward method above, where D_0^2 would have been involved. But the most significant difference is that the Lax–Wendroff scheme is stable for

$$k \max_x (a(x)) \leq h,$$

while (6.3) is unstable for all k . (The stability domain for the second order Taylor expansion does not include the imaginary axis.)

By adding more terms in the Taylor expansion (6.2), the accuracy can be raised to any order. For convenience, we let $a(x)$ be a constant a . The third order correction term in time is obtained by applying the Lax–Wendroff principle as

$$\frac{k^3}{6} u_{ttt} = \frac{k^3}{6} a^3 u_{xxx} = \frac{k^3}{6} a^3 D_0 D_+ D_- u + \mathcal{O}(h^2 k^3).$$

It is reasonable to raise the order of accuracy in space as well. In principle it would be enough to go to third order, but the scheme is kept symmetric in space if fourth order operators are used. We get

$$u^{n+1} = \left(I + kaD_0(I - \frac{h^2}{6}D_+D_-) + \frac{k^2 a^2}{2}D_+D_-(I - \frac{h^2}{12}D_+D_-) + \frac{k^3 a^3}{6}D_0D_+D_-\right) u^n.$$

The local truncation error is $k\mathcal{O}(h^4 + kh^4 + k^2h^2 + k^3)$, and if k/h is constant, we have an $\mathcal{O}(h^3)$ approximation.

The amplification factor is

$$\hat{Q} = 1 + \lambda i \sin \xi \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} \right) - 2\lambda^2 \sin^2 \frac{\xi}{2} \left(1 + \frac{1}{3} \sin^2 \frac{\xi}{2} \right) - \frac{2}{3} \lambda^3 i \sin \xi \sin^2 \frac{\xi}{2},$$

where $\lambda = ka/h$. The von Neumann condition is $\lambda \leq 0.865$, i.e., there is a slightly stronger stability restriction compared to the classic second order Lax–Wendroff scheme. However, the third order accuracy allows for a coarser grid, and there should be a clear gain in efficiency.

Since 4th order operators were used in space, it is easy to raise the accuracy one more step. We have only to add a term corresponding to the fourth term in the Taylor expansion:

$$\frac{k^4}{24}u_{ttt} = \frac{k^4}{24}a^4u_{xxxx} = \frac{k^4}{24}a^4(D_+D_-)^2u + \mathcal{O}(h^2k^4).$$

The resulting scheme is

$$\begin{aligned} u^{n+1} = & \left(I + kaD_0(I - \frac{h^2}{6}D_+D_-) + \frac{k^2 a^2}{2}D_+D_-(I - \frac{h^2}{12}D_+D_-) \right. \\ & \left. + \frac{k^3 a^3}{6}D_0D_+D_- + \frac{k^4}{24}a^4(D_+D_-)^2 \right) u^n. \end{aligned}$$

The von Neumann condition is in this case $\lambda \leq 0.919$. Obviously, this scheme is superior to the 3rd order one, since the computational stencil is not any wider, and the stability limit is actually slightly less restrictive.

6.2 Implicit Fourth Order Methods

Consider the equation

$$u_t + a(x)u_x = 0, \quad (6.5)$$

and discretize first in time by the trapezoidal rule:

$$\frac{u^{n+1} - u^n}{k} + a \frac{\partial}{\partial x} \frac{u^{n+1} + u^n}{2} = 0.$$

The approximation is centered at $t = t_{n+1/2}$, and the truncation error is

$$\frac{k^2}{24}u_{ttt} + a\frac{k^2}{8}u_{xtt} + \mathcal{O}(k^4). \quad (6.6)$$

The standard Crank–Nicholson scheme is obtained by substituting $\partial/\partial x$ by D_0 . For constant coefficients $a_j = a$, the amplification factor is

$$\hat{Q} = \frac{b - ic}{b + ic}, \quad (6.7)$$

where $b = 1$ and $c = \lambda i \sin \xi$, $\lambda = ka/h$, and unconditional stability follows.

In the next version we accept the second order error in time, but raise the accuracy in space by replacing D_0 by the operator $Q_4 = D_0(I - \frac{h^2}{6}D_+D_-)$. For convenience, we define the operator (or matrix) A by

$$A = \text{diag}(a_1 a_2 \dots a_N),$$

resulting in

$$(I + \frac{k}{2}AQ_4)u^{n+1} = (I - \frac{k}{2}AQ_4)u^n.$$

The amplification factor still has the form (6.7), but now with

$$c = \lambda i \sin \xi \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2}\right),$$

and we have again an unconditionally stable scheme.

In order to keep the scheme more compact, we replace Q_4 by the Padé type approximation

$$Q_4^{[P]} = P^{-1}D_0 := (I + \frac{h^2}{6}D_+D_-)^{-1}D_0$$

as discussed in Section 4.3. The scheme is

$$(I + \frac{k}{2}AP^{-1}D_0)u^{n+1} = (I - \frac{k}{2}AP^{-1}D_0)u^n. \quad (6.8)$$

Since the amplification factor has the form (6.7) with real b and c , unconditional stability holds even in this case.

There is one complication here. The coefficient matrix multiplying u^{n+1} is dense, and it is not a good idea to solve the system (6.8) as it stands. Furthermore, the trick to multiply the equation (6.8) from the left by P doesn't work. A factor PAP^{-1} occurs, and since P and A do not commute, we don't get rid of the matrix P^{-1} . However, by a slight reformulation, we can solve the equation without iteration. When dropping the superscript on u^{n+1} , the system to be solved has the form

$$(I + \frac{k}{2}AP^{-1}D_0)u = F,$$

where the right hand side is known. For periodic problems the operators P and D_0 commute, i.e., $D_0P - PD_0 = 0$. Since

$$P^{-1}D_0 - D_0P^{-1} = P^{-1}(D_0P - PD_0)P^{-1} = 0,$$

it follows that also P^{-1} and D_0 commute. Therefore we can write the system as

$$(I + \frac{k}{2}AD_0P^{-1})u = F.$$

With $v = P^{-1}u$, the equation is

$$(P + \frac{k}{2}AD_0)v = F.$$

This tridiagonal system is solved for v , and $u = Pv$ is then computed. The system is well conditioned, since the magnitude of the Fourier transform of the coefficient matrix for constant A is

$$|\hat{P} + \frac{k}{2}A\hat{D}_0| = \left| \frac{2 + \cos \xi}{3} + \frac{\lambda}{2}i \sin \xi \right| \geq \frac{1}{3}.$$

In the next version, we aim for 4th order accuracy also in time, while still keeping the two-level structure of the scheme. Therefore, the time derivatives in the truncation error (6.6) are converted into space derivatives:

$$\begin{aligned} u_{tt} &= -au_{xt} = a(au_x)_x, \\ u_{ttt} &= a(au_{tx})_x. \end{aligned}$$

The truncation error is

$$\frac{k^2}{24}u_{ttt} + \frac{k^2}{8}au_{xtt} + \mathcal{O}(k^4) = -\frac{k^2}{12}a(au_{xt})_x + \mathcal{O}(k^4).$$

Since there is a factor k^2 in the leading part of the truncation error, it is enough to use a second order approximation of the time derivative to obtain 4th order. The resulting discretization in time is

$$\frac{u^{n+1} - u^n}{k} + a \frac{\partial}{\partial x} \frac{u^{n+1} + u^n}{2} + \frac{k^2}{12}a \frac{\partial}{\partial x} \left(a \frac{\partial}{\partial x} \frac{u^{n+1} - u^n}{k} \right) = 0.$$

It remains to approximate in space. One possibility is to use the 4th order operator Q_4 for $\partial/\partial x$ as shown above, and the standard 2nd order formula $a_jD_+ + a_{j-1/2}D_-$ for $(\partial/\partial x)a\partial/\partial x$. We get

$$(I + \frac{k^2}{12}a_jD_+ + a_{j-1/2}D_-)(u_j^{n+1} - u_j^n) + \frac{k}{2}a_jD_0(I - \frac{h^2}{6}D_+D_-)(u_j^{n+1} + u_j^n) = 0.$$

For a constant coefficient $a_j = a$, the scheme simplifies to

$$(I + \frac{k^2}{12}a^2D_+D_-)(u^{n+1} - u^n) + \frac{k}{2}aD_0(I - \frac{h^2}{6}D_+D_-)(u^{n+1} + u^n) = 0,$$

where the space index j has been omitted. It is easy to see that the amplification factor in Fourier space has the form (6.7) with real b and c , i.e., $|\hat{Q}| = 1$ independent of $\lambda = ak/h$. Accordingly, the scheme is unconditionally stable and energy conserving also in this case.

In order to construct a formula that is even more compact, we use the Padé difference operator $Q_4^{[P]}$ for $\partial/\partial x$ as above, and obtain

$$(I + \frac{k^2}{12} a_j D_+ a_{j-1/2} D_- P^{-1})(u_j^{n+1} - u_j^n) + \frac{k}{2} a_j P^{-1} D_0 (u_j^{n+1} + u_j^n) = 0. \quad (6.9)$$

Note that an extra operator P^{-1} has been introduced in the $\mathcal{O}(k^2)$ correction term. This can be done without altering the order of accuracy, since $P^{-1} = I + \mathcal{O}(h^2)$ when applied to smooth functions. At a first glance it seems like an extra complication. However, as will be shown below, it actually simplifies the solution procedure.

For constant coefficients a , the scheme is

$$(I + \frac{k^2}{12} a^2 D_+ D_- P^{-1})(u^{n+1} - u^n) + \frac{k}{2} a P^{-1} D_0 (u^{n+1} + u^n) = 0.$$

The Fourier transform of $D_+ D_- P^{-1}$ is real, while the transform of $P^{-1} D_0$ is purely imaginary, so even in this case the amplification factor has the form (6.7) with real b and c , showing unconditional stability and conservation of the norm.

We are again encountering the same problem as above, when it comes to solving for u^{n+1} , but there is an easy fix also in this case. In addition to the matrix A used above, we define the shifted matrix

$$A_{-1/2} = \text{diag}(a_{1/2} a_{3/2} \dots a_{N-1/2}),$$

with periodicity conditions applied. We also replace $P^{-1} D_0$ by $D_0 P^{-1}$, just as above. With $v = P^{-1} u^{n+1}$ and

$$R = \frac{k^2}{12} A D_+ A_{-1/2} D_-,$$

the solution is obtained in two steps as

$$\begin{aligned} (P + R + \frac{k}{2} A D_0) v &= F, \\ u^{n+1} &= Pv. \end{aligned}$$

The Fourier transform of the coefficient matrix for constant A is

$$\hat{q}(\xi, \lambda) = \frac{1}{6}(4 - \lambda^2) + \frac{1}{6}(2 + \lambda^2) \cos \xi + i \frac{\lambda}{2} \sin \xi.$$

We have $\hat{q} = 0$ at $\xi = \pi$ if $\lambda = 1$, so for the periodic case, we must stay away from this λ -value. For initial-boundary value problems this difficulty may not be present, but on the other hand, we must make sure that P and D_0 commute in that case.

We shall now make some comparisons of accuracy. For $a = 1$ and $u(x, 0) = e^{i\omega x}$, the true solution of (6.5) is $e^{i\omega(x-t)}$, and during one time step, the solution changes to $e^{i\omega(x-t-k)} = e^{-i\omega k} e^{i\omega(x-t)}$. This can be interpreted as advancing the solution in Fourier space according to the relation

$$\hat{u}(\omega, t+k) = e^{-i\omega k} \hat{u}(\omega, t) = e^{-i\lambda\omega h} \hat{u}(\omega, t).$$

On the discrete side, the corresponding relation is

$$\hat{u}^{n+1}(\xi) = \hat{Q}(\xi) \hat{u}^n(\xi),$$

where $\hat{Q} \approx e^{-i\lambda\omega h}$. We know already that $|\hat{Q}| = 1$ for all the difference schemes in this section. Therefore we can write

$$\hat{Q} = e^{-i\lambda\phi(\xi)},$$

where $\phi(\xi)$ is real. For the discrete case we have $|\xi| = |\omega h| \leq \pi$, and therefore it is natural to compare ξ with $\phi(\xi)$, just as we did in Section 1.2. We compare the standard Crank–Nicholson second order scheme with the two schemes (6.8) and (6.9) based on Padé approximations. We have in Fourier space

$$\begin{aligned}\hat{Q}_{2,2}(\xi) &= \frac{1 - i\frac{\lambda}{2} \sin \xi}{1 + i\frac{\lambda}{2} \sin \xi}, \\ \hat{Q}_{4,2}(\xi) &= \frac{\frac{1}{5}(4 + 2 \cos \xi) - i\frac{\lambda}{2} \sin \xi}{\frac{1}{5}(4 + 2 \cos \xi) + i\frac{\lambda}{2} \sin \xi}, \\ \hat{Q}_{4,4}(\xi) &= \frac{\frac{1}{5}(4 - \lambda^2 + (2 + \lambda^2) \cos \xi) - i\frac{\lambda}{2} \sin \xi}{\frac{1}{5}(4 - \lambda^2 + (2 + \lambda^2) \cos \xi) + i\frac{\lambda}{2} \sin \xi}.\end{aligned}\tag{6.10}$$

Here the subscript (p, q) denotes the order of accuracy in (space, time) for the particular scheme. Figure 6.1 shows $\phi_{p,q}(\xi)$ for $\lambda = 0.8$ and $\lambda = 1.5$.

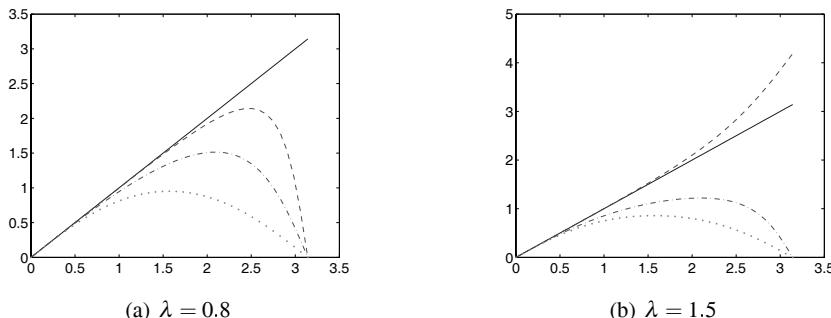


Fig. 6.1 $\phi_{p,q}(\xi)$, $(p, q) = (2, 2) (\cdots), (4, 2) (-\cdots), (4, 4) (---), (2, 4) (\cdot)$

For a real computation, it is reasonable to assume that there are at least 4 grid points per wave length for the highest wave number we are interested in. This means that the analysis can be limited to the interval $0 \leq \xi \leq \pi/2$. As shown in the figure above, the accuracy depends on λ . Furthermore, for each λ , the error is increasing with increasing ξ , i.e., the maximum occurs at $\xi = \pi/2$. Figure 6.2 shows the quotient $\phi(\pi/2)/(\pi/2)$ for $0.5 < \lambda \leq 2.5$; ideally it should be one.

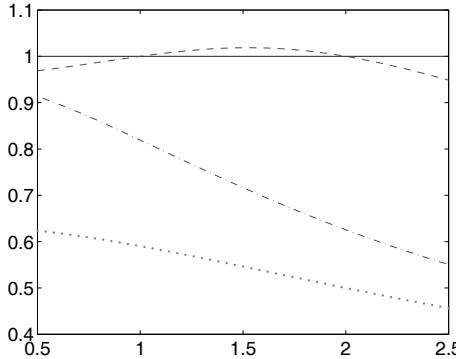


Fig. 6.2 $\phi_{p,q}(\pi/2)/(\pi/2)$, $(p, q) = (2, 2) (\cdots)$, $(4, 2) (-\cdot)$, $(4, 4) (--)$

There is the expected significant differences between the three schemes, and the $(4, 4)$ scheme is remarkably accurate for a wide range of λ -values.

In Section 1.2 higher order approximations in space were compared for different degrees of smoothness of the solution. For time integration, the MATLAB ode45 solver was used. That solver tries to use the best choice of time step for keeping the accuracy, and it is at least 4th order in time. It is of interest to see how these new schemes behave in comparison, when the time-step is fixed, so they were run for the same problem (1.5) as in Figures 1.2 and 1.3. Figures 6.3 and 6.4 show the result of the $(4, 2)$ and $(4, 4)$ schemes (6.8) and (6.9) respectively with $\lambda = 0.8$. As expected, the accuracy of the $(4, 2)$ scheme is somewhere in between the second and fourth order (space accuracy) schemes for both cases $r = 1$ and $r = 3$ in Figures 1.2 and 1.3. Furthermore, the $(4, 4)$ scheme is at least as good as the fourth order scheme with the ode45 solver. In fact, it is hard to distinguish the numerical solution from the true one for $r = 3$. Note that we have not picked the best λ -value, which is $\lambda = 1$. In that case the solution is exact for constant coefficients a_j , since $u_j^{n+1} = u_{j-1}^n$.

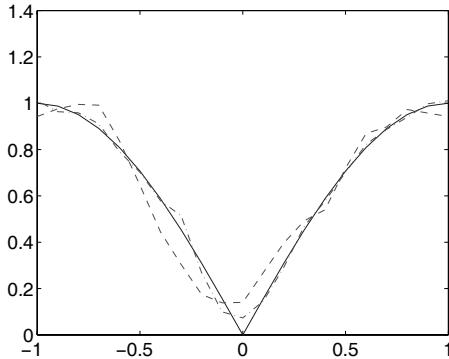


Fig. 6.3 $u(x, 6)$, $r = 1$ (—), $(4, 2)$ scheme (---), $(4, 4)$ scheme (—·), $N = 20$, $\lambda = 0.8$

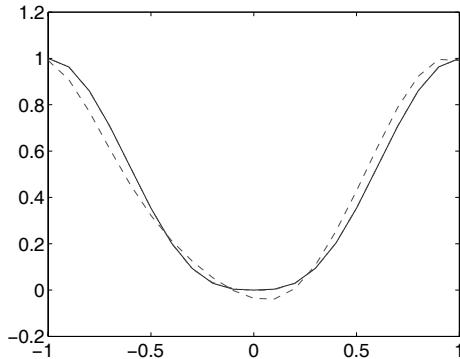


Fig. 6.4 $u(x, 6)$, $r = 3$ (—), $(4, 2)$ scheme (---), $(4, 4)$ scheme (—·), $N = 20$, $\lambda = 0.8$

In order to further demonstrate the strength of the $(4, 4)$ scheme, we ran a pulse for 30 periods, i.e., until $t = 60$, and compare it to the $(4, 2)$ scheme. The result is shown in Figure 6.5. There is now a dramatic difference between the two. The 2nd order error in time causes the solution to become completely out of phase. Both solutions are computed in 1500 steps. A similar accuracy with the $(4, 2)$ scheme requires a finer grid in space and a much smaller time step. A rough estimate based on experiments shows that an order of magnitude more work is required in the 1-D case, and the difference would of course be even larger in several space dimensions.

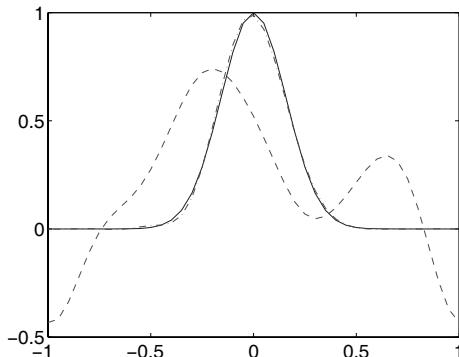


Fig. 6.5 True solution (—), (4, 2) scheme (---), (4, 4) scheme (-·-), $t = 60$, $\lambda = 0.8$, $N = 40$

6.3 Summary

The principle of discretizing first in space and then in time, does not give all possibilities. Here we have shown how to use a more flexible construction principle. Starting out from a Taylor expansion in time, we get high order derivatives involved for high order accuracy. Aiming for a one step scheme, the time derivatives are substituted by space derivatives by using the differential equation. This principle leads in the first step to the 2nd order Lax–Wendroff scheme [Lax and Wendroff, 1964], which has been successfully used for quite general, even nonlinear, problems. The MacCormack scheme is different, but uses essentially the same principle, see [MacCormack, 1969] and [MacCormack and Pauly, 1972]. This scheme was for a long time the dominating one for the Euler and Navier–Stokes equations in aerodynamics, see further Chapter 11.

Here we have generalized the same principle to higher order accuracy. It was shown that a 4th order scheme in space and time obtained this way, has a quite favorable stability limit.

These explicit schemes are of interest mainly for hyperbolic equations. In our example, the equation has variable coefficients in space, but not in time, and this is a large class of problems by itself. For time dependent coefficients, the higher order versions can still be applied, but more terms enter into the difference approximation.

We have also derived an implicit scheme which is 4th order accurate in space and time. The version that uses compact Padé approximations in space has remarkable accuracy. The gain when adding the extra terms for raising the accuracy in time to the same level as in space, is significant, and the (4, 4) scheme should clearly be used for long time integrations.

Our discussion in this chapter has been limited to periodic solutions, but the schemes can be generalized to initial-boundary value problems as well. However, there are some details that have to be considered. For example, we have been using

the commutation property $PD_0 = D_0P$ for the Padé approximations, and the construction near the boundaries must be constructed carefully when the solutions are no longer periodic.

Chapter 7

Boundary Treatment

The construction of spatial difference approximations on structured grids is quite straightforward as long as we stay away from the boundaries; this was demonstrated in Chapter 4. The boundary conditions are more complicated, in particular for higher order approximations. Even if a high order approximation of a given boundary condition is easy to construct, we need extra numerical boundary conditions to define a unique numerical solution, since the computational stencils are wide and introduce extra boundary points. These have to be accurate enough, and they have to be such that the full approximation is stable. In this chapter we shall discuss various ways of generating numerical boundary conditions, concentrating mainly on first order hyperbolic systems, where the difficulties are most distinctive.

7.1 Numerical Boundary Conditions

We shall start with a fourth order approximation to a simple parabolic example, where we give a recipe for constructing numerical boundary conditions. We choose a semi-discrete quarter space problem such that we can concentrate on one boundary, and for convenience it is assumed that $u(x, t) = 0$ for $x \geq \bar{x}$. The problem is

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(0, t) &= g(t), \\ u(x, 0) &= f(x), \end{aligned}$$

and investigate the fourth order approximation

$$\begin{aligned} \frac{du_j}{dt} &= (D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2) u_j, \quad j = 1, 2, \dots, \\ u_0(t) &= g(t), \\ u_j(0) &= f(x_j). \end{aligned}$$

The five-point formula applied at $x = x_1$ involves the points x_{-1} and x_0 , but we have only one boundary condition. The idea is to use the differential equation to generate the extra boundary condition. The boundary condition can be differentiated with respect to t , giving $du_0/dt = g'$, and the differential equation then implies

$$u_{xx}(0, t) = g'(t).$$

Going one step further, we get

$$u_{xxxx}(0, t) = g''(t).$$

We know that

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

and therefore we use the extra boundary condition

$$D_+ D_- u_0(t) = g'(t) + \frac{h^2}{12} g''.$$

A smooth solution requires that there is a certain compatibility at the corner ($x = 0, t = 0$), and here we assume that $d^r f / dx^r = d^r g / dt^r = 0$ for $r = 0, 1, \dots, r^*$ for r^* sufficiently large. The error $w_j(t) = u(x_j, t) - u_j(t)$ satisfies

$$\begin{aligned} \frac{dw_j}{dt} &= (D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2) w_j + h^4 \tilde{F}_j, \quad j = 1, 2, \dots, \\ w_0(t) &= 0, \\ D_+ D_- w_0(t) &= h^4 \tilde{g}(t), \\ w_j(0) &= 0, \end{aligned}$$

where \tilde{F} and \tilde{g} are bounded. It can be shown that the difference operator is semi-bounded with the homogeneous boundary conditions, but strong stability is not known. Therefore, in order to derive an error estimate, we try to eliminate the boundary data by constructing a function $\phi_j(t)$ that satisfies

$$\begin{aligned} \phi_0(t) &= 0, \\ D_+ D_- \phi_0(t) &= \tilde{g}(t), \\ \phi_j(0) &= 0, \\ ||\phi(t)||_h &< \infty. \end{aligned} \tag{7.1}$$

The second condition of (7.1) is normalized properly since $\tilde{g}(t)$ is bounded and smooth as a function of t , and ϕ can therefore be chosen and extended easily for increasing x_j such that $d\phi/dt$ and $(D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2)\phi$ are bounded. The difference $v = w - h^4 \phi$ satisfies

$$\begin{aligned}\frac{dv_j}{dt} &= (D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2) v_j + h^4 \tilde{G}_j, \quad j = 1, 2, \dots, \\ v_0(t) &= 0, \\ D_+ D_- v_0(t) &= 0, \\ v_j(0) &= 0,\end{aligned}$$

where \tilde{G} is a bounded function that depends on ϕ . The forcing function $h^4 \tilde{G}_j$ is driving the solution by itself, no other data are involved. By the results in Section 2.3.2 it follows that here is an $\mathcal{O}(h^4)$ estimate of the solution v . Since $w = v + h^4 \phi$, where ϕ is bounded, the final estimate

$$\|w\|_h \leq K_1 h^4$$

follows by the triangle inequality.

With the boundary condition $u_x(0, t) = g(t)$, we can proceed in the same way, and obtain a fourth order error estimate with

$$\begin{aligned}D_+ u_0(t) &= g(t) + \frac{h^2}{24} g'(t), \\ D_+^2 D_- u_0(t) &= g'(t) + \frac{h^2}{8} g''(t).\end{aligned}$$

Here it is assumed that the grid is located such that $x_0 = -h/2$, $x_1 = h/2$. The principle used for this example can be used for any order of accuracy for parabolic problems. The boundary conditions are differentiated with respect to t , and the derivatives of the solution u are then converted into derivatives with respect to x by using the differential equation. In that way we obtain a relation between these space derivatives and known boundary data, and that relation is discretized. For a real world problem, this procedure is of course technically more complicated, since the conversion from time to space derivatives introduces more terms. However, it is a one time effort.

We next turn to the hyperbolic case, and we choose an example with variable coefficients:

$$\begin{aligned}u_t &= a(x) u_x, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(0, t) &= g(t), \\ u(x, 0) &= f(x).\end{aligned}$$

Here it is assumed that $a(0) < 0$, such that the problem is well posed. The standard centered fourth order approximation uses 5 points, and we get the same kind of situation as in the previous parabolic case with the need for an extra numerical boundary condition. The differentiation procedure provides the relations

$$\begin{aligned} u_x &= a^{-1} u_t, \\ u_{xx} &= (a^{-1})_x u_t + a^{-2} u_{tt}, \\ u_{xxx} &= (a^{-1})_{xx} u_t + (a^{-1}(a^{-1})_x + (a^{-2})_x) u_{tt} + a^{-3} u_{ttt}. \end{aligned}$$

We use a fourth order approximation of the first equation at the boundary. The approximation of u_x introduces a term containing u_{xxx} , and we use the third equation to obtain the complete approximation

$$\begin{aligned} \frac{du_j}{dt} &= D_0(I - \frac{h^2}{6}D_+D_-)u_j, \quad j = 1, 2, \dots, \\ u_0(t) &= g(t), \\ D_0u_0(t) &= a^{-1}g'(t) + \frac{h^2}{6}\left((a^{-1})_{xx}g'(t) + (a^{-1}(a^{-1})_x + (a^{-2})_x)g''(t) + a^{-3}g'''(t)\right), \\ u_j(0) &= f(x_j). \end{aligned}$$

The procedure for estimating the error $w_j(t) = u(x_j, t) - u_j(t)$ is the same as for the previous example with subtraction of a certain function $h^4\phi$ from w . The crucial equation is this time

$$D_0\phi_0(t) = \tilde{g}(t),$$

which again can be satisfied by a smooth function such that $d\phi/dt$ and $D_0(I - \frac{h^2}{6}D_+D_-)\phi$ are bounded. The final error w is $\mathcal{O}(h^4)$.

In contrast to the parabolic case, the procedure cannot be generalized to hyperbolic systems. The reason is that there are fewer boundary conditions than the number of dependent variables, unless all the eigenvalues of the coefficient matrix have the same sign with only ingoing characteristics. Therefore, we shall describe another general method in the next section.

7.2 Summation by Parts (SBP) Difference Operators

The model “outflow” problem

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(x, 0) &= f(x), \end{aligned}$$

has no boundary condition at $x = 0$. (Also here we assume for convenience that $u(x, t) = 0$ for $x \geq \bar{x}$.) When using the standard scalar product

$$(u, v) = \int_0^\infty u(x)v(x)dx,$$

The differential operator $\partial/\partial x$ is semibounded, and

$$(v, \frac{\partial}{\partial x} v) = -\frac{1}{2}|v(0)|^2 \quad (7.2)$$

for all functions $v(x)$ with $v(x) = 0$ for $x \geq \bar{x}$. This leads to the energy estimate

$$\frac{d}{dt}||u||^2 = -|u(0,t)|^2.$$

The basis for summation by parts operators D is the requirement that they satisfy the property (7.2) in the discrete sense:

$$(v, Dv)_h = -\frac{1}{2}|v_0|^2. \quad (7.3)$$

We say that the difference operator D is *strictly semibounded*. No inner points occur at the right hand side of (7.3), and therefore the difference operator must be based on a centered skewsymmetric operator at inner points. Because of the centered structure, the order of accuracy at inner points is even, and we denote it by $p = 2s$. Furthermore, in order to recover the full order of accuracy, we would like it to be accurate at least of order $2s - 1$ near the boundary. Then there are three questions that are fundamental in the construction:

- i) Do such operators exist?
- ii) How should the discrete scalar product $(\cdot, \cdot)_h$ be constructed near the boundary?
- iii) How should the centered difference operator be modified near the boundary?

In Section 2.3.2 we used the example $p = 2$ for illustration of stability. With the right boundary removed, the problem is

$$\begin{aligned} \frac{du_j}{dt} &= Du_j, \quad j = 0, 1, \dots, \\ u_j(0) &= f_j, \end{aligned}$$

where

$$Du_j = \begin{cases} D_+ u_j & j = 0, \\ D_0 u_j & j = 1, 2, \dots. \end{cases}$$

With the scalar product defined by

$$(u, v)_h = \frac{h}{2}u_0v_0 + \sum_{j=1}^{\infty} u_jv_jh.$$

the condition (7.3) is satisfied.

The centered difference operator and the scalar product are both modified at the boundary to achieve the desired property (7.3). For higher order operators, it is natural to generalize this construction, and allow for modifications at more than one point near the boundary. The scalar product and the norm have the form

$$(u, v)_h = \sum_{i,j=0}^{r-1} h_{ij}u_iv_jh + \sum_{j=r}^{\infty} u_jv_jh, \quad ||u||_h^2 = (u, u)_h^2, \quad (7.4)$$

where the coefficients h_{ij} are the elements of a positive definite Hermitian matrix H_0 . The number of points r that are modified in the scalar product, is yet to be determined. In the second order case above, we could choose $r = s = 1$, but in general we must expect that $r \geq s$. In the remaining part of this section we shall consider grid functions u as vectors \mathbf{u} . Difference operators D are considered as matrices, but we use the same notation D . The scalar product is denoted by

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_j u_j v_j h,$$

where the sum is taken over all elements in the vectors.

The structure of H and D is illustrated by an example with $r = 5$ and $s = 2$:

$$H = \begin{bmatrix} H_0 & 0 \\ 0 & I \end{bmatrix} = \left[\begin{array}{cc|c} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right] \quad (7.5)$$

$$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \left[\begin{array}{cc|ccccc} \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \times & \times \\ \hline & & \times & \times & \times & \times & & \\ & & & \times & \times & \times & & \\ & & & & \times & \times & \times & \times \\ & & & & & \times & \times & \times \\ & & & & & & \ddots & \ddots & \ddots \end{array} \right]$$

Here D_{11} is an $r \times r$ matrix, D_{12} has r rows, D_{21} has s nonzero rows, and D_{22} represents the centered difference operator (with zeros in the diagonal), but cut off to the left.

The following theorem shows the existence of the difference operators we are looking for:

Theorem 7.1. *For every order of accuracy $2s$ in the interior, there is a scalar product with a positive definite matrix H_0 , such that strictly semibounded difference operators D exist with order of accuracy $2s - 1$ near the boundary.* \square

These operators are not unique except for $p = 2$, and there are several parameters that can be selected. There are various criteria that can be applied for the choice of these parameters. One criterion is minimal bandwidth, another one is to minimize the coefficient for the leading error term.

For the simple scalar outflow problem the problem is now completely solved. However, already for scalar inflow problems, we must ask ourselves how the boundary condition should be implemented for this type of difference operators. This is a nontrivial problem, and we shall show in the next section that all the operators above cannot be used in a straightforward way. For systems of PDE, further restrictions are required.

One type of approach is based on a modified form of the matrix H_0 . In particular, we need that the first row and column is zero, except the first element, such that

$$H_0 = \begin{bmatrix} h_{00} & 0 & \cdots & 0 \\ 0 & h_{11} & \cdots & h_{1,r-1} \\ \vdots & \vdots & & \vdots \\ 0 & h_{r-1,1} & \cdots & h_{r-1,r-1} \end{bmatrix}. \quad (7.6)$$

We call this a *restricted form* of H_0 . Also in this case one can prove

Theorem 7.2. *For every order of accuracy $2s$ in the interior, there is a scalar product with a positive definite matrix H_0 of restricted form, such that strictly semibounded difference operators D exist with order of accuracy $2s - 1$ near the boundary.* \square

The generalization of the theory to domains with corners in several space dimensions is difficult if H_0 is not diagonal. Therefore, it is interesting to know if H_0 can be further restricted. It turns out that we then have to restrict the accuracy:

Theorem 7.3. *For order of accuracy $2s$, $1 \leq s \leq 4$ in the interior, there are strictly semibounded difference operators D with order of accuracy s near the boundary, where H_0 in the scalar product is diagonal.* \square

The global accuracy is $s + 1$, which is less than optimal except for the case $p = 2$. Still these operators are frequently used in practice. After all, the global accuracy increases with increasing $p = 2s$.

In Appendix B the elements of some of the SBP operators are given. Here we show the structure of three of the operators.

$p = 4, r = 5$, global accuracy 4, H of restricted form (7.6):

$$D = \left[\begin{array}{cccc|ccccc} \times & \times & \times & \times & & & & & \\ \times & \times & \times & \times & \times & \times & & & \\ \times & \times & \times & \times & \times & \times & & & \\ \times & & \\ \times & & \\ \hline & & & & \times & \times & \times & \times & \\ & & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times & \times \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{array} \right].$$

$p = 4, r = 4$, global accuracy 3, H diagonal:

$$D = \left[\begin{array}{cccc|ccccc} \times & \times & \times & \times & & & & & \\ \times & \times & \times & & & & & & \\ \times & \times & \times & \times & \times & & & & \\ \times & & \\ \hline & & & & \times & \times & \times & \times & \\ & & & & \times & \times & \times & \times & \times \\ & & & & \times & \times & \times & \times & \times \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{array} \right]$$

$p = 6, r = 6$, global accuracy 4, H diagonal:

$$D = \left[\begin{array}{cccccc|cccccc} \times & \times & \times & \times & \times & \times & & & & & & \\ \times & \times & \times & \times & \times & \times & & & & & & \\ \times & \times & \times & \times & \times & \times & & & & & & \\ \times & & & & & \\ \times & & & & \\ \times & & & \\ \hline & & & & & & \times & & & & & \\ & & & & & & \times & \times & \times & \times & & \\ & & & & & & \times & \times & \times & \times & \times & \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{array} \right]$$

Let us now turn to the second derivative $\partial^2/\partial x^2$, and the relation

$$(v, v_{xx}) = -||v_x||^2 - v(0)v_x(0). \quad (7.7)$$

The question is if we can construct difference operators $D^{(2)}$ that approximate $\partial^2/\partial x^2$ with high order accuracy, and at the same time satisfy the relation (7.7) in the discrete sense. There is an immediate positive answer to that question, since we can use $D^{(2)} = D^2$, where D is one of the SBP operators constructed above for approximation of $\partial/\partial x$. In order to see this, we prove the following lemma:

Lemma 7.1. *A matrix D is strictly semibounded if and only if*

$$\langle \mathbf{u}, HD\mathbf{v} \rangle = -\langle D\mathbf{u}, H\mathbf{v} \rangle - u_0 v_0 \quad (7.8)$$

for all vectors \mathbf{u}, \mathbf{v} that vanish for $x_j \geq \bar{x}$.

Proof: Strict semiboundedness follows immediately by letting $\mathbf{v} = \mathbf{u}$ in (7.8). The proof in the other direction follows by expanding the relation

$$\langle \mathbf{u} + \mathbf{v}, HD(\mathbf{u} + \mathbf{v}) \rangle = -\frac{1}{2}|u_0 + v_0|^2,$$

which gives

$$\langle \mathbf{u}, HD\mathbf{u} \rangle + \langle \mathbf{v}, HD\mathbf{v} \rangle + \langle \mathbf{u}, HD\mathbf{v} \rangle + \langle \mathbf{v}, HD\mathbf{u} \rangle = -\frac{1}{2}(|u_0|^2 + |v_0|^2 + 2u_0 v_0).$$

The equality (7.8) then follows by (7.3). \square

If D is strictly semibounded, we get by using Lemma 7.1

$$\langle \mathbf{v}, HD^2\mathbf{v} \rangle = -\langle D\mathbf{v}, HD\mathbf{v} \rangle - u_0(D\mathbf{v})_0,$$

which corresponds exactly to (7.7). However, there are drawbacks with this type of operator. In order to realize this, it is enough to take a look at the inner points. The standard second order approximation $D_2^{(2)} = D_+ D_-$ uses only 3 points, but the second order accurate $D_2^2 = D_0^2$ uses 5 points. (D_0 is the notation used through the whole book for the standard centered difference operator, and should not be confused with a zero order approximation.) For general order of accuracy p , the width is $p+1$ points for the compact operators compared to $2p+1$ points for D_p^2 . By construction, the operators D_p^2 are well defined near the boundary, but they require more work than necessary. There is still another drawback. We have seen in Section 2.2.2 that the Fourier transform of the standard approximation $D_+ D_-$ is $-4 \sin^2(\xi/2)/h^2$. The Fourier coefficients for the time dependent ODE system will be proportional to $\exp(-4 \sin^2(\xi/2)t/h^2)$, i.e., there will be strong damping for the highest wave number $\xi = \pi$, just as for the PDE problem. The Fourier transform of D_0^2 on the other hand, is $-(\sin^2 \xi)/h^2$, and there will be no damping at all for $\xi = \pi$.

We prefer to work with vectors \mathbf{u} defined as above, and matrices corresponding to the operators discussed above (using the same notation). We drop the subscript p

denoting the order of accuracy, and denote by $D^{(2)}$ a general matrix corresponding to a discretization of $\partial^2/\partial x^2$. This means that if it is applied to a discretization of a smooth function $u(x)$, the truncation error is $\mathcal{O}(h^p)$ at inner points. The matrix H is crucial for the construction, and is defined as for the first derivative above. In fact, if we want to solve a PDE problem where there are both first and second derivatives present, then H not only has to be of the same type as for the first derivative, it has to be identical. The reason is that in order to guarantee stability for the time dependent problem, we must use the same norm for both terms.

In order to obtain a discrete version of (7.7), we look for a positive definite matrix H of the form discussed above, such that

$$HD^{(2)} = -M + S,$$

where M is a positive semidefinite matrix. The matrix S has the form

$$S = \frac{1}{h^2} \begin{bmatrix} s_0 & s_1 & \cdots & s_q & 0 & \cdots \\ 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \quad (7.9)$$

where the first row represents an approximation of $h^{-1}\partial/\partial x$ at $x = 0$. Let us now consider an approximation of $au_x + bu_{xx}$, where a and b are positive constants. If $D = D^{(1)}$ is an SBP approximation of $\partial/\partial x$, then we have

$$\begin{aligned} \langle \mathbf{u}, H(aD^{(1)} + bD^{(2)})\mathbf{u} \rangle &= -a \frac{|u_0|^2}{2} + b \langle \mathbf{u}, (-M + S)\mathbf{u} \rangle \\ &= -a \frac{|u_0|^2}{2} - b \langle \mathbf{u}, M\mathbf{u} \rangle + \frac{bu_0}{h} \sum_{j=0}^q s_j u_j. \end{aligned} \quad (7.10)$$

The boundary terms on the right hand side correspond exactly to those obtained after integration by parts for the continuous case. The term $-\langle \mathbf{u}, M\mathbf{u} \rangle$ corresponds to $-||u_x||^2$. In order to demonstrate the connection between the continuous and discrete problems, we show the matrices obtained for the second order case $p = 2$:

$$H = \text{diag}(1/2 \ 1 \ 1 \ 1 \ \dots \ \dots),$$

$$S = \frac{1}{h^2} \begin{bmatrix} 3/2 & -2 & 1/2 & 0 & \cdots \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots \end{bmatrix},$$

$$M = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \dots \\ -1 & 2 & -1 & 0 & 0 & \dots \\ 0 & -1 & 2 & -1 & 0 & \dots \\ 0 & 0 & -1 & 2 & -1 & \dots \\ 0 & 0 & 0 & -1 & 2 & -1 \\ & & & & \ddots & \ddots \end{bmatrix},$$

$$D^{(2)} = \frac{1}{h^2} \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & 0 & \dots \\ & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Obviously $D^{(2)}$ is an approximation of $\partial^2/\partial x^2$ at every point, and the order of accuracy is down to one at the boundary, since it is not centered at $x = 0$. The matrix M is an approximation of $-\partial^2/\partial x^2$, but only at the inner points. For the continuous case we know that (u, u_{xx}) equals $-||u_x||^2$ except for the boundary terms. Therefore it is reasonable that the term $-b\langle \mathbf{u}, M\mathbf{u} \rangle$ in (7.10) represents $-b||u_x||^2$.

Let us next consider an example with variable coefficients:

$$u_t = (au)_x + au_x,$$

where $a = a(x) > 0$. The differential operator is semibounded, and we have

$$(u, (au)_x) + (u, au_x) = -a(0)|u(0, t)|^2 - (u_x, au) + (u, au_x) = -a(0)|u(0, t)|^2.$$

This leads to the equality

$$\|u(\cdot, t)\|^2 = \|u(\cdot, 0)\|^2 - \int_0^t a(0)|u(0, \tau)|^2 d\tau, \quad (7.11)$$

which describes the exact energy loss at the boundary. A discretization with an SBP operator D yields

$$\frac{du_j}{dt} = D(au)_j + a_j Du_j, \quad j = 1, 2, \dots.$$

Let the matrix A be defined by $A = \text{diag}(a_0 a_1 \dots)$. When taking the discrete scalar product with u , and using vector notation and Lemma 7.1, we get

$$\begin{aligned} \langle \mathbf{u}, HDA\mathbf{u} \rangle + \langle \mathbf{u}, HAD\mathbf{u} \rangle &= -\langle D\mathbf{u}, HA\mathbf{u} \rangle + \langle \mathbf{u}, HAD\mathbf{u} \rangle - a_0|u_0(t)|^2 \\ &= -\langle AHD\mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, HAD\mathbf{u} \rangle - a_0|u_0(t)|^2, \end{aligned} \quad (7.12)$$

where in the last equality we have used the fact that A and H are symmetric. We would like the first two terms of the right hand side to cancel, but they do that only if $HA = AH$. The matrix A is diagonal, but the commuting property holds only if H

is diagonal as well. So, here we find another advantage with the diagonal norms. To be sure that we have exactly the same properties as the differential equation, we should avoid the nondiagonal norms for variable coefficients.

Fortunately, the situation is not as bad as it might look at a first glance. We recall that the matrices H are almost diagonal, i.e., the deviation occurs only at a fixed and low number of points near the boundaries. The matrix $HA - AH$ is symmetric and zero everywhere except in the upper $r \times r$ block. Therefore, if $a(x)$ is a Lipschitz continuous function such that $|u(x + \delta) - u(x)| \leq K\delta$, we have

$$-\langle AHD\mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, HAD\mathbf{u} \rangle = \langle \mathbf{u}, (HA - AH)D\mathbf{u} \rangle = h\phi(\mathbf{u}_L). \quad (7.13)$$

Here $h\phi(\mathbf{u}_L) = \langle \mathbf{u}_L, B\mathbf{u}_L \rangle$, where B is a bounded $r \times r$ matrix, and

$$\mathbf{u}_L = (u_0 \ u_1 \ \dots \ u_{r-1})^T.$$

(Recall that the scalar product $\langle \cdot, \cdot \rangle$ contains a factor h multiplying each term.) Therefore, instead of an exact discrete version of (7.11), we get when using the notation $\|u\|_h^2 = \langle \mathbf{u}, H\mathbf{u} \rangle$

$$\begin{aligned} \|u(t)\|_h^2 &= \|u(0)\|_h^2 - \int_0^t a_0 |u_0(\tau)|^2 d\tau + h \int_0^t \phi(\mathbf{u}_L(\tau)) d\tau, \\ |\phi(\mathbf{u}_L(\tau))| &\leq K \sum_{j=0}^{r-1} |u_j(\tau)|^2, \end{aligned} \quad (7.14)$$

where K is independent of h . The extra term in (7.14) does not influence the order of accuracy, which is determined solely by the accuracy of the difference operator D as given by the theorems 7.1, 7.2 and 7.3.

We make still another observation concerning (7.14). There is a slight smoothness restriction on $a(x)$, but not on the grid function u_j for this estimate to hold. Assume that also u_j satisfies the analogue of the Lipschitz condition, i.e., $|u_{j+1} - u_j| \leq Kh$. Since D represents an approximation of $\partial/\partial x$, the first r elements of $D\mathbf{u}$ are of order $|u_j|$, and it follows that $|\langle \mathbf{u}, (HA - AH)D\mathbf{u} \rangle|$ is of order h^2 , i.e., the last term in (7.14) is also $\mathcal{O}(h^2)$. In other words, as long as the solution u has some smoothness, the strict semiboundedness is violated only by an extra $\mathcal{O}(h^2)$ -term.

In Table 7.1 these results are clearly illustrated for the 4th order accurate operator $D = D_4^{(1)}$ as defined in Section B.2 of Appendix B, and applied in the interval $-1 \leq x \leq 1$.

Table 7.1 $h|\phi(\mathbf{u}_L)|$ as defined in (7.13)

N	u nonsmooth	u smooth
10	0.0173	0.00224
20	0.0086	0.00055
40	0.0043	0.00014

The table shows $h|\phi(\mathbf{u}_L)|$ for the Lipschitz continuous function $a(x) = \max(1 - 2x, 0)$ and three different values of N . The second column is for the nonsmooth function $u_j = (-1)^j$, and the third column for the discretized Lipschitz continuous grid function $u(x) = 1 + x$.

Finally, we shall discuss a nonlinear equation. Consider the Burgers' equation

$$u_t + uu_x = 0,$$

which can also be written in the split form

$$u_t + \frac{1}{3}((u^2)_x + uu_x) = 0. \quad (7.15)$$

This equation easily gives rise to discontinuous solutions called shocks, even if the initial function is smooth. Therefore one has to be careful when defining solutions to this equation. However, assuming that u is such that integration by parts can be carried out, and that $u(x, t) = 0$ for $x \geq \bar{x}$, we get

$$\frac{d}{dt}||u||^2 = \frac{1}{3}u(0, t)^3.$$

If D is an SBP operator, we use the semidiscrete approximation

$$\frac{du_j}{dt} = \frac{1}{3}(D(u^2)_j + u_j Du_j), \quad j = 1, 2, \dots$$

Summation by parts is always well defined, and if H is diagonal, we get

$$\frac{d}{dt}||u||_h^2 = \frac{1}{3}u_0(t)^3.$$

Here it is a more severe restriction than for the linear equation to assume that u is Lipschitz continuous in space, but if it is, we get for general H just as in the linear case

$$\begin{aligned} \frac{d}{dt}||u||_h^2 &= \frac{1}{3}u_0(t)^3 + h\phi(\mathbf{u}_L). \\ |\phi(\mathbf{u}_L)| &\leq K \sum_{j=0}^{r-1} |u_j|^3, \end{aligned}$$

where K is independent of h . If H is diagonal, then the constant K is zero. In practice one can say that, as long as no shocks are approaching the boundary, the splitting of the equation together with the SPB operators guarantee that the proper norm conservation holds almost exactly, regardless of the particular structure of the matrix H .

It should be said that the type of centered approximations at inner points, which is the foundation for the SBP operators, don't work well for solutions with discontinuities. Artificial viscosity has to be added in order to avoid the oscillations that

otherwise occur around the discontinuity. This problem will be discussed further in Chapter 11.

The SBP operators we have discussed until now are based on the standard centered approximations at inner points as described in Section 4.1. It is also possible to construct SBP operators based on the Padé type approximations described in Section 4.3. These operators have the form $P^{-1}Q$, where P is a positive definite nondiagonal operator. Again, thinking of the grid functions as vectors and of operators as matrices, we write the semidiscrete approximation of $u_t = u_x$ as

$$\frac{d\mathbf{u}}{dt} = P^{-1}Q\mathbf{u},$$

or equivalently

$$P \frac{d\mathbf{u}}{dt} = Q\mathbf{u}.$$

We are looking for a positive definite matrix H such that

$$\frac{d}{dt} \langle \mathbf{u}, HP\mathbf{u} \rangle = 2 \langle \mathbf{u}, HP \frac{d\mathbf{u}}{dt} \rangle = 2 \langle \mathbf{u}, HQ\mathbf{u} \rangle = -|u_0|^2.$$

As an example, consider the 4th order approximation given in Table 4.5. At inner points, the matrices P and Q are both tridiagonal, and we allow for a tridiagonal H as well. It turns out that 3rd order accuracy near the boundary is obtained with $r = 4$. The matrices corresponding to this operator, as well as a number of other SBP operators, are given in appendix B.

7.3 SBP Operators and Projection Methods

The application of the SPB operators to the hyperbolic outflow problem has already been discussed. We get exactly the same loss of energy at the boundary for the numerical solution as for the PDE problem. Let us now turn to the inflow problem

$$\begin{aligned} u_t + u_x &= 0, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(0, t) &= 0, \\ u(x, 0) &= f(x). \end{aligned}$$

The SBP operator can be applied here as well. Assume for example that the semi-discrete system is approximated by an explicit one step method in time. Then it is natural to advance the scheme one step using the SBP operator, and then enforce the true boundary condition $u_0 = 0$. This procedure is called the *injection method*. At a first glance, it may look like stability will follow trivially, since the boundary term $|u_0|^2/2$ vanishes. But it does not. In particular, time stability does not follow, and this is of course a fundamental property for this equation. In fact, in [Strand, 1994] an example of a strictly semibounded sixth order operator D_6 was presented which

leads to a semidiscrete inflow problem with growing solutions. Since it is important to understand the mechanism, we shall take a closer look at this case.

Here we work with the vectors

$$\mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{u}} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

In $\tilde{\mathbf{u}}$ the first element u_0 has been eliminated, and we use the notation \tilde{D} for matrices where the first row and column have been eliminated.

With the vector notation, strict semiboundedness can be expressed as

$$\langle \mathbf{u}, H D \mathbf{u} \rangle = -\frac{1}{2} u_0^2, \quad (7.16)$$

where D approximates $\partial/\partial x$, and

$$H = \begin{bmatrix} H_0 & 0 \\ 0 & I \end{bmatrix}.$$

For $u_0 = 0$ we can write (7.16) as

$$\langle \tilde{\mathbf{u}}, \widetilde{H D \mathbf{u}} \rangle = 0. \quad (7.17)$$

The injection method can be written as

$$\frac{d\tilde{\mathbf{u}}}{dt} = -\tilde{D}\tilde{\mathbf{u}}$$

for the inner points, and we have for $u_0 = 0$

$$\frac{d}{dt} \langle \mathbf{u}, H \mathbf{u} \rangle = 2 \langle \mathbf{u}, H \frac{d\mathbf{u}}{dt} \rangle = 2 \langle \tilde{\mathbf{u}}, \tilde{H} \frac{d\tilde{\mathbf{u}}}{dt} \rangle = -2 \langle \mathbf{u}, \tilde{H} \tilde{D} \tilde{\mathbf{u}} \rangle.$$

Here is where the derivation stalls. We would like the right hand side to vanish, but unfortunately it doesn't follow from (7.17). For this we need the condition $\widetilde{H D} = \widetilde{H D}$, which is not always satisfied.

In the summary section of Chapter 2 we quoted a theorem by Goldberg and Tadmor, which says that one can always prescribe the values at all missing points and still have a stable scheme. But this is for the case where the difference operator at inner points is applied as far to the left as possible. In the counter example referred to above, this is not the case.

Another type of implementation is the *projection method*. We define the projection matrix

$$P = \text{diag}(0 \ 1 \ 1 \ 1 \ \dots \ \dots), \quad (7.18)$$

and consider the semidiscrete system

$$\begin{aligned} \frac{d\mathbf{u}}{dt} + PD\mathbf{u} &= 0, \\ \mathbf{u}(0) &= P\mathbf{f}. \end{aligned} \tag{7.19}$$

Here the boundary component u_0 is included in the system. The initial condition restricts the data such that $u_0(0) = 0$, and the first differential equation therefore guarantees that $u_0(t) = 0$ for all t . This shows that the solution of this system is identical to the one obtained with the injection method. Therefore, this projection method is in general no good either, and we have to come up with a better projection P . Olsson did that in [Olsson, 1995a]. The physical boundary condition is formulated for the vector \mathbf{u} as

$$L\mathbf{u} = 0,$$

where

$$L = [1 \ 0 \ 0 \ 0 \ \dots \ \dots].$$

The projection is then defined by

$$P = I - H^{-1}L^T(LH^{-1}L^T)^{-1}L. \tag{7.20}$$

With this projection P , one can show that (7.19) satisfies an energy estimate and is time stable. But note that the initial data must satisfy the boundary condition.

For the sixth order operator $D = D_6$ mentioned above, the new projection matrix (7.20) is

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots \\ 0.0676 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & & \\ 0.0806 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & & \\ -0.0048 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & & \\ -0.1137 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & & \\ 0.1517 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & & \\ \vdots & & & & & & & & \ddots & \\ \vdots & & & & & & & & & \ddots \end{bmatrix},$$

which should be compared to the simple diagonal projection matrix (7.18) corresponding to direct injection. This type of projection method can be generalized to systems of PDE

$$\begin{aligned} u_t &= Au_x, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ Bu(0,t) &= 0, \\ u(x,0) &= f(x), \end{aligned}$$

where A is a real and symmetric matrix. The vector representing the numerical solution is now

$$\mathbf{u} = \begin{bmatrix} u_0^{(1)} & u_0^{(2)} & \cdots & u_0^{(m)} & u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(m)} & \cdots \cdots \end{bmatrix}.$$

The matrix Q approximating $A\partial/\partial x$ corresponding to the strictly semidiscrete difference operator is obtained by substituting each element d_{ij} in D by the matrix by $d_{ij}A$, i.e., each \times in the matrix (7.5) represents an $m \times m$ submatrix. The matrix H is generalized in the same way. The physical boundary condition

$$Bu = 0$$

is expressed in terms of \mathbf{u} as

$$L\mathbf{u} = 0,$$

and the projection matrix has the same form as in (7.20). The projection method is then defined by

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= P Q \mathbf{u}, \\ \mathbf{u}(0) &= Pf. \end{aligned} \tag{7.21}$$

If the PDE system satisfies an energy estimate, then the approximation can be shown to do that as well, i.e.,

$$\frac{d}{dt} \langle \mathbf{u}, H\mathbf{u} \rangle = 0.$$

There is of course an advantage by using a simpler form of projection, that does not depend on the matrix H_0 in the norm. If the matrix A has q negative eigenvalues, we assume for convenience that the physical boundary condition has the form

$$u^I(0, t) = 0,$$

where u^I contain the first q components of u . Furthermore, we assume that the differential operator $A\partial/\partial x$ is semibounded, i.e.,

$$\mathbf{v}^T A \mathbf{v} \geq 0$$

for all vectors \mathbf{v} with $\mathbf{v}^I = \mathbf{0}$. The generalization of the simple projection for the inflow scalar equation above, is obtained by enforcing the boundary condition at the point $x_0 = 0$, but otherwise leaving the vector unaltered:

$$\begin{aligned} P\mathbf{u} &= (I - L)\mathbf{u} \\ &= \begin{bmatrix} 0 & \cdots & 0 & u_0^{(q+1)} & u_0^{(q+2)} & \cdots & u_0^{(m)} & u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(m)} & \cdots \cdots \end{bmatrix}. \end{aligned} \tag{7.22}$$

Here is where the restricted form of the matrix H_0 comes in. Under the assumption that $A\partial/\partial x$ is semibounded, one can prove

Theorem 7.4. *The approximation (7.21) satisfies an energy estimate if P is defined by (7.22) and H_0 in the norm has the restricted form (7.6).* \square

The crucial point in the proof, which can be found in [Gustafsson et al., 1995] (Sec 11.4), is that P is symmetric if H_0 has the restricted form, i.e., $\langle \mathbf{u}, HP\mathbf{v} \rangle = \langle P\mathbf{u}, H\mathbf{v} \rangle$.

In the previous section we mentioned the diagonal form of H_0 , which is more convenient for application to multidimensional problems. This diagonal form is a special case of the restricted form (7.6), and theorem 7.4 applies. We also note that the H -dependent projection (7.20) reduces to the simple form (7.22) for diagonal H_0 , which we already know leads to stability. Recall, however, that the order of accuracy deteriorates with these difference operators for $p \geq 4$.

So far, we have assumed homogeneous boundary conditions. Even if that is natural when discussing semibounded operators and time stable methods, we must know how to implement nonzero data for a practical computation.

Assume that the boundary condition for the PDE system is

$$u^I(x, t) = g(t).$$

For injection methods the implementation is trivial, we simply prescribe $\mathbf{u}_0^I = g(t)$. For projection methods it is different. The first q differential equations (for \mathbf{u}_0^I in (7.21)) must be substituted by

$$\frac{d\mathbf{u}_0^I}{dt} = g'(t),$$

i.e., the boundary condition is substituted by its differentiated form. We have assumed that the initial data satisfies the boundary condition, i.e., for the inhomogeneous case we require $\mathbf{f}_0^I = g(0)$. Hence, the solution will satisfy the true condition $\mathbf{u}_0^I = g(t)$ for all time. However, when discretizing in time, we must expect that the ODE solver doesn't produce the correct boundary values $g(t_n)$. The error will of course be on the truncation error level, and it will tend to zero as $h \rightarrow 0$. Still it is sometimes important to satisfy the boundary conditions exactly, and in such a case, the method should be modified.

Actually, there is still another reason for modifying the projection method. Recall that the initial data are restricted, such that they satisfy the exact boundary condition. When looking at the complete operator PQ by itself, and using the normal mode analysis according to Section 2.4.1, it turns out that there is a generalized eigenvalue $\tilde{s} = 0$ of the kind that leads to an instability. The reason is that the operator by itself doesn't notice any boundary condition of the kind that is required for problems with ingoing characteristics, since it is not enforced. As we have just described, this condition is taken care of by the restriction of the initial data, and that restriction is not included in the normal mode analysis based on the Laplace transform. Note that this result of the normal mode analysis is not an artifact. If the condition $\mathbf{f}_0^I = g(0)$ is perturbed, then we must expect an instability.

As a result of the arguments above, it seems like we are in trouble here, since we have already shown that pure injection leads to growing solutions for some of the SBP operators. However, we shall construct a modified projection method, which eliminates the problem. We illustrate the method by a (2×2) PDE system

$$\begin{aligned} u_t &= \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} u_x, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u^{(1)}(0, t) &= u^{(2)}(0, t), \\ u(x, 0) &= f(x), \end{aligned}$$

and the standard second order SBP operator. With the vector/matrix notation used above we have

$$\mathbf{u} = [u_0^{(1)} \ u_0^{(2)} \ u_1^{(1)} \ u_1^{(2)} \ \dots \ \dots]^T,$$

and

$$Q = \frac{1}{h} \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & -1/2 & 0 \\ 0 & -1/2 & 0 & 0 & 0 & 1/2 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

The matrix H in the norm is

$$H = \text{diag} \left[\frac{1}{2} \ \frac{1}{2} \ 1 \ 1 \ \dots \ \dots \right].$$

The boundary condition $Bu = 0$ has not the standard form $u^I = 0$ treated earlier, since there is a coupling between the elements in u . For the numerical solution we have

$$L\mathbf{u} = 0, \quad L = [1 \ -1 \ 0 \ 0 \ 0 \ 0 \ \dots \ \dots],$$

and P defined by (7.20) is

$$P = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \\ & & \ddots & \\ & & & \ddots \end{bmatrix}.$$

If the initial data satisfy $f_0^{(1)} = f_0^{(2)}$, then the approximation (7.21) is stable, and PQ has the form

$$PQ = \frac{1}{2h} \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

For the reasons above, we now introduce a modification. Without discretization in time, the solution satisfies the true boundary condition, and we want to enforce that one without the restriction on the initial data. Therefore we eliminate $u_0^{(1)}$, and work with the vector

$$\tilde{\mathbf{u}} = [u_0^{(2)} \ u_1^{(1)} \ u_1^{(2)} \ \dots \ \dots]^T.$$

The boundary condition $u_0^{(1)} = u_0^{(2)}$ is used to eliminate the first row and column of PQ . By adding the first column to the second one, we obtain the approximation

$$\begin{aligned} \frac{d\tilde{\mathbf{u}}}{dt} &= \widetilde{PQ}\tilde{\mathbf{u}}, \\ \tilde{\mathbf{u}}(0) &= \tilde{\mathbf{f}}, \end{aligned} \tag{7.23}$$

where

$$\widetilde{PQ} = \frac{1}{2h} \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Note that, in contrast to the projection method, the boundary condition is now enforced. Also, in contrast to the injection method, the formulation (7.23) is obtained by applying the projection before the boundary condition is enforced. There are cases where the injection and modified projection methods are equivalent, but in general they are not.

7.4 SBP Operators and Simultaneous Approximation Term (SAT) Methods

In this section we shall introduce another way of implementing the SBP operators, that avoids the particular difficulties that we have demonstrated in the previous section. In analogy with the projection method, the main idea is to construct the difference operator in the main approximation such that it takes the boundary condition into account, but now by adding a penalty term.

We begin also here by discussing the simple scalar equation, but now with two boundaries included:

$$\begin{aligned} u_t &= u_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(1,t) &= g(t), \\ u(x,0) &= f(x). \end{aligned}$$

The grid is defined by $x_j = jh$, $j = 0, 1, \dots, N$, $Nh = 1$, and we introduce the vector

$$\mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_N \end{bmatrix}.$$

The matrix H in the norm and the difference operator matrix D are both $(N+1) \times (N+1)$ matrices, and they have the form

$$H = \begin{bmatrix} H_0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & H_N \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} & 0 \\ D_{21} & D_{22} & D_{N-1,N} \\ 0 & D_{N,N-1} & D_{NN} \end{bmatrix}.$$

Here the submatrices in the lower right corner are defined by symmetry rules from the ones already defined in the upper left corner. The elements are defined by the coordinate transformation $\xi_j = 1 - x_{N-j}$. For example, the elements \tilde{h}_{ij} of H_N and \tilde{d}_{ij} of D_{NN} are defined by

$$\begin{aligned} \tilde{h}_{ij} &= h_{r-1-i, r-1-j}, \quad 0 \leq i, j \leq r-1, \\ \tilde{d}_{ij} &= -d_{r-1-i, r-1-j}, \quad 0 \leq i, j \leq r-1, \end{aligned}$$

where h_{ij} and d_{ij} are the elements of H_0 and D_{11} respectively. The strict semiboundedness implies

$$\langle \mathbf{u}, HD\mathbf{u} \rangle = \frac{1}{2}(|u_N|^2 - |u_0|^2).$$

Next we define the new vector

$$\mathbf{w} = H^{-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

i.e., it is the last column of the matrix H^{-1} . The SAT method is defined by

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= D\mathbf{u} - \tau(u_N - g(t))\mathbf{w}, \\ \mathbf{u}(0) &= \mathbf{f}. \end{aligned} \tag{7.24}$$

If H is diagonal, only the last equation is affected by the extra penalty term. However, for more general matrices, more equations are affected.

The stability condition

$$\tau \geq \frac{1}{2}$$

is easily obtained from

$$\frac{d}{dt} \langle \mathbf{u}, H\mathbf{u} \rangle = 2 \langle \mathbf{u}, H(D\mathbf{u} - \tau u_N \mathbf{w}) \rangle = -u_0^2 + u_N^2 - 2\tau u_N^2$$

when assuming $g(t) = 0$.

As a numerical test, we solve the nonlinear Burgers' equation with two boundaries:

$$\begin{aligned} u_t + uu_x &= 0, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= g(t), \\ u(x, 0) &= f(x). \end{aligned}$$

We limit ourselves to the case with smooth solutions, and solve the problem for $0 \leq t \leq 0.5$ and

$$\begin{aligned} g(t) &= 1 - 0.1 \sin(4\pi t), \\ f(x) &= 1. \end{aligned}$$

We use the split form (7.15) of the equation, and approximate $\partial/\partial x$ by the three different SBP operators $D_2^{(1)}$, $D_{4,diag}^{(1)}$ with diagonal H_0 , and $D_{4,dense}^{(1)}$ with dense H_0 as given in Appendix B. The trapezoidal rule is used for time discretization, with a small time step corresponding to $\lambda = 0.01$. In that way the discretization error in time is negligible. Figure 7.1 shows the result for a very coarse grid, N=25. (The “exact” solution is obtained numerically on a very fine grid.) The two 4th order methods give quite good results, while the second order one is way off in the central part of the interval.

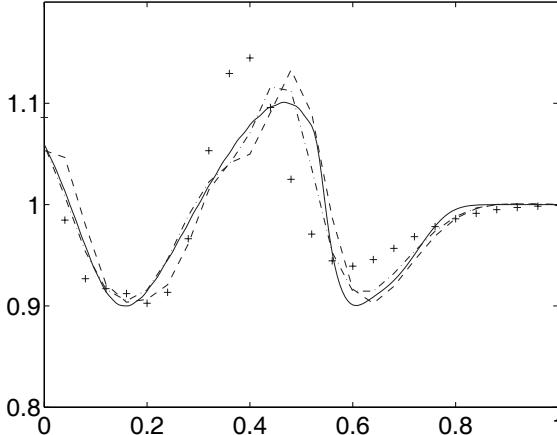


Fig. 7.1 $u(x, 0.5)$, $N = 20$, Exact (—), $D_2^{(1)}$ (++) , $D_{4,diag}^{(1)}$ (−·), $D_{4,dense}^{(1)}$ (---)

Next we consider hyperbolic systems of PDE. For convenience, we assume that the coefficient matrix is diagonal, such that the coupling between the variables is stated in the boundary conditions:

$$\begin{aligned} u_t &= \Lambda u_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u^I(1, t) &= R u^{II}(1, t) + g^I(t), \\ u^{II}(0, t) &= L u^I(0, t) + g^{II}(t), \\ u(x, 0) &= f(x). \end{aligned}$$

Here u^I and g^I are vectors with q components and u^{II} and g^{II} have $m - q$ components corresponding to the matrix $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ with

$$\begin{aligned} \Lambda^I &= \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_q) > 0, \\ \Lambda^{II} &= \text{diag}(\lambda_{q+1}, \lambda_{q+2}, \dots, \lambda_m) < 0. \end{aligned}$$

The matrices L and R represent the reflection at the boundaries. Let the norm be defined by

$$E(t) = |L|(u^I, (\Lambda^I)^{-1}u^I) + |R|(u^{II}, -(\Lambda^{II})^{-1}u^{II}). \quad (7.25)$$

For homogeneous boundary conditions, one can show that $E(t) \leq E(0)$ if $|L| |R| \leq 1$, and we assume that this condition is satisfied. It is most convenient to state the approximation in terms of each component

$$\mathbf{u}^{(\nu)} = \begin{bmatrix} u_0^{(\nu)} \\ u_1^{(\nu)} \\ \vdots \\ u_N^{(\nu)} \end{bmatrix}, \quad \nu = 1, 2, \dots, m.$$

Define the vectors

$$\mathbf{w}^{(\nu)} = H^{-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \nu = 1, 2, \dots, q, \quad \mathbf{w}^{(\nu)} = H^{-1} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \quad \nu = q+1, q+2, \dots, m,$$

where H is the matrix corresponding to the scalar case. If D is the corresponding SBP matrix approximating $\partial/\partial x$, the SAT approximation is

$$\begin{aligned} \frac{d\mathbf{u}^{(\nu)}}{dt} &= \lambda_\nu D\mathbf{u}^{(\nu)} - \tau\lambda_\nu(u_N^{(\nu)} - (Ru_N^H)^{(\nu)} - g^{(\nu)}(t))\mathbf{w}, \quad \nu = 1, 2, \dots, q, \\ \frac{d\mathbf{u}^{(\nu)}}{dt} &= \lambda_\nu D\mathbf{u}^{(\nu)} - \tau\lambda_\nu(u_N^{(\nu)} - (Lu_0^I)^{(\nu)} - g^{(\nu)}(t))\mathbf{w}, \quad \nu = q+1, q+2, \dots, m, \\ \mathbf{u}^{(\nu)}(0) &= \mathbf{f}^{(\nu)}, \quad \nu = 1, 2, \dots, m. \end{aligned}$$

(The notation sometimes becomes a little complicated. Recall that u_N^H is the approximation of the vector $u^H(x_N, t)$ with $m-q$ components, Ru_N^H is a vector with q components, and $(Ru_N^H)^{(\nu)}$ is component no. ν of this vector. Similarly for $(Lu_0^I)^{(\nu)}$.) In [Carpenter et al., 1994] it is shown that this approximation is time stable if

$$1 - \sqrt{1 - |L||R|} \leq \tau|L||R| \leq 1 + \sqrt{1 - |L||R|}.$$

The norm is the discrete analogue of (7.25), i.e., (7.25) with $(u, v) \rightarrow \langle \mathbf{u}, H\mathbf{v} \rangle$.

The SBP operators were constructed under the assumption that the accuracy near the boundary is $p-1$. That criterion is based on the results that were discussed in Section 3.2. If the semidiscrete approximation of the time dependent problem is strongly stable, then it follows easily that the global error is of order h^p . Furthermore, if centered difference operators are used at the inner points, strong stability follows if the Kreiss condition is satisfied. However, there is a catch here, regardless of the particular type of implementation. Consider for example the simple scalar outflow problem. Then the centered approximations of order p require $p/2$ numerical boundary conditions, and in order to apply the normal mode analysis, there can be no more. We can also consider the boundary conditions as a modification of the difference operator at $p/2$ points. The SBP operators on the other hand, are modified at more than $p/2$ points, except in the case $p=2$. Therefore, the normal mode analysis cannot be applied in a direct way. Referring back to Section 2.3.2, we recall that the problem is Laplace transformed in time, and then the resulting difference equation

is solved in space. Before doing that, we now have to eliminate a few variables \hat{u}_j near the boundaries, such that there are $p/2$ remaining boundary conditions. Then we can derive the matrix $C(\tilde{s})$ in (2.53) and investigate the Kreiss condition. This was done in [Gustafsson, 1998], where the sixth order approximation was shown to be strongly stable for a (2×2) PDE system.

Another way of proving full global accuracy is to use a direct application of the energy method. This was done for the second order SBP operator applied to a scalar PDE in Section 3.2, but for higher order SBP operators it is more difficult. It was shown in that section how to derive an error estimate of order $p - 1/2$ by a direct application of the energy method. We illustrate here how the same technique can be used for the SAT method. The error $v_j(t) = u(x_j, t) - u_j(t)$ satisfies

$$\begin{aligned} \frac{d\mathbf{v}}{dt} &= D\mathbf{v} - \tau v_N \mathbf{w} + \mathbf{F}, \\ \mathbf{u}(0) &= \mathbf{f}. \end{aligned}$$

By assumption the elements of the vector \mathbf{F} satisfy

$$F_j = \begin{cases} \mathcal{O}(h^{p-1}), & j = 0, 1, \dots, r-1 \\ \mathcal{O}(h^p), & j = r, r+1, \dots, N-r \\ \mathcal{O}(h^{p-1}), & j = N-r+1, N-r+2, \dots, N. \end{cases}$$

Note that the term $u_N - g(t)$ in (7.24) doesn't contribute anything extra to the truncation error, since $u(x_N, t) - g(t) = 0$. We get

$$\frac{d}{dt} \langle \mathbf{v}, H\mathbf{v} \rangle \leq \sum_{i,j=0}^{r-1} (h_{ij} v_i F_j + \tilde{h}_{ij} v_{N-r+1+i} F_{N-r+1+j}) h + 2 \sum_{j=r}^{N-r} v_j F_j h.$$

By using the inequalities

$$v_i F_j h \leq \delta |v_i|^2 h + \frac{1}{4\delta} |F_j|^2 h,$$

we obtain

$$\begin{aligned} \frac{d}{dt} \langle \mathbf{v}, H\mathbf{v} \rangle &\leq K \sum_{j=0}^{r-1} (|v_j|^2 + |F_j|^2 + |v_{N-r+1+j}|^2 + |F_{N-r+1+j}|^2) h + \sum_{j=r}^{N-r} (|v_j|^2 + |F_j|^2) h \\ &\leq K \langle \mathbf{v}, H\mathbf{v} \rangle + \mathcal{O}(h^{2p-1}) + \mathcal{O}(h^{2p}), \end{aligned}$$

which leads to

$$\langle \mathbf{v}, H\mathbf{v} \rangle^{1/2} = \|v\|_h \leq K_1(t) h^{p-1/2}.$$

The order of the F -terms near the boundary is raised half a step since there are only the finite number $2r$ such terms.

In [Gustafsson, 1981] a more powerful method is used for obtaining full order of accuracy p , and it is based on the energy method. Also in this case, it is necessary to eliminate a few boundary variables u_j in order to get standard form of the

approximation. We have already indicated above how to get the same result by using normal mode analysis. Both of these methods have been used to obtain optimal error estimates for particular examples, but there is no general result for the whole class of SBP operators applied to hyperbolic first order PDE. However, all numerical experiments that we have studied, indicate full order of accuracy.

Let us next consider the scalar parabolic model problem

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u_x(0, t) - \alpha u(0, t) &= g(t), \\ u(x, 0) &= f(x), \end{aligned}$$

where $\alpha \geq 0$. For $g(t) = 0$ there is an energy estimate

$$\frac{1}{2} \frac{d}{dt} \|u\|^2 = -\|u_x\|^2 - \alpha |u(0, t)|^2.$$

The SBP-SAT approximation is

$$\begin{aligned} \frac{d\mathbf{u}}{dt} &= D^{(2)}\mathbf{u} + \tau H^{-1}((Su)_0 - \alpha u_0 - g(t))\mathbf{w}, \\ \mathbf{u}(0) &= \mathbf{f}, \end{aligned}$$

where $D^{(2)} = H^{-1}(-M + S)$ is the type of operator discussed in Section 7.2, and $\mathbf{w} = [1 \ 0 \ 0 \ \dots]^T$. Recall that S defined in (7.9) represents an approximation of $-h^{-1}\partial/\partial x$ in its first row. For $g(t) = 0$ we get

$$\frac{1}{2} \frac{d}{dt} \langle \mathbf{u}, H\mathbf{u} \rangle = -\langle \mathbf{u}, M\mathbf{u} \rangle - u_0(S\mathbf{u})_0 + \tau u_0((Su)_0 - \alpha u_0),$$

i.e., there is an energy estimate

$$\frac{1}{2} \frac{d}{dt} \langle \mathbf{u}, H\mathbf{u} \rangle = -\langle \mathbf{u}, M\mathbf{u} \rangle - \alpha |u_0|^2$$

also here if $\tau = 1$.

Let us next take a look at the accuracy. At inner points the accuracy is p . There are two other types of truncation errors. The first one is the local approximations of $\partial^2/\partial x^2$ near the boundary, and we assume that it is of order p_1 . The other type is the accuracy of the penalty term in the first differential equation, and we assume that the true solution satisfies

$$(Su(x_j, t))_{x_j=0} - \alpha u(0, t) - g(t) = \mathcal{O}(h^{p_2}).$$

The error vector with elements $v_j(t) = u(x_j, t) - u_j(t)$ satisfies

$$\frac{d\mathbf{v}}{dt} = D^{(2)}\mathbf{v} + H^{-1}((S\mathbf{v})_0 - \alpha v_0)\mathbf{w} + \mathbf{F} + h^{p_2}\tilde{\mathbf{g}}.$$

By assumption the elements of the vector \mathbf{F} satisfy

$$F_j = \begin{cases} \mathcal{O}(h^{p_1}), & j = 0, 1, \dots, r-1 \\ \mathcal{O}(h^p), & j = r, r+1, \dots, \end{cases}$$

and $\tilde{\mathbf{g}}$ is a bounded vector with the only nonzero elements in the first r positions. After taking the weighted scalar product with \mathbf{v} , we get

$$\frac{d}{dt} \langle \mathbf{v}, H\mathbf{v} \rangle = -2 \langle \mathbf{v}, M\mathbf{v} \rangle - 2\alpha |v_0|^2 + T_1 + T_2 + T_3,$$

where

$$\begin{aligned} T_1 &= 2 \sum_{i,j=0}^{r-1} h_{ij} v_i F_j h, \\ T_2 &= 2 \sum_{j=r}^{\infty} v_j F_j h, \\ T_3 &= 2 \langle \mathbf{v}, H\tilde{\mathbf{g}} \rangle. \end{aligned}$$

For the terms in T_1 we first use the inequality

$$2v_i F_j h \leq \delta |v_i|^2 + \frac{1}{\delta} |h F_j|^2, \quad 0 \leq i, j \leq r-1,$$

and then the discrete Sobolev inequality (see [Gustafsson et al., 1995], Chapter 11)

$$|v_i|^2 \leq \varepsilon \|D_+ v\|_h^2 + \frac{1}{\varepsilon} \|v\|_h^2 \leq K_1 (\varepsilon \langle \mathbf{v}, M\mathbf{v} \rangle + \frac{1}{\varepsilon} \langle \mathbf{v}, H\mathbf{v} \rangle).$$

Here the parameters δ and ε are positive, but otherwise arbitrary. For T_2 we have

$$T_2 \leq K_2 \langle \mathbf{v}, H\mathbf{v} \rangle + \mathcal{O}(h^{2p}).$$

For T_3 we get in the same way as for T_1

$$|v_0|^2 \leq K_3 \left(\varepsilon \langle \mathbf{v}, M\mathbf{v} \rangle + \frac{1}{\varepsilon} \langle \mathbf{v}, H\mathbf{v} \rangle \right).$$

The parameters δ and ε are now chosen such that

$$T_1 + T_3 \leq 2 \langle \mathbf{v}, M\mathbf{v} \rangle + \mathcal{O}(h^{2(p_1+1)}) + \mathcal{O}(h^{2(p_2+1)}).$$

We now have a differential inequality

$$\frac{d}{dt} \langle \mathbf{v}, H\mathbf{v} \rangle \leq K \langle \mathbf{v}, H\mathbf{v} \rangle + \mathcal{O}(h^{2p_0}),$$

which leads to the final error estimate

$$\|v\|_h \leq \mathcal{O}(h^{p_0}),$$

where $p_0 = \min(p, p_1 + 1, p_2 + 1)$. The global order of accuracy is raised one step not only for the approximation of $\partial^2/\partial x^2$ near the boundary, but also for the approximation of the physical boundary condition. This is somewhat surprising, but is an effect of the inherent damping properties for parabolic equations together with the SAT formulation.

So far, we have been discussing the SAT method applied to physical or computational boundaries. Another important application is internal boundaries caused by changes in the grid structure. The geometry and/or the behavior of the solution may be such that a refinement of the grid is necessary in certain subdomains. As a model example, we consider the equation

$$u_t = au_x, \quad -\infty < x < \infty, \quad 0 \leq t,$$

where different step sizes h_L and h_R are used for $x \leq 0$ and $x \geq 0$ respectively. We define the vectors

$$\mathbf{v} = \begin{bmatrix} \vdots \\ \vdots \\ v_{-2} \\ v_{-1} \\ v_0 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ \vdots \end{bmatrix},$$

where the components are defined by $\{v_{-j} = -jh_L, w_j = jh_R\}$, $j = 0, 1, 2, \dots$. The solution $u(0, t)$ is represented by two approximations v_0 and w_0 , and both of them are carried through the computation. The vectors \mathbf{v} and \mathbf{w} have infinite length, but for convenience we assume that they are made finite by assuming zero values for $|x_j| \geq \bar{x}$.

Let $D_L = H_L^{-1}Q_L$ and $D_R = H_R^{-1}Q_R$ be two SBP operators. The SAT method is defined by

$$\begin{aligned} \frac{d\mathbf{v}}{dt} &= aD_L\mathbf{v} + \tau_L(v_0 - w_0)\mathbf{e}_L, \\ \frac{d\mathbf{w}}{dt} &= aD_R\mathbf{w} + \tau_R(w_0 - v_0)\mathbf{e}_R, \end{aligned}$$

where

$$\mathbf{e}_L = H_L^{-1} \begin{bmatrix} \vdots \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{e}_R = H_R^{-1} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}.$$

The energy method is applied with the scalar products $\langle \mathbf{v}^{(1)}, H_L \mathbf{v}^{(2)} \rangle$, $\langle \mathbf{w}^{(1)}, H_R \mathbf{w}^{(2)} \rangle$, and we have

$$\begin{aligned}
& \frac{d}{dt} (\langle \mathbf{v}, H_L \mathbf{v} \rangle + \langle \mathbf{w}, H_R \mathbf{w} \rangle) \\
&= 2 \langle \mathbf{v}, H_L (a D_L \mathbf{v} + \tau_L (v_0 - w_0) \mathbf{e}_L) \rangle + 2 \langle \mathbf{w}, H_R (a D_R \mathbf{w} + \tau_R (w_0 - v_0) \mathbf{e}_R) \rangle \\
&= (a + 2\tau_L)|v_0|^2 - 2\tau_L v_0 w_0 - (a - 2\tau_R)|w_0|^2 - 2\tau_R v_0 w_0 = [v_0 \ w_0] C \begin{bmatrix} v_0 \\ w_0 \end{bmatrix},
\end{aligned} \tag{7.26}$$

where

$$C = \begin{bmatrix} a + 2\tau_L & -(\tau_L + \tau_R) \\ -(\tau_L + \tau_R) & -a + 2\tau_R \end{bmatrix}.$$

This is a negative semidefinite matrix if

$$\tau_L \leq -\frac{a}{2} \quad \text{and} \quad \tau_R = \tau_L + a.$$

Consequently, this is the stability condition.

7.5 Summary

For high order approximations, the physical boundary conditions are not enough to define the complete difference scheme, and there is a need for numerical boundary conditions. In the first section we showed how such conditions can be constructed by differentiation of the physical boundary condition and the PDE. This is another application of the same principle as we used in Chapter 6 to obtain higher order approximations at inner points.

For first order hyperbolic systems, there are not enough physical boundary conditions to allow for this technique, and here is where the SBP operators come in. They are constructed such that they mimic the energy loss property at outflow boundaries for the PDE. The question then is how to enforce the physical boundary conditions that are prescribed if there is at least one characteristic entering the domain. We have presented three different ways of doing this: injection, projection and SAT. The first option is the easiest one to implement, but we have shown that there are cases where it leads to growing solutions, see also [Gustafsson, 1998]. The projection is a reformulation (7.19) of the problem, and by choosing the form (7.20) of the projection operator, stability is guaranteed. With a simpler choice of the projection operator, stability requires some restrictions on the approximation, and we have stated some theorems about this.

The third possibility is the SAT-method, which introduces a penalty term in the approximation of the differential equation. This method is quite simple to implement, and is gaining popularity. Practical cases show that full accuracy is obtained even if the local accuracy near the boundaries is decreased, but a general proof is still missing without strengthening the conditions on the difference scheme.

The development of SBP operators started in the middle of the seventies. Galerkin finite element methods were strongly on advance, and they satisfy the crucial con-

dition (7.3) automatically. Kreiss wanted to show that one could construct high order finite difference approximations with the same property, and together with his student Godela Scherer he developed the first such approximations of $\partial/\partial x$, see [Kreiss and Scherer, 1974] and [Kreiss and Scherer, 1977]. Later, Pelle Olsson and Bo Strand made further extensions and analysis of these operators, and also used them for nontrivial problems, see [Olsson, 1995a], [Olsson, 1995b] and [Strand, 1994]. The main actors in more recent developments are Jan Nordström and his students and colleagues, among them Mark Carpenter at NASA Langley Research Center. Important papers from the last decade are [Carpenter et al., 1999], [Nordström and Carpenter, 1999], [Mattsson, 2003], [Mattsson and Nordström, 2004], [Mattsson and Nordström, 2006]. Here there is more emphasis on the implementation of the SBP operators and how to enforce the physical boundary conditions. Furthermore, the construction of approximations of the second derivative $\partial^2/\partial x^2$ is considered in these papers. An extra difficulty arises when the PDE contains both first and second order derivatives, as for example the Navier-Stokes equations, since the same norm must be used for all terms in the equation.

The SAT method was originally developed for pseudospectral Chebyshev methods in [Funaro and Gottlieb, 1988], and was then extended to difference methods by Carpenter et.al in [Carpenter et al., 1994].

Chapter 8

The Box Scheme

If one is strict about the meaning of high order methods, this chapter doesn't belong in this book, because the box scheme is only second order accurate. The reason for bringing it in, is that the error coefficient is very small compared to other second order methods because of the very compact structure. The small error coefficient makes it competitive with higher order schemes for many applications. Furthermore, the box structure allows for a very effective implementation, since it can be advanced in time with very little work, and still being unconditionally stable.

8.1 The Original Box Scheme

We begin by considering the equation

$$u_t + au_x = 0, \quad (8.1)$$

where a is a constant. We define the average operator by

$$Au_j = \frac{1}{2}(u_{j+1} + u_j),$$

and for convenience we use the notation D instead of D_+ for the forward difference operator. The box scheme is

$$A(u_j^{n+1} - u_j^n) + \frac{ak}{2}D(u_j^{n+1} + u_j^n) = 0. \quad (8.2)$$

For periodic solutions we have on the Fourier side

$$\hat{u}^{n+1} = \hat{Q}\hat{u}^n = \frac{\cos(\xi/2) - i\lambda \sin(\xi/2)}{\cos(\xi/2) + i\lambda \sin(\xi/2)} \hat{u}^n,$$

where $\lambda = ak/h$, i.e., the method is unconditionally stable. Furthermore, since $|\hat{Q}| = 1$, there is exact energy conservation.

Let us next investigate the accuracy. In Section 6.2 we compared different fully discrete schemes, and we use the same type of comparison here. For $a = 1$ and $f(x) = e^{i\omega x}$, we write

$$\hat{Q}(\xi) = e^{-i\lambda\phi(\xi)},$$

where $\phi(\xi)$ is real, and compare $\phi(\xi)$ with ξ . In 6.2 the amplification factor was derived for the standard second order Crank-Nicholson scheme CN_{22} ($\hat{Q} = \hat{Q}_{2,2}$) and the version with fourth order accuracy in space CN_{42} ($\hat{Q} = \hat{Q}_{4,2}$). Figure 8.1 shows $\phi(\xi)$ for the box scheme and for the Crank-Nicholson schemes.

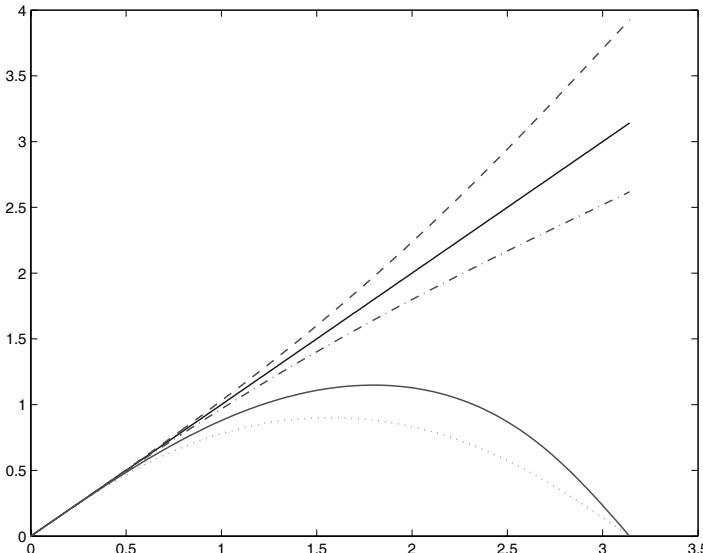


Fig. 8.1 Phase speed $\phi(\xi)$, in descending order: Box $\lambda = 0.8$, Exact (and Box $\lambda = 1.0$), Box $\lambda = 1.2$, CN_{42} $\lambda = 1.2$, CN_{22} $\lambda = 1.2$

The very good accuracy of the box scheme for a wide range of wave numbers is clearly seen from this figure. Note that, in contrast to standard centered schemes, the phase speed for the box scheme is higher than the true one for $\lambda < 1$.

Also here we make a special comparison for $\xi = \pi/2$, which corresponds to 4 points per wavelength. The error in the interval $0 \leq \xi \leq \pi/2$ is maximal at $\xi = \pi/2$ for all the approximations, and Figure 8.2 shows $\phi(\pi/2)$ for $0.5 \leq \lambda \leq 1.5$.

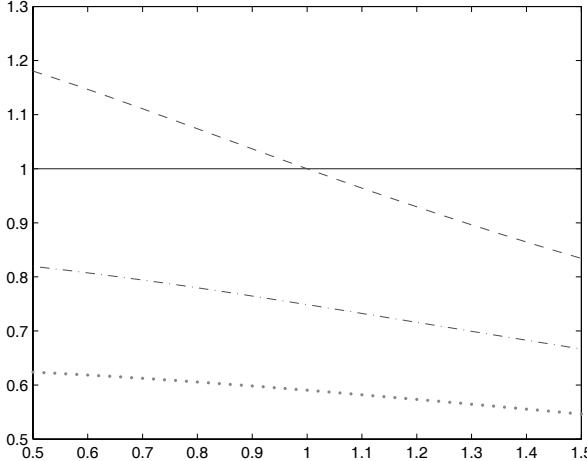


Fig. 8.2 Phase speed $\phi(\pi/2)$, $0.5 \leq \lambda \leq 1.5$ in descending order: Box scheme, CN_{42} , CN_{22}

For real application problems, a is of course not constant. However, from an accuracy point of view, one should try to choose the time step such that $ka(x)/h \approx 1$ in some average sense.

Let us next turn to the initial-boundary value problem. If $a > 0$, then we specify the boundary condition $u(0, t) = g(t)$, and the box scheme is

$$\begin{aligned} (A + \frac{ak}{2}D)u_j^{n+1} &= (A - \frac{ak}{2}D)u_j^n, \quad j = 0, 1, \dots, N-1, \\ u_0^{n+1} &= g^{n+1}, \\ u_j^0 &= f_j, \quad j = 0, 1, \dots, N. \end{aligned} \tag{8.3}$$

The scheme is very convenient to implement, since we can compute the new values u_j^{n+1} from left to right as shown in Figure 8.3.

Considering stability, we define the discrete scalar product and norm by

$$(u, v)_h = \sum_{j=0}^{N-1} u_j v_j h, \quad \|u\|_h^2 = (u, u)_h,$$

and obtain

$$\begin{aligned} (Au, Du)_h &= \frac{1}{2} \sum_{j=0}^{N-1} (u_{j+1} + u_j)(u_{j+1} - u_j) = \frac{1}{2} \sum_{j=0}^{N-1} (u_{j+1}^2 - u_j^2) \\ &= \frac{1}{2}(u_N^2 - u_0^2) = \frac{1}{2}u_N^2. \end{aligned} \tag{8.4}$$

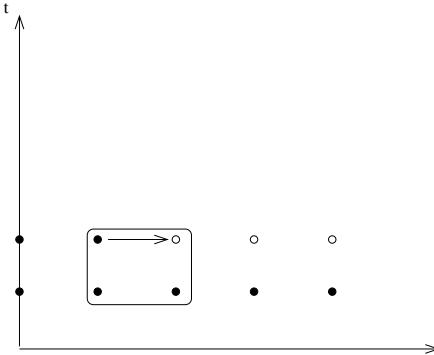


Fig. 8.3 Upper right point computed from the other three in the rectangle

By taking the squared norm of both sides of the first equation in (8.3) and using (8.4), we get

$$\|Au^{n+1}\|_h^2 + \left\| \frac{ak}{2} Du^{n+1} \right\|_h^2 = \|Au^n\|_h^2 + \left\| \frac{ak}{2} Du^n \right\|_h^2 - \frac{ak}{2} (|u_N^n|^2 + |u_N^{n+1}|^2). \quad (8.5)$$

The last term represents the energy leakage through the boundary $x = 1$. It remains to show that we get an estimate in terms of $\|u^{n+1}\|_h$. With $\lambda = ak/h$, we have for any u satisfying the boundary condition

$$\begin{aligned} \|Au\|_h^2 + \left\| \frac{ak}{2} Du \right\|_h^2 &= \frac{1+\lambda^2}{4} \sum_{j=0}^{N-1} (u_j^2 + u_{j+1}^2)h + \frac{1-\lambda^2}{4} \sum_{j=0}^{N-1} 2u_j u_{j+1} h \\ &= \frac{1+\lambda^2}{2} \|u\|_h^2 + \frac{1+\lambda^2}{4} u_N^2 h + \frac{1-\lambda^2}{4} \sum_{j=0}^{N-1} 2u_j u_{j+1} h. \end{aligned}$$

But

$$\left| \sum_{j=0}^{N-1} 2u_j u_{j+1} h \right| \leq \sum_{j=0}^{N-1} (u_j^2 + u_{j+1}^2)h = 2\|u\|_h^2 - u_0^2 h + u_N^2 h,$$

and by defining

$$\|u\|_h^2 = \|u\|_h^2 + \frac{h}{2} u_N^2$$

we get

$$c_1^2 \|u\|_h^2 \leq \|Au\|_h^2 + \left\| \frac{ak}{2} Du \right\|_h^2 \leq c_2^2 \|u\|_h^2,$$

where

$$\begin{aligned} c_1 &= \min(\lambda, 1), \\ c_2 &= \max(\lambda, 1). \end{aligned} \quad (8.6)$$

Therefore we get from (8.5)

$$\|u^n\|_{\tilde{h}} \leq \frac{c_2}{c_1} \|f\|_{\tilde{h}}$$

independent of $\lambda = ak/h > 0$. If a approaches zero, the estimate breaks down, and this has severe implications if $a = a(x)$ as we shall see.

For variable and Lipschitz continuous coefficients $a(x)$, one can show by generalizing the energy method in the usual way, that the solutions satisfy

$$\|u^n\|_{\tilde{h}} \leq K e^{\alpha t_n} \|f\|_{\tilde{h}}. \quad (8.7)$$

Here K depends on \tilde{c}_2/\tilde{c}_1 , where

$$\begin{aligned}\tilde{c}_1 &= \min\left\{\min_j(|a_j|k/h), 1\right\}, \\ \tilde{c}_2 &= \max\left\{\max_j(|a_j|k/h), 1\right\}.\end{aligned}$$

The constant K tends to infinity as $\tilde{c}_1 \rightarrow 0$ (this would be the case also for periodic boundary conditions). This leads to the requirement that a does not change sign anywhere in the domain, and for our initial-boundary value problem, a has to be positive. This condition is real as we shall see.

Assume that $a(0) > 0$, $a(1) > 0$, but negative in parts of the inner domain, such that $a(x) = 0$ at some points. As noted above, the solution is obtained by solving from left to right. Even if this is technically possible even for negative a , the algorithm cannot be expected to work in practice. For simplicity, assume that $a(x) = \bar{a} < 0$ in some interval. Then

$$u_{j+1}^{n+1} = -\frac{1-k\bar{a}/h}{1+k\bar{a}/h} u_j^{n+1} + R(u^n), \quad j = j_0, j_0 + 1, \dots,$$

where $R(u^n)$ contain values from the previous time level only. The amplification factor in the x -direction is obviously larger than one in magnitude, showing that the scheme is unstable. Indeed, the scheme cannot be advanced a single time-step before it blows up if the mesh is fine enough.

This is of course a serious drawback with the scheme. It can be used for hyperbolic systems, but only if all the waves propagate in the same direction in the whole domain. In the next section we propose a remedy for this difficulty.

8.2 The Shifted Box Scheme

In applications the coefficients are in general not constant. Furthermore, we are most often dealing with problems with some inherent energy conserving property. Therefore we consider the model problem

$$\begin{aligned}u_t + \frac{1}{2}((au)_x + au_x) &= 0, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(x, 0) &= f(x),\end{aligned} \quad (8.8)$$

where $a = a(x, t)$. For periodic solutions, the differential operator in space is skew-symmetric, since

$$(u, (av)_x + av_x) = -((au)_x + au_x, v)$$

for all u, v , and this implies energy conservation $\|u(\cdot, t)\| = \|f(\cdot)\|$.

The box scheme is

$$\begin{aligned} Au_j^{n+1} &+ \frac{k}{4} ((a_j^{n+1} u_j^{n+1}) + (A \frac{a_j^n + a_j^{n+1}}{2}) Du_j^{n+1}) \\ &= Au_j^n - \frac{k}{4} ((a_j^n u_j^n) + (A \frac{a_j^n + a_j^{n+1}}{2}) Du_j^n), \quad j = 0, 1, \dots, N-1, \\ u_j^0 &= f_j, \quad j = 0, 1, \dots, N, \end{aligned}$$

which can also be solved as an explicit scheme from left to right if $a(x, t) > 0$. Furthermore, there are no stability problems in that case, and we would like to take advantage of that. We make the variable transformation $\xi = x + ct$, where c is a constant such that $c + a(x, t) > 0$ for all x, t , and define $v(\xi, t) = u(x, t)$, $\tilde{a}(\xi, t) = a(x, t)$. The new differential equation is

$$v_t + \frac{1}{2} ((c + \tilde{a}) v)_\xi + \frac{1}{2} (c + \tilde{a}) v_\xi = 0, \quad (8.9)$$

which represents a right-going wave. When applying the box scheme to this equation, the solution is well behaved. However, in one time step, the solution v has propagated a distance ck longer in the positive ξ -direction than u , i.e., $\xi = x + ck$. Therefore, staying in the (x, t) space, the true solution u at a certain point (\bar{x}, t_{n+1}) is obtained by the shift

$$u(\bar{x}, t_{n+1}) = v(\bar{x} + ck, t_{n+1}).$$

In order to avoid interpolation, the added velocity c is chosen such that $ck = \mu h$, or equivalently, $\lambda c = \mu$, where μ is an integer and $\lambda = k/h$. Hence, after computing v_j^{n+1} , the solution u_j^{n+1} is obtained by $u_j^{n+1} = v_{j+\mu}^{n+1}$, which is used as initial values for computing v and u at the next time level.

A particularly convenient choice is $\mu = 1$, i.e., $\lambda = 1/c$. However, in some cases this λ -value may be too large. In such a case λ is chosen as $1/(rc)$, where r is an integer. Then the scheme for v is advanced r steps before u is obtained by a shift of v .

Note that $\tilde{a}(\xi, t) = a(\xi - ct, t)$ has to be centered correctly to keep second order accuracy. The shifted box scheme for marching from time step $n+s$ to $n+s+1$, where $0 \leq s \leq r-1$, is

$$\begin{aligned} \tilde{v}_{j+1}^{n+s+1} (1 + \frac{\lambda}{2} (\omega_{j+1}^{n+s+1} + \omega_c)) + \tilde{v}_j^{n+s+1} (1 - \frac{\lambda}{2} (\omega_j^{n+s+1} + \omega_c)) &= \\ \tilde{v}_{j+1}^{n+s} (1 - \frac{\lambda}{2} (\omega_{j+1}^{n+s} + \omega_c)) + \tilde{v}_j^{n+s} (1 + \frac{\lambda}{2} (\omega_j^{n+s} + \omega_c)), & \\ j = \mu, 1 + \mu, \dots, N + \mu - 1, \quad \tilde{v}_\mu^{n+s+1} &= g^{n+s+1}, \end{aligned} \quad (8.10)$$

where ω is defined as:

$$\begin{aligned}\omega_j^{n+s} &= c + \left(1 - \frac{s}{r}\right)a_j^n + \frac{s}{r}a_j^{n+r}, \\ \omega_c &= \frac{1}{2}(A\omega_j^{n+s} + A\omega_j^{n+s+1}).\end{aligned}$$

Since the solution is shifted every r th step, v_j^n does not approximate the solution of (8.9), except for the first step, and therefore we use the notation \tilde{v}_j^n . The complete algorithm is

1. Define initial data $u_j^0 = \tilde{v}_j^0 = f_j$, $j = 0, 1, \dots, N + \mu$ (for extra outside points, use extrapolation, see 4)
2. Compute \tilde{v}_j^{n+r} , $j = 1 + \mu, 2 + \mu, \dots, N + \mu$ by the box scheme
3. $u_j^{n+r} = \tilde{v}_{j+\mu}^{n+r}$, $j = 1, 2, \dots, N$
4. Compute extra boundary points u_{N+j}^{n+r} by extrapolation

$$u_{N+j}^{n+r} = 2u_{N+j-1}^{n+r} - u_{N+j-2}^{n+r}, \quad j = 1, 2, \dots, \mu$$
5. $n + r \rightarrow n$
6. $\tilde{v}_j^n = u_j^n$, $j = 0, 1, \dots, N + \mu$
7. If $t \leq T$, go to 2

For the periodic case, the stability is not affected, since the norm of the solution does not change during a cyclic shift. In the nonperiodic case, the extrapolation in 4. is necessary, and the stability proof above is not valid anymore. However, there is no experimental evidence of any instabilities.

Next we present results from a test case, with the data

$$\begin{aligned}a(x) &= 0.5 + 0.6 \cos(2\pi x), \\ f(x) &= e^{-800(x-0.6)^2},\end{aligned}\tag{8.11}$$

as shown in Figure 8.4.

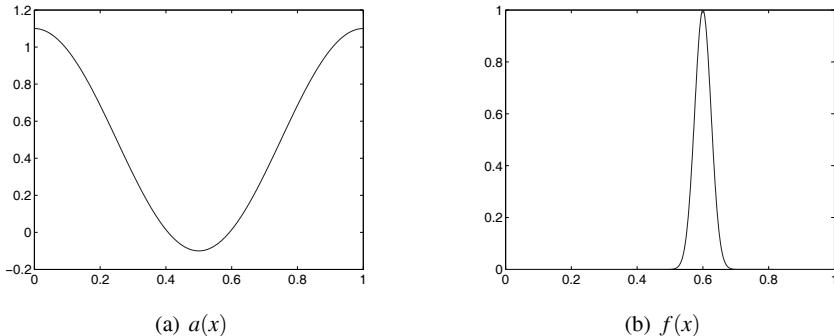
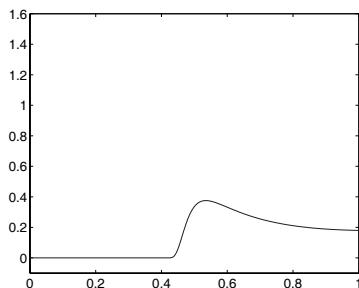


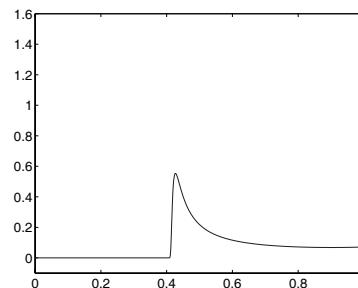
Fig. 8.4 Velocity $a(x)$ and initial function $f(x)$

This is a challenging problem, since the characteristics are converging near the zero velocity point $\bar{x} = 0.4068$, where they become vertical. This means that the left part of the pulse is moving towards the left, while at the same time the zero values are retained for $x \leq \bar{x}$. The result is that after some time a very sharp spike is formed with a vertical left side at $x = \bar{x}$.

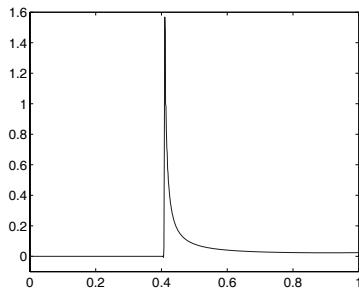
Figure 8.5, shows the solution u obtained by the shifted box scheme at three different times for $N = 800$. In the first three figures we have used $c = 1$ and $\lambda = 1$, and in the last one $c = 0.5$ and $\lambda = 2$. The mesh is rather fine, but it is still a quite amazing result considering the very thin spike at $t = 3$. Indeed, only 5 grid points are representing the upper half ($u > 0.8$) of the spike. Note also, that the larger CFL-number in the last figure gives virtually the same result.



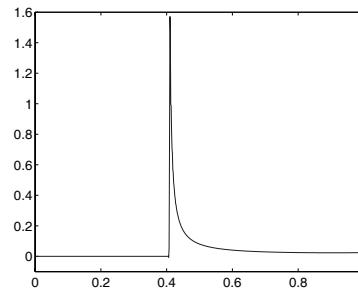
(a) $t=1, c = 1, \lambda = 1$



(b) $t=2, c = 1, \lambda = 1$



(c) $t=3, c = 1, \lambda = 1$



(d) $t=3, c = 0.5, \lambda = 2$

Fig. 8.5 $u(x,t)$, $N = 800$

8.3 Two Space Dimensions

Here we consider the problem

$$u_t + (au)_x + (bu)_y = 0, \quad (8.12)$$

where the coefficients a and b satisfy

$$a_x + b_y = 0.$$

This extra divergence condition makes the equation (8.12) energy conserving, i.e.,

$$\frac{d}{dt} \int \int u^2 dx dy = 0.$$

For incompressible flow, a and b represent the velocity components, and the equation is a model describing the distribution of some physical component u in the given flow (scalar transport).

Because of the divergence condition, one can also use the equation

$$u_t + au_x + bu_y = 0.$$

The complete initial-boundary value problem is

$$\begin{aligned} u_t + (au)_x + (bu)_y &= 0, \quad 0 \leq x, y < 1, \quad 0 \leq t, \\ u(0, y, t) &= g_1(y, t), \quad 0 \leq y \leq 1, \\ u(x, 0, t) &= g_2(x, t), \quad 0 \leq x \leq 1, \\ u(x, y, 0) &= f(x, y), \end{aligned}$$

where it is assumed that

$$a(0, y) > 0, \quad a(1, y) > 0, \quad b(x, 0) > 0, \quad b(x, 1) > 0.$$

This means that the flow is entering the domain at the south and west boundaries, and it is leaving the domain at the east and north boundaries. The grid in space is defined by

$$\begin{aligned} x_i &= ih_1, \quad i = 0, 1, \dots, M, \quad Mh_1 = 1, \\ y_j &= jh_2, \quad j = 0, 1, \dots, N, \quad Nh_2 = 1, \end{aligned}$$

where h_1 h_2 are the step sizes in the x - and y -direction, respectively. With the notation u_{ij}^n for the grid function at (x_i, y_j, t_n) , the average and difference operators are defined by

$$\begin{aligned} A_x u_{ij} &= (u_{i+1,j} + u_{ij})/2, \quad A_y u_{ij} = (u_{i,j+1} + u_{ij})/2, \\ D_x u_{ij} &= (u_{i+1,j} - u_{ij})/h_1, \quad D_y u_{ij} = (u_{i,j+1} - u_{ij})/h_2. \end{aligned}$$

The box scheme is

$$A_x A_y (u_{ij}^{n+1} - u_{ij}^n) + \frac{k}{2} \left(A_y D_x ((au)_{ij}^{n+1} + (au)_{ij}^n) + A_x D_y ((bu)_{ij}^{n+1} + (bu)_{ij}^n) \right) = 0,$$

and we assume first periodic solutions. The Fourier transforms of the average and difference operators are (after a half step shift)

$$\hat{A}_x = \cos \xi, \quad \hat{A}_y = \cos \eta, \quad \hat{D}_x = \frac{2}{h_1} i \sin \xi, \quad \hat{D}_y = \frac{2}{h_2} i \sin \eta.$$

Here $\xi = \omega_x h_1 / 2$, $\eta = \omega_y h_2 / 2$, where ω_x , ω_y are the wave numbers corresponding to the x - and y -direction, respectively. Assuming that a and b are constants, the Fourier transform of the box scheme is

$$\begin{aligned} & (\cos \xi \cos \eta + i \lambda_x \cos \eta \sin \xi + i \lambda_y \cos \xi \sin \eta) \hat{u}^{n+1} \\ &= (\cos \xi \cos \eta - i \lambda_x \cos \eta \sin \xi - i \lambda_y \cos \xi \sin \eta) \hat{u}^n, \end{aligned}$$

where $\lambda_1 = ak/h_1$, $\lambda_2 = bk/h_2$. If $|\cos \xi| + |\cos \eta| \neq 0$, then obviously

$$|\hat{u}^{n+1}| = |\hat{u}^n|,$$

and we have energy conservation. The case $\cos \xi = \cos \eta = 0$ is special, since the coefficient of \hat{u}^{n+1} is zero, corresponding to a singularity of the system. This singularity is removed by introducing nonperiodic boundary conditions (or by using a grid with an odd number of grid points N in the periodic case).

The initial-boundary value problem is

$$\begin{aligned} & A_x A_y (u_{ij}^{n+1} - u_{ij}^n) + \frac{k}{2} \left(A_y D_x ((au)_{ij}^{n+1} + (au)_{ij}^n) + A_x D_y ((bu)_{ij}^{n+1} + (bu)_{ij}^n) \right) = 0, \\ & i = 0, 1, \dots, M-1, \quad j = 0, 1, \dots, N-1, \\ & u_{0j}^{n+1} = g_1(y_j, t_{n+1}), \quad j = 0, 1, \dots, N, \\ & u_{i0}^{n+1} = g_2(x_i, t_{n+1}), \quad i = 0, 1, \dots, M, \\ & u_{ij}^0 = f_{ij}, \quad i = 0, 1, \dots, M, \quad j = 0, 1, \dots, N. \end{aligned} \tag{8.13}$$

As for the 1-D case, it is possible to advance the algorithm using no more work than for an explicit scheme. With u_{ij}^n given, the solution at the next time-level is obtained from left to right for each y_j in the (x, y) plane (or alternatively, in the vertical direction for each x_i).

The scalar product and norm are defined by

$$(u, v)_h = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} u_{ij} v_{ij} h_1 h_2, \quad \|u\|_h^2 = (u, u)_h.$$

For $g_1 = g_2 = 0$ and with positive and constant a, b , one can show in analogy with the 1-D case that

$$\begin{aligned} & \|A_x A_y u^{n+1}\|_h^2 + \left\| \frac{k}{2} (a A_y D_x + b A_x D_y) u^{n+1} \right\|_h^2 \\ &= \|A_x A_y u^n\|_h^2 + \left\| \frac{k}{2} (a A_y D_x + b A_x D_y) u^n \right\|_h^2 \\ &\quad - \frac{\lambda_1}{16} \sum_{j=0}^{N-1} ((u_{Mj}^n + u_{M,j+1}^n)^2 + (u_{Mj}^{n+1} + u_{M,j+1}^{n+1})^2) h_2 \\ &\quad - \frac{\lambda_2}{16} \sum_{i=0}^{M-1} ((u_{iN}^n + u_{i+1,N}^n)^2 + (u_{iN}^{n+1} + u_{i+1,N}^{n+1})^2) h_1. \end{aligned}$$

As in the 1-D case, it can be shown that this leads to a stability estimate for the l_2 -norm. Accordingly, even in the 2-D case, we have an unconditionally stable scheme, which can be advanced in time at the same cost as an explicit scheme.

For variable coefficients, we must require that the coefficients are positive in the whole domain, and therefore we need to generalize the shifted box scheme to 2-D for the general case. A constant c is added to the a -velocity, and a constant d to the b -velocity giving the new differential equation

$$v_t + ((c + \tilde{a})v)_\xi + ((d + \tilde{b})v)_\eta = 0, \quad (8.14)$$

where $c + \tilde{a} > 0$, $d + \tilde{b} > 0$. In this way the solution becomes shifted a distance ck in the x -direction, and a distance dk in the y -direction for each time step. We choose $\lambda_x c = \mu$, and $\lambda_y d = \nu$, where μ and ν are integers. As in the 1-D case, it may also be necessary to choose a λ -value such that r time steps have to be taken before the solution is shifted a whole space step. The algorithm is completely analogous to the 1-D algorithm presented in Section 8.2, and it is omitted here.

We shall now present a few numerical experiments. We first define a divergence free velocity field with positive a and b , where there is no need for the shift operation:

$$\begin{aligned} a &= 1.0 + 0.1 \cos(2\pi x) \cos(2\pi y), \\ b &= 0.4 + 0.1 \sin(2\pi x) \sin(2\pi y). \end{aligned} \quad (8.15)$$

This field is divergence free also in the box-scheme approximation sense:

$$A_y D_x a_{ij} + A_x D_y b_{ij} = 0, \quad i = 0, 1, \dots, M-1, \quad j = 0, 1, \dots, N-1.$$

In the first experiment, we compute the movement of a sharp pulse, which initially has the form

$$u(x, y, 0) = e^{-800((x-0.2)^2 + (y-0.2)^2)}.$$

Figure 8.6, shows the solution at $t = 0$, $t = 0.6$, $t = 0.82$, where in the last case the pulse is half way through the computational boundary. The compact nature of the scheme is such that no numerical boundary conditions are required at the outflow boundaries. Therefore the pulse is passing through $x = 1$ without any trouble. The

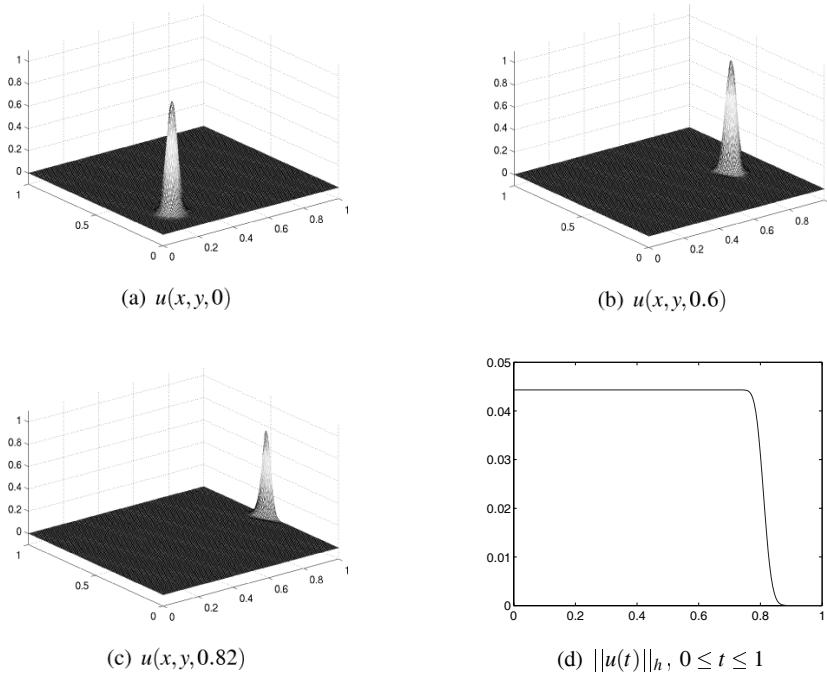


Fig. 8.6 Box scheme solution, $M = N = 200$, $\lambda = 1$

lower right figure shows the norm as a function of time. After a constant phase, it goes down smoothly to zero as the pulse passes out through the boundary.

As a test of the robustness of the scheme, we have also run a case with random data distributed on $[0, 1] \times [0, 1]$ for the initial function f and for the boundary functions g_1, g_2 in (8.13). The scheme was run for a long time interval ($0 \leq t \leq 100$), and with two quite different CFL-numbers $\lambda = \lambda_1 = \lambda_2 = 1$ and $\lambda = 20$. The norm is not constant in this case, since energy is added through the inflow boundaries in a random, but bounded fashion. In both cases the solutions stay bounded, see Figure 8.7.

Next we present the shifted box scheme for a case where the a -velocity is negative in part of the domain. The velocity field is

$$\begin{aligned} a &= 0.06 + 0.08 \cos(2\pi x) \cos(0.4y), \\ b &= 0.5 + 0.4 \sin(2\pi x) \sin(0.4y), \end{aligned} \tag{8.16}$$

and in the v -equation (8.14) we choose $c = 1$, $d = 0$.

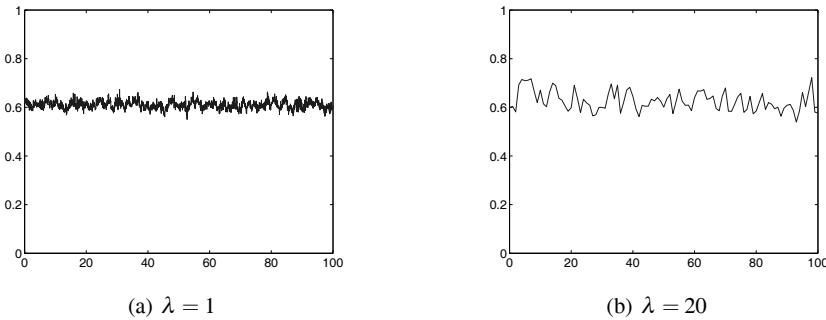


Fig. 8.7 $\|u(t)\|_h$, $0 \leq t \leq 100$, random initial and boundary data, $M = N = 20$

The initial data are

$$u(x, y, 0) = e^{-400((x-0.5)^2 + (y-0.2)^2)}.$$

Figure 8.8 shows the solution at $t = 0$ and at $t = 1.5$, where the pulse is half way through the outflow boundary $y = 1$.

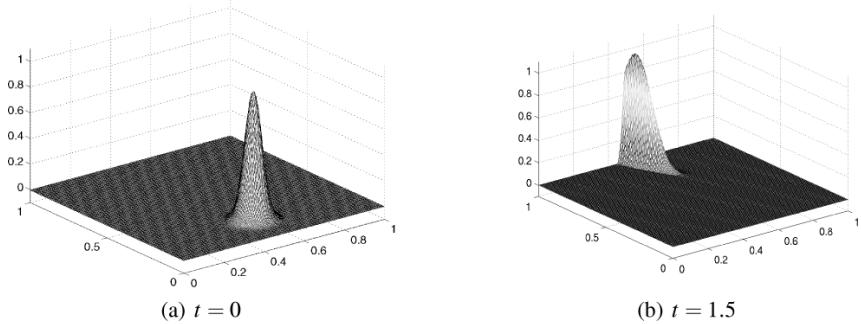


Fig. 8.8 $\|u(t)\|$, $M = N = 200$, $c = 1$, $d = 0$, $\lambda = 0.5$

8.4 Nonuniform Grids

In this section we shall discuss the box scheme on a nonuniform grid with grid points x_j , $j = 0, 1, \dots, N$, and step sizes

$$h_{j+1/2} = x_{j+1} - x_j, \quad j = 0, 1, \dots, N-1.$$

Here we don't make any assumption on smoothness of the grid, i.e., the grid points have a completely arbitrary distribution. In Section 8.2 we considered an energy

conserving 1-D model problem. In order to indicate the connection between the norm in the continuous and discrete case, we consider first the different problem

$$\begin{aligned} u_t + a(x)u_x &= 0, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= g(t), \\ u(x, 0) &= f(x). \end{aligned} \tag{8.17}$$

Here we assume that $a(x)$ is a given positive time-independent function satisfying $0 < a_{\min} \leq a(x) \leq a_{\max}$. Later we shall treat the case where $a(x)$ may be negative in part of the domain, which requires the use of the shifted box scheme.

The differential equation in (8.17) does not preserve energy. However, there is a bound on the L_2 -norm, which is independent of t . Let the modified norm be defined by

$$\|u\|_a^2 = \int_0^1 |u(x, t)|^2 \frac{1}{a(x)} dx.$$

Then, for $g(t) = 0$,

$$\frac{d}{dt} \|u\|_a^2 = \int_0^1 2uu_t \frac{1}{a} dx = - \int_0^1 2uu_x dx = -|u(1, t)|^2,$$

i.e.,

$$\|u(\cdot, t)\|_a^2 + \int_0^t |u(1, \tau)|^2 d\tau = \|f(\cdot)\|_a^2.$$

For the standard L_2 -norm, we get

$$\|u(\cdot, t)\|^2 \leq \frac{a_{\max}}{a_{\min}} \|f(\cdot)\|^2 - a_{\max} \int_0^t |u(1, \tau)|^2 d\tau. \tag{8.18}$$

We note here, that if $a(x)$ approaches zero, then the estimate breaks down. This doesn't mean that the problem is ill posed, but there is in general no time-independent bound on the solution.

With the average and difference operators defined by

$$\begin{aligned} Au_j &= \frac{1}{2}(u_{j+1} + u_j), \\ Du_j &= \frac{1}{h_{j+1/2}}(u_{j+1} - u_j), \end{aligned}$$

the box scheme is

$$\begin{aligned} Au_j^{n+1} + \frac{k}{2}(Aa_j)Du_j^{n+1} &= Au_j^n - \frac{k}{2}(Aa_j)Du_j^n, \quad j = 0, 1, \dots, N-1, \\ u_0^{n+1} &= g^{n+1}, \\ u_j^0 &= f_j, \quad j = 0, 1, \dots, N. \end{aligned} \tag{8.19}$$

Here it is assumed that $a(x)$ is stored at the grid points x_j ; if $a(x)$ is available at the half-points $x_{j+1/2}$, then Aa_j is substituted by $a(x_{j+1/2})$. Actually, we will later use the notation $a_{j+1/2}$ for Aa_j .

Also in this nonuniform case, there is the great advantage that u_{j+1}^{n+1} is obtained as a combination of $u_j^n, u_{j+1}^n, u_j^{n+1}$, i.e., the work per time step is comparable to an explicit scheme.

Concerning accuracy, we note that the computational stencil involves only two grid points in the x -direction, and each space interval is independent of the neighboring intervals. Hence, the scheme is second order accurate, and the truncation error is $\mathcal{O}(h^2)$, where $h = \max_j(h_{j+1/2})$. This property is not shared by usual central difference schemes, where the order of accuracy drops from second to first order if applied directly on a nonuniform grid.

For the stability proof, we define the discrete scalar product and norm by

$$(u, v)_* = \sum_{j=0}^{N-1} u_j v_j d_{j+1/2}, \quad \|u\|_*^2 = (u, u)_*,$$

where $d_{j+1/2} = h_{j+1/2}/a_{j+1/2}$. We first note that

$$\begin{aligned} (Au, (Aa)Du)_* &= \frac{1}{2} \sum_{j=0}^{N-1} (u_{j+1} + u_j)(u_{j+1} - u_j) \\ &= \frac{1}{2} \sum_{j=0}^{N-1} (u_{j+1}^2 - u_j^2) = \frac{1}{2}(u_N^2 - u_0^2) = \frac{1}{2}u_N^2. \end{aligned}$$

By taking the squared norm of both sides of the first equation in (8.19), we get the relation

$$\|Au^{n+1}\|_*^2 + \left\| \frac{k(Aa)}{2} Du^{n+1} \right\|_*^2 = \|Au^n\|_*^2 + \left\| \frac{k(Aa)}{2} Du^n \right\|_*^2 - \frac{k}{2}(|u_N^n|^2 + |u_N^{n+1}|^2).$$

It remains to derive the final estimate for the standard l_2 -norm. It is practical to first define the norm without the right end point included:

$$\|u\|_h^2 = \sum_{j=0}^{N-1} |u_j|^2 h_{j+1/2}.$$

Let $\beta_{j+1/2} = ka_{j+1/2}/h_{j+1/2}$ be the local CFL-numbers. We get

$$\begin{aligned}
& \|Au\|_*^2 + \left\| \frac{k(Aa)}{2} Du \right\|_*^2 \\
&= \frac{1}{4} \sum_{j=0}^{N-1} ((1 + \beta_{j+1/2}^2)(u_j^2 + u_{j+1}^2) + (1 - \beta_{j+1/2}^2)2u_j u_{j+1}) d_{j+1/2} \\
&\leq \frac{1}{4} \sum_{j=0}^{N-1} \max_j (1 + \beta_{j+1/2}^2 + |1 - \beta_{j+1/2}^2|)(u_j^2 + u_{j+1}^2) d_{j+1/2} \\
&\leq c_2 (\|u\|_h^2 + \frac{1}{2} u_N^2 h_{N-1/2}),
\end{aligned}$$

where

$$c_2 = \frac{1}{2} \max_j \left(\frac{1}{a_{j+1/2}} (1 + \beta_{j+1/2}^2 + |1 - \beta_{j+1/2}^2|) \right).$$

In the same way we get

$$\begin{aligned}
& \|Au\|_*^2 + \left\| \frac{k(Au)}{2} Du \right\|_*^2 \\
&\geq \frac{1}{4} \sum_{j=0}^{N-1} \min_j (1 + \beta_{j+1/2}^2 - |1 - \beta_{j+1/2}^2|)(u_j^2 + u_{j+1}^2) d_{j+1/2} \\
&\geq c_1 (\|u\|_h^2 + \frac{1}{2} u_N^2 h_{N-1/2}),
\end{aligned}$$

where

$$c_1 = \frac{1}{2} \min_j \left(\frac{1}{a_{j+1/2}} (1 + \beta_{j+1/2}^2 - |1 - \beta_{j+1/2}^2|) \right).$$

By defining

$$\begin{aligned}
\lambda_{\min} &= \min_j (\beta_{j+1/2}^2 / a_{j+1/2}), \\
\lambda_{\max} &= \max_j (\beta_{j+1/2}^2 / a_{j+1/2}),
\end{aligned}$$

the definition of c_1 and c_2 can be given as

$$\begin{aligned}
c_1 &= \begin{cases} 1/a_{\max} & \text{if } \max_j \beta_{j+1/2} \leq 1 \\ \lambda_{\min} & \text{if } \max_j \beta_{j+1/2} > 1, \end{cases} \\
c_2 &= \begin{cases} 1/a_{\min} & \text{if } \max_j \beta_{j+1/2} \leq 1 \\ \lambda_{\max} & \text{if } \max_j \beta_{j+1/2} > 1. \end{cases}.
\end{aligned}$$

With the norm defined by

$$\|u\|_h^2 = \|u\|_h^2 + \frac{1}{2} u_N^2 h_{N-1/2},$$

we get the final estimate

$$\begin{aligned}
\|u^n\|_h^2 &\leq \frac{1}{c_1} (\|Au^n\|_*^2 + \left\| \frac{kAa}{2} Du^n \right\|_*^2) \\
&\leq \frac{1}{c_1} \left(\|Af\|_*^2 + \left\| \frac{kAa}{2} Df \right\|_*^2 - \sum_{\nu=1}^{n-1} |u_N^\nu|^2 k - \frac{k}{2} (|f_N|^2 + |u_N^n|^2) \right) \\
&\leq \frac{c_2}{c_1} \|f\|_h^2 - \frac{1}{c_1} \sum_{\nu=1}^{n-1} |u_N^\nu|^2 k - \frac{k}{2c_1} (|f_N|^2 + |u_N^n|^2).
\end{aligned}$$

For CFL-numbers $\beta_{j+1/2} \leq 1$, this estimate is a direct discretization of (8.18).

To verify the accuracy, the scheme was run for $a = 1$, $\beta = 0.8$, with the initial function

$$f(x) = e^{-800(x-0.2)^2}.$$

The solution at $t=0.5$ is

$$u(x, 0.5) = e^{-800(x-0.7)^2}.$$

The step size for the nonuniform grid was taken as

$$h_{j+1/2} = \frac{1}{N}(1 + 2\text{rand}_j)/2, \quad (8.20)$$

where rand_j are random numbers in the interval $(0, 1)$. (The first step is always taken as $1/N$ for practical reasons.) The figures 8.9 and 8.10 show the step size and the solution for the nonuniform grids.

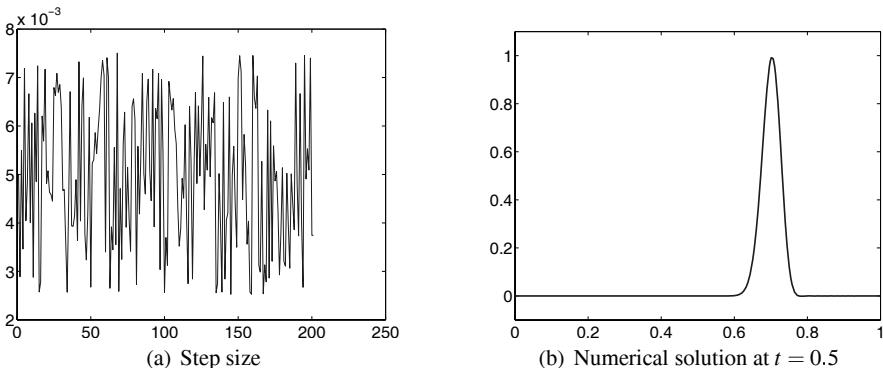


Fig. 8.9 Nonuniform grid, $N = 200$

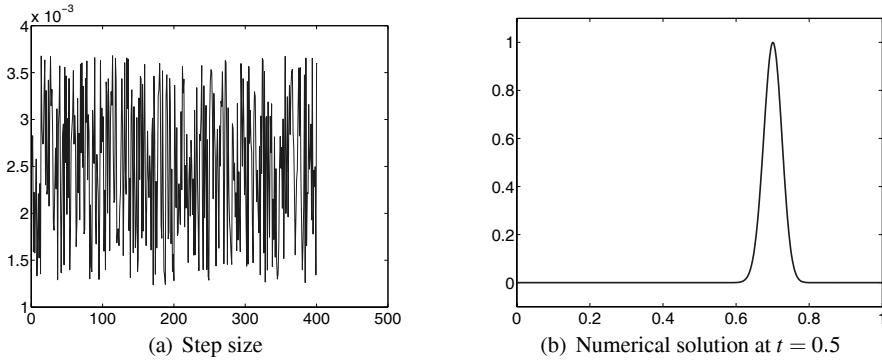


Fig. 8.10 Nonuniform grid, $N = 400$

Table 8.1 shows the l_2 -error for both the uniform and nonuniform grids.

Table 8.1 The l_2 -error

N	Uniform grid	Nonuniform grid
200	0.0069	0.0126
400	0.0017	0.0029

The second order accuracy is clearly seen for both cases, but the error constant is larger for the nonuniform case as expected.

If $a(x)$ is negative in part of the domain, we must use the shifted box scheme also for nonuniform grids. Also here we add a constant c to the coefficient $a(x)$. However, the simple procedure of applying a simple shift of the solution does not work, since the uniform grid is a fundamental pre-requisite for that procedure. However, we can substitute the shift by a second computation using the box scheme with wave speed $-c$. In this way, the complete algorithm can be considered as a split scheme with two stages. In the uniform case, the second stage is the shift operation.

The first stage can be seen as a variable transformation $\xi = x + ct$, just as we saw for the uniform grid case. Therefore we don't have to repeat the algorithm.

We now go back to the energy conserving problem (8.8). The box scheme is

$$\begin{aligned}
 Au_j^{n+1} + \frac{k}{4} \left(D(a_j^{n+1} u_j^{n+1}) + \left(A \frac{a_j^n + a_j^{n+1}}{2} \right) Du_j^{n+1} \right) \\
 = Au_j^n - \frac{k}{4} \left(D(a_j^n u_j^n) + \left(A \frac{a_j^n + a_j^{n+1}}{2} \right) Du_j^n \right), \quad j = 0, 1, \dots, N-1, \\
 u_0^{n+1} = g^{n+1}, \\
 u_j^0 = f_j, \quad j = 0, 1, \dots, N.
 \end{aligned}$$

In the second stage of each time step, we solve the equation

$$u_t - cu_x = 0$$

from right to left on the nonuniform grid with the computed solution from the first stage as initial data.

For the numerical experiment we solve the same problem as for the uniform case, with the data (8.11). Figure 8.11 shows the initial function and the solution at $t = 3$ on a uniform grid with 800 points.

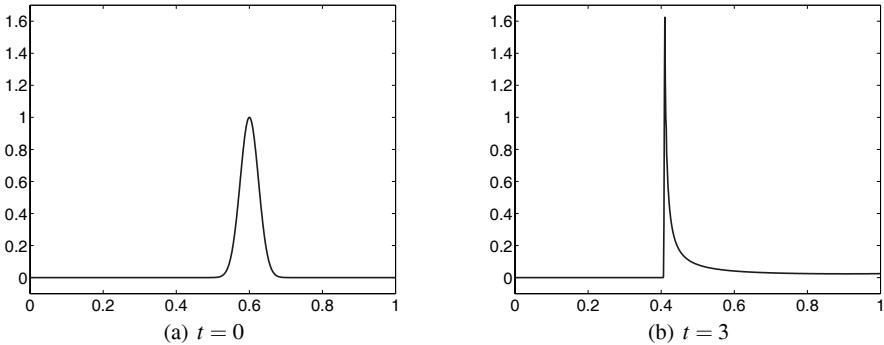


Fig. 8.11 Initial function and solution u at $t = 3$, $N = 800$

In the next experiment, the step size h is chosen as a piecewise constant function, with more points concentrated in the right part of the domain. This is shown in Figure 8.12 together with the solution u at $t = 3$.

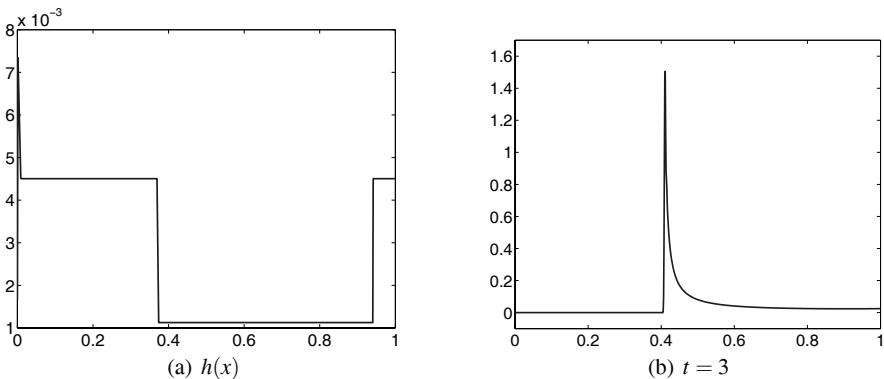


Fig. 8.12 Step size $h(x)$ and solution u at $t = 3$, $N = 600$

Finally, Figure 8.13 shows the result of a computation with random step size according to (8.20). The result is amazingly accurate for this nonsmooth solution and very rough grid.

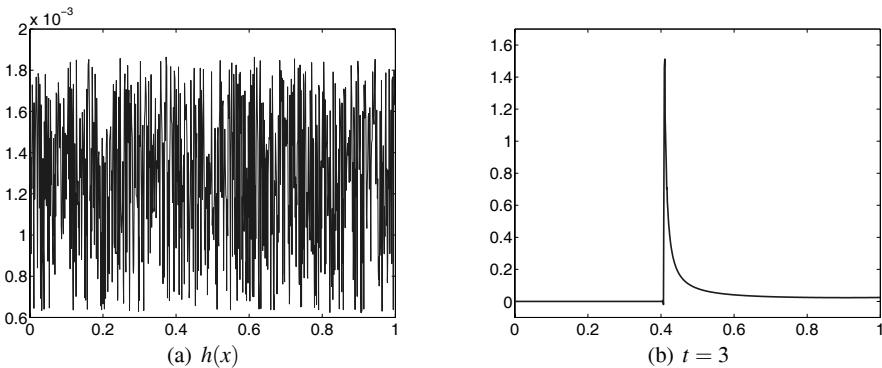


Fig. 8.13 Step size $h(x)$ and solution u at $t = 3$, $N = 800$

8.5 Summary

The box scheme is included in this book because its compact nature makes it very accurate. It is particularly effective for scalar transport problems, and it retains almost exactly the conservation properties of the differential equation. It is unconditionally stable, and is advanced at no more cost than an explicit scheme.

The original form can be used only for problems where the propagation direction does not change sign for any coordinate x , y or z , and this is of course too restrictive. We have shown how to modify it by a splitting technique where the characteristic speed is shifted by a constant, followed by a computation and then a shift back. The extra work is negligible.

Another advantage is that it can be easily extended to nonuniform grids with tensor product structure. This means that the point distribution in each one of the coordinate directions $\{x_1, x_2, \dots, x_d, t\}$ is arbitrary, and independent of all the remaining coordinates.

The original box scheme has an old history, and was first introduced by Wendroff [Wendroff, 1960]. Later it was developed further for parabolic problems by Keller, see for example [Keller, 1971] and [Keller, 1978]. The work presented in this book is based on the recent papers [Gustafsson and Khalighi, 2006a] and [Gustafsson and Khalighi, 2006b].

Chapter 9

Wave Propagation

In this chapter we shall discuss wave propagation applications, where the benefit of high order methods is clearly pronounced. Most applications in acoustics and electromagnetics are very demanding problems, with large computational domains both in space and time compared to the typical wavelength. From the analysis in Chapter 1, we know that the use of lower order methods should be limited to quite small problems. The equations are energy conserving, and since there is no damping, the errors accumulate over time.

Problems in acoustics and electromagnetics are governed by the wave equation in some form, and we begin by developing high order methods for that one. We shall use the first order system form and concentrate on the methods obtained by applying the Lax–Wendroff principle discussed in Section 6.1, since they turn out to be very effective when used on staggered grids. Already the basic second order Yee scheme has quite good properties, and we shall present some recent results obtained for that one, in particular for problems with discontinuous coefficients. For such problems, the theoretical analysis in Chapter 1 does not apply, since it was based on the assumption of smooth solutions. In this chapter we shall present a more detailed analysis and show that, even if the Yee scheme is good also in this case, the higher order methods are more effective.

In the final section we shall take a look at the initial–boundary value problem, and demonstrate how the boundary conditions can be embedded in the problem by interpreting them as an extreme case of discontinuous coefficients. In that way, very effective implementations of the algorithm can be done.

9.1 The Wave Equation

In this section we shall discuss the properties of the wave equation and develop a class of high order one step methods. We postpone the treatment of the boundary conditions until later, and assume that the equation is defined in the whole space

$-\infty < x < \infty$, possibly with periodic solutions. For ease of presentation, we begin by the one-dimensional case.

9.1.1 One Space Dimension

The wave equation in one space dimension written as a first order hyperbolic system is

$$\begin{bmatrix} p \\ u \end{bmatrix}_t = \begin{bmatrix} 0 & a(x) \\ b(x) & 0 \end{bmatrix} \begin{bmatrix} p \\ u \end{bmatrix}_x, \quad (9.1)$$

and we assume first 2π -periodic solutions in space. The notation p and u refers to the acoustics case described in the next section, where p is the pressure and u is the particle velocity. The coefficients a and b are defined by $a = -\rho c^2$, $b = -1/\rho$, where ρ is the density of the fluid and c is the speed of sound. In electromagnetics the equations are identical to the Maxwell equations with the following interpretation of the variables:

$$\begin{aligned} p &\rightarrow E \text{ electric field,} \\ u &\rightarrow H \text{ magnetic field,} \\ a &= -1/\epsilon, \\ b &= -1/\mu, \\ \mu &= \text{permeability,} \\ \epsilon &= \text{permittivity.} \end{aligned}$$

It should be mentioned that the system (9.1) represents only the principle part of the wave equation, and various physical applications may require more terms. For example, acoustic sources introduce forcing functions in the equations. In electromagnetics there may be charges and currents. However, the principal part determines the basic properties of the approximation, and by limiting the analysis to (9.1), we simplify the presentation.

Furthermore, we assume that the coefficients a and b depend on x but not on t , and this is the situation in most practical applications. They must have the same sign in order to retain the hyperbolic character, and the waves are propagating with the speed $\pm\sqrt{ab}$. For convenience we assume that a and b are positive. (In the examples above they are actually negative, but the simple transformation $x \rightarrow -x$ makes them positive.)

By differentiation, the first order system (9.1) can be reformulated as a second order scalar PDE. If there is no coupling between p and u through the boundary conditions, we can use either

$$p_{tt} = a(b p_x)_x \quad (9.2)$$

or

$$u_{tt} = b(a u_x)_x. \quad (9.3)$$

This form is often thought of as the original form of the wave equation, and is frequently used as the basis for numerical methods, at least for acoustic problems. Here, however, we shall use the first order system form, and later comment on the connection to the scalar form.

The usual L_2 -norm is defined by

$$\begin{aligned} ||p||^2 &= ||p(\cdot, t)||^2 = \int_0^{2\pi} |p(x, t)|^2 dx, \\ ||u||^2 &= ||u(\cdot, t)||^2 = \int_0^{2\pi} |u(x, t)|^2 dx. \end{aligned}$$

For the system (9.1) we use a weighted norm that is conserved with time:

$$\begin{aligned} \frac{d}{dt} \left(\left| \frac{1}{\sqrt{a}} p \right|^2 + \left| \frac{1}{\sqrt{b}} u \right|^2 \right) &= 2 \int_0^{2\pi} \left(\frac{1}{a} pp_t + \frac{1}{b} uu_t \right) dx \\ &= 2 \int_0^{2\pi} (pu_x + up_x) dx = 2 \int_0^{2\pi} pu_x dx - 2 \int_0^{2\pi} u_x p dx = 0. \end{aligned}$$

Note that even if $a(x) = b(x)$ such that the system is symmetric, the norm has to be weighted in order to provide conservation.

In order to compute numerical solutions, we use a grid that is staggered in both space and time as shown in Figure 9.1.

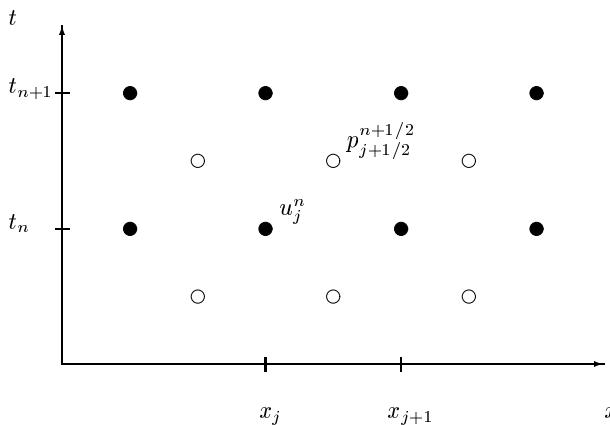


Fig. 9.1 The staggered grid

The Yee scheme

The standard second order scheme on this grid is due to Yee [Yee, 1966], and was developed for electromagnetics:

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + ka_{j+1/2}D_+u_j^n, \\ u_j^{n+1} &= u_j^n + kb_jD_-p_{j+1/2}^{n+1/2}. \end{aligned} \quad (9.4)$$

Even if p and u are stored at different levels, it is a one step scheme. If $p^{n-1/2}$ and u^n are known, the solution is advanced one step by first computing $p^{n+1/2}$ and then u^{n+1} .

We rewrite (9.4) as

$$\begin{aligned} \frac{p_{j+1/2}^{n+1/2} - p_{j+1/2}^{n-1/2}}{a_{j+1/2}} &= kD_+u_j^n, \\ \frac{u_j^{n+1} - u_j^n}{b_j} &= kD_-p_{j+1/2}^{n+1/2}, \end{aligned} \quad (9.5)$$

and introduce the discrete scalar product and norm by

$$(p, q)_h = \sum_j p_{j+1/2} q_{j+1/2} h, \quad \|p\|_h^2 = (p, p)_h$$

for grid functions stored at half points, and by

$$(u, v)_h = \sum_j u_j v_j h, \quad \|u\|_h^2 = (u, u)_h$$

for grid functions stored at integer points. By multiplying the first equation by $p_{j+1/2}^{n+1/2} + p_{j+1/2}^{n-1/2}$, the second equation by $u_j^{n+1} + u_j^n$, and then summing up, we obtain

$$\begin{aligned} \left\| \frac{1}{\sqrt{a}} p^{n+1/2} \right\|_h^2 &= \left\| \frac{1}{\sqrt{a}} p^{n-1/2} \right\|_h^2 + k(D_+u^n, p^{n+1/2} + p^{n-1/2})_h, \\ \left\| \frac{1}{\sqrt{b}} u^{n+1} \right\|_h^2 &= \left\| \frac{1}{\sqrt{b}} u^n \right\|_h^2 + k(D_-p^{n+1/2}, u^{n+1} + u^n)_h. \end{aligned}$$

By defining the modified discrete energy

$$E^n = \left\| \frac{1}{\sqrt{a}} p^{n-1/2} \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u^n \right\|_h^2 - k(D_-p^{n-1/2}, u^n)_h, \quad (9.6)$$

and using the relation

$$(D_+u^n, p^{n+1/2})_h = -(u^n, D_-p^{n+1/2})_h$$

(see (2.36)), we get exact energy conservation:

$$E^{n+1} = E^n.$$

It remains to show that E^n is a positive quantity for nonzero p and u . By using standard inequalities, one can prove that

$$k|(D_- p, u)_h| < \left\| \frac{1}{\sqrt{a}} p \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u \right\|_h^2$$

if

$$\frac{k}{h} \sqrt{c_1 c_2} < 1, \quad (9.7)$$

where

$$\begin{aligned} c_1 &= \frac{1}{2} \max_j (\sqrt{a_{j+1/2} b_j} + \sqrt{a_{j+1/2} b_{j+1}}), \\ c_2 &= \frac{1}{2} \max_j (\sqrt{a_{j+1/2} b_j} + \sqrt{a_{j-1/2} b_j}). \end{aligned} \quad (9.8)$$

Note that there is no smoothness requirement on the coefficients a and b except boundedness. This means that one can use the method for layered media, i.e., for problems where a and b are discontinuous. The scheme can be applied without any special procedure at the discontinuities, and the stability is still retained.

If

$$\frac{k}{h} \sqrt{c_1 c_2} \leq 1 - \delta, \quad \delta > 0,$$

then it follows that

$$k|(D_- p, u)_h| \leq (1 - \delta_1) \left(\left\| \frac{1}{\sqrt{a}} p \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u \right\|_h^2 \right), \quad \delta_1 > 0.$$

Hence,

$$\begin{aligned} \left\| \frac{1}{\sqrt{a}} p^{n-1/2} \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u^n \right\|_h^2 &\leq \frac{1}{\delta_1} E^n = \frac{1}{\delta_1} E^{n-1} = \dots = \frac{1}{\delta_1} E^0 \\ &\leq K \left(\left\| \frac{1}{\sqrt{a}} p^{-1/2} \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u^0 \right\|_h^2 \right), \end{aligned}$$

where $K = (2 - \delta_1)/\delta_1$. Note that K is independent of t . This means that the discrete norm, corresponding to the conserved norm for the differential equation, may not be exactly conserved, but it is always bounded independent of the length of the integration time interval.

A fourth order one step scheme

Next we derive a fourth order approximation by using the same principle as for the Lax–Wendroff scheme, which was discussed in Section 6.1. As we have seen in Section 4.2, we have

$$\begin{aligned} \frac{\partial u}{\partial x}(x_{j+1/2}) &= D_+ u_j - \frac{h^2}{24} D_+^2 D_- u_j + \mathcal{O}(h^4), \\ \frac{\partial p}{\partial x}(x_j) &= D_- p_{j+1/2} - \frac{h^2}{24} D_-^2 D_+ p_{j+1/2} + \mathcal{O}(h^4). \end{aligned}$$

By using these approximations in the right hand side of (9.4), we obtain fourth order accuracy in space. It remains to add correction terms such that we get fourth order

accuracy in time as well. The approximation of the time derivative of p is centered at time level $t_{n+1/2}$, and the truncation error in the Yee scheme is $k^2 p_{ttt}/24$. We use the differential equations to transform p_{ttt} into space derivatives. By differentiation of (9.1), (9.2) and (9.3) we get

$$p_{ttt} = a(bp_{xt})_x = a(b(au_x)_x)_x, \quad (9.9)$$

and

$$u_{ttt} = b(a(bp_x)_x)_x. \quad (9.10)$$

Since p_{ttt} and u_{ttt} both have a factor $k^2/24$ multiplying them in the truncation error, it is enough to approximate the expressions in the right hand sides of (9.9) and (9.10) by second order approximations. These are easily obtained by just centering the basic difference operators correctly. Therefore, the scheme

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + ka_{j+1/2}D_+u_j^n \\ &\quad + \frac{k}{24}(k^2a_{j+1/2}D_+b_jD_-a_{j+1/2}D_+ - h^2a_{j+1/2}D_+^2D_-)u_j^n, \\ u_j^{n+1} &= u_j^n + kb_jD_-p_{j+1/2}^{n+1/2} \\ &\quad + \frac{k}{24}(k^2b_jD_-a_{j+1/2}D_+b_jD_- - h^2b_jD_+D_-^2)p_{j+1/2}^{n+1/2} \end{aligned} \quad (9.11)$$

is fourth order accurate in both space and time, and the truncation error is $\mathcal{O}(h^4)$ if $k = \mathcal{O}(h)$. The computational stencil is shown in Figure 9.2.

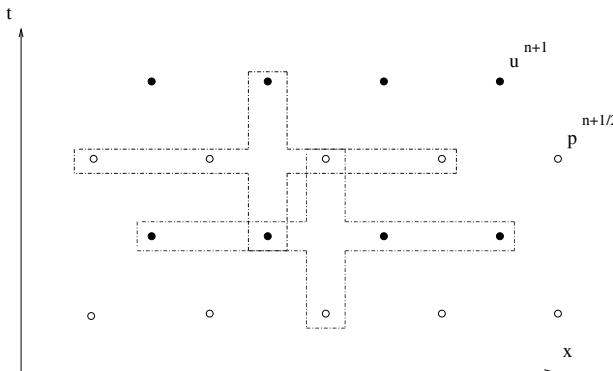


Fig. 9.2 Computational stencil for the fourth order one step scheme

For constant coefficients a, b , it can be shown that the stability condition is

$$\lambda c \leq 1,$$

where $c = \sqrt{ab}$ and $\lambda = k/h$. This condition is equivalent to the stability condition for the second order scheme. For variable coefficients $a(x), b(x)$, one can show by using the energy method, that a sufficient condition for stability is

$$\begin{aligned} \lambda^2 \max_j (\max(a_{j+1/2}b_j, a_{j-1/2}b_j)) &< 2, \\ \lambda \sqrt{c_1 c_2} &< \sqrt{\frac{12}{13+\beta}}, \end{aligned} \quad (9.12)$$

where c_1, c_2 are defined in (9.8), and

$$\beta = \max_j \max(\sqrt{a_{j+3/2}/a_{j+1/2}}, \sqrt{a_{j-1/2}/a_{j+1/2}}).$$

It was shown above how the scalar form of the wave equation was obtained by differentiating the first order system. The analogous procedure can be applied to the discrete system both for the second and fourth order case. Instead of differentiating with respect to t and x , we apply difference operators in the t and x direction in order to eliminate one of the variables. When eliminating p from the Yee scheme, we obtain

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = k^2 b_j D_- a_{j+1/2} D_+ u_j^n,$$

which is the standard five point approximation of (9.3).

The same method can be applied to the 4th order approximation (9.11), and we obtain

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = k^2 b_j D_- Q_1 a_{j+1/2} D_+ Q_2 u_j^n,$$

where

$$Q_1 = I + \frac{1}{24}(k^2 a_{j+1/2} D_+ b_j D_- - h^2 D_+ D_-),$$

$$Q_2 = I + \frac{1}{24}(k^2 b_j D_- a_{j+1/2} D_+ - h^2 D_+ D_-).$$

This is a 9-point computational stencil, and it is shown in Figure 9.3 together with the 5-point stencil. This scheme could be used instead of (9.11), and the stability and accuracy properties are the same. This is a two step scheme, which is of course necessary for an approximation of a PDE with second order derivatives in time. Therefore two levels of initial data are required, and it can be obtained from the initial conditions for the PDE, which must be given not only for u but also for $\partial u / \partial t$.

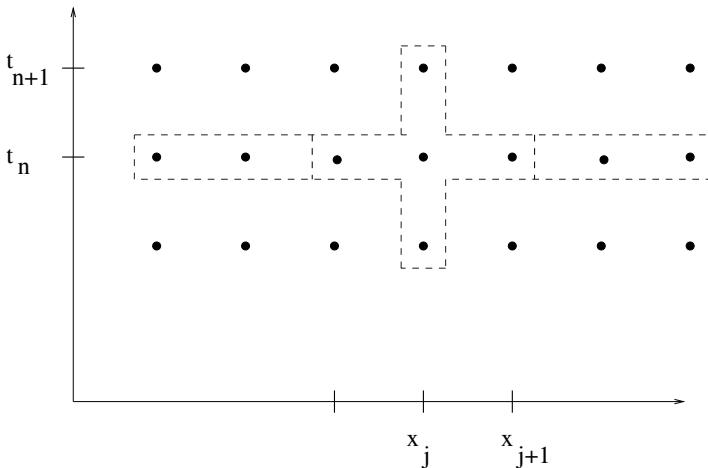


Fig. 9.3 Computational second and fourth order stencil for the scalar wave equation

Comparison of accuracy

In Chapter 6, we compared the accuracy between different approximations by computing the symbol $\hat{Q}(\xi)$. In the present case, $\hat{Q}(\xi)$ is a 2×2 matrix, and the eigenvalues $z_1(\xi)$ and $z_2(\xi)$ represent the two waves going in different directions. The magnitude is one, and for $c = 1$ we write

$$z_1(\xi) = e^{i\lambda\phi(\xi)},$$

where $\phi(\xi)$ is real. The exact solution of the differential equation corresponds to $\phi(\xi) = \xi = \omega h$, and the question is how well ξ is approximated by $\phi(\xi)$. Figure 9.4 shows $\phi(\xi) = \phi_{2,2}(\xi)$ for the Yee scheme and $\phi_{4,4}(\xi)$ for the new 4th order scheme for the two cases $\lambda = 0.5$ and $\lambda = 0.8$.

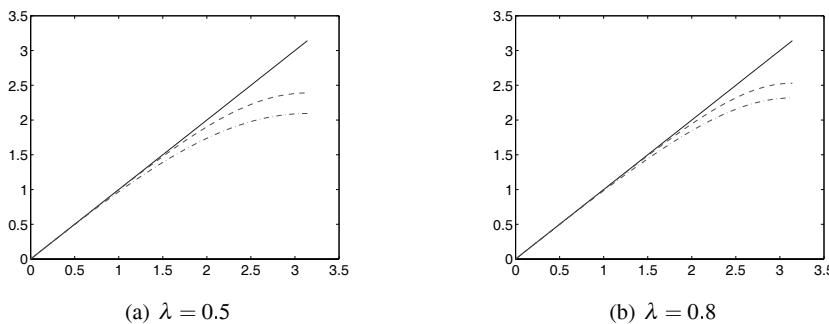


Fig. 9.4 $\phi_{p,q}(\xi)$, $(p, q) = (2, 2) (-\cdot)$, $(4, 4) (--)$

Assuming that we have 4 points per wavelength for the highest wave number of interest, the maximum phase error is $\max_{0 \leq \xi \leq \pi/2} |\xi - \phi(\xi)| = |\pi/2 - \phi(\pi/2)|$ for all λ with $0 < \lambda \leq 1$. Figure 9.5 shows $\phi_{p,q}(\pi/2)/(\pi/2)$ as a function of λ .

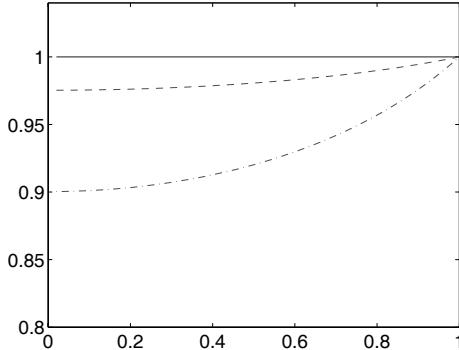


Fig. 9.5 $\phi_{p,q}(\pi/2)/(\pi/2)$, $(p, q) = (2, 2)$ (\cdots), $(4, 4)$ ($--$), $0 < \lambda \leq 1$

The construction of one step schemes can be generalized to any order of accuracy. The truncation error in space for the fourth order scheme is

$$3ah^4 u_{xxxx}/640 \text{ for the } p\text{-equation}$$

and

$$3bh^4 p_{xxxx}/640 \text{ for the } u\text{-equation.}$$

The truncation error in time is

$$\begin{aligned} \frac{k^4}{1920} p_{tttt} &= \frac{k^4}{1920} a \left(b \left(a(b(a u_x)_x)_x \right)_x \right)_x, \\ \frac{k^4}{1920} u_{tttt} &= \frac{k^4}{1920} b \left(a \left(b(a(b p_x)_x)_x \right)_x \right)_x. \end{aligned}$$

When substituting these expressions by standard second order 6-point formulas, a sixth order scheme is obtained. This procedure can be continued to any order just by keeping track of the truncation error for the latest approximation.

9.1.2 Two Space Dimensions

Here we consider the wave equation in 2-D

$$\begin{bmatrix} p \\ u \\ v \end{bmatrix}_t = \begin{bmatrix} 0 & a(x,y) & 0 \\ b(x,y) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p \\ u \\ v \end{bmatrix}_x + \begin{bmatrix} 0 & 0 & a(x,y) \\ 0 & 0 & 0 \\ b(x,y) & 0 & 0 \end{bmatrix} \begin{bmatrix} p \\ u \\ v \end{bmatrix}_y + \begin{bmatrix} F \\ G \\ H \end{bmatrix}.$$

The variables u and v are the two velocity components in the acoustic case, and the two magnetic field components in the electromagnetic case. The functions a and b are defined exactly as for the 1-D case, and for convenience they are assumed to be positive also here. Forcing functions have been introduced here, just to indicate how they enter in the truncation errors.

The scalar product and norm are defined by

$$(f, g) = \int_0^{2\pi} \int_0^{2\pi} f(x, y)g(x, y) dx dy, \quad \|f\|^2 = (f, f).$$

The conservation property

$$\frac{d}{dt} \left(\left\| \frac{1}{\sqrt{a}} p(\cdot, \cdot, t) \right\|^2 + \left\| \frac{1}{\sqrt{b}} u(\cdot, \cdot, t) \right\|^2 + \left\| \frac{1}{\sqrt{b}} v(\cdot, \cdot, t) \right\|^2 \right) = 0$$

is then obtained by integration by parts:

$$\begin{aligned} \left(\frac{1}{a} p, au_x \right) &= (p, u_x) = -(u, p_x) = -\left(\frac{1}{b} u, bp_x \right), \\ \left(\frac{1}{a} p, av_y \right) &= (p, v_y) = -(v, p_y) = -\left(\frac{1}{b} v, bp_y \right). \end{aligned}$$

The second and fourth order schemes

The staggered grid is shown in Figure 9.6, and we denote the step size in the x - and y -direction by h_1 and h_2 respectively.

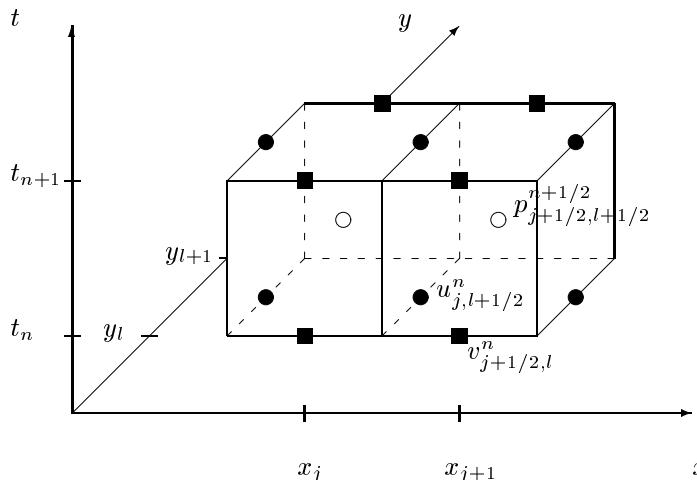


Fig. 9.6 Staggered grid for p , u and v

The variables u and v are stored at the same time level, while p is stored in between. All three variables are stored at different points in space.

The difference scheme is obtained as a direct generalization of the one-dimensional form. With the forward and backwards difference operators in space denoted by D_{+x} , D_{+y} , D_{-x} , D_{-y} respectively, the standard second order Yee scheme is

$$\begin{aligned} p_{j+1/2,l+1/2}^{n+1/2} &= p_{j+1/2,l+1/2}^{n-1/2} + ka_{j+1/2,l+1/2}(D_{+x}u_{j,l+1/2}^n + D_{+y}v_{j+1/2,l}^n) \\ &\quad + kF_{j+1/2,l+1/2}^n, \\ u_{j,l+1/2}^{n+1} &= u_{j,l+1/2}^n + kb_{j,l+1/2}D_{-x}p_{j+1/2,l+1/2}^{n+1/2} + kG_{j,l+1/2}^{n+1/2}, \\ v_{j+1/2,l}^{n+1} &= v_{j+1/2,l}^n + kb_{j+1/2,l}D_{-y}p_{j+1/2,l+1/2}^{n+1/2} + kH_{j+1/2,l}^{n+1/2}. \end{aligned}$$

The truncation errors in space are eliminated as in the one-dimensional case by approximating the third order derivatives in each direction by four point formulas. For the truncation error in time, we use the relations

$$\begin{aligned} p_{ttt} &= a \left((b(au_x)_x)_x + (b(av_y)_x)_x + (b(au_x)_y)_y + (b(av_y)_y)_y \right) \\ &\quad + a((bF_x)_x + G_{xt} + (bF_y)_y + H_{yt}) + F_{tt}, \\ u_{ttt} &= b \left((a(bp_x)_x)_x + (a(bp_y)_y)_x \right) \\ &\quad + b((aG_x)_x + (aH_y)_x + F_{xt}) + G_{tt}, \\ v_{ttt} &= b \left((a(bp_x)_y)_y + (a(bp_y)_y)_y \right) \\ &\quad + b((aG_x)_y + (aH_y)_y + F_{yt}) + H_{tt}, \end{aligned}$$

and centered second order accurate formulas for approximation of the space derivatives. If the derivatives of the forcing functions F, G, H are not available, then they are approximated by standard 2nd order formulas. To keep the presentation simple, we omit those terms from now on.

The complete fourth order approximation is

$$\begin{aligned} p_{j+1/2,l+1/2}^{n+1/2} &= p_{j+1/2,l+1/2}^{n-1/2} + ka_{j+1/2,l+1/2}(D_{+x}u_{j,l+1/2}^n + D_{+y}v_{j+1/2,l}^n) \\ &\quad + \frac{1}{24}(k^2 D_{+x}b_{j,l+1/2}D_{-x}a_{j+1/2,l+1/2}D_{+x}u_{j,l+1/2}^n \\ &\quad + k^2 D_{+y}b_{j+1/2,l}D_{-y}a_{j+1/2,l+1/2}D_{+x}u_{j,l+1/2}^n \\ &\quad + k^2 D_{+x}b_{j,l+1/2}D_{-x}a_{j+1/2,l+1/2}D_{+y}v_{j+1/2,l}^n \\ &\quad + k^2 D_{+y}b_{j+1/2,l}D_{-y}a_{j+1/2,l+1/2}D_{+y}v_{j+1/2,l}^n \\ &\quad - h_1^2 D_{+x}^2 D_{-x}u_{j,l+1/2}^n - h_2^2 D_{+y}^2 D_{-y}v_{j+1/2,l}^n), \end{aligned}$$

$$\begin{aligned} u_{j,l+1/2}^{n+1} &= u_{j,l+1/2}^n + kb_{j,l+1/2}(D_{-x} + \frac{1}{24}(k^2 D_{-x}a_{j+1/2,l+1/2}D_{+x}b_{j,l+1/2}D_{-x} \\ &\quad + k^2 D_{-x}a_{j+1/2,l+1/2}D_{+y}b_{j+1/2,l}D_{-y} - h_1^2 D_{+x}^2 D_{-x}^2))p_{j+1/2,l+1/2}^{n+1/2}, \end{aligned}$$

$$\begin{aligned}
v_{j+1/2,l}^{n+1} &= v_{j+1/2,l}^n + kb_{j+1/2,l}(D_{-y} + \frac{1}{24}(k^2 D_{-y} a_{j+1/2,l+1/2} D_{+x} b_{j,l+1/2} D_{-x} \\
&\quad + k^2 D_{-y} a_{j+1/2,l+1/2} D_{+y} b_{j+1/2,l} D_{-y} - h_2^2 D_{+y} D_{-y}^2)) p_{j+1/2,l+1/2}^{n+1/2}.
\end{aligned} \tag{9.13}$$

For the stability analysis we first assume constant coefficients a, b , and use Fourier analysis. We define

$$\begin{aligned}
p_{j+1/2,l+1/2}^{n-1/2} &= \hat{p}^n e^{i\omega_1(j+1/2)h_1} e^{i\omega_2(l+1/2)h_2}, \\
u_{j,l+1/2}^n &= \hat{u}^n e^{i\omega_1 j h_1} e^{i\omega_2(l+1/2)h_2}, \\
v_{j+1/2,l}^n &= \hat{v}^n e^{i\omega_1(j+1/2)h_1} e^{i\omega_2 l h_2},
\end{aligned}$$

and get for the homogeneous case

$$\begin{bmatrix} 1 & 0 & 0 \\ -b\hat{Q}_{q1} & 1 & 0 \\ -b\hat{Q}_{q2} & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{p}^{n+1} \\ \hat{u}^{n+1} \\ \hat{v}^{n+1} \end{bmatrix} = \begin{bmatrix} 1 & a\hat{Q}_{q1} & a\hat{Q}_{q2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{p}^n \\ \hat{u}^n \\ \hat{v}^n \end{bmatrix}.$$

With the notation

$$\begin{aligned}
s_\nu &= \sin \frac{\xi_\nu}{2} \\
\lambda_\nu &= k/h_\nu
\end{aligned}, \quad \nu = 1, 2,$$

the Fourier transformed difference operators are

$$\begin{aligned}
\hat{Q}_{21} &= 2\lambda_1 i s_1, \\
\hat{Q}_{22} &= 2\lambda_2 i s_2, \\
\hat{Q}_{41} &= \lambda_1 i s_1 \left(2 + \frac{1}{3} s_1^2 - \frac{ab}{3} (\lambda_1^2 s_1^2 + \lambda_2^2 s_2^2) \right), \\
\hat{Q}_{42} &= \lambda_2 i s_2 \left(2 + \frac{1}{3} s_2^2 - \frac{ab}{3} (\lambda_1^2 s_1^2 + \lambda_2^2 s_2^2) \right).
\end{aligned}$$

The explicit form of one time step in Fourier space is

$$\begin{bmatrix} \hat{p}^{n+1} \\ \hat{u}^{n+1} \\ \hat{v}^{n+1} \end{bmatrix} = \begin{bmatrix} 1 & a\hat{Q}_{q1} & a\hat{Q}_{q2} \\ b\hat{Q}_{q1} & 1 + ab\hat{Q}_{q1}^2 & ab\hat{Q}_{q1}\hat{Q}_{q2} \\ b\hat{Q}_{q2} & ab\hat{Q}_{q1}\hat{Q}_{q2} & 1 + ab\hat{Q}_{q2}^2 \end{bmatrix} \begin{bmatrix} \hat{p}^n \\ \hat{u}^n \\ \hat{v}^n \end{bmatrix}, \quad q = 2, 4.$$

One eigenvalue of the amplification matrix is $z = 1$, and the other two are given by

$$z^2 - (2 + ab\hat{Q}_{q1}^2 + ab\hat{Q}_{q2}^2)z + 1 = 0,$$

leading to the condition

$$ab(|\hat{Q}_{q1}|^2 + |\hat{Q}_{q2}|^2) < 4, \quad q = 2, 4.$$

A straightforward calculation shows that this condition is satisfied under the stability condition

$$ab(\lambda_1^2 + \lambda_2^2) < 1, \quad \lambda_1 = \frac{k}{h_1}, \quad \lambda_2 = \frac{k}{h_2}$$

for both the second and fourth order case.

For variable coefficients we use the energy method. The derivation of an energy estimate is straightforward. However, we include it here as an illustration of the energy method for a nontrivial problem (as an exercise to Section 2.3.3). The approximation is written in the form

$$(p_{j+1/2,l+1/2}^{n+1/2} - p_{j+1/2,l+1/2}^{n-1/2})/a_{j+1/2,l+1/2} = R_1(u^n, v^n), \quad (9.14)$$

$$(u_{j,l+1/2}^{n+1} - u_{j,l+1/2}^n)/b_{j,l+1/2} = R_2(p^{n+1/2}), \quad (9.15)$$

$$(v_{j+1/2,l}^{n+1} - v_{j+1/2,l}^n)/b_{j+1/2,l} = R_3(p^{n+1/2}), \quad (9.16)$$

where $R_1(u^n, v^n)$, $R_2(p^{n+1/2})$, $R_3(p^{n+1/2})$ are defined by the right hand sides of (9.13). The first equation is multiplied by $(p_{j+1/2,l+1/2}^{n+1/2} + p_{j+1/2,l+1/2}^{n-1/2})$ and summed up over j and l . An analogous operation on (9.15) and (9.16) gives as a result

$$\begin{aligned} & \left\| \frac{1}{\sqrt{a}} p^{n+1/2} \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} u^{n+1} \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} v^{n+1} \right\|_h^2 = \left\| \frac{1}{\sqrt{a}} p^{n-1/2} \right\|_h^2 \\ & + \left\| \frac{1}{\sqrt{b}} u^n \right\|_h^2 + \left\| \frac{1}{\sqrt{b}} v^n \right\|_h^2 - \left((R_1(u^n, v^n), p^{n+1/2} - p^{n-1/2})_h \right. \\ & \left. + (R_2(p^{n+1/2}), u^{n+1} - u^n)_h + (R_3(p^{n+1/2}), v^{n+1} - v^n)_h \right). \end{aligned}$$

Here we have used the notation $(f, g)_h$, $\|f\|_h = \sqrt{(f, f)_h}$ for the usual l_2 -scalar product and norm with the summation defined over the staggered grid-points. The weighting factors in the norms are taken at the same points as the corresponding dependent variable, such that $a_{j+1/2,l+1/2}$ is used in the first term, $b_{j,l+1/2}$ in the second one, and $b_{j+1/2,l}$ in the third one.

By using the simple summation by parts formulas for periodic grid-functions

$$(D_{+x}f, g)_h = -(f, D_{-x}g)_h,$$

$$(D_{+y}f, g)_h = -(f, D_{-y}g)_h,$$

we find that the modified energy

$$\begin{aligned}
E(p^{n+1/2}, u^{n+1}, v^{n+1}) = & \\
& \|\frac{1}{\sqrt{a}} p^{n+1/2}\|_h^2 + \|\frac{1}{\sqrt{b}} u^{n+1}\|_h^2 + \|\frac{1}{\sqrt{b}} v^{n+1}\|_h^2 \\
& - k \left((D_{-x} + \frac{k^2}{24} D_{-x} a D_{+x} b D_{-x} + \frac{k^2}{24} D_{-x} a D_{+y} b D_{-y} \right. \\
& \left. - \frac{h_1^2}{24} D_{+x} D_{-x}^2) p^{n+1/2}, u^{n+1} \right)_h - k \left((D_{-y} + \frac{k^2}{24} D_{-y} a D_{+y} b D_{-y} \right. \\
& \left. + \frac{k^2}{24} D_{-y} a D_{+x} b D_{-x} - \frac{h_2^2}{24} D_{+y} D_{-y}^2) p^{n+1/2}, v^{n+1} \right)_h
\end{aligned} \tag{9.17}$$

is conserved, i.e.,

$$E(p^{n+1/2}, u^{n+1}, v^{n+1}) = E(p^{n-1/2}, u^n, v^n).$$

It remains to be shown that the energy is nonnegative. We define the operators

$$\begin{aligned}
Q_1 &= I + \frac{k^2}{24} a D_{+x} b D_{-x} + \frac{k^2}{24} a D_{+y} b D_{-y} - \frac{h_1^2}{24} D_{+x} D_{-x}, \\
Q_2 &= I + \frac{k^2}{24} a D_{+y} b D_{-y} + \frac{k^2}{24} a D_{+x} b D_{-x} - \frac{h_2^2}{24} D_{+y} D_{-y},
\end{aligned}$$

and note that we need to estimate the terms $k|(D_{-x} Q_1 p, u)_h|$ and $k|(D_{-y} Q_2 p, v)_h|$.

With the notation

$$r = p/\sqrt{a}, \quad s = u/\sqrt{b}, \quad q = v/\sqrt{b},$$

the expressions to estimate are $k|(D_{-x} Q_1 \sqrt{a} r, \sqrt{b} s)_h|$ and $k|(D_{-y} Q_2 \sqrt{a} r, \sqrt{b} q)_h|$.

We will make use of the standard inequality

$$\alpha x y \leq \frac{1}{2} (\beta^2 x^2 + \gamma^2 y^2),$$

where the constants α, β, γ are positive with $\beta\gamma = \alpha$.

The expression $k D_{-x} Q_1 \sqrt{a_{j+1/2, l+1/2}} r_{j+1/2, l+1/2}$ can be written as

$$\lambda_1 \left(\sqrt{a_{j+1/2, l+1/2}} f(j+1/2, l) - \sqrt{a_{j-1/2, l+1/2}} f(j-1/2, l) \right),$$

where

$$f(\nu, l) = d_1 r_{\nu, l+3/2} + d_2 r_{\nu+1, l+1/2} + d_3 r_{\nu, l+1/2} + d_4 r_{\nu-1, l+1/2} + d_5 r_{\nu, l-1/2},$$

with $d_i = d_i(\nu, l)$ defined as

$$\begin{aligned}
d_1 &= \frac{1}{24} \sqrt{\frac{a_{\nu,l+3/2}}{a_{\nu,l+1/2}}} \lambda_2^2 b_{\nu,l+1} a_{\nu,l+1/2}, \\
d_2 &= \frac{1}{24} \sqrt{\frac{a_{\nu+1,l+1/2}}{a_{\nu,l+1/2}}} (\lambda_1^2 b_{\nu+1/2,l+1/2} a_{\nu,l+1/2} - 1), \\
d_3 &= 1 + \frac{1}{24} (2 - (\lambda_1^2 (b_{\nu+1/2,l+1/2} + b_{\nu-1/2,l+1/2}) + \lambda_2^2 (b_{\nu,l+1} + b_{\nu,l})) a_{\nu,l+1/2}), \\
d_4 &= \frac{1}{24} \sqrt{\frac{a_{\nu-1,l+1/2}}{a_{\nu,l+1/2}}} (\lambda_1^2 b_{\nu-1/2,l+1/2} a_{\nu,l+1/2} - 1), \\
d_5 &= \frac{1}{24} \sqrt{\frac{a_{\nu,l-1/2}}{a_{\nu,l+1/2}}} \lambda_2^2 b_{\nu,l} a_{\nu,l+1/2}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&k|(D_{-x}Q_1 \sqrt{ar}, \sqrt{bs})_h| \\
&= \sum_{j,l} \lambda_1 (\sqrt{a_{j+1/2,l+1/2}} f(j+1/2, l+1/2) \\
&\quad - \sqrt{a_{j-1/2,l+1/2}} f(j-1/2, l+1/2)) \sqrt{b_{j,l+1/2}} s_{j,l+1/2} h_1 h_2 \\
&\leq \sum_{j,l} \lambda_1 (\sqrt{a_{j+1/2,l+1/2}} |f(j+1/2, l+1/2)| \\
&\quad + \sqrt{a_{j-1/2,l+1/2}} |f(j-1/2, l+1/2)|) \sqrt{b_{j,l+1/2}} |s_{j,l+1/2}| h_1 h_2 \\
&\leq \frac{1}{2} \sum_{j,l} (\lambda_1^2 a_{j+1/2,l+1/2} b_{j,l+1/2} f(j+1/2, l+1/2)^2 + s_{j,l+1/2}^2 + \\
&\quad + \lambda_1^2 a_{j-1/2,l+1/2} b_{j,l+1/2} f(j-1/2, l+1/2)^2 + s_{j,l+1/2}^2) h_1 h_2 \\
&= \frac{1}{2} \sum_{j,l} (\lambda_1^2 a_{j+1/2,l+1/2} (b_{j,l+1/2} + b_{j+1,l+1/2}) f(j+1/2, l+1/2)^2 \\
&\quad + 2s_{j,l+1/2}^2) h_1 h_2 \\
&\leq \lambda_1^2 \max_{j,l} (a_{j+1/2,l+1/2} (b_{j,l+1/2} + b_{j+1,l+1/2}) / 2) \|f\|_h^2 + \|s\|_h^2.
\end{aligned}$$

An analogous estimate is done for the term containing Q_2 . It will contain an expression $g(j, \nu)$ which is a sum of five terms in analogy with $f(\nu, l)$. We get

$$\begin{aligned}
&k|(D_{-x}Q_1 \sqrt{ar}, \sqrt{bs})_h| \leq \lambda_1^2 c_1^2 \|f\|_h^2 + \|s\|_h^2, \\
&k|(D_{-y}Q_2 \sqrt{ar}, \sqrt{bq})_h| \leq \lambda_2^2 c_2^2 \|g\|_h^2 + \|q\|_h^2,
\end{aligned}$$

where

$$\begin{aligned} c_1^2 &= \frac{1}{2} \max_{j,l} (a_{j+1/2,l+1/2}(b_{j,l+1/2} + b_{j+1,l+1/2})) , \\ c_2^2 &= \frac{1}{2} \max_{j,l} (a_{j+1/2,l+1/2}(b_{j+1/2,l} + b_{j+1/2,l+1})) \end{aligned} \quad (9.18)$$

To estimate $\|f\|_h$ (and similarly for g) we make the following assumptions

$$\begin{aligned} \max_{j,l} (\max(\lambda_1^2 a_{j+1/2,l+1/2} b_{j+1/2\pm 1/2,l+1/2})) &< 1 , \\ \max_{j,l} (\max(\lambda_2^2 a_{j+1/2,l+1/2} b_{j+1/2,l+1/2\pm 1/2})) &< 1 . \end{aligned}$$

Let

$$\beta = \max_{j,l} \left(\max \left(\sqrt{a_{j+1/2\pm 1/2,l+1/2}/a_{j+1/2,l+1/2}}, \sqrt{a_{j+1/2,l+1/2\pm 1/2}/a_{j+1/2,l+1/2}}} \right) \right) . \quad (9.19)$$

Then we have

$$\begin{aligned} \|f\|_h^2 &\leq \sum_{j,l} \left(|r_{j+1/2,l+1/2}| \right. \\ &\quad \left. + \frac{1}{24} (2|r_{j+1/2,l+1/2}| + \beta(|r_{j+3/2,l+1/2}| + |r_{j+1/2,l+3/2}| \right. \\ &\quad \left. + |r_{j+1/2,l-1/2}| + |r_{j-1/2,l+1/2}|)) \right)^2 h_1 h_2 . \end{aligned}$$

Since we are dealing with periodic grid functions, the sum can be computed using Fourier transforms and Parseval's relation. We get

$$\begin{aligned} \|f\|_h &\leq \frac{13+2\beta}{12} \|r\|_h , \\ \|g\|_h &\leq \frac{13+2\beta}{12} \|r\|_h . \end{aligned}$$

For the energy in (9.17) to be nonnegative we need

$$\lambda_1^2 c_1^2 \|f\|_h^2 + \lambda_2^2 c_2^2 \|g\|_h^2 < \|r\|_h^2 ,$$

which gives the final sufficient stability condition

$$\lambda_1^2 c_1^2 + \lambda_2^2 c_2^2 < \left(\frac{12}{13+2\beta} \right)^2 , \quad (9.20)$$

where c_1 and c_2 are defined in (9.18) and β in (9.19).

9.2 Discontinuous Coefficients

In this section we shall consider a case with discontinuous coefficients, and for convenience we assume that there is only one discontinuity, which is located at

$x = \bar{x}$. In 1-D the variables p and u are continuous, but the x -derivatives are not. The jump conditions follows immediately from the PDE:

$$\begin{aligned}[au_x] &= 0, \\ [bp_x] &= 0,\end{aligned}\tag{9.21}$$

where

$$[f] = \lim_{x \rightarrow \bar{x}+} f(x) - \lim_{x \rightarrow \bar{x}-} f(x).$$

In order to avoid any bad effects on the accuracy, it is very common to introduce some special procedure near the discontinuity, where the jump conditions (9.21) are taken into account. This is a complication in the sense that the implementation of the algorithm becomes more intricate when it comes to parallelization. We shall first analyze the behavior of the one step schemes that were derived above with no special procedure applied at the discontinuity. Then we shall discuss a method for the improvement of the accuracy by a technique that is such that the parallelization difficulty is avoided.

9.2.1 The Original One Step Scheme

In this section we shall discuss the behavior of the schemes derived above when they are applied across the discontinuity without any special procedure applied there, and with no modification of the coefficients.

One space dimension

For convenience we assume that the coefficients are piecewise constant with

$$\begin{aligned}a(x) &= a^L, \quad b(x) = b^L \quad \text{if } x \leq \bar{x}, \\ a(x) &= a^R, \quad b(x) = b^R \quad \text{if } \bar{x} < x.\end{aligned}$$

The coefficients $a(x)$ and $b(x)$ are well defined everywhere, and therefore the methods developed in Section 9.1.1 can be applied without any special procedure near $x = \bar{x}$. In order to estimate the error, the explicit form of the exact and numerical solutions are calculated. By periodicity in space, it follows that the solutions are periodic in time as well. Therefore we make the ansatz

$$\begin{aligned}p(x, t) &= \hat{p}(x, \omega) e^{i\omega t}, \\ u(x, t) &= \hat{u}(x, \omega) e^{i\omega t},\end{aligned}$$

where ω is the frequency. In this way, we obtain an ODE system in space for the coefficients \hat{p} and \hat{u} . On each side of the interface it is a system with constant coefficients, and the form of the solutions are well known. By matching the solutions at the interface, we obtain the explicit form of the solution everywhere.

For the numerical solution, the same procedure can be applied. The difference is that we now get a system of difference equations in space, but the general form of the solutions are also here well known, see Appendix A. The solutions contain a number of linearly independent components $\sigma_\nu(\hat{p}_{j+1/2}^{(\nu)} \hat{u}_j^{(\nu)})^T$ on each side of the interface. The coefficients σ_ν are obtained by the matching conditions near the interface, which lead to an algebraic system. This system becomes larger for higher order methods, and solving it is a little complicated. The whole procedure has been carried out for the 2nd and 4th order methods, and we present the results here without the details of the derivation.

We emphasize that our interest is in the solution of the problem over large time intervals compared to the typical wavelength. In other words, we consider the case where many waves are passing the interface. Therefore we can as well consider high frequency solutions on a fixed time interval of order one.

Consider the problem on the interval $0 < t_n \leq T$, where T is fixed, and denote the maximum error in space and time by $w^{(p)}$, where p is the formal order of accuracy for smooth solutions. The procedure described above then leads to the error estimate

$$w^{(p)} \leq C_p(\omega^{p+1} h^p + \omega h),$$

where the constant C_p depends on p , but not on ω or h . The first part is the usual error that is present even for smooth solutions, while the second part arises from the crude first order treatment of the internal boundary. One should note the important difference between the two parts. Even if the second part is first order in h , it is also a function of $\xi = \omega h$, which can never be larger than π in magnitude. On the contrary, the first part has the form $\omega \xi^p$, which means that for larger number of waves hitting the interface, the error increases. In fact, if we define the error level ε , and the number of grid points per wavelength $N_\omega = N/\omega$, we have

$$\varepsilon = \frac{\gamma_p \omega}{N_\omega^p} + \frac{\gamma_1}{N_\omega},$$

where the constant γ_p depends on p but not on ω . For large ω , the first term will dominate, but the number of grid points can be kept down by choosing higher order accuracy p .

For the numerical experiments, we choose an example from acoustics with

$$\begin{aligned} a(x) &= -c^2(x)\rho(x), \\ b(x) &= -1/\rho(x), \end{aligned}$$

where c is the speed of sound, and ρ is the density. The piecewise constant coefficients are

$$\begin{aligned} \rho &= \rho_L, \quad c = c_L \quad \text{if } x \leq \bar{x}, \\ \rho &= \rho_R, \quad c = c_R \quad \text{if } x > \bar{x}. \end{aligned}$$

We choose the special solution

$x \leq \bar{x}$:

$$p(x, t) = \sin \left(\omega \left(t - \frac{x - \bar{x}}{c_L} \right) \right) - \frac{\rho_{LC_L} - \rho_{RC_R}}{\rho_{LC_L} + \rho_{RC_R}} \sin \left(\omega \left(t + \frac{x - \bar{x}}{c_L} \right) \right),$$

$$u(x, t) = \frac{1}{\rho_{LC_L}} \left(\sin \left(\omega \left(t - \frac{x - \bar{x}}{c_L} \right) \right) + \frac{\rho_{LC_L} - \rho_{RC_R}}{\rho_{LC_L} + \rho_{RC_R}} \sin \left(\omega \left(t + \frac{x - \bar{x}}{c_L} \right) \right) \right),$$

$\bar{x} < x$:

$$p(x, t) = \frac{2\rho_{RC_R}}{\rho_{LC_L} + \rho_{RC_R}} \sin \left(\omega \left(t - \frac{x - \bar{x}}{c_R} \right) \right),$$

$$u(x, t) = \frac{2}{\rho_{LC_L} + \rho_{RC_R}} \sin \left(\omega \left(t - \frac{x - \bar{x}}{c_R} \right) \right), \quad (9.22)$$

i.e., the right-going waves pass through the interface with modified amplitude, while there are no left-going waves on the right side of the interface. For problems in several space dimensions, one may want to use uniform grids, where the grid lines are not aligned with the internal boundaries. Figure 9.7 shows a typical case in 2-D, where the square and circle points represent the velocity components for a problem in acoustics. (The dashed lines represent an imaginary staircase boundary to the right of the true boundary.)

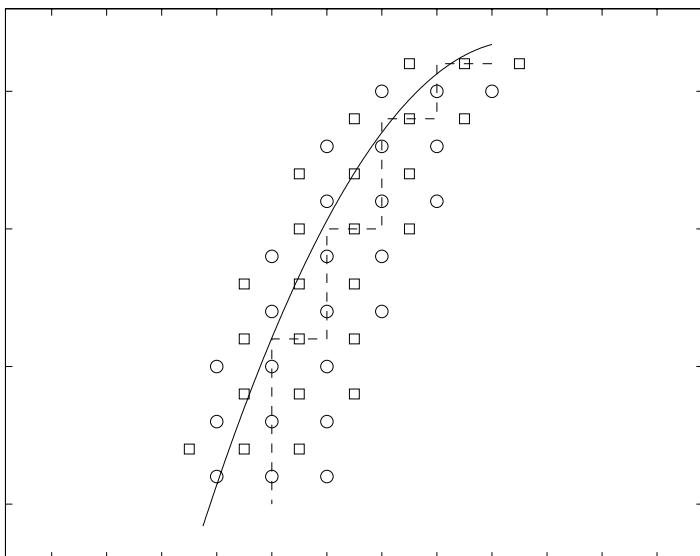


Fig. 9.7 2-D boundary, u and v points

In order to simulate this situation in 1-D, we place the discontinuity such that it does not necessarily fall on a grid point, see Figure 9.8.

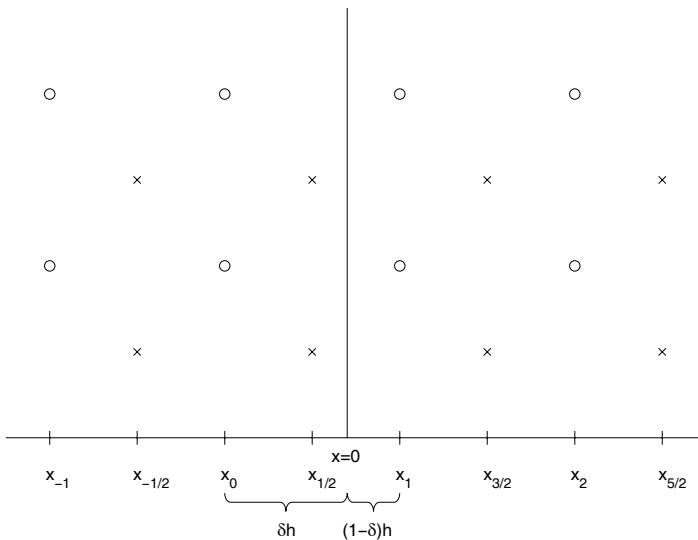


Fig. 9.8 Staggered grid and discontinuity at $x = 0$

We want to simulate the case where many waves are hitting the interface $x = \bar{x}$. Instead of running the scheme over long time intervals, we stay with a shorter interval $0 \leq t < T$ and choose large ω -values. We use $\omega = \omega_0 c_L$, where ω_0 is an integer. In the figure captions we are giving the parameters ω_0 and the number of points per shortest wavelength N_ω . The parameters are

$$\begin{aligned} \rho_L &= 0.55556, & \rho_R &= 1.00000, \\ c_L &= 0.87879, & c_R &= 1.00000, \end{aligned}$$

corresponding to sediment and water respectively. The Courant number is $\lambda = k/h = 0.8$.

In the following, we denote the two schemes by Q_2 and Q_4 . We first present a relatively short run with $T = 0.12\pi$. Figure 9.9 shows the error

$$\epsilon^n = \max_j |p(x_{j+1/2}, t_{n+1/2}) - p_{j+1/2}^{n+1/2}|$$

as a function of time for both schemes with $N_\omega = 7.5$ and $N_\omega = 15$. The second order part of the error dominates for Q_2 , while the first order part of the error dominates for Q_4 . The error is first order already at $t = 0$ due to the fact that the discontinuity is represented with a first order error, but it doesn't grow significantly with time. The figures demonstrate very clearly that the error growth in time is much weaker with the 4th order scheme as predicted by the analysis above.

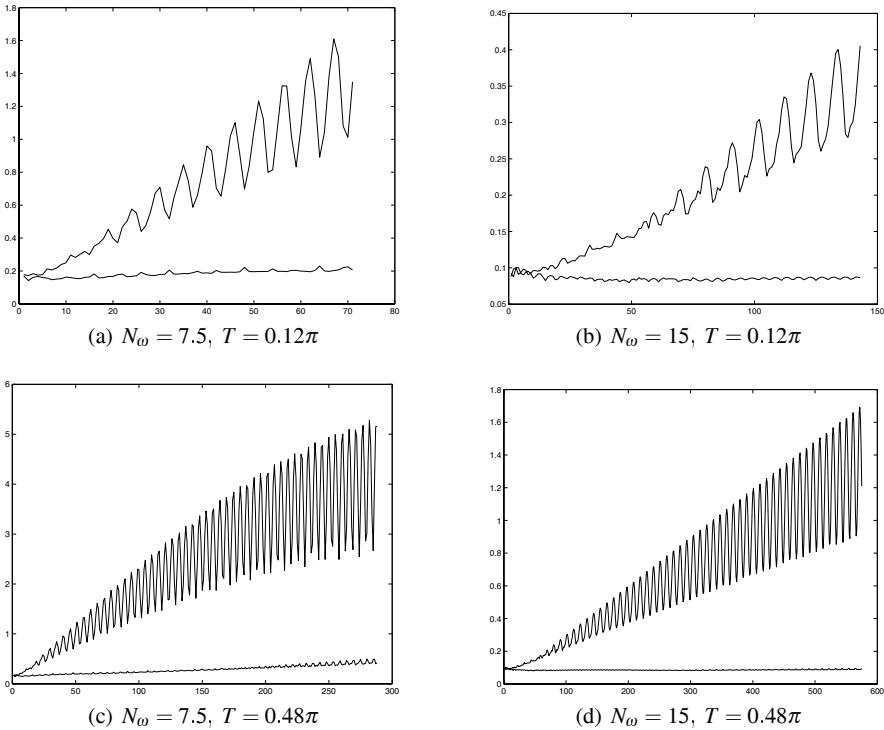


Fig. 9.9 The error ε^n as a function of time for Q_2 and Q_4 (lower curve), $\omega_0 = 128$

Next we show the solution as a function of x at $t = 0.48\pi$ for the higher frequency $\omega_0 = 256$. Figure 9.10 shows the result of the 2nd and 4th order scheme for 10 points per wave length. The solution is shown in an interval of length $\pi/20$ around the interface \bar{x} (located at point no. 81 in the figure). The Q_2 -scheme is completely out of phase, while the Q_4 -scheme is almost exact.

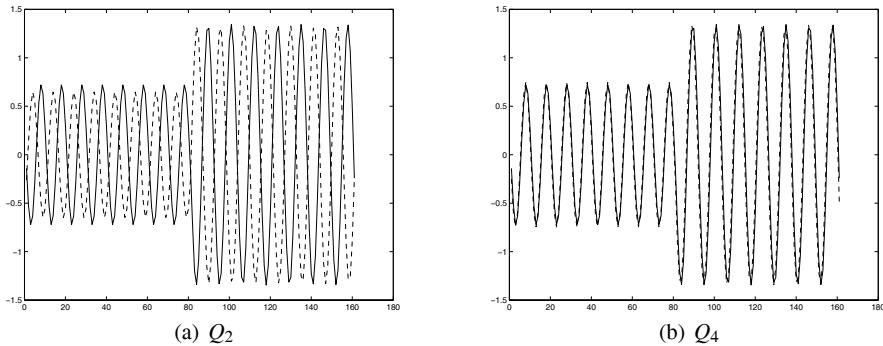


Fig. 9.10 p as a function of x , $\omega_0 = 256$, $N_\omega = 10$. Exact solution (—), numerical solution (---)

When decreasing the number of points per wavelength further, even Q_4 goes out of phase, and the sixth order scheme Q_6 should be used. Figure 9.11 shows the solution u for $N_\omega = 7.5$

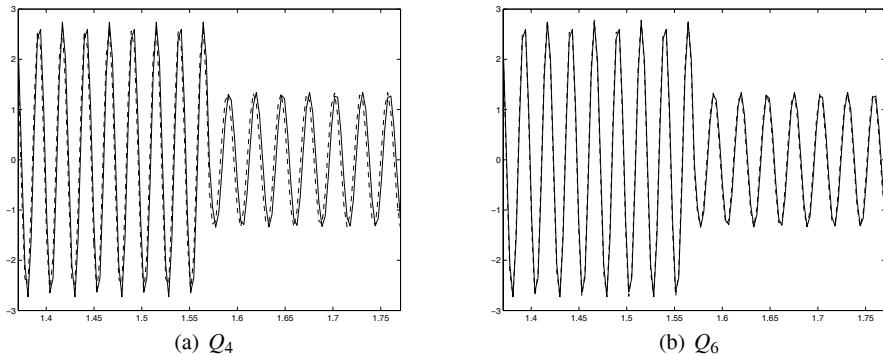


Fig. 9.11 u as a function of x , $\omega_0 = 256$, $N_\omega = 10$. Exact solution (—), numerical solution (---)

Two space dimensions

For the one-dimensional problem it was possible to derive error estimates for the case with discontinuous coefficients by making a direct comparison between the exact solutions for the differential equation and for the difference approximation. For the general two-dimensional case this is technically very complicated, and probably impossible in practice. Furthermore, the solution is no longer continuous if the coefficients are discontinuous. In Section 9.2.3 we shall set up a simple 1-D model problem with discontinuous solutions, and show how the coefficients can be modified based on a theoretical analysis. Here we shall present some numerical exper-

iments in 2-D with a straightforward application of the original coefficients, where we use the numerical solution on a very fine grid as the exact one. The model problem is a tilted square within a larger periodic square with side length 2π such that the grid points don't fall on the boundary. Figure 9.12 shows the grid near a corner of the tilted square.

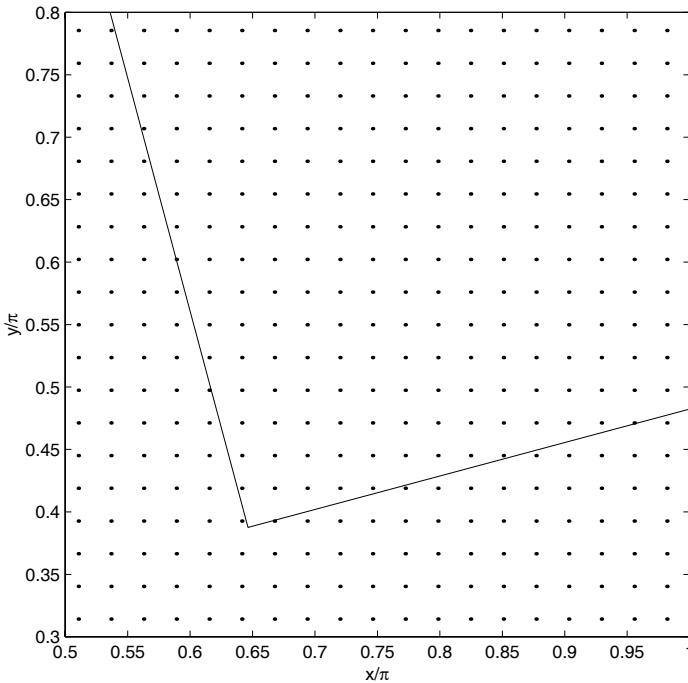


Fig. 9.12 The grid near the corner of the tilted square

As for the 1-D case we use the acoustic interpretation of the variables with $a = -\rho c^2$, $b = -1/\rho$, where $\rho = c = 1$ inside the tilted square and $\rho = 0.55556$, $c = 0.87879$ outside. The initial conditions are

$$p(x, y, k/2) = \begin{cases} \sin(c_L \omega k/2 - \omega x), & -\pi/4 \leq x \leq \pi/4 \\ 0 & \text{elsewhere,} \end{cases}$$

$$u(x, y, 0) = \begin{cases} \sin(-\omega x)/(c_L \rho_L), & -\pi/4 \leq x \leq \pi/4 \\ 0 & \text{elsewhere,} \end{cases}$$

$$v(x, y, 0) = 0 \quad \text{in the whole region}$$

for some wave number ω . This is a right-going wave train which is fully contained in the outer region. When it hits the inner square, the waves are speeded up, and

there are reflections. Figure 9.13 shows the pressure p at $t = 0$ and at $t = 0.98\pi$ on a grid with 336×336 points and $k/h = 0.392$. The inner square is tilted 15 degrees, and the wave number is $\omega = 32$.

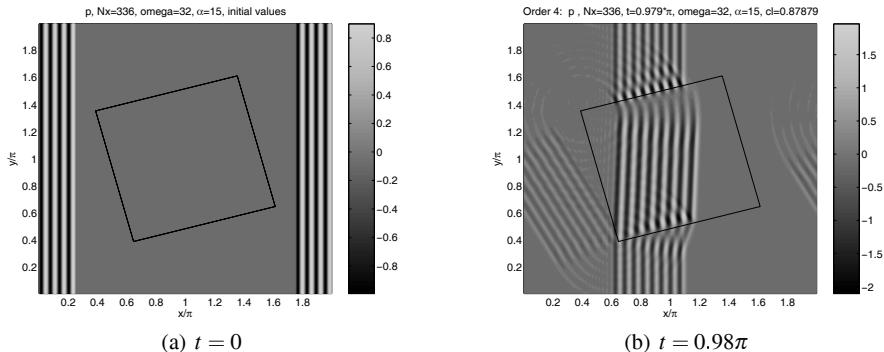


Fig. 9.13 p at $t = 0$ and $t = 0.98\pi$

The error in the solution p was computed as the l_2 -norm in the x -direction for a fixed y , where the fourth order solution on a 1680×1680 grid represents the true solution. Figure 9.14 shows the error as a function of time for two different grids. In all of the figures, the upper curve is the error for the second order method, the lower one for the fourth order.

The one-dimensional analysis in Section 9.1 shows that the error has one part of high order corresponding to the usual truncation error for smooth solutions, and another part of first order arising from the coarse approximation of the internal boundary. The numerical 2-D results show a very similar behavior. The uncertain location of the boundary relative to the grid is the same for both the second and fourth order method, but the latter one produces much more accurate results. Even if the computing time is not measured here, it seems clear that for long time integrations, the fourth order scheme is more effective than the second order one.

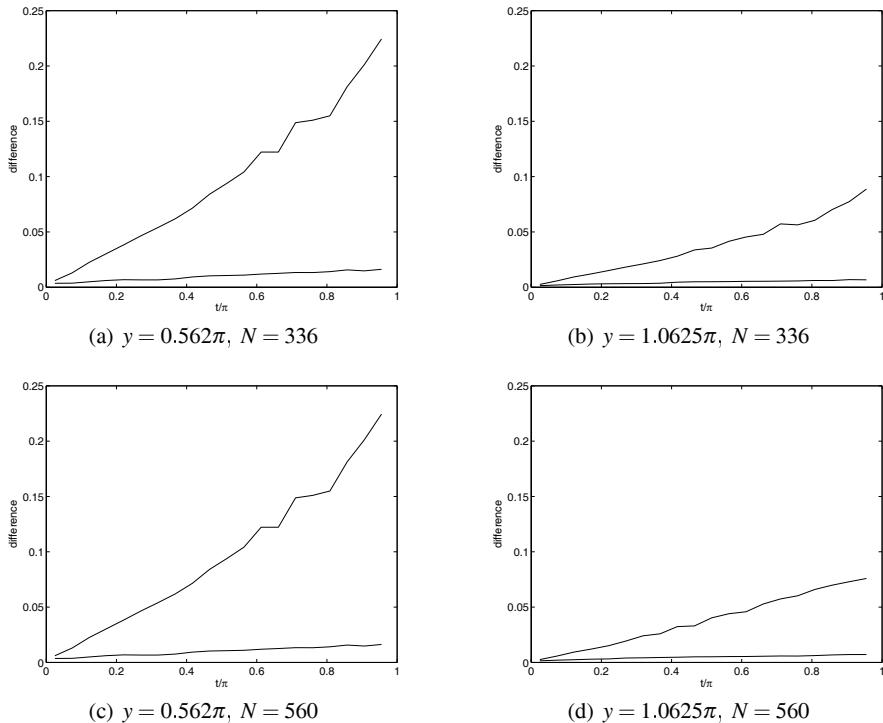


Fig. 9.14 The error in p for the 2nd (upper curve) and 4th (lower curve) order schemes on an $N \times N$ grid

9.2.2 Modified Coefficients

The results in the previous section show that the very coarse treatment of the internal boundary is less severe than could be expected. In contrast to the ordinary high order error, the first order error arising at the discontinuity does not increase very much with time. Still, one may want to decrease the level of this error, but without causing any extra implementation difficulties. The advantage of the original scheme is that once the coefficients are defined in the whole domain, the scheme is applied everywhere without restrictions. Therefore, if one could modify the coefficients and define them uniquely before the computation in such a way that the error is decreased, then the effectiveness would be retained.

In Section 9.2.1 it was shown that the formal accuracy is first order for the whole class of one step schemes considered there. The theoretical results were based on an analysis where the true solutions of the PDE and of the numerical method were explicitly calculated. Here we shall take a different approach, and study the local

accuracy of the approximation near the internal boundary located at \bar{x} , and then use the general stability theory for deriving an error estimate.

The Yee scheme

We begin by considering the Yee scheme, and assume that $(a(x), b(x))$ is constant on each side of \bar{x} , and takes the values (a^L, b^L) and (a^R, b^R) respectively. Since the derivatives in space are discontinuous, the difficulty is the approximation in space. We consider first the right hand side of the first equation and ask: how big is the truncation error

$$\tau = a(x_{j+1/2})D_+u(x_j) - a(x_{j+1/2})u_x(x_{j+1/2})$$

when $x_j \leq \bar{x} < x_{j+1}$? By Taylor expansion around \bar{x} we get

$$D_+u(x_j) = \frac{1}{h} \left((x_{j+1} - \bar{x})u_x^R - (x_j - \bar{x})u_x^L + \mathcal{O}(h^2) \right),$$

where u_x^L and u_x^R are the left and right limit values of u_x at $x = \bar{x}$. The expansion of au_x takes different forms depending on the location of the grid:

$$a(x_{j+1/2})u_x(x_{j+1/2}) = \begin{cases} a^L u_x^L + \mathcal{O}(h) & \text{if } x_{j+1/2} \leq \bar{x}, \\ a^R u_x^R + \mathcal{O}(h) & \text{if } x_{j+1/2} > \bar{x}. \end{cases}$$

The truncation error is

$$\tau = \begin{cases} a^L \left(\frac{1}{h} \left((x_{j+1} - \bar{x})u_x^R - (x_j - \bar{x})u_x^L \right) - u_x^L \right) + \mathcal{O}(h) & \text{if } x_{j+1/2} \leq \bar{x}, \\ a^R \left(\frac{1}{h} \left((x_{j+1} - \bar{x})u_x^R - (x_j - \bar{x})u_x^L \right) - u_x^R \right) + \mathcal{O}(h) & \text{if } x_{j+1/2} > \bar{x}. \end{cases}$$

Obviously $\tau = \mathcal{O}(1)$, since $u_x^R - u_x^L = \mathcal{O}(1)$.

This truncation error occurs only at one point, and it doesn't contradict the first order error in the solution that was derived in Section 9.2.1. It is the same type of mechanism here as for the numerical boundary conditions that were discussed in Section 3.2, where it was shown that the order of accuracy can be lowered one step for hyperbolic problems.

In order to eliminate the $\mathcal{O}(1)$ part of the error, we modify the definition of $a(x)$ in the difference scheme such that the new function \tilde{a} satisfies

$$\tilde{a}(x) = \frac{a^L a^R}{a^R + (a^L - a^R)H_h(x - \bar{x})},$$

where

$$H_h(x) = \begin{cases} 0 & \text{if } x \leq -h/2, \\ 1/2 + x/h & \text{if } -h/2 < x \leq h/2, \\ 1 & \text{if } x > h/2 \end{cases} \quad (9.23)$$

is the regularized Heaviside function. Let us now verify the accuracy for the interval of interest, which is $-h/2 < x_{j+1/2} - \bar{x} \leq h/2$. Since $a^L u_x^L = a^R u_x^R$, the new

truncation error is

$$\tilde{\tau} = \frac{\tilde{a}(x_{j+1/2})}{h} ((x_{j+1} - \bar{x})u_x^R - (x_j - \bar{x})u_x^L) - a^L u_x^L + \mathcal{O}(h)$$

in the whole interval. We have

$$\begin{aligned}\tilde{\tau} &= \frac{a^L a^R ((x_{j+1} - \bar{x})u_x^R - (x_j - \bar{x})u_x^L)}{h a^R + (a^L - a^R)(h/2 + x_{j+1/2} - \bar{x})} - a^L u_x^L + \mathcal{O}(h) \\ &= \frac{a^L (x_{j+1} - \bar{x}) a^L u_x^L - a^R (x_j - \bar{x}) a^L u_x^L}{a^L (x_{j+1} - \bar{x}) - a^R (x_j - \bar{x})} - a^L u_x^L + \mathcal{O}(h) = \mathcal{O}(h).\end{aligned}$$

For the second PDE, the same type of analysis is done, and the truncation error becomes $\mathcal{O}(h)$ if $b(x)$ is substituted by

$$\tilde{b}(x) = \frac{b^L b^R}{b^R + (b^L - b^R) H_h(x - \bar{x})},$$

where $H_h(x)$ is defined in (9.23). With the first order approximations at the internal boundary obtained in this way, one can now prove that the global accuracy is $\mathcal{O}(h^2)$ by using the technique from Chapter 3. We omit this part here.

It should be noted that the modification of the coefficients is not a polynomial regularization in the traditional sense of $a(x)$ and $b(x)$ themselves, but rather of $1/a(x)$ and $1/b(x)$. The modification of $a(x)$ for the case $a^L = 2$, $a^R = 1$ is shown in Figure 9.15.

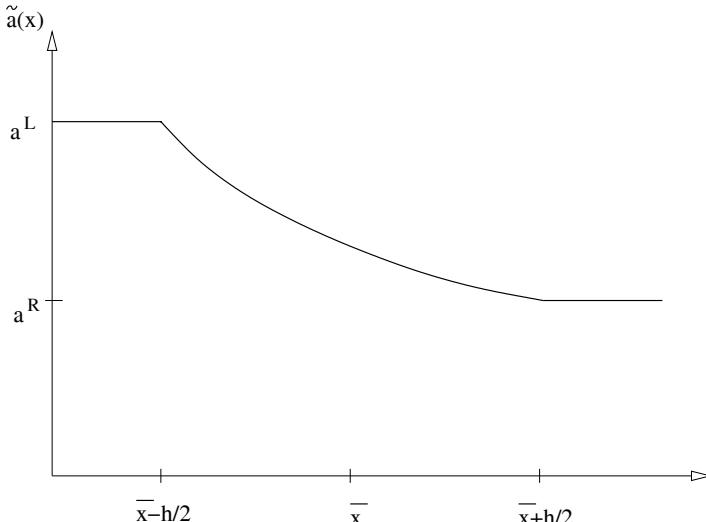


Fig. 9.15 The function $\tilde{a}(x)$ for $a^L = 2$, $a^R = 1$

The 4th order scheme

We next turn to the 4th order scheme (called Q_4 here). Full accuracy would require local 3rd order accuracy near the discontinuity. It turns out that the principle used above for modification of the coefficients does not work to achieve this, even if the interval around the discontinuity is extended. However, we have shown in Section 9.2.1 that Q_4 beats Q_2 even for zero order local accuracy of the approximation at the interface. Therefore it is worthwhile raising the local accuracy to first order, just as for Q_2 . However, taking the original Q_4 and substitute $a_{j+1/2}$ by $\tilde{a}_{j+1/2}$ does not work. The reason is that the second order correction terms in Q_4 are $\mathcal{O}(1)$ across the interface, and the modified coefficients were not constructed to take care of this error. The remedy is to switch off these terms locally.

We write the modified scheme in the form

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + k\tilde{a}_{j+1/2}D_+u_j^n \\ &\quad + \frac{k}{24}(k^2\tilde{a}_{j+1/2}\alpha_{j+1/2}D_+\tilde{b}_jD_-\tilde{a}_{j+1/2}D_+ - h^2\tilde{a}_{j+1/2}\alpha_{j+1/2}D_+^2D_-)u_j^n, \\ u_j^{n+1} &= u_j^n + k\tilde{b}_jD_-p_{j+1/2}^{n+1/2} \\ &\quad + \frac{k}{24}(k^2\tilde{b}_j\alpha_jD_-\tilde{a}_{j+1/2}D_+\tilde{b}_jD_- - h^2\tilde{b}_j\alpha_jD_+D_-^2)p_{j+1/2}^{n+1/2}, \end{aligned} \tag{9.24}$$

where

$$\alpha(x) = \begin{cases} 1 & \text{if } |x - \bar{x}| \geq 3h/2, \\ 0 & \text{if } |x - \bar{x}| < 3h/2. \end{cases}$$

In this way, we still have a local $\mathcal{O}(h)$ approximation near the interface, and an $\mathcal{O}(h^4)$ approximation in the smooth part.

We shall now present numerical results to compare the different schemes, and the example is the same as in Section 9.2.1 with $\omega_0 = 16$. The interface is located at $\bar{x} = \pi/2$, and all the computations were done for $0 \leq t \leq 0.48\pi$. The error depends on the location of the interface relative to the grid, and therefore several experiments were carried out with different small perturbations of the location:

$$\bar{x}_\nu = \left(\frac{1}{2} + \frac{\beta}{M}\nu h\right)\pi, \quad \nu = 1, 2, \dots, M,$$

where $\beta = \sqrt{2}/1.42$. In order to compute the error at the same points, the grids are refined by a factor 3 with $N = 480, 1440, 4320$.

From the theoretical analysis, we know that the error should be $\mathcal{O}(h)$ with the original coefficients, and $\mathcal{O}(h^2)$ with modified coefficients. Figure 9.16 shows the error in max-norm (over the interval $[0, \pi]$) as a function of time, and it agrees well with the theoretical results. Note also that for the modified coefficients, the error is small initially, but increases with time. This is an effect of the second order phase error, which will dominate already from the beginning.

For the fourth order scheme, the original coefficients give first order accuracy. For the modified coefficients implemented as in (9.24) we expect one order higher. Also here, as shown in Figure 9.17, there is quite good agreement with the theory.

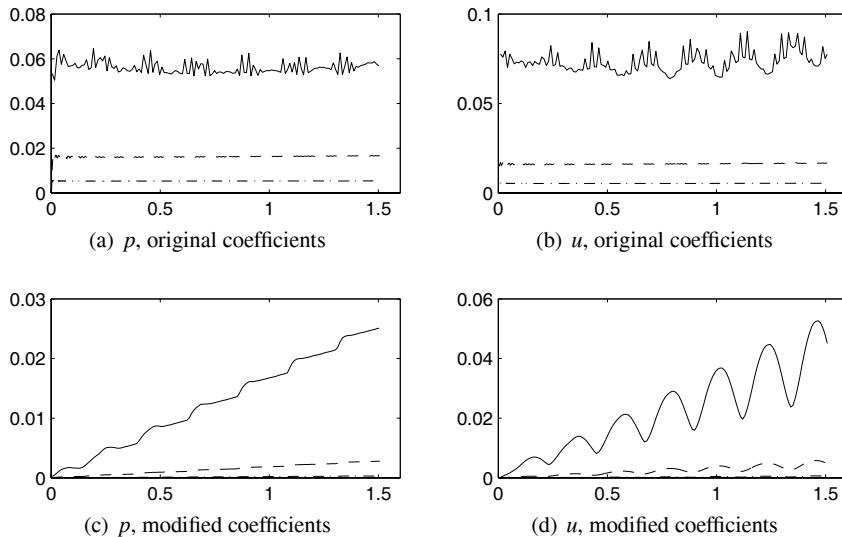


Fig. 9.16 The error in the solution versus t for the Yee scheme, $N = 480, 1440, 4320$

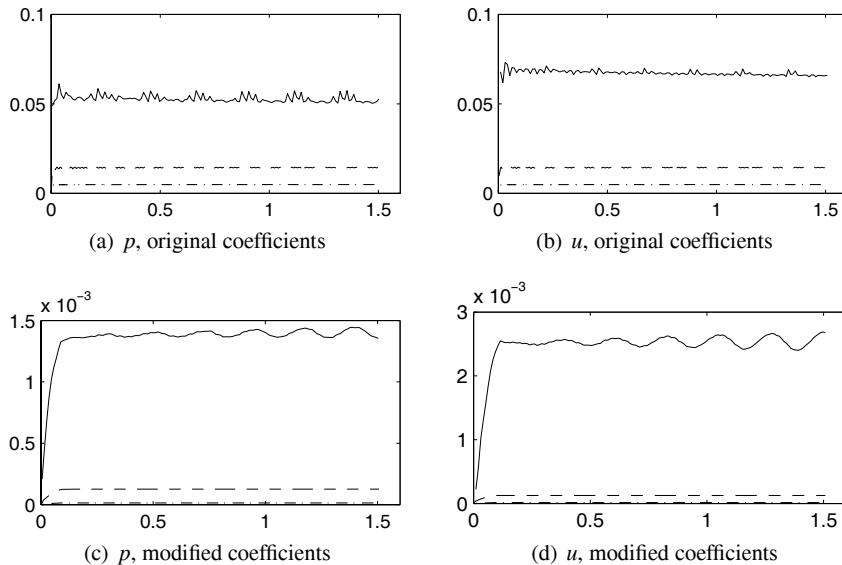


Fig. 9.17 The error in the solution versus t for the 4th order scheme, original coefficients, $N = 480, 1440, 4320$

9.2.3 An Example with Discontinuous Solution

The solution of the 1-D wave equation is continuous even for the case with discontinuous coefficients. For problems in several space dimensions this is not true. Therefore we shall construct and analyze a slightly modified 1-D problem, where the solution is discontinuous. Let $a(x)$ and $b(x)$ be piecewise constant with a discontinuity at $x = \bar{x}$ and consider the system

$$\begin{aligned} p_t &= (au)_x, \\ u_t &= bp_x, \end{aligned}$$

where the solution u is discontinuous at $x = \bar{x}$. With $\tilde{u} = au$, we get the system

$$\begin{aligned} p_t &= \tilde{u}_x, \\ \tilde{u}_t &= abp_x. \end{aligned} \tag{9.25}$$

This system has the same structure as the one studied previously in this chapter. The solutions p and \tilde{u} are continuous, and the jump conditions corresponding to (9.21) are

$$\begin{aligned} [\tilde{u}_x] &= 0, \\ [abp_x] &= 0. \end{aligned}$$

Expressed in the original variables, we have

$$\begin{aligned} p^L &= p^R, \\ a^L u^L &= a^R u^R, \\ a^L u_x^L &= a^R u_x^R, \\ a^L b^L p_x^L &= a^R b^R p_x^R. \end{aligned} \tag{9.26}$$

The Yee scheme for the original system is

$$\begin{aligned} p_{j+1/2}^{n+1/2} &= p_{j+1/2}^{n-1/2} + kD_+(a_j u_j^n), \\ u_j^{n+1} &= u_j^n + kb_j D_- p_{j+1/2}^{n+1/2}. \end{aligned}$$

Assuming that $x_j \leq \bar{x} < x_{j+1}$, a Taylor expansion gives

$$\begin{aligned} \frac{1}{h} (a(x_{j+1})u(x_{j+1}) - a(x_j)u(x_j)) \\ = \frac{1}{h} \left(a^R (u^R + (x_{j+1} - \bar{x})u_x^R) - a^L (u^L + (x_j - \bar{x})u_x^L) \right) + \mathcal{O}(h), \end{aligned}$$

where the superscripts L and R denote the left and right limits respectively at $x = \bar{x}$. This expression is supposed to approximate $(au)_x$, and since a is piecewise constant,

we have

$$(au)_x(x_{j+1/2}) = \begin{cases} a^L u_x^L + \mathcal{O}(h) & \text{if } x_{j+1/2} < \bar{x} \\ a^R u_x^R + \mathcal{O}(h) & \text{if } x_{j+1/2} \geq \bar{x}. \end{cases}$$

By using the jump conditions (9.26), we get

$$\begin{aligned} \frac{1}{h} (a(x_{j+1})u(x_{j+1}) - a(x_j)u(x_j)) &= a^R u_x^R + \mathcal{O}(h), \\ (au)_x(x_{j+1/2}) &= a^R u_x^R + \mathcal{O}(h), \end{aligned}$$

on both sides of the interface, i.e., the truncation error is $\mathcal{O}(h)$.

When considering the second differential equation, the interval of interest is $x_{j-1/2} \leq \bar{x} < x_{j+1/2}$. A Taylor expansion gives, when taking the jump conditions (9.26) into account

$$\begin{aligned} \frac{b(x_j)}{h} (p(x_{j+1/2}) - p(x_{j-1/2})) &= \frac{b(x_j)}{h} (p^R + (x_{j+1/2} - \bar{x})p_x^R - p^L - (x_{j-1/2} - \bar{x})p_x^L) \\ &\quad + \mathcal{O}(h) = \frac{b(x_j)p_x^L}{h} ((x_{j+1/2} - \bar{x}) \frac{a^L b^L}{a^R b^R} - (x_{j-1/2} - \bar{x})) + \mathcal{O}(h) \\ &= \frac{b(x_j)p_x^R}{h} ((x_{j+1/2} - \bar{x}) - (x_{j-1/2} - \bar{x}) \frac{a^R b^R}{a^L b^L}) + \mathcal{O}(h). \end{aligned}$$

For consistency, the last expression must be equal to $b^L p_x^L + \mathcal{O}(h)$ if $x_j \leq \bar{x}$, and equal to $b^R p_x^R + \mathcal{O}(h)$ if $x_j > \bar{x}$. This is achieved if $b(x)$ is modified to

$$\tilde{b} = \begin{cases} b^L & \text{if } x \leq \bar{x} - h/2 \\ \Theta((x - \bar{x})/h, a^L b^L, a^R b^R) / a^L & \text{if } \bar{x} - h/2 < x \leq \bar{x} \\ \Theta((x - \bar{x})/h, a^L b^L, a^R b^R) / a^R & \text{if } \bar{x} < x \leq \bar{x} + h/2 \\ b^R & \text{if } \bar{x} + h/2 < x, \end{cases} \quad (9.27)$$

where

$$\Theta(\xi, \phi^L, \phi^R) = \frac{\phi^L \phi^R}{(\phi^L + \phi^R)/2 + \xi(\phi^L - \phi^R)}.$$

In conclusion, the coefficient $a(x)$ is not modified, but $b(x)$ is substituted by \tilde{b} defined in (9.27). In that way, the local truncation error is $\mathcal{O}(h)$, and we can expect overall second order accuracy. Note that \tilde{b} is not a regularization of b in the traditional sense. There is still a discontinuity, but the shape is modified in the neighborhood, see Figure 9.18.

The numerical experiments were done with the parameters

$$\begin{aligned} a^L &= -1/2, b^L = -1.8, \\ a^R &= -1, \quad b^R = -1. \end{aligned}$$

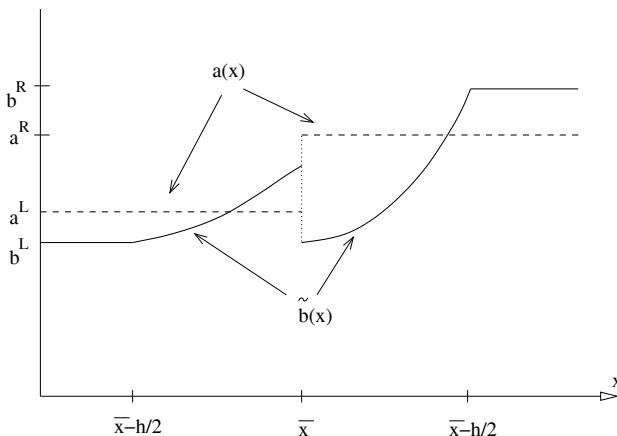


Fig. 9.18 The coefficients $a(x)$ and $\tilde{b}(x)$ defined in (9.27)

By the transformation $\tilde{u} = au$ leading to the system (9.25), the true solution p, \tilde{u} can be calculated as in Section 9.2.1. We choose the form (9.22) with u substituted by \tilde{u} and with c_L, ρ_L, c_R, ρ_R chosen such that they match the coefficients 1 and ab in (9.25). The original discontinuous true solution p, u is then obtained by $u = \tilde{u}/a$. Figure 9.19 shows the error as a function of time for the original and modified coefficients.

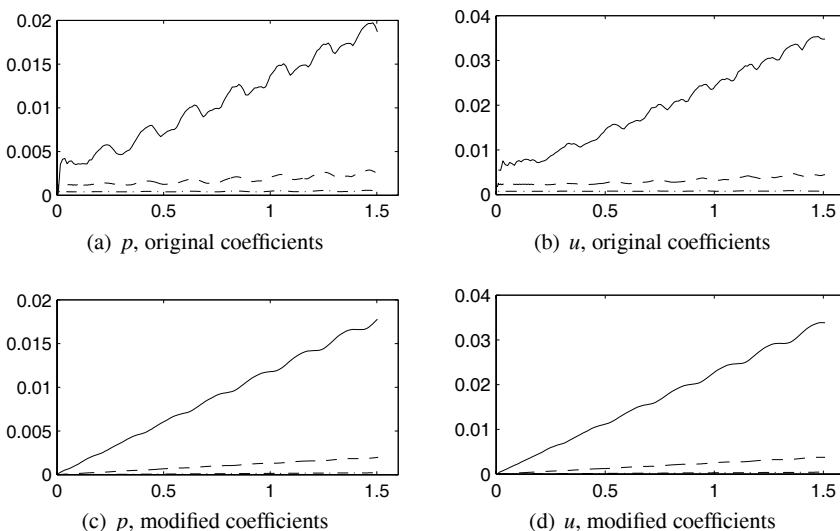


Fig. 9.19 The error in the solution versus t for the Yee scheme, $N = 480, 1440, 4320$

Here we make an interesting observation. The order of the error agrees roughly with the theoretical prediction in the beginning of the computation. However, at the end of the integration, the error is approximately the same for the original and modified coefficient case for each grid, and it is of second order. The explanation is that even if the time integration is not very long compared to the typical wavelength, the time dependent phase error takes over very quickly for this problem. This indicates that the 4th order scheme would do a much better job also here.

9.3 Boundary Treatment

So far we have disregarded the boundaries in this chapter. Initial-boundary value problems require not only approximation of the physical boundary conditions, but also construction of numerical boundary conditions. In Sections 2.3 and 2.4 it was demonstrated how these conditions are constructed and analyzed with respect to stability, and in Chapter 3 the order of accuracy was discussed. Here we shall first apply these principles to the wave equation. Then we shall discuss how the technique presented in the previous section for treating internal boundaries can be generalized to external physical boundaries.

9.3.1 High Order Boundary Conditions

Consider the 1-D wave equation in a semi-infinite domain with a solid wall boundary at $x = 0$:

$$\begin{aligned} \begin{bmatrix} p_t \\ u_t \end{bmatrix} &= \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} p_x \\ u_x \end{bmatrix}, \quad 0 \leq x < \infty, \quad 0 \leq t, \\ u(0, t) &= 0, \\ p(x, 0) &= f^{(1)}(x), \\ u(x, 0) &= f^{(2)}(x). \end{aligned}$$

For the Yee scheme on a staggered grid with $x_j = jh$, $x_{j+1/2} = (j + 1/2)h$, the boundary condition is obvious:

$$u_0^{n+1} = 0.$$

No numerical boundary condition is required because of the compact nature of the scheme. The fourth order scheme requires two extra conditions. Differentiation of the boundary condition shows that all derivatives of u with respect to t are zero at $x = 0$. Differentiation of the PDE system then leads to the conclusion that all even x -derivatives of u and all odd x -derivatives of p are zero. Therefore, we define the grid such that $x_{-1/2}$ is the leftmost point for p and x_{-1} for u , and use the extra conditions

$$\begin{aligned} D_+ p_{-1/2}^{n+1/2} &= 0, \\ D_+ D_- u_0^{n+1} &= 0. \end{aligned}$$

(The last condition simplifies to $u_{-1}^{n+1} + u_1^{n+1} = 0$.) Note that these equations are exact, since the truncation error contains only odd derivatives of p and even derivatives of u . Assume that

$$(p_{j+1/2}^{n-1/2}, u_j^n), \quad j = -1, 0, \dots$$

are known. Then the advancement one step is described by the algorithm

1. Compute $p_{j+1/2}^{n+1/2}, \quad j = 0, 1, \dots$ from (9.11)
2. Put $p_{-1/2}^{n+1/2} = p_{1/2}^{n+1/2}$
3. Compute $u_j^{n+1}, \quad j = 1, 2, \dots$ from (9.11)
4. Put $u_0^{n+1} = 0$
5. Put $u_{-1}^{n+1} = -u_1^{n+1}$

(In practice there is of course a boundary to the right as well, where the same principle is used.) Another way of implementation is to substitute $p_{-1/2}^{n+1/2}$ and u_{-1}^{n+1} as defined by step 2 and 5 into the general formulas in step 1 and 3.

The procedure can be generalized to the 6th and higher order versions of the scheme by adding more points to the left, and use approximations of the higher order odd derivatives of p and of the higher order even derivatives of u centered at $x = 0$.

9.3.2 Embedded Boundaries

For internal boundaries we have demonstrated how the methods presented in this chapter can be applied across the interface without enforcing any extra conditions that depend on the solution. This means that problems with irregular interfaces can be solved fast on Cartesian grids. For real physical boundaries, like solid walls, the same advantage is attained if we can extend the computational domain such that the boundary becomes embedded in a regular domain, and if the coefficients can be defined in such a way, that the true boundary condition will be satisfied. This is what we shall discuss in this section.

We change the domain from the previous section, such that the boundary is located at $x = \bar{x}$, with the domain of interest to the left:

$$\begin{aligned} \begin{bmatrix} p_t \\ u_t \end{bmatrix} &= \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} p_x \\ u_x \end{bmatrix}, \quad -\infty < x \leq \bar{x}, \quad 0 \leq t, \\ u(\bar{x}, t) &= 0, \\ p(x, 0) &= f^{(1)}(x), \\ u(x, 0) &= f^{(2)}(x). \end{aligned} \tag{9.28}$$

This is the problem to solve, but we change it by adding the quarter-space to the right of \bar{x} . In order to guarantee that the solutions agree for $x \leq \bar{x}$, we extend the initial data such that $u(x, 0) = 0$ for $x > \bar{x}$, while the extension of p is arbitrary. The coefficient b is defined as zero for $x > \bar{x}$. For simplicity, we assume also here piecewise constant coefficients:

$$a(x) = a^L, \quad -\infty < x < \infty,$$

$$b(x) = \begin{cases} b^L & \text{if } x < \bar{x} \\ 0 & \text{if } x \geq \bar{x}. \end{cases}$$

The new problem is

$$\begin{aligned} \begin{bmatrix} p_t \\ u_t \end{bmatrix} &= \begin{bmatrix} 0 & a^L \\ b(x) & 0 \end{bmatrix} \begin{bmatrix} p_x \\ u_x \end{bmatrix}, \quad -\infty < x < \infty, \quad 0 \leq t, \\ p(x, 0) &= f^{(1)}(x), \\ u(x, 0) &= f^{(2)}(x), \end{aligned} \tag{9.29}$$

where we have used the same notation for the extended functions $f^{(1)}(x)$, $f^{(2)}(x)$ as for the original ones. Since $u_t(\bar{x}, t) = 0$, and $u(\bar{x}, 0) = 0$, the solution (p, u) of (9.29) satisfies (9.28).

The Yee scheme is now well defined for this problem. However, the stability proof from Section 9.1.1 does not apply. Since $b = 0$ to the right of \bar{x} , the norm which has b_j in the denominator, must be modified. Assuming that \bar{x} is located in the interval $(x_m, x_{m+1}]$, the squared norm is defined in analogy with (9.6) as

$$E^n = \sum_{j=-\infty}^{\infty} \frac{1}{a_{j+1/2}} |p_{j+1/2}^{n-1/2}|^2 h + \sum_{j=-\infty}^m \frac{1}{b_j} |u_j^n|^2 h - \lambda \sum_{j=-\infty}^m (p_{j+1/2}^{n-1/2} - p_{j-1/2}^{n-1/2}) u_j^n h.$$

By using the same basic technique as in Section 9.1.1, one can show that $E^{n+1} = E^n$ if

$$\lambda \max_j \left(\max(a_{j+1/2} b_j, a_{j-1/2} b_j) \right)^{1/2} < 1.$$

(This condition implies (9.7), (9.7).)

For the numerical experiments, we use $a^L = b^L = -1$, and define the initial pulse

$$p(x, 0) = u(x, 0) = e^{-800(x-0.4\pi)^2},$$

which starts at $x = 0.4\pi$ and moves to the right until it hits the boundary at $x = \pi/2$, where it is reflected, see Figure 9.20.

For the numerical solution, we need numerical values also for a^R and b^R , and we use $a^R = a^L$. From an implementation point of view, it is convenient to use nonzero values for b^R , and we choose $b^R = 10^{-10}$. Figure 9.21 shows the computed solution obtained by the Yee scheme, first when the pulse hits the boundary and then after the reflection is complete.

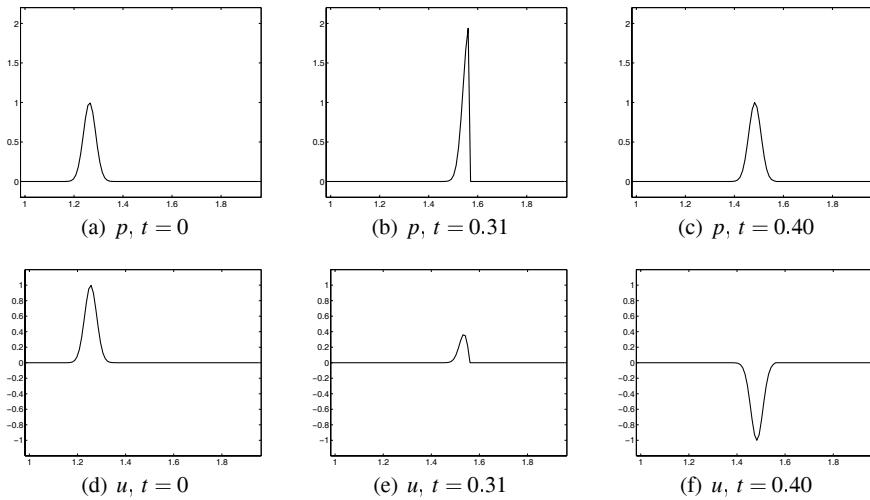


Fig. 9.20 p and u as a function of t

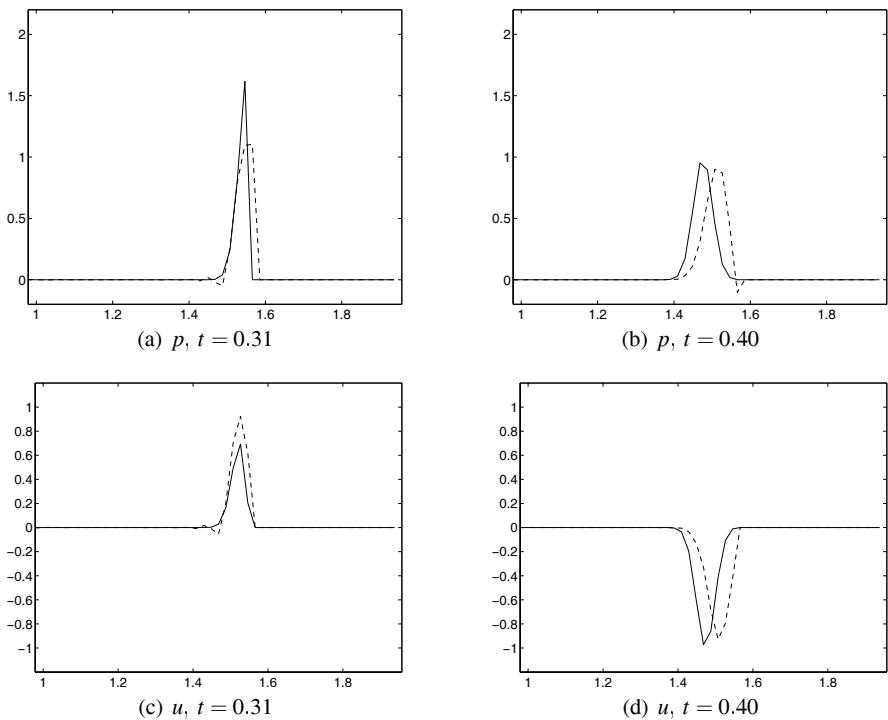


Fig. 9.21 Exact solution (—) and Yee scheme solution(---), $N = 321$

The scheme behaves quite well even near the boundary, and there is no sign of any instability. However, the exact location of the boundary is unknown to the scheme, and therefore the error in the solution is $\mathcal{O}(h)$. With a twice as fine grid, the computed solution is better as shown in Figure 9.22. (The points of time are not exactly the same for different grids, since the full and half time-step doesn't exactly match the two points $t = 0.31$ and $t = 0.40$.)

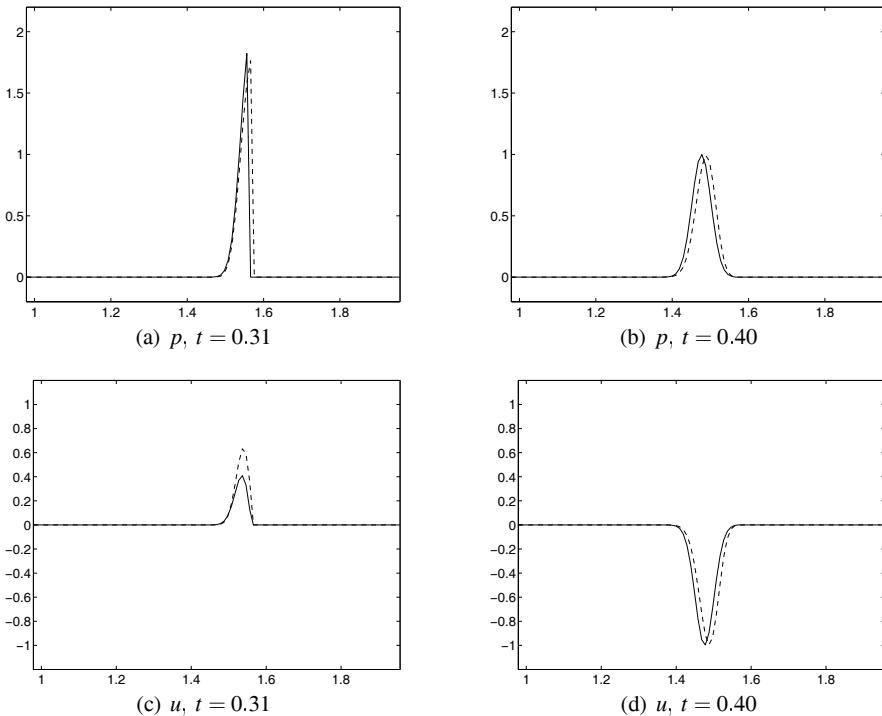


Fig. 9.22 Exact solution (—) and Yee scheme solution(— —), $N = 642$

In accordance with the principles derived for the internal boundary case discussed in Section 9.2.1, we expect the fourth order scheme to do a much better job for long time integration. However, for the short run presented here, there is not much difference as shown in Figure 9.23 for the coarse grid with $N = 321$.

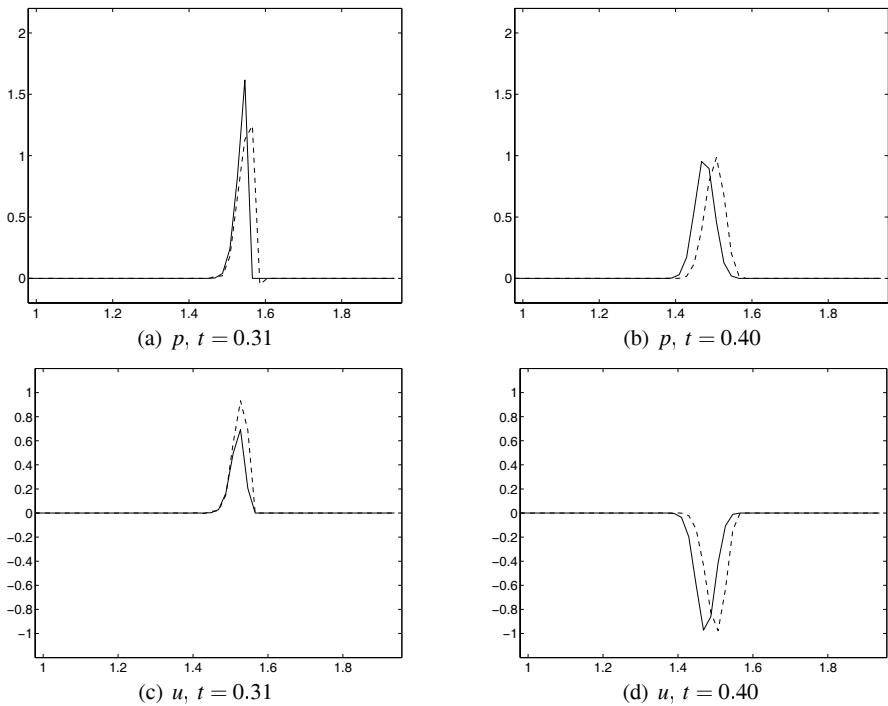


Fig. 9.23 Exact solution (—) and numerical solution, 4th order scheme (---), $N = 321$

Modified coefficients

By modifying the coefficients around internal boundaries, the accuracy could be improved as demonstrated in Section 9.2.2. The same principle can be applied here. Consider the situation shown in Figure 9.24, where $u_{m+1}^n = 0$ is the first point to the right of the boundary. When computing p at the nearest point to the left, the formula without modification is

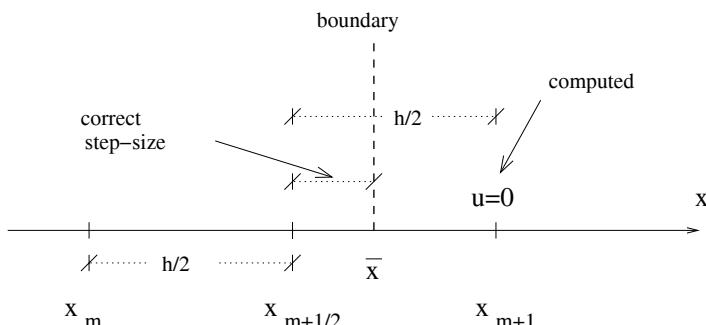


Fig. 9.24 The grid near the boundary

$$p_{m+1/2}^{n+1/2} = p_{m+1/2}^{n-1/2} + \frac{ka^L}{h}(0 - u_m^n).$$

Since the true value $u = 0$ is attained at the boundary $x = \bar{x}$, the inconsistency in the formula can be seen as an error in the length h of the step size. This error will be present in the whole interval between $\bar{x} - h$ and \bar{x} , and this is where the coefficient a^L must be changed for higher accuracy. Therefore we define

$$a_h(x) = \begin{cases} a^L & \text{if } x \leq \bar{x} - h, \\ \frac{a^L h}{\bar{x} - x + h/2} & \text{if } \bar{x} - h < x \leq \bar{x}, \\ a^L & \text{if } \bar{x} < x. \end{cases} \quad (9.30)$$

For $x = x_{m+1/2}$ we get

$$p_{m+1/2}^{n+1/2} = p_{m+1/2}^{n-1/2} + \frac{ka^L}{\bar{x} - x_m}(0 - u_m^n),$$

which is a first order approximation. The new function $a_h(x)$ is actually a “deregularization” of $a(x)$, since a constant function is modified such that it becomes discontinuous, see Figure 9.25.

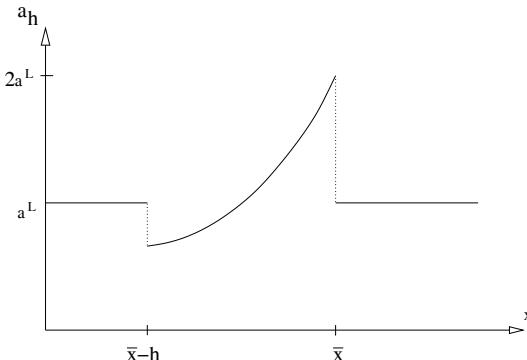


Fig. 9.25 The modified function $a_h(x)$

The modification of $b(x)$ is simpler. If the grid is located as in Figure 9.24, the original b -values are correct since $b(x_m) = b^L$ and $b(x_{m+1}) = 0$. However, if $x_m < \bar{x} \leq x_{m+1/2}$, then the unmodified formula sees the coefficient b^L as representing the whole interval $[x_{m-1/2}, x_{m+1/2}]$, which is not correct. By modifying b such that it is zero at x_m , we get $u_m = 0$. Since $u(\bar{x}, t) = 0$ implies $u(x_m, t) = \mathcal{O}(h)$ for the true solution, the error is $\mathcal{O}(h)$. The modified $b(x)$ is

$$b_h(x) = \begin{cases} b^L & \text{if } x \leq \bar{x} - h/2, \\ 0 & \text{if } x > \bar{x} - h/2, \end{cases} \quad (9.31)$$

i.e., the discontinuity has simply been relocated from \bar{x} to $\bar{x} - h/2$. Figure 9.26 shows the result for the same traveling pulse as presented earlier. The improvement compared to the original coefficient case shown in Figure 9.21 is clearly seen.

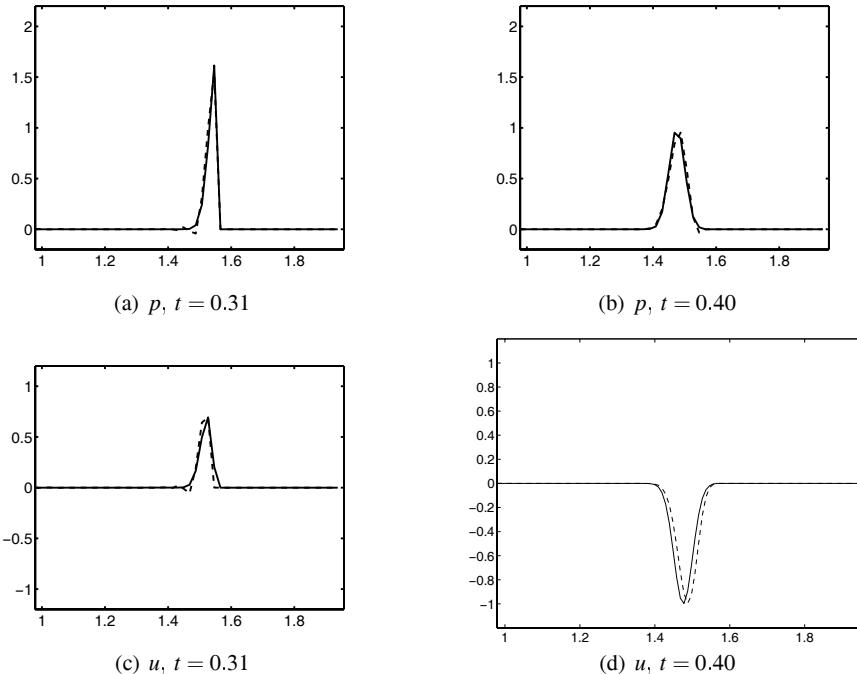


Fig. 9.26 Exact solution (—) and numerical solution with the Yee scheme and modified coefficients a and b (—), $N = 321$

9.4 Summary

This whole chapter is centered around a class of one step methods for the wave equation formulated as a first order system. The methods are generalizations of the Yee method on a staggered grid, and can be extended to any even order of accuracy. The numerical experiments are concentrated on problems with discontinuous coefficients, where the derivatives of higher order do not exist. The analysis and the experiments demonstrate that, even if no special treatment is applied at the internal boundaries, the higher order methods are still superior for typical problems, in particular for long time integrations. The error from the internal boundaries is first order, but increases very little with time. Therefore, the phase error dominates after

some time, and the usual considerations based on smooth solutions and accounted for in Chapter 1, apply also here.

Even if the experiments for our model problems show good solution behavior, it is clear that difficulties may occur if the internal boundaries are disregarded by the algorithm. In 1-D there is no problem, but in several space dimensions, the so called staircase problem has been shown to cause trouble for certain problems with curved boundaries. As shown in Figure 9.7, the method interprets the internal or external boundary as a staircase, and since the solution is discontinuous, the error may be too large. This problem has been treated by several authors, see for example [Ditkovski et al., 2001] and [Rylander and Bondesson, 2002]. The great advantage when avoiding special procedures at the internal boundaries, is that the implementation of the scheme can made effectively on parallel computers. In Section 9.2.2 we showed how the local first order accuracy can be raised one step by a certain a priori modification of the coefficients, which means that the implementation advantage is retained. The characteristic properties of the scheme are the same, but the error is kept smaller from the beginning. However, for discontinuous solutions, the theoretical analysis is limited to the 1-D case.

In the final section, it is shown how the physical boundaries can be embedded in the computational domain. By prescribing extreme values for the coefficients, almost exact boundary conditions are enforced without significant effects on the stability.

Much work on numerical solution of wave propagation problems has been done over the years, and this is not the place to give an extensive review. The standard second order scheme for the scalar wave equation is perhaps the most common model example in textbooks for illustration of difference schemes in general. Later work on this equation with particular emphasis on the boundary conditions is found in [Kreiss et al., 2004]. A new approach to the interface problem is give in [Mattsson and Nordström, 2006], where the projection operators described in Chapter 7 are used to match the solutions on each side. The fundamental Yee scheme for the first order system formulation was first presented in [Yee, 1966]. There has been several generalizations to high order accuracy, a good overview is found in the article by Turkel in the book [Taflove, 1998], where the Law–Wendroff principle is also applied. Later work by Turkel and his coworkers is found in [Turkel and Yefet, 2000], [Yefet and Turkel, 2000] and [Kashdan and Turkel, 2006]. Zheng et.al. constructed an unconditionally stable method for the Maxwell’s equations in [Zheng et al., 1999] and [Zheng et al., 2000], which was later generalized to higher order by Lee and Fornberg in [Lee and Fornberg, 2004], see also [Fornberg, 2003]. Certain parts of the material in this chapter are direct reprintings from [Gustafsson and Mossberg, 2004] and [Gustafsson and Wahlund, 2004], copyright ©2004 Society for Industrial and Applied Mathematics, reprinted with permission. Other papers forming the basis for the work presented here are [Gustafsson and Wahlund, 2005], [Tornberg and Engquist, 2003], [Tornberg and Engquist, 2004], [Tornberg and Engquist, 2006] and [Tornberg et al., 2006].

Finally it should be mentioned that there is another way of treating wave propagation problems, while still using finite difference methods. By first Fourier trans-

forming the equation in time, a boundary value problem is obtained for each frequency ω . We used this technique for the *analysis* in Section 9.2.1, but not for the actual computation. If the scalar 2nd order wave equation is used as a model, the new equation is the *Helmholtz equation*. This class of methods goes under the name *frequency domain methods*, and the major problem here is the construction of fast iterative solvers for the algebraic system that is obtained after discretization.

Chapter 10

A Problem in Fluid Dynamics

In this chapter, we shall discuss a problem in fluid dynamics and the application of a fourth order method. We shall concentrate on incompressible flow and the Navier–Stokes equations. This is a quite wide research area by itself, and an abundance of methods have been developed over the years. Many packages for numerical simulation are available, and they are usually based on finite element or finite volume methods. Here we shall describe a 4th order difference method that has been developed recently, and which is designed to be very effective, at least for the case where the computational domain can be transformed to a rectangle via a coordinate transformation.

10.1 Large Scale Fluid Problems and Turbulent Flow

One of the big challenges is turbulent flow, where the smallest scale of variation is extremely small compared to the computational domain. For example, full scale aircraft design requires the simulation of aerodynamic flow around a whole aircraft, or at least around a whole wing, at very high Reynolds numbers corresponding to the low viscosity of air. This means that one needs to resolve the flow on the millimeter level in a three-dimensional domain of several meters in each direction. This is still a too large computation on today’s powerful computers regardless of the numerical method that is used. There are many ways of substituting the small scale simulation by the introduction of turbulence models. In fact, this has been an area of intensive research for several decades in the fluid dynamics community. For special applications, many quite good results have been obtained this way, but there are still no established models that are sufficiently accurate for general flow problems.

For laminar flow, very large problems can be solved with good resolution of the smallest scales. However, most methods in practical use are of first or second order accuracy, and there are still many unsolved problems, where higher order methods are required. There are also some simpler turbulent flow problems, where there is a possibility of achieving meaningful results by direct numerical simulation all the

way to the smallest scale. But these problems require high order approximations for sure. In other words, there is a need of constructing fully developed high order methods for the standard PDE models in fluids.

A pervading theme in this book has been to make sure that the numerical methods are based on a sound theoretical basis. But our analysis has been limited to linear problems, and here we now have a genuinely nonlinear problem. However, unlike the compressible flow case, there are no shocks in the true sense for incompressible flow, and if the variations in the flow is resolved by the grid, there is good hope that the linear analysis is sufficient to get a good understanding of the methods. In the end, we have of course to rely on numerical experiments to confirm the relevance of the numerical results.

The numerical method described here for the Navier–Stokes equations is such that the Stokes part of the equations play a central role. Therefore we shall begin by considering these equations by itself.

10.2 Stokes Equations for Incompressible Flow

As we shall see later, the numerical method for the Navier–Stokes equations that we are going to develop, have a semi-implicit form with the implicit part corresponding to the Stokes equations. Therefore, we shall analyze these equations first, to get a better understanding of the properties.

Let $\mathbf{w} = (u, v, w)^T$ be the velocity components in the three coordinate directions respectively, and let p be the pressure. With the gradient and Laplace operator defined in the usual way, the Stokes equations for incompressible flow are

$$\begin{aligned}\mathbf{w}_t + \nabla p &= \varepsilon \Delta \mathbf{w}, \\ \nabla \cdot \mathbf{w} &= 0.\end{aligned}\tag{10.1}$$

Here $\varepsilon = 1/Re$, where Re is the *Reynolds number* defined by $Re = \bar{u}\bar{l}/\tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is the kinematic viscosity, \bar{u} is a typical velocity magnitude, and \bar{l} is a typical length scale.

Even if most flow problems of interest are stated in three space dimensions, we shall here limit ourselves to problems in two space dimensions. The reason is that the notation is simplified, and the principles for construction of the numerical method can quite easily be generalized to 3-D.

In Cartesian coordinates, the equations are

$$\begin{aligned}u_t + p_x &= \varepsilon(u_{xx} + u_{yy}), \\ v_t + p_y &= \varepsilon(v_{xx} + v_{yy}), \\ u_x + v_y &= 0.\end{aligned}\tag{10.2}$$

By differentiation of the first equation with respect to x and the second equation with respect to y and then adding the two equations, we obtain the alternative formulation

$$\begin{aligned} u_t + p_x &= \varepsilon(u_{xx} + u_{yy}), \\ v_t + p_y &= \varepsilon(v_{xx} + v_{yy}), \\ p_{xx} + p_{yy} &= 0. \end{aligned} \quad (10.3)$$

This form is a basis for many numerical methods, since p occurs explicitly also in the third equation. Here we shall stick to the first form for reasons that will be clear later.

For the purpose of analysis, we assume first that the solutions to these equations are 2π -periodic in the y -direction, and consider the domain $\{0 \leq x \leq 1, 0 \leq y \leq 2\pi\}$. The standard type of boundary conditions is specification of the velocity components. We shall modify these slightly to the form

$$\begin{aligned} u(0, y, t) - \frac{1}{2\pi} \int_0^{2\pi} u(0, y, t) dy &= w_L(y, t), \quad u(1, y, t) = u_R(y, t), \\ v(0, y, t) &= v_L(y, t), \quad v(1, y, t) = v_R(y, t), \\ \int_0^{2\pi} p(0, y, t) dy &= q_L(t), \end{aligned} \quad (10.4)$$

where it is assumed that $\int_0^{2\pi} w_L(y, t) dy = 0$. We shall make a few comments on the form of the boundary conditions. Assume that the conditions on u are substituted by the simpler conditions

$$u(0, y, t) = u_L(y, t), \quad u(1, y, t) = u_R(y, t)$$

under the restriction

$$\int_0^{2\pi} u_L(y, t) dy = \int_0^{2\pi} u_R(y, t) dy.$$

The latter condition results from an integration of the divergence condition $u_x + v_y = 0$ together with periodicity in the y -direction. Let us next introduce a perturbation in the equation such that

$$\int_0^{2\pi} u_L(y, t) dy = \int_0^{2\pi} u_R(y, t) dy + \delta, \quad \delta \neq 0.$$

For the analysis we Fourier transform the equations in the y -direction, and study the Fourier coefficients $\hat{u}(x, \omega, t)$. For $\omega = 0$, the divergence condition simplifies to

$$\hat{u}_x(x, 0, t) = 0$$

with the boundary conditions

$$\hat{u}(0, 0, t) = \hat{u}_R(0, t) + \delta, \quad \hat{u}(1, 0, t) = \hat{u}_R(0, t),$$

and obviously no solution exists.

Next we introduce a perturbation in the version (10.4), such that $w_L(y, t)$ is substituted by $w_L(y, t) + \delta(y, t)$, i.e.,

$$\int_0^{2\pi} w_L(y, t) dy = - \int_0^{2\pi} \delta(y, t) dy.$$

When Fourier transforming the problem, we have as before

$$\hat{u}_x(x, 0, t) = 0$$

for $\omega = 0$. However, for the transformed boundary conditions at $x = 0$, the coefficient $\hat{u}(0, 0, t)$ disappears, and we are left with the single condition

$$\hat{u}(1, 0, t) = \hat{u}_R(0, t),$$

which allows for a unique solution.

The significance of this analysis becomes apparent when constructing the numerical method. With the form (10.4) of boundary conditions, we have a sound basis for obtaining a nonsingular and well conditioned discrete system of algebraic equations.

We shall now state a theorem giving an estimate of the solution in terms of the given data, and we prescribe the initial condition

$$\mathbf{w}(x, y, 0) = \mathbf{f}(x, y). \quad (10.5)$$

With the variable t omitted, the norms are defined by

$$\begin{aligned} \|\mathbf{w}\|^2 &= \int_0^{2\pi} \int_0^1 |\mathbf{w}(x, y)|^2 dx dy, \quad |\mathbf{w}|^2 = u^2 + v^2, \\ \|\mathbf{w}_B\|^2 &= \int_0^{2\pi} |\mathbf{w}_B(y)|^2 dy, \quad |\mathbf{w}_B|^2 = |w_L|^2 + |u_R|^2 + |v_L|^2 + |v_R|^2, \\ \|p\|^2 &= \int_0^{2\pi} \int_0^1 |p(x, y)|^2 dx dy, \end{aligned}$$

and one can prove

Theorem 10.1. *Assume that the boundary data w_L, u_R, v_L, v_R are 2π -periodic in y and $\int_0^{2\pi} w_L(y, t) dy = 0$. On any finite time interval $0 \leq t \leq T$, the solution of the system (10.2) with boundary conditions (10.4) satisfies*

$$\begin{aligned} \|\mathbf{w}\|^2 &\leq K_1 (\|\mathbf{f}\| + \max_t |q_L|^2 + \max_t \|\mathbf{w}_B\|^2), \\ \|p\|^2 &\leq K_2 (\|\mathbf{f}\| + \max_t |q_L|^2 + \max_t \|\mathbf{w}_B\|^2 + \max_t \|\frac{\partial \mathbf{w}_B}{\partial y}\|^2), \end{aligned}$$

where K_1 and K_2 are constants depending only on T . □

In general one can expect that the velocity components are more regular than the pressure. The estimate of $\|\mathbf{w}\|$ requires that the boundary data are in L_2 , while the estimate of $\|p\|$ requires that the derivative of the boundary velocity data are in L_2 as well.

There are many more ways of prescribing boundary conditions, that lead to well posed problems, see [Gustafsson and Nilsson, 2002], and we come back to this later.

Later we shall also consider the nonperiodic case, where boundary conditions are prescribed also in the y -direction.

10.3 A Fourth Order Method for Stokes Equations

The presence of the divergence condition is a complication, since no time derivative of p occurs in the equations. There are several ways of avoiding this complication. One way is to introduce an artificial pressure time derivative in the third equation. Another way is to use the pressure form (10.3) instead of the original form. The advantage is that two stage numerical methods can be constructed based on this form. In the first stage the first two equations are advanced by some time stepping scheme. In the second stage, the Laplace equation is solved for the pressure. However, this requires boundary conditions for p , which introduces an extra difficulty. To avoid this, we shall stick with the original form, which also has certain other advantages.

The structure of the equations allows for a compact approximation if we use a staggered grid, see Figure 10.1.

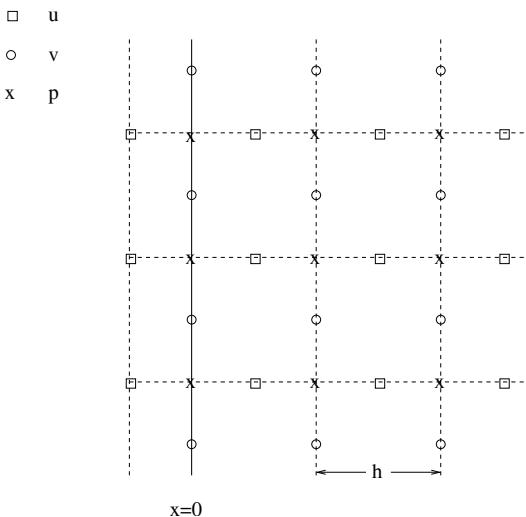


Fig. 10.1 The staggered (x, y) -grid for the Stokes equations

The second order approximation in space is immediately obtained by centering each equation properly. The first derivatives to be approximated are u_x, v_y, p_x, p_y , and each approximation is centered in the middle between two grid points for the corresponding variable. The second derivatives are centered at the grid points. Note that no extra boundary conditions are required for this approximation if u and v are specified at the boundaries. Referring to Figure 10.1, the v -points are located

on the boundary at $x = 0$, and the leftmost u -points one half step to the left of the boundary. The boundary values for v are prescribed for $v_{0,j}$, while the values for u are prescribed for $(u_{-1/2,j} + u_{1/2,j})/2$. For nonperiodic problems, the same principle is applied in the y -direction. The approximations are then well defined for all three differential equations, and no boundary condition on p is required.

We shall now construct a fourth order scheme based on Padé approximations. If h represents the step size, the fourth order formulas are

$$\begin{aligned}\frac{1}{24}f'_{j-1} + \frac{11}{12}f'_j + \frac{1}{24}f'_{j+1} &= \frac{1}{h}(f_{j+1/2} - f_{j-1/2}), \\ \frac{1}{10}f''_{j-1} + f''_j + \frac{1}{10}f''_{j+1} &= \frac{6}{5h^2}(f_{j+1} - 2f_j + f_{j-1}).\end{aligned}$$

For the implementation, it is necessary to impose boundary conditions such that we obtain closed systems of the form $Pf' = Qf$ and $Rf'' = Sf$, where P, Q, R, S are square matrices. Even if the formulas at inner points are no wider than the second order ones, one should note that three points of the derivative f' are coupled to each other. Hence, even if the physical boundary conditions are used for f , extra numerical boundary conditions on f' are needed. In analogy with the SBP operators, it is convenient for the implementation if f' can be computed independently, and then enforce the physical boundary conditions. Therefore we construct numerical boundary conditions also for f . Since P represents a tridiagonal matrix, we shouldn't have more than two nonzero elements in the first row. We have to distinguish between the forms of P and Q for p on one hand and u and v on the other hand. Consider the points near the boundary $x = 0$ in Figure 10.1. The function p is represented at x_0, x_1, x_2 , while p_x is represented at $x_{-1/2}, x_{1/2}$. On the other hand, the corresponding points for u are $x_{-1/2}, x_{1/2}, x_{3/2}$, while for u_x they are x_0, x_1 . In the y -direction, we have the same situation for p and v .

The extra conditions are constructed with 3rd order accuracy. When including also the right boundary, the complete one-dimensional formulas are

$$(P_1 f')_{j-1/2} = \begin{cases} \frac{1}{552}(f'_{-1/2} + 23f'_{1/2}), & j = 0, \\ \frac{1}{24}(f'_{j-1/2} + 22f'_{j+1/2} + f'_{j+3/2}), & j = 0, 1, \dots, N, \\ \frac{1}{552}(23f'_{N-1/2} + f'_{N+1/2}), & j = N + 1, \end{cases}$$

$$(Q_1 f)_{j-1/2} = \begin{cases} \frac{1}{552h}(-25f_0 + 26f_1 - f_2), & j = 0, \\ \frac{1}{h}(f_{j+1} - f_j), & j = 0, 1, \dots, N, \\ \frac{1}{552h}(f_{N-2} - 26f_{N-1} + 25f_N), & j = N + 1, \end{cases}$$

for $f = p$, and

$$(P_2 f')_j = \begin{cases} \frac{1}{24}(-f'_0 + f'_1), & j=0, \\ \frac{1}{24}(f'_{j-1} + 22f'_j + f'_{j+1}), & j=1, 2, \dots, N-1, \\ \frac{1}{24}(f'_{N-2} - f'_N), & j=N, \end{cases}$$

$$(Q_2 f)_j = \begin{cases} \frac{1}{24h}(f_{-1/2} - 2f_{1/2} + f_{3/2}), & j=0, \\ \frac{1}{h}(f_{j+1/2} - f_{j-1/2}), & j=1, 2, \dots, N-1, \\ \frac{1}{24h}(-f_{N-3/2} + 2f_{N-1/2} - f_{N+1/2}), & j=N, \end{cases}$$

for $f = u$ and $f = v$.

For the second derivatives there are two alternatives. One is to use Padé approximations directly derived for f'' , the other one is to apply the first derivative approximations $P^{-1}Q$ twice. For the first alternative, we want to limit the number of nonzero elements in the first row of the left hand matrix R to two in order to keep the band width small. Then five points are required on the right hand side Sf to keep up the accuracy:

$$(Rf'')_j = \begin{cases} \frac{1}{100}(f''_0 + 10f''_1), & j=0, \\ \frac{1}{12}(f''_{j-1} + 10f''_j + f''_{j+1}), & j=1, 2, \dots, N, \\ \frac{1}{100}(10f''_{N-1} + f''_N), & j=N+1, \end{cases}$$

$$(Sf)_j = \begin{cases} \frac{1}{1200h^2}(145f_0 - 304f_1 + 174f_2 - 16f_3 + f_4), & j=0, \\ \frac{1}{h^2}(f_{j-1} - 2f_j + f_{j+1}), & j=1, 2, \dots, N, \\ \frac{1}{1200h^2}(f_{N-4} - 16f_{N-3} + 174f_{N-2} - 304f_{N-1} + 145f_N), & j=N+1. \end{cases}$$

These formulas are applied at integer points. Looking back at the staggered grid in Figure 10.1, we see that they should be used in the y -direction for u , and in the x -direction for v . For u in the x -direction and v in the y -direction, the same formulas are used, but now applied at half-points. We denote these operators by \tilde{R} and \tilde{S} .

We recall from Chapter 3 that the physical boundary conditions must be approximated to the same order as the differential equations themselves. Therefore the boundary conditions on u must be approximated to fourth order. For the function itself we use the average defined by four points. For the integral we have

$$\int_{y_0}^{y_M} f(y)dy = h\left(\frac{f(y_0)}{2} + f(y_1) + \dots + f(y_{M-1}) + \frac{f(y_M)}{2}\right) - \frac{h^2}{12}(f'(y_M) - f'(y_0)) + \mathcal{O}(h^4).$$

When approximating the derivatives f' with second order one-sided formulas, we obtain a fourth order approximation $I(f)$ of the integral. This leads to the following formulas to be used in the boundary conditions:

$$\begin{aligned} f_0 &= \frac{1}{16}(5f_{-1/2} + 15f_{1/2} - 5f_{3/2} + f_{5/2}), \\ f_N &= \frac{1}{16}(5f_{N+1/2} + 15f_{N-1/2} - 5f_{N-3/2} + f_{-5/2}), \\ I(f) &= h\left(\frac{f_0}{2} + f_1 + \dots + f_{M-1} + \frac{f_M}{2}\right) - \frac{h}{24}(3f_0 - 4f_1 + f_2 + 3f_M - 4f_{M-1} + f_{M-2}). \end{aligned}$$

The space discretizations are defined in one space dimension above, and we introduce the subscripts x and y to indicate the coordinate directions. We now have the following approximations

$$\begin{aligned} p_x &\approx P_{1x}^{-1}Q_{1x}p \text{ at all points } (x_{i-1/2}, y_j), \\ p_y &\approx P_{1y}^{-1}Q_{1y}p \text{ at all points } (x_i, y_{j-1/2}), \\ u_x &\approx P_{2x}^{-1}Q_{2x}u \text{ at all points } (x_{i-1/2}, y_j), \\ v_y &\approx P_{2y}^{-1}Q_{2y}v \text{ at all points } (x_i, y_{j-1/2}), \\ u_{xx} &\approx \tilde{R}_x^{-1}\tilde{S}_xu \text{ at all points } (x_{i-1/2}, y_j), \\ u_{yy} &\approx R_y^{-1}S_yu \text{ at all points } (x_{i-1/2}, y_j), \\ v_{xx} &\approx R_x^{-1}S_xv \text{ at all points } (x_i, y_{j-1/2}), \\ v_{yy} &\approx \tilde{R}_y^{-1}\tilde{S}_yv \text{ at all points } (x_i, y_{j-1/2}). \end{aligned} \tag{10.6}$$

The approximations (10.6) are well defined for u and v at all grid points. This allows for advancement of the scheme without using the physical boundary conditions, but this possibility can of course not be used. We have simply included the extra points such that it is possible to compute the derivatives independent of the specific problem. The physical boundary conditions are implemented afterwards for each time step.

For the time discretization we use a regular grid without staggering, and here we stay with the standard second order backward differentiation BDF. The approximation of (10.1) is

$$\begin{aligned} \frac{3}{2}\mathbf{w}^{n+1} + k(\nabla p^{n+1} - \varepsilon \Delta \mathbf{w}^{n+1}) &= 2\mathbf{w}^n - \frac{1}{2}\mathbf{w}^{n-1}, \\ \nabla \cdot \mathbf{w}^{n+1} &= 0, \end{aligned}$$

i.e., on a 2-D Cartesian grid (subscripts omitted)

$$\begin{aligned} \frac{3}{2}u^{n+1} + k(P_{1x}^{-1}Q_{1x}p^{n+1} - \varepsilon(\tilde{R}_x^{-1}\tilde{S}_x + R_y^{-1}S_y)u^{n+1}) &= 2u^n - \frac{1}{2}u^{n-1}, \\ \frac{3}{2}v^{n+1} + k(P_{1y}^{-1}Q_{1y}p^{n+1} - \varepsilon(R_x^{-1}S_x + \tilde{R}_y^{-1}\tilde{S}_y)v^{n+1}) &= 2v^n - \frac{1}{2}v^{n-1}, \\ P_{2x}^{-1}Q_{2x}u^{n+1} + P_{2y}^{-1}Q_{2y}v^{n+1} &= 0. \end{aligned} \tag{10.7}$$

In the next section we shall describe an effective iterative method for solving this system. Here, we present the results for a simple numerical experiment, where a standard GMRES solver was used for the algebraic system.

The test is a straight channel flow problem. The domain is $\{0 \leq x \leq 1, -1 \leq y \leq 1\}$ with initial conditions

$$\begin{aligned} u(x, y, 0) &= U(y)e^{\alpha x}, \\ v(x, y, 0) &= V(y)e^{\alpha x}. \end{aligned}$$

In the analysis above the solutions were 2π -periodic in the y -direction, but here we consider the nonperiodic case, and prescribe the boundary conditions

$$\begin{aligned} u(0, y, t) - \frac{1}{2} \int_{-1}^1 u(0, y, t) dy &= (U(y) - \frac{1}{2} \int_{-1}^1 U(y) dy) e^{-\omega t}, \\ v(0, y, t) &= V(y)e^{-\omega t}, \end{aligned}$$

$$\begin{aligned} u(1, y, t) &= U(y)e^{\alpha - \omega t}, \\ v(1, y, t) &= V(y)e^{\alpha - \omega t}, \end{aligned}$$

$$\begin{aligned} u(x, -1, t) &= 0, & u(x, 1, t) &= 0, \\ v(x, -1, t) &= 0, & v(x, 1, t) &= 0. \end{aligned}$$

Here

$$\begin{aligned} U(y) &= c_1 \sin(\alpha y) + \frac{2\kappa}{\alpha} c_2 \sin(\kappa y), \\ V(y) &= c_1 \cos(\alpha y) + 2c_2 \cos(\kappa y), \\ P(y) &= \frac{\omega}{\alpha} c_1 \sin(\alpha y), \end{aligned}$$

where

$$\kappa = \frac{1}{\varepsilon} \sqrt{\varepsilon^2 + \varepsilon \omega \alpha^2}.$$

The solution to the problem is

$$\begin{aligned} u(x, y, t) &= U(y)e^{\alpha x - \omega t}, \\ v(x, y, t) &= V(y)e^{\alpha x - \omega t}, \\ p(x, y, t) &= P(y)e^{\alpha x - \omega t}. \end{aligned}$$

For a real life problem, the first time level has to be computed by a one-step scheme in order to get the main scheme started. Here, however, we use the true solution at $t = k$ as a second set of initial data.

The fourth order interpolation is used for u at the boundaries $x = 0$ and $x = 1$, and for v at the boundaries $y = -1$ and $y = 1$.

The numerical values

$$\begin{aligned}\varepsilon &= 1, \\ \alpha &= 1, \\ \omega &= 11.6347883720355431, \\ c_1 &= 0.9229302839450678, \\ c_2 &= 0.2722128679701572\end{aligned}$$

are used. The error is measured in the l_2 -norm for each variable at the end of the integration $t = T$. Table 10.1 shows the error on an $N \times N$ grid, for $N = 20$ and $N = 40$.

Table 10.1 The l_2 -error in u, v, p

N	Variable	T=0.1	T=0.2	T=0.3	T=0.4	T=0.5
20	u	$2.3 \cdot 10^{-5}$	$7.1 \cdot 10^{-6}$	$2.2 \cdot 10^{-6}$	$7.0 \cdot 10^{-7}$	$2.2 \cdot 10^{-7}$
	v	$1.8 \cdot 10^{-4}$	$5.7 \cdot 10^{-5}$	$1.8 \cdot 10^{-5}$	$5.6 \cdot 10^{-6}$	$1.7 \cdot 10^{-6}$
	p	$8.5 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$	$8.3 \cdot 10^{-5}$	$2.6 \cdot 10^{-5}$	$8.1 \cdot 10^{-6}$
40	u	$1.4 \cdot 10^{-6}$	$4.4 \cdot 10^{-7}$	$1.4 \cdot 10^{-7}$	$4.3 \cdot 10^{-8}$	$1.3 \cdot 10^{-8}$
	v	$9.9 \cdot 10^{-6}$	$3.1 \cdot 10^{-6}$	$9.6 \cdot 10^{-7}$	$3.0 \cdot 10^{-7}$	$9.4 \cdot 10^{-8}$
	p	$5.8 \cdot 10^{-5}$	$1.9 \cdot 10^{-5}$	$5.7 \cdot 10^{-6}$	$1.8 \cdot 10^{-6}$	$5.6 \cdot 10^{-7}$

10.4 Navier–Stokes Equations for Incompressible Flow

With the notation

$$\mathcal{N}(\mathbf{w}) = (\mathbf{w} \cdot \nabla) \mathbf{w}, \quad \mathcal{L}(\mathbf{w}, p) = \nabla p - \varepsilon \Delta \mathbf{w},$$

the Navier–Stokes equations for incompressible flow can be written as

$$\begin{aligned}\mathbf{w}_t + \mathcal{N}(\mathbf{w}) + \mathcal{L}(\mathbf{w}, p) &= 0, \\ \nabla \cdot \mathbf{w} &= 0.\end{aligned}\tag{10.8}$$

Also here we limit ourselves to problems in two space dimensions. In Cartesian coordinates, the equations are

$$\begin{aligned}u_t + uu_x + vu_y + p_x &= \varepsilon(u_{xx} + u_{yy}), \\ v_t + uv_x + vv_y + p_y &= \varepsilon(v_{xx} + v_{yy}), \\ u_x + v_y &= 0.\end{aligned}\tag{10.9}$$

We shall first consider flow in a straight channel $\{0 \leq x \leq 1, -1 \leq y \leq 1\}$ with solid walls at $y = -1$ and $y = 1$, with the conditions $u = v = 0$ applied there. In

the x -direction the same boundary conditions as for the Stokes equations can be used. However, in order to optimize the solver of the implicit part, it is convenient to change one of the conditions slightly and use

$$\begin{aligned} u(0,y,t) - \frac{1}{2} \int_{-1}^1 u(0,y,t) dy &= w_L(y,t), \quad u(1,y,t) = u_R(y,t), \\ v(0,y,t) &= v_L(y,t), \quad v(1,y,t) = v_R(y,t), \\ \int_{-1}^1 p(0,y,t) dy + \int_{-1}^1 u(0,y,t) dy &= q_L(t), \end{aligned} \quad (10.10)$$

where the last condition is the modified one.

Many problems have open boundaries where no velocity data are available. Most often this is the case when the fluid is passing out through the boundary, and the problem is to construct “outflow boundary conditions” or “downstream boundary conditions”. This is a fundamental problem that has generated much research during the last decades, but it is outside the scope of this book. One common method is to prescribe zero derivatives, which is a good approximation if the boundary is sufficiently far downstream. For our model problem we suggest the following set of conditions which will be used in the numerical experiments to be presented in the next section:

$$\begin{aligned} u(0,y,t) &= u_L(y,t), \quad u_x(1,y,t) - \frac{1}{2} \int_{-1}^1 u_x(1,y,t) dy = 0, \\ v(0,y,t) &= v_L(y,t), \quad v_x(1,y,t) = 0, \\ \int_{-1}^1 p(1,y,t) dy + \int_{-1}^1 u_x(1,y,t) dy &= q_R(t). \end{aligned} \quad (10.11)$$

Nonrectangular domains

For fluid problems, we must keep up the formal accuracy at the boundaries. The type of embedding technique that was described for wave propagation problems on non-rectangular domains does not work well here. One reason is that for high Reynolds numbers Re , the fluid exhibits boundary layers, i.e., there are sharp gradients near the boundaries, and small perturbations in the geometry has fundamental effects on the flow. For problems with curved boundaries, uniform rectangular grids cannot be used. There are several techniques for overcoming this difficulty. One of them is the use of overlapping grids. The idea is to construct a body fitted grid along the boundary, another Cartesian grid in the rest of the domain, and then to connect the two grids via interpolation. There are several software packages available for these methods.

Here we shall describe in some detail the simpler case, where a transformation is used for the whole domain, and furthermore we shall use orthogonal transformations. This technique can be applied only for simple model problems, but it contains several important parts of interest for the general case.

A coordinate transformation changes the differential equations. Let the transformation be defined by

$$x = x(\xi, \eta), \quad y = y(\xi, \eta),$$

where ξ and η are the new independent variables. The new pressure is defined by

$$\tilde{p}(\xi, \eta) = p(x(\xi, \eta), y(\xi, \eta)),$$

and we have

$$\begin{aligned}\tilde{p}_\xi(\xi, \eta) &= p_\xi(x(\xi, \eta), y(\xi, \eta)) = p_x x_\xi + p_y y_\xi, \\ \tilde{p}_\eta(\xi, \eta) &= p_\eta(x(\xi, \eta), y(\xi, \eta)) = p_x x_\eta + p_y y_\eta.\end{aligned}$$

We can now solve for p_x and p_y , and then substitute those expressions in the differential equations. One step further in the differentiation process gives us the second derivatives p_{xx} , p_{xy} , p_{yy} . This is a standard procedure, but when it comes to the velocity components, we have to make a choice. One possibility is to keep the old variables u and v that represent the velocity components in the x - and y -direction (global formulation). The other option is to introduce new velocity components, that in contrast to the original ones, represent the velocity in the ξ - and η -direction (local formulation). We shall choose the second option, and denote the local components by \tilde{u} , \tilde{v} . One advantage with this choice is that \tilde{u} and \tilde{v} still represent the perpendicular and tangential velocity components at the boundaries.

With the scale factors defined by

$$\begin{aligned}n_1 &= \sqrt{x_\xi^2 + y_\xi^2}, \\ n_2 &= \sqrt{x_\eta^2 + y_\eta^2},\end{aligned}$$

we get for the new coordinate system

$$\begin{aligned}\nabla \tilde{p} &= \left[\begin{array}{c} \frac{1}{n_1} \tilde{p}_\xi \\ \frac{1}{n_2} \tilde{p}_\eta \end{array} \right], \\ \nabla \cdot \mathbf{w} &= \frac{1}{n_1 n_2} ((n_2 \tilde{u})_\xi + (n_1 \tilde{v})_\eta).\end{aligned}$$

Unfortunately, the Laplacian takes a more complicated form. The two components are

$$\begin{aligned}
(\Delta \mathbf{w})^{(1)} &= \frac{1}{n_1 n_2} \left(\left(\frac{n_2}{n_1} \tilde{u}_\xi \right)_\xi + \left(\frac{n_1}{n_2} \tilde{u}_\eta \right)_\eta - \frac{(n_1)_\eta^2 + (n_2)_\xi^2}{n_1 n_2} \tilde{u} \right. \\
&\quad \left. + \frac{2(n_1)_\eta \tilde{v}_\xi}{n_1} - \frac{2(n_2)_\xi \tilde{v}_\eta}{n_2} + \left(\frac{(n_1)_\eta}{n_1} - \frac{(n_2)_\eta}{n_2} \right)_\xi \tilde{v} \right), \\
(\Delta \mathbf{w})^{(2)} &= \frac{1}{n_1 n_2} \left(\left(\frac{n_2}{n_1} \tilde{v}_\xi \right)_\xi + \left(\frac{n_1}{n_2} \tilde{v}_\eta \right)_\eta - \frac{(n_1)_\eta^2 + (n_2)_\xi^2}{n_1 n_2} \tilde{v} \right. \\
&\quad \left. - \frac{2(n_1)_\eta \tilde{u}_\xi}{n_1} + \frac{2(n_2)_\xi \tilde{u}_\eta}{n_2} - \left(\frac{(n_1)_\eta}{n_1} + \frac{(n_2)_\eta}{n_2} \right)_\xi \tilde{u} \right).
\end{aligned}$$

The boundary conditions have the form (10.10), but now in the (ξ, η) space.

10.5 A Fourth Order Method for Navier–Stokes Equations

The numerical method for the Navier–Stokes equations is an extension of the method for the Stokes equations. The idea is to approximate the Stokes part as described in the previous section, and then take the nonlinear advective part in (10.8) explicitly as a second order extrapolation. The scheme is

$$\begin{aligned}
\frac{3}{2} \mathbf{w}^{n+1} + k \mathcal{L}(\mathbf{w}^{n+1}, p^{n+1}) &= 2 \mathbf{w}^n - \frac{1}{2} \mathbf{w}^{n-1} - 2k \mathcal{N}(\mathbf{w}^n) + k \mathcal{N}(\mathbf{w}^{n-1}), \\
\nabla \cdot \mathbf{w}^{n+1} &= 0.
\end{aligned}$$

We first discuss the 2-D equations (10.9) in a straight channel with boundary conditions (10.10). Concerning the numerical boundary conditions for the Padé approximations, we have a case here, which is interesting from a theoretical point of view. The whole system is not parabolic because of the divergence condition, and the theory in chapters 2 and 3 does not apply directly. For large Reynolds numbers it is interesting to consider the limit case $Re \rightarrow \infty$. Again, there is no clean classification of the PDE-system. However, if we add a term εp_t to the third equation, where ε is small corresponding to slightly compressible flow, the linearized system is hyperbolic. Disregarding the y -direction, it represents two fast sound waves, and one advection wave moving with velocity u . In Chapter 3 it was shown that under certain conditions one can relax the order of accuracy one step near the boundary, but that result does not apply in this case. The reason is that $u = 0$ at the boundary, and this case with vertical characteristics was not included in the theoretical discussion. In fact, it was shown in [Ferm and Lötstedt, 2002] that it is necessary to keep the same approximation order near the boundary as at inner points. Accordingly, we construct 4th order noncentered approximations.

When including also the right boundary, the complete one-dimensional formulas are

$$(P_1 f')_{j-1/2} = \begin{cases} \frac{1}{12672}(24f'_{-1/2} + 528f'_{1/2}), & j=0, \\ \frac{1}{24}(f'_{j-1/2} + 22f'_{j+1/2} + f'_{j+3/2}), & j=1,2,\dots,N, \\ \frac{1}{12672}(528f'_{N-1/2} + 24f'_{N+1/2}), & j=N+1, \end{cases}$$

$$(Q_1 f)_{j-1/2} = \begin{cases} \frac{1}{12672h}(-577f_0 + 603f_1 - 27f_2 + f_3), & j=0, \\ \frac{1}{h}(f_{j+1} - f_j), & j=1,2,\dots,N, \\ \frac{1}{12672h}(-f_{N-3} + 27f_{N-2} - 603f_{N-1} + 577f_N), & j=N+1, \end{cases}$$

$$(P_2 f')_j = \begin{cases} \frac{1}{24}f'_1, & j=0, \\ \frac{1}{24}(f'_{j-1} + 22f'_j + f'_{j+1}), & j=1,2,\dots,N-1, \\ \frac{1}{24}f'_N, & j=N, \end{cases}$$

$$(Q_2 f)_j = \begin{cases} \frac{1}{576h}(f_{-1/2} - 27f_{1/2} + 27f_{3/2} - f_{5/2}), & j=0, \\ \frac{1}{h}(f_{j+1/2} - f_{j-1/2}), & j=1,2,\dots,N-1, \\ \frac{1}{576h}(f_{N-5/2} - 27f_{N-3/2} + 27f_{N-1/2} - f_{N+1/2}), & j=N. \end{cases}$$

The advective terms don't fit into the staggered grid structure. Therefore, we need Padé approximations also on regular grids:

$$(P_3 f')_j = \begin{cases} \frac{1}{108}(6f'_0 + 18f'_1), & j=0, \\ \frac{1}{6}(f'_{j-1} + 4f'_j + f'_{j+1}), & j=1,2,\dots,N-1, \\ \frac{1}{108}(18f'_{N-1} + 6f'_N), & j=N, \end{cases}$$

$$(Q_3 f)_j = \begin{cases} \frac{1}{108h}(-17f_0 + 9f_1 + 9f_2 - f_3), & j=0, \\ \frac{1}{2h}(f_{j+1} - f_{j-1}), & j=1,2,\dots,N-1, \\ \frac{1}{108h}(f_{N-3} - 9f_{N-2} - 9f_{N-1} + 17f_N), & j=N. \end{cases}$$

The same formulas hold also for half-points $x_{j+1/2}$ with obvious modifications, and we denote these operators by \tilde{P}_3 and \tilde{Q}_3 .

In the previous section we saw that the Laplacian takes a different form after coordinate transformation. For example, the second derivative u_{xx} gives rise to a term $a(\xi, \eta)(b(\xi, \eta)u_\xi)_\xi$, where $a(\xi, \eta)$ and $b(\xi, \eta)$ are known functions. Therefore, it is appropriate to use Padé approximations of the first derivatives twice, and we use the $P_3^{-1}Q_3$ version above.

We also need fourth order interpolation formulas for u in the term uv_y , and for v in the term vu_y , and we use the Padé approximation

$$\frac{1}{4}f_{j-1}^* + \frac{3}{2}f_j^* + \frac{1}{4}f_{j+1}^* = f_{j-1/2} + f_{j+1/2}$$

for interpolated values f^* at inner points. We call this interpolation operator E , and add a subscript x or y depending on the coordinate direction. Since these operators

are used on the explicit side, all the values are known at the boundaries, and there is no need to construct any one-sided formulas there. The final scheme is

$$\begin{aligned} \frac{3}{2}u^{n+1} + k(P_{1x}^{-1}Q_{1x}p^{n+1} - \varepsilon(\tilde{P}_{3x}^{-1}\tilde{Q}_{3x}\tilde{P}_{3x}^{-1}\tilde{Q}_{3y} + P_{3y}^{-1}Q_{3y}P_{3y}^{-1}Q_{3y})u^{n+1}) \\ = 2(I - ku^n\tilde{P}_{3x}^{-1}\tilde{Q}_{3x} - k(E_xE_yv^n)P_{3y}^{-1}Q_{3y})u^n \\ - \frac{1}{2}(I + ku^{n-1}\tilde{P}_{3x}^{-1}\tilde{Q}_{3x} + k(E_xE_yv^{n-1})P_{3y}^{-1}Q_{3y})u^{n-1}, \\ \frac{3}{2}v^{n+1} + k(P_{1y}^{-1}Q_{1y}p^{n+1} - \varepsilon(P_{3x}^{-1}Q_{3x}P_{3x}^{-1}Q_{3y} + \tilde{P}_{3x}^{-1}\tilde{Q}_{3x}\tilde{P}_{3x}^{-1}\tilde{Q}_{3y})v^{n+1}) \\ = 2(I - k(E_xE_yu^n)P_{3x}^{-1}Q_{3x} - kv^n\tilde{P}_{3y}^{-1}\tilde{Q}_{3y})v^n \\ - \frac{1}{2}(I + k(E_xE_yu^{n-1})P_{3x}^{-1}Q_{3x} + kv^{n-1}\tilde{P}_{3y}^{-1}\tilde{Q}_{3y})v^{n-1}, \\ P_{2x}^{-1}Q_{2x}u^{n+1} + P_{2y}^{-1}Q_{2y}v^{n+1} = 0, \end{aligned} \quad (10.12)$$

with the approximation of the physical boundary conditions (10.10) imposed after each completed time step.

The von Neumann condition

We shall now investigate the von Neumann condition for the approximation of the constant coefficient linearized Navier–Stokes equations

$$\begin{aligned} u_t + au_x + bu_y + p_x &= \varepsilon(u_{xx} + u_{yy}), \\ v_t + av_x + bv_y + p_y &= \varepsilon(v_{xx} + v_{yy}), \\ u_x + v_y &= 0 \end{aligned}$$

with periodic solutions in both space directions. The difference operators in (10.12) are modified by removing the special formulas at the boundaries, such that they fit the periodic solutions.

The standard second order explicit operators are obtained by simply putting $P_1 = P_2 = P_3 = I$, and this allows for a general stability analysis including both the second and fourth order operators. With the parameters

$$\begin{aligned} \xi_1 &= \omega_1 h_1, \quad \xi_2 = \omega_2 h_2, \\ \lambda_1 &= \frac{k}{h_1}, \quad \lambda_2 = \frac{k}{h_2}, \end{aligned}$$

the Fourier transform of the approximation is

$$\begin{aligned}
& \frac{3}{2}\hat{u}^{n+1} - 2\hat{u}^n + \frac{1}{2}\hat{u}^{n-1} + a\lambda_1 ir_1(2\hat{u}^n - \hat{u}^{n-1}) + b\lambda_2 ir_2(2\hat{u}^n - \hat{u}^{n-1}) \\
& \quad + 2\lambda_1 is_1 \hat{p}^{n+1} + 4\varepsilon k \left(\frac{d_1}{h_1^2} + \frac{d_2}{h_2^2} \right) \hat{u}^{n+1} = 0, \\
& \frac{3}{2}\hat{v}^{n+1} - 2\hat{v}^n + \frac{1}{2}\hat{v}^{n-1} + a\lambda_1 ir_1(2\hat{v}^n - \hat{v}^{n-1}) + b\lambda_2 ir_2(2\hat{v}^n - \hat{v}^{n-1}) \quad (10.13) \\
& \quad + 2\lambda_2 is_1 \hat{p}^{n+1} + 4\varepsilon k \left(\frac{d_1}{h_1^2} + \frac{d_2}{h_2^2} \right) \hat{v}^{n+1} = 0, \\
& \frac{1}{h_1} is_1 \hat{u}^{n+1} + \frac{1}{h_2} is_2 \hat{v}^{n+1} = 0.
\end{aligned}$$

Here the parameters are defined by

$$\begin{aligned}
s_1 &= \sin \frac{\xi_1}{2}, \quad s_2 = \sin \frac{\xi_2}{2}, \\
r_1 &= \sin \xi_1, \quad r_2 = \sin \xi_2, \\
d_1 &= \sin^2 \frac{\xi_1}{2}, \quad d_2 = \sin^2 \frac{\xi_2}{2}
\end{aligned} \quad (10.14)$$

in the second order case, and by

$$\begin{aligned}
s_1 &= \frac{12 \sin(\xi_1/2)}{11 + \cos \xi_1}, \quad s_2 = \frac{12 \sin(\xi_2/2)}{11 + \cos \xi_2}, \\
r_1 &= \frac{3 \sin \xi_1}{2 + \cos \xi_1}, \quad r_2 = \frac{3 \sin \xi_2}{2 + \cos \xi_2}, \\
d_1 &= \frac{6 \sin^2(\xi_1/2)}{5 + \cos \xi_1}, \quad d_2 = \frac{6 \sin^2(\xi_2/2)}{5 + \cos \xi_2}
\end{aligned} \quad (10.15)$$

in the fourth order case.

We solve first the last equation for \hat{v}^{n+1} assuming $\xi_2 \neq 0$. With $\hat{\mathbf{w}} = (\hat{u} \ \hat{p})^T$, we get a system of difference equations in time

$$\begin{bmatrix} \gamma & 2i\lambda_1 s_1 \\ \gamma \frac{s_1}{s_2} - \frac{2i\lambda_2 s_2}{\lambda_1} \end{bmatrix} \hat{\mathbf{w}}^{n+1} + \begin{bmatrix} \alpha & 0 \\ \alpha \frac{s_1}{s_2} & 0 \end{bmatrix} \hat{\mathbf{w}}^n + \begin{bmatrix} \beta & 0 \\ \beta \frac{s_1}{s_2} & 0 \end{bmatrix} \hat{\mathbf{w}}^{n-1} = 0,$$

where

$$\begin{aligned}
\alpha &= -2 + 2i(a\lambda_1 r_1 + b\lambda_2 r_2), \\
\beta &= \frac{1}{2} - i(a\lambda_1 r_1 + b\lambda_2 r_2), \\
\gamma &= \frac{3}{2} + 4\varepsilon k \left(\frac{d_1}{h_1^2} + \frac{d_2}{h_2^2} \right).
\end{aligned}$$

Referring back to Section 2.2.2, the amplification factor z satisfies

$$\operatorname{Det} \begin{bmatrix} \gamma z^2 + \alpha z + \beta & 2i\lambda_1 s_1 z^2 \\ \frac{s_1}{s_2}(\gamma z^2 + \alpha z + \beta) - \frac{2i\lambda_2^2 s_2}{\lambda_1} z^2 \end{bmatrix} = -2iz^2 q(z) \frac{\lambda_1^2 s_1^2 + \lambda_2^2 s_2^2}{\lambda_1 s_2} = 0,$$

where

$$q(z) = \gamma z^2 + \alpha z + \beta.$$

By assumption, $s_2 \neq 0$, which means that the last factor is nonzero. Accordingly, the roots are given by $z_{1,2} = 0$ together with the solutions to the scalar quadratic equation $q(z) = 0$.

The assumption $s_2 \neq 0$ is no restriction. If instead $s_1 \neq 0$, then we just switch the directions, and solve first for \hat{u}^{n+1} instead of \hat{v}^{n+1} . The trivial case $s_1 = s_2 = 0$ produces the roots $z_1 = 1$, $z_2 = 1/3$, and is of no concern.

The fourth order case is included in the derivation above, we just have to remember the different parameter definitions (10.14) and (10.15).

The quadratic equation $q(z) = 0$ can now be solved. If we define $\theta = \varepsilon\lambda/h$ for $h = h_1 = h_2$, the contours in Figure 10.2 are the outer boundaries $|z| = 1$ of the stability domains for different values of θ as a function of $a\lambda$ and $b\lambda$ for the 2nd and 4th order schemes.

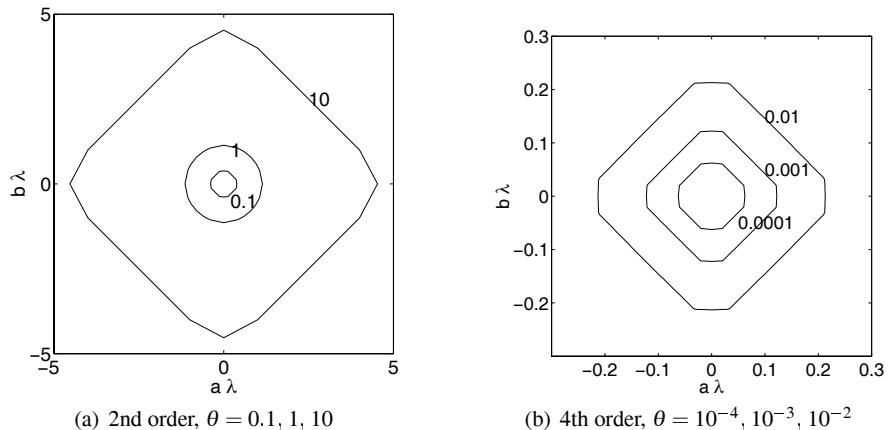


Fig. 10.2 Stability domains for (10.12)

The time step restriction is more severe in the 4th order case. In [Kress and Lötstedt, 2006], other time discretizations are considered, leading to relaxed conditions on the time step.

Most presentations in the literature on Navier–Stokes solvers limit the Fourier analysis to the scalar advection diffusion equation

$$\phi_t + a\phi_x + b\phi_y = \varepsilon(\phi_{xx} + \phi_{yy}). \quad (10.16)$$

For the difference schemes considered here, this is actually correct. The equation $q(z) = 0$ is

$$\begin{aligned} \left(\frac{3}{2} + 4\epsilon k \left(\frac{d_1}{h_1^2} + \frac{d_2}{h_2^2}\right)\right) \hat{\phi}^{n+1} + (-2 + 2a\lambda_1 ir_1 + 2b\lambda_2 ir_2) \hat{\phi}^n \\ + \left(\frac{1}{2} - a\lambda_1 ir_1 - b\lambda_2 ir_2\right) \hat{\phi}^{n-1} = 0, \end{aligned}$$

which is exactly the Fourier transform of the discrete form of (10.16) for both the second and fourth order case.

Let us next discuss so called *parasitic solutions*. For convenience, we limit ourselves to the steady state equations, and begin by the trivial problem $d\phi/dx = 0$. The solution of the standard difference approximation $D_0\phi_j = 0$ is

$$\phi_j = \sigma_1 + \sigma_2(-1)^j,$$

where the constants σ_1 and σ_2 are to be determined by the boundary conditions. On the other hand, the compact approximation $D_+\phi_j = 0$ has the solution

$$\phi_j = \sigma_1.$$

The oscillating component $(-1)^j$ for the first approximation is the parasitic solution, and it is completely disconnected from the differential equation. If the extra numerical boundary condition is formulated right, the coefficient σ_2 is $\mathcal{O}(h^2)$. This means that the parasitic solution is part of the error with the right order, but it is usually considered as a troublemaker.

Another way of characterizing such a solution is to consider the Fourier transformed equation

$$(\sin \xi) \hat{\phi}(\xi) = 0.$$

Nontrivial solutions $\hat{\phi}(\xi)$ are obtained for $\xi = 0$ and $\xi = \pi$, where $\hat{\phi}(\pi)e^{j\pi} = \hat{\phi}(\pi)(-1)^j = \sigma_2(-1)^j$ is the parasitic solution.

Turning to our problem, one could possibly expect parasitic solutions, since the approximation contains noncompact operators for the advection terms. We look for nontrivial solutions of the Fourier transformed steady state system obtained from (10.13)

$$C(\xi, \eta) \begin{bmatrix} \hat{u} \\ \hat{v} \\ \hat{p} \end{bmatrix} = 0, \quad C(\xi, \eta) = \begin{bmatrix} \eta & 0 & 2\lambda_1 i \sin \frac{\xi_1}{2} \\ 0 & \eta & 2\lambda_2 i \sin \frac{\xi_2}{2} \\ \lambda_1 i \sin \frac{\xi_1}{2} & \lambda_2 i \sin \frac{\xi_2}{2} & 0 \end{bmatrix},$$

where

$$\eta = a\lambda_1 i \sin \xi_1 + b\lambda_2 i \sin \xi_2 + 4\epsilon \left(\frac{\lambda_1 \sin^2(\xi_1/2)}{h_1} + \frac{\lambda_2 \sin^2(\xi_2/2)}{h_2} \right).$$

The determinant is

$$\text{Det } C(\xi, \eta) = 2\eta(\lambda_1 \sin^2 \frac{\xi_1}{2} + \lambda_2 \sin^2 \frac{\xi_2}{2}).$$

This expression is zero only if $\xi_1 = \xi_2 = 0$, which shows that there is no parasitic solution for any one of the two approximations.

Iterative solution of the system of equations

Here we shall describe the solution procedure for the system of equations to be solved for each time step. We separate the inner and outer velocity points, and denote them by \mathbf{w} and \mathbf{w}_B respectively. With p denoting the vector containing all the pressure points, the final system of equations has the form

$$\begin{bmatrix} A_0 & A_1 & G_0 \\ A_2 & C & B_0 \\ D_0 & D_1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_B \\ p \end{bmatrix} = \begin{bmatrix} b \\ b_B \\ 0 \end{bmatrix}.$$

Here the first row represents the first two equations in (10.12), where the notation G_0 refers to the gradient. The second row corresponds to the physical boundary conditions, and the last row is the divergence condition. The matrices A_0, G_0, D_0 are dense, due to the Padé approximations of the type $P^{-1}Q$, and they should certainly not be computed explicitly. The iterative method is such that it requires the multiplication of a vector with $P^{-1}Q$, and this can be achieved by solving tri-diagonal systems.

Before constructing the iterative method, the velocity variables at the boundary are eliminated. With

$$A = A_0 - A_1 C^{-1} A_2,$$

$$G = G_0 - A_1 C^{-1} B_0,$$

$$D = D_0 - D_1 C^{-1} A_2,$$

$$B = -D_1 C^{-1} B_0,$$

$$c = b - A_1 C^{-1} b_B,$$

$$d = -D_1 C^{-1} b_B,$$

the system becomes

$$\begin{bmatrix} A & G \\ D & B \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ p \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix}.$$

For convenience, we introduce the notation

$$M = \begin{bmatrix} A & G \\ D & B \end{bmatrix}, \quad x = \begin{bmatrix} \mathbf{w} \\ p \end{bmatrix}, \quad z = \begin{bmatrix} c \\ d \end{bmatrix},$$

and define the approximate factorization of M :

$$\tilde{M} = \begin{bmatrix} A & 0 \\ D & I \end{bmatrix} \begin{bmatrix} I & \frac{2}{3}G \\ 0 & B - \frac{2}{3}DG \end{bmatrix} = \begin{bmatrix} A & \frac{2}{3}AG \\ D & B \end{bmatrix}. \quad (10.17)$$

Only the upper right block is nonzero in the difference

$$M - \tilde{M} = \begin{bmatrix} 0 & (I - \frac{2}{3}A)G \\ 0 & 0 \end{bmatrix}.$$

Instead of solving $Mx = z$ by iteration, we solve the factored system within an outer iteration

$$x^{(\nu+1)} = x^{(\nu)} + \tilde{M}^{-1}r^{(\nu)},$$

where $r^{(\nu)} = z - Mx^{(\nu)}$. For each such iteration, systems with coefficient matrix

$$\begin{bmatrix} A & 0 \\ D & I \end{bmatrix}$$

are solved first, and systems with coefficient matrix

$$\begin{bmatrix} I & 0 \\ 0 & B - \frac{2}{3}DG \end{bmatrix}$$

next. The heavy parts are represented by the systems with matrices A and $B - \frac{2}{3}DG$, where DG is an approximation of the Laplacian Δ , and A approximates $\frac{3}{2}I - k\varepsilon\Delta$. Consequently, if we can construct effective solvers for these two subsystems, we have a fast iterative solver for the whole system.

With the partitioning

$$\delta x = \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix}, \quad r^{(\nu)} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix},$$

the algorithm is

1. Solve $A\delta_1 = r_1$
2. Compute $\delta_2 = r_2 - D\delta_1$
3. Solve $(\frac{2}{3}DG - B)\delta x_2 = -\delta_2$
4. Compute $\delta x_1 = \delta_1 - \frac{2}{3}G\delta x_2$

The systems in step 1 and 3 are solved by preconditioning and iteration. With a preconditioner \tilde{A} , we use the iteration

$$\delta_1^{(\nu+1)} = \delta_1^{(\nu)} + \tilde{A}^{-1}(r_1 - A\delta_1^{(\nu)}),$$

where the convergence is governed by the iteration matrix $I - \tilde{A}^{-1}A$. With the simple choice $\tilde{A} = \frac{3}{2}I$, we get

$$I - \tilde{A}^{-1}A = \frac{2}{3}k\varepsilon L_h,$$

where L_h is an approximation of the Laplacian Δ . Therefore we get convergence if $k\varepsilon||L_h||_h < 3/2$, and since we are thinking of applications where ε is small, we can expect fast convergence. Not only that, but we also have an easy iteration procedure. The main operation is the multiplication of a vector by A , which contain Padé

type operators. But this is a fast operation, since the main work is the solution of tridiagonal systems. Note that A_0 in A acts on the inner velocity points, but it contains Padé operators of the type $P^{-1}Q$. These are defined above with the boundary points included. Therefore, we use the values $f^{(m)}$ of the latest iteration on the right hand side of the systems $Pf' = Qf$, but the boundary values of f' are disregarded. The correct physical boundary conditions are included in the elimination procedure leading to the algorithm and the form of A .

The reason for the simple solution procedure in step 1 is that the time derivative in the equation introduces the diagonal $\frac{3}{2}I$, which allows for the simple structure of the preconditioner. The solution in step 3 is a little worse. The matrix DG represents the Laplacian, and aiming for a the standard Krylov space iteration method like Bi-CGSTAB, we need a good preconditioner. Incomplete LU factorization is often used, but it requires the explicit form of DG . This is no good, since the Padé type approximations lead to dense matrices. Therefore we use instead a sparse factorization of the standard second order approximations \hat{D} and \hat{G} of the divergence and gradient operators. With this preconditioner, each iteration step is fast, since Bi-CGSTAB requires only matrix-vector multiplications, which is fast also with the Padé operators included as noted above.

Numerical experiments

The first test is the Navier–Stokes equations and the well known driven cavity problem. The fluid is put in motion by the upper wall, which has the velocity $u_0 = 1$. Figure 10.3 shows the steady state streamlines for two different Reynolds numbers $Re = 400$ and $Re = 5000$.

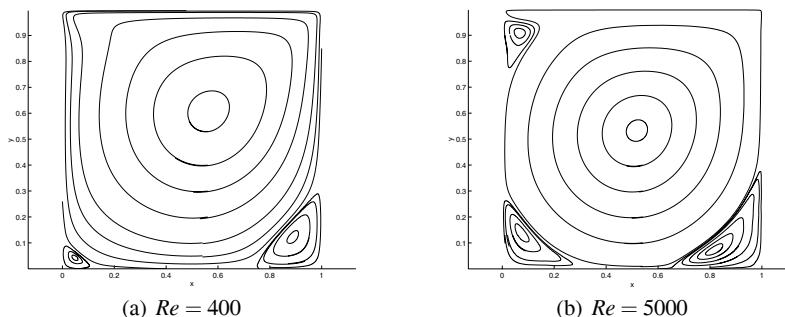


Fig. 10.3 Streamlines for the driven cavity problem

This problem has solutions that have been computed by others to very high accuracy. Figure 10.4 shows the 4th order solution u and v at the center line $x = 0.5$ on a 41×41 grid together with the solution obtained by Ghia et.al. [Ghia et al., 1982] on a 129×129 grid. The results are almost equal.

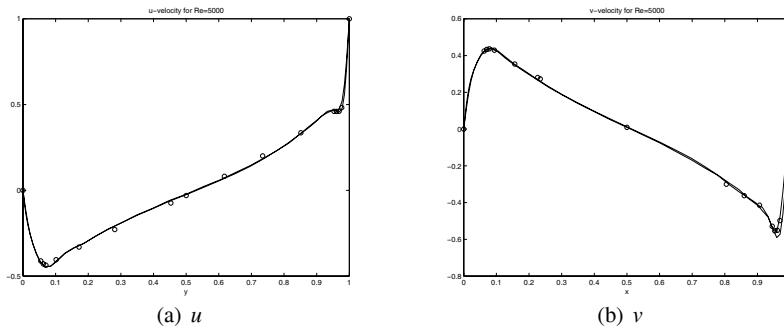


Fig. 10.4 The computed solution u and v at $x = 0.5$ on a 41×41 grid, and the Ghia solution (o) on a 129×129 grid, $Re = 5000$

The next test problem is flow in a constricting channel as shown in Figure 10.5 and presented in [Mancera and Hunt, 1997]. The orthogonal transformation between the physical (x, y) space and the computational (ξ, η) space is obtained by a conformal mapping $z = z(\zeta)$, where $z = x + iy$, $\zeta = \xi + i\eta$. The mapping is

$$z = \zeta(A + B \tanh \zeta),$$

i.e., in component form

$$\begin{aligned} x &= A\xi + \frac{B(\xi \sinh(2\xi) - \eta \sin(2\eta))}{\cosh(2\xi) + \cos(2\eta)}, \\ y &= A\eta + \frac{B(\eta \sinh(2\xi) + \xi \sin(2\eta))}{\cosh(2\xi) + \cos(2\eta)}. \end{aligned}$$

With the limit values of the channel width to the left and right denoted by $2a$ and $2b$ respectively, the constants A and B are

$$A = \frac{a+b}{2\lambda}, \quad B = \frac{b-a}{2\lambda},$$

where λ is the so called shape factor.

A direct mapping as described above, with a uniform grid in (ξ, η) -space gives the grid in physical space as shown in Figure 10.5.

This is not an optimal grid, since the action of the flow is concentrated around the constriction. Therefore, an a priori stretching is applied in the ξ -direction by the transformation

$$\xi(\xi_1) = \frac{\lambda}{2} \sinh\left(\frac{2}{\lambda} \xi_1\right) + \xi_0,$$

where ξ_0 is the singular point of the conformal mapping. The resulting final grid in physical space is shown in Figure 10.6.

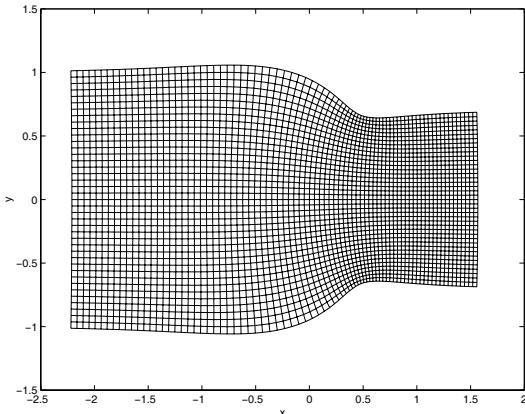


Fig. 10.5 (x, y) -grid in physical space corresponding to a uniform (ξ, η) grid

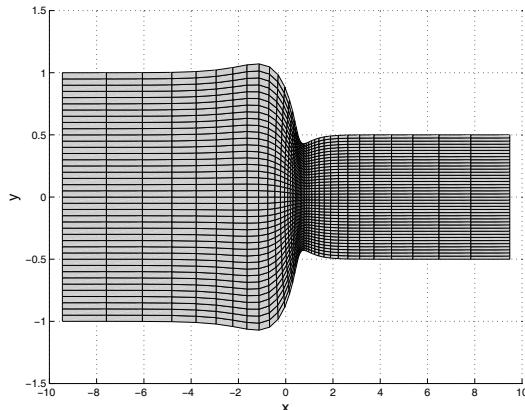


Fig. 10.6 (x, y) -grid in physical space corresponding to a stretched (ξ, η) grid

(The last figure shows a different case compared to Figure 10.5; the length and width of the channel are different.)

The numerical test presented here is for the geometry shown in Figure 10.6. The shape factor is $\lambda = 0.7$, and the Reynolds number is $Re = 150$ (with half the channel width as reference length scale), which produces laminar flow without separation. The boundary conditions have the form (10.11) with

$$u_L = 1 - y^2, \quad v_L = 0, \quad q_L = 1.$$

The true solution is not available for this problem, and in order to estimate the convergence rate \tilde{p} , we use the common technique of using three grids G_ν with step

size $h_\nu = 2^\nu h$, $\nu = 0, 1, 2$. Let $u^{[\nu]}$ denote the numerical solution on grid G_ν for the variable u at a certain point $(\bar{x}, \bar{y}, \bar{t})$. We assume that the error has the form

$$u^{[\nu]} - u(\bar{x}, \bar{y}, \bar{t}) = c(2^\nu h)^{\tilde{p}}, \quad \nu = 0, 1, 2. \quad (10.18)$$

We have three equations, and therefore the unknowns c and $u(\bar{x}, \bar{y}, \bar{t})$ can be eliminated. We get

$$\frac{u^{[2]} - u^{[1]}}{u^{[1]} - u^{[0]}} = 2^{\tilde{p}},$$

i.e.,

$$\tilde{p} = {}^2 \log \frac{u^{[2]} - u^{[1]}}{u^{[1]} - u^{[0]}}.$$

The equality should be taken approximately, since (10.18) does not hold exactly.

Instead of measuring the differences at a certain point, we measure it in the l_2 -norm for each of the three variables, and use grids with 21×21 , 41×41 , 81×81 points. Table 10.2 shows the result.

Table 10.2 Estimated convergence rate \tilde{p} and l_2 -error for the constricting channel computation

Component	\tilde{p}	l_2 -norm of error
u	4.0	$1.0 \cdot 10^{-5}$
v	3.5	$1.0 \cdot 10^{-5}$
p	4.2	$1.9 \cdot 10^{-5}$

The last column shows the estimated error which is based on the average convergence rate 4, giving the formula

$$u^{[0]} - u \approx \frac{u^{[1]} - u^{[0]}}{15},$$

which is also used for v and p .

10.6 Summary

Fluid dynamics is probably the field where the most advanced development of computational methods has been seen over the years, and numerous books have been written about it. In this chapter, we had no intention of giving any overview of the field, but rather to treat an example and show that a fourth order difference method works very well for geometries with structured grids. We chose incompressible flow governed by the Stokes and Navier–Stokes equations, and use the original form with the divergence condition as one of the equations. In this way we avoid the difficulty with boundary conditions for the pressure, which is encountered when the

divergence condition is substituted by the pressure equation. The implicit backward differentiation formula (BDF) is used for discretization in time, and the discrete divergence condition is then part of the algebraic system of equations to be solved at each time step.

The method uses compact Padé type approximations on a staggered grid, and we have given a fairly detailed description of the algorithm. The Padé approximations are defined for each derivative independently, and the original physical boundary conditions are then enforced to obtain the solution at each time step. No real comparison of effectiveness with other methods is done, but the results certainly look promising.

We have included a complete Fourier analysis for the linearized equations, both for the 2nd and 4th order case. It turns out that the stability limit on k is actually determined by a scalar advection-diffusion difference equation. This confirms that the frequently used simplified type of analysis is correct in our case.

The presentation in this chapter is based on the work in [Gustafsson and Nilsson, 2002], [Nilsson et al., 2003], [Gustafsson et al., 2003] (including Fig 10.2), [Brüger et al., 2005a] (including Fig 10.3–10.6) and [Brüger et al., 2005b]. Several other time discretizations are analyzed in [Kress and Lötstedt, 2006] with stability domains determined on the basis of the von Neumann condition.

Chapter 11

Nonlinear Problems with Shocks

The construction of difference methods for shock problems is a vast topic, and several books have been written about it. Even more than for linear problems, the computational methods used in practice are dominated by low order ones. In this chapter we shall discuss some basic ideas, and illustrate them by well known low order methods. We shall give particular emphasis on those techniques where generalizations to high order are possible. We shall limit ourselves to first order conservation laws; this is also where the mathematical efforts have been concentrated over the years.

We begin with a short section about the application of difference methods to nonlinear problems with smooth solutions and the connection to linear theory.

11.1 Difference Methods and Nonlinear Equations

Even if the PDE is nonlinear, the mechanical application of difference methods is straightforward, and we have seen a few examples in the previous chapters. For example, if a linear term $a(x)u_x$ is approximated by $a_j Qu_j^n$, where Q is a difference operator, then the nonlinear term $a(u)u_x$ is approximated by $a(u_j^n)Qu_j^n$. As long as the solution is smooth, there is a good possibility that the numerical solution behaves well. A necessary condition is that the linearized approximation is stable. If the stability condition for the linear problem is

$$\frac{k}{h} \max_j |a_j| \leq \lambda_0,$$

we require

$$\frac{k}{h} \max_{j,n} |a(u_j^n)| \leq \lambda_0$$

for the nonlinear problem. Since the solution u_j^n is not known a priori, it may be difficult to choose the time step once and for all. A common way to overcome this

difficulty, is to check the condition during the solution process. A typical algorithm computes $\max_j |a(u_j^n)|$ every m :th step, where m is a small integer. The time step is chosen with a good margin such that k is computed as

$$k = \frac{(1 - \delta)h\lambda_0}{\max_j |a(u_j^n)|},$$

where δ is of the order 0.2, say. A local change of time step is easy for one step schemes, but for multistep schemes it is not. Some sort of interpolation must be done in order to define the solution at the required time levels behind. Furthermore, if the changes are made too frequently, the stability may be affected, even for linear problems.

If the solution is sufficiently smooth, the analysis of the linearized problem may be enough to ensure stability and convergence theoretically even for the nonlinear problem. There is a paper by Strang [Strang, 1987], where the precise conditions and rate of convergence are derived with the linearized problem as a basis for the analysis.

Stability by itself can be ensured also for nonlinear problems by using the energy method. This was demonstrated in Section 7.2, where the equation (7.15) is cast in so called selfadjoint form allowing for an energy estimate. If the solution is smooth, we can expect convergence of the numerical solution to the right solution, but some extra conditions are required for a theoretical proof.

If the solution is not smooth, the situation is completely different. Nonlinear problems may well be such that discontinuities are developed after some time, even if the initial function is smooth. In such a case, the first challenge is to define precisely what is meant by a solution, and the second challenge is to construct a numerical method that produces this solution. A direct application of the methods that we have discussed above will fail in most cases.

11.2 Conservation Laws

We begin by discussing some basic concepts that are necessary in order to understand the mechanism underlying the construction of numerical methods for shock problems. We shall consider a so called scalar conservation law

$$u_t + f_x(u) = 0, \quad (11.1)$$

where $f(u)$ is a nonlinear function of u . Here we shall assume that f is a convex function of u , but there are also interesting applications where this is not the case. The name conservation law refers to the fact that the integral of the solution is independent of time. Assuming that u vanishes outside an interval (x_1, x_2) we have

$$\frac{d}{dt} \int_{x_1}^{x_2} u dx = - \int_{x_1}^{x_2} f_x(u) dx = f(u(x_1, t)) - f(u(x_2, t)) = 0.$$

Let us next study Burgers' equation, which is the simplest but most well known conservation law for the purpose of analysis. We use the interval $-1 \leq x \leq 1$, and prescribe boundary conditions on both sides:

$$\begin{aligned} u_t + \left(\frac{u^2}{2}\right)_x &= 0, \quad -1 \leq x \leq 1, \quad 0 \leq t, \\ u(-1, t) &= 1, \\ u(1, t) &= 0, \\ u(x, 0) &= u_0(x). \end{aligned} \tag{11.2}$$

The so called primitive form of the equation is

$$u_t + uu_x = 0.$$

This looks very much like the linear advection equation that we have used many times in the earlier part of this book. Indeed, as long as the solution is smooth, it makes sense to use the same arguments as for the linear equation. The coefficient u can be seen as the characteristic velocity, such that a certain feature in the solution is traveling with the velocity $dx/dt = u(x, t)$. Assume that the initial function u_0 is linear from one to zero in the left half of the interval and zero in the right part, see Figure 11.1. The value one at $x = -1$ is traveling with the speed one to the right, while the lower values are traveling with a slower speed. The result is a profile that becomes steeper with increasing time, and at $t = 1$ the solution has formed a shock at $x = 0$, see Figure 11.1.

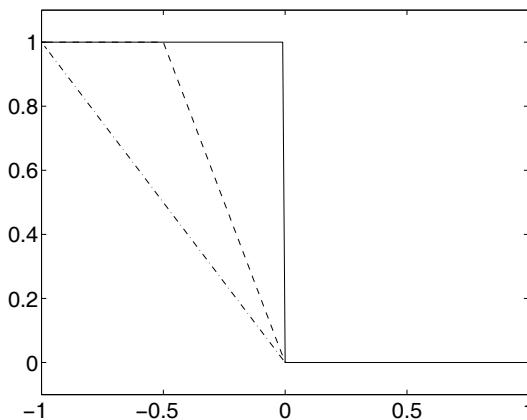


Fig. 11.1 Solution of Burgers' equation, $t = 0$ (\cdots), $t = 0.5$ ($--$), $t = 1$ ($-$)

The problem is to define a unique solution for $t > 1$.

At this point we modify the equation by introducing a “viscous” term, and obtain the *viscous Burgers' equation*

$$v_t + \left(\frac{v^2}{2}\right)_x = \varepsilon v_{xx}. \quad (11.3)$$

Here ε is a small positive number, and the term on the right hand side represents some kind of viscosity or smoothing. For this equation we compute the solution up to $t = 2$, and the result is shown in Figure 11.2.

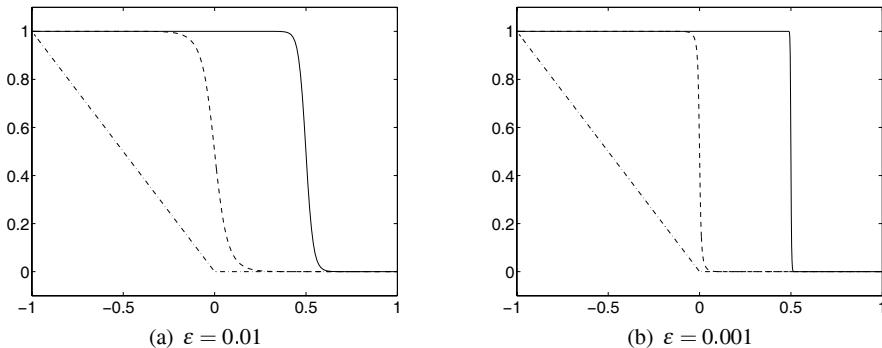


Fig. 11.2 Solution of the viscous Burgers' equation, $t = 0$ (\cdots), $t = 1$ ($--$), $t = 2$ ($-$)

The viscosity term smoothes the solution such that the profile at $t = 1$ is rounded a little bit. There is no special difficulty in computing the solution for $t > 1$, since the equation is now parabolic and has a uniquely defined solution for all t . Furthermore, when the viscosity becomes smaller, the solution is sharpened up. A discontinuity is formed at $t = 1$, just as for the nonviscous equation, and this shock is then traveling to the right. We make two interesting observations. It seems like a unique solution is obtained in the limit when $\varepsilon \rightarrow 0$, and it makes sense to define the solution of the conservation law as the limit solution of the viscous equation. The second observation is that the shock travels with the speed 0.5 once it has formed, and this is the average of the max- and min-values at the shock.

Let us now go back to the general conservation law (11.1). To avoid the difficulty with definition of the derivative across shocks, we define a weak solution. Let $\phi(x, t)$ be test functions that are differentiable and have compact support, i.e., they vanish outside some finite domain. The equation is multiplied by these test functions and integrated:

$$\int_0^\infty \int_{-\infty}^\infty (u_t + f_x(u)) \phi \, dx \, dt = 0.$$

After integration by parts we get

$$\int_0^\infty \int_{-\infty}^\infty (u\phi_t + f(u)\phi_x) \, dx \, dt = - \int_{-\infty}^\infty u(x, 0)\phi(x, 0) \, dx. \quad (11.4)$$

We say that u is a *weak solution* of the conservation law if it satisfies (11.4) for all differentiable test functions ϕ with compact support. The trouble is that there may

be many weak solutions to the same problem, and we must decide which one is of interest.

In order to extract the proper solution, we define the viscous solution $v(x, t, \varepsilon)$ as the unique solution to

$$v_t + f_x(v) = \varepsilon v_{xx},$$

with the same initial and boundary conditions. The correct solution to the original problem is then defined as

$$u(x, t) = \lim_{\varepsilon \rightarrow 0} v(x, t, \varepsilon).$$

There is a reasonable physical explanation to this definition. We take gas dynamics as an example, with inviscid flow governed by the *Euler equations*

$$\begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(E + p) \end{bmatrix}_x = 0. \quad (11.5)$$

The variables are density ρ , velocity u , energy E and pressure p . Since there are four unknowns, we also need an equation of state $p = p(\rho, u, E)$. The equation (11.5) is a nonlinear system of conservation laws, and if the velocity is greater than the speed of sound, there are shocks. However, the assumption of no viscosity is a simplification that is never satisfied exactly in nature. There is always some viscous forces, and the true state is governed by the *Navier–Stokes equations*

$$\begin{bmatrix} \rho \\ \rho u \\ E \end{bmatrix}_t + \begin{bmatrix} \rho u \\ \rho u^2 + p \\ u(E + p) \end{bmatrix}_x = \varepsilon \mathbf{g}_x(\mathbf{u}, \mathbf{u}_x).$$

Here \mathbf{u} is the vector containing the state variables ρ, u, E , and the right hand side contains second derivatives corresponding to viscosity. The coefficient $\varepsilon = 1/Re$ is small, since the *Reynolds number* Re is large for many important applications, typically of the order 10^7 for flow around aircrafts. The solution is practically discontinuous, but for an extremely fine grid, one should be able to observe a slightly rounded profile. For practical shock computations, such a resolution is impossible, and therefore it is necessary to define the proper solution to the Euler equations. This solution is defined as the limit solution when $\varepsilon \rightarrow 0$.

Going back to the scalar case, it can be shown that the shock speed s is given by the *Rankine–Hugoniot condition*

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R}, \quad (11.6)$$

where u_L and u_R denote the states on the left and right sides of the shock respectively. For Burgers' equation, we have $s = (u_L + u_R)/2$, which agrees with the observation we made for the numerical experiment above.

The so called primitive form of the conservation law is

$$u_t + f'(u)u_x = 0.$$

With reference to the wave (or advection) equation considered earlier in this book, we can identify $f'(u)$ with the wave speed or characteristic speed. The proper shock solution can now be defined by the *entropy condition*

$$f'(u_L) > s > f'(u_R).$$

This condition is simply a requirement that the local characteristics point into the shock from both sides. Figure 11.3 shows the situation for the example above with Burgers' equation, where the shock moves with the speed 1/2.

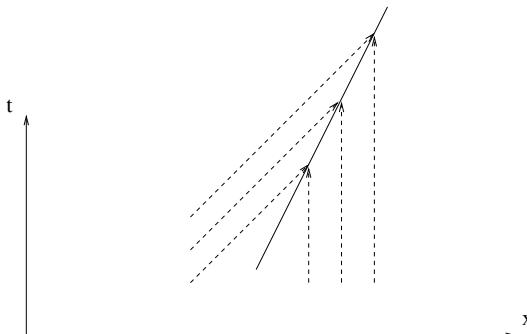


Fig. 11.3 Shock location (−) and characteristics (−−)

We have already given the Euler equations as an example of systems of conservation laws. A general system has the form

$$\mathbf{u}_t + \mathbf{f}_x(\mathbf{u}) = \mathbf{0},$$

where $\mathbf{f}(\mathbf{u})$ is a nonlinear vector function of the vector \mathbf{u} . The Rankine–Hugoniot condition for the shock speed s is now given by

$$s(\mathbf{u}_L - \mathbf{u}_R) = \mathbf{f}(\mathbf{u}_L) - \mathbf{f}(\mathbf{u}_R).$$

The states on both sides of the shock cannot be arbitrarily given. Since \mathbf{u} and \mathbf{f} are vectors, the left and right hand sides must be linearly dependent.

The Jacobian \mathbf{f}' of \mathbf{f} plays an important role for systems. If \mathbf{u} has m elements, there are m families of possible shocks, each one associated with one of the eigenvalues $\lambda_\nu(\mathbf{u})$ of the matrix $\mathbf{f}'(\mathbf{u})$. For a shock corresponding to the eigenvalue λ_ν , the *Lax entropy condition* is

$$\lambda_\nu(\mathbf{u}_L) > s > \lambda_\nu(\mathbf{u}_R),$$

i.e., the local characteristics point into the shock.

Finally we note that the solution may well have discontinuities that are not shocks in the mathematical sense. If the characteristics do not point into the discontinuity as described above, but are parallel, then we have a so called contact discontinuity. Such solutions are of course possible also for linear problems. In such cases, the situation is actually worse than for shocks. The reason is that possible oscillations that are created around the shock are spread wider. For real shocks, the characteristics help pushing the oscillations back towards the discontinuity, while this is not the case for contact discontinuities.

11.3 Shock Fitting

The main difficulty with shock problems and difference methods has its roots in the quite unnatural way of using grid points on both sides of the discontinuity in an attempt to approximate a derivative that is not even well defined. It is hardly surprising that the result is not satisfactory, and the main content of this chapter is about techniques that help to overcome this difficulty. Another way to avoid this trouble is to separate the regions on either side of the shock, and to treat them separately with the shock location as an internal boundary in between. We have shown above that the shock speed is well defined by the Rankine–Hugoniot conditions. In this way we can keep track of the shock location as time progresses.

Let us assume that we are using an explicit method for a scalar conservation law on the interval $0 \leq x \leq 1$ and that there is one shock with location $\xi(t)$ known for $t = t_n$ with $x_m \leq \xi(t_n) < x_{m+1}$. The difference scheme is applied for all grid points on the left hand side. If the local characteristic is pointing to the right, some kind of one-sided procedure is applied near the shock, without involving any point to the right of x_m . This internal boundary treatment must be constructed such that it is stable as discussed earlier in this book. The analogous procedure is applied to the right of the shock without involving any point to the left of x_{m+1} . The condition (11.6) with $u_L = u_m^{n+1}$, $u_R = u_{m+1}^{n+1}$ is used to compute $\xi(t_{n+1})$. If $\xi(t_{n+1})$ has passed a grid point, the index m is adjusted correspondingly, and this completes the computation over one time step.

There is of course nothing in this procedure that prohibits the use of high order methods. The principles for construction discussed earlier can be applied within each smooth region and with the stable boundary procedures applied near the shocks. If the computation of $\xi(t)$ is done to the same accuracy as the main scheme, the presence of shocks does not destroy the full accuracy.

In several space dimensions there are conditions corresponding to the Rankine–Hugoniot conditions. They are considerably more complicated, and the character of the discontinuities has a wider spectrum. One type of discontinuity may interfere with another type, and we must keep track of the exact behavior at the intersection. This is the reason for the rare use of shock fitting for real life multidimensional problems. However, such computations exist. The benefit of having accurate shock

representation with neither nonphysical oscillations or nonphysical smoothing may outweigh the more extensive and complicated programming effort.

11.4 Artificial Viscosity

In the first section of this chapter we saw that with a viscosity term added to the conservation law, we get a solution that approaches the correct solution when the viscosity coefficient tends to zero. This suggests a possible numerical technique based on artificial viscosity added to the basic scheme. This allows for so called *shock capturing*, where the scheme hopefully provides an approximation without keeping track of the exact location of the shock. However, the results shown in Figure 11.2 are obtained on grids that are fine enough to resolve the viscous solution. In other words, there must be a certain number of points in the transition layer from one side of the shock to the other side. For nontrivial problems, this leads to unrealistic computation times and memory requirements. With a coarser grid, one has to be very careful with the construction to avoid bad oscillations. For example, let us assume that we want to solve the problem (11.2) for the inviscid Burgers' equation, but use an artificial viscosity term corresponding to the viscous Burger's equation (11.3). With the standard second order centered difference operators for $\partial/\partial x$ and $\partial^2/\partial x^2$ and a viscosity coefficient $\varepsilon = 0.001$, we get the result shown in Figure 11.4 with 250 grid points.

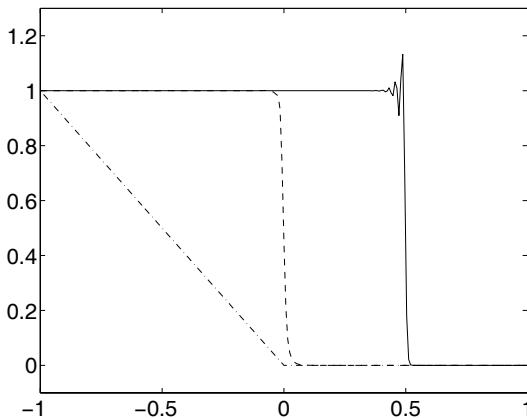


Fig. 11.4 Solution of the viscous Burgers' equation, $\varepsilon = 0.001$, $N = 250$ (compare Figure 11.2(b))

The oscillations behind the shock are sometimes accepted, in particular if they are limited to the immediate neighborhood of the shock as here, but in general they cause severe trouble. In gas dynamics, for example, the pressure may go negative locally, which may cause the algorithm to break down. On the other hand, if the

viscosity coefficient is increased, the result is a more smoothed solution as shown in Figure 11.2(a), which may be unacceptable as well. This example shows that the introduction of artificial viscosity does not guarantee plain sailing.

Some methods introduce viscosity as an effect of the truncation error. One such example is the *Lax–Friedrichs scheme*

$$u_j^{n+1} = \frac{1}{2}(u_{j-1}^n + u_{j+1}^n) - kD_0 f(u_j^n).$$

The averaging of a smooth function $u(x)$ gives

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

i.e., the scheme approximates a parabolic equation, and it is only first order accurate. This is an example of a dissipative scheme as introduced in Section 2.2.2. This particular method introduces a quite strong damping, in particular for small Courant numbers $\lambda = k/h$. The dissipative term is independent of k , so for smaller time steps, it acts more frequently on a fixed time interval $[0, T]$. Figure 11.5 shows this effect on the computed Burger solution for two different values of λ .

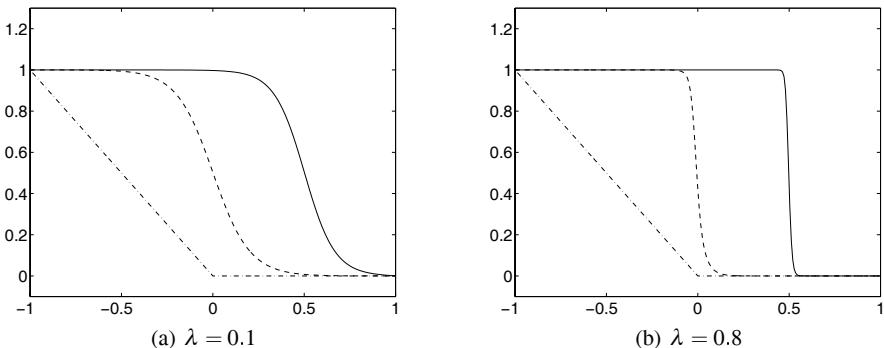


Fig. 11.5 Solution of Burgers' equation with the Lax–Friedrichs scheme, $N = 250$

Viscosity may also be introduced as an effect of the construction when achieving second order accuracy. The most famous methods based on this principle is the Lax–Wendroff scheme and the MacCormack scheme. The latter one is

$$\begin{aligned} u_j^* &= u_j^n - kD_- f(u_j^n), \\ u_j^{**} &= u_j^* - kD_+ f(u_j^*), \\ u_j^{n+1} &= \frac{1}{2}(u_j^n + u_j^{**}). \end{aligned}$$

For a linear problem $u_t + u_x = 0$, the two schemes are equivalent, and reduce to

$$u_j^{n+1} = (I - kD_0 + \frac{k^2}{2}D_+D_-)u_j^n.$$

The last term raises the accuracy to second order in time, and introduces at the same time a dissipative term. The scheme has less damping than the Lax-Friedrichs scheme for nonlinear problems, and produces sharper shock profiles. However, one has to pay for that by accepting some oscillations around the shock (also called “overshoot”), see Figure 11.6.

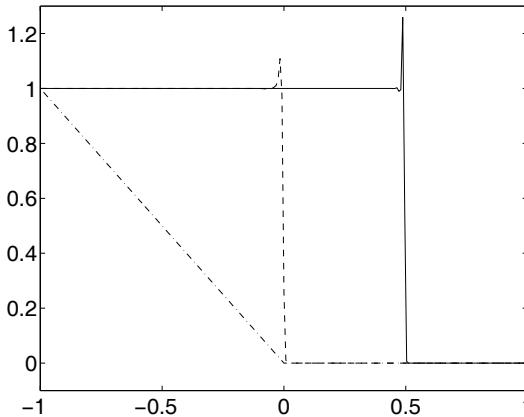


Fig. 11.6 Solution of Burgers' equation with the MacCormack scheme, $N = 250$, $k/h = 0.8$

In this case, the dominating truncation error is a third derivative, and this one is the source of the oscillations.

At this point we shall discuss the monotonicity concept. Assume that we have an entropy satisfying solution $u(x, t)$ to a scalar conservation law with $u(x, 0) = u_0(x)$, and another one with $v(x, 0) = v_0(x)$. Then it can be shown that if

$$u_0(x) \geq v_0(x) \text{ for all } x,$$

then

$$u(x, t) \geq v(x, t) \text{ for all } x \text{ and } t.$$

The numerical method is *monotone* if it has the same property, i.e., if

$$u_j^n \geq v_j^n \text{ for all } j,$$

then

$$u_j^{n+1} \geq v_j^{n+1} \text{ for all } j.$$

A numerical method is *monotonicity preserving* if the solution u_j^n is monotone as a function of j for all n provided u_j^0 is monotone. One can show that a monotone method is monotonicity preserving.

Our model problem above is an example where the solution is monotone for all t . In order to avoid oscillations, we would like the numerical solution to be monotonicity preserving. The numerical methods we have been considering here are linear, i.e., the coefficients in the scheme are independent of the solution. For such methods there is a severe accuracy limitation as given by

Theorem 11.1. *Any monotonicity preserving linear method can be at most first order accurate.* \square

This explains the behavior of the second order accurate MacCormack scheme for the example above.

It is possible to design methods with very sharp shocks based on centered approximations and artificial viscosity. For steady state shocks the theory is simpler. One can take advantage of that by transforming the time dependent problem to a stationary one, and we shall indicate how it can be done. As an example we consider the so called *Riemann problem* for a scalar conservation law:

$$u_t + f_x(u) = 0, \quad -\infty < x < \infty, \quad 0 \leq t,$$

$$u(x, 0) = \begin{cases} u_L, & x < 0 \\ u_R, & x \geq 0. \end{cases}$$

After a coordinate transformation

$$y = x - st,$$

$$\tau = t,$$

the problem becomes a stationary one

$$(f - su)_y = 0, \quad -\infty < y < \infty,$$

$$u(y, 0) = \begin{cases} u_L, & y < 0 \\ u_R, & y \geq 0. \end{cases}$$

In order to design a scheme for sharp and monotone shock profiles, we add a viscous term to the standard centered difference approximation:

$$D_0(f_j - su_j) = \varepsilon h D_+ D_- u_j,$$

$$\lim_{j \rightarrow -\infty} u_j = u_L,$$

$$\lim_{j \rightarrow \infty} u_j = u_R.$$

The difference method can be written in conservative form as

$$D_+ \left(\frac{1}{2} (f_j - su_j + f_{j-1} - su_{j-1}) - \varepsilon (u_j - u_{j-1}) \right) = 0.$$

This leads to

$$\begin{aligned} & \frac{1}{2}(f_{j+1} - su_{j+1} + f_j - su_j) - \varepsilon(u_{j+1} - u_j) \\ &= \frac{1}{2}(f_j - su_j + f_{j-1} - su_{j-1}) - \varepsilon(u_j - u_{j-1}) = \dots = f_L - su_L, \end{aligned}$$

i.e.,

$$f_{j+1} - f_L - s(u_{j+1} - u_L) - 2\varepsilon u_{j+1} = -(f_j - f_L) + s(u_j - u_L) - 2\varepsilon u_j.$$

With the definitions

$$\begin{aligned} F(u) &= f(u) - f_L - s(u - u_L) - 2\varepsilon u, \\ G(u) &= f_L - f(u) + s(u - u_L) - 2\varepsilon u, \end{aligned}$$

the difference equation can be written

$$F(u_{j+1}) = G(u_j). \quad (11.7)$$

The question is now if the two states u_L and u_R can be connected by a grid function u_j via a transition layer with very few points, and this is now a pure algebraic problem. Indeed, it can be shown that a one point transition layer can be achieved by a proper choice of ε . However, it turns out that a more robust method is obtained by using the linearized form of (11.7). The linearization is done around a perfect shock solution, and the value of ε is then obtained under the condition that the perturbation is almost zero. For our model problem, it was shown in [Gustafsson and Olsson, 1996] that the choice

$$\varepsilon = \frac{1}{4}(f'(u_L) - f'(u_R)) \quad (11.8)$$

annihilates the oscillations completely.

For the real time dependent problem, we cannot expect the resulting method to work as nicely as for the transformed problem, but numerical evidence shows that the method is quite robust.

An attractive feature with this procedure of deriving the viscosity coefficient is that it can be generalized to higher order methods. The 4th order centered approximation is

$$\partial/\partial x \approx RD_0, \quad R = I - \frac{h^2}{6}D_+D_-.$$

The trick is to involve the difference operator R in the viscosity term. For the transformed problem we use

$$RD_0(f_j - su_j) = \varepsilon h RD_+ D_- u_j.$$

Since the operator R is nonsingular, the 4th order approximation gives the same solution as the 2nd order approximation. For a real problem, we work in the original physical x, t -space, and there is time discretization to take into account.

For a numerical test we choose again Burgers' equation as used throughout this chapter. We have $u_L = 1$ and $u_R = 0$, and (11.8) gives $\varepsilon = 0.25$. Figure 11.7 shows the result for both schemes.

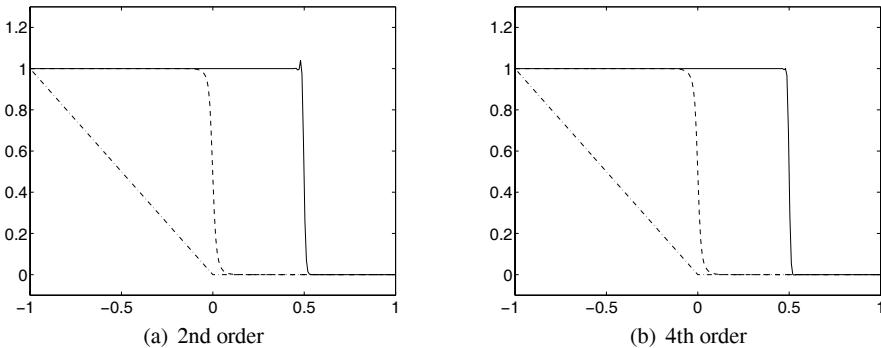


Fig. 11.7 Solution of Burgers' equation with the 2nd and 4th order schemes with viscosity terms, $\varepsilon = 0.25$, $N = 250$

The solutions are quite accurate, and the 4th order one has a particularly nice profile at the shock.

The procedure is easily generalized to any even order of accuracy. All the centered approximations of order $2p$ have the form $R_{2p}D_0$, where R_{2p} is a nonsingular operator. The viscosity term is taken as $\varepsilon h R_{2p} D_+ D_- u_j$, and the conclusion is the same as for the second order approximation. Note that we can use the implicit compact operators of the Padé type as well, since they have the same form with R_{2p} taken as the inverse of a sparse operator.

Formally, the viscosity term takes the accuracy down to first order, even for the smooth part of the solution. The cure for this is to activate it only in the immediate neighborhood of the shock. One way of implementing this is to use the modified viscosity term

$$\varepsilon h D_+ r_j D_- u_j,$$

where r_j is a grid function with $r_j = 1$ close to the shock, and $r_j = 0$ outside. In between, there is a smooth transition between the two values in a narrow region.

In the next section we shall see how to construct the scheme in the transition zone via so called flux limiters applied directly to the conservative form of the scheme.

11.5 Upwind Methods

As we have seen, the trouble with shock computations is the differencing across the shock, since there is a jump of order one. It is therefore natural to construct methods that change the shape of the stencil when it passes from one side of the shock

to the other. Since shock computations are strongly associated with gas dynamics, these methods are called *upwind methods*. The stencil picks up values only from the upwind side, i.e., the side where the characteristics are coming from. This principle is used for shock fitting methods as described above, but those methods require that the shock position is kept as a state variable. When talking about shock capturing methods, we must have some mechanism that automatically changes the stencil, only with knowledge of $f(u)$ and $f'(u)$ as a basis.

In order to construct high order monotonicity preserving methods, we conclude from the previous section that we must aim for nonlinear methods. Such methods are nonlinear also for linear problems. The last decades have been a very active research period when it comes to so called high resolution methods for shock problems. We don't have the ambition here to survey all these methods, but we shall discuss the basic principles.

We limit ourselves to one step schemes, and require that they are conservative. This means that they can be written in conservative form

$$u_j^{n+1} = u_j^n - \frac{k}{h} (F(u_{j-r+1}^n, u_{j-r+2}^n, \dots, u_{j+p}^n) - F(u_{j-r}^n, u_{j-r+1}^n, \dots, u_{j+p-1}^n)). \quad (11.9)$$

Here F is called the *flux function*. For the conservation law, we have when disregarding boundary terms

$$\frac{d}{dt} \int u dx = 0.$$

In the discrete case we have

$$\sum_j u_j^{n+1} - \sum_j u_j^n = 0.$$

One can prove that if the scheme converges to a solution, the conservation property guarantees that this solution is the correct one satisfying the entropy condition. In particular it guarantees that the shock speed is correct.

The simplest example of a conservative scheme is the first order upwind scheme for the Burgers' equation

$$u_j^{n+1} = u_j^n - \frac{k}{2h} ((u_j^n)^2 - (u_{j-1}^n)^2),$$

where it is assumed that u_j^n is non-negative everywhere. Here the numerical flux function is the same as the true flux function:

$$F(u_{j-r+1}^n, u_{j-r+2}^n, \dots, u_{j+p}^n) = F(u_j^n) = \frac{(u_j^n)^2}{2}.$$

The Lax–Friedrichs and the MacCormack schemes are conservative as well, and we have seen above that they produce the correct shock speed.

Upwind methods are easily generalized to general conservation laws if f' has the same sign in the whole computational domain. They can also be generalized to

systems of conservation laws if none of the eigenvalues of the matrix \mathbf{f}' changes sign anywhere. This is a very unusual situation, and we must design the upwind methods such that they work with changing signs.

We go back to the scalar case. The basic question is how to switch from one stencil to another one when f' changes sign. We shall describe how the switch is constructed in the Engquist–Osher scheme. We split f into the two functions

$$\begin{aligned} f^+(u) &= f(0) + \int_0^u \max(f'(\theta), 0) d\theta, \\ f^-(u) &= \int_0^u \min(f'(\theta), 0) d\theta, \end{aligned} \quad (11.10)$$

that satisfy $f(u) = f^+(u) + f^-(u)$. The numerical flux is defined as

$$F(u, v) = f^+(u) + f^-(v), \quad (11.11)$$

and the complete conservative method is

$$u_j^{n+1} = u_j^n - \frac{k}{h} (F(u_j^n, u_{j+1}^n) - F(u_{j-1}^n, u_j^n)). \quad (11.12)$$

Another principle for the construction of upwind methods is used by the *Godunov method*, which we shall now describe. We define first the cell average

$$U(j, t) = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t) dx,$$

and we are going to construct a numerical approximation U_j^n to this one for all j and n . Assuming that U_j^n is known, we define a piecewise constant function $\tilde{u}^n(x)$ with

$$\tilde{u}^n(x) = U_j^n, \quad x_{j-1/2} \leq x < x_{j+1/2}$$

for all j . There is now a discontinuity in each interval (x_j, x_{j+1}) , and with $\tilde{u}^n(x)$ as initial data, we solve $\tilde{u}_t + f_x(\tilde{u}) = 0$ exactly for $t_n \leq t \leq t_{n+1}$. This is a Riemann problem, and its solution $\tilde{u}(x, t)$ is well defined as long as discontinuities from neighboring intervals do not interfere. This provides a limit on the time step, just as for ordinary explicit difference methods. Note that $\tilde{u}(x, t)$ is in general not piecewise constant at $t = t_{n+1}$. The new approximation U_j^{n+1} is obtained as the average

$$U_j^{n+1} = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} \tilde{u}(x, t_{n+1}) dx.$$

The algorithm over one time step is now complete, and the process is continued with new initial data for the Riemann problem constructed from U_j^{n+1} .

By using the conservation properties of the differential equation, it is easily shown that the algorithm can be written in conservative form as

$$U_j^{n+1} = U_j^n - \frac{k}{h} (F(U_j^n, U_{j+1}^n) - F(U_{j-1}^n, U_j^n)) ,$$

where the flux function is defined by

$$F(U_j^n, U_{j+1}^n) = \frac{1}{k} \int_{t_n}^{t_{n+1}} f(\tilde{u}(x_{j+1/2}, t)) dt .$$

The upwind nature of the Godunov method is due to the fact that the PDE is solved exactly. For a right-going wave governed by $u_t + u_x = 0$, the cell averages U_j^n are moving to the right with speed one, and U_j^{n+1} is not affected by U_{j+1}^n .

The method is only first order accurate, and strongly dissipative. However, it is quite robust, and for systems it also has the advantage that there is no need to find the eigensystem of \mathbf{f}' provided that we can solve the Riemann problem exactly.

The obvious approach to raise the accuracy, is to substitute the piecewise constant $\tilde{u}(x, t_n)$ by a more refined procedure. The first choice would be to construct it as a piecewise linear function

$$\tilde{u}(x, t_n) = U_j^n + s_j^n(x - x_j) , \quad x_{j-1/2} \leq x < x_{j+1/2} ,$$

where the slope s_j^n depends on the neighboring cell averages. For a linear PDE with right-going characteristics, it is natural to use

$$s_j^n = \frac{1}{h} (U_{j+1}^n - U_j^n) .$$

It is easily shown that we get the Lax-Wendroff method in this way, which is second order accurate. However, we are now back with the oscillation trouble that was discussed above. There is still another difficulty: the exact solution $\tilde{u}(x, t)$ of the conservation law with the new type of initial data may no longer be available. This problem may be overcome by introducing approximate solutions, but we shall not pursue this issue here.

But how do we handle the oscillation problem? One way is to introduce *slope limiters*. The basic idea is to have some mechanism that modifies the slopes if they are becoming too large, and many slope limiters have been suggested.

Usually the construction is done in such a way that the scheme becomes *total variation diminishing* (TVD). The *total variation* TV of a grid function u_j is defined by

$$TV(u) = \sum_j |u_{j+1} - u_j| .$$

For a TVD method we require

$$TV(u^{n+1}) \leq TV(u^n)$$

for all grid functions u_j^n . The corresponding property defined by integrals holds for the conservation law, and it is therefore a natural property to require. Furthermore it

can be shown that a TVD method is monotonicity preserving, and we should expect nonoscillatory and well behaving solutions.

The upwind methods or other specially designed methods for shock computation are almost always slower on the computer compared to standard methods. Furthermore, they have lower accuracy. Therefore, they are not effective if they are used in the whole computational domain. In order to overcome this problem, the application of a certain shock computation method can be limited to the immediate neighborhood of the shocks, and coupled to a more effective high order method in the rest of the domain. This is called a *hybrid method*. Such a procedure is closely connected to applying a viscosity term locally as discussed in the previous section. For a scheme written in conservative form, we can achieve the localization by using a flux function of the form

$$F = F_L + \phi(F_H - F_L),$$

or equivalently

$$F = F_H - (1 - \phi)(F_H - F_L).$$

Here the subscripts L and H indicate low and high order, respectively. The function ϕ is called a *flux limiter*, and it is constructed as a function of the smoothness of the solutions, with $\phi \approx 0$ around the shock. One way to achieve this is to look at two consecutive differences, and define

$$\theta_j = \frac{u_j - u_{j-1}}{u_{j+1} - u_j}.$$

If θ_j is near one, the solution is smooth, otherwise not. The remaining problem is to define the function ϕ , such that the low order flux becomes activated when the argument is not close to one. There is a rich collection of flux limiters that have been used. One of them is the “superbee” limiter suggested by Roe [Roe, 1985]

$$\phi(\theta) = \max(0, \min(1, 2\theta), \min(\theta, 2)).$$

11.6 ENO and WENO Schemes

As we have seen above, the generalization of a certain method to high order accuracy is quite complicated and has often bad side effects. Perhaps the most interesting ones from this point of view are the *essentially nonoscillating* (ENO) schemes, and we shall now describe one version of them.

The basis is again the cell averages $U(j, t)$ and their approximations U_j^n . The true cell averages satisfy the integrated version of the conservation law:

$$\frac{d}{dt} U(j, t) + \frac{1}{h} \left(f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t)) \right) = 0. \quad (11.13)$$

The trouble with this form is that it contains both cell averages $U(j, t)$ and point values $u(x_{j+1/2})$. If we can derive accurate approximations of $u(x_{j+1/2})$ in terms of $U(j, t)$, we can use these directly in (11.13).

Define the primitive function $w(x)$ at a fixed time t by

$$w(x) = \int_{x_{m-1/2}}^x u(\xi, t) d\xi,$$

where the point $x_{m-1/2}$ is an arbitrary midpoint. With W_j defined as

$$W_j = w(x_{j+1/2}),$$

we have

$$W_j = h \sum_{\nu=m}^j U(\nu, t).$$

(For convenience we use a half step shift and use the notation W_j instead of $w_{j+1/2}$.) We can now use the point values W_ν in the neighborhood of x_j to construct a polynomial $P_j(x)$ that approximates $w(x)$ for $x_{j-1/2} \leq x \leq x_{j+1/2}$. The derivative $P'_j(x)$ is then an approximation of $u(x, t)$. If the polynomial has degree p , then

$$P_j(x) = w(x) + \mathcal{O}(h^{p+1}), \quad x_{j-1/2} \leq x \leq x_{j+1/2}$$

if $w(x)$ is smooth. For the derivative we lose one order and get

$$P'_j(x) = u(x, t) + \mathcal{O}(h^p), \quad x_{j-1/2} \leq x \leq x_{j+1/2}.$$

At each cell interface we now have two values to choose from:

$$\begin{aligned} u_{j+1/2}^L(t) &= P'_j(x_{j+1/2}) \text{ and} \\ u_{j+1/2}^R(t) &= P'_{j+1}(x_{j+1/2}). \end{aligned}$$

In the smooth part, these are very close, and both accurate of order p . In the non-smooth part they differ more, and we need a flux function $F(u, v)$ built on a good switch. One choice is the Enquist–Osher flux (11.10), (11.11). Another one is the Roe flux defined by

$$F(u, v) = \begin{cases} f(u) & \text{if } f'(w) \geq 0 \text{ for } \min(u, v) \leq w \leq \max(u, v), \\ f(v) & \text{if } f'(w) \leq 0 \text{ for } \min(u, v) \leq w \leq \max(u, v), \\ \tilde{h}(u, v) & \text{else,} \end{cases} \quad (11.14)$$

where

$$\begin{aligned} \tilde{h}(u, v) &= \frac{1}{2} (f(u) + f(v) - \beta(v - u)), \\ \beta &= \max_{\min(u, v) \leq w \leq \max(u, v)} |f'(w)|. \end{aligned} \quad (11.15)$$

The final approximation then is

$$\frac{d}{dt}U_j(t) + \frac{1}{h}\left(F(u_{j+1/2}^L(t), u_{j+1/2}^R(t)) - F(u_{j-1/2}^L(t), u_{j-1/2}^R(t))\right) = 0. \quad (11.16)$$

With proper boundary conditions, this is a closed system, since each u^L and u^R is a function of a few of the values $U_\nu(t)$ as described above.

So far we have left out the important question of how to select the points for interpolation of $W_j(t)$. We recall that ENO means that there are essentially no oscillations in the solution. Several strategies have been proposed, the following is one of them. The algorithm is iterative and is based on the Newton interpolation formula. For convenience we assume here that W_{j-1} and W_j represent the values at the end points of the scaled interval $[0, 1]$, and the formula can then be written in the form

$$P_j^{[p]}(x) = W_{j-1} + (\Delta_+ W_{j-1})x + \sum_{r=2}^p (\Delta_+^r W_{\nu(r)}) g(\nu(r), x).$$

Here $\nu(r)$ indicates the leftmost point for each difference operator $\Delta_+^r = (E - I)^r$, and the determination of this point is the crucial part of the ENO scheme.

The first degree polynomial $P_j^{[1]}$ is obtained as the straight line through W_{j-1} and W_j . The question is now if we should extend the polynomial to second degree by including W_{j-2} or W_{j+1} . We compute the two second order differences $\Delta_+^2 W_{j-2}$ and $\Delta_+^2 W_{j-1}$, and find out which one is smallest in magnitude. If it is the first one we include W_{j-2} in the interpolation formula, otherwise W_{j+1} . We continue in this way by computing the two possible third order differences and so on, and end up with a polynomial of degree p that hopefully has very small oscillations. Note that all the different polynomials interpolate correctly at the endpoints of the interval, but we need the derivatives of the polynomial, and therefore possible oscillations will influence the solution significantly.

For hyperbolic problems like these, we have shown earlier in this book that for high order discretizations in space, one should use high order methods in time as well for solving the system (11.16). We leave out that part here, with the only remark that high order Runge–Kutta methods are often used.

There is another variant of the ENO schemes called *weighted essentially nonoscillating* schemes (WENO). They are very similar, using the differences $\Delta_+^r W_{\nu(r)}$ as indicators in the choice of points for interpolation. In the ENO case, only one difference is chosen for each order r resulting in a unique stencil as a basis for the final interpolating polynomial. In the WENO case, the approximation is obtained as a weighted combination of all the possible polynomials.

If we are aiming for a final p th degree polynomial $P_j^{[p]}$ associated with the cell $[x_{j-1/2}, x_{j+1/2}]$, there are p possible polynomials $\tilde{P}_{[\mu]}^{[p]}$, $\mu = 1, 2, \dots, p$ to choose from. These are built on $(p+1)$ -point stencils, shifting from a completely left-sided one to a completely right-sided one. There are all together $2p-1$ differences $\Delta_+ W_\mu$, $2p-2$ differences $\Delta_+^2 W_\mu$ and so on, until the p differences $\Delta_+^p W_\mu$. The final choice is then obtained as

$$P_j^{[p]} = \sum_{\mu=1}^p \alpha_\mu \tilde{P}_{[\mu]}^{[p]}.$$

The coefficients α_ν are chosen such that the sum is a convex combination of the polynomials. Furthermore the coefficients are weighted according to the size of $|\Delta_+^r W_\mu|^2$ with smaller weight for larger difference. The exact formulas are obtained from [Liu et al., 1994].

We end this section by presenting a computation made by Johan Larsson at CTR, Stanford University. The so called Shu–Osher problem presented in [Shu and Osher, 1988] is governed by the Euler equations (11.5) with $p = (\gamma - 1)\rho e$, where e is the internal energy and $\gamma = 1.4$. The problem is often used as a model for shock-turbulence interaction. It describes a Mach 3 shock that is moving to the right through a density wave, and the interaction creates sound waves behind the shock. The computation of the numerical solution is done by using a hybrid method based on a 7th order WENO scheme around the shocks with the Roe flux function (11.14), (11.15). It is a direct generalization of the 5th order method described in [Jiang and Shu, 1996]. The method is coupled to an 8th order centered method in the smooth parts, and the classic 4th order Runge–Kutta method is used for discretization in time. The choice between the two methods is determined for each time step by computing the WENO weights, but only for the density ρ . The complete WENO scheme is then applied for each stage of the Runge–Kutta step.

The computational domain is $-5 \leq x \leq 5$, and the initial data are

$$x < -4 :$$

$$\rho = 3.857143$$

$$u = 2.629369$$

$$p = 10.333333$$

$$x \geq 4 :$$

$$\rho = 1 + 0.2 \sin(5x)$$

$$u = 0$$

$$p = 1 .$$

Figure 11.8 shows the result for 400 grid points at $t = 1.8$. The WENO scheme has been applied where the solid line is thick, and the centered scheme where it is thin.

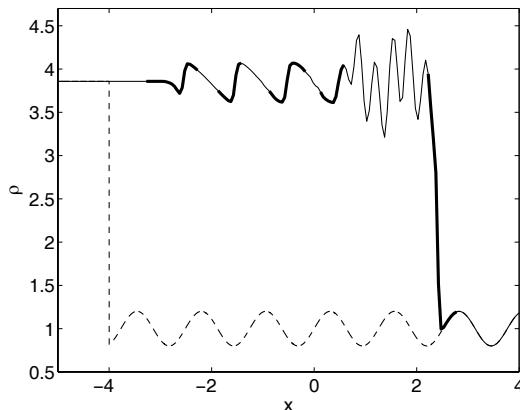


Fig. 11.8 The Shu-Osher problem, WENO method, density ρ , $N = 400$, $t = 0$ (—), $t = 1.8$ (—)

11.7 Summary

The construction of high order methods for shock problems is more difficult than for problems with smooth solutions. In this chapter we have discussed the basic principles. For systems of conservation laws, there are essentially two main classes of methods, namely the ones that are applied directly to the systems, and the ones which require a transformation of the variables for identification of the eigensystem of the Jacobian of the flux function (field by field decomposition). Obviously, the latter ones are more expensive, but should be expected to give more accurate results.

ENO and WENO methods may up to now be the most interesting class of methods for high order approximations, since they have the same basic structure for any order of accuracy. ENO was introduced by Harten et.al. in [Harten et al., 1987], and WENO by Liu et.al. in [Liu et al., 1994]. Interestingly, there are ways of designing such methods such that the field by field decomposition is avoided, see [Liu and Osher, 1998].

Since smooth solutions are always a condition for high accuracy, also these methods deteriorate in accuracy around the shock, but as soon as the solution becomes smooth, their accuracy rises automatically. There is one word of warning here, that we have not discussed. For systems of conservation laws, there are characteristics that go out from the shock area. If the accuracy is low at the shock, do these characteristics transport the larger error into other parts of the computational domain? There is an interesting paper about this by Engquist and Sjögren [Engquist and Sjögren, 1998], but in general it is not clear when it happens.

The semidiscrete methods require time discretizations. Since the properties around the shock are important to preserve, high order discretizations may be tricky to construct. For example, how can we keep the nonoscillating or TVD property? In [Shu and Osher, 1988] a class of Runge–Kutta type methods was developed in this spirit. The basic idea is to assume that the Euler forward method is strongly stable in the total variation norm, and then develop high order methods that are also strongly stable, possibly with a reduction of the time step. These methods were called *TVD time discretizations*. This idea has later been developed further, and a survey is given by Gottlieb et.al. in [Gottlieb et al., 2001]. The methods are there relabeled as *strong stability-preserving(SSP)*.

We have left out multi-dimensional problems in this chapter, and these are of course the interesting ones for real applications. Viscous methods that don't require any transformations, can be generalized without much more trouble than for linear problems. The additional difficulty is associated with the transition mechanism that is switching off the viscosity. The difficulty is much more pronounced for upwind methods. One approach is to split the methods such that they are applied in each space direction sequentially. However, it is difficult to achieve more than second order accuracy this way. Furthermore, the 1-D theory is easiest to generalize if one of the local 1-D directions is perpendicular to the shock, and such coordinate systems are difficult to construct.

Some of the methods described above were partly based on cell averages, which are closely associated with *finite volume methods*. These are based on the integrated

form of the conservation law, and lead automatically to conservative methods, also in several space dimensions. Finite volume methods are described in the final chapter of this book.

Much of the mathematical theory for shocks is due to Lax, and the essential part of it is found in [Lax, 1957] and [Lax, 1972]. Actually, he won the prestigious Abel price 2005, with part of the motivation given to his work on conservation laws. Regarding early computational methods, there was the important paper [Godunov, 1959] by Godunov, where the Godunov method was introduced. During the fifties and sixties, the computations were dominated by dissipative (or viscous) methods like the Lax–Wendroff method [Lax and Wendroff, 1960] and the MacCormack method [MacCormack, 1969]. Among all the people who have made significant contributions during the very active last decades, we find (in alphabetic order) P. Collella, B. Engquist, J.P. Goodman, A. Harten, R. LeVeque, S. Osher, P.L. Roe, C. Shu, P-K. Sweby, E. Tadmor, B. van Leer and P. Woodward. A large part of the important references to the work of these people and others are found in the book by LeVeque [LeVeque, 1992].

The development of computational methods for shocks has been driven largely by applications, particularly in gas dynamics. Even with a considerable amount of theory available at this time, significant parts of the theoretical basis is often missing. This is particularly true when it comes to systems of conservation laws, and to problems in several space dimensions. Shock problems are a striking example of a class of mathematical problems where the practical use of computational methods plays a significant rule in science and engineering well ahead of the theoretical mathematical/numerical analysis.

Chapter 12

Introduction to Other Numerical Methods

This book is about difference methods. However, in this chapter we give a brief introduction to other methods, without going into theoretical details. The chapter covers the most common and general methods, which are finite element methods (FEM), discontinuous Galerkin methods (DG), spectral methods and finite volume methods (FVM). More specialized methods, like the method of characteristics, are not included here.

12.1 Finite Element Methods

Finite element methods were originally developed for elliptic problems in a variational formulation, i.e., the solution is required to minimize an integral representing an energy for the system. For time dependent problems, FEM are usually derived from the classic *Galerkin formulation*, but there is also the *Petrov–Galerkin formulation*. Recently, a new class of approximations called *Discontinuous Galerkin methods* has been developed, and we devote a separate section to that one.

12.1.1 Galerkin FEM

For illustration we use the model problem

$$\begin{aligned} u_t &= (au_x)_x, \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= 0, \\ u_x(1, t) &= 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{12.1}$$

where $a = a(x) \geq a_0 > 0$. The scalar product and norm are defined in the usual way by

$$(u, v) = \int_0^1 uv dx, \quad \|u\|^2 = (u, u).$$

Instead of requiring that the differential equation be satisfied for all x in the domain, we take the scalar product with a *test function* $v = v(x)$:

$$(u_t - (au_x)_x, v) = 0,$$

and require that this equation holds for all test functions v belonging to some function space \mathcal{S} . In order to make these equations meaningful when the solution is not smooth, an integration by parts is carried out. Assuming that u and v satisfy the boundary conditions, we obtain

$$\int_0^1 (u_t - (au_x)_x) v dx = \int_0^1 (u_t v + au_x v_x) dx.$$

The function space \mathcal{S} containing the solution u and the test function v is defined as

$$\mathcal{S} = \{v / \|v\|^2 + \|v_x\|^2 < \infty, v(0) = 0\}.$$

Note that only the *essential boundary condition* $v(0) = 0$ is imposed. We shall later discuss the boundary condition at the right. The Galerkin formulation can now be defined:

Find $u \in \mathcal{S}$ for each $t > 0$, such that

$$(u_t, v) + (au_x, v_x) = 0 \tag{12.2}$$

for all $v \in \mathcal{S}$, and such that $u(x, 0) = f(x)$. \square

The Galerkin formulation is the same as the weak form that was introduced for conservation laws in Section 11.2. It is of course no numerical method, it is just a weaker form of the original problem, in the sense that it allows for less smooth solutions. However, it is also a good foundation for a numerical method, by substituting \mathcal{S} by a finite dimensional subspace $\mathcal{S}_h \subset \mathcal{S}$. The functions in \mathcal{S}_h are defined everywhere, but they are associated with the finite dimension $N = 1/h$, where h is some typical step size in the x -direction. We denote the test functions by $v^h(x)$, and the approximate solution by $u^h(x, t)$. The discrete Galerkin method is:

Find $u^h \in \mathcal{S}_h$ for each $t > 0$, such that

$$(u_t^h, v^h) + (au_x^h, v_x^h) = 0 \tag{12.3}$$

for all $v^h \in \mathcal{S}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

Here it is assumed that f^h is an approximation of f .

Let us next see how (12.3) can be expressed in a more concrete way. Assume that the set of functions $\{\phi_j(x)\}_{j=1}^N$ is a basis in \mathcal{S}_h , i.e., all functions $v^h \in \mathcal{S}_h$ can

be expressed as a linear combination of the functions $\{\phi_j\}$. The solution $u^h(x, t)$ is written in the form

$$u^h(x, t) = \sum_{j=1}^N \alpha_j(t) \phi_j(x),$$

where $\{\alpha_j(t)\}$ are the coefficients to be determined by (12.3). Since v^h also can be represented as a linear combination of $\{\phi_j\}$, (12.3) is satisfied for all $v^h \in \mathcal{S}_h$ if and only if it is satisfied for all ϕ_j . Accordingly, the coefficients α_j are given by the system

$$\begin{aligned} \sum_{j=1}^N ((\phi_j, \phi_\nu) \frac{d\alpha_j}{dt} + (a \frac{d\phi_j}{dx}, \frac{d\phi_\nu}{dx}) \alpha_j) &= 0, \quad \nu = 1, 2, \dots, N, \\ \alpha_j(0) &= \hat{f}_j, \quad j = 1, 2, \dots, N, \end{aligned} \tag{12.4}$$

where $f^h = \sum \hat{f}_j \phi_j$ is an approximation of f . This is an initial value problem for a system of ordinary differential equations, and when the functions ϕ_j have been selected, it still remains to solve this ODE system numerically.

Still nothing has been said that motivates the name finite element method. The formulation (12.3) is very general, it just requires that the subspace \mathcal{S}_h is an approximation of \mathcal{S} in the sense that for each function $v \in \mathcal{S}$ there is a function $v^h \in \mathcal{S}_h$ such that $\|v - v^h\|$ is small. A finite element method is obtained if the basis functions ϕ_j are zero everywhere except in a small interval I_j . Usually these basis functions are piecewise polynomials. Since $\{\phi_j\}$ is a basis in \mathcal{S}_h , it is necessary that the subintervals I_j are distributed over the whole interval $[0, 1]$, and therefore only a few of the coefficients multiplying $d\alpha_j/dt$ and α_j in (12.4) are nonzero.

We consider next a simple choice of \mathcal{S}_h , namely the space of continuous and piecewise linear functions. In contrast to finite difference methods, FEM is well suited for use on nonuniform grids, and we define the basis functions by

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/h_{j-1}, & x_{j-1} \leq x < x_j \\ (x_{j+1} - x)/h_j, & x_j \leq x < x_{j+1} \\ 0 & \text{otherwise.} \end{cases} \tag{12.5}$$

Here x_j , $j = 0, 1, \dots, N$ are the grid points (usually called nodes, when dealing with FEM) distributed in the interval $[0, 1]$, and h_j is the length of the subinterval $[x_j, x_{j+1}]$. The basis functions are often called the roof or hat functions, and they are shown in Figure 12.1.

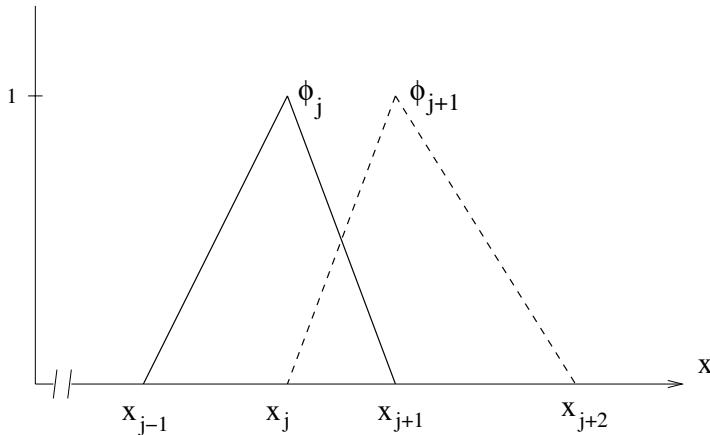


Fig. 12.1 Piecewise linear basis functions

Since $\phi_j(x_j) = 1$, we have $u^h(x_j, t) = \alpha_j(t)$, which shows that $\alpha_j(t)$ is the value of the numerical solution $u^h(x, t)$ at the node $x = x_j$. With the vector $\boldsymbol{\alpha}$ defined as $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_N)^T$, the ODE system can be written in the form

$$\begin{aligned} M \frac{d\boldsymbol{\alpha}}{dt} + K\boldsymbol{\alpha} &= 0, \\ \boldsymbol{\alpha}(0) &= \hat{\boldsymbol{\alpha}}, \end{aligned} \tag{12.6}$$

where M is called the mass matrix and K the stiffness matrix. Since the nonzero part of a given basis function ϕ_j overlaps the nonzero part of only the two adjacent functions ϕ_{j-1} and ϕ_{j+1} , both matrices M and K are tridiagonal. For the more general case, they may be wider, but they are still band matrices. This is similar to semidiscrete difference approximations as we have seen earlier in this book.

For the piecewise linear case and $a \equiv 1$, integration gives the matrix elements

$$\begin{aligned} M_{j,j+1} &= (\phi_{j+1}, \phi_j) = \frac{h_j}{6}, \quad j = 1, 2, \dots, N-1, \\ M_{j,j} &= (\phi_j, \phi_j) = \frac{2(h_{j-1} + h_j)}{6}, \quad j = 1, 2, \dots, N-1, \\ M_{j-1,j} &= (\phi_{j-1}, \phi_j) = \frac{h_{j-1}}{6}, \quad j = 2, 3, \dots, N, \\ M_{N,N} &= (\phi_N, \phi_N) = \frac{2h_{N-1}}{6}, \end{aligned}$$

$$\begin{aligned} K_{j,j+1} &= \left(\frac{d\phi_{j+1}}{dx}, \frac{d\phi_j}{dx} \right) = -\frac{1}{h_j}, \quad j = 1, 2, \dots, N-1, \\ K_{j,j} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_j}{dx} \right) = \frac{1}{h_{j-1}} + \frac{1}{h_j}, \quad j = 1, 2, \dots, N-1, \\ K_{j-1,j} &= \left(\frac{d\phi_j}{dx}, \frac{d\phi_{j-1}}{dx} \right) = -\frac{1}{h_{j-1}}, \quad j = 2, 3, \dots, N, \\ K_{N,N} &= \left(\frac{d\phi_N}{dx}, \frac{d\phi_N}{dx} \right) = \frac{1}{h_{N-1}}. \end{aligned}$$

Note that only the left half of the rightmost basis function ϕ_N is involved in the integration. For a uniform grid with $h_j = h$, we get after division by h

$$M = \frac{1}{6} \begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{bmatrix}, \quad K = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}.$$

There is a direct similarity with a difference approximation, since K represents the standard 3-point difference formula for $-u_{xx}$. However, the mass matrix M represents a weighted average over three points, instead of the natural identity matrix.

For discretization in time, a difference method is usually chosen. For an explicit method, there is a fundamental difference compared to the case where a difference method is used in space. Since the matrix M is not diagonal, FEM require the solution of a system of algebraic equations in each step also in the explicit case.

Stability for the Galerkin FEM is easy to prove by using the energy method. Consider the Galerkin equation (12.2), and choose $v(x)$ as the solution $u(x, t)$, where t is considered as a parameter. This can be done since both functions belong to the same function space \mathcal{S} . We get

$$\frac{d}{dt} \|u\|^2 = 2(u, u_t) = -2(a u_x, u_x) \leq -2a_0 \|u_x\|^2 \leq 0,$$

which shows that the problem is well posed.

The same technique is used to prove stability for the approximation (12.3). The test function v^h is chosen as the solution u^h , and we get also here

$$\frac{d}{dt} \|u^h\|^2 = 2(u^h, u_t^h) = -2(a u_x^h, u_x^h) \leq -2a_0 \|u_x^h\|^2 \leq 0,$$

which leads to

$$\|u^h(\cdot, t)\| \leq \|u^h(\cdot, 0)\| = \|f^h(\cdot)\|.$$

This is a great advantage with the Galerkin FEM: if there is an energy estimate for the original problem, stability follows automatically for the approximation. (It remains of course to discretize in time.)

But what happened to the boundary conditions, which we have seen playing such an important role for stability? The condition $u(0, t) = 0$ was imposed by including it into the definition of the solution space \mathcal{S} , but we have not even bothered about the condition $u_x(1, t) = 0$! Let us investigate this and assume that there is a smooth solution u of (12.2) which belongs to \mathcal{S} for all t . After integration by parts, we get

$$0 = \int_0^1 (u_t v + au_x v_x) dx = \int_0^1 (u_t - (au_x)_x) v dx + a(1)u_x(1, t)v(1),$$

which holds for all $v \in \mathcal{S}$. Obviously the original differential equation in (12.1) must be satisfied. Furthermore, since $v(1)$ is arbitrary, the solutions must satisfy the boundary condition $u_x(1, t) = 0$. This condition is therefore called a *natural boundary condition*. It is not included in the Galerkin formulation, but it is satisfied for all smooth solutions u .

Let us next see, how the Galerkin FEM handles this. For convenience we assume again that $a(x) \equiv 1$, and furthermore that the grid is uniform. When using the finite difference notation u_j for α_j , the last equation of the ODE system in (12.6) reads

$$\frac{h}{6} \left(\frac{du_{N-1}}{dt} + 2 \frac{du_N}{dt} \right) = \frac{u_{N-1} - u_N}{h}. \quad (12.7)$$

In the usual finite difference manner, we substitute a smooth solution u of the differential equation into (12.7). A Taylor expansion around $x = 1$ gives

$$\frac{h}{6} \left(3u_t - hu_{tx} + \frac{h^2}{2}u_{txx} + \mathcal{O}(h^3) \right) = -u_x + \frac{h}{2}u_{xx} - \frac{h^2}{6}u_{xxx} + \frac{h^3}{24}u_{xxxx} + \mathcal{O}(h^4).$$

By differentiating the equation $u_t = u_{xx}$, we get

$$u_{tx} = u_{xxx}, \quad u_{txx} = u_{xxxx},$$

giving

$$0 = -u_x - \frac{h^3}{24}u_{xxxx} + \mathcal{O}(h^4).$$

This shows that if the finite element is considered as a difference method, it produces a third order accurate approximation of the boundary condition $u_x(1, t) = 0$. Accordingly, this boundary condition is "natural" also for the Galerkin FEM.

In applications, the boundary conditions are in general not homogeneous, and we consider next the problem

$$\begin{aligned} u_t &= (au_x)_x + F(x, t), \quad 0 \leq x \leq 1, \quad 0 \leq t, \\ u(0, t) &= g_0(t), \\ u_x(1, t) + \beta u(1, t) &= g_1(t), \\ u(x, 0) &= f(x). \end{aligned} \quad (12.8)$$

Here we have also included a forcing function in the differential equation, as well as a zero order term in the Neumann boundary condition at $x = 1$. In contrast to difference schemes, the generalization of FEM to this problem is not immediately obvious. The function space \mathcal{S} for our previous example is denoted by H_0^1 , where the subscript refers to the boundary condition, and the superscript indicates that the first derivatives are in L_2 . We now define the new space H_E^1 by

$$H_E^1 = \{v(x, t) / \|v(\cdot, t)\|^2 + \|v_x(\cdot, t)\|^2 < \infty, v(0, t) = g_0(t)\},$$

where t is a parameter. Obviously H_E^1 is not a linear space, since the sum w of any two functions in H_E^1 satisfies the boundary condition $w(0, t) = 2g_0(t)$ instead of $w(0, t) = g_0(t)$. The nonzero data in the PDE and in the boundary condition at $x = 1$ require extra terms in the weak formulation:

Find $u \in H_E^1$ for each $t > 0$, such that

$$(u_t, v) + (au_x, v_x) = (F, v) - \beta a(1, t)u(1, t)v(1) + g_1(t)a(1, t)v(1)$$

for all $v \in H_0^1$, and such that $u(x, 0) = f(x)$. \square

If u is smooth, integration by parts results in the relation

$$(u_t, v) = ((au_x)_x, v) + (F, v) - a(1, t)(u_x(1, t) + \beta u(1, t) - g_1(t))v(1).$$

The consistency with the original problem is seen by multiplying the PDE in (12.8) by $v(x)$ and integrating, and then add the boundary expression $u_x(1, t) + \beta u(1, t) - g_1(t)$ (which is zero) multiplied by $-a(1, t)v(1)$.

To define the Galerkin numerical method, we work with the subspace $\mathcal{S}_h \subset H_0^1$:

Find $u^h \in \mathcal{S}_h$ for each $t > 0$, such that

$$(u_t^h, v^h) + (au_x^h, v_x^h) = (F, v^h) - \beta a(1, t)u_x^h(1, t)v^h(1) + g_1(t)a(1, t)v^h(1)$$

for all $v^h \in \mathcal{S}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

For a fourth order PDE

$$u_t = -(au_{xx})_{xx},$$

integration by parts is carried out twice in order to get the weak formulation

$$(u_t, v) = -(au_{xx}, v_{xx}).$$

As an example of boundary conditions, we take

$$\begin{aligned} u(0, t) &= 0, & u_{xx}(1, t) &= 0, \\ u_x(0, t) &= 0, & u_{xxx}(1, t) &= 0. \end{aligned}$$

The function space for the solution u and the test functions is then defined by

$$H_0^2 = \{v(x) / ||v_{xx}||^2 + ||v_x||^2 + ||v||^2 < \infty, v(0) = 0, v_x(0) = 0\}.$$

One can show that the conditions at $x = 1$ are natural boundary conditions in the sense discussed above, and they are not included explicitly in the definition of the solution space. The Galerkin approximation is

Find $u^h \in \mathcal{S}_h \subset H_0^2$ for each $t > 0$, such that

$$(u_t^h, v^h) = -(au_{xx}^h, v_{xx}^h)$$

for all $v^h \in \mathcal{S}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

For this problem, the requirement of smoothness is stricter, since also the second derivatives must be in L_2 . This means that the piecewise linear elements above cannot be used, we must make sure that the functions have continuous first derivatives. We shall discuss different types of finite elements later, but first derive some basic results concerning accuracy.

A general parabolic equation has the form

$$u_t + Pu = F,$$

where P is a positive definite differential operator of order $2q$ in the space \mathcal{S} , where the essential boundary conditions are imposed. Using integration by parts in analogy with the derivation for the model problems above, the weak form for $u \in \mathcal{S}$ is

$$\begin{aligned} (u_t, v) + (u, v)_* &= (F, v) \quad \text{for all } v \in \mathcal{S}, \\ u(x, 0) &= f(x). \end{aligned}$$

Here $(u, v)_*$ is a generalized scalar product; for the first model example above we had

$$(u, v)_* = \int_0^1 au_x v_x dx,$$

where a is a positive function. The Galerkin approximation is defined by

Find $u^h \in \mathcal{S}_h \subset \mathcal{S}$ for each $t > 0$, such that

$$(u_t^h, v^h) + (u^h, v^h)_* = (F, v^h) \tag{12.9}$$

for all $v^h \in \mathcal{S}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

When deriving error estimates for difference methods, the true solution is inserted into the difference scheme, and the truncation error occurs as a forcing function. The error estimate then follows by stability. This procedure does not work for FEM. The reason is that the true solution u cannot be inserted into (12.9), since in general it is not in the space \mathcal{S}_h . Therefore we must make a detour via the least square approximation Qv in \mathcal{S}_h to functions v in \mathcal{S} with distances measured in the

special norm $(v, v)_*^{1/2}$. In fact, one can show that the error estimate for the Galerkin FEM can be converted to a pure function approximation problem. Let Q be the *projection operator* such that $Qv \in \mathcal{S}_h$ and

$$(v - Qv, v^h)_* = 0 \text{ for all } v^h \in \mathcal{S}_h.$$

This is the condition for a least square approximation of v , i.e., Qv minimizes $\|v - v^h\|_*^2$ over all $v^h \in \mathcal{S}_h$. A fundamental result from approximation theory is the following lemma:

Lemma 12.1. *Assume that h is the largest subinterval in the partition of $[0, 1]$, and that $\mathcal{S}_h \subset \mathcal{S}$ consists of all piecewise polynomials of degree $r - 1$. Then the estimate*

$$\|v - Qv\| \leq Kh^r \|v\|_r$$

holds, where

$$\|v\|_r^2 = \sum_{\nu=0}^r \left\| \frac{d^\nu v}{dx^\nu} \right\|^2.$$

□

It is then possible to prove that if the solution u is smooth enough, and $\mathcal{S}_h \subset \mathcal{S}$ consists of all piecewise polynomials of degree $r - 1$, then there is an error estimate

$$\|u(\cdot, t) - u^h(\cdot, t)\| \leq Kh^r.$$

These results reduces the construction of FEM to the problem of constructing the space \mathcal{S}_h , under the restriction that the functions are at least as smooth as the functions in \mathcal{S} . For a parabolic problem of order $2q$, the functions must have continuous derivatives of at least order $q - 1$.

The basis functions for piecewise polynomials of a given degree $r - 1$ on each interval I_j are not unique, but some bases are more convenient than others. In the example above, we saw that the hat functions with the peak value 1 at the nodes provide a convenient basis. Since we are interested in higher order methods, we shall discuss a few other spaces \mathcal{S}_h .

We begin with *piecewise quadratics*. The nodes x_j , $j = 0, 1, \dots, N$ are the points where the derivatives may be discontinuous. On each subinterval

$$I_j = [x_{j-1}, x_j],$$

v^h is determined by three parameters. The essential boundary condition $v^h(0) = 0$ leaves two parameters undetermined in I_1 ; for $j = 2, 3, \dots, N$, continuity at x_j leaves two parameters undetermined in each subinterval I_j . Accordingly, $\text{Dim}(\mathcal{S}_h) = 2N$. This means that there are two basis functions associated with each subinterval. This requires an inner node, and we choose the midpoint $x_{j-1/2}$. The basis functions $\phi_{j-1/2}, \phi_j$ are defined as

$$\phi_j(x) = \begin{cases} 1 & \text{at } x = x_j \\ 0 & \text{at all other nodes} \\ \text{quadratic polynomial in each } I_\nu & \end{cases}$$

$$\phi_{j-1/2}(x) = \begin{cases} 1 & \text{at } x = x_{j-1/2} \\ 0 & \text{at all other nodes} \\ \text{quadratic polynomial in each } I_\nu & \end{cases}$$

These functions are shown in Figure 12.2.

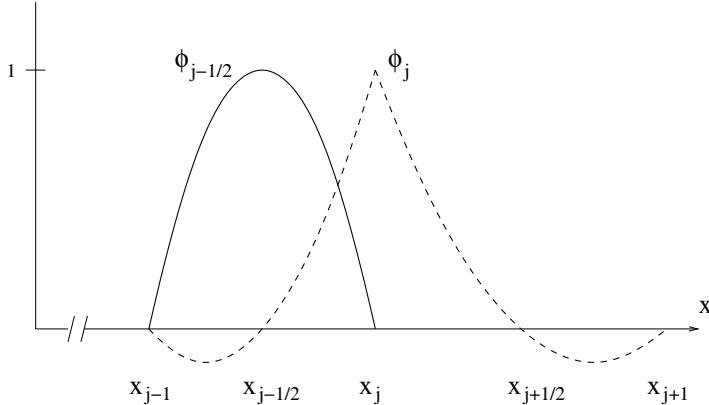


Fig. 12.2 Piecewise quadratic basis functions

Note the difference between the original nodes x_j and the extra nodes $x_{j-1/2}$. Any function

$$v^h = \sum_{j=1}^N (\alpha_j \phi_j + \alpha_{j-1/2} \phi_{j-1/2})$$

has continuous derivatives of any order across $x = x_{j-1/2}$, but in general the first derivatives are discontinuous across $x = x_j$.

For *piecewise cubics* the same kind of argument as above leads to $\text{Dim}(\mathcal{S}_h) = 3N$. Two extra nodes $x_{j-2/3}, x_{j-1/3}$ are introduced in each subinterval I_j , and the basis functions $\phi_{j-2/3}, \phi_{j-1/3}, \phi_j$ equal 1 at the corresponding nodes. With the additional requirement of being zero at all other nodes and cubic in each subinterval I_j , they are uniquely determined. They are shown in Figure 12.3.

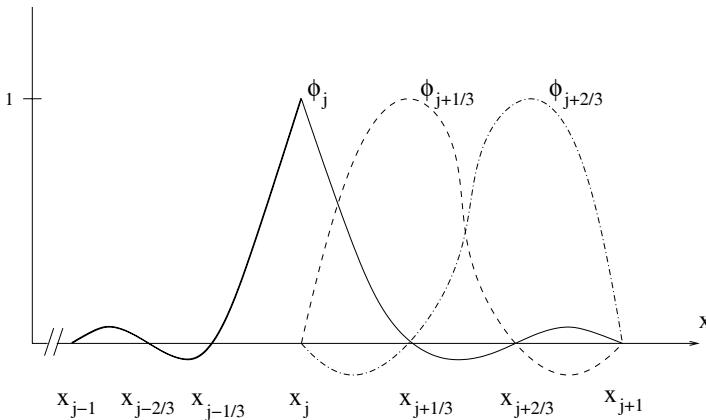


Fig. 12.3 Piecewise cubic basis functions

For PDE with differential operators of order 4 in space, we saw above that the first derivatives of the functions in the approximating space must be continuous. This means that for a cubic, we add still another restriction at each node, and the dimension of the space \mathcal{S}_h goes down to $2N + 1$. The basis functions include also the node values of $d\phi_j/dx$, and we arrive at the classic *Hermite polynomials*.

The generalization of Galerkin FEM to several space dimensions is easy at an abstract level. Consider the heat equation in the unit square Ω with boundary $\partial\Omega$:

$$\begin{aligned} u_t &= u_{xx} + u_{yy} \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \\ u(x, y, 0) &= f(x, y). \end{aligned}$$

The scalar product is defined by

$$(u, v) = \int_0^1 \int_0^1 uv dx dy,$$

and we get using integration by parts

$$(u_{xx} + u_{yy}, v) = -(u_x, v_x) - (u_y, v_y).$$

The space \mathcal{S} for the weak form is

$$\mathcal{S} = \{u / \|u\|^2 + \|u_x\|^2 + \|u_y\|^2 < \infty, \quad u = 0 \text{ on } \partial\Omega\}.$$

The Galerkin approximation is then formulated in analogy with the 1-D case:

Find $u^h \in \mathcal{S}_h \subset \mathcal{S}$ for each $t > 0$, such that

$$(u_t^h, v^h) + (u_x^h, v_x)^h + (u_y^h, v_y^h) = 0$$

for all $v^h \in \mathcal{S}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

The remaining question is how to construct the subspace \mathcal{S}_h . For our example, where the computational domain is a square, the 1-D basis functions can be used for the construction. We define the nodes x_i , $i = 0, 1, \dots, M$, y_j , $j = 0, 1, \dots, N$, and let $\phi_i(x)$, $i = 1, 2, \dots, M - 1$, $\phi_j(y)$, $j = 1, 2, \dots, N - 1$, be defined as above. Then we use

$$\psi_{ij}(x, y) = \phi_i(x)\phi_j(y), \quad i = 1, 2, \dots, M - 1, \quad j = 1, 2, \dots, N - 1$$

as the 2-D basis functions. These are piecewise bilinear, i.e., they are piecewise linear in one variable if the other variable is held constant. Since the basis functions are zero in most of the domain, the system of algebraic equations to be solved for each time step has a sparse coefficient matrix also in this case.

A general linear function in 2-D has the form $v^h(x, y) = a_0 + a_1x + a_2y$, i.e., it is defined by 3 values. Therefore, triangles are more natural, with the nodes at the corners. Furthermore, a great advantage with triangles is that they allow for greater flexibility in the partitioning of the domain, particularly if the domain is irregular. One basis function is associated with each corner node, where it is defined as one. At all other nodes in the domain it is zero, and this gives a piecewise linear function which is nonzero only at the triangles with one corner at the particular node, see Figure 12.4.

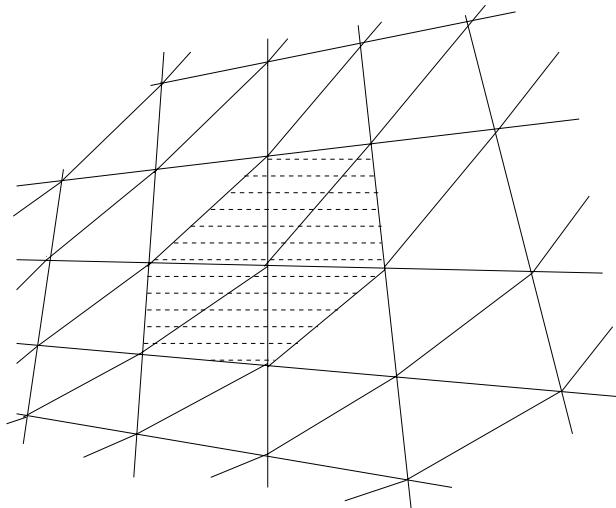


Fig. 12.4 Domain where one basis function is nonzero

This enforces continuity at nodes, but is it continuous also along the edges of the triangles, which is required by the theory? Consider two neighboring triangles with a common edge defined by

$$y = \alpha + \beta x, \quad x_0 \leq x \leq x_1.$$

On this line segment the piecewise linear functions have the form

$$v^h(x, y) = a_0 + a_1 x + a_2(\alpha + \beta x) = b_0 + b_1 x, \quad x_0 \leq x \leq x_1.$$

Since the values at the end points uniquely determine the values all along this line, we have continuity across the boundary.

The generalization to higher degree polynomials is analogous to the 1-D case. For quadratics, we have 6 coefficients to determine on each triangle, and we introduce one new node at the midpoint of each side, see Figure 12.5. Since the functions are quadratic along each side, it is uniquely determined by the three node values, and continuity is secured.

For cubics, there are 10 coefficients to determine on each triangle. Besides the corners, we introduce 2 nodes on each side partitioning it into 3 equal pieces. One coefficient remains, and we introduce the 10th node at the center of gravity, see Figure 12.5.

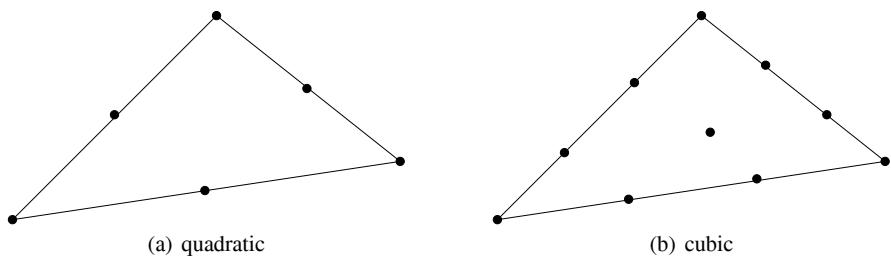


Fig. 12.5 Nodes for quadratic and cubic functions

By the same type of arguments as used for quadratics, we conclude that the resulting functions are continuous across element boundaries.

If the boundary Γ is piecewise linear, the triangular elements allow for a perfect representation of the domain. However, for curved boundaries, we must make some approximation. If the domain is convex, the extension is easy. The space \mathcal{S}_h consists of all piecewise linear functions v^h which are extended to be zero in the area outside Γ_h , see Figure 12.6.

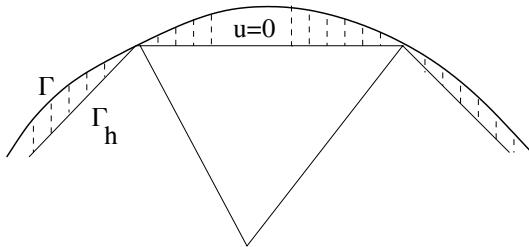


Fig. 12.6 Domain for extended space S_h

In this way continuity is secured, and the whole theory leading to the final error estimate can be applied. However, we have to make sure that the basic approximation property holds. In other words, how well do the functions in \mathcal{S}_h approximate the functions in \mathcal{S} ? The critical part is near the boundary, where v^h is zero at Γ . But the maximal distance between Γ and Γ_h is of the order h^2 if the triangle sides are of order h . Therefore, if the true solution u is smooth in Ω and zero at Γ , it cannot be larger in magnitude than $\mathcal{O}(h^2)$ at Γ_h . Accordingly, for linear elements, the simple procedure described here is sufficient to obtain second order accuracy. For higher order elements, more sophisticated procedures must be used, with one edge of the triangles made nonlinear by using higher degree polynomials.

In three space dimensions, the natural choice is to define the piecewise polynomials on tetrahedrons. For example, a linear function has the form $u^h = a_0 + a_1x + a_2y + a_3z$, and since there are four corners, the nodes are located there.

For the piecewise linear 1-D case we used the explicit form of the basis functions when deriving the form of the mass and stiffness matrices. For real computations, these matrices are found in a more direct way without knowing the specific form of the basis functions. The general form of the piecewise polynomial is known on each subdomain, and the elements of the matrices can be computed by a summation of all neighboring contributions.

Let us finally make an interesting comparison with the compact Padé type approximations on uniform grids treated in Section 4.3. Consider the equation $u_t = u_x$ and the piecewise linear basis functions $\phi_j(x)$ defined in (12.5). The Galerkin formulation is

$$\sum_{j=1}^N \frac{d\alpha_j(t)}{dt} \int_0^1 \phi_j(x) \phi_\nu(x) dx = \sum_{j=1}^N \alpha_j(t) \int_0^1 \phi'_j(x) \phi_\nu(x) dx, \quad \nu = 1, 2, \dots, N.$$

The integrals on the left hand side are the same as for the parabolic model problem above, leading to the same tridiagonal mass matrix. On the right side we have

$$\int_0^1 \phi'_{j-1} \phi_j dx = -\frac{1}{2}, \quad \int_0^1 \phi'_j \phi_j dx = 0, \quad \int_0^1 \phi'_{j+1} \phi_j dx = \frac{1}{2},$$

leading to

$$\frac{1}{6} \left(\frac{d\alpha_{j+1}}{dt} + 4 \frac{d\alpha_j}{dt} + \frac{d\alpha_{j-1}}{dt} \right) = \frac{1}{2h} (\alpha_{j+1} - \alpha_{j-1}).$$

Since $u^h(x_j, t) = \alpha_j(t)$, we can consider the approximation as a difference scheme, and it has the form

$$\frac{d\alpha_j}{dt} = (I + \frac{h^2}{6} D_+ D_-)^{-1} D_0 \alpha_j. \quad (12.10)$$

But this is exactly the fourth order Padé type approximation derived above.

From a finite element point of view, this is an interesting observation. From a piecewise linear subspace spanned by the basis functions above, one expects a second order accurate solution. Obviously, we get a fourth order approximation at the nodes, and this is called *superconvergence*, see [Thomée and Wendroff, 1974]. However, it is obtained only for the very special case with constant coefficients, while for the finite difference approximation, it can be generalized to more general equations as demonstrated in Section 4.3.

12.1.2 Petrov–Galerkin FEM

If the PDE has an energy conserving property, the Galerkin approximations are also energy conserving. A simple example of this is the equation $u_t = u_x$. For the special case of a uniform grid, we get the approximation (12.10), where α_j is the solution at $x = x_j$. For periodic solutions, we have the Fourier transform

$$\frac{d\hat{\alpha}}{dt} = \frac{3i \sin \xi}{h(2 + \cos \xi)} \hat{\alpha},$$

i.e., $|\hat{\alpha}(t)| = |\hat{\alpha}(0)|$. For certain applications it is necessary to have some dissipation in the approximation, for example if the true solution has discontinuities. In such a case, the numerical solution will be highly oscillatory since there is no damping of the high frequencies. For difference methods, we saw in the previous chapter that artificial viscosity can be introduced by adding extra terms. For FEM, one can instead generalize the Galerkin formulation by dropping the requirement that the test space is the same as the solution space \mathcal{S}_h . If the test space is chosen different from \mathcal{S}_h , we get a *Petrov–Galerkin* method. For a general PDE $u_t = Pu$, the formulation is

Find $u^h \in S_h$ for each $t > 0$, such that

$$(u_t^h, v^h) = (Pu^h, v^h)$$

for all $v^h \in \mathcal{R}_h$, and such that $u^h(x, 0) = f^h(x)$. \square

For our model problem on a uniform grid, the space \mathcal{R}_h can be defined as all functions of the form $v^h - hv_x^h$, where $v^h \in \mathcal{S}_h$. The ODE system becomes

$$\sum_{i=1}^N \left(\frac{d\alpha_i}{dt} \phi_i, \phi_j - h\phi'_j \right) = \sum_{i=1}^N (\alpha_i \phi'_i, \phi_j - h\phi'_j), \quad j = 1, 2, \dots, N.$$

By using the earlier computed integrals, we obtain

$$\frac{1}{6} \frac{d}{dt} (\alpha_{j-1} + 4\alpha_j + \alpha_{j+1}) - hD_0 \frac{d\alpha_j}{dt} = D_0 \alpha_j + hD_+ D_- \alpha_j.$$

After Fourier transformation, we get

$$\left(\frac{2 + \cos \xi}{3} - ih \sin \xi \right) \frac{d\hat{\alpha}}{dt} = \frac{1}{h} \left(i \sin \xi - 4 \sin^2 \frac{\xi}{2} \right) \hat{\alpha},$$

or, equivalently

$$\frac{d\hat{\alpha}}{dt} = \hat{Q} \hat{\alpha}, \quad \hat{Q} = \frac{1}{h} \frac{i \sin \xi - 4h \sin^2 \frac{\xi}{2}}{1 - \frac{2}{3} \sin^2 \frac{\xi}{2} - ih \sin \xi}.$$

It is easily shown that

$$\operatorname{Re} \hat{Q} \leq -\delta |\xi|^2, \quad \delta > 0,$$

i.e., there is a damping of the amplitudes corresponding to nonzero wave numbers. $\hat{Q}(\xi)$ is shown in Figure 12.7 for discrete ξ -values and $h = 1$.

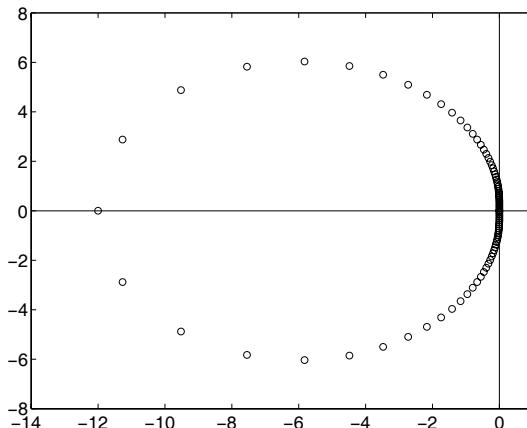


Fig. 12.7 $\hat{Q}(\xi)$, $0 \leq \xi \leq 2\pi$

For more general problems in several space dimensions, the Petrov–Galerkin concept is generalized by choosing the space \mathcal{S}_h as for Galerkin methods described in the previous section, and then choosing \mathcal{R}_h . This latter choice is a more delicate question. One type of generalization used in fluid dynamics is the *streamline-diffusion* method developed by Tom Hughes, Claes Johnson and their coworkers.

There are many papers on this topic, one of them is [Johnson et al., 1990], where a convergence analysis is presented.

12.2 Discontinuous Galerkin Methods

We have demonstrated for Galerkin methods how the subspaces \mathcal{S}_h are constructed such that the numerical solutions u^h have the right regularity properties. If these regularity requirements are relaxed, even allowing for discontinuous solutions, then we get *Discontinuous Galerkin methods*. The nice machinery leading to stability and error estimates are lost in the process, but other advantages are gained.

The starting point is a conservation law

$$u_t + f_x(u) = 0,$$

where the flux function $f(u)$ may be a nonlinear function of u . Referring back to the previous chapter, we consider an interval $I_j = [x_{j-1/2}, x_{j+1/2}]$, multiply by a test function $\phi(x)$ and integrate over I_j :

$$\int_{I_j} u_t \phi \, dx + \int_{I_j} f_x(u) \phi \, dx = 0.$$

Integration by parts gives

$$\int_{I_j} u_t \phi \, dx + f(u(x_{j+1/2})) \phi(x_{j+1/2}) - f(u(x_{j-1/2})) \phi(x_{j-1/2}) - \int_{I_j} f(u) \phi_x \, dx = 0.$$

Already here there is a deviation from the Galerkin formulation, since the integration by parts is not done over the whole domain. The next deviation is that the flux function will be substituted by a numerical flux function \tilde{f} that depends on both the left and right limits at each point. :

$$f(u) \rightarrow \tilde{f}(u^-, u^+).$$

At all points where u is continuous, we construct \tilde{f} such that $\tilde{f}(u^-, u^+) = f(u)$, but here we are going to deal also with discontinuous functions u . In this case we get a numerical solution u^h satisfying

$$\begin{aligned} \int_{I_j} u_t^h \phi \, dx + \tilde{f}(u^h(x_{j+1/2})^-, u^h(x_{j+1/2})^+) \phi(x_{j+1/2}) \\ - \tilde{f}(u^h(x_{j-1/2})^-, u^h(x_{j-1/2})^+) \phi(x_{j-1/2}) - \int_{I_j} f(u^h) \phi_x \, dx = 0. \end{aligned}$$

It remains to choose the subspace with its basis functions and \tilde{f} . For simplicity we choose piecewise linear functions, and pick the basis functions

$$\begin{aligned}\phi_j &= \frac{1}{h_j}(x_{j+1/2} - x), \quad \psi_j = \frac{1}{h_j}(x - x_{j-1/2}), \quad x \in I_j, \\ \phi_j &= \psi_j = 0 \text{ else},\end{aligned}$$

where h_j is the length of I_j . These functions are discontinuous as shown in Figure 12.8.

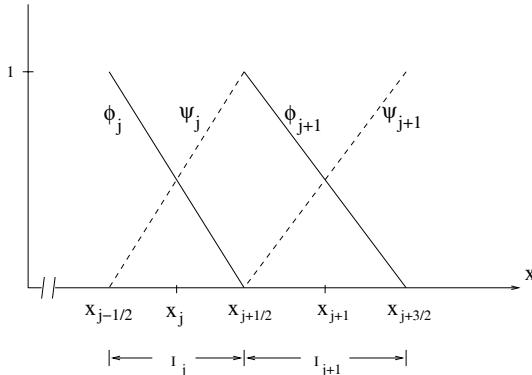


Fig. 12.8 Piecewise linear basis functions

The numerical solution has the form

$$u^h(x, t) = \alpha_j(t)\phi_j(x) + \beta_j(t)\psi_j(x), \quad x \in I_j,$$

where the coefficients α_j and β_j are to be determined. With N intervals I_j in the domain, the weak form of the problem is now obtained as

$$\begin{aligned}&\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\frac{d\alpha_j}{dt} \phi_j + \frac{d\beta_j}{dt} \psi_j \right) \phi_j dx + \tilde{f}(u^h(x_{j+1/2})^-, u^h(x_{j+1/2})^+) \phi_j(x_{j+1/2}) \\&- \tilde{f}(u^h(x_{j-1/2})^-, u^h(x_{j-1/2})^+) \phi_j(x_{j-1/2}) - \int_{x_{j-1/2}}^{x_{j+1/2}} f(\alpha_j \phi_j + \beta_j \psi_j)(\phi_j)_x dx = 0, \\&\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\frac{d\alpha_j}{dt} \phi_j + \frac{d\beta_j}{dt} \psi_j \right) \psi_j dx + \tilde{f}(u^h(x_{j+1/2})^-, u^h(x_{j+1/2})^+) \psi_j(x_{j+1/2}) \\&- \tilde{f}(u^h(x_{j-1/2})^-, u^h(x_{j-1/2})^+) \psi_j(x_{j-1/2}) - \int_{x_{j-1/2}}^{x_{j+1/2}} f(\alpha_j \phi_j + \beta_j \psi_j)(\psi_j)_x dx = 0, \\&j = 1, 2, \dots, N.\end{aligned}$$

The basis functions are locally defined as nonzero only in one subinterval. Hence, at a first glance, it looks like there is no coupling between the different subintervals. However, this coupling is obtained by the choice of flux function \tilde{f} . Let us study the simple equation $u_t + u_x = 0$, i.e., $f(u) = u$. Since information is flowing from left to right, it is natural to make the solution in I_j dependent on what is going on in I_{j-1} .

A simple choice is

$$\tilde{f}(u^-, u^+) = u^-,$$

and the form of the basis functions then imply

$$\begin{aligned}\tilde{f}(u^h(x_{j-1/2})^-, u^h(x_{j-1/2})^+) &= u^h(x_{j-1/2})^- = \beta_{j-1}, \\ \tilde{f}(u^h(x_{j+1/2})^-, u^h(x_{j+1/2})^+) &= u^h(x_{j+1/2})^- = \beta_j.\end{aligned}$$

This leads to the approximation

$$\begin{aligned}\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\frac{d\alpha_j}{dt} \phi_j + \frac{d\beta_j}{dt} \psi_j \right) \phi_j dx + \beta_{j-1} - \int_{x_{j-1/2}}^{x_{j+1/2}} (\alpha_j \phi_j + \beta_j \psi_j) (\phi_j)_x dx &= 0, \\ \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\frac{d\alpha_j}{dt} \phi_j + \frac{d\beta_j}{dt} \psi_j \right) \psi_j dx + \beta_j - \int_{x_{j-1/2}}^{x_{j+1/2}} (\alpha_j \phi_j + \beta_j \psi_j) (\psi_j)_x dx &= 0,\end{aligned}\quad j = 1, 2, \dots, N.$$

The integrals involving the basis functions can now be computed. For convenience we assume a uniform node distribution with subinterval length h . When dropping the subscripts, we get

$$\begin{aligned}\int \phi^2 dx &= \frac{h}{3}, \quad \int \phi \psi dx = \frac{h}{6}, \quad \int \psi^2 dx = \frac{h}{3}, \\ \int \phi \phi_x dx &= -\frac{1}{2}, \quad \int \psi \phi_x dx = -\frac{1}{2}, \quad \int \phi \psi_x dx = \frac{1}{2}, \quad \int \psi \psi_x dx = \frac{1}{2},\end{aligned}$$

and the ODE system becomes

$$\begin{aligned}\frac{h}{3} \frac{d\alpha_j}{dt} + \frac{h}{6} \frac{d\beta_j}{dt} &= \beta_{j-1} - \frac{1}{2} \beta_j, \\ \frac{h}{6} \frac{d\alpha_j}{dt} + \frac{h}{3} \frac{d\beta_j}{dt} &= \frac{1}{2} \alpha_j - \frac{1}{2} \beta_j, \quad j = 1, 2, \dots, N.\end{aligned}\tag{12.11}$$

We have avoided the boundary conditions here. If boundary data are specified at the left, then β_0 is known. For the periodic case, we take $\beta_0 = \beta_N$.

At this point we notice a significant difference when comparing with Galerkin methods. There is no coupling of $d\alpha_j/dt$ and $d\beta_j/dt$ to the corresponding coefficients at neighboring intervals. In other words, the mass matrix has a simple block diagonal form, and each 2×2 system is easily solved before the computation. We get

$$\begin{aligned}\frac{d\alpha_j}{dt} &= \frac{1}{h} (4\beta_{j-1} - 3\alpha_j - \beta_j), \\ \frac{d\beta_j}{dt} &= \frac{1}{h} (-2\beta_{j-1} + 3\alpha_j - \beta_j), \quad j = 1, 2, \dots, N.\end{aligned}$$

This allows for a fast explicit difference method when solving the ODE-system in time, just as for finite difference methods.

In contrast to Galerkin methods, the stability does not follow automatically. The reason is that the weak form of the original problem cannot be transferred to the numerical approximation, since we have introduced a special numerical flux function. Therefore other techniques for proving stability must be introduced. Assuming periodic solutions, we can do a standard von Neumann analysis. The Fourier transformed system for α and β is

$$\frac{d}{dt} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \frac{1}{h} \begin{bmatrix} -3 & 4e^{-i\xi} - 1 \\ 3 & -2e^{-i\xi} - 1 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}.$$

The eigenvalues of the coefficient matrix \hat{Q} are shown in Figure 12.9 for discrete ξ -values and $h = 1$.

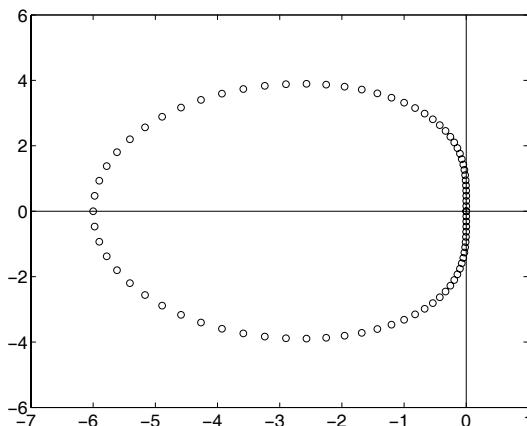


Fig. 12.9 Eigenvalues of $\hat{Q}(\xi)$, $0 \leq \xi \leq 2\pi$

All of them are located in the left half plane, with the single eigenvalue zero at the imaginary axis. Since \hat{Q} can be diagonalized by a bounded similarity transformation, the conclusion is that the approximation is stable, and furthermore, it is dissipative. This latter property is an effect of the numerical flux function. The choice we made introduces the value β_{j-1} on the left side of the interval I_j , which corresponds to an upwind scheme.

We solved the problem for $-1 \leq x \leq 1$, $0 \leq t \leq 6$ with the same initial data

$$|\sin \frac{\pi x}{2}|^r$$

as was used in Section 1.2 (the equation was $u_t = u_x$ there). The nodes are defined such that $x_j = -1 + jh$. At $t = 6$ the wave has made 3 laps, and is back to its original position.

Figure 12.10, 12.11, 12.12 show the solution

$$u^h = \sum_{j=1}^N (\alpha_j \phi_j + \beta_j \psi_j)$$

for $r = 1, 3, 5$ and $N = 20$.

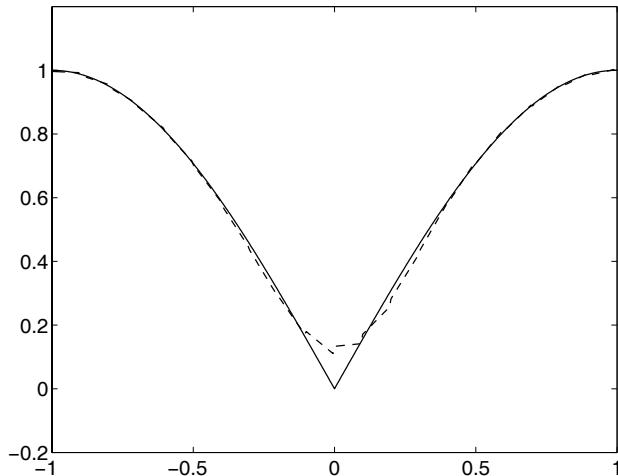


Fig. 12.10 $u(x, 6)$ (—), $u^h(x, 6)$ (---), $r = 1$, $N = 20$

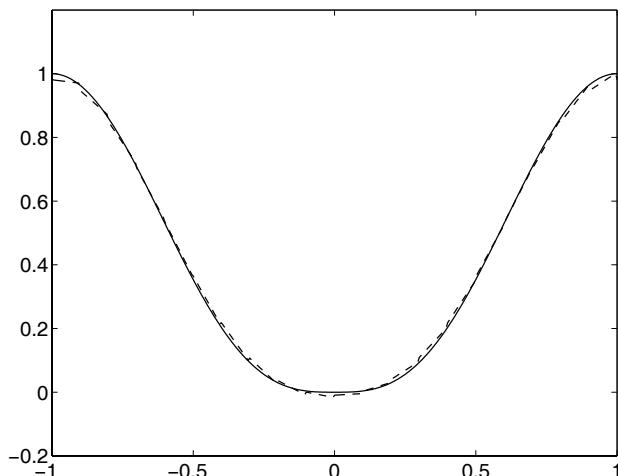


Fig. 12.11 $u(x, 6)$ (—), $u^h(x, 6)$ (---), $r = 3$, $N = 20$

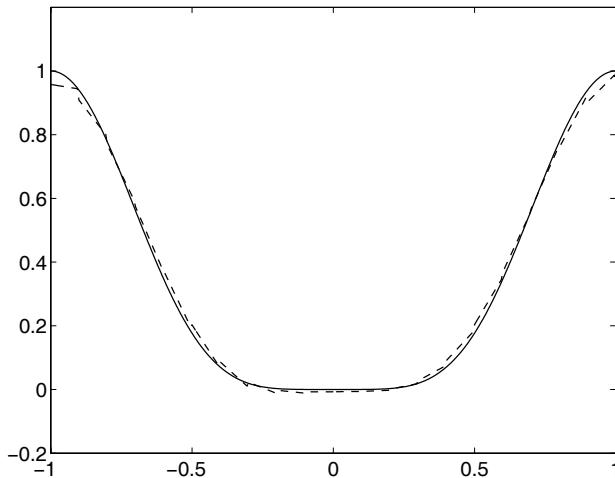


Fig. 12.12 $u(x, 6)$ (—), $u^h(x, 6)$ (---), $r = 5$, $N = 20$

In comparison with the results for the second order difference method displayed in Figure 1.2, 1.3, 1.4, it is quite good. The sharp cusp at $x = 0$ for $r = 1$ causes trouble, but it did so also for the difference methods. On the scale used in the pictures, it is hard to see the discontinuities at the nodes $x_{j+1/2}$. In Figure 12.13, showing the solution near $x = -1$ at another scale, they are clearly visible.

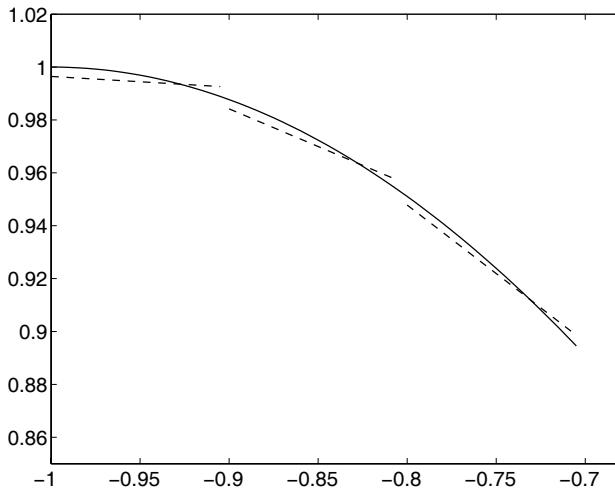


Fig. 12.13 $u(x, 6)$, $r = 1$ (—), $u^h(x, 6)$, $N = 20$

The basis functions ϕ_j and ψ_j were chosen in order to illustrate the block diagonal structure of the mass matrix. In our case, we could as well have chosen them such that a diagonal mass matrix is obtained without solving the 2×2 systems. The off-diagonal elements are obtained as $\int \phi \psi$, and therefore it is better to choose the orthogonal linear functions $\phi_j = 1$ and $\psi_j = 2(x - x_j)/h_j$.

Higher order methods are obtained by simply raising the degree of the polynomials in each subinterval. There is still no coupling between the time derivatives in different subintervals, but the local systems become larger. However, as noted above, by choosing orthogonal polynomials, even this minor problem is eliminated.

The generalization to several space dimensions is done in analogy with the Galerkin methods. The polynomials are constructed locally in each element such as triangles in two space dimensions. The basis functions associated with a certain triangle T_j are zero at all other triangles, and the solution u^h becomes discontinuous at the edges of each triangle.

The discontinuity of the numerical solutions is not of much concern. In fact, if the solution is represented by the values at the midpoints x_j of the intervals I_j , it can be seen as a discrete solution in analogy with difference methods. Another possibility is to add penalty terms of the type $\tau(\beta_j - \beta_{j-1})$, just as for SAT-methods described in Section 7.4.

DG methods is currently a very active area of research, and the methods are now being used for an increasing number of applications. The survey article [Cockburn et al., 2000] has an extensive reference list on this class of methods.

12.3 Spectral Methods

All methods that have been considered so far are based on some kind of discretization that is associated with a grid, or a set of nodes. Another avenue is to write the numerical solution as a finite expansion in terms of some known type of functions. This is a classical approach that has been used in applied mathematics for a long time, usually for some special cases with a small parameter involved. This general approach can be called spectral methods, since the basis functions are associated with the spectrum of some operator. However, for computational purposes, one must be able to compute the coefficients in an effective way, and then it helps to introduce some kind of grid also in this case. Furthermore, we limit the spectral expansion to an approximation of the derivatives occurring in the time dependent PDE, and then insert these approximations in a difference scheme for solution in time. This procedure is sometimes called a *pseudospectral method*, and it is a special case of a *collocation methods*. For this class of methods, we require that the differential equation is satisfied at the grid points only. This is in contrast to Galerkin methods, which require that the weak formulation is approximated in the whole domain. Indeed, we can consider collocation methods as a special case of Petrov–Galerkin methods. We want to approximate the solution of $u_t = P(\partial/\partial x)u$ and make the ansatz $u^h(x, t) = \sum_j \alpha_j(t) \phi_j(x)$. Let x_j be a set of grid points and let

$\delta_j(x)$ be the Dirac δ -function such that

$$\int v(x)\delta_j(x)dx = v(x_j), \quad j = 1, 2, \dots, N.$$

The Petrov–Galerkin method is

$$\int u_t^h(x, t)\delta_j(x)dx = \int (P(\partial/\partial x)u^h(x, t))\delta_j(x)dx,$$

which leads to

$$u_t^h(x_j, t) = (P(\partial/\partial x)u^h)(x_j, t), \quad j = 1, 2, \dots, N.$$

Note the difference compared to finite difference methods. In that case we are dealing solely with the values at the grid points, while for collocation methods we work with functions $u^h(x, t)$ in a certain subspace. These are defined everywhere in the computational domain, which makes it possible to apply the differential operator $P(\partial/\partial x)$ to them.

For problems with periodic solutions, the natural choice of basis functions is trigonometric polynomials. This leads to *Fourier methods*, which will be discussed in the first part of this section. For nonperiodic solutions, other types of basis functions, like the classic orthogonal polynomials, are used, and they will be discussed in the second part.

12.3.1 Fourier Methods

We shall begin by discussing trigonometric interpolation of 2π -periodic functions $v(x)$ that are known at grid points $x_j = jh$ as $v_j = v(x_j)$, $j = 0, 1, \dots, N$. Assuming that N is an even number with $(N+1)h = 2\pi$, we write the approximation as

$$\tilde{v}(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{v}_\omega e^{i\omega x}. \quad (12.12)$$

Under the interpolation condition

$$\tilde{v}(x_j) = v_j, \quad j = 0, 1, \dots, N,$$

the coefficients are given by

$$\hat{v}_\omega = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^N v_j e^{-i\omega x_j h}, \quad \omega = -N/2, -N/2 + 1, \dots, N/2. \quad (12.13)$$

This is the discrete Fourier transform, that was defined in Section 2.2.2 for grid functions u_j^n . The difference here is that we define the interpolating function $\tilde{v}(x)$ everywhere.

For difference methods we have seen that the accuracy depends on the smoothness of the solution, and in the error estimates it was assumed that the solution is “sufficiently smooth”. Here we shall be somewhat more precise. One can prove the following lemma:

Lemma 12.2. *Assume that $v(x)$ has q continuous derivatives, and let $\|\cdot\|_\infty$ denote the maximum norm. Then there are constants K_p such that*

$$\left\| \frac{d^p v}{dx^p}(\cdot) - \frac{d^p \tilde{v}}{dx^p}(\cdot) \right\|_\infty \leq \frac{K_p}{N^{q-p}}, \quad p < q.$$

□

This estimate shows that trigonometric interpolation is very accurate for smooth functions, both for the function itself and for its derivatives. Let us now see how this can be used for numerical solution of PDE.

Let $\tilde{w} = d\tilde{v}/dx$ be the derivative of the interpolating function. Then we have

$$\tilde{w}(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{w}_\omega e^{i\omega x} = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} i\omega \hat{v}_\omega e^{i\omega x}.$$

If $v(x)$ is known only at the grid points, we want to compute approximations of dv/dx at the grid points as well. The algorithm is

1. Compute \hat{v}_ω , $\omega = -N/2, -N/2+1, \dots, N/2$ by the discrete Fourier transform
2. Compute $\hat{w}_\omega = i\omega \hat{v}_\omega$, $\omega = -N/2, -N/2+1, \dots, N/2$
3. Compute w_j , $j = 0, 1, \dots, N$ by the inverse discrete Fourier transform

When the discrete Fourier transform is carried out by using the fast Fourier transform, the operation count of the numerical differentiation is $\mathcal{O}(N \log N)$. This fast computation is one of the reasons why numerical methods based on trigonometric interpolation became popular from the very beginning.

If \mathbf{v} and \mathbf{w} are the vectors containing the grid values v_j and w_j , we can also consider the three step algorithm as a linear transformation

$$\mathbf{w} = S\mathbf{v},$$

where S is an $(N+1) \times (N+1)$ matrix, which represents the approximation of $\partial/\partial x$. It can be shown that the elements $S_{\mu\nu}$ of the matrix S are real and have the simple form

$$S_{\mu\nu} = \begin{cases} \frac{(-1)^{\mu+\nu}}{2} \left(\sin \frac{x_\mu - x_\nu}{2} \right)^{-1}, & \mu \neq \nu \\ 0, & \mu = \nu, \end{cases}$$

$\mu, \nu = 0, 1, \dots, N.$

If N is not too large, one can use the simple matrix–vector multiplication $S\mathbf{v}$ for the actual computation. Obviously, the matrix is skewsymmetric, and it easily shown that the eigenvalues are $i\omega$, $\omega = -N/2, -N/2 + 1, \dots, N/2$, with corresponding eigenvectors $\mathbf{e}_\omega = [1, e^{i\omega h}, e^{i\omega 2h}, \dots, e^{i\omega Nh}]^T$.

We have assumed here that there is an odd number of given numbers v_j . The only reason for this assumption is the convenience of a symmetric expansion, such that the extreme values of ω are $\pm N/2$. If there is an even number of points x_j , $j = 0, 1, \dots, N-1$, $Nh = 2\pi$, the trigonometric polynomial is still (12.12), but now with the coefficients given by

$$\begin{aligned}\hat{v}_\omega &= \frac{1}{c_\omega \sqrt{2\pi}} \sum_{j=0}^{N-1} v_j e^{-i\omega x_j} h, \quad \omega = -N/2, -N/2 + 1, \dots, N/2, \\ c_\omega &= \begin{cases} 1, & |\omega| < N/2 \\ 2, & |\omega| = N/2. \end{cases}\end{aligned}\tag{12.14}$$

At a first glance it looks like we have a mistake here, since there are only N points given, and we have $N+1$ coefficients \hat{v}_ω . However, the full information is contained in the N coefficients \hat{v}_ω , $\omega = -N/2, -N/2 + 1, \dots, N/2 - 1$, since $\hat{v}_{N/2} = \hat{v}_{-N/2}$.

The elements $S_{\mu\nu}$ of the matrix S are also here real and are explicitly given by

$$S_{\mu\nu} = \begin{cases} \frac{(-1)^{\mu+\nu}}{2} \cot \frac{x_\mu - x_\nu}{2}, & \mu \neq \nu \\ 0, & \mu = \nu, \end{cases}$$

$$\mu, \nu = 0, 1, \dots, N-1.$$

In Chapter 5 we discussed the stability domains for various ODE solvers determined by the test equation. The Fourier method applied to $u_t = u_x$ is $d\mathbf{u}/dt = S\mathbf{u}$, and the Fourier transformed system is very simple:

$$\frac{d\hat{u}_\omega}{dt} = i\omega \hat{u}_\omega, \quad |\omega| \leq N/2.$$

If the stability domain for a certain ODE solver has a cutoff α_0 at the imaginary axis, the time step k has the stability limit

$$k \frac{N}{2} \leq \alpha_0.$$

Since $(N+1)h = 2\pi$ we get the stability limit

$$k \left(\frac{\pi}{h} - \frac{1}{2} \right) \leq \alpha_0,$$

and since h is arbitrarily small, we get the final condition

$$\frac{k}{h} \leq \frac{\alpha_0}{\pi}.$$

For problems with variable coefficients, the situation is more complicated than for difference methods. We showed in Chapter 2, that difference operators D and a smooth variable coefficient $a(x)$ have the commuting property $aD = Da + \mathcal{O}(h)$, which leads to an energy estimate if the constant coefficient approximation satisfies such an estimate. This commuting property is not shared by the Fourier differential operator, since it does not have the necessary local character. In fact, there are examples where variable coefficients trigger instabilities with Fourier methods. One way of stabilizing these, is to use extra dissipation operators, see for example [Kreiss and Oliger, 1979] and [Tadmor, 1986]. Another way is to formulate the PDE in “quasi-selfadjoint” form. If $a(x)$ is differentiable, the equation $u_t + a(x)u_x = 0$ can be written as

$$u_t + \frac{1}{2}au_x + \frac{1}{2}(au)_x - \frac{1}{2}a_xu = 0.$$

With the notation

$$A = \text{diag}(a_0, a_1, \dots, a_N), \quad B = \text{diag}((a_x)_0, (a_x)_1, \dots, (a_x)_N),$$

we use the approximation

$$\frac{d\mathbf{u}}{dt} + \frac{1}{2}AS\mathbf{u} + \frac{1}{2}SA\mathbf{u} - \frac{1}{2}B\mathbf{u} = 0.$$

Since S is skewsymmetric, summation by parts give the same type of estimate as the PDE satisfies.

For various difference and other methods we have used the simple equation $u_t + u_x = 0$ to demonstrate the numerical behavior. For the Fourier method there is no need to do that, which we will now explain.

The initial function can be interpolated such that it has the representation

$$f(x_j) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{f}_\omega e^{i\omega x_j}$$

at the grid points. Then we know that the true solution at the grid points is

$$u(x_j, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{u}_\omega(t) e^{i\omega x_j},$$

where

$$\hat{u}_\omega(t) = \hat{f}_\omega e^{-i\omega t}. \quad (12.15)$$

With the vectors

$$\mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}, \quad \mathbf{e}_\omega = \begin{bmatrix} 1 \\ e^{i\omega h} \\ e^{i\omega 2h} \\ \vdots \\ e^{i\omega Nh} \end{bmatrix},$$

(assuming $(N+1)h = 2\pi$) we write the numerical solution as

$$\mathbf{u}(t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{u}_\omega(t) \mathbf{e}_\omega,$$

$$\mathbf{u}(0) = \mathbf{f} = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{f}_\omega \mathbf{e}_\omega.$$

The equation $d\mathbf{u}/dt + S\mathbf{u} = 0$ implies

$$\sum_{\omega=-N/2}^{N/2} \frac{d\hat{u}_\omega(t)}{dt} \mathbf{e}_\omega + \sum_{\omega=-N/2}^{N/2} \hat{u}_\omega(t) S \mathbf{e}_\omega = 0,$$

and since \mathbf{e}_ω is an eigenvector of S with eigenvalue $i\omega$, we get

$$\frac{d\hat{u}_\omega(t)}{dt} + i\omega \hat{u}_\omega(t) = 0,$$

$$\hat{u}_\omega(0) = \hat{f}_\omega, \quad \omega = -N/2, -N/2 + 1, \dots, N/2.$$

This equation has the solution (12.15), which shows that the numerical solution is exact at the grid points. The conclusion is that if the Fourier method is applied to an equation with constant coefficients, then the errors at the grid points of the solution arise solely as an effect of the time discretization.

In order to get a nontrivial test problem, we use the nonlinear Burgers' equation $u_t + uu_x = 0$. As in Section 7.2 we write it in the split form

$$u_t + \frac{1}{3}((u^2)_x + uu_x) = 0,$$

which is the natural basis for obtaining an energy estimate. With the notation $U = \text{diag}(u_0, u_1, \dots, u_N)$ and $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=0}^N u_j v_j h$, the Fourier method is

$$\frac{d\mathbf{u}}{dt} + \frac{1}{3}(SU\mathbf{u} + US\mathbf{u}) = 0.$$

Since S is skewsymmetric, we have

$$\langle \mathbf{u}, SU\mathbf{u} \rangle = -\langle S\mathbf{u}, U\mathbf{u} \rangle = -\langle US\mathbf{u}, \mathbf{u} \rangle,$$

leading to norm conservation $d|\mathbf{u}|^2/dt = 0$.

In the numerical test, the Matlab ODE solver `ode45` was used, and the result is shown in Figure 12.14. The numerical solution is remarkably good with just 21 points in the whole interval.

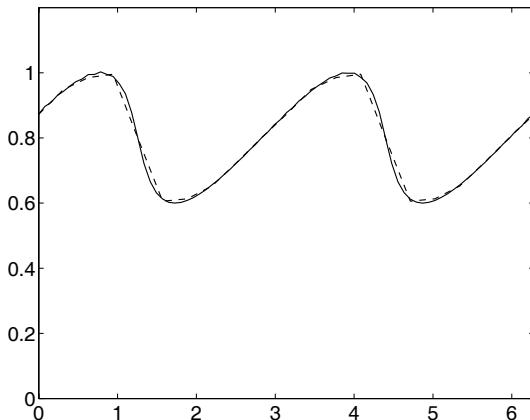


Fig. 12.14 The Fourier method for Burgers' equation. True solution (—), numerical solution $N = 20$ (---)

The Fourier method can of course be used for PDE of any order. When computing approximations of $\partial^q u / \partial x^q$, the multiplication of the Fourier coefficients \hat{v}_ω for the interpolating polynomial $v(x)$ by $i\omega$ in the algorithm above, is simply substituted by multiplication by $(i\omega)^q$.

12.3.2 Polynomial Methods

Spectral methods were originally introduced with trigonometric polynomials as the basis. For nonperiodic problems, these functions are not well suited. The reason is that Fourier modes by construction are periodic, and if the function to be interpolated is nonperiodic, it is interpreted as discontinuous at the boundaries. The result is that the well known Gibbs' phenomenon occurs as nonphysical oscillations.

The next step in the development of spectral methods, was therefore to introduce Chebyshev polynomials as the basis. The main reason for picking this particular class of classic polynomials is that it can be expressed as a cosine transform if the interpolating points are chosen right, and the fast Fourier transform can therefore be used as the central part of the algorithm. Later other types of orthogonal polynomials were introduced.

Chebyshev polynomials $T_n(x)$ are defined in the interval $[-1, 1]$ as the solutions to the *Sturm–Liouville problem*

$$\begin{aligned} \frac{d}{dx} \left(\sqrt{1-x^2} \frac{dT_n(x)}{dx} \right) + \frac{n^2}{\sqrt{1-x^2}} T_n(x) &= 0, \\ T_n(-1) &= (-1)^n, \quad T_n(1) = 1. \end{aligned} \tag{12.16}$$

They are orthogonal in the sense that

$$\int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_m(x) T_n(x) dx = 0 \quad \text{if } m \neq n.$$

Furthermore, they satisfy the two step recurrence relation

$$T_{n+1} = 2xT_n(x) - T_{n-1}(x), \quad T_0(x) = 1, \quad T_1(x) = x,$$

which is very convenient for computation. Another property that makes them convenient for computation is obtained by the transformation $x = \cos \xi$, $\tilde{T}_n(\xi) = T_n(\cos \xi)$, which reduces (12.16) to the simple form

$$\frac{d^2 \tilde{T}_n(\xi)}{d\xi^2} + n^2 \tilde{T}_n(\xi) = 0$$

with the solution

$$\tilde{T}_n(\xi) = \cos(n\xi), \quad 0 \leq \xi \leq \pi,$$

or equivalently,

$$T_n(x) = \cos(n \arccos x), \quad -1 \leq x \leq 1.$$

Obviously we have $|T_n(x)| \leq 1$ for all n .

The interpolation problem is stated as for trigonometric polynomials, but the interpolation points are chosen differently. With the interpolating function expressed as

$$\tilde{v}(x) = \sum_{n=0}^N \hat{v}_n T_n(x),$$

we require

$$\tilde{v}(x_j) = v_j, \quad x_j = -\cos\left(\frac{j\pi}{N}\right), \quad j = 0, 1, \dots, N.$$

The points x_j are called the *Gauss–Lobatto* points. They were originally derived for highly accurate quadrature formulas, and they arise here in the process of approximating the continuous Chebyshev transform. With $h = 2/N$, the Chebyshev coefficients are then obtained as

$$\begin{aligned} \hat{v}_n &= \frac{1}{c_n} \sum_{j=0}^N \frac{1}{c_j} v_j T_n(x_j) h, \quad n = 0, 1, \dots, N, \\ c_n &= \begin{cases} 1, & n = 1, 2, \dots, N-1 \\ 2, & n = 0, N. \end{cases} \end{aligned}$$

We can also use the equivalent expansion

$$\begin{aligned}\tilde{v}(x) &= \sum_{n=0}^N \hat{v}_n \cos \frac{j n \pi}{N}, \\ \hat{v}_n &= \frac{1}{c_n} \sum_{j=0}^N \frac{1}{c_j} v_j \cos \frac{j n \pi}{N} h, \quad n = 0, 1, \dots, N.\end{aligned}$$

This is the cosine transformation, which can be computed fast by a special version of the FFT.

In the pseudospectral (or collocation) method, we need to compute the derivative of the interpolating polynomial. For this we can use the two recurrence relations:

$$\begin{aligned}T'_{n+1}(x) &= 2(n+1)T_n(x) + \frac{n+1}{n-1} T'_{n-1}(x), \\ (1-x^2)T'_n(x) &= \frac{n}{2} (T_{n-1}(x) - T_{n+1}(x)).\end{aligned}$$

In analogy with Fourier methods, one can also here derive an explicit representation of the $(N+1) \times (N+1)$ differentiation matrix S_T in the relation $\mathbf{w} = S_T \mathbf{v}$, where \mathbf{w} contains the grid values of the differentiated interpolating function. The elements are

$$(S_T)_{\mu\nu} = \begin{cases} -(2N^2 + 1)/6, & \mu = \nu = 0, \\ -x_\mu / (2(1 - x_\mu^2)), & \mu = \nu = 1, 2, \dots, N-1, \\ (2N^2 + 1)/6, & \mu = \nu = N, \\ c_\mu (-1)^{\mu+\nu} / (c_\nu (x_\mu - x_\nu)), & \mu \neq \nu. \end{cases}$$

It is possible to show that all the eigenvalues of this matrix are located in the left half plane, and the semidiscrete approximation is stable.

The computation is done in physical x -space, and the grid points are determined by the mapping $\cos(\xi)$, where the ξ -points are uniformly distributed. This means that we get a compression of points near the boundaries $x = \pm 1$. Figure 12.15 shows an example with $N = 40$.

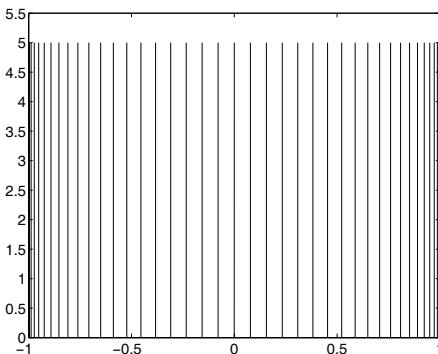


Fig. 12.15 Chebyshev grid for $N = 40$

If there are no special features of the solution near the boundaries, this point distribution is not optimal, and this may be a disadvantage of the method.

Let us now take a closer look at the behavior of the Chebyshev polynomials near the boundaries. Differentiation gives

$$\frac{dT_n(x)}{dx} = \frac{n}{\sqrt{1-x^2}} \sin(n \arccos x),$$

and at a first glance it looks like we are running into trouble near the boundaries. However, with $x = 1 - \varepsilon$ and ε small, we have $\arccos x \approx \sqrt{2\varepsilon}$, and for any fixed n

$$\frac{dT_n(x)}{dx} \approx \frac{n}{\sqrt{2\varepsilon}} \sin(n\sqrt{2\varepsilon}) \approx n^2.$$

This shows that the derivatives are well defined also at the boundaries. On the other hand, the derivative becomes large when n increases, i.e., when the grid is refined. The norm of the differentiation matrix is $\mathcal{O}(N^2)$, and for a first order PDE of the type $u_t + a(x)u_x = 0$, this leads to a severe restriction on the time step of the type $k = \mathcal{O}(1/N^2)$ for an explicit method. This is natural when considering the structure of the grid. If $h = 2/N$ is a typical step size for a uniform grid in space, the Chebychev grid has a step size of order h^2 near the boundaries.

One remedy for this complication is to divide the domain into smaller subdomains, and then couple them in some way across the inner boundaries. In this way the degree of the polynomials can be kept low, and the grid points become more evenly distributed. This technique is an example of what has become known as a *spectral element method*. There are also other techniques to relax the time step restriction using certain transformations, but these are obtained at the price of extra computations.

One reason for introducing Chebyshev polynomials is to make it possible to handle nonperiodic boundaries, but we have not yet discussed how the boundary conditions should be implemented. Consider the example

$$\begin{aligned} u_t &= u_x, \quad -1 \leq x \leq 1, \quad t \geq 0, \\ u(1,t) &= g(t), \\ u(x,0) &= f(x), \end{aligned}$$

and assume that we are using an explicit ODE solver. One way of implementing the boundary condition is to advance the solution u_j^n one step at all points x_j , $j = 0, 1, \dots, N$ including the boundaries, and then enforce the condition $u_N^{n+1} = g(t_n)$. However, a more convenient, and recently more popular method, is to use a penalty term based on the SAT-technique that was described in Section 7.4. In analogy with difference methods we let $\mathbf{u} = [u_0 \ u_1 \ \dots \ u_N]^T$ be the solution vector containing all grid values, and $\mathbf{w} = [0 \ 0 \ \dots \ 1]^T$. The semidiscrete SAT method is

$$\begin{aligned}\frac{d\mathbf{u}}{dt} &= S_T \mathbf{u} - \tau(u_N - g(t)) \mathbf{w}, \\ \mathbf{u} &= \mathbf{f}.\end{aligned}$$

It can be shown that it is stable if $\tau = \alpha N^2$, where α is a constant that is big enough.

This type of penalty methods is also used to enforce (almost) continuity across internal boundaries that arise with spectral elements as mentioned above.

Several different kinds of classical orthogonal polynomials have been revived as a consequence of the interest in spectral methods. We shall give some brief comments on still another one of them.

Legendre polynomials $P_n(x)$ are defined in the interval $[-1, 1]$ as the solutions to the *Sturm–Liouville problem*

$$\begin{aligned}\frac{d}{dx} \left((1-x^2) \frac{dP_n(x)}{dx} \right) + n(n+1)P_n(x) &= 0, \\ P_n(-1) &= (-1)^n, \quad P_n(1) = 1.\end{aligned}$$

Orthogonality:

$$\int_{-1}^1 P_m(x) P_n(x) dx = 0 \quad \text{if } m \neq n.$$

Recurrence relations:

$$\begin{aligned}P_{n+1} &= \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x), \quad P_0(x) = 1, \quad P_1(x) = x, \\ P'_{n+1} &= (2n+1) P_n(x) - P'_{n-1}(x).\end{aligned}$$

The interpolation points x_j corresponding to the Gauss–Lobatto points for Legendre polynomials are the roots of

$$(1-x^2) \frac{dP_N(x)}{dx} = 0.$$

Discrete Legendre transform ($h = 2/N$):

$$\begin{aligned}\tilde{v}(x) &= \sum_{n=0}^N \hat{v}_n P_n(x), \\ \hat{v}_n &= \frac{1}{c_n} \sum_{j=0}^N v_j P_n(x_j) h, \quad n = 0, 1, \dots, N, \\ c_n &= \begin{cases} \frac{2}{2n+1}, & n = 0, 1, \dots, N-1 \\ \frac{2}{N}, & n = N. \end{cases}\end{aligned}$$

Differentiation matrix:

$$(S_L)_{\mu\nu} = \begin{cases} -N(N+1)/4, & \mu = \nu = 0, \\ 0, & \mu = \nu = 1, 2, \dots, N-1, \\ N(N+1)/4, & \mu = \nu = N, \\ P_N(x_\mu)/(P_N(x_\nu)(x_\mu - x_\nu)), & \mu \neq \nu. \end{cases}$$

The procedure for implementation of the boundary conditions is in principle the same as for Chebyshev polynomials, and the penalty term approach is often used also for Legendre polynomials.

The formulas presented here are based on interpolation at the Gauss–Lobatto points. There are other points that can be used as the basis for interpolation, and they give different formulas.

For a more comprehensive presentation of spectral methods, we recommend the recent book [Hesthaven et al., 2007] by Hesthaven et.al.

12.4 Finite Volume Methods

In Chapter 11 conservation laws were discussed, and the concept of cell averages was introduced. These are closely connected to finite volume methods (FVM), and in this section we shall describe these methods. The practical use is for problems in several space dimensions, but we start by 1-D problems, which have the form

$$u_t + f_x(u) = 0,$$

where the flux function $f(u)$ may be a nonlinear function of u . The domain is divided into subintervals, or *finite volumes*, $I_j = [x_{j-1/2}, x_{j+1/2}]$ with length h_j . Note that we don't assume a uniform distribution of the endpoints $x_{j+1/2}$, i.e., the length of the subintervals I_j may vary from one interval to the next. In analogy with cell averages, the value $U_j(t)$ is considered as an approximation of the average of the solution over I_j :

$$U_j(t) \approx U(j, t) = \frac{1}{h_j} \int_{I_j} u(x, t) dx.$$

Integration over one interval yields

$$\frac{dU(j, t)}{dt} + f(u(x_{j+1/2}, t)) - f(u(x_{j-1/2}, t)) = 0.$$

We are now looking for an approximation of $u(x, t)$ at the half-points expressed in terms of $U(j, t)$. The problem of finding point values in terms of cell averages was discussed in Section 11.6. Here we limit ourselves to second order methods, and then the connection is easy. Let $x_j = (x_{j-1/2} + x_{j+1/2})/2$ be the midpoint of the interval I_j . Since

$$\frac{1}{h_j} \int_{I_j} u(x, t) dx = u(x_j, t) + \mathcal{O}(h_j^2)$$

for smooth solutions, we can as well work with point values instead of cell averages. We switch to the usual notation $u_j(t)$ for point values, and use the approximations

$$f(u(x_{j+1/2}, t)) \approx \frac{f(u_j(t)) + f(u_{j+1}(t))}{2},$$

$$f(u(x_{j-1/2}, t)) \approx \frac{f(u_{j-1}(t)) + f(u_j(t))}{2}.$$

This results in the *centered finite volume method*

$$\frac{du_j}{dt} + \frac{f(u_{j+1}) - f(u_{j-1})}{2h_j} = 0. \quad (12.17)$$

When disregarding the boundary terms, a summation over the whole computational domain gives

$$\frac{d}{dt} \sum_j u_j h_j = 0,$$

i.e., the conservation property is preserved in the discrete sense. As we saw in Chapter 11, this is an important property when dealing with shocks, and this is one reason for the popularity of finite volume methods. On the other hand, we also saw in Chapter 11 that these centered methods introduce oscillations around a shock, and they require some kind of dissipation mechanism for this type of problems. This is done in analogy with difference methods.

Let us next take a look at the accuracy of (12.17). Sometimes it is possible to generate the points $x_{j+1/2}$ by a smooth transformation $x = x(\xi)$, where the grid in ξ -space is uniform with step size h such that $\xi_{j+1/2} - \xi_{j-1/2} = h$. However, in multi-dimensional applications, the geometry is sometimes such that one has to accept more abrupt changes of the grid size from one finite volume to the next. In such a case, the form (12.17) may not even be consistent. For example, consider the equation $u_t + u_x = 0$, and a grid where the intervals I_{j-1}, I_j, I_{j+1} have lengths $h, h, 2h$. The approximation becomes

$$\frac{du_j}{dt} + \frac{u_{j+1} - u_{j-1}}{2h} = 0,$$

which is no good, since $x_{j+1} - x_{j-1} = 2.5h$. Therefore, it is better to store the points x_j and change the approximation to

$$\frac{du_j}{dt} + \frac{f(u_{j+1}) - f(u_{j-1})}{x_{j+1} - x_{j-1}} = 0. \quad (12.18)$$

We still have conservation, but now with different weights in the sum:

$$\frac{d}{dt} \sum_j u_j \tilde{h}_j = 0, \quad \tilde{h}_j = (x_{j+1} - x_{j-1})/2.$$

Assume next that there is a uniform grid ξ_j in ξ -space with step size h , and that the grid in x -space is generated by a smooth transformation $x = x(\xi)$ such that $x_j = x(\xi_j)$. For convenience, we assume that $f(u) = u$, such that the approximation is

$$\frac{du_j}{dt} + \frac{u_{j+1} - u_{j-1}}{x_{j+1} - x_{j-1}} = 0.$$

We have

$$\begin{aligned} u(x_{j+1}) &= u(x_j) + (x_{j+1} - x_j)u_x(x_j) + \frac{(x_{j+1} - x_j)^2}{2}u_{xx}(x_j) + \mathcal{O}(h^3), \\ u(x_{j-1}) &= u(x_j) - (x_j - x_{j-1})u_x(x_j) + \frac{(x_j - x_{j-1})^2}{2}u_{xx}(x_j) + \mathcal{O}(h^3), \end{aligned}$$

where we have used that $|x_{j+1} - x_j| = \mathcal{O}(h)$ for all j . We get

$$\frac{u(x_{j+1}) - u(x_{j-1})}{x_{j+1} - x_{j-1}} = u_x(x_j) + \frac{(x_{j+1} - x_j)^2 - (x_j - x_{j-1})^2}{2(x_{j+1} - x_{j-1})}u_{xx}(x_j) + \mathcal{O}(h^2).$$

By assumption we have

$$\begin{aligned} x_{j+1} &= x_j + h \frac{dx}{d\xi}(\xi_j) + \frac{h^2}{2} \frac{d^2x}{d\xi^2}(\xi_j) + \mathcal{O}(h^3), \\ x_{j-1} &= x_j - h \frac{dx}{d\xi}(\xi_j) + \frac{h^2}{2} \frac{d^2x}{d\xi^2}(\xi_j) + \mathcal{O}(h^3), \end{aligned}$$

i.e.,

$$\begin{aligned} (x_{j+1} - x_j)^2 &= h^2 \left(\frac{dx}{d\xi}(\xi_j) \right)^2 + \mathcal{O}(h^3), \\ (x_j - x_{j-1})^2 &= h^2 \left(\frac{dx}{d\xi}(\xi_j) \right)^2 + \mathcal{O}(h^3). \end{aligned}$$

This shows that

$$\frac{u(x_{j+1}) - u(x_{j-1})}{x_{j+1} - x_{j-1}} = u_x(x_j) + \mathcal{O}(h^2),$$

i.e., (12.18) is a second order approximation.

This conclusion holds for difference approximations as well. The transformation from ξ -space introduces a slight shift from the properly centered scheme in physical space, but the shift is not large enough to destroy the order of accuracy.

We now turn to the two-dimensional case and the conservation law

$$u_t + f_x(u) + g_y(u) = 0.$$

The finite volumes I_j now become quadrilaterals V_j . In Figure 12.16 the central finite volume is denoted by V_0 with boundary ∂V_0 and corner points A, B, C, D . The center of gravity points are denoted by $0, 1, 2, 3, 4$ for the five quadrilaterals that become coupled to each other.

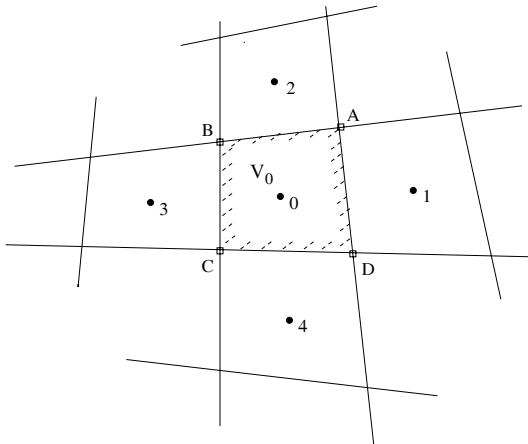


Fig. 12.16 Finite volume grid in 2-D

Integration by parts is generalized to the Green's formula

$$\int \int_{V_j} u_t dx dy + \int_{\partial V_j} f(u) dy - \int_{\partial V_j} g(u) dx = 0.$$

The line integrals are approximated by the average of the two u -values on each side, multiplied by the corresponding distance. For example, for the integral along the eastern side we use

$$\int_D^A f(u) dy \approx (y_A - y_D) \frac{f(u_0) + f(u_1)}{2},$$

where $u_0 = u(x_0, y_0)$, and y_A denotes the y -coordinate at the point A etc. With the approximation

$$\int \int_{V_0} u_t dx dy \approx \Delta V_0 \frac{du_0}{dt},$$

where ΔV_0 is the area of V_0 , and with the notation $f_0 = f(u_0)$ etc, the method becomes

$$\begin{aligned}
2\Delta V_0 \frac{du_0}{dt} + (y_A - y_D)(f_0 + f_1) + (y_B - y_A)(f_0 + f_2) \\
+ (y_C - y_B)(f_0 + f_3) + (y_D - y_C)(f_0 + f_4) \\
- (x_A - x_D)(f_0 + f_1) - (x_B - x_A)(f_0 + f_2) \\
- (x_C - x_B)(f_0 + f_3) - (x_D - x_C)(f_0 + f_4) = 0.
\end{aligned} \tag{12.19}$$

When summing up over the whole computational domain, all terms cancel each other, and we get conservation in the discrete sense

$$\frac{d}{dt} \sum_j u_j \Delta V_j = 0,$$

where the boundary terms have been disregarded.

In the special case of a uniform rectangular grid, we have

$$y_B - y_A = y_D - y_C = x_A - x_D = x_C - x_B = 0.$$

With the finite difference notation $u(x_0, y_0) \rightarrow u_{ij}$, $\Delta x = x_A - x_B$, $\Delta y = y_A - y_D$, the finite volume method reduces to the standard centered second order difference scheme

$$\frac{du_{ij}}{dt} + \frac{f_{i+1,j} - f_{i-1,j}}{2\Delta x} + \frac{g_{i,j+1} - g_{i,j-1}}{2\Delta y} = 0.$$

The method (12.19) is the 2-D analogue of (12.17). We showed that if the grid is not generated by a smooth transformation, then the scheme may no longer be consistent. This is of course true also in the 2-D case. The scheme uses only the coordinates of the corner points of the finite volume (corresponding to $x_{j-1/2}$ and $x_{j+1/2}$ in the 1-D case), and the coordinates of the center points have no influence. Accordingly, for more general grids, the boundaries of the finite volumes should be modified. One way of doing that is to involve the center points of the remaining surrounding four quadrilaterals and furthermore introduce a new extended quadrilateral \tilde{V}_0 as shown in Figure 12.17 with dashed boundaries. The extra center points are labeled 1.5, 2.5, 3.5, 4.5, and the area of the modified dotted quadrilateral is denoted by $\Delta \tilde{V}_0$ (which is four times larger than ΔV_0 in the uniform case). In the formula (12.19), we now make the following changes:

$$\begin{aligned}
x_A - x_D &\rightarrow (x_{1.5} + x_2 - x_4 - x_{4.5})/4 & y_A - y_D &\rightarrow (y_{1.5} + y_2 - y_4 - y_{4.5})/4, \\
x_B - x_A &\rightarrow (x_{2.5} + x_3 - x_1 - x_{1.5})/4 & y_B - y_A &\rightarrow (y_{2.5} + y_3 - y_1 - y_{1.5})/4, \\
x_C - x_B &\rightarrow (x_{3.5} + x_4 - x_2 - x_{2.5})/4 & y_C - y_B &\rightarrow (y_{3.5} + y_4 - y_2 - y_{2.5})/4, \\
x_D - x_C &\rightarrow (x_1 + x_{4.5} - x_3 - x_{3.5})/4 & y_D - y_C &\rightarrow (y_1 + y_{4.5} - y_3 - y_{3.5})/4, \\
\Delta V_0 &\rightarrow \Delta \tilde{V}_0/4.
\end{aligned}$$

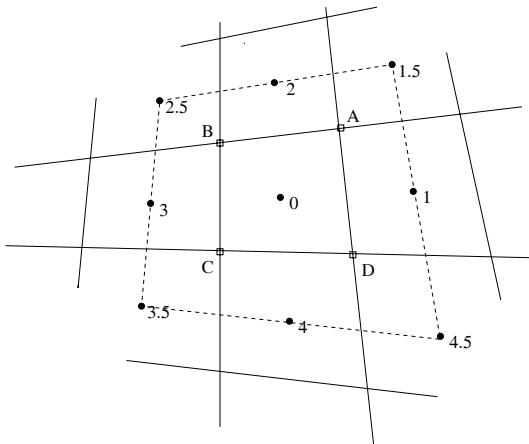


Fig. 12.17 Finite volume grid. The area of the dashed quadrilateral is $\Delta\tilde{V}_0$

We have limited ourselves to quadrilaterals here. However, the methods can be worked out also for triangular grids, where the line integrals along the sides are approximated by using the function values at neighboring triangles.

The generalization of finite volume methods to higher order accuracy is not trivial. For 1-D problems we can use some of the tools described in Section 11.6 generalized to nonuniform grids. However, in 2-D, the problem goes from accurate approximation of point values to accurate approximation of line integrals. Therefore, finite volume methods are usually thought of as belonging to the class of robust, but low order methods. When it comes to unstructured grids and higher order accuracy, finite element methods are more convenient.

Appendix A

Solution of Difference Equations

In this appendix we derive the general form of difference equations with constant coefficients in one space dimension. We begin by the scalar case, and consider the difference equation

$$u_{j+q} + \alpha_{q-1}u_{j+q-1} + \dots + \alpha_0u_j = 0, \quad j = 0, 1, \dots,$$

where α_ν are constants. With $\mathbf{u}_j = [u_{j+q-1} \ u_{j+q-2} \ \dots \ u_j]^T$ the difference equation can be written as

$$\mathbf{u}_{j+1} = Q\mathbf{u}_j, \quad (\text{A.1})$$

where

$$Q = \begin{bmatrix} -\alpha_{q-1} & -\alpha_{q-2} & \dots & \dots & -\alpha_0 \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

The eigenvalues κ and eigenvectors $\boldsymbol{\phi} = [\phi_1 \ \phi_2 \ \dots \ \phi_q]^T$ satisfy

$$Q = \begin{bmatrix} -(\alpha_{q-1} + \kappa) & -\alpha_{q-2} & \dots & \dots & -\alpha_0 \\ 1 & -\kappa & \dots & \dots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & 1 & -\kappa \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \vdots \\ \phi_q \end{bmatrix} = 0.$$

With ϕ_q as an arbitrary constant, we get

$$\begin{aligned}
\phi_{q-1} &= \kappa \phi_q, \\
\phi_{q-2} &= \kappa \phi_{q-1} = \kappa^2 \phi_q, \\
&\vdots \\
\phi_1 &= \kappa^{q-1} \phi_q, \\
(-\kappa^q - \alpha_{q-1} \kappa^{q-1} - \alpha_{q-2} \kappa^{q-2} - \dots - \alpha_0) \phi_q &= 0.
\end{aligned}$$

The last equation leads to the *characteristic equation*

$$\kappa^q + \alpha_{q-1} \kappa^{q-1} + \dots + \alpha_0 = 0. \quad (\text{A.2})$$

We assume first that all roots κ are distinct. Then there is a matrix T such that

$$T^{-1} Q T = \text{diag}(\kappa_1 \kappa_2 \dots \kappa_q) = \Lambda,$$

and with $\mathbf{v}_j = T^{-1} \mathbf{u}_j$, we get

$$\mathbf{v}_{j+1} = T^{-1} Q T T^{-1} \mathbf{u}_j = \Lambda \mathbf{v}_j,$$

i.e.,

$$\mathbf{v}_j = \Lambda^j \mathbf{v}_0,$$

and

$$\mathbf{u}_j = T \Lambda^j \mathbf{v}_0 = T \begin{bmatrix} v_{q-1} \kappa_1^j \\ v_{q-2} \kappa_2^j \\ \vdots \\ v_0 \kappa_q^j \end{bmatrix},$$

where v_ν are constants, yet to be determined. The last element of the vector \mathbf{u}_j has the form

$$u_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j + \dots + \sigma_q \kappa_q^j, \quad (\text{A.3})$$

where the constants σ_ν are to be determined by the initial or boundary conditions. This is the general form of the solution, when the roots κ_ν of the characteristic equation (A.2) are distinct.

Assume next that there is a double root κ_1 . Then there is a matrix T that takes Q into the *Jordan canonical form*

$$T^{-1} Q T = \begin{bmatrix} \kappa_1 & 1 & 0 & \dots & 0 \\ 0 & \kappa_1 & 0 & \dots & 0 \\ \vdots & & \kappa_3 & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \kappa_q \end{bmatrix}.$$

With $\mathbf{v}_j = T^{-1}\mathbf{u}_j$ we denote by $\mathbf{v}_j^I = [v_j^{(1)} \ v_j^{(2)}]^T$ the first two elements of \mathbf{v}_j , and the Jordan box is

$$R = \begin{bmatrix} \kappa_1 & 1 \\ 0 & \kappa_1 \end{bmatrix}.$$

We get

$$\mathbf{v}_j^I = R^j \mathbf{v}_0^I = R^j \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \quad R^j = \begin{bmatrix} \kappa_1^j & j\kappa_1^{j-1} \\ 0 & \kappa_1^j \end{bmatrix}.$$

This gives

$$\mathbf{v}_j^I = \begin{bmatrix} c_1 \kappa_1^j + \frac{c_2}{\kappa_1} j \kappa_1^j \\ c_2 \kappa_1^j \end{bmatrix} = \begin{bmatrix} (c_1 + \tilde{c}_2 j) \kappa_1^j \\ c_2 \kappa_1^j \end{bmatrix},$$

and we get the final form of the solution

$$u_j = (\sigma_1 + \sigma_2 j) \kappa_1^j + \sigma_3 \kappa_3^j + \dots + \sigma_q \kappa_q^j.$$

It is now easy to see the form of the solution in the general case. If the root κ_ν has multiplicity m_ν , the corresponding Jordan box is an upper bidiagonal $m_\nu \times m_\nu$ matrix. The solution has the form

$$u_j = \sum_{\{\kappa_\nu \text{ distinct}\}} (\sigma_1^{(\nu)} + j\sigma_2^{(\nu)} + \dots + j^{m_\nu-1} \sigma_{m_\nu}^{(\nu)}) \kappa_\nu^j, \quad (\text{A.4})$$

where $\sigma_\mu^{(\nu)}$ are constants to be determined by initial or boundary conditions.

For a given case there is no need to do the transformations that were carried out for the derivation above. It is enough to solve the characteristic equation (A.2), and then write down the form (A.3) or (A.3) of the solution.

Let us next discuss the generalizations to systems of difference equations

$$u_{j+q} + A_{q-1} u_{j+q-1} + \dots + A_0 u_j = 0, \quad j = 0, 1, \dots,$$

where u_j are vectors with m elements and A_j are $m \times m$ matrices. The big ($mq \times mq$) matrix Q in (A.1) now takes the form

$$Q = \begin{bmatrix} -A_{q-1} & -A_{q-2} & \dots & \dots & -A_0 \\ I & 0 & \dots & \dots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & I & 0 \end{bmatrix}.$$

The components ϕ_ν in the eigenvectors ϕ are now vectors, and the eigenvalues are given in analogy with the scalar case by

$$(-I\kappa^p - A_{q-1}\kappa^{q-1} - \dots - A_0)\phi_q = 0, \quad (\text{A.5})$$

where ϕ_q is a vector with m components. The coefficient matrix must be singular, and we get the characteristic equation

$$\text{Det}(I\kappa^p + A_{q-1}\kappa^{q-1} + \dots + A_0) = 0. \quad (\text{A.6})$$

This equation together with (A.5) gives the final general form of the solution. There are mq roots of the characteristic equation (A.6), but there can be only m linearly independent “eigenvectors” $\phi_{q\mu} = \psi_\mu$. If there is a full set of such vectors, and the corresponding roots $\kappa_{\mu\nu}$ are all distinct, then the solution has the general form

$$u_j = \sum_{\mu=1}^m \sum_{\nu=1}^q \sigma_{\mu\nu} \kappa_{\mu\nu}^j \psi_\mu,$$

where $\sigma_{\mu\nu}$ are scalar constants to be determined by the initial or boundary conditions.

In the general case the solution may have a more complicated form. However, it is not enough to identify the multiple roots to obtain the general form of the solution. For example, if two identical scalar difference equations are artificially combined to a system, there will obviously be a double root κ . But this fact should not change the two original scalar forms of the solution. There are two linearly independent vectors ψ_ν belonging to the double root κ .

For a given problem, the system (A.5),(A.6) is solved for κ and ϕ_q , and then we trace the transformation procedure back to obtain the solution. The most general form of the solution is

$$u_j = \sum_{\nu} P_\nu(j) \kappa_\nu^j,$$

where $P_\nu(j)$ are polynomials in j with vector coefficients that in all contain mq undetermined scalar coefficients σ_ν .

It should be said that in general the normal mode analysis based on the Laplace transformed difference equations is usually much simpler than might be indicated here, even for systems of PDE. The reason is that in most cases there are only a few special points \tilde{s} or z in the complex plane that must be investigated, and the equations (A.5),(A.6) are easy to solve.

Appendix B

The Form of SBP Operators

In this appendix we list some of the SBP operators and the corresponding matrix H in the norm. We use the notation $D_p^{(1)}$ for approximations of $\partial/\partial x$ with order of accuracy $p = 2s$ at inner points. For approximations of the second derivative $\partial^2/\partial x^2$ we use the notation $D_p^{(2)} = H^{-1}(-M + S)$.

We use a slightly different notation compared to Section 7.2. The matrix elements are numbered in the traditional way, i.e., the numbering of the rows and columns starts at one instead of zero.

B.1 Diagonal H -norm

We begin by the operators based on diagonal matrices H_0 in the norm. The local order of accuracy near the boundary is $p/2$ for $D_p^{(1)}$ and $D_p^{(2)}$. The first row in S has local order of accuracy $p/2 + 1$.

p = 2

$$H = \begin{bmatrix} \frac{1}{2} & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{bmatrix}, \quad D_2^{(1)} = \frac{1}{h} \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & -\frac{1}{2} & 0 & \frac{1}{2} & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

$$D_2^{(2)} = \frac{1}{h^2} \begin{bmatrix} 1 & -2 & 1 & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \quad S = \frac{1}{h^2} \begin{bmatrix} \frac{3}{2} & -2 & \frac{1}{2} & & & \\ 0 & 0 & & & & \\ & 0 & & & & \\ & & & & & \\ & & & & & \ddots \end{bmatrix}.$$

p = 4

$$H = \begin{bmatrix} \frac{17}{48} & & & & & & & \\ & \frac{59}{48} & & & & & & \\ & & \frac{43}{48} & & & & & \\ & & & \frac{49}{48} & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & \ddots \end{bmatrix},$$

$$D_4^{(1)} = \frac{1}{h} \begin{bmatrix} -\frac{24}{17} & \frac{59}{34} & -\frac{4}{17} & -\frac{3}{34} & & & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & & & & & \\ \frac{4}{43} & -\frac{59}{86} & 0 & \frac{59}{86} & -\frac{4}{43} & & & & \\ \frac{3}{98} & 0 & -\frac{59}{98} & 0 & \frac{32}{49} & -\frac{4}{49} & & & \\ & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$D_4^{(2)} = \frac{1}{h^2} \begin{bmatrix} 2 & -5 & 4 & -1 & & & & & \\ 1 & -2 & 1 & 0 & & & & & \\ -\frac{4}{43} & \frac{59}{43} & -\frac{110}{43} & \frac{59}{43} & -\frac{4}{43} & & & & \\ -\frac{1}{49} & 0 & \frac{59}{49} & -\frac{1}{49} & \frac{64}{49} & -\frac{4}{49} & & & \\ & & & -\frac{1}{12} & \frac{4}{3} & -\frac{5}{2} & \frac{4}{3} & \frac{1}{2} & \frac{1}{12} \\ & & & & -\frac{1}{12} & \frac{4}{3} & -\frac{5}{2} & \frac{4}{3} & \frac{1}{2} \\ & & & & & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$S = \frac{1}{h^2} \begin{bmatrix} \frac{11}{6} & -3 & \frac{3}{2} & -\frac{1}{3} \\ 0 & 0 & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix}.$$

p = 6

$$H = \begin{bmatrix} \frac{13649}{43200} & & & & & \\ & \frac{12013}{8640} & & & & \\ & & \frac{2711}{4320} & & & \\ & & & \frac{5359}{4320} & & \\ & & & & \frac{7877}{8640} & \\ & & & & & \frac{43801}{43200} \\ & & & & & 1 \\ & & & & & \ddots \\ & & & & & \end{bmatrix}.$$

$$D_6^{(1)} = \frac{1}{h} \begin{bmatrix} -\frac{21600}{13649} & \frac{104009}{54596} & \frac{30443}{81894} & -\frac{33311}{27298} & \frac{16863}{27298} & -\frac{15025}{163788} \\ -\frac{104009}{240260} & 0 & -\frac{311}{72078} & \frac{2029}{24026} & -\frac{24337}{48052} & \frac{36661}{360390} \\ -\frac{30443}{162660} & \frac{311}{32532} & 0 & -\frac{11155}{16266} & \frac{41287}{32532} & -\frac{21999}{54220} \\ \frac{33311}{107180} & -\frac{2029}{21436} & \frac{485}{1398} & 0 & \frac{4147}{21436} & \frac{25427}{321540} & \frac{72}{5359} \\ -\frac{16863}{78770} & \frac{24337}{31508} & -\frac{41287}{47262} & -\frac{4147}{15754} & 0 & \frac{342523}{472620} & -\frac{1296}{7877} & \frac{144}{7877} \\ \frac{15025}{525612} & -\frac{36661}{262806} & \frac{21999}{87602} & -\frac{25427}{262806} & -\frac{342523}{525612} & 0 & \frac{32400}{43801} & -\frac{6480}{43801} & \frac{720}{43801} \\ & & & & & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} \\ & & & & & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$D_6^{(2)} = \frac{1}{h^2} \left[\begin{array}{cccccc} \frac{114170}{40947} & -\frac{438107}{54596} & \frac{336409}{40947} & -\frac{276997}{81894} & \frac{3747}{13649} & \frac{21035}{163788} \\ \frac{6173}{5860} & -\frac{2066}{879} & \frac{3283}{1758} & -\frac{303}{293} & \frac{2111}{3516} & -\frac{601}{4395} \\ -\frac{52391}{81330} & \frac{134603}{32532} & -\frac{21982}{2711} & \frac{112915}{16266} & -\frac{46969}{16266} & \frac{30409}{54220} \\ \frac{68603}{321540} & -\frac{12423}{10718} & \frac{112915}{32154} & -\frac{75934}{16077} & \frac{53369}{21436} & -\frac{54899}{160770} & \frac{48}{5359} \\ -\frac{7053}{39385} & \frac{86551}{94524} & -\frac{46969}{23631} & \frac{53369}{15754} & -\frac{87904}{23631} & \frac{820271}{472620} & -\frac{1296}{7877} & \frac{96}{7877} \\ \frac{21035}{525612} & -\frac{24641}{131403} & \frac{30409}{87602} & -\frac{54899}{131403} & \frac{820271}{525612} & -\frac{117600}{43801} & \frac{64800}{43801} & -\frac{6480}{43801} & \frac{480}{43801} \\ & & & & \frac{1}{90} & -\frac{3}{20} & \frac{3}{2} & -\frac{49}{18} & \frac{3}{2} & -\frac{3}{20} & \frac{1}{90} \\ & & & & \frac{1}{90} & -\frac{3}{20} & \frac{3}{2} & -\frac{49}{18} & \frac{3}{2} & -\frac{3}{20} & \frac{1}{90} \\ & & & & & \ddots \end{array} \right],$$

$$S = \frac{1}{h^2} \left[\begin{array}{ccccc} \frac{25}{12} & -4 & 3 & -\frac{4}{3} & \frac{1}{4} \\ 0 & & 0 & & \\ & & & \ddots & \end{array} \right].$$

p = 8

We denote the elements

- of H_0 by h_{ij} ,
- of $hD_8^{(1)}$ by $D1_{i,j}$,
- of $hD_8^{(2)}$ by $D2_{i,j}$, $i, j = 1, 2, \dots$

$$h_{11} = \frac{1498139}{5080320} \quad h_{33} = \frac{20761}{80640} \quad h_{55} = \frac{299527}{725760} \quad h_{77} = \frac{670091}{725760}$$

$$h_{22} = \frac{1107307}{725760} \quad h_{44} = \frac{1304999}{725760} \quad h_{66} = \frac{103097}{80640} \quad h_{88} = \frac{5127739}{5080320}.$$

Interior stencil of $hD_8^{(1)}$:

$$\begin{aligned} hD_8^{(1)} v_j = & \frac{1}{280} v_{j-4} - \frac{4}{105} v_{j-3} + \frac{1}{5} v_{j-2} - \frac{4}{5} v_{j-1} + \frac{4}{5} v_{j+1} - \frac{1}{5} v_{j+2} \\ & + \frac{4}{105} v_{j+3} - \frac{1}{280} v_{j+4}. \end{aligned}$$

$$\begin{aligned} D1_{1,1} &= \frac{-2540160}{1498139} & D1_{3,1} &= \frac{-66264997}{8719620} & D1_{5,1} &= \frac{-20708767}{2096689} & D1_{7,1} &= \frac{-27390659}{56287644} \\ D1_{1,2} &= \frac{5544277}{5992556} & D1_{3,2} &= \frac{944709}{415220} & D1_{5,2} &= \frac{165990199}{3594324} & D1_{7,2} &= \frac{7568311}{2680364} \\ D1_{1,3} &= \frac{198794991}{29962780} & D1_{3,3} &= 0 & D1_{5,3} &= \frac{-96962637}{1198108} & D1_{7,3} &= \frac{-22524966}{3350455} \\ D1_{1,4} &= \frac{-256916579}{17977668} & D1_{3,4} &= \frac{-20335981}{249132} & D1_{5,4} &= \frac{68748371}{1198108} & D1_{7,4} &= \frac{66558305}{8041092} \\ D1_{1,5} &= \frac{20708767}{1498139} & D1_{3,5} &= \frac{32320879}{249132} & D1_{5,5} &= 0 & D1_{7,5} &= \frac{-14054993}{2680364} \\ D1_{1,6} &= \frac{-41004357}{5992556} & D1_{3,6} &= \frac{-35518713}{415220} & D1_{5,6} &= \frac{-27294549}{1198108} & D1_{7,6} &= \frac{2084949}{2680364} \\ D1_{1,7} &= \frac{27390659}{17977668} & D1_{3,7} &= \frac{2502774}{103805} & D1_{5,7} &= \frac{14054993}{1198108} & D1_{7,7} &= 0 \\ D1_{1,8} &= \frac{-2322531}{29962780} & D1_{3,8} &= \frac{-3177073}{1743924} & D1_{5,8} &= \frac{-42678199}{25160268} & D1_{7,8} &= \frac{70710683}{93812740} \\ D1_{1,9} &= 0 & D1_{3,9} &= 0 & D1_{5,9} &= \frac{-2592}{299527} & D1_{7,9} &= \frac{-145152}{670091} \\ D1_{1,10} &= 0 & D1_{3,10} &= 0 & D1_{5,10} &= 0 & D1_{7,10} &= \frac{27648}{670091} \\ D1_{1,11} &= 0 & D1_{3,11} &= 0 & D1_{5,11} &= 0 & D1_{7,11} &= -\frac{2592}{670091} \\ D1_{1,12} &= 0 & D1_{3,12} &= 0 & D1_{5,12} &= 0 & D1_{7,12} &= 0 \\ D1_{2,1} &= \frac{-5544277}{31004596} & D1_{4,1} &= \frac{256916579}{109619916} & D1_{6,1} &= \frac{13668119}{8660148} & D1_{8,1} &= \frac{2323531}{102554780} \\ D1_{2,2} &= 0 & D1_{4,2} &= \frac{-49607267}{5219996} & D1_{6,2} &= \frac{-850651}{103097} & D1_{8,2} &= \frac{-48319961}{307664340} \\ D1_{2,3} &= \frac{-85002381}{22146140} & D1_{4,3} &= \frac{61007943}{5219996} & D1_{6,3} &= \frac{35518713}{2061940} & D1_{8,3} &= \frac{9531219}{20510956} \\ D1_{2,4} &= \frac{49607267}{4429228} & D1_{4,4} &= 0 & D1_{6,4} &= \frac{-21696041}{1237164} & D1_{8,4} &= \frac{-3870214}{5127739} \\ D1_{2,5} &= \frac{-165990199}{13287684} & D1_{4,5} &= \frac{-68748371}{5219996} & D1_{6,5} &= \frac{9098183}{1237164} & D1_{8,5} &= \frac{2246221}{3238572} \\ D1_{2,6} &= \frac{7655859}{1107307} & D1_{4,6} &= \frac{65088123}{5219996} & D1_{6,6} &= 0 & D1_{8,6} &= \frac{-21360021}{102554780} \\ D1_{2,7} &= \frac{-7568311}{4429228} & D1_{4,7} &= \frac{-66558305}{15659988} & D1_{6,7} &= \frac{-231661}{412388} & D1_{8,7} &= \frac{-70710683}{102554780} \\ D1_{2,8} &= \frac{48319961}{465068940} & D1_{4,8} &= \frac{3870214}{9134993} & D1_{6,8} &= \frac{7120007}{43300740} & D1_{8,8} &= 0 \\ D1_{2,9} &= 0 & D1_{4,9} &= 0 & D1_{6,9} &= \frac{3072}{103097} & D1_{8,9} &= \frac{4064256}{5127739} \\ D1_{2,10} &= 0 & D1_{4,10} &= 0 & D1_{6,10} &= \frac{-288}{103097} & D1_{8,10} &= \frac{-1016064}{5127739} \\ D1_{2,11} &= 0 & D1_{4,11} &= 0 & D1_{6,11} &= 0 & D1_{8,11} &= \frac{193536}{5127739} \\ D1_{2,12} &= 0 & D1_{4,12} &= 0 & D1_{6,12} &= 0 & D1_{8,12} &= \frac{-18144}{5127739}. \end{aligned}$$

Interior stencil of $h^2 D_8^{(2)}$:

$$\begin{aligned} h^2 D_8^{(2)} v_j = & -\frac{1}{560} v_{j-4} + \frac{8}{315} v_{j-3} - \frac{1}{5} v_{j-2} + \frac{8}{5} v_{j-1} - \frac{205}{72} v_j \\ & + \frac{8}{5} v_{j+1} - \frac{1}{5} v_{j+2} + \frac{8}{315} v_{j+3} - \frac{1}{560} v_{j+4}. \end{aligned}$$

$D2_{1,1} = \frac{4870382994799}{1358976868290}$	$D2_{3,1} = \frac{7838984095}{52731029988}$	$D2_{5,1} = \frac{1455067816}{21132528431}$	$D2_{7,1} = \frac{-135555328849}{8509847458140}$
$D2_{1,2} = \frac{-893640087518}{75498714905}$	$D2_{3,2} = \frac{1168338040}{5649753213}$	$D2_{5,2} = \frac{-171562838}{3018932633}$	$D2_{7,2} = \frac{11904122576}{101307707835}$
$D2_{1,3} = \frac{926594825119}{60398971924}$	$D2_{3,3} = \frac{-88747895}{144865467}$	$D2_{5,3} = \frac{-43205598281}{36227191596}$	$D2_{7,3} = \frac{-5124426509}{13507694378}$
$D2_{1,4} = \frac{-1315109406200}{135897686829}$	$D2_{3,4} = \frac{423587231}{627750357}$	$D2_{5,4} = \frac{48242560214}{9056797899}$	$D2_{7,4} = \frac{43556319241}{60784624701}$
$D2_{1,5} = \frac{39126983272}{15099742981}$	$D2_{3,5} = \frac{-43205598281}{22599012852}$	$D2_{5,5} = \frac{-52276055645}{6037865266}$	$D2_{7,5} = \frac{-80321706377}{8104616268}$
$D2_{1,6} = \frac{12344491342}{75498714905}$	$D2_{3,6} = \frac{4876378562}{1883251071}$	$D2_{5,6} = \frac{57251587238}{9056797899}$	$D2_{7,6} = \frac{73790130002}{33769235945}$
$D2_{1,7} = \frac{-451560522577}{2717953736580}$	$D2_{3,7} = \frac{-5124426509}{3766502142}$	$D2_{5,7} = \frac{-80321706377}{36227191596}$	$D2_{7,7} = \frac{-950494905688}{303923123505}$
$D2_{1,8} = 0$	$D2_{3,8} = \frac{10496900965}{39548272491}$	$D2_{5,8} = \frac{8078087158}{21132528431}$	$D2_{7,8} = \frac{239073018673}{141830790969}$
$D2_{1,9} = 0$	$D2_{3,9} = 0$	$D2_{5,9} = \frac{-1296}{299527}$	$D2_{7,9} = \frac{-145152}{670091}$
$D2_{1,10} = 0$	$D2_{3,10} = 0$	$D2_{5,10} = 0$	$D2_{7,10} = \frac{18432}{670091}$
$D2_{1,11} = 0$	$D2_{3,11} = 0$	$D2_{5,11} = 0$	$D2_{7,11} = \frac{-1296}{670091}$
$D2_{1,12} = 0$	$D2_{3,12} = 0$	$D2_{5,12} = 0$	$D2_{7,12} = 0$
$D2_{2,1} = \frac{333806012194}{390619153855}$	$D2_{4,1} = \frac{-94978241528}{828644350023}$	$D2_{6,1} = \frac{10881504334}{327321118845}$	$D2_{8,1} = 0$
$D2_{2,2} = \frac{-116464627209}{111605472530}$	$D2_{4,2} = \frac{82699112501}{157837019052}$	$D2_{6,2} = \frac{-28244698346}{14028479505}$	$D2_{8,2} = \frac{-2598164715}{206729925524}$
$D2_{2,3} = \frac{1168338040}{33481641759}$	$D2_{4,3} = \frac{1270761693}{13153084921}$	$D2_{6,3} = \frac{4876378562}{9352031967}$	$D2_{8,3} = \frac{10496900965}{155047444143}$
$D2_{2,4} = \frac{82699112501}{133926567036}$	$D2_{4,4} = \frac{-167389605005}{11837764289}$	$D2_{6,4} = \frac{-1057998671}{12469375956}$	$D2_{8,4} = \frac{-44430275135}{310094888286}$
$D2_{2,5} = \frac{-171562838}{1160547253}$	$D2_{4,5} = \frac{48242560214}{39459254763}$	$D2_{6,5} = \frac{57251587238}{28056095951}$	$D2_{8,5} = \frac{425162482}{2720130599}$
$D2_{2,6} = \frac{-28244698346}{167408208795}$	$D2_{4,6} = \frac{-31673996013}{52612339684}$	$D2_{6,6} = \frac{-137531401019}{93520319670}$	$D2_{8,6} = \frac{-137529995233}{6201897765752}$
$D2_{2,7} = \frac{11904122576}{167408208795}$	$D2_{4,7} = \frac{43556319241}{11837764289}$	$D2_{6,7} = \frac{73790130002}{46760159835}$	$D2_{8,7} = \frac{239073018673}{155047444143}$
$D2_{2,8} = \frac{-2598164715}{312495323084}$	$D2_{4,8} = \frac{-44430275135}{552429566682}$	$D2_{6,8} = \frac{-137529995233}{785570685228}$	$D2_{8,8} = \frac{-14464800000}{51682481381}$
$D2_{2,9} = 0$	$D2_{4,9} = 0$	$D2_{6,9} = \frac{2048}{103097}$	$D2_{8,9} = \frac{8128512}{5127739}$
$D2_{2,10} = 0$	$D2_{4,10} = 0$	$D2_{6,10} = \frac{-144}{103097}$	$D2_{8,10} = \frac{-1016064}{5127739}$
$D2_{2,11} = 0$	$D2_{4,11} = 0$	$D2_{6,11} = 0$	$D2_{8,11} = \frac{129024}{5127739}$
$D2_{2,12} = 0$	$D2_{4,12} = 0$	$D2_{6,12} = 0$	$D2_{8,12} = \frac{-9072}{5127739},$

$$S = \frac{1}{h^2} \left[\begin{array}{ccccccccc} \frac{4723}{2100} & -\frac{839}{175} & \frac{157}{35} & -\frac{278}{105} & \frac{103}{140} & \frac{1}{175} & -\frac{6}{175} & & \\ & 0 & & & & & & \ddots & \\ & & & & & & & & \end{array} \right].$$

B.2 Full H_0 -norm

In this section there is no restriction on the structure of the submatrix H_0 in the norm. The local order of accuracy near the boundary is $p - 1$ for $D_p^{(1)}$ and $p - 2$ for $D_p^{(2)}$. The first row in S has local order of accuracy $p - 1$.

$$\underline{\mathbf{p} = 4}$$

Let $r1$ and $r2$ be defined by

$$r1 = -\frac{2177\sqrt{295369}-1166427}{25488}, \quad r2 = \frac{66195\sqrt{53}\sqrt{5573}-35909375}{101952}.$$

The elements h_{ij} of the symmetric matrix H_0 are

$$\begin{aligned} h_{11} &= -\frac{216r2+2160r1-2125}{12960} & h_{23} &= \frac{1836r2+14580r1+7295}{2160} \\ h_{12} &= \frac{81r2+675r1+415}{540} & h_{24} &= -\frac{216r2+2160r1+655}{4320} \\ h_{13} &= -\frac{72r2+720r1+445}{1440} & h_{33} &= -\frac{4104r2+32400r1+12785}{4320} \\ h_{14} &= -\frac{108r2+756r1+421}{1296} & h_{34} &= \frac{81r2+675r1+335}{540} \\ h_{22} &= -\frac{4104r2+32400r1+11225}{4320} & h_{44} &= -\frac{216r2+2160r1-12085}{12960}. \end{aligned}$$

The approximation of $\partial/\partial x$ is defined as $D_4^{(1)} = H^{-1}Q$, where the elements of Q are defined by

$$\begin{aligned} q_{11} &= -\frac{1}{2} & q_{23} &= -\frac{864r2+6480r1+2315}{1440} \\ q_{12} &= -\frac{864r2+6480r1+305}{4320} & q_{24} &= \frac{108r2+810r1+415}{270} \\ q_{13} &= \frac{216r2+1620r1+725}{540} & q_{34} &= -\frac{864r2+6480r1+785}{4320} \\ q_{14} &= -\frac{864r2+6480r1+3335}{4320} \end{aligned}$$

and

$$Q = \frac{1}{h} \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ -q_{12} & 0 & q_{23} & q_{24} \\ -q_{13} & -q_{23} & 0 & q_{34} & -\frac{1}{12} \\ -q_{14} & -q_{24} & -q_{34} & 0 & \frac{2}{3} \\ & & \frac{1}{12} & -\frac{2}{3} & 0 \\ & & & \frac{1}{12} & -\frac{2}{3} \\ & & & & 0 \\ & & & & \frac{2}{3} \\ & & & & -\frac{1}{12} \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Interior stencil of $h^2 D_4^{(2)}$:

$$h^2 D_4^{(2)} v_j = -\frac{1}{12} v_{j-3} + \frac{4}{3} v_{j-1} - \frac{5}{2} v_j + \frac{4}{3} v_{j+1} - \frac{1}{12} v_{j+2}.$$

Elements $D_{2,i,j}$ of $h^2 D_4^{(2)}$ near the boundary:

$$\begin{aligned} D_{2,1,1} &= 2.D0 & D_{2,2,3} &= 0.4451860342366326D0 & D_{2,3,5} &= -0.370366153552986D - 2 \\ D_{2,1,2} &= -5.D0 & D_{2,2,4} &= 0.3889216118347602D0 & D_{2,3,6} &= -0.4126377895574793D - 2 \\ D_{2,1,3} &= 4.D0 & D_{2,2,5} &= -0.1115146289530765D0 & D_{2,4,1} &= -0.1972518376035006D - 1 \\ D_{2,1,4} &= -1.D0 & D_{2,2,6} &= 0.5713690397754591D - 2 & D_{2,4,2} &= -0.3132697803201818D - 2 \\ D_{2,1,5} &= 0.D0 & D_{2,3,1} &= -0.2020917311782903D - 1 & D_{2,4,3} &= 0.1209782628816308D1 \\ D_{2,1,6} &= 0.D0 & D_{2,3,2} &= 0.1076710314575741D1 & D_{2,4,4} &= -0.2413299862026212D1 \\ D_{2,2,1} &= 0.9113401326379418D0 & D_{2,3,3} &= -0.2104749527124675D1 & D_{2,4,5} &= 0.1308408547618058D1 \\ D_{2,2,2} &= -0.1639646840154013D1 & D_{2,3,4} &= 0.1056078425097867D1 & D_{2,4,6} &= -0.8203343284460206D - 1, \end{aligned}$$

$$S = \frac{1}{h^2} \left[\begin{array}{cccc} \frac{11}{6} & -3 & \frac{3}{2} & -\frac{1}{3} \\ & 0 & & \\ & & 0 & \\ & & & \ddots \end{array} \right].$$

p = 6

Let $r1, r2, r3$ be defined by

$$r1 = -3.6224891259957, r2 = 96.301901955532, r3 = -609.05813881563.$$

Then the elements of H_0 are

$$\begin{aligned} h_{11} &= -\frac{14400 * r2 + 302400 * r1 - 7420003}{3.6288e7} & h_{33} &= -\frac{51300 * r3 + 1094400 * r2 + 9585000 * r1 - 39593423}{64800} \\ h_{12} &= -\frac{75600 * r3 + 1497600 * r2 + 11944800 * r1 - 59330023}{2.17728e7} & h_{34} &= \frac{120960 * r3 + 2584800 * r2 + 226800000 * r1 - 93310367}{181440} \\ h_{13} &= -\frac{9450 * r3 + 202050 * r2 + 1776600 * r1 - 7225847}{340200} & h_{35} &= \frac{5400 * r3 + 104400 * r2 + 810000 * r1 - 3766003}{129600} \\ h_{14} &= \frac{900 * r2 + 18900 * r1 - 649}{226800} & h_{36} &= \frac{900 * r2 + 18900 * r1 - 37217}{226800} \\ h_{15} &= \frac{86400 * r3 + 1828800 * r2 + 15854400 * r1 - 66150023}{3110400} & h_{44} &= -\frac{17100 * r3 + 364800 * r2 + 3195000 * r1 - 13184701}{21600} \\ h_{16} &= \frac{378000 * r3 + 7747200 * r2 + 65167200 * r1 - 279318239}{1.08864e8} & h_{45} &= \frac{3780 * r3 + 82575 * r2 + 741825 * r1 - 2976857}{34020} \\ h_{22} &= \frac{302400 * r3 + 6091200 * r2 + 49896000 * r1 - 210294289}{7257600} & h_{46} &= -\frac{1890 * r3 + 40410 * r2 + 355320 * r1 - 1458223}{68040} \\ h_{23} &= \frac{3780 * r3 + 82575 * r2 + 741825 * r1 - 2991977}{34020} & h_{55} &= \frac{302400 * r3 + 6091200 * r2 + 49896000 * r1 - 213056209}{7257600} \\ h_{24} &= \frac{5400 * r3 + 104400 * r2 + 810000 * r1 - 3756643}{129600} & h_{56} &= -\frac{75600 * r3 + 1497600 * r2 + 11944800 * r1 - 54185191}{2.17728e7} \\ h_{25} &= -\frac{529200 * r3 + 11107200 * r2 + 95508000 * r1 - 400851749}{2419200} & h_{66} &= -\frac{14400 * r2 + 302400 * r1 - 36797603}{3.6288e7} \\ h_{26} &= \frac{86400 * r3 + 1828800 * r2 + 15854400 * r1 - 65966279}{3110400}. \end{aligned}$$

$D_6^{(1)}$ has the form $D_6^{(1)} = H^{-1}Q$, where the elements of Q are:

$$\begin{aligned}
q_{11} &= \frac{-1}{2} q_{12} = \frac{415800 \cdot r^3 + 8604000 \cdot r^2 + 72954000 \cdot r^1 - 283104553}{3.26592e7} & q_{33} &= 0 \\
q_{13} &= \frac{120960 \cdot r^3 + 2672640 \cdot r^2 + 24192000 \cdot r^1 - 100358119}{6531840} & q_{34} &= \frac{-6993000 \cdot r^3 + 148096800 \cdot r^2 + 1286334000 \cdot r^1 - 5353075351}{8164800} \\
q_{14} &= \frac{-25200 \cdot r^3 + 542400 \cdot r^2 + 4788000 \cdot r^1 - 19717139}{403200} & q_{35} &= \frac{21168000 \cdot r^3 + 449049600 \cdot r^2 + 3907008000 \cdot r^1 - 16212561187}{3.26592e7} \\
q_{15} &= \frac{604800 \cdot r^3 + 13363200 \cdot r^2 + 120960000 \cdot r^1 - 485628701}{3.26592e7} & q_{36} &= \frac{-75600 \cdot r^3 + 1627200 \cdot r^2 + 14364000 \cdot r^1 - 58713721}{1209600} \\
q_{16} &= \frac{41580 \cdot r^3 + 860400 \cdot r^2 + 7295400 \cdot r^1 - 31023481}{3265920} & q_{44} &= 0 \\
q_{22} &= 0 & q_{45} &= \frac{-9450000 \cdot r^3 + 200635200 \cdot r^2 + 1747116000 \cdot r^1 - 7263657599}{3.26592e7} \\
q_{23} &= \frac{-9450000 \cdot r^3 + 200635200 \cdot r^2 + 1747116000 \cdot r^1 - 7286801279}{3.26592e7} & q_{46} &= \frac{604800 \cdot r^3 + 13363200 \cdot r^2 + 120960000 \cdot r^1 - 485920643}{3.26592e7} \\
q_{24} &= \frac{21168000 \cdot r^3 + 449049600 \cdot r^2 + 3907008000 \cdot r^1 - 16231108387}{3.26592e7} & q_{55} &= 0 \\
q_{25} &= -\frac{165375 \cdot r^3 + 3516300 \cdot r^2 + 30656250 \cdot r^1 - 126996371}{453600} & q_{56} &= \frac{415800 \cdot r^3 + 8604000 \cdot r^2 + 72954000 \cdot r^1 - 286439017}{3.26592e7} \\
q_{26} &= \frac{604800 \cdot r^3 + 13363200 \cdot r^2 + 120960000 \cdot r^1 - 482536157}{3.26592e7} & q_{66} &= 0
\end{aligned}$$

$$Q = \frac{1}{h} \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} & q_{15} & q_{16} \\ -q_{12} & q_{22} & q_{23} & q_{24} & q_{25} & q_{26} \\ -q_{13} & -q_{23} & q_{33} & q_{34} & q_{35} & q_{36} \\ -q_{14} & -q_{24} & -q_{34} & q_{44} & q_{45} & q_{46} & \frac{1}{60} \\ -q_{15} & -q_{25} & -q_{35} & -q_{45} & q_{55} & q_{56} & -\frac{3}{20} & \frac{1}{60} \\ -q_{16} & -q_{26} & -q_{36} & -q_{46} & -q_{56} & q_{66} & \frac{3}{4} & -\frac{3}{20} \\ & & & & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{5}{4} & -\frac{3}{20} & \frac{1}{60} \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Interior stencil of $h^2 D_6^{(2)}$:

$$h^2 D_6^{(2)} v_j = \frac{1}{90} v_{j-3} - \frac{3}{20} v_{j-2} + \frac{3}{2} v_{j-1} - \frac{49}{18} v_j + \frac{3}{2} v_{j+1} - \frac{3}{120} v_{j+2} + \frac{1}{90} v_{j+3}.$$

$$\begin{aligned}
D_{2,1,1} &= 0.3548420602490798D1 & D_{2,3,1} &= -0.5393093966319141D - 1 & D_{2,5,1} &= 0.1623318041994786D - 1 \\
D_{2,1,2} &= -0.1162385694827807D2 & D_{2,3,2} &= 0.1153943542621719D1 & D_{2,5,2} &= -0.8794616833597996D - 1 \\
D_{2,1,3} &= 0.1480964237069501D2 & D_{2,3,3} &= -0.2040716873611299D1 & D_{2,5,3} &= 0.103577624811612D0 \\
D_{2,1,4} &= -0.8968412049815223D1 & D_{2,3,4} &= 0.698739734417074D0 & D_{2,5,4} &= 0.114967901600216D1 \\
D_{2,1,5} &= 0.2059642370694317D1 & D_{2,3,5} &= 0.421429883414006D0 & D_{2,5,5} &= -0.2443599523155367D1 \\
D_{2,1,6} &= 0.3761430517226221D0 & D_{2,3,6} &= -0.2262171762222378D0 & D_{2,5,6} &= 0.1375113224609842D1 \\
D_{2,1,7} &= -0.2015793975095019D0 & D_{2,3,7} &= 0.5090670369467911D - 1 & D_{2,5,7} &= -0.1218565837960692D0 \\
D_{2,1,8} &= 0.5117538641997827D - 13 & D_{2,3,8} &= -0.4371323842747547D - 2 & D_{2,5,8} &= 0.8668492495883396D - 2 \\
D_{2,1,9} &= -0.3386357570016522D - 15 & D_{2,3,9} &= 0.2245491919975288D - 3 & D_{2,5,9} &= 0.1307369479706344D - 3 \\
D_{2,2,1} &= 0.857883182233682D0 & D_{2,4,1} &= -0.2032638843942139D - 1 & D_{2,6,1} &= -0.3185308684167192D - 2 \\
D_{2,2,2} &= -0.1397247220064007D1 & D_{2,4,2} &= 0.4181668262047738D - 1 & D_{2,6,2} &= 0.1943844988205038D - 1 \\
D_{2,2,3} &= 0.3461647289468133D - 1 & D_{2,4,3} &= 0.1009041221554696D1 & D_{2,6,3} &= -0.3865422059089032D - 1 \\
D_{2,2,4} &= 0.6763679122231971D0 & D_{2,4,4} &= -0.2044119911750601D1 & D_{2,6,4} &= -0.8123817099768654D - 1 \\
D_{2,2,5} &= -0.1325900419870384D0 & D_{2,4,5} &= 0.9609112011420257D0 & D_{2,6,5} &= 0.1445296692538394D1 \\
D_{2,2,6} &= -0.6345391502339508D - 1 & D_{2,4,6} &= 0.9142374273488277D - 1 & D_{2,6,6} &= -0.2697689107917306D1 \\
D_{2,2,7} &= 0.244383001412735D - 1 & D_{2,4,7} &= -0.4316909959745465D - 1 & D_{2,6,7} &= 0.1494463382995396D1 \\
D_{2,2,8} &= -0.2800316968929196D - 4 & D_{2,4,8} &= 0.4668725019017949D - 2 & D_{2,6,8} &= -0.1495167135596915D0 \\
D_{2,2,9} &= 0.1331275129575954D - 4 & D_{2,4,9} &= -0.2461732836225921D - 3 & D_{2,6,9} &= 0.110849963339009D - 1
\end{aligned}$$

$$S = \frac{1}{h^2} \begin{bmatrix} -s_1 & -s_2 & -s_3 & -s_4 & -s_5 & -s_6 & -s_7 \\ 0 & & & & & & \\ & 0 & & & & & \\ & & \ddots & & & & \\ & & & & & & \end{bmatrix},$$

where

$$\begin{aligned} s_1 &= \frac{278586692617}{123868739203} & s_3 &= -\frac{555639772335}{123868739203} & s_5 &= -\frac{91132000935}{123868739203} & s_7 &= \frac{386084381}{11260794473} \\ s_2 &= \frac{593862126054}{123868739203} & s_4 &= \frac{327957232980}{123868739203} & s_6 &= -\frac{707821338}{123868739203}. \end{aligned}$$

p = 8

The elements of H_0 are

$$\begin{aligned} h_{11} &= .17278828151304213131 & h_{36} &= -.86170509476217520096 \\ h_{12} &= .26442559491473446335 & h_{37} &= .23272822377785868580 \\ h_{13} &= -.31178196884312775720 & h_{38} &= -.27718561649605047808e - 1 \\ h_{14} &= .30943600415290628943 & h_{44} &= 4.5048009713827725943 \\ h_{15} &= -.22687886379977231568 & h_{45} &= -2.8194294132418145464 \\ h_{16} &= .10699051704359999240 & h_{46} &= 1.4139447698567646313 \\ h_{17} &= -.29198436836537438395e - 1 & h_{47} &= -.40168231609234325290 \\ h_{18} &= .35313167405158916567e - 2 & h_{48} &= .49803339945923016666e - 1 \\ h_{22} &= 1.7526198409167495775 & h_{55} &= 3.3650057528838514157 \\ h_{23} &= -1.1935018761521814908 & h_{56} &= -1.2261739857055893831 \\ h_{24} &= 1.3954360831111377900 & h_{57} &= .35734819363435995274 \\ h_{25} &= -1.0175504328863858993 & h_{58} &= -.45187886555639925953e - 1 \\ h_{26} &= .47894397921583602822 & h_{66} &= 1.6529897067349637657 \\ h_{27} &= -.13175544973723205885 & h_{67} &= -.19435314011455700920 \\ h_{28} &= .16150503562667126340e - 1 & h_{68} &= .24990727889886053755e - 1 \\ h_{33} &= 2.9309103416555983901 & h_{77} &= 1.0588289336073241051 \\ h_{34} &= -2.4589576423781345782 & h_{78} &= -.76680474804950146879e - 2 \\ h_{35} &= 1.8303365981930377633 & h_{88} &= 1.0010106996984233357. \end{aligned}$$

Interior stencil of $hD_8^{(1)}$:

$$\begin{aligned} hD_8^{(1)} v_j = & \frac{1}{280} v_{j-4} - \frac{4}{105} v_{j-3} + \frac{1}{5} v_{j-2} - \frac{4}{5} v_{j-1} \\ & + \frac{4}{5} v_{j+1} - \frac{1}{5} v_{j+2} + \frac{4}{105} v_{j+3} - \frac{1}{280} v_{j+4}. \end{aligned}$$

Elements $D1_{i,j}$ of $hD_8^{(1)}$ near the boundary:

$D1_{1,1} = -2.592857142857143d0$	$D1_{3,9} = -0.0010055472165286575d0$	$D1_{6,5} = -1.1039047461210942d0$
$D1_{1,2} = 7.0d0$	$D1_{3,10} = 3.8794759110886745d-4$	$D1_{6,6} = 0.27359438161151506d0$
$D1_{1,3} = -10.500000000000002d0$	$D1_{3,11} = -6.8704635776930775d-5$	$D1_{6,7} = 0.63460627616441168d0$
$D1_{1,4} = 11.66666666666667d0$	$D1_{3,12} = 5.0082327516880903d-6$	$D1_{6,8} = -0.1334700237623892d0$
$D1_{1,5} = -8.7500000000000053d0$	$D1_{4,1} = -0.01374973021859196d0$	$D1_{6,9} = 0.020553664597517285d0$
$D1_{1,6} = 4.2000000000000064d0$	$D1_{4,2} = 0.13486302290702401d0$	$D1_{6,10} = -5.2448374888321939d-4$
$D1_{1,7} = -1.1666666666666703d0$	$D1_{4,3} = -0.72696910885431376d0$	$D1_{6,11} = -3.5364415168681655d-4$
$D1_{1,8} = 0.14285714285714438d0$	$D1_{4,4} = 0.017812686448076227d0$	$D1_{6,12} = 1.9848170711942382d-5$
$D1_{1,9} = 0.$	$D1_{4,5} = 0.6393775076060112d0$	$D1_{7,1} = 0.0012022691922812573d0$
$D1_{1,10} = 0.$	$D1_{4,6} = 0.022147920488592564d0$	$D1_{7,2} = -0.0099404639226380095d0$
$D1_{1,11} = 0.$	$D1_{4,7} = -0.12565765302041595d0$	$D1_{7,3} = 0.039882284654387555d0$
$D1_{1,12} = 0.$	$D1_{4,8} = 0.068585850834641665d0$	$D1_{7,4} = -0.1150048724090931d0$
$D1_{2,1} = -0.14298292410192714d0$	$D1_{4,9} = -0.019233942555101392d0$	$D1_{7,5} = 0.30419289144259271d0$
$D1_{2,2} = -1.4489927866116712d0$	$D1_{4,10} = 0.0031486152221289702d0$	$D1_{7,6} = -0.89394403806786826d0$
$D1_{2,3} = 2.9964729239178829d0$	$D1_{4,11} = -3.4332575058906574d-4$	$D1_{7,7} = 0.056932295488355524d0$
$D1_{2,4} = -2.4929493873234359d0$	$D1_{4,12} = 1.8156892537446339d-5$	$D1_{7,8} = 0.77700125713081891d0$
$D1_{2,5} = 1.657878083897146d0$	$D1_{5,1} = 0.008862667686782706d0$	$D1_{7,9} = -0.19389096736102715d0$
$D1_{2,6} = -0.7430309325660327d0$	$D1_{5,2} = -0.081997207069678529d0$	$D1_{7,10} = 0.037021801268937171d0$
$D1_{2,7} = 0.19660315690279093d0$	$D1_{5,3} = 0.36016066918405126d0$	$D1_{7,11} = -0.0034453239617136533d0$
$D1_{2,8} = -0.022921081922677878d0$	$D1_{5,4} = -1.138622537383166d0$	$D1_{7,12} = -7.1586232168436659d-6$
$D1_{2,9} = -5.9765761149596457d-5$	$D1_{5,5} = 0.45753588027993108d0$	$D1_{8,1} = -1.7018888206312392d-4$
$D1_{2,10} = -2.086075830489387d-5$	$D1_{5,6} = 0.38914191081194799d0$	$D1_{8,2} = 0.001407457714362741d0$
$D1_{2,11} = 3.0132094578465812d-6$	$D1_{5,7} = 0.047353759946214852d0$	$D1_{8,3} = -0.0051428241139400005d0$
$D1_{2,12} = -6.3374647316878056d-8$	$D1_{5,8} = -0.06070785441402261d0$	$D1_{8,4} = 0.014469260937249902d0$
$D1_{3,1} = 0.024035178364099158d0$	$D1_{5,9} = 0.022163172460071972d0$	$D1_{8,5} = -0.052868823562666749d0$
$D1_{3,2} = -0.33511996288598189d0$	$D1_{5,10} = -0.004385536290694581d0$	$D1_{8,6} = 0.21333416109695102d0$
$D1_{3,3} = -0.77719249965062054d0$	$D1_{5,11} = 4.989505469285538d-4$	$D1_{8,7} = -0.80809452640150636d0$
$D1_{3,4} = 1.6547851227840236d0$	$D1_{5,12} = -2.7483320646974942d-5$	$D1_{8,8} = 0.0032794907268712007d0$
$D1_{3,5} = -0.81942145154188328d0$	$D1_{6,1} = -0.0035154238926503028d0$	$D1_{8,9} = 0.79912442343041246d0$
$D1_{3,6} = 0.32374316200170666d0$	$D1_{6,2} = 0.032627145467311741d0$	$D1_{8,10} = -0.19984491904967425d0$
$D1_{3,7} = -0.080342166504192805d0$	$D1_{6,3} = -0.1441805719128838d0$	$D1_{8,11} = 0.038076865351646491d0$
$D1_{3,8} = 0.010193913461294177d0$	$D1_{6,4} = 0.42454757757811995d0$	$D1_{8,12} = -0.0035703772476434379d0.$

Interior stencil of $h^2 D_8^{(2)}$:

$$\begin{aligned} h^2 D_8^{(2)} = & -\frac{1}{560} v_{j-4} + \frac{8}{315} v_{j-3} - \frac{1}{5} v_{j-2} + \frac{8}{5} v_{j-1} - \frac{205}{72} v_j \\ & + \frac{8}{5} v_{j+1} - \frac{1}{5} v_{j+2} + \frac{8}{315} v_{j+3} - \frac{1}{560} v_{j+4}. \end{aligned}$$

Elements $D_{2,i,j}$ of $hD_8^{(2)}$ near the boundary:

$$\begin{aligned} D_{2,1,1} &= 0.459559486573298D1 & D_{2,3,9} &= -0.3982789723630851D-1 & D_{2,5,5} &= 0.9837460406594241D0 \\ D_{2,1,2} &= -0.1737587003697495D2 & D_{2,3,10} &= 0.3613749551360845D-3 & D_{2,6,6} &= -0.2324548401306039D1 \\ D_{2,1,3} &= 0.2671554512941234D2 & D_{2,3,11} &= -0.4325584278035533D-4 & D_{2,6,7} &= 0.1311841356539494D1 \\ D_{2,1,4} &= -0.182533124810469D2 & D_{2,3,12} &= 0.2504116375844045D-5 & D_{2,6,8} &= -0.1004519515199941D0 \\ D_{2,1,5} &= -0.208613717646957D2 & D_{2,4,1} &= 0.41383452956972D-2 & D_{2,6,9} &= 0.5233628499353206D-2 \\ D_{2,1,6} &= 0.1436890974117532D2 & D_{2,4,2} &= -0.9317382863354466D-1 & D_{2,6,10} &= 0.545827493505434D-3 \\ D_{2,1,7} &= -0.115733475947655D2 & D_{2,4,3} &= 0.129627891150047D-3 & D_{2,6,11} &= -0.2121077126646392D-3 \\ D_{2,1,8} &= 0.4224129963025046D1 & D_{2,4,4} &= -0.2301454617412467D1 & D_{2,6,12} &= 0.9924085355971193D-5 \\ D_{2,1,9} &= -0.6155162453781307D0 & D_{2,4,5} &= 0.9496907454250803D0 & D_{2,7,1} &= 0.279876168104461D-2 \\ D_{2,1,10} &= 0.0D0 & D_{2,4,6} &= 0.3211654316623705D0 & D_{2,7,2} &= -0.2273831582632708D-1 \\ D_{2,1,11} &= 0.0D0 & D_{2,4,7} &= -0.2522929099188442D0 & D_{2,7,3} &= 0.7941253606056085D-1 \\ D_{2,1,12} &= 0.0D0 & D_{2,4,8} &= 0.93004304984534D-1 & D_{2,7,4} &= -0.1414647886625625D0 \\ D_{2,2,1} &= 0.76663483674837219D0 & D_{2,4,9} &= -0.1915026759150047D-1 & D_{2,7,5} &= 0.1676244442372628D-1 \\ D_{2,2,2} &= -0.9196803752511684D0 & D_{2,4,10} &= 0.2348749011339006D-2 & D_{2,7,6} &= 0.1416153480495742D1 \\ D_{2,2,3} &= -0.842207972743573D0 & D_{2,4,11} &= -0.2039417953611042D-3 & D_{2,7,7} &= -0.2745855902808057D1 \\ D_{2,2,4} &= 0.1034353161256694D1 & D_{2,4,12} &= 0.9078446268723169D-5 & D_{2,7,8} &= 0.156496049973078D1 \\ D_{2,2,5} &= 0.9224566003381633D0 & D_{2,5,1} &= 0.9702030996040648D-2 & D_{2,7,9} &= -0.192894978753717D0 \\ D_{2,2,6} &= -0.1915904204050834D1 & D_{2,5,2} &= -0.6794130076783107D-1 & D_{2,7,10} &= 0.2457122826770361D-1 \\ D_{2,2,7} &= 0.1358106271667133D1 & D_{2,5,3} &= 0.1333345846555384D0 & D_{2,7,11} &= -0.1709935539582438D-2 \\ D_{2,2,8} &= -0.4098818319034839D0 & D_{2,5,4} &= 0.9149526892124103D0 & D_{2,7,12} &= -0.3579311608421833D-5 \\ D_{2,2,9} &= 0.6642475553490265D-1 & D_{2,5,5} &= -0.1952769586741238D1 & D_{2,8,1} &= -0.4040249839128132D-3 \\ D_{2,2,10} &= -0.163599150014082D-4 & D_{2,5,6} &= 0.8452193170274778D0 & D_{2,8,2} &= 0.3282528546989832D-2 \\ D_{2,2,11} &= 0.1619270768597741D-5 & D_{2,5,7} &= 0.2160399635643034D0 & D_{2,8,3} &= -0.17215310517751D-1 \\ D_{2,2,12} &= -0.182303495391327D0 & D_{2,5,8} &= -0.1182303495391327D0 & D_{2,8,4} &= 0.2230119016828276D-1 \\ D_{2,3,1} &= -0.9930472508132245D-1 & D_{2,5,9} &= 0.2723622874009121D-1 & D_{2,8,5} &= -0.589253331879865D-2 \\ D_{2,3,2} &= 0.1494543277053003D1 & D_{2,5,10} &= -0.3328174251388732D-2 & D_{2,8,6} &= -0.1734623416535851D0 \\ D_{2,3,3} &= -0.31702882257908D1 & D_{2,5,11} &= 0.2983387640521609D-3 & D_{2,8,7} &= 0.1585367510775726D1 \\ D_{2,3,4} &= 0.2864206231232483D1 & D_{2,5,12} &= -0.1374160632348747D-4 & D_{2,8,8} &= -0.2842164305683002D1 \\ D_{2,3,5} &= -0.2207907708688962D1 & D_{2,6,1} &= -0.7955850076892026D-2 & D_{2,8,9} &= 0.15989732398365099D1 \\ D_{2,3,6} &= 0.1847596328819334D1 & D_{2,6,2} &= 0.137416063234801436D-1 & D_{2,8,10} &= -0.199880296017579D0 \\ D_{2,3,7} &= -0.9882048151539512D0 & D_{2,6,3} &= -0.2054352689917416D0 & D_{2,8,11} &= 0.2538577000496714D-1 \\ D_{2,3,8} &= 0.29886691124072D0 & D_{2,6,4} &= 0.2743737688500514D0 & D_{2,8,12} &= -0.1785188623821719D-2 \end{aligned}$$

$$S = \frac{1}{h^2} \begin{bmatrix} -s_1 & -s_2 & -s_3 & -s_4 & -s_5 & -s_6 & -s_7 & -s_8 & -s_9 \\ 0 & & & & & & & & \\ & 0 & & & & & & & \\ & & \ddots & & & & & & \\ s_1 & = & \frac{26605318914871}{10574000000000} s_6 & = & -\frac{259035026131}{2643500000000} \\ s_2 & = & \frac{16881394988747}{26435000000000} s_7 & = & \frac{5193568357271}{52870000000000} \\ s_3 & = & -\frac{44151764954129}{52870000000000} s_8 & = & -\frac{1245462146053}{26435000000000} \\ s_4 & = & \frac{19479098298429}{26435000000000} s_9 & = & \frac{76749811}{10000000000} \\ s_5 & = & -\frac{7142764970579}{2114800000000} \end{bmatrix}$$

B.3 A Padé Type Operator

The following 4th order operator (3rd order near the boundary) was given in [Carpenter et al., 1994].

$$P = \begin{bmatrix} \frac{211}{429} & 1 & 0 & 0 \\ 1 & \frac{3563}{1688} & -\frac{1}{8} & 0 \\ 0 & \frac{43}{17} & \frac{1893}{1054} & \frac{139}{186} \\ & & \frac{1}{4} & 1 \\ & & \frac{1}{4} & \frac{1}{4} \\ & & \frac{1}{4} & \frac{1}{4} \\ & & \frac{1}{4} & 1 \\ & & & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$Q = \frac{1}{h} \begin{bmatrix} -\frac{289}{234} & \frac{279}{286} & \frac{75}{1851} & -\frac{7}{2574} \\ -\frac{8635}{3376} & \frac{6987}{3376} & -\frac{286}{3376} & -\frac{203}{3376} \\ -\frac{15043}{18972} & -\frac{4089}{2108} & \frac{124}{18972} & \frac{29353}{18972} \\ 0 & 0 & -\frac{3}{4} & 0 \\ & & -\frac{3}{4} & \frac{3}{4} \\ & & -\frac{3}{4} & 0 \\ & & & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$H = \begin{bmatrix} \frac{70282007653}{7658388480} & -\frac{9426299}{2268480} & -\frac{192913}{1067520} & 0 \\ -\frac{55530689643}{2552796160} & \frac{8051589}{756160} & \frac{149823}{355840} & 0 \\ \frac{63842626133}{2552796160} & -\frac{9153739}{756160} & -\frac{4433}{355840} & -\frac{1}{8} \\ -\frac{1498870443}{7658388480} & \frac{10110149}{2268480} & \frac{102703}{1067520} & 1 \\ 0 & 0 & 0 & -\frac{1}{8} \\ & & & -\frac{1}{8} \\ & & & -\frac{1}{8} \\ & & & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

References

- [Abarbanel et al., 2000] Abarbanel, S., Ditkowsky, A., and Gustafsson, B. (2000). On error bounds of finite difference approximations to partial differential equations - temporal behavior and rate of convergence. *J. Scient. Comp.*, 15:79–116.
- [Brüger et al., 2005a] Brüger, A., Gustafsson, B., Lötstedt, P., and Nilsson, J. (2005a). High order accurate solution of the incompressible Navier–Stokes equations. *J. Comp. Phys.*, 203:49–71.
- [Brüger et al., 2005b] Brüger, A., Gustafsson, B., Lötstedt, P., and Nilsson, J. (2005b). Splitting methods for high order solution of the incompressible Navier–Stokes equations in 3D. *Int. J. for Numer. Meth. Fluids*, 47:1157–1163.
- [Butcher, 2003] Butcher, J. (2003). *Numerical Methods for ordinary differential equations. Runge-Kutta and general linear methods*. John Wiley & Sons.
- [Carpenter et al., 1994] Carpenter, M., Gottlieb, D., and Abarbanel, S. (1994). Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes. *J. Comp. Phys.*, 111:220–236.
- [Carpenter et al., 1995] Carpenter, M., Gottlieb, D., Abarbanel, S., and Don, W.-S. (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.*, 16:1241–1252.
- [Carpenter et al., 1999] Carpenter, M., Nordström, J., and Gottlieb, D. (1999). A stable and conservative interface treatment of arbitrary spatial accuracy. *J. Comp. Phys.*, 148:341–365.
- [Cockburn et al., 2000] Cockburn, B., Karniadakis, G., and Shu, C.-W. (2000). *Discontinuous Galerkin methods: theory, computation and applications*. Lecture Notes in Computational Science and engineering, Vol 11. Springer.
- [Collatz, 1966] Collatz, L. (1966). *The numerical treatment of differential equations*. Springer-Verlag, New York.
- [Courant et al., 1928] Courant, R., Friedrichs, K., and Levy, H. (1928). Über die partielle Differentialgleichungen der Mathematischen Physik. *Mathematische Annalen*, 100:32–74.
- [Dahlquist, 1963] Dahlquist, G. (1963). A special stability problem for linear multistep methods. *BIT*, 3:27–43.
- [Ditkovski et al., 2001] Ditkovski, A., Dridi, K., and Hesthaven, J. (2001). Convergent Cartesian grid methods for Maxwell’s equations in complex geometries. *J. Comp. Phys.*, 170:39–80.
- [Efraimsson, 1998] Efraimsson, G. (1998). A numerical method for the first-order wave equation with discontinuous initial data. *Numer. Methods Partial Differential Equations*, 14:353–365.
- [Engquist and Sjögreen, 1998] Engquist, B. and Sjögreen, B. (1998). The convergence rate of finite difference schemes in the presence of shocks. *SIAM J. Numer. Anal.*, 35:2464–2485.
- [Ferm and Lötstedt, 2002] Ferm, L. and Lötstedt, P. (2002). On numerical errors in the boundary conditions of the Euler equations. *Appl. Math. Comput.*, 128:129–140.
- [Fornberg, 1988] Fornberg, B. (1988). Generation of finite difference formulas on arbitrarily spaced grids. *Math. Comp.*, 51:699–706.

- [Fornberg, 1990] Fornberg, B. (1990). High-order finite differences and the pseudospectral method on staggered grids. *SIAM J. Numer. Anal.*, 27:904–918.
- [Fornberg, 1998] Fornberg, B. (1998). Calculation of weights in finite difference formulas. *SIAM Review*, 40:685–691.
- [Fornberg, 2003] Fornberg, B. (2003). Some numerical techniques for Maxwell's equations in different types of geometries. *Topics in Computational Wave Propagation. Lectures Notes in Computational Science and Engineering*, 31:265–299.
- [Funaro and Gottlieb, 1988] Funaro, D. and Gottlieb, D. (1988). A new method of imposing boundary conditions for hyperbolic equations. *Math. Comp.*, 51:599–613.
- [Ghia et al., 1982] Ghia, U., Ghia, K., and Shin, C. (1982). High Re solutions for incompressible flow using the Navier–Stokes equation and multigrid methods. *J. Comp. Phys.*, 48:387–411.
- [Godunov, 1959] Godunov, S. (1959). A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47:271–.
- [Godunov and Ryabenkii, 1963] Godunov, S. and Ryabenkii, V. (1963). Spectral criteria for the stability of boundary problems for non-self-adjoint difference equations. *Uspekhi Mat. Nauk*, 18.
- [Goldberg, 1991] Goldberg, M. (1991). Simple stability criteria for difference approximations of hyperbolic initial–boundary value problems. II. *Third International Conference on Hyperbolic Problems*, Eds., B. Engquist and B. Gustafsson, Studentlitteratur, Lund, Sweden, pages 519–527.
- [Goldberg and Tadmor, 1981] Goldberg, M. and Tadmor, E. (1981). Scheme-independent stability criteria for difference approximations of hyperbolic initial–boundary value problems. II. *Math. Comp.*, 36:605–626.
- [Goldberg and Tadmor, 1985] Goldberg, M. and Tadmor, E. (1985). Convenient stability criteria for difference approximations of hyperbolic initial–boundary value problems. *Math. Comp.*, 44:361–377.
- [Goldberg and Tadmor, 1987] Goldberg, M. and Tadmor, E. (1987). Convenient stability criteria for difference approximations of hyperbolic initial–boundary value problems. II. *Math. Comp.*, 48:503–520.
- [Goldberg and Tadmor, 1989] Goldberg, M. and Tadmor, E. (1989). Simple stability criteria for difference approximations of hyperbolic initial–boundary value problems. *Nonlinear Hyperbolic Equations - Theory, Numerical Methods and Applications*, Eds., J. Ballmann and R. Jeltsch, Vieweg Verlag, Braunschweig, Germany, pages 179–185.
- [Gottlieb et al., 2001] Gottlieb, S., Shu, C.-W., and Tadmor, E. (2001). Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43:89–112.
- [Gustafsson, 1975] Gustafsson, B. (1975). The convergence rate for difference approximations to mixed initial boundary value problems. *Mat. Comp.*, 29:396–406.
- [Gustafsson, 1981] Gustafsson, B. (1981). The convergence rate for difference approximations to general mixed initial boundary value problems. *SIAM J. Numer. Anal.*, 18:179–190.
- [Gustafsson, 1998] Gustafsson, B. (1998). On the implementation of boundary conditions for the method of lines. *BIT*, 38:293–314.
- [Gustafsson and Khalighi, 2006a] Gustafsson, B. and Khalighi, Y. (2006a). The shifted box scheme for scalar transport problems. *J. Scient. Comp.*, 28:319–335.
- [Gustafsson and Khalighi, 2006b] Gustafsson, B. and Khalighi, Y. (2006b). The shifted box scheme on nonuniform grids. *Submitted to J. Scient. Comp.*
- [Gustafsson et al., 1995] Gustafsson, B., Kreiss, H.-O., and Oliger, J. (1995). *Time dependent problems and difference methods*. John Wiley & Sons.
- [Gustafsson et al., 1972] Gustafsson, B., Kreiss, H.-O., and Sundström, A. (1972). Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26:649–686.
- [Gustafsson and Kress, 2001] Gustafsson, B. and Kress, W. (2001). Deferred correction methods for initial value problems. *BIT*, 41:986–995.
- [Gustafsson et al., 2003] Gustafsson, B., Lötstedt, P., and Göran, A. (2003). A fourth order difference method for the incompressible Navier–Stokes equations. *Numerical simulations of incompressible flows*, Editor: M.M. Hafez. World Scientific, pages 263–276.

- [Gustafsson and Mossberg, 2004] Gustafsson, B. and Mossberg, E. (2004). Time compact high order difference methods for wave propagation. *SIAM J. Sci. Comput.*, 26:259–271.
- [Gustafsson and Nilsson, 2002] Gustafsson, B. and Nilsson, J. (2002). Fourth order methods for the Stokes and Navier–Stokes equations on staggered grids. *Frontiers of Computational Fluid Dynamics 2002*, Eds: D.A. Caughey and M.M. Hafez, World Scientific, pages 165–178.
- [Gustafsson and Olsson, 1996] Gustafsson, B. and Olsson, P. (1996). High-order centered difference methods with sharp shock resolution. *J. Sci. Comput.*, 11:229–260.
- [Gustafsson and Wahlund, 2004] Gustafsson, B. and Wahlund, P. (2004). Time compact difference methods for wave propagation in discontinuous media. *SIAM J. Sci. Comput.*, 26:272–293.
- [Gustafsson and Wahlund, 2005] Gustafsson, B. and Wahlund, P. (2005). Time compact high order difference methods for wave propagation, 2D. *J. Sci. Comput.*, 25:195–211.
- [Hairer et al., 1993] Hairer, E., Nørset, S., and Wanner, G. (1993). *Solving ordinary differential equations I. (2nd rev. ed.)*. Springer Series in Computational Mathematics, Vol 8.
- [Hairer and Wanner, 1996] Hairer, E. and Wanner, G. (1996). *Solving ordinary differential equations II. (2nd rev. ed.)*. Springer Series in Computational Mathematics, Vol 8.
- [Harten et al., 1987] Harten, A., Engquist, B., Osher, S., and Chakravarthy, S. (1987). Uniformly high order accurate essentially nonoscillatory schemes, III. *J. Comp. Phys.*, 71:231–303.
- [Hesthaven et al., 2007] Hesthaven, J., Gottlieb, D., and Gottlieb, S. (2007). *Spectral methods for time-dependent problems*. Cambridge Monographs on Applied and Computational Mathematics 21, Cambridge University Press.
- [Hörmander, 1985] Hörmander, L. (1985). *The analysis of linear partial differential operators Vol I-IV*. Springer.
- [Jakobsson, 2006] Jakobsson, S. (2006). Frequency optimized computation methods. *J. Sci. Comput.*, 26:329–362.
- [Jeltsch and Nevanlinna, 1981] Jeltsch, R. and Nevanlinna, O. (1981). Stability of explicit time discretizations for solving initial value problems. *Numer. Math.*, 37:61–91.
- [Jeltsch and Nevanlinna, 1982] Jeltsch, R. and Nevanlinna, O. (1982). Stability and accuracy of time discretizations for initial value problems. *Numer. Math.*, 40:245–296.
- [Jeltsch and Nevanlinna, 1983] Jeltsch, R. and Nevanlinna, O. (1983). Stability of semidiscretizations of hyperbolic problems. *SIAM J. Numer. Anal.*, 20:1210–1218.
- [Jeltsch and Smit, 1987] Jeltsch, R. and Smit, J. (1987). Accuracy barriers of difference schemes for hyperbolic equations. *SIAM J. Numer. Anal.*, 24:1–11.
- [Jeltsch and Smit, 1998] Jeltsch, R. and Smit, J. (1998). Accuracy barriers for stable three-time-level difference schemes for hyperbolic equations. *IMA J. Numer. Anal.*, 18:445–484.
- [Jiang and Shu, 1996] Jiang, G.-S. and Shu, C.-W. (1996). Efficient implementation of weighted ENO schemes. *J. Comp. Phys.*, 126:202–228.
- [Johnson et al., 1990] Johnson, C., Szepessy, A., and Hansbo, P. (1990). On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws. *Math. Comp.*, 54:107–129.
- [Kashdan and Turkel, 2006] Kashdan, E. and Turkel, E. (2006). High-order accurate modeling of electromagnetic wave propagation across media - Grid conforming bodies. *J. Comp. Phys.*, 218:816–835.
- [Keller, 1971] Keller, H. (1971). A new difference scheme for parabolic problems. *Numerical solution of partial differential equations (Ed: B. Hubbard)*, Academics New York, 2:327–350.
- [Keller, 1978] Keller, H. (1978). Numerical methods in boundary-layer theory. *Ann. Rev. Fluid Mech.*, 10:417–433.
- [Kreiss, 1962] Kreiss, H.-O. (1962). Über die stabilitätsdefinition für differenzengleichungen die partielle differentialgleichungen approximieren. *BIT*, 2:153–181.
- [Kreiss, 1964] Kreiss, H.-O. (1964). On difference approximations of the dissipative type for hyperbolic differential equations. *Comm. Pure Appl. Math.*, 17:335–353.
- [Kreiss, 1968] Kreiss, H.-O. (1968). Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22:703–714.
- [Kreiss, 1970] Kreiss, H.-O. (1970). Initial boundary value problems for hyperbolic systems. *Comm. Pure Appl. Meth.*, 23:277–298.

- [Kreiss and Oliger, 1972] Kreiss, H.-O. and Oliger, J. (1972). Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, 24:199–215.
- [Kreiss and Oliger, 1979] Kreiss, H.-O. and Oliger, J. (1979). Stability of the Fourier method. *SIAM J. Num. Anal.*, 16:421–433.
- [Kreiss et al., 2004] Kreiss, H.-O., Petersson, A., and Yström, J. (2004). Difference approximations of the Neumann problem for the second order wave equation. *SIAM J. Numer. Anal.*, 42:1292–1323.
- [Kreiss and Scherer, 1974] Kreiss, H.-O. and Scherer, G. (1974). Finite element and finite difference methods for hyperbolic partial differential equations. *Mathematical aspects of finite elements in partial differential equations*, Academic Press, Orlando, FL.
- [Kreiss and Scherer, 1977] Kreiss, H.-O. and Scherer, G. (1977). On the existence of energy estimates for difference approximations for hyperbolic systems. *Technical report, Uppsala University, Dept of Scientific Computing, Uppsala, Sweden*.
- [Kreiss and Wu, 1993] Kreiss, H.-O. and Wu, L. (1993). On the stability definition of difference approximations for the initial boundary value problem. *Appl. Num. Math.*, 12:213–227.
- [Kress and Gustafsson, 2002] Kress, W. and Gustafsson, B. (2002). Deferred correction methods for initial boundary value problems. *J. Sci. Comput.*, 17:241–251.
- [Kress and Lötstedt, 2006] Kress, W. and Lötstedt, P. (2006). Time step restrictions using semi-explicit methods for the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 195:4433–4447.
- [Kutta, 1901] Kutta, W. (1901). Beitrag zur näherungsweisen integration totaler differentialgleichungen. *Zeitschr. für Math. u. Phys.*, 46:435–453.
- [Lax, 1957] Lax, P. (1957). Hyperbolic systems of conservation laws, II. *Comm. Pure Appl. Math.*, 10:537–566.
- [Lax, 1972] Lax, P. (1972). *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, II. SIAM Regional Conference Series in Applied Mathematics, No 11.
- [Lax and Richmyer, 1956] Lax, P. and Richmyer, R. (1956). Survey of the stability of linear finite difference equations. *Comm. Pure Appl. Math.*, 9:267–.
- [Lax and Wendroff, 1960] Lax, P. and Wendroff, B. (1960). Systems of conservation laws. *Comm. Pure Appl. Math.*, 13:217–237.
- [Lax and Wendroff, 1964] Lax, P. and Wendroff, B. (1964). Difference schemes for hyperbolic equations with high order accuracy. *Comm. Pure Appl. Math.*, 17:381–398.
- [Lee and Fornberg, 2004] Lee, J. and Fornberg, B. (2004). Some unconditionally stable time stepping methods for the 3D Maxwell's equations. *J. Comput. Appl. Math.*, 166:497–523.
- [Lele, 1992] Lele, S. (1992). Compact finite difference schemes with spectral-like resolution. *J. Comp. Phys.*, 103:16–42.
- [LeVeque, 1992] LeVeque, R. (1992). *Numerical methods for conservation laws*. Birkhäuser Verlag, Basel.
- [Levy and Tadmor, 1998] Levy, D. and Tadmor, E. (1998). From semidiscrete to fully discrete: stability of Runge-Kutta schemes by the energy method. *SIAM Review*, 40:40–73.
- [Liu and Osher, 1998] Liu, X.-D. and Osher, S. (1998). Convex eno high order multi-dimensional schemes without field by field decomposition or staggered grids. *J. Comp. Phys.*, 142:304–330.
- [Liu et al., 1994] Liu, X.-D., Osher, S., and Chan, T. (1994). Weighted essentially non-oscillatory schemes. *J. Comp. Phys.*, 115:200–212.
- [MacCormack, 1969] MacCormack, R. (1969). The effect of viscosity in hypervelocity impact cratering. *AIAA Paper No. 69-354*.
- [MacCormack and Pauly, 1972] MacCormack, R. and Pauly, A. (1972). Computational efficiency achieved by time splitting of finite difference operators. *AIAA Paper No. 72-154*.
- [Mancera and Hunt, 1997] Mancera, P. and Hunt, R. (1997). Fourth order method for solving the Navier–Stokes equations. *Int. J. Numer. Math. Fluids*, 25:1119–1135.
- [Mattsson, 2003] Mattsson, K. (2003). Boundary procedure for summation-by-parts operators. *J. Scient. Comp.*, 18:133–153.
- [Mattsson and Nordström, 2004] Mattsson, K. and Nordström, J. (2004). Summation by parts operators for finite difference approximations of second derivatives. *J. Comp. Phys.*, 199:503–540.

- [Mattsson and Nordström, 2006] Mattsson, K. and Nordström, J. (2006). High order finite difference methods for wave propagation in discontinuous media. *J. Comp. Phys.*, 220:249–269.
- [Michelson, 1983] Michelson, D. (1983). Stability theory of difference approximations for multidimensional initial-boundary value problems. *Math. Comp.*, 40:1–46.
- [Milne, 1926] Milne, W. E. (1926). Numerical integration of ordinary differential equations. *Amer. Math. Monthly*, 33:455–460.
- [Nehrbass et al., 1998] Nehrbass, J. W., Jevtić, J. O., and Lee, R. (1998). Reducing the phase error for finite-difference methods without increasing the order. *IEEE Trans. Antennas and Propagation*, 46:1194–1201.
- [Nilsson et al., 2003] Nilsson, J., Gustafsson, B., Lötstedt, P., and Brüger, A. (2003). High order difference method on staggered, curvilinear grids for the incompressible Navier–Stokes equations. *Proceedings of Second MIT Conference on Computational Fluid and Solid Mechanics 2003*. Editor: Bathe K.J. Elsevier, Amsterdam, pages 1057–1061.
- [Nordström and Carpenter, 1999] Nordström, J. and Carpenter, M. (1999). Boundary and interface conditions for high order finite difference methods applied to the Euler and Navier–Stokes equations. *J. Comp. Phys.*, 148:621–645.
- [Olsson, 1995a] Olsson, P. (1995a). Summation by parts, projections, and stability: I. *Math. Comp.*, 64:1035–1065.
- [Olsson, 1995b] Olsson, P. (1995b). Summation by parts, projections, and stability. II. *Math. Comp.*, 64:1473–1493.
- [Osher, 1969] Osher, S. (1969). Stability of difference approximations of dissipative type for mixed initial boundary value problems. I. *Math. Comp.*, 23:335–340.
- [Reddy and Trefethen, 1990] Reddy, S. C. and Trefethen, L. N. (1990). Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues. *Comp. Meth. Appl. Mech. Eng.*, 80:147–164.
- [Reddy and Trefethen, 1992] Reddy, S. C. and Trefethen, L. N. (1992). Stability of the method of lines. *Numer. Math.*, 62:235–267.
- [Richardson, 1922] Richardson, L. (1922). *Weather predictions by numerical process*. Cambridge University Press.
- [Richtmyer and Morton, 1967] Richtmyer, R. and Morton, K. (1967). *Difference methods for initial-value problems*. Interscience Publishers.
- [Roe, 1985] Roe, P. (1985). Some contributions to the modeling of discontinuous flows. *Lect. Notes Appl. Math.*, 22:163–193.
- [Runge, 1895] Runge, C. (1895). Ueber die numerische auflösung von differentialgleichungen. *Math. Ann.*, 46:167–178.
- [Rylander and Bondesson, 2002] Rylander, T. and Bondesson, A. (2002). Stability of explicit-implicit hybrid time-stepping schemes for Maxwell’s equations. *J. Comp. Phys.*, 179:426–438.
- [Shu and Osher, 1988] Shu, C.-W. and Osher, S. (1988). Efficient implementation of essentially non-oscillatory shock-capturing schemes. *J. Comp. Phys.*, 77:439–471.
- [Strand, 1994] Strand, B. (1994). Summation by parts for finite difference approximations for d/dx . *J. Comp. Phys.*, 110:47–67.
- [Strang, 1987] Strang, W. G. (1987). Accurate partial difference methods II: non-linear problems. *SIAM J. Sci. Statist. Comput.*, 8:135–151.
- [Strikwerda, 1980] Strikwerda, J. (1980). Initial boundary value problems for the method of lines. *J. Comp. Phys.*, 34:94–107.
- [Strikwerda, 1989] Strikwerda, J. (1989). *Finite difference schemes and partial differential equations*. Wadsworth & Brooks/Cole.
- [Svärd and Nordström, 2006] Svärd, M. and Nordström, J. (2006). On the order of accuracy for difference approximations of initial-boundary value problems. *J. Comp. Phys.*, 218:333–352.
- [Swartz and Wendroff, 1974] Swartz, B. and Wendroff, B. (1974). The relative efficiency of finite difference and finite element methods I: Hyperbolic problems and splines. *SIAM J. Num. Anal.*, 11:979–993.
- [Tadmor, 1986] Tadmor, E. (1986). The exponential accuracy of Fourier and Chebyshev differencing methods. *SIAM, J. Num. Anal.*, 23:1–10.

- [Tadmor, 2002] Tadmor, E. (2002). From semidiscrete to fully discrete: stability of runge-kutta schemes by the energy method. II. *SIAM, Proceedings in Applied Mathematics*, 109:25–49.
- [Taftove, 1998] Taftove, A. (1998). *Advances in Computational Electrodynamics*. Artec House.
- [Tam and Webb, 1993] Tam, C. K. W. and Webb, J. C. (1993). Dispersion-relation-preserving finite difference schemes for computational acoustics. *J. Comp. Phys.*, 107:262–281.
- [Thomée and Wendroff, 1974] Thomée, V. and Wendroff, B. (1974). Converegence estimates for Galerkin methods for variable coefficient initial value problems. *SIAM J. Numer.Anal.*, 11:1059–1068.
- [Tornberg and Engquist, 2003] Tornberg, A.-K. and Engquist, B. (2003). Regularization techniques for numerical approximation of PDEs with singularities. *J. Sci. Comput.*, 18:527–552.
- [Tornberg and Engquist, 2004] Tornberg, A.-K. and Engquist, B. (2004). Numerical approximations of singular source terms in differential equations. *J. Comp. Phys.*, 200:462–488.
- [Tornberg and Engquist, 2006] Tornberg, A.-K. and Engquist, B. (2006). High order difference methods for wave propagation in discontinuous media. *To appear in BIT*.
- [Tornberg et al., 2006] Tornberg, A.-K., Engquist, B., B.Gustafsson, and Wahlund, P. (2006). A new type of boundary treatment for wave propagation. *BIT Num. Math.*, 46, Suppl. 5:145–170.
- [Trefethen, 1983] Trefethen, L. (1983). Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundström. *J. Comp. Phys.*, 49:199–217.
- [Turkel and Yefet, 2000] Turkel, E. and Yefet, A. (2000). On the construction of a high order difference scheme for complex domains in a cartesian grid. *Appl. Numer. Anal.*, 33:113–124.
- [von Neumann, 1944] von Neumann, J. (1944). Proposal and analysis of a numerical method for the treatment of hydro-dynamical shock problems. *Nat. Def. Res. Com., Report AM-551*.
- [Wendroff, 1960] Wendroff, B. (1960). On centered difference equations for hyperbolic systems. *J. Soc. Ind. Appl. Math.*, 8:549–555.
- [Yee, 1966] Yee, K. (1966). Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas and Prop.*, 14:302–307.
- [Yefet and Turkel, 2000] Yefet, A. and Turkel, E. (2000). Fourth order compact implicit method for the Maxwell equations with discontinuous coefficients. *Appl. Numer. Anal.*, 33:125–134.
- [Zheng et al., 1999] Zheng, F., Cheng, Z., and Zhang, J. (1999). A finite-difference time-domain method without the Courant stability conditions. *IEEE Microwave Guided Wave Lett.*, 9:441–443.
- [Zheng et al., 2000] Zheng, F., Cheng, Z., and Zhang, J. (2000). Toward the development of a three-dimensional unconditionally stable finite-difference time-domain method. *IEEE Transactions on Microwave Theory and Techniques*, 48:1550–1558.
- [Zingg, 2000] Zingg, D. W. (2000). Comparison of high-accuracy finite-difference methods for linear wave propagation. *SIAM J. Sci. Comput.*, 22:476–502.
- [Zingg et al., 1993] Zingg, D. W., Lomax, H., and Jurgens, H. (1993). An optimized finite-difference scheme for wave propagation problems. *AIAA paper*, 93(0459).
- [Zingg et al., 1996] Zingg, D. W., Lomax, H., and Jurgens, H. (1996). High-accuracy finite-difference schemes for linear wave propagation. *SIAM J. Sci. Comput.*, 17:328–346.

Index

- acoustics, 178, 194, 199
- Adams methods, 104, 105
- Adams–Bashforth method, 104
- Adams–Moulton method, 105
- advection diffusion equation, 235
- amplification matrix, 25
- artificial viscosity, 252–257
- average operator, 157
- BDF methods, 105, 226
- boundary
 - curved, 279
 - embedded, 210
 - function, 13
 - layer, 229
 - operator, 13
- boundary condition
 - downstream, 229
 - essential, 268
 - high order, 209
 - natural, 272
 - outflow, 229
- box scheme, 157–176
 - original, 157–161
 - shifted, 161–164
- Burgers’ equation, 247, 257, 258
 - viscous, 247
- Cartesian coordinates, 228
- Cauchy problem, 16, 51
- CFL-number, 97, 168, 173
 - local, 171
- characteristic equation, 24, 42, 60, 308, 310
- characteristic speed, 250
- Chebyshev polynomials, 295
- collocation methods, 289
- condition number, 18, 23, 24, 40, 96
- conservation laws, 246–251
- conservative form, 259
- consistent, 70
- constant coefficients, 19, 26
- contact discontinuity, 251
- convergence, 71
- convergence rate, 69–80, 91
- convergent method, 71
- cosine transformation, 297
- Courant number, 97
- Crank–Nicholson scheme, 39, 72, 118, 121
- deferred correction, 108–111
- determinant condition, 50, 62
- diagonal norm, 133, 138, 311
- difference equations, 307–310
- difference operator, 2
 - backward, 2
 - centered, 2
 - forward, 2
 - optimized, 91
- difference scheme, 38
- differential inequality, 29, 30
- differential operator, 13, 19, 30, 38
 - linear, 17
- discontinuous coefficients, 192–209
- discontinuous Galerkin method, 283–289
- discontinuous solution, 206
- discrete Fourier transform, 291
- dispersion error, 4
- dissipative, 26
- distinct roots, 308
- divergence condition, 165, 237
- domain
 - convex, 279
 - nonrectangular, 229

- eigenvalue problem, 42, 49, 60
- electromagnetics, 178
- energy conserving, 165
- energy method, 29–41
- Engquist–Osher scheme, 259
- ENO schemes, 261–263
- Enquist–Osher flux, 262
- entropy condition, 250
- equation of state, 249
- error equation, 70, 74
- essentially nonoscillating schemes, 261
- Euler equations, 249, 264
- Euler method
 - backward, 40, 102, 105
 - forward, 71, 98
- field by field decomposition, 265
- finite element methods, 267–283
- finite volume methods, 265, 300–305
- flux function, 258
- flux limiter, 261
- forcing function, 13
- Fourier analysis, 16–28
- Fourier methods, 290–295
- Fourier series, 19
- Fourier transform, 3, 9, 19, 23, 25, 26
 - discrete, 21, 96
 - fast, 21
- frequency domain methods, 218
- full norm, 317
- Galerkin FEM, 267–281
- Galerkin formulation, 268
- Galerkin method, 268
- Gauss–Lobatto points, 296
- generalized solution, 31
- generalized eigenvalue, 49, 50, 55, 61–63
- generalized eigenvector, 44
- Gibbs’ phenomenon, 295
- GKS theory, 68
- global formulation, 230
- Godunov method, 259
- Godunov–Ryabenkii condition, 44, 46, 60, 61, 68
- grid
 - function, 2
 - nonuniform, 169–176
 - staggered, 85–87, 179
 - standard type, 81–85
 - tensor product, 176
 - uniform, 1
- hat function, 269
- heat equation, 8, 15
- backward, 16
- Helmholtz equation, 218
- Hermite polynomials, 277
- Hermitian matrix, 132
- Heun’s method, 98
- high resolution methods, 258
- hybrid method, 261
- hyperbolic system, 20, 99, 178
- ill posed, 16
- incompressible flow, 165, 220, 228
- initial function, 13
- initial-boundary value problem, 13, 29–66, 72–79
- injection method, 140
- iterative solution, 237
- Jacobian, 250
- Jordan box, 309
- Jordan canonical form, 308
- Laplace equation, 223
- Laplace operator, 220
- Laplace transform, 46–48, 59, 60
- Laplacian, 230
- Lax entropy condition, 250
- Lax equivalence theorem, 72
- Lax–Friedrichs scheme, 253
- Lax–Wendroff principle, 116, 181
- Lax–Wendroff scheme, 116, 124, 253
- leap-frog scheme, 24, 63
- least square approximation, 275
- least square problem, 92
- Legendre polynomials, 299
- linear multistep methods, 102–108
- Lipschitz continuous, 138
- local formulation, 230
- MacCormack scheme, 124, 253
- maximally semibounded, 31
- Maxwell equations, 178
- method of lines, 33, 40, 81
- Milne method, 106
- modified coefficients, 201, 214–216
- monotone, 254
- monotonicity preserving, 254
- multistep scheme, 25
- Navier–Stokes equations, 228–231, 249
 - linearized, 233
- Navier–Stokes solvers, 235
- nodes, 269
- nonlinear methods, 258
- nonlinear problems, 245–266

- normal mode analysis, 41–66
- one sided formulas, 83
- one step scheme, 23–25, 181
- order of accuracy, 69–80
- orthogonal transformation, 229
- outflow problem, 130
- Padé approximation, 232
- Padé difference operator, 87–91, 118, 120, 224, 232, 323
- parabolic equations, 8
- parasitic solution, 236
- Parseval's relation, 15, 46
- discrete, 22
- particle velocity, 178
- periodic problems, 15–28, 69–72
- Petrov–Galerkin FEM, 281–283
- Petrovski condition, 19
- piecewise bilinear, 278
- piecewise linear, 269, 279
- points per wavelength, 4
- polynomial methods, 295–300
- predictor, 108
- predictor–corrector, 108
- pressure, 178
- primitive form, 247, 249
- projection matrix, 141
- projection methods, 140–146
- projection operator, 275
- pseudospectral methods, 289
- quarter space problem, 42
- Rankine–Hugoniot condition, 249, 250
- restricted norm, 133
- Reynolds number, 220, 231, 249
- Richardson extrapolation, 94, 111–113
- Riemann problem, 255, 259
- Roe flux, 262
- roof function, 269
- Runge–Kutta methods, 97–102, 116, 264
- explicit, 101
- implicit, 101
- SAT method, 147–155, 298
- SBP operators, 84, 130–156, 311–323
- scalar transport, 165
- Schrödinger equation, 11
- semibounded, 30, 31, 35–39, 51, 53, 54
- semidiscrete approximation, 28, 33–38, 40–59, 81
- shift operator, 2
- shock capturing, 252, 258
- shock fitting, 251
- shock profiles, 255
- shocks, 245–266
- Shu–Osher problem, 264
- similarity transformation, 17
- skewsymmetric, 38, 131
- slope limiter, 260
- Sobolev inequality, 153
- space variables, 13
- spectral element method, 298
- spectral methods, 289–300
- stability, 21–28, 33, 66, 95
- stability domain, 96, 98
- stable, 23, 37, 39
- A-stable, 96, 102, 105
- boundary stable, 50
- generalized sense, 52, 65
- strongly, 37, 39
- strongly, generalized sense, 52, 66
- time stable, 37
- zero-stable, 104
- steady state system, 236
- step size, 1
- Stokes equations, 220–223
- streamline–diffusion method, 282
- strictly semibounded, 131, 132, 135
- strong stability-preserving methods, 265
- Sturm–Liouville problem, 295
- summation by parts, 36
- superconvergence, 281
- symbol, 19, 25
- Taylor expansion, 2, 115
- test equation, 95
- test function, 248, 268, 283
- total variation, 260
- total variation diminishing, 260
- trapezoidal rule, 39, 102, 117
- triangular elements, 278
- tridiagonal systems, 88
- truncation error, 2, 70
- local, 70
- turbulent flow, 219
- TVD method, 260
- TVD time discretizations, 265
- unitary transformation, 96
- upwind methods, 257–261
- variable coefficients, 85
- variable coefficients, 26, 167, 293
- von Neumann condition, 25, 26, 28, 233
- for semidiscrete scheme, 28

- wave equation, 177–179
- wave number, 2
- wave propagation, 2, 177–218
- wave speed, 4, 250
- weak form, 268
- weak solution, 248
- well posed, 14, 18, 32
- problem, 13–16
- strongly, 32
- WENO schemes, 263–264
- Yee scheme, 179–181, 187, 202, 206, 209, 211
- z-transform, 60