

TIME
DEPENDENT
PROBLEMS
— AND —
DIFFERENCE
METHODS

BERTIL GUSTAFSSON

HEINZ-OTTO KREISS

JOSEPH OLIGER



PURE AND APPLIED MATHEMATICS:
A Wiley-Interscience Series of Texts,
Monographs, and Tracts

Time dependent problems frequently pose challenges in areas of science and engineering dealing with numerical analysis, scientific computation, mathematical models, and most importantly—numerical experiments intended to analyze physical behavior and test design. *Time Dependent Problems and Difference Methods* addresses these various industrial considerations in a pragmatic and detailed manner, giving special attention to time dependent problems in its coverage of the derivation and analysis of numerical methods for computational approximations to Partial Differential Equations (PDEs).

The book is written in two parts. Part I discusses problems with periodic solutions; Part II proceeds to discuss initial boundary value problems for partial differential equations and numerical methods for them. The problems with periodic solutions have been chosen because they allow the application of Fourier analysis without the complication that arises from the infinite domain for the corresponding Cauchy problem. Furthermore, the analysis of periodic problems provides necessary conditions when constructing methods for initial boundary value problems. Much of the material included in Part II appears for the first time in this book.

The authors draw on their own interests and combined extensive experience in applied mathematics and computer science to bring about this practical and useful guide. They provide complete discussions of the pertinent theorems and back them up with examples and illustrations.

(continued from front flap)

For physical scientists, engineers, or anyone who uses numerical experiments to test designs or to predict and investigate physical phenomena, this invaluable guide is destined to become a constant companion. *Time Dependent Problems and Difference Methods* is also extremely useful to numerical analysts, mathematical modelers, and graduate students of applied mathematics and scientific computations.

About the authors

BERTIL GUSTAFSSON is a professor with the Department of Scientific Computing at Uppsala University, Sweden.

HEINZ-OTTO KREISS is a professor with the UCLA Department of Mathematics. He is the coauthor of *Initial-Boundary Value Problems and the Navier-Stokes Equations*.

JOSEPH OLIGER is a professor with the Department of Computer Science at Stanford University.

Cover Design: David Levy

(continued on back flap)

What Every Physical Scientist and Engineer Needs to Know About Time Dependent Problems . . .

Time Dependent Problems and Difference Methods covers the analysis of numerical methods for computing approximate solutions to partial differential equations for time dependent problems. This original book includes for the first time a concrete discussion of initial boundary value problems for partial differential equations. The authors have redone many of these results especially for this volume, including theorems, examples, and over one hundred illustrations.

The book takes some less-than-obvious approaches to developing its material:

- Treats differential equations and numerical methods with a parallel development, thus achieving a more useful analysis of numerical methods
- Covers hyperbolic equations in particularly great detail
- Emphasizes error bounds and estimates, as well as the sufficient results needed to justify the methods used for applications

Time Dependent Problems and Difference Methods is written for physical scientists and engineers who use numerical experiments to test designs or to predict and investigate physical phenomena. It is also extremely useful to numerical analysts, mathematical modelers, and graduate students of applied mathematics and scientific computations.

WILEY-INTERSCIENCE

John Wiley & Sons, Inc.

Professional, Reference and Trade Group

605 Third Avenue, New York, N.Y. 10158-0012

New York • Chichester • Brisbane • Toronto • Singapore

ISBN 0-471-50734-2

90000



9 780471 507345

PURE AND APPLIED MATHEMATICS
A WILEY-INTERSCIENCE SERIES OF
TEXTS, MONOGRAPHS & TRACTS

**BERTIL GUSTAFSSON
HEINZ-OTTO KREISS
JOSEPH OLIGER**

**TIME DEPENDENT PROBLEMS
AND DIFFERENCE METHODS**

PURE AND APPLIED MATHEMATICS

A Wiley-Interscience Series of Texts, Monographs, and Tracts

Founded by RICHARD COURANT

Editor Emeritus: PETER HILTON

Editors: MYRON B. ALLEN III, DAVID A. COX,
HARRY HOCHSTADT, PETER LAX, JOHN TOLAND

A complete list of the titles in this series appears at the end of this volume.

TIME DEPENDENT PROBLEMS AND DIFFERENCE METHODS

BERTIL GUSTAFSSON

*Department of Scientific Computing
Uppsala University, Sweden*

HEINZ-OTTO KREISS

*Department of Mathematics
University of California at Los Angeles*

JOSEPH OLIGER

*Department of Computer Science
Stanford University*



A Wiley-Interscience Publication

JOHN WILEY & SONS, INC.

New York • Chichester • Brisbane • Toronto • Singapore

This text is printed on acid-free paper.

Copyright © 1995 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission of further information should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012.

Library of Congress Cataloging in Publication Data:

Gustafsson, Bertil, 1939-

Time dependent problems and difference methods / Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.

p. cm. — (Pure and applied mathematics)

Includes bibliographical references and index.

ISBN 0-471-50734-2 (acid-free)

1. Differential equations, Partial—Numerical solutions.

I. Kreiss, Heinz-Otto II. Oliger, Joseph, 1941- . III. Title.

IV. Series: Pure and applied mathematics (John Wiley & Sons: Unnumbered)

QA374.0974 1995

51 .353—dc20

94-44176

CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface	ix
Acknowledgments	xiii
PART I Problems with Periodic Solutions	1
1. Fourier Series and Trigonometric Interpolation	3
1.1 Some results from the theory of Fourier series,	3
1.2 Periodic gridfunctions and difference operators,	17
1.3 Trigonometric interpolation,	24
1.4 The S operator for differentiation,	30
1.5 Generalizations,	33
Bibliographic Notes,	37
2. Model Equations	38
2.1 First-order wave equation, convergence, and stability,	38
2.2 The Leap-frog scheme,	50
2.3 Implicit methods,	55
2.4 Truncation error,	59
2.5 Heat equation,	61
2.6 Convection-diffusion equation,	70
2.7 Higher order equations,	74
2.8 Generalization to several space dimensions,	76
Bibliographic Notes,	78
3. Higher Order Accuracy	80
3.1 Efficiency of higher order accurate difference approximations,	80
3.2 Fourier method,	95
Bibliographic Notes,	104
4. Well-Posed Problems	106
4.1 Well-posedness,	106

4.2	Scalar differential equations with constant coefficients in one space dimension, 113	
4.3	First-order systems with constant coefficients in one space dimension, 116	
4.4	Parabolic systems with constant coefficients in one space dimension, 122	
4.5	General systems with constant coefficients, 127	
4.6	Semibounded operators with variable coefficients, 134	
4.7	The solution operator and Duhamel's principle, 142	
4.8	Generalized solutions, 149	
4.9	Well-posedness of nonlinear problems, 152	
	Bibliographic Notes, 156	
5.	Stability and Convergence for Numerical Approximations of Linear and Nonlinear Problems	157
5.1	Stability and convergence, 157	
5.2	Stability for approximations with constant coefficients, 171	
5.3	Approximations with variable coefficients: The energy method, 182	
5.4	Splitting methods, 195	
5.5	Stability for nonlinear problems, 201	
	Bibliographic Notes, 210	
6.	Hyperbolic Equations and Numerical Methods	211
6.1	Systems with constant coefficients in one space dimension, 211	
6.2	Systems with variable coefficients in one space dimension, 214	
6.3	Systems with constant coefficients in several space dimensions, 218	
6.4	Systems with variable coefficients in several space dimensions, 221	
6.5	Approximations with constant coefficients, 222	
6.6	Approximations with variable coefficients, 235	
6.7	The method of lines, 238	
6.8	The finite-volume method, 254	
6.9	The Fourier method, 262	
	Bibliographic Notes, 267	
7.	Parabolic Equations and Numerical Methods	270
7.1	General parabolic systems, 270	
7.2	Stability for difference and Fourier methods, 275	

7.3	Difference approximations in several space dimensions, 282 Bibliographic Notes, 289	
8.	Problems with Discontinuous Solutions	290
8.1	Difference methods for linear hyperbolic equations, 290	
8.2	Method of characteristics, 297	
8.3	Method of characteristics in several space dimensions, 304	
8.4	Method of characteristics on a regular grid, 305	
8.5	Regularization using viscosity, 313	
8.6	The inviscid Burgers' equations, 316	
8.7	The viscous Burgers' equation and traveling waves, 320	
8.8	Numerical methods for scalar equations based on regularization, 328	
8.9	Regularization for systems of equations, 336	
8.10	High-resolution methods, 345 Bibliographic Notes, 355	
PART II	Initial-Boundary-Value Problems	357
9.	The Energy Method for Initial-Boundary-Value Problems	359
9.1	Characteristics and boundary conditions for hyperbolic systems in one space dimension, 359	
9.2	Energy estimates for hyperbolic systems in one space dimension, 368	
9.3	Energy estimates for parabolic differential equations in one space dimension, 375	
9.4	Well-posed problems, 381	
9.5	Semibounded operators, 385	
9.6	Quarter-space problems in more than one space dimension, 390 Bibliographic Notes, 397	
10.	The Laplace Transform Method for Initial-Boundary-Value Problems	398
10.1	Solution of hyperbolic systems, 398	
10.2	Solution of parabolic problems, 404	
10.3	Generalized well-posedness, 410	
10.4	Systems with constant coefficients in one space dimension, 419	
10.5	Hyperbolic systems with constant coefficients in several space dimensions, 427	

10.6	Parabolic systems in more than one space dimension,	440
10.7	Systems with variable coefficients in general domains,	443
	Bibliographic Notes,	444
11.	The Energy Method for Difference Approximations	445
11.1	Hyperbolic problems,	445
11.2	Parabolic differential equations,	458
11.3	Stability, consistency, and order of accuracy,	465
11.4	Higher order approximations,	471
11.5	Several space dimensions,	484
	Bibliographic Notes,	491
	Appendix 11,	492
12.	The Laplace Transform Method for Difference Approximations	496
12.1	Necessary conditions for stability,	496
12.2	Sufficient conditions for stability,	506
12.3	A fourth-order accurate approximation for hyperbolic differential equations,	524
12.4	Stability in the generalized sense for hyperbolic systems,	526
12.5	An example that does not satisfy the Kreiss condition but is stable in the generalized sense,	538
12.6	Parabolic equations,	552
12.7	The convergence rate,	567
	Bibliographic Notes,	575
13.	The Laplace Transform Method for Fully Discrete Approximations: Normal Mode Analysis	576
13.1	General theory for approximations of hyperbolic systems,	576
13.2	The method of lines and generalized stability,	599
13.3	Several space dimensions,	611
13.4	Domains with irregular boundaries and overlapping grids,	615
	Bibliographic Notes,	620
	Appendix A.1 Results from Linear Algebra,	622
	Appendix A.2 Laplace Transform,	625
	Appendix A.3 Iterative Methods,	629
	References,	633
Index		639

PREFACE

In this preface, we discuss the material to be covered, the point of view we take, and our emphases. Our primary goal is to discuss material relevant to the derivation and analysis of numerical methods for computing approximate solutions to partial differential equations for time-dependent problems arising in the sciences and engineering. It is our intention that this book should be useful for graduate students interested in applied mathematics and scientific computation as well as physical scientists and engineers whose primary interests are in carrying out numerical experiments to investigate physical behavior and test designs.

We carry out a parallel development of material for differential equations and numerical methods. Our motivation for this approach is twofold: the usual treatment of partial differential equations does not follow the lines that are most useful for the analysis of numerical methods, and the derivation of numerical methods is increasingly utilizing and benefiting from following the detailed development for the differential equations.

Most of our development and analysis is for linear equations, whereas most of the calculations done in practice are for nonlinear problems. However, this is not so fruitless as it may sound. If the nonlinear problem of interest has a smooth solution, then it can be linearized about this solution and the solution of the nonlinear problem will be a solution of the linearized problem with a perturbed forcing function. Errors of numerical approximations for the nonlinear problem can thus be estimated locally, and justified in terms of the linearized equations. A problem often arises in this scenario; the mathematical properties required to guarantee that the solution is smooth *a priori* may not be known or verifiable. So we often perform calculations whose results we cannot justify *a priori*. In this situation, we can proceed rationally, if not rigorously, by using a method that we could justify for the corresponding linearized problems and can be justified *a posteriori*, at least in principle, if the obtained solution satisfies certain smoothness properties. The smoothness properties of our computed solutions can be observed to experimentally verify the needed smoothness requirements and justify our computed results *a posteriori*. However, this procedure is not without its limitations. There are many problems that do not have smooth solutions. There are genuinely nonlinear phenomena, such as

shocks, rarefaction waves, and nonlinear instability, that we must study in a nonlinear framework, and we discuss such issues separately. There are a few general results for nonlinear problems that generally are justifications of the linearization procedure mentioned above and that we include when available.

The material covered in this book emphasizes our own interests and work. In particular, our development of hyperbolic equations is more complete and detailed than our development of parabolic equations and equations of other types. Similarly, we emphasize the construction and analysis of finite difference methods, although we do discuss Fourier methods. We devote a considerable portion of this book to initial boundary value problems and numerical methods for them. This is the first book to contain much of this material and quite a lot of it has been redone for this presentation. We also tend to emphasize the sufficient results needed to justify methods used in applications rather than necessary results, and to stress error bounds and estimates which are valid for finite values of the discretization parameters rather than statements about limits.

We have organized this book in two parts: Part I discusses problems with periodic solutions and Part II discusses initial-boundary-value problems. It is simpler and more clear to develop the general concepts and to analyze problems and methods for the periodic boundary problems where the boundaries can essentially be ignored and Fourier series or trigonometric interpolants can be used. This same development is often carried out elsewhere for the Cauchy, or pure initial-value, problem. These two treatments are dual to each other, one relying upon Fourier series and the other upon Fourier integrals. We have chosen periodic boundary problems, because we are, in this context, dealing with a finite, computable method without any complications arising from the infinite domains of the corresponding Cauchy problems. Periodic boundary problems do arise naturally in many physical situations such as flows in toroids or on the surface of spheres; for example, the separation of periodic boundary and initial-boundary-value problems is also natural, because the results for initial-boundary-value problems often take the following form: If the problem or method is good for the periodic boundary problem and if some additional conditions are satisfied, then the problem or method is good for a corresponding initial-boundary-value problem. So an analysis and understanding of the corresponding periodic boundary problem is often a necessary condition for results for more general problems.

In Part I, we begin with a discussion in **Chapter 1 of Fourier series and trigonometric interpolation**, which is central to this part of the book. In Chapter 2, we discuss model equations for **convection and diffusion**. Throughout the book, we often rely upon a model equation approach to our material. Equations typifying various phenomena, such as convection, diffusion, and dispersion, that distinguish the difficulties inherent in approximating equations of different types are central to our analysis and development. Difference methods are first introduced in this chapter and discussed in terms of the model equations. In Chapter 3, we consider the efficiencies of using **higher order accurate methods**, which, in a natural limit, lead to the **Fourier or pseudospectral method**. The

concept of a well-posed problem is introduced in Chapter 4 for general linear and nonlinear problems for partial differential equations. The general stability and convergence theory for difference methods is presented in Chapter 5. Sections are devoted to the tools and techniques needed to establish stability for methods for linear problems with constant coefficients and then for those with variable coefficients. Splitting methods are introduced, and their analysis is carried out. These methods are very useful for problems in several space dimensions and to take advantage of special solution techniques for particular operators. The chapter closes with a discussion of stability for nonlinear problems. Chapters 6 and 7 are devoted to specific results and methods for hyperbolic and parabolic equations, respectively. Nonlinear problems with discontinuous solutions, in particular, hyperbolic conservation laws with shocks and numerical methods for them are discussed in Chapter 8, which concludes Part I of the book and our basic treatment of partial differential equations and methods in the periodic boundary setting.

Part II is devoted to the discussion of the initial boundary value problem for partial differential equations and numerical methods for these problems. Chapter 9 discusses the energy method for initial-boundary-value problem for hyperbolic and parabolic equations. Chapter 10 discusses Laplace transform techniques for these problems. Chapter 11 treats stability for difference approximations using the energy method and follows the treatment of the differential equations in Chapter 9. Chapter 12 follows from Chapter 10 in terms of development—here the Laplace transform is used for difference approximations. This treatment is carried out for the semidiscretized problem: Only the spacial part of the operator is discretized. Finally, the fully discretized problem is treated in Chapter 13 using the Laplace transform. The so-called “normal mode analysis” technique is used and developed in these last two chapters. In particular, sufficient stability conditions for the fully discretized problem are obtained in terms of stability results for the semidiscretized problem, which are much easier to obtain.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance we have gotten from our students and colleagues who have worked through various versions of this material and have supplied us with questions and suggestions that have been very helpful. We want to give special thanks to Barbro Kreiss and Mary Washburn who have expertly handled the preparation of our manuscript and have carried out most of the computations we have included. Their patience and good humor through our many versions and revisions is much appreciated. Pelle Olsson has gone over our manuscript carefully with us and a number of sections have been improved by his suggestions.

Finally, we acknowledge the Office of Naval Research and the National Aeronautics and Space Administration for their support of our work.

I

**PROBLEMS WITH
PERIODIC SOLUTIONS**

1

FOURIER SERIES AND TRIGONOMETRIC INTERPOLATION

Expansions of functions in Fourier series are particularly useful for both the analysis and construction of numerical methods for partial differential equations. In this chapter, we present the main results of this theory which are used as the basis for most of the analysis in Part I of this book.

1.1 SOME RESULTS FROM THE THEORY OF FOURIER SERIES

To begin, we consider the representation of continuous complex valued functions by Fourier series. We assume throughout this chapter that functions are 2π -periodic and defined for all real numbers. If a function is only defined on a bounded interval, then we can make it 2π -periodic by means of a change of scale and extend it periodically. However, the number of derivatives that the extended function has will crucially impact the results. The basic result is shown by Theorem 1.1.1.

Theorem 1.1.1. *Let $f(x) \in C^1(-\infty, \infty)^*$ be 2π -periodic. Then $f(x)$ has a Fourier series representation*

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \quad (1.1.1)$$

* $C^n(-\infty, \infty)$ [or $C^n(a, b)$] denotes the class of n times continuously differentiable functions for $-\infty < x < \infty$ [or $-\infty < a \leq x \leq b < \infty$].

where the Fourier coefficients $\hat{f}(\omega)$ are given by

$$\boxed{\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-i\omega x} f(x) dx.} \quad (1.1.2)$$

The series converges uniformly to $f(x)$.

One can weaken the assumptions.

Theorem 1.1.2. Assume that $f(x)$ is 2π -periodic and piecewise in C^1 . If $f(x) \in C^1(a, b)$ in some interval $a < x < b$, then the Fourier series Eq. (1.1.1) converges uniformly to $f(x)$ in any subinterval $a < \alpha \leq x \leq \beta < b$. At a point of discontinuity x the Fourier series converges to $\frac{1}{2}(f(x+0) + f(x-0))$.

As an example we consider the saw-tooth function, see Figure 1.1.1 below,

$$v(x) = \frac{1}{2}(\pi - x) \quad \text{for } 0 < x \leq 2\pi, \quad v(x) = v(x + 2\pi). \quad (1.1.3)$$

Its Fourier coefficients are given by

$$\hat{v}(\omega) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \frac{1}{2}(\pi - x) e^{-i\omega x} dx = \begin{cases} 0, & \text{for } \omega = 0, \\ \sqrt{\frac{\pi}{2}} \frac{1}{i\omega}, & \text{for } \omega \neq 0. \end{cases}$$

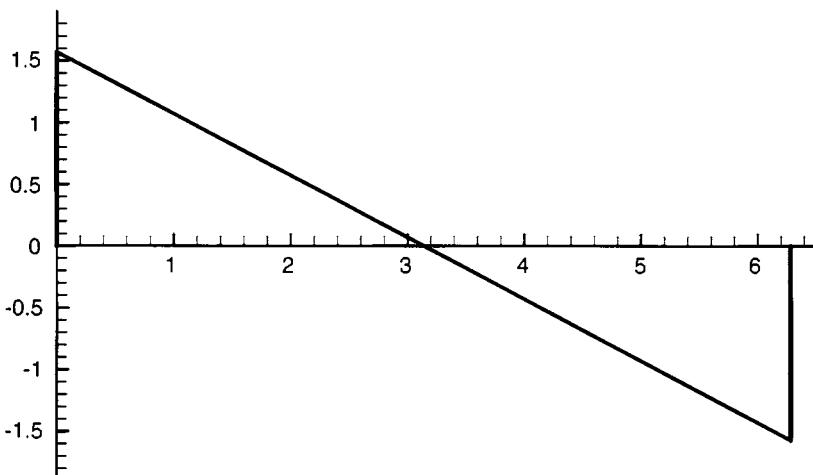


Figure 1.1.1.

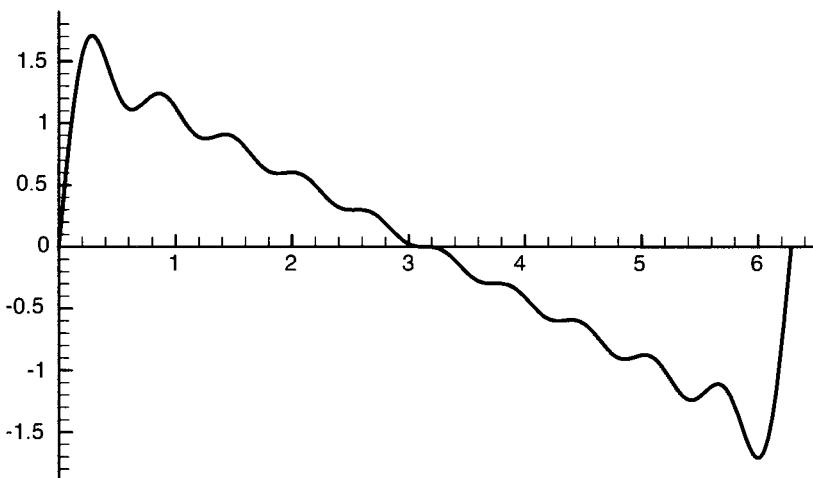


Figure 1.1.2.

Therefore,

$$v(x) = \sum_{\omega \neq 0} \frac{e^{i\omega x}}{2i\omega} = \sum_{\omega=1}^{\infty} \frac{\sin \omega x}{\omega}. \quad (1.1.4)$$

By Theorem 1.1.2, the series converges uniformly to $f(x)$ in every interval $0 < \alpha \leq x \leq \beta < 2\pi$. In the neighborhood of $x = 0, 2\pi$, the so-called Gibb's

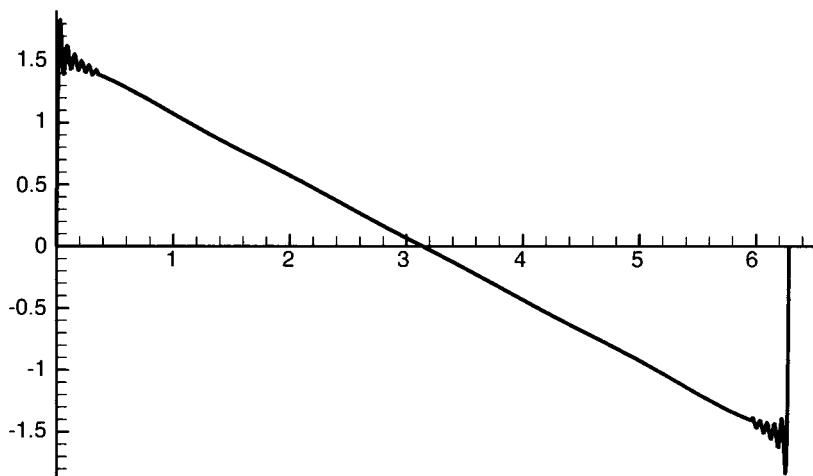


Figure 1.1.3.

phenomenon is evident. In Figures 1.1.2 and 1.1.3 we show the graphs of the partial sums

$$v_N(x) = \sum_{\omega=1}^N \frac{\sin \omega x}{\omega}, \quad N = 10, 100,$$

respectively.

Near the jumps, there are rapid oscillations that narrow, but do not converge, to zero. Analytically one can show that

$$v(x) - v_N(x) = R((N + 1/2)x) + \mathcal{O}\left(\frac{|x| + 1/N}{N}\right),$$

where

$$R(y) = \frac{\pi}{2} - \int_0^y \frac{\sin t}{t} dt.$$

We show $v(x) - v_N(x)$ for $N = 10$ in Figure 1.1.4.

The partial sum $v_N(x)$ is obtained by simply setting the coefficients equal to zero for $\omega > N$. One can also decrease the coefficients smoothly as ω increases. For example, the partial sum can be modified as follows:

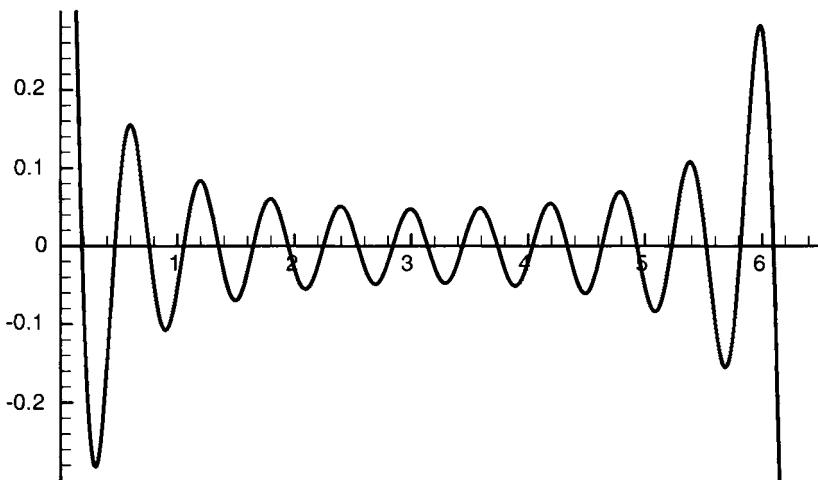


Figure 1.1.4.

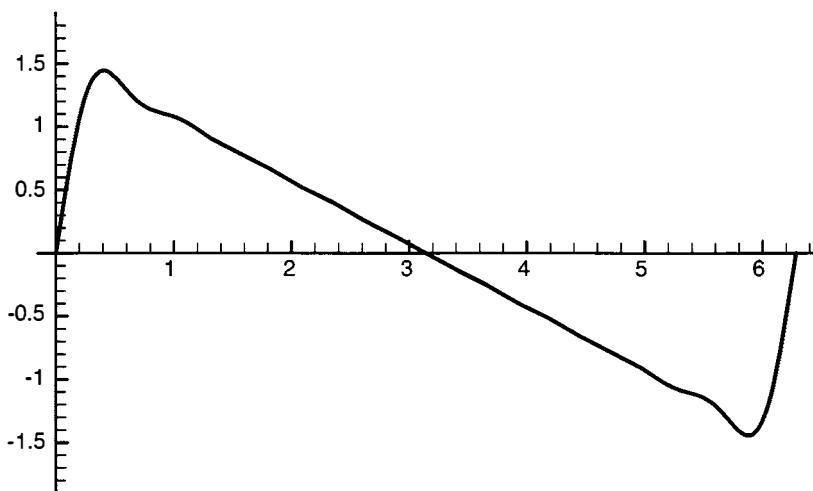


Figure 1.1.5.

$$\tilde{v}_N(x) = \sum_{\omega=1}^N \frac{\sin(\omega\pi/N)}{\omega\pi/N} \frac{\sin \omega x}{\omega}. \quad (1.1.5)$$

In Figures 1.1.5 and 1.1.6, we have calculated $\tilde{v}_N(x)$ for $N = 10, 100$.

There are fewer oscillations, but the jump is not as sharp. One can show that $\tilde{v}_N(x)$ converges to $v(x)$ in every interval $0 < \alpha \leq x \leq \beta < 2\pi$ as $N \rightarrow \infty$.

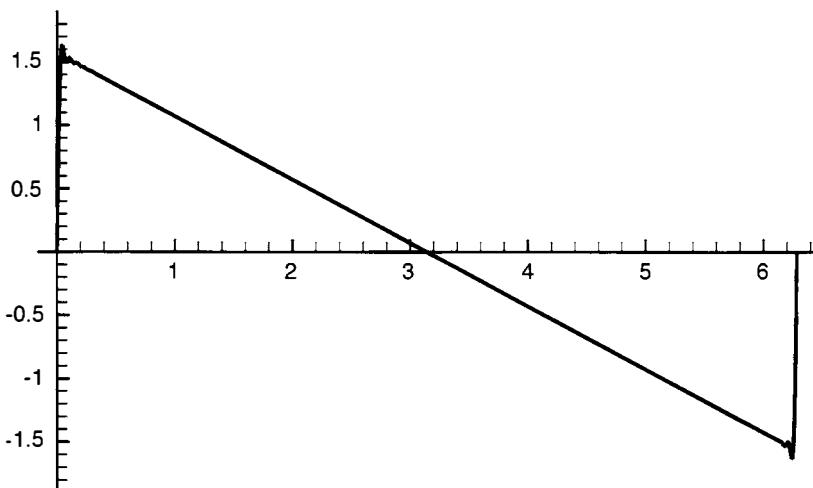


Figure 1.1.6.

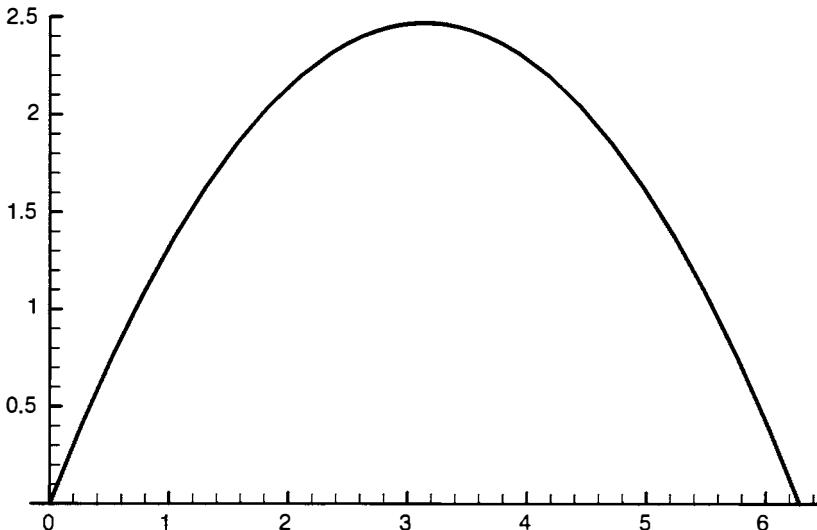


Figure 1.1.7.

If we integrate Eq. (1.1.3), we obtain the smoother function

$$v^{(1)}(x) := \frac{1}{2} \left(\pi x - \frac{1}{2} x^2 \right) = \frac{1}{2} \int_0^x (\pi - t) dt = a_0 - \sum_{\omega=1}^{\infty} \frac{\cos \omega x}{\omega^2},$$

$$a_0 = \sum_{\omega=1}^{\infty} \frac{1}{\omega^2}, \quad 0 \leq x \leq 2\pi, \quad (1.1.6)$$

where $v^{(1)}(x)$ is Lipschitz continuous and its first derivative $dv^{(1)}/dx = v$ has a jump at $x = 0, \pm 2\pi, \pm 4\pi, \dots$. The series converges uniformly for all x . The graph of $v^{(1)}(x)$ is shown in Figure 1.1.7.

In Figures 1.1.8 and 1.1.9, we have calculated $v_N^{(1)} = a_0 - \sum_{\omega=1}^N (\cos \omega x / \omega^2)$ for $N = 10, 100$, respectively. The rate of convergence is much better.

If we integrate Eq. (1.1.3) p times choosing the constant of integration judiciously, we obtain a 2π -periodic function $v^{(p)}$, which has $p-1$ continuous derivatives and whose p th derivative has a jump. Its Fourier coefficients satisfy the estimate

$$|\hat{v}^{(p)}(\omega)| \leq \text{constant} / (|\omega|^p + 1). \quad (1.1.7)$$

We want to show that the decay rate in Eq. (1.1.7) is typical. We say that a 2π -periodic function $f(x)$ is a piecewise C^1 function if we can divide the interval $0 \leq x \leq 2\pi$ into subintervals $\alpha_j \leq x \leq \alpha_{j+1}$, $0 = \alpha_0 < \alpha_1 < \dots < \alpha_n = 2\pi$, so that $f(x)$ is a C^1 function on every open subinterval.

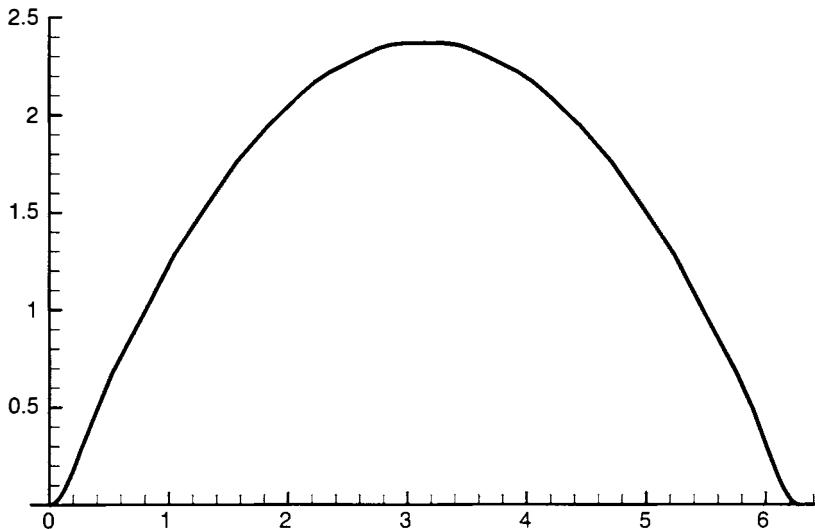


Figure 1.1.8.

Theorem 1.1.3. Let $f(x)$ be a 2π -periodic function and assume that its p th derivative is a piecewise C^1 -function. Then

$$|\hat{f}(\omega)| \leq \text{constant}/(|\omega|^{p+1} + 1).$$

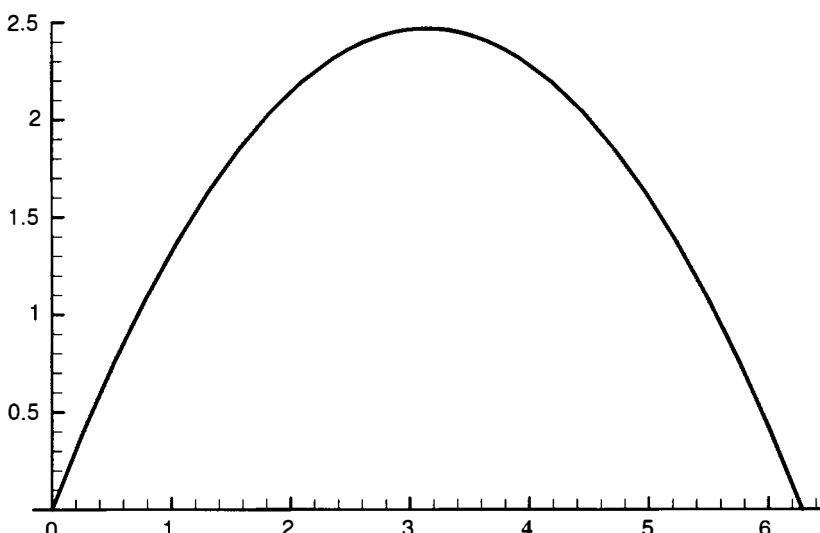


Figure 1.1.9.

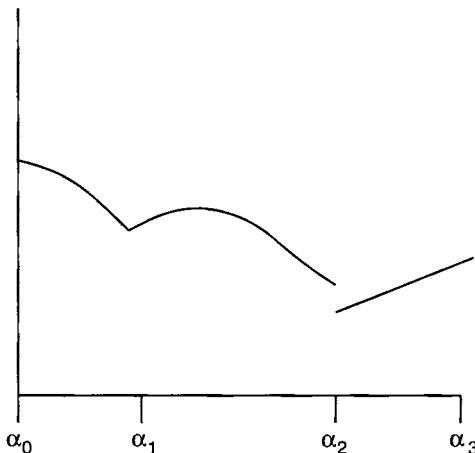


Figure 1.1.10.

Proof. Integration by parts for $\omega \neq 0$ yields

$$\begin{aligned}\sqrt{2\pi}\hat{f}(\omega) &= \int_0^{2\pi} f(x)e^{-i\omega x} dx = \sum_{j=0}^{n-1} \int_{\alpha_j}^{\alpha_{j+1}} f(x)e^{-i\omega x} dx \\ &= -\sum_{j=0}^{n-1} \left(\frac{1}{i\omega} f(x)e^{-i\omega x} \Big|_{\alpha_j+0}^{\alpha_{j+1}-0} - \int_{\alpha_j}^{\alpha_{j+1}} \frac{1}{i\omega} \frac{df(x)}{dx} e^{-i\omega x} dx \right).\end{aligned}$$

If $p = 0$, we obtain the estimate

$$\begin{aligned}|\sqrt{2\pi}\hat{f}(\omega)| &\leq \frac{1}{|\omega|} \left(\sum_{j=0}^{n-1} (|f(\alpha_j + 0)| + |f(\alpha_{j+1} - 0)|) \right. \\ &\quad \left. + \int_{\alpha_j}^{\alpha_{j+1}} \left| \frac{df(x)}{dx} \right| dx \right) \leq \frac{\text{constant}}{|\omega|}.\end{aligned}$$

If $p > 0$, then we obtain

$$\sqrt{2\pi}\hat{f}(\omega) = \frac{1}{i\omega} \int_0^{2\pi} \frac{df(x)}{dx} e^{-i\omega x} dx.$$

We can apply the process to df/dx . The general result follows by induction.

Let $f(x)$ be a piecewise smooth 2π -periodic function with jumps $f(\alpha_j + 0) - f(\alpha_j - 0)$ at $x = \alpha_j$, $0 \leq \alpha_0 < \alpha_1 < \dots < \alpha_{n-1} < 2\pi$. Then

$$f^{(1)} = f(x) - \frac{1}{\pi} \sum_{j=0}^{n-1} (f(\alpha_j + 0) - f(\alpha_j - 0)) v(x - \alpha_j)$$

belongs to C . Here $v(x - \alpha_j)$ is the saw-tooth function shown in Eq. (1.1.3) with jump π at $x = \alpha_j$. $df^{(1)}/dx$ can have jumps $df^{(1)}(\beta_j + 0)/dx - df^{(1)}(\beta_j - 0)/dx$ at $x = \beta_j$, $0 \leq \beta_0 < \beta_1 < \dots < \beta_{m-1} < 2\pi$. Then

$$f^{(2)} = f^{(1)}(x) - \frac{1}{\pi} \sum_{j=0}^{m-1} \left(\frac{df^{(1)}(\beta_j + 0)}{dx} - \frac{df^{(1)}(\beta_j - 0)}{dx} \right) v^{(1)}(x - \beta_j)$$

belongs to C^1 . This process can be continued and the study of general piecewise smooth functions can be reduced to the study of the saw-tooth function.

It is often more convenient to study convergence in the L_2 norm rather than pointwise. Let \bar{f} denote the conjugate complex value of f . We define the L_2 scalar product and norm by

$$(f, g) = \int_0^{2\pi} \bar{f}g \, dx,$$

$$\|f\| = (f, f)^{1/2} = \left(\int_0^{2\pi} |f|^2 dx \right)^{1/2},$$

and say that a sequence f_μ converges to f in the mean, or in L_2 , if

$$\lim_{\mu \rightarrow \infty} \|f_\mu - f\| = 0.$$

The scalar product is a bilinear form that satisfies the following equalities:

$$(f, g) = \overline{(g, f)}, \quad (f + g, h) = (f, h) + (g, h),$$

$$(\lambda f, g) = \bar{\lambda}(f, g), \quad (f, \lambda g) = \lambda(f, g), \quad (1.1.8)$$

where λ is a scalar. The norm satisfies

$$\|\lambda f\| = |\lambda| \|f\|, \quad (1.1.9a)$$

$$|(f, g)| \leq \|f\| \cdot \|g\|, \quad (1.1.9b)$$

and the triangle inequalities

$$\|f + g\| \leq \|f\| + \|g\|, \quad (1.1.9c)$$

$$\|\|f\| - \|g\|\| \leq \|f - g\|. \quad (1.1.9d)$$

The inequality in Eq. (1.1.9b) is a generalization of the usual Cauchy-Schwarz inequality for finite dimensional vector spaces and is obtained by considering the integrals as limits of Riemann sums.

We can now state the fundamental

Lemma 1.1.1. *The exponential functions $(1/\sqrt{2\pi})e^{inx}$, $n = 0, \pm 1, \pm 2, \dots$ are orthonormal with respect to the L_2 scalar product, that is,*

$$\left(\frac{1}{\sqrt{2\pi}} e^{inx}, \frac{1}{\sqrt{2\pi}} e^{imx} \right) = \begin{cases} 1 & \text{for } n = m, \\ 0 & \text{for } n \neq m. \end{cases} \quad (1.1.10)$$

Proof. Equation (1.1.10) follows directly from the definition

$$\left(\frac{1}{\sqrt{2\pi}} e^{inx}, \frac{1}{\sqrt{2\pi}} e^{imx} \right) = \frac{1}{2\pi} \int_0^{2\pi} e^{i(m-n)x} dx.$$

If

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} \hat{f}(j) e^{ijx},$$

then we formally obtain Eq. (1.1.2) from

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-i\omega x} f(x) dx &= \frac{1}{\sqrt{2\pi}} (e^{i\omega x}, f(x)) \\ &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \hat{f}(j) (e^{i\omega x}, e^{ijx}) = \hat{f}(\omega). \end{aligned}$$

We will now investigate the convergence of

$$S_N = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N}^N \hat{f}(\omega) e^{i\omega x}, \quad \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} (e^{i\omega x}, f(x))$$

to f . From Lemma 1.1.1,

$$\|S_N\|^2 = \frac{1}{2\pi} \left(\sum_{\omega=-N}^N \hat{f}(\omega) e^{i\omega x}, \sum_{\nu=-N}^N \hat{f}(\nu) e^{i\nu x} \right) = \sum_{\omega=-N}^N |\hat{f}(\omega)|^2.$$

Also,

$$\begin{aligned} (f, S_N) + (S_N, f) &= \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N}^N ((f, \hat{f}(\omega) e^{i\omega x}) + (\hat{f}(\omega) e^{i\omega x}, f)) \\ &= \sum_{\omega=-N}^N (\hat{f}(\omega) \bar{\hat{f}}(\omega) + \bar{\hat{f}}(\omega) \hat{f}(\omega)) = 2 \sum_{\omega=-N}^N |\hat{f}(\omega)|^2, \end{aligned}$$

that is,

$$\|f - S_N\|^2 = \|f\|^2 - (f, S_N) - (S_N, f) + \|S_N\|^2 = \|f\|^2 - \sum_{\omega=-N}^N |\hat{f}(\omega)|^2,$$

and we obtain Theorem 1.1.4.

Theorem 1.1.4 (Bessel's inequality). *The inequality*

$$\boxed{\sum_{\omega=-N}^N |\hat{f}(\omega)|^2 \leq \|f\|^2} \tag{1.1.11}$$

holds for all N . Furthermore,

$$\boxed{\lim_{N \rightarrow \infty} \|f - S_N\|^2 = \lim_{N \rightarrow \infty} \left(\|f\|^2 - \sum_{\omega=-N}^N |\hat{f}(\omega)|^2 \right) = 0}$$

if and only if Parseval's relation

$$\boxed{\sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = \|f\|^2} \quad (1.1.12)$$

holds.

If $f \in C^1(-\infty, \infty)$ is 2π -periodic, then, by Theorem 1.1.1, S_N converges uniformly to f , that is,

$$\lim_{N \rightarrow \infty} \max_{0 \leq x \leq 2\pi} |f(x) - S_N(x)| = 0,$$

and, therefore, also

$$\lim_{N \rightarrow \infty} \|f - S_N\| = 0.$$

Therefore, Parseval's relation holds for all 2π -periodic $f \in C^1(-\infty, \infty)$. It also holds for much more general functions. Let f be a piecewise continuous function. It is known that it can be approximated arbitrarily well in the L_2 norm by 2π -periodic functions in C^1 , that is, there exists a sequence $\{f_\nu\}$ such that

$$\lim_{\nu \rightarrow \infty} \|f - f_\nu\| = 0.$$

For example, the saw-tooth function can easily be approximated by C^1 functions, see Figure 1.1.11.

By Eq. (1.1.9),

$$|\|f\| - \|f_\nu\|| \leq \|f - f_\nu\|.$$

Therefore,

$$\lim_{\nu \rightarrow \infty} \|f_\nu\| = \|f\|.$$

Also, using Eq. (1.1.11) and the Cauchy–Schwarz inequality for absolutely convergent sums we get

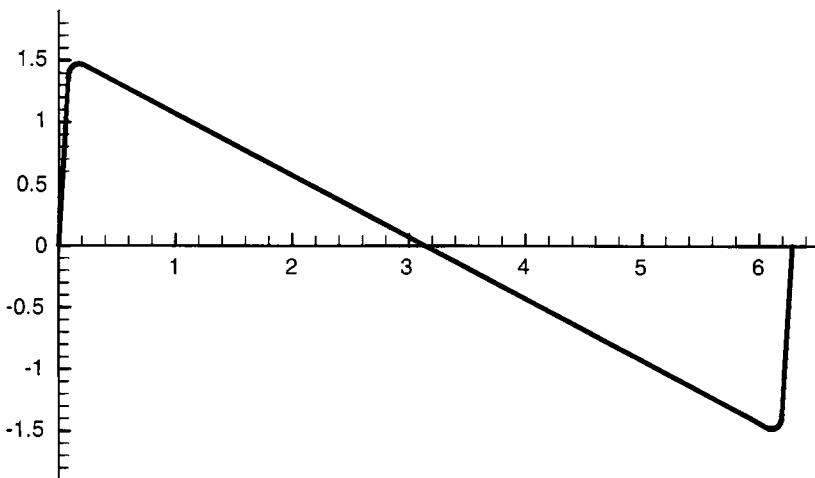


Figure 1.1.11.

$$\begin{aligned}
 & \left| \sum_{\omega = -\infty}^{\infty} (|\hat{f}(\omega)|^2 - |\hat{f}_\nu(\omega)|^2) \right| \\
 & \leq \sum_{\omega = -\infty}^{\infty} (|\hat{f}(\omega)| + |\hat{f}_\nu(\omega)|) \left| |\hat{f}(\omega)| - |\hat{f}_\nu(\omega)| \right| \\
 & \leq \sum_{\omega = -\infty}^{\infty} (|\hat{f}(\omega)| + |\hat{f}_\nu(\omega)|) |\hat{f}(\omega) - \hat{f}_\nu(\omega)| \\
 & \leq \left(\sum_{\omega = -\infty}^{\infty} (|\hat{f}(\omega)| + |\hat{f}_\nu(\omega)|)^2 \right)^{1/2} \\
 & \quad \cdot \left(\sum_{\omega = -\infty}^{\infty} |\hat{f}(\omega) - \hat{f}_\nu(\omega)|^2 \right)^{1/2} \\
 & \leq (2(\|f\|^2 + \|f_\nu\|^2))^{1/2} \|f - f_\nu\|.
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \|f\|^2 - \sum_{\omega = -\infty}^{\infty} |\hat{f}(\omega)|^2 \\
 = \lim_{\nu \rightarrow \infty} \left(\|f\|^2 - \|f_\nu\|^2 - \sum_{\omega = -\infty}^{\infty} (|\hat{f}(\omega)|^2 - |\hat{f}_\nu(\omega)|^2) \right) = 0,
 \end{aligned}$$

and we have proved the following theorem.

Theorem 1.1.5. Any piecewise continuous function f can be expanded into a Fourier series

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

which converges to f in the L_2 norm. [Parseval's relation (1.1.12) holds.]

We define the L_2 space to consist of those functions f that can be approximated arbitrarily well in the L_2 norm by functions in C^1 . Integration is now performed in the Lebesgue sense. By the same argument as before, they can be expanded into Fourier series that converge to f . Therefore, the L_2 space can also be characterized as the space of all convergent Fourier series

$$f = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \quad \|f\|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 < \infty.$$

Instead of C^1 functions, we could also use C^∞ functions, because any C^1 function can be approximated arbitrarily well by C^∞ functions. This is a very useful principle. A “bad” function can be approximated by a sequence of smooth functions. Estimates and other properties are derived for the smooth functions. Then the limit is taken and hopefully the desired properties are also valid in the limit.

A slight generalization of Parseval's relation is shown by Theorem 1.1.6.

Theorem 1.1.6. Let $f, g \in L_2$, then

$$(f, g) = \sum_{\omega=-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{g}(\omega),$$

where

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

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

Proof. If $f, g \in C^1(-\infty, \infty)$, then their Fourier series converge uniformly and, therefore, by Lemma 1.1.1,

$$(f, g) = \frac{1}{2\pi} \left(\sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \sum_{\nu=-\infty}^{\infty} \hat{g}(\nu) e^{i\nu x} \right) = \sum_{\omega=-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{g}(\omega).$$

For general functions in L_2 the relation follows taking limits as before.

EXERCISES

1.1.1. Prove Eqs. (1.1.8) and (1.1.9) for the L_2 scalar product and norm.

1.1.2. Let f be a real function with the Fourier series

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

Prove that

$$S_N = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N}^N \hat{f}(\omega) e^{i\omega x}$$

is real for all N .

1.1.3. Write a program for computing $v(x) - v_N(x)$ as in Figure 1.1.3. Verify the theoretical error estimate numerically.

1.2. PERIODIC GRIDFUNCTIONS AND DIFFERENCE OPERATORS

Let $h = 2\pi/(N+1)$, where N is a natural number, denote a grid interval. A grid on the x axis is defined to be the set of gridpoints

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

A discrete, possibly complex valued, function u defined on the grid is called a gridfunction, see Figure 1.2.1. Here we are only interested in 2π -periodic gridfunctions [i.e., using the notation $u_j = u(x_j)$],

$$u_j = u(x_j) = u(x_j + 2\pi) = u_{j+N+1}.$$

Clearly, the product and sum of gridfunctions are again gridfunctions. Their gridvalues are

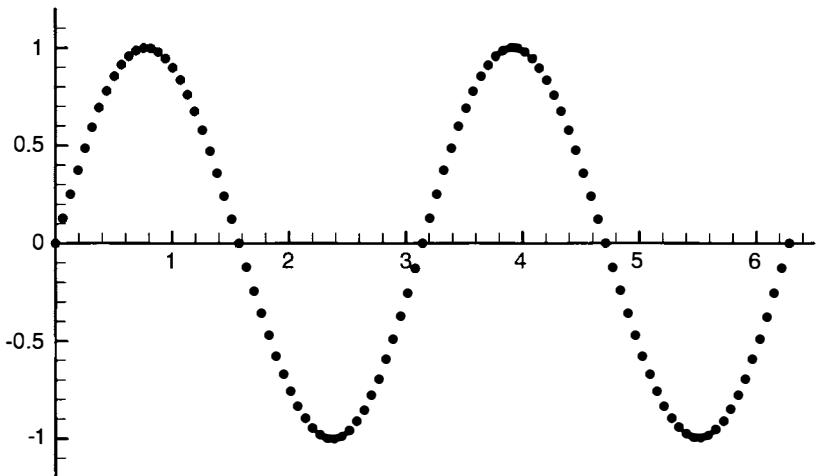


Figure 1.2.1.

$$(uv)_j = u_j v_j, \quad (u + v)_j = u_j + v_j.$$

We denote the set of all 2π -periodic gridfunctions by P_h . If $u, v \in P_h$, then $uv, u + v \in P_h$.

We now introduce difference operators. They play a fundamental role throughout the whole book. We start with the translation operator E . It is defined by

$$(Ev)_j = v_{j+1}.$$

If $v \in P_h$, then $Ev \in P_h$. Powers of E are defined recursively,

$$E^p v = E^{p-1}(Ev).$$

Thus,

$$(E^p v)_j = v_{j+p}. \tag{1.2.1}$$

The inverse also exists and

$$(E^{-1}v)_j = v_{j-1}.$$

If we define E^0 by $E^0 v = v$, then Eq. (1.2.1) holds for all integers p . E is a linear operator and

$$(aE^p + bE^q)v = aE^p v + bE^q v.$$

The forward, backward, and central difference operators are defined by

$$\begin{aligned} D_+ &= (E - E^0)/h, & D_- &= (E^0 - E^{-1})/h = E^{-1}D_+, \\ D_0 &= (E - E^{-1})/(2h) = \frac{1}{2}(D_+ + D_-), \end{aligned} \quad (1.2.2)$$

respectively. In particular, consider these operators acting on the functions $e^{i\omega x}$. Then we have for all $x = x_j$

$$\begin{aligned} hD_+e^{i\omega x} &= (e^{i\omega h} - 1)e^{i\omega x} = (i\omega h + \mathcal{O}(\omega^2 h^2))e^{i\omega x}, \\ hD_-e^{i\omega x} &= (1 - e^{-i\omega h})e^{i\omega x} = (i\omega h + \mathcal{O}(\omega^2 h^2))e^{i\omega x}, \\ hD_0e^{i\omega x} &= i \sin(\omega h)e^{i\omega x} = (i\omega h + \mathcal{O}(\omega^3 h^3))e^{i\omega x}. \end{aligned} \quad (1.2.3)$$

Thus,

$$\begin{aligned} \left| \left(D_+ - \frac{\partial}{\partial x} \right) e^{i\omega x} \right| &= \mathcal{O}(\omega^2 h), & \left| \left(D_- - \frac{\partial}{\partial x} \right) e^{i\omega x} \right| &= \mathcal{O}(\omega^2 h), \\ \left| \left(D_0 - \frac{\partial}{\partial x} \right) e^{i\omega x} \right| &= \mathcal{O}(\omega^3 h^2). \end{aligned} \quad (1.2.4)$$

Consequently, one says that D_+, D_- are first-order accurate approximations of $\partial/\partial x$ since the error is proportional to h . D_0 is second-order accurate.

Higher derivatives are approximated by products of the above operators. For example,

$$\begin{aligned} (D_+ D_- v)_j &= (D_- D_+ v)_j = h^{-2}((E - 2E^0 + E^{-1})v)_j \\ &= h^{-2}(v_{j+1} - 2v_j + v_{j-1}). \end{aligned}$$

In particular,

$$\begin{aligned} h^2 D_+ D_- e^{i\omega x} &= (e^{i\omega h} - 2 + e^{-i\omega h})e^{i\omega x} = -4 \sin^2 \left(\frac{\omega h}{2} \right) e^{i\omega x} \\ &= (-\omega^2 h^2 + \mathcal{O}(\omega^4 h^4))e^{i\omega x}. \end{aligned} \quad (1.2.5)$$

Therefore,

$$\left| \left(D_+ D_- - \frac{\partial^2}{\partial x^2} \right) e^{i\omega x} \right| = \mathcal{O}(\omega^4 h^2),$$

and $D_+ D_-$ is a second-order accurate approximation of $\partial^2/\partial x^2$. Note that all of the above operators commute, since they are all defined in terms of powers of E .

We need to define norms for finite-dimensional vector spaces and discuss some of their properties. We begin with the usual Euclidean inner product and norm. Consider the m -dimensional vector space consisting of all $u = (u^{(1)}, \dots, u^{(m)})^T$ where $u^{(j)}$, $j = 1, \dots, m$, are complex numbers. We denote the conjugate transpose of u by u^* ($u^* = u^T$ if u is real). The scalar product and norm are defined by

$$\langle u, v \rangle = u^* v = \sum_{j=1}^m \bar{u}^{(j)} v^{(j)}, \quad \text{and} \quad |u| = \langle u, u \rangle^{1/2},$$

respectively. The scalar product is a bilinear form that satisfies the following equalities:

$$\langle u, v \rangle = \overline{\langle v, u \rangle}, \tag{1.2.6a}$$

$$\langle u + w, v \rangle = \langle u, v \rangle + \langle w, v \rangle, \tag{1.2.6b}$$

$$\langle \lambda u, v \rangle = \bar{\lambda} \langle u, v \rangle, \quad \langle u, \lambda v \rangle = \lambda \langle u, v \rangle, \lambda \text{ a complex number.} \tag{1.2.6c}$$

The following inequalities hold:

$$|\langle u, v \rangle| \leq |u| |v|, \tag{1.2.7a}$$

$$|u + v| \leq |u| + |v|, \tag{1.2.7b}$$

$$||u| - |v|| \leq |u - v|. \tag{1.2.7c}$$

$$\langle u, v \rangle \leq |u| \cdot |v| \leq \delta |u|^2 + \frac{1}{4\delta} |v|^2 \quad \text{for } \delta > 0. \tag{1.2.7d}$$

Let $A = (a_{ij})$ be a complex $m \times m$ matrix. Then its transpose is denoted by $A^T = (a_{ji})$ and its conjugate transpose by $A^* = (\bar{a}_{ji})$. The Euclidean norm of the matrix A is defined by

$$|A| = \max_{|u|=1} |Au|,$$

where the norm on the right-hand side is the vector norm defined above. If A and B are matrices, then

$$|Au| \leq |A| |u|, \quad (1.2.8a)$$

$$|A + B| \leq |A| + |B|, \quad (1.2.8b)$$

$$|AB| \leq |A| |B|. \quad (1.2.8c)$$

If the scalar λ and vector u satisfy $Au = \lambda u$, then λ is an eigenvalue of A and u is the corresponding eigenvector. The spectral radius, $\rho(A)$, of a matrix A is defined by

$$\rho(A) = \max_j |\lambda_j|,$$

where the λ_j are the eigenvalues of A . $\rho(A)$ satisfies the inequality

$$\rho(A) \leq |A|. \quad (1.2.8d)$$

We next define a scalar product and norm for our periodic gridfunctions of length $N+1$. For fixed h and $N+1$, these functions form a vector space. However, we are interested in these functions as $h \rightarrow 0$ and $N(h)+1 \rightarrow \infty$. The Euclidean inner product and norm defined above would not necessarily be finite in this limit, so we must use a different definition.

We define a discrete scalar product and norm for periodic gridfunctions by

$$(u, v)_h = \sum_{j=0}^N \bar{u}_j v_j h \quad \text{and} \quad \|u\|_h^2 = (u, u)_h,$$

respectively.

The scalar product is also a bilinear form and satisfies the same equalities as the Euclidean inner product above in (1.2.6):

$$(u, v)_h = \overline{(v, u)_h}, \quad (1.2.9a)$$

$$(u + w, v)_h = (u, v)_h + (w, v)_h, \quad (1.2.9b)$$

$$(\lambda u, v)_h = \bar{\lambda}(u, v)_h, \quad (u, \lambda v)_h = \lambda(u, v)_h, \quad \lambda \text{ a complex number.} \quad (1.2.9c)$$

The following inequalities also hold in analogy with (1.2.7):

$$|(u, v)_h| \leq \|u\|_h \|v\|_h, \quad (1.2.10a)$$

$$|(u, av)_h| \leq \|a\|_\infty \|u\|_h \|v\|_h, \quad \|a\|_\infty = \max_j |a_j|, \quad (1.2.10b)$$

$$\|u + v\|_h \leq \|u\|_h + \|v\|_h, \quad (1.2.10c)$$

$$|\|u\|_h - \|v\|_h| \leq \|u - v\|_h. \quad (1.2.10d)$$

If u, v are the projections of continuous functions onto the grid, then

$$\lim_{h \rightarrow 0} (u, v)_h = (u, v), \quad \lim_{h \rightarrow 0} \|u\|_h^2 = \|u\|^2,$$

converge to the L_2 scalar product and norm.

Therefore, the above estimates are also valid for the L_2 scalar product and norm applied to C^1 functions. Since any function $\in L_2$ can be approximated arbitrarily well by a C^1 function they are valid for all L_2 functions. This provides a proof for Eq. (1.1.9).

The norm of an operator is defined in the usual way,

$$\|Q\|_h = \sup_{u \neq 0} \|Qu\|_h / \|u\|_h = \sup_{\|u\|_h=1} \|Qu\|_h.$$

From this definition it follows that $\|Qu\|_h \leq \|Q\|_h \|u\|_h$. Thus,

$$\|E^p u\|_h^2 = \sum_{j=0}^N |u_{j+p}|^2 h = \sum_{j=0}^N |u_j|^2 h = \|u\|_h^2$$

implies

$$\|E^p\|_h = 1, \quad p = 0, \pm 1, \pm 2, \dots \quad (1.2.11)$$

Also,

$$\|D_+ u\|_h = \frac{1}{h} \|(E - E^0)u\|_h \leq \frac{2}{h} \|u\|_h,$$

that is,

$$\|D_+\|_h \leq 2/h.$$

The general inequalities

$$\|P + Q\|_h \leq \|P\|_h + \|Q\|_h, \quad \|PQ\|_h \leq \|P\|_h \|Q\|_h \quad (1.2.12)$$

give us

$$\|D_-\|_h = \|E^{-1} D_+\|_h \leq \frac{2}{h}, \quad \|D_0\|_h = \frac{1}{2h} \|E - E^{-1}\|_h \leq \frac{1}{h}.$$

Actually, these inequalities for the norms of D_+ , D_- , D_0 can be replaced by

equalities. For D_+ we define $u_j = (-1)^j$ and obtain

$$\|u\|_h^2 = (N + 1)h,$$

$$\|D_+ u\|_h^2 = \sum_{j=0}^N ((-1)^{j+1} - (-1)^j)^2 h^{-1} = 4(N + 1)h^{-1} = \frac{4}{h^2} \|u\|_h^2,$$

which yields

$$\|D_+\|_h = 2/h. \quad (1.2.13)$$

Using the same gridfunction u_j again, we get

$$\|D_-\|_h = 2/h. \quad (1.2.14)$$

For D_0 we choose $u_j = i^j$ (where $i = \sqrt{-1}$) and obtain

$$\|u\|_h^2 = (N + 1)h,$$

$$\|D_0 u\|_h^2 = \sum_{j=0}^N \frac{1}{4h} ((-i)^{j+1} - (-i)^{j-1})(i^{j+1} - i^{j-1})$$

$$= \frac{N+1}{h} = \frac{1}{h^2} \|u\|_h^2,$$

so

$$\|D_0\|_h = 1/h. \quad (1.2.15)$$

We now consider systems of partial differential equations and consequently need to define a norm and scalar product for vector-valued gridfunctions $u = (u^{(1)}, \dots, u^{(m)})^T$. Let u and v be two such vector-valued gridfunctions, then we define

$$(u, v)_h = \sum_{j=0}^N \langle u_j, v_j \rangle h \quad (1.2.16a)$$

and

$$\|u\|_h = (u, u)_h^{1/2}. \quad (1.2.16b)$$

The properties shown in Eqs. (1.2.9) and (1.2.10) are still valid. We can also generalize (1.2.10b) when α is replaced by an $(m \times m)$ matrix A . If A is a

constant matrix we have

$$|(Au, v)_h| \leq |A| \|u\|_h \|v\|_h. \quad (1.2.17)$$

If $A = A_j$ is a matrix valued gridfunction, then

$$|(Au, v)_h| \leq \max_j |A_j| \|u\|_h \|v\|_h. \quad (1.2.18)$$

EXERCISES

1.2.1. Derive estimates for

$$\left| \left(D - \frac{\partial^3}{\partial x^3} \right) e^{i\omega x} \right|$$

where $D = D_+^3, D_- D_+^2, D_-^2 D_+, D_-^3, D_0 D_+ D_-$.

1.2.2. The difference operators D_+, D_0 both approximate $\partial/\partial x$, but they have different norms. Explain why this is not a contradiction.

1.2.3. Compute $\|D_+ D_-\|_h$.

1.3. TRIGONOMETRIC INTERPOLATION

Let $u \in P_h$ be a 2π -periodic gridfunction and assume that N is even. We want to show that there is a unique trigonometric polynomial

$$\text{Int}_N u(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}(\omega) e^{i\omega x},$$

which interpolates u , that is,

$$u_j = \text{Int}_N u(x_j) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}(\omega) e^{i\omega x_j}, \quad j = 0, 1, 2, \dots, N. \quad (1.3.1)$$

Equation (1.3.1) represents a system of $N+1$ equations for the $N+1$ unknowns $\tilde{u}(\omega)$. We want to show that this system has a unique solution. In fact, we can derive an explicit representation of $\tilde{u}(\omega)$ in terms of u_j . This representation is a consequence of Lemma 1.3.1.

Lemma 1.3.1. *The exponential functions $e^{i\nu x}$, $\nu = 0, \pm 1, \pm 2, \pm N/2$, are orthogonal with respect to the discrete scalar product, that is,*

$$(e^{i\nu x}, e^{i\mu x})_h = \begin{cases} 2\pi, & \text{if } \nu = \mu, \\ 0, & \text{if } 0 < |\mu - \nu| \leq N. \end{cases}$$

Proof. If $\nu = \mu$ we have

$$(e^{i\nu x}, e^{i\nu x})_h = \sum_{j=0}^N h = (N + 1)h = 2\pi.$$

If $0 < |\nu - \mu| \leq N$, then

$$(e^{i\nu x}, e^{i\mu x})_h = \sum_{j=0}^N e^{i(\mu - \nu)jh} h = \frac{1 - e^{i(\mu - \nu)2\pi}}{1 - e^{i(\mu - \nu)h}} h = 0.$$

This proves the lemma.

We can now prove the following theorem.

Theorem 1.3.1. *The interpolation problem (1.3.1) has the unique solution*

$$\tilde{u}(\omega) = \frac{1}{\sqrt{2\pi}} (e^{i\omega x}, u)_h, \quad |\omega| \leq N/2. \quad (1.3.2)$$

Proof. Assume that Eq. (1.3.1) has a solution. We multiply it by $e^{-i\nu x}h$ and sum to obtain

$$(e^{i\nu x}, u)_h = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} (e^{i\nu x}, e^{i\omega x})_h \tilde{u}(\omega) = \sqrt{2\pi} \tilde{u}(\nu).$$

In particular, if $u_j = 0$, $j = 0, 1, 2, \dots, N$, then the homogeneous equations only have the trivial solution. Therefore, Eq. (1.3.1) has the unique solution (1.3.2).

The process of representing the sequence $\{u(x_j)\}_0^N$ using the Fourier coefficients $\tilde{u}(\omega)$ in Eq. (1.3.1) is often called the discrete Fourier transform. This transform became very important with the advent of the so called fast Fourier transform (FFT). The FFT makes it possible to compute the coefficients $\tilde{u}(\omega)$ in $\mathcal{O}(N \log N)$ operations, compared to the $\mathcal{O}(N^2)$ operations required for their straightforward computation. This transformation can, of course, be defined for any sequence of data. The important fact here is that if the discrete function $\{u_j\}$ can be considered as the restriction of a smooth function u , then the trigonometric polynomial $\text{Int}_N u$ is a very accurate approximation of u . This statement will be made more precise by deriving an error estimate.

REMARK. We assume that N is even and, consequently, that the mesh consists of an odd number of points in the interval $[0, 2\pi]$. This is a consequence of the assumption that our trigonometric polynomials are symmetric, that is, ω goes from $-N/2$ to $N/2$. If N is odd, one can use nonsymmetric polynomials, where $-(N+1)/2 + 1 \leq \omega \leq (N+1)/2$. This corresponds to an even number of points in the interval $[0, 2\pi]$ with $h = 2\pi/(N+1)$ and $x_j = jh$, $j = 0, 1, \dots, N$. All the results we present in this section can also be derived for this case by changing the summation limits. We restrict ourselves to symmetric forms for convenience.

We now discuss properties of the interpolant. We obtain a discrete version of Parseval's relation from Lemma 1.3.1 using the same argument used in the proof of Theorem 1.1.6.

Theorem 1.3.2. *Let*

$$\text{Int}_N u^{(j)} = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}^{(j)}(\omega) e^{i\omega x}, \quad j = 1, 2,$$

interpolate the two gridfunctions. Then

$$(u^{(1)}, u^{(2)})_h = \sum_{\omega=-N/2}^{N/2} \overline{\tilde{u}^{(1)}(\omega)} \tilde{u}^{(2)}(\omega) = (\text{Int}_N u^{(1)}, \text{Int}_N u^{(2)}). \quad (1.3.3)$$

We now prove that the derivatives of an interpolant can be estimated in terms of the difference quotients of the corresponding gridfunction.

Theorem 1.3.3. *Let $\text{Int}_N u$ be the interpolant of a gridfunction u . Then*

$$\|\text{Int}_N u\|^2 = \sum_{\omega=-N/2}^{N/2} |\tilde{u}(\omega)|^2 = \|u\|_h^2, \quad (1.3.4)$$

$$\|D_+^l u\|_h^2 \leq \left\| \frac{d^l}{dx^l} \text{Int}_N u \right\|^2 \leq \left(\frac{\pi}{2} \right)^{2l} \|D_+^l u\|_h^2, \\ l = 0, 1, \dots. \quad (1.3.5)$$

Proof. Equation (1.3.4) follows from Parseval's relation (1.3.3). We also have

$$\begin{aligned}\|D_+^l u\|_h^2 &= \|D_+^l \text{Int}_N u\|_h^2 = \frac{1}{2\pi} \left\| \sum_{\omega} \tilde{u}(\omega) \left(\frac{e^{i\omega h} - 1}{h} \right)^l e^{i\omega x} \right\|_h^2 \\ &= \sum_{\omega} |\tilde{u}(\omega)|^2 \left| \frac{e^{i\omega h} - 1}{h} \right|^{2l} = \sum_{\omega} |\tilde{u}(\omega)|^2 \left(\frac{2 \sin(\omega h/2)}{h} \right)^{2l} \\ &\leq \sum_{\omega} |\tilde{u}(\omega)|^2 \omega^{2l} = \left\| \frac{d^l}{dx^l} \text{Int}_N u \right\|^2.\end{aligned}$$

We obtain the upper bound in Eq. (1.3.5) by using the inequality $2|x|/\pi \leq |\sin x|$ for $|x| \leq \pi/2$.

$$\begin{aligned}\|D_+^l u\|_h^2 &\geq \left(\frac{2}{\pi} \right)^{2l} \sum_{\omega} |\tilde{u}(\omega)|^2 \omega^{2l} \\ &= \left(\frac{2}{\pi} \right)^{2l} \left\| \frac{d^l}{dx^l} \text{Int}_N u \right\|^2,\end{aligned}$$

which proves Eq. (1.3.5).

Now consider a 2π -periodic function u and assume that we can represent it as a Fourier series

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

Consider the restriction of u to the grid. Then we can interpolate the gridvalues and obtain

$$\text{Int}_N u(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}(\omega) e^{i\omega x}.$$

The relationship between the Fourier coefficients $\hat{u}(\omega)$ and the coefficients $\tilde{u}(\omega)$ of the interpolant is computed in Lemma 1.3.2.

Lemma 1.3.2.

Aliasing

$$\tilde{u}(\omega) = \sum_{l=-\infty}^{\infty} \hat{u}(\omega + l(N+1)), \quad |\omega| \leq N/2. \quad (1.3.6)$$

In particular, if $\hat{u}(\omega) = 0$ for $|\omega| > N/2$, then $\text{Int}_N u \equiv u$.**Proof.** We can write any integer μ in the form

$$\mu = \omega + l(N+1), \quad |\omega| \leq N/2, \text{ where } l \text{ is an integer.}$$

As in Section 1.2, $h = 2\pi/(N+1)$ and, therefore,

$$e^{i\mu x_j} = e^{i\omega x_j} e^{il(N+1)x_j} = e^{i\omega x_j} e^{il(N+1)jh} = e^{i\omega x_j} e^{2\pi ilj} = e^{i\omega x_j}$$

implies

$$\begin{aligned} u(x_j) &= \frac{1}{\sqrt{2\pi}} \sum_{\mu=-\infty}^{\infty} \hat{u}(\mu) e^{i\mu x_j}, \\ &= \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \left(\sum_{l=-\infty}^{\infty} \hat{u}(\omega + l(N+1)) \right) e^{i\omega x_j}, \\ &\equiv \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{u}(\omega) e^{i\omega x_j}. \end{aligned}$$

Equation (1.3.6) follows from the uniqueness of the interpolant.

We can now prove the fundamental approximation theorem for trigonometric interpolation. Let

$$\|u(\cdot)\|_{\infty} = \sup_{0 \leq x \leq 2\pi} |u(x)|$$

denote the L_{∞} norm.**Theorem 1.3.4.** Let u be a 2π -periodic function and assume that its Fourier coefficients satisfy an estimate

$$|\hat{u}(\omega)| \leq \frac{C}{|\omega|^m}, \quad \omega \neq 0, m > 1. \quad (1.3.7)$$

Then

$$\|u(\cdot) - \text{Int}_N u(\cdot)\|_\infty \leq \frac{2C}{\sqrt{2\pi}} (N/2)^{1-m} \left(\frac{1}{m-1} + \frac{2(N+1)}{N} B_m \right),$$

$$B_m = \sum_{j=1}^{\infty} \frac{1}{(2j-1)^m}. \quad (1.3.8)$$

Proof. We write $u(x) = u_N(x) + u_R(x)$, where

$$\begin{aligned} u_N(x) &= \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} \hat{u}(\omega) e^{i\omega x}, \\ u_R(x) &= \frac{1}{\sqrt{2\pi}} \sum_{|\omega| > N/2} \hat{u}(\omega) e^{i\omega x}. \end{aligned}$$

We also write $\text{Int}_N u = \text{Int}_N u_N + \text{Int}_N u_R$. $\text{Int}_N u_N$ is the trigonometric interpolant of $u_N(x)$ and from Lemma 1.3.2 we have $\text{Int}_N u_N = u_N$. $\text{Int}_N u_R$ is the interpolant of $u_R(x)$, so Lemma 1.3.2 implies that

$$\tilde{u}_R(\omega) = \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} \hat{u}(\omega + j(N+1)), \quad |\omega| \leq N/2, \quad (1.3.9)$$

and Eq. (1.3.7) gives us

$$|u_R(x)| \leq \frac{2C}{\sqrt{2\pi}} \sum_{\omega=N/2+1}^{\infty} \omega^{-m} \leq \frac{2C}{\sqrt{2\pi}} \frac{(N/2)^{1-m}}{m-1}.$$

From Eqs. (1.3.7) and (1.3.9), we obtain

$$\begin{aligned} |\tilde{u}_R(\omega)| &\leq \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} |\hat{u}(\omega + j(N+1))| \\ &\leq \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} \frac{C}{|\omega + j(N+1)|^m}, \quad |\omega| \leq N/2. \end{aligned}$$

Since $|\omega + j(N+1)| \geq |-N/2 + j(N+1)| > |N(2j-1)/2|$ for $j > 0$ and $|\omega + j(N+1)| \geq |N/2 + j(N+1)| > |N(2j+1)/2|$ for $j < 0$, we can estimate the

last sum to obtain

$$|\tilde{u}_R(\omega)| \leq 2C(N/2)^{-m} \sum_{j=1}^{\infty} \frac{1}{(2j-1)^m}.$$

Therefore,

$$\|\text{Int}_N u_R\|_{\infty} \leq \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} |\tilde{u}_R(\omega)| \leq \frac{4}{\sqrt{2\pi}} \frac{N+1}{N} C(N/2)^{1-m} B_m,$$

and Eq. (1.3.8) follows from

$$\|u(x) - \text{Int}_N u(x)\|_{\infty} \leq \|u_R(x)\|_{\infty} + \|\text{Int}_N u_R(x)\|_{\infty}.$$

Theorem 1.3.4 establishes uniform convergence for $m > 1$ and, by Theorem 1.1.3, applies to continuous functions that are piecewise smooth. The term $\text{Int}_N u_R$ is often referred to as the aliasing error. The final estimate of $\text{Int}_N u_R$ shows that this error is quite small if the function u is smooth, that is, if m is large. Also, it is of the same order as the error we commit by truncating the Fourier series.

The next result follows immediately from the last theorem.

Corollary 1.3.1. *There are constants C_l such that*

$$\left\| \frac{d^l}{dx^l} u(x) - \frac{d^l}{dx^l} \text{Int}_N u(x) \right\|_{\infty} < C_l (N/2)^{1+l-m}, \quad 1 + l < m.$$

EXERCISES

- 1.3.1. Write a program that computes $\text{Int}_N u(x)$, where $u(x)$ is the smoothed saw-tooth function as defined in Section 1.1. Verify numerically that the error is $\mathcal{O}(N^{1-m})$ according to Eq. (1.3.8).
- 1.3.2. Formulate and prove the theorems in Section 1.3 for an even number of points in the interval $[0, 2\pi]$.

1.4. THE S-OPERATOR FOR DIFFERENTIATION

Theorem 1.3.4 shows that we can obtain very accurate approximations of smooth functions $u(x)$ using trigonometric interpolation. If the function $u(x)$

is well represented in terms of Fourier components that correspond to wave numbers ω less than $N/2$ in magnitude, then $\text{Int}_N u$ is a very accurate approximation. In particular, $\text{Int}_N u \equiv u(x)$ if $\hat{u}_\omega = 0$ for $|\omega| > N/2$. This is a good reason for using trigonometric polynomials as a basis for approximating derivatives accurately.

Denote by $w(x)$ the derivative of the interpolating polynomial $\text{Int}_N u$, that is,

$$\begin{aligned} w(x) &= \frac{d}{dx} \text{Int}_N u = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} \tilde{w}(\omega) e^{i\omega x} \\ &=: \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} i\omega \tilde{u}(\omega) e^{i\omega x}. \end{aligned} \quad (1.4.1)$$

First interpolation,
then take derivative,
just like find zeros

Let $u(x)$ be a smooth function and assume that we only know its values at the gridpoints x_j . We want to compute approximations of du/dx at the gridpoints x_j . We proceed by calculating the trigonometric interpolant and then evaluate the derivative of this interpolant in the points x_j . In detail, the $\{\tilde{u}(\omega)\}$ are computed by using the FFT, and the $\{\tilde{w}(\omega)\}$ are obtained by $N+1$ multiplications. Then w can be calculated in the gridpoints using the inverse FFT. If we define the gridvalues of u as a vector $\mathbf{u} = (u_0, u_1, \dots, u_N)^T$, and correspondingly, $\mathbf{w} = (w_0, w_1, \dots, w_N)^T$, then the relation between these two vectors can be written as

$$\mathbf{w} = S\mathbf{u} \quad (1.4.2)$$

Differential Matrix in Spectral method

where S is an $(N+1) \times (N+1)$ matrix. The matrix S is a discrete form of the differential operator $\partial/\partial x$, as were the difference operators D_+ and D_0 defined in Section 1.2. Since it uses information from all gridpoints, it is natural to expect good accuracy. In fact, Corollary 1.3.1 gives us an error estimate. If there are many gridpoints, the computation can be performed very efficiently using the FFT.

We will now analyze some basic properties of the matrix S . Denote by \mathbf{e}_ω the vector

$$\mathbf{e}_\omega = (1, e^{i\omega h}, e^{i\omega 2h}, \dots, e^{i\omega Nh})^T, \quad |\omega| \leq N/2. \quad (1.4.3)$$

In the previous section we have proved that \mathbf{u} can be expressed in terms of these vectors, that is,

$$\mathbf{u} = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} \tilde{u}(\omega) \mathbf{e}_\omega \quad (1.4.4)$$

Also,

$$S\mathbf{e}_\omega = i\omega \mathbf{e}_\omega, \quad |\omega| \leq N/2, \quad (1.4.5)$$

because the interpolant corresponding to \mathbf{e}_ω is $e^{i\omega x}$. Therefore, \mathbf{e}_ω is an eigenvector of S with the eigenvalue $i\omega$. The Euclidean scalar product of two eigenvectors is

$$\langle \mathbf{e}_\omega, \mathbf{e}_\nu \rangle = \sum_{j=0}^N e^{i(\nu - \omega)jh} = \begin{cases} N + 1, & \text{if } \omega = \nu, \\ 0, & \text{if } 0 < |\omega - \nu| \leq N, \end{cases} \quad (1.4.6)$$

which was proved in Lemma 1.3.1.

This shows that the eigenvectors of S form an orthogonal basis. If U is the matrix whose columns are these eigenvectors, then we can write $S = U\Lambda U^*$ where Λ is a diagonal matrix whose entries are the purely imaginary eigenvalues of S . Consequently $S^* = (U\Lambda U^*)^* = U\Lambda^* U^* = U(-\Lambda)U^* = -S$, that is, S is skew-Hermitian. The eigenvalue of maximum modulus is $iN/2$, hence $\rho(S) = N/2$, and, since S is skew-Hermitian,

$$|S| := \max_{|\mathbf{u}|=1} |S\mathbf{u}| = N/2.$$

We collect these results in the following lemma.

Lemma 1.4.1. *The matrix S , defined in Eq. (1.4.2), has the following properties:*

1. It is skew-Hermitian.
2. The eigenvalues are $i\omega$, $\omega = -N/2, -N/2 + 1, \dots, N/2$.
3. $|S| = N/2$.

REMARK. We can also think of S as an operator on P_h (see Section 1.2). Then the above properties become

$$(\mathbf{u}, S\mathbf{v})_h = -(S\mathbf{u}, \mathbf{v})_h, \quad \|S\|_h = N/2.$$

EXERCISES

- 1.4.1.** Write a program that computes $\mathbf{w} = S\mathbf{u}$ according to Eq. (1.4.2) for some function u , where $\partial u / \partial x$ is analytically known. Find the number of grid-points $N + 1$ such that the error

$$\max_{-N/2 \leq j \leq N/2} \left| \frac{\partial u(x_j)}{\partial x} - w_j \right|$$

is equal to

$$\max_{-50 \leq j \leq 50} \left| \frac{\partial u(x_j)}{\partial x} - D_0 u_j \right|.$$

1.5. GENERALIZATIONS

Let $f(x) = f(x_1, x_2)$, $g(x) = g(x_1, x_2)$ denote functions that are **2π -periodic** in both x_1 and x_2 . We define the L_2 scalar product and norm by

$$(f, g) = \int_0^{2\pi} \int_0^{2\pi} \bar{f}g \, dx_1 dx_2, \quad \|f\| = (f, f)^{1/2}. \quad (1.5.1)$$

The trigonometric functions

$$\begin{aligned} e^{i\langle \omega, x \rangle}, \quad \omega &= (\omega_1, \omega_2), \quad \omega_j \text{ integers,} \\ x &= (x_1, x_2), \quad \langle \omega, x \rangle = \omega_1 x_1 + \omega_2 x_2, \end{aligned} \quad (1.5.2)$$

are again orthogonal, that is,

$$(e^{i\langle \omega, x \rangle}, e^{i\langle \nu, x \rangle}) = \begin{cases} (2\pi)^2, & \text{for } \omega = \nu, \\ 0, & \text{for } \omega \neq \nu. \end{cases} \quad (1.5.3)$$

Therefore, we obtain a formal Fourier series

$$f(x) = \frac{1}{2\pi} \sum_{\omega_1=-\infty}^{\infty} \sum_{\omega_2=-\infty}^{\infty} \hat{f}(\omega) e^{i\langle \omega, x \rangle}, \quad (1.5.4)$$

where

$$\hat{f}(\omega) = \frac{1}{2\pi} (e^{i\langle \omega, x \rangle}, f) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} f(x) e^{-i\langle \omega, x \rangle} dx_1 dx_2.$$

If $f(x_1, x_2)$ is a piecewise C^p function, then we can again use integration by

parts to prove for $\omega_j \neq 0$,

$$|\hat{f}(\omega)| \leq K_{jp}/|\omega_j|^p, \quad j = 1, 2,$$

where

$$\begin{aligned} K_{1p} &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} \left| \frac{\partial^p f}{\partial x_1^p} \right| dx_1 dx_2, \\ K_{2p} &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} \left| \frac{\partial^p f}{\partial x_2^p} \right| dx_1 dx_2. \end{aligned}$$

Therefore, there are constants K and K_p , such that

$$|\hat{f}(\omega)| \leq \begin{cases} K, & \text{for } \omega = 0, \\ \frac{K_p}{|\omega_1|^p + |\omega_2|^p}, & \text{for } \omega \neq 0. \end{cases} \quad (1.5.5)$$

We will now show that the Fourier series (1.5.4) converges uniformly to f for smooth functions. If $f(x_1, x_2)$ is a smooth function of x_1, x_2 , then it is a smooth function of x_2 for every fixed x_1 . Therefore, the Fourier series

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\sqrt{2\pi}} \sum_{\omega_2=-\infty}^{\infty} \hat{f}_2(x_1, \omega_2) e^{i\omega_2 x_2}, \\ \hat{f}_2(x_1, \omega_2) &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(x_1, x_2) e^{-i\omega_2 x_2} dx_2, \end{aligned} \quad (1.5.6)$$

converge uniformly to $f(x_1, x_2)$. Every Fourier coefficient is a smooth function of x_1 and, therefore, can be expanded into a uniformly convergent series

$$\hat{f}_2(x_1, \omega_2) = \frac{1}{\sqrt{2\pi}} \sum_{\omega_1=-\infty}^{\infty} \hat{f}(\omega_1, \omega_2) e^{i\omega_1 x_1}, \quad (1.5.7)$$

with

$$\begin{aligned}
 \hat{f}(\omega_1, \omega_2) &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \hat{f}_2(x_1, \omega_2) e^{-i\omega_1 x_1} dx_1 \\
 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{2\pi} f(x_1, x_2) e^{-i\langle \omega, x \rangle} dx_1 dx_2 \\
 &= \frac{1}{2\pi} (e^{i\langle \omega, x \rangle}, f) = \hat{f}(\omega).
 \end{aligned}$$

Using Eq. (1.5.5), we obtain Eq. (1.5.4) if we substitute Eq. (1.5.7) into Eq. (1.5.6), and the series converges uniformly to f .

Again, we can relax the smoothness requirement if we are only interested in convergence in the L_2 norm. Let

$$S_N = \frac{1}{2\pi} \sum_{\omega_1=-N}^N \sum_{\omega_2=-N}^N \hat{f}(\omega) e^{i\langle \omega, x \rangle}.$$

As before,

$$\|f - S_N\|^2 = \|f\|^2 - \sum_{\omega_1=-N}^N \sum_{\omega_2=-N}^N |\hat{f}(\omega)|^2,$$

and, therefore,

$$\lim_{N \rightarrow \infty} \|f - S_N\| = 0$$

if, and only if, Parseval's relation,

$$\|f\|^2 = \sum_{\omega_1=-\infty}^{\infty} \sum_{\omega_2=-\infty}^{\infty} |\hat{f}(\omega)|^2, \quad (1.5.8)$$

holds. As before, Parseval's relation holds for all smooth functions and all functions that can be approximated arbitrarily well by smooth functions in the L_2 norm. This class of functions is called L_2 . The generalization of Parseval's relation

$$(f, g) = \sum_{\omega_1=-\infty}^{\infty} \sum_{\omega_2=-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{g}(\omega) \quad (1.5.9)$$

is also valid.

A two-dimensional grid is defined by

$$x_j := h j := h(j_1, j_2), \quad h = 2\pi/(N + 1), \quad j_\nu = 0 \pm 1, \pm 2, \dots \quad (1.5.10)$$

Gridfunctions are denoted by

$$u_j := u(x_j) = u(hj_1, hj_2).$$

We assume that the gridfunctions are 2π -periodic in both directions.

E_{x_1}, E_{x_2} will denote the translation operators in the x_1 and x_2 directions, respectively, that is,

$$E_{x_1} u(x_j) = u(h(j_1 + 1), hj_2), \quad E_{x_2} u(x_j) = u(hj_1, h(j_2 + 1)).$$

The difference operators $D_{+x_1}, D_{+x_2}, \dots$ are defined in terms of E_{x_1}, E_{x_2} as before. The discrete scalar product and norm are now defined by

$$(u, v)_h = \sum_{j_1=0}^N \sum_{j_2=0}^N \bar{u}_j v_j h^2, \quad \|u\|_h^2 = (u, u)_h^{1/2}.$$

As we can develop a Fourier series in two dimensions using a one-dimensional Fourier series, we can develop the two-dimensional interpolating polynomial using one-dimensional ones. Let $u(x) = u(x_1, x_2)$ be a gridfunction. For every fixed x_1 , we determine

$$\text{Int}_{x_2} u(x_1, x_2) = \frac{1}{\sqrt{2\pi}} \sum_{\omega_2=-N/2}^{N/2} \tilde{u}_2(x_1, \omega_2) e^{i\omega_2 x_2}. \quad (1.5.11)$$

Then we interpolate $\tilde{u}_2(x_1, \omega_2)$ for every fixed ω_2 and obtain

$$\text{Int}_{x_1} \tilde{u}_2(x_1, \omega_2) = \frac{1}{\sqrt{2\pi}} \sum_{\omega_1=-N/2}^{N/2} \tilde{u}(\omega_1, \omega_2) e^{i\omega_1 x_1}. \quad (1.5.12)$$

Substituting Eq. (1.5.12) into Eq. (1.5.11) gives us the desired two-dimensional interpolating polynomial which, therefore, can be obtained by solving $2(N + 1)$ one-dimensional interpolation problems.

We can consider Fourier expansions of vector-valued functions instead of scalar functions. We can also consider vector-valued functions $f = (f^{(1)}, \dots, f^{(m)})$ in d space dimensions. We can use a different discretization interval in each space dimension $h_l = 2\pi/(N_l + 1)$, $l = 1, 2, \dots, d$. Then $x_j := (h_1 j_1, \dots, h_d j_d)$ and $f_j = f(x_j)$. The Fourier expansion of f is then of

the form

$$f = \frac{1}{(2\pi)^{d/2}} \sum_{\omega} \hat{f}(\omega) e^{i\langle \omega, x \rangle}, \quad \hat{f} = (\hat{f}^{(1)}, \dots, \hat{f}^{(m)}),$$

where $\hat{f}^{(\nu)}$ are the Fourier coefficients of $f^{(\nu)}$. The discrete scalar product becomes

$$(f, g)_h = \sum_{j_1=0}^{N_1} \dots \sum_{j_d=0}^{N_d} \langle f_j, g_j \rangle h_1 \dots h_d,$$

where

$$\langle f_j, g_j \rangle = \sum_{\nu=1}^m \bar{f}_j^{(\nu)} g_j^{(\nu)}.$$

We will usually assume that $h_1 = \dots = h_d$ to simplify the notation.

EXERCISES

- 1.5.1. Formulate and prove the generalization of Theorems 1.3.1, 1.3.3 in two space dimensions.
- 1.5.2. Compute $\|D_{+x_j}\|_h$, $\|D_{-x_j}\|_h$, $\|D_{0x_j}\|_h$, $j = 1, 2$, on a rectangular grid with gridsize h_j in the x_j direction, $j = 1, 2$.
- 1.5.3. Discuss the generalization of the S operator to two space dimensions.

BIBLIOGRAPHIC NOTES

There is an extensive literature on Fourier series. Two of the books containing most of the basic theory are by Churchill and Brown (1978) and by Zygmund (1977). Some of the theory is also found in Courant-Hilbert (1953). The basic Theorem 1.1.1 is proven there and the examples in Section 1.1 are also discussed. Theorem 1.1.2 is proven by Titchmarsh (1937).

Although most of the components of the Fast Fourier Transform (FFT) were known about 1920, it was the basic paper by Cooley and Tukey (1965), which created the method as we know it today. A general description of the FFT is found in Conte de Boor (1972).

2

MODEL EQUATIONS

In this chapter, we examine several model equations to introduce some basic properties of differential equations and difference approximations by example. Generalizations of these ideas are discussed throughout the remainder of this book.

2.1. FIRST-ORDER WAVE EQUATION, CONVERGENCE, AND STABILITY

The equation $u_t = u_x$ is the simplest *hyperbolic* equation; the general definition of the class of hyperbolic equations is given in Chapter 6. We consider the initial value problem

$$\begin{aligned} u_t &= u_x, & -\infty < x < \infty, 0 \leq t, \\ u(x, 0) &= f(x), & -\infty < x < \infty, \end{aligned} \quad (2.1.1)$$

where $f(x) = f(x + 2\pi)$ is a smooth 2π -periodic function. To begin, we assume that

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{f}(\omega)$$

consists of one wave. We try to find a solution of the same type

$$u(x, t) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{u}(\omega, t). \quad (2.1.2)$$

Substituting Eq. (2.1.2) into Eq. (2.1.1) yields the ordinary differential equation

$$\frac{d\hat{u}}{dt} = i\omega \hat{u}, \quad \hat{u}(\omega, 0) = \hat{f}(\omega),$$

which is called the Fourier transform of Eq. (2.1.1). Therefore,

$$\hat{u}(\omega, t) = e^{i\omega t} \hat{u}(\omega, 0) = e^{i\omega t} \hat{f}(\omega).$$

It follows that

$$u(x, t) = \frac{1}{\sqrt{2\pi}} e^{i\omega(x+t)} \hat{f}(\omega) = f(x + t) \quad (2.1.3)$$

is a solution of our problem. Now consider the general case

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{f}(\omega). \quad (2.1.4)$$

By the **superposition principle**

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega(x+t)} \hat{f}(\omega) = f(x + t) \quad (2.1.5)$$

is a solution to our problem. For every fixed t Parseval's relation yields

$$\|u(\cdot, t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{i\omega t} \hat{f}(\omega)|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = \|f(\cdot)\|^2. \quad (2.1.6)$$

$\|u\|^2$ is often called the **energy** of u . Therefore, Eq. (2.1.1) is said to be **energy conserving**—the obvious phrase, **norm conserving**, is often used in this context as well. Clearly, any method of approximation must be nearly norm conserving to be useful. We also note that there is a **finite speed of propagation** associated with this problem. The expression (2.1.5) shows that the solution is constant along the lines $x + t = \text{constant}$ which are called **characteristics** (see Figure 2.1.1). Any particular feature of the initial data, such as a wave crest, is propagated along these characteristics. In our case, **the speed of propagation (or wave speed) is $dx/dt = -1$** . For general hyperbolic systems, there may be many families of characteristics corresponding to different wave speeds of different components. The important thing is that these speeds are always finite.

We now solve the problem using a difference approximation. We introduce a **space step $h = 2\pi/(N+1)$** , with N a natural number, and a **time step $k > 0$** . h, k define a grid in x, t space, consisting of the **gridpoints** $(x_j, t_n) := (jh, nk)$. Gridfunctions will be denoted by $u_j^n = u(x_j, t_n)$. A simple approximation based on forward differences in time and centered differences in space is

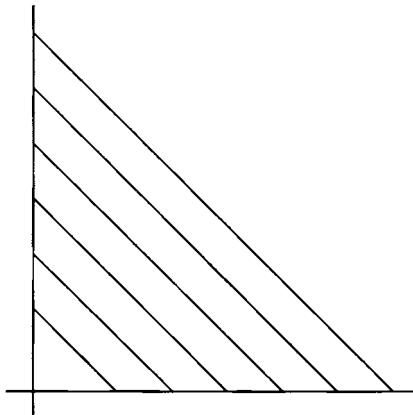


Figure 2.1.1.

$$\begin{aligned} v_j^{n+1} &= (I + kD_0)v_j^n =: Qv_j^n, \\ v_j^0 &= f_j, \quad j = 0, \pm 1, \pm 2, \dots \end{aligned} \tag{2.1.7}$$

If v^n is known at time $t_n = nk$, then we can use Eq. (2.1.7) to calculate v_j^{n+1} for all j . Thus, the initial data determine a unique solution. Also, if v^n is 2π -periodic, then v^{n+1} is too. Therefore, we can restrict the calculation to $j = 0, 1, 2, \dots, N$ and use periodicity conditions to extend the solution and provide the extra needed values for Eq. (2.1.7) at $j = 0, N$, that is, $v_{-1}^n = v_N^n, v_{N+1}^n = v_0^n$.

We will now calculate the solution analytically. First consider the case where f consists of one single wave, that is,

$$f_j = \frac{1}{\sqrt{2\pi}} e^{i\omega x_j} \hat{f}(\omega), \quad j = 0, 1, 2, \dots, N.$$

As in the continuous case we make the ansatz

$$v_j^n = \frac{1}{\sqrt{2\pi}} \hat{v}^n(\omega) e^{i\omega x_j}, \tag{2.1.8}$$

that is, we assume that the solution can also be expressed in terms of one single Fourier component. Substituting Eq. (2.1.8) into Eq. (2.1.7) yields

$$e^{i\omega x_j} \hat{v}^{n+1}(\omega) = \left(e^{i\omega x_j} + \frac{\lambda}{2} (e^{i\omega x_{j+1}} - e^{i\omega x_{j-1}}) \right) \hat{v}^n(\omega),$$

where $\lambda = k/h$. This equation can be rewritten as

$$e^{i\omega x_j} \hat{v}^{n+1}(\omega) = (1 + i\lambda \sin \xi) e^{i\omega x_j} \hat{v}^n(\omega),$$

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

$$\hat{v}^{n+1}(\omega) = \hat{Q} \hat{v}^n(\omega), \quad \hat{Q} = 1 + i\lambda \sin \xi. \quad (2.1.9)$$

\hat{Q} is called the *symbol*, or *amplification factor*, of $(I + kD_0)$. We refer to it as the Fourier transform of $(I + kD_0)$ and Eq. (2.1.9) as the Fourier transform of Eq. (2.1.7). The solution of Eq. (2.1.9) is

$$\hat{v}^n(\omega) = \hat{Q}^n \hat{v}^0(\omega) = \hat{Q}^n \hat{f}(\omega),$$

and it is clear that

$$v_j^n = \frac{1}{\sqrt{2\pi}} \hat{Q}^n e^{i\omega x_j} \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \left(1 + i \frac{k}{h} \sin(\omega h) \right)^n e^{i\omega x_j} \hat{f}(\omega)$$

solves our problem.

Now we consider a sequence of mesh intervals $k, h \rightarrow 0$. We want to show that v_j^n converges to the corresponding solution of the differential equation. We have

$$\begin{aligned} \left(1 + i \frac{k}{h} \sin(\omega h) \right)^n &= (1 + i\omega k + \mathcal{O}(kh^2\omega^3))^n \\ &= (e^{i\omega k} + \mathcal{O}(k^2\omega^2 + kh^2\omega^3))^n \\ &= (1 + \mathcal{O}((k\omega^2 + h^2\omega^3)t_n)) e^{i\omega t_n}. \end{aligned}$$

Therefore,

$$v_j^n = \frac{1}{\sqrt{2\pi}} (1 + \mathcal{O}((k\omega^2 + h^2\omega^3)t_n)) e^{i\omega(x_j + t_n)} \hat{f}(\omega).$$

Thus, for every fixed ω , we obtain

$$\lim_{k, h \rightarrow 0} v_j^n = u(x_j, t_n)$$

in any finite interval $0 \leq t \leq T$.

Now assume that the initial data are represented by a trigonometric polynomial

$$u(x, 0) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-M}^M e^{i\omega x} \hat{f}(\omega).$$

By the superposition principle, the above result implies that the solution of the difference approximation will converge to the solution of the differential equation as $k, h \rightarrow 0$. Thus, one might think that the approximation could be useful in practice. However, consider Eq. (2.1.1) with initial data $f(x) \equiv 0$ which has the trivial solution $u(x, t) \equiv 0$. Now consider the problem with **perturbed data**

$$\hat{f}(\omega) = \begin{cases} \epsilon, & \text{for } \omega = N/4, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding solution of the transformed difference approximation is

$$\hat{v}^n(N/4) = \left(1 + i \frac{k}{h} \sin \left(\frac{2\pi}{N+1} \frac{N}{4} \right) \right)^n \epsilon \sim \left(1 + i \frac{k}{h} \right)^n \epsilon,$$

that is,

$$|\hat{v}^{t_n/k}(N/4)|^2 \sim \left(1 + \frac{k^2}{h^2} \right)^{t_n/k} \epsilon^2.$$

For $t_n = 1$, that is $n = 1/k$

$$|\hat{v}^{1/k}(N/4)|^2 \sim \left(1 + \frac{k^2}{h^2} \right)^{1/k} \epsilon^2.$$

Now consider any sequence $k, h \rightarrow 0$ with $k/h = \lambda > 0$ fixed. Then

$$\lim_{k \rightarrow 0} |\hat{v}^{1/k}(N/4)| = \infty.$$

This “explosion,” or growth, can be arbitrarily fast. For example, if we consider $\lambda = 10$, $k = 10^{-5}$, then

$$|\hat{v}^{1/k}(N/4)|^2 \sim 100^{10^5} \epsilon^2.$$

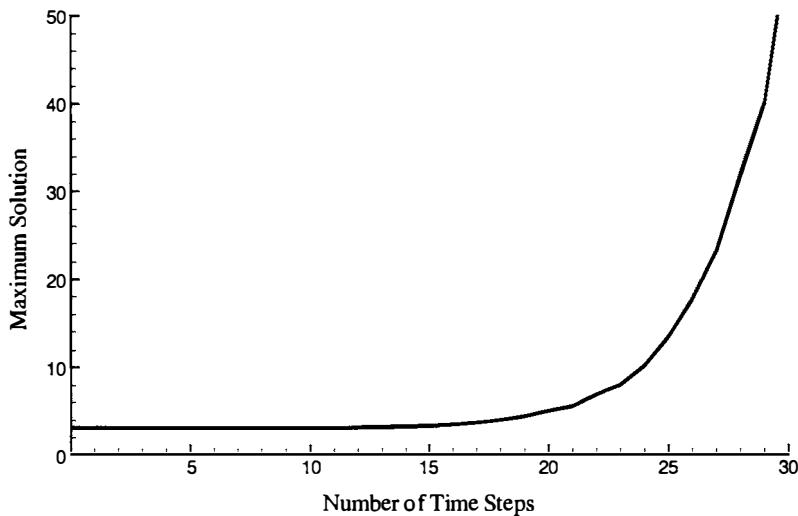


Figure 2.1.2.

The numerical calculation is therefore worthless. In Figures 2.1.2 and 2.1.3 we have calculated the maximum of the solutions of Eq. (2.1.7) with initial data

$$f_j = \begin{cases} x_j, & \text{for } 0 \leq x_j \leq \pi, \\ 2\pi - x_j, & \text{for } \pi \leq x_j \leq 2\pi, \end{cases}$$

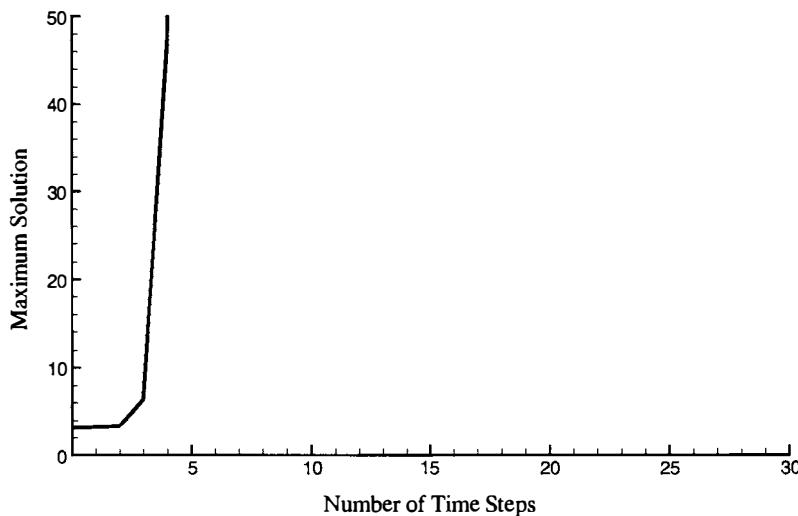


Figure 2.1.3.

and stepsizes $h = 10^{-2}$, $k = 10^{-2}$ and $h = 10^{-2}$, $k = 10^{-1}$, respectively. The analytic results lead us to expect that the solutions will grow like $2^{n/2}$ and $101^{n/2}$, respectively. The numerical results confirm that prediction.

In realistic computations one must always expect perturbations, either from measurement errors in the data or from rounding errors due to the finite representation of numbers in the computer. Therefore, we consider only such sequences $k, h \rightarrow 0$, for which

$$\sup_{0 \leq t_n \leq T, \omega, k, h} |\hat{Q}^n| \leq K(T), \quad (2.1.10)$$

and we call the method *stable* for such a sequence.

If we choose a sequence with

$$k = ch^2, \quad c > 0 \text{ constant},$$

then

$$|\hat{Q}^n|^2 = \left| 1 + \frac{k^2}{h^2} \operatorname{sch}^2(\omega h) \right|^n \leq (1 + \lambda^2)^n = (1 + ck)^n \leq e^{ct_n},$$

and Eq. (2.1.10) holds.

However, this method is not practical to use for the following reasons:

1. As we will see, there are stable methods with $k = ch$. Therefore, the amount of work required to calculate v at $t = 1$ with this method is unnecessarily large.
2. If we want to calculate v for large t , then the exponential growth factor e^{ct} can amplify small perturbations so much that the solution is worthless.

We modify our previous difference equation by adding *artificial viscosity*, that is, we consider

$$v_j^{n+1} = (I + kD_0)v_j^n + \sigma khD_+D_-v_j^n, \quad v_j^0 = f_j. \quad (2.1.11)$$

Here $\sigma > 0$ is a constant, which we will choose later. We can write Eq. (2.1.11) in the form

$$\frac{v_j^{n+1} - v_j^n}{k} = D_0v_j^n + \sigma hD_+D_-v_j^n, \quad (2.1.12)$$

which approximates the differential equation

$$u_t = u_x + \sigma h u_{xx}.$$

As $h \rightarrow 0$ we obtain Eq. (2.1.1). Thus, Eq. (2.1.11) is a *consistent* difference approximation of Eq. (2.1.1), that is, the difference approximation converges formally to the differential equation as $k, h \rightarrow 0$.

We will now choose σ, k , and h so that

$$|\hat{Q}| \leq 1. \quad (2.1.13)$$

In this case, Eq. (2.1.10) is certainly satisfied.

From Eqs. (1.2.3) and (1.2.5), \hat{Q} is of the form

$$\hat{Q} = 1 + i\lambda \sin \xi - 4\sigma \lambda \sin^2 \frac{\xi}{2}, \quad \xi = \omega h, \lambda = k/h.$$

Therefore,

$$\begin{aligned} |\hat{Q}|^2 &= \left(1 - 4\sigma \lambda \sin^2 \frac{\xi}{2} \right)^2 + \lambda^2 \sin^2 \xi, \\ &= 1 - 8\sigma \lambda \sin^2 \frac{\xi}{2} + 16\sigma^2 \lambda^2 \sin^4 \frac{\xi}{2} + 4\lambda^2 \sin^2 \frac{\xi}{2} \left(1 - \sin^2 \frac{\xi}{2} \right), \\ &= 1 - (8\sigma \lambda - 4\lambda^2) \boxed{\sin^2 \frac{\xi}{2}} + (16\sigma^2 - 4)\lambda^2 \boxed{\sin^4 \frac{\xi}{2}}. \end{aligned}$$

There are two ways we can satisfy Eq. (2.1.13):

1. Suppose $2\sigma \leq 1$. If $0 \leq 8\sigma \lambda - 4\lambda^2$, that is,

$$0 < \lambda \leq 2\sigma \leq 1 \quad (2.1.14)$$

then $|\hat{Q}| \leq 1$. By letting $|\xi|$ be small, we see that these conditions are also necessary.

2. Suppose $1 \leq 2\sigma$. If we replace $\sin^4(\xi/2)$ by $\sin^2(\xi/2)$ it follows that $|\hat{Q}| \leq 1$ if

$$0 \leq 8\sigma \lambda - 4\lambda^2 - 16\sigma^2 \lambda^2 + 4\lambda^2,$$

that is,

$$1 \leq 2\sigma, \quad 2\sigma \lambda \leq 1. \quad (2.1.15)$$

By letting $\sin(\xi/2) = 1$ we see that these conditions are also necessary.

There are two particular schemes of the above type which have been used extensively:

1. *The Lax–Friedrichs Method ($\sigma = h/2k = 1/(2\lambda)$).*

$$\begin{aligned} v_j^{n+1} &= \frac{1}{2}(v_{j+1}^n + v_{j-1}^n) + kD_0v_j^n, \\ &= (I + kD_0)v_j^n + \frac{1}{2}h^2D_+D_-v_j^n. \end{aligned} \quad (2.1.16)$$

In this case, Eq. (2.1.15) is satisfied if $k/h \leq 1$, that is, $|\hat{Q}| \leq 1$. It is remarkable that the simple change $v_j^n \rightarrow \frac{1}{2}(v_{j+1}^n + v_{j-1}^n)$ has such an effect on the solution.

2. *The Lax–Wendroff Method ($\sigma = k/2h = \lambda/2$).*

$$v_j^{n+1} = v_j^n + kD_0v_j^n + \frac{k^2}{2} D_+D_-v_j^n. \quad (2.1.17)$$

Now Eq. (2.1.14) is satisfied if $k/h \leq 1$.

In Figures 2.1.4 and 2.1.5, we have used the Lax–Friedrichs method and the Lax–Wendroff method to calculate the solution of Eq. (2.1.1) with initial data

$$f(x) = \begin{cases} x, & \text{for } 0 \leq x \leq \pi, \\ 2\pi - x, & \text{for } \pi \leq x \leq 2\pi, \end{cases}$$

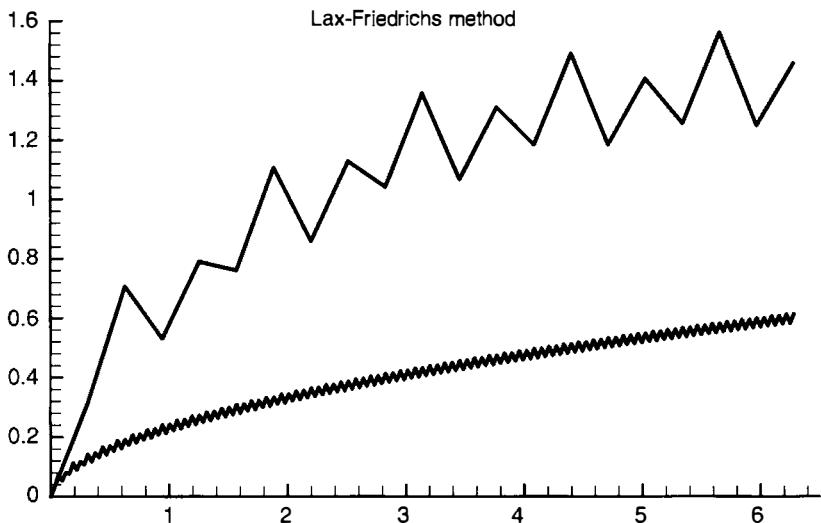


Figure 2.1.4. The Lax–Friedrichs method.

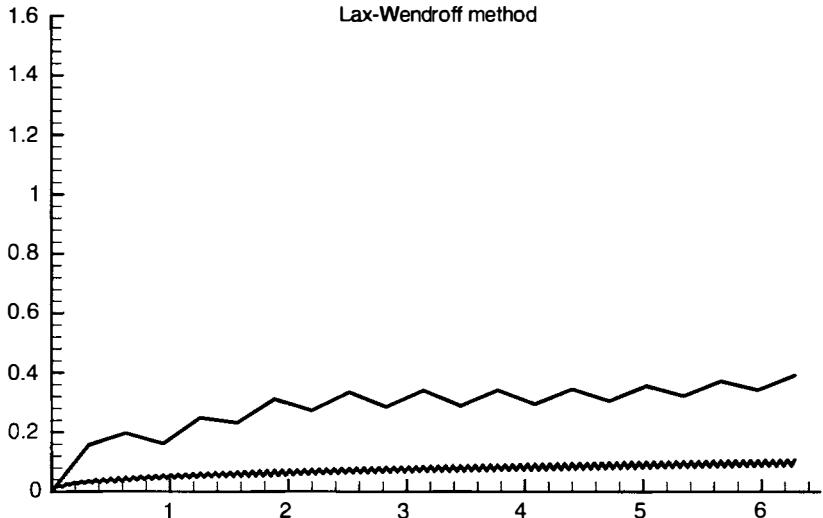


Figure 2.1.5. The Lax-Wendroff method.

and $k/h = 1/2$, $h = 2\pi/10, 2\pi/100$, respectively. We show here the absolute maximum error plotted against time. The accuracy is not impressive, but there is no explosion.

We now consider a rather general difference approximation of Eq. (2.1.1):

$$v_j^{n+1} = Q v_j^n, \quad Q = \sum_{\nu=-r}^s A_\nu(k, h) E^\nu, \quad v_j^0 = f_j. \quad (2.1.18)$$

Here the A_ν are rational functions of k and h , and $r, s \geq 0$ are integers. Thus, we use the $s+r+1$ values $v_{j-r}^n, \dots, v_{j+s}^n$ to calculate v_j^{n+1} . We again consider simple wave solutions

$$v_j^n = \frac{1}{\sqrt{2\pi}} e^{i\omega x_j} \hat{v}^n(\omega).$$

By observing that $E e^{i\omega x} = e^{i\xi} e^{i\omega x}$, we obtain

$$\hat{v}^{n+1}(\omega) = \hat{Q}(\xi) \hat{v}^n(\omega), \quad \hat{Q} = \sum_{\nu=-r}^s A_\nu e^{i\nu\xi},$$

that is,

$$\hat{v}^n(\omega) = \hat{Q}^n(\xi)\hat{v}^0(\omega). \quad (2.1.19)$$

We assume that the initial data belongs to L_2 , that is, $f(x)$ can be expanded as a Fourier series

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \quad \sum_{\omega} |\hat{f}(\omega)|^2 < \infty. \quad (2.1.20)$$

For the difference approximation we use the restriction of $f(x)$ to the grid. We denote by

$$\text{Int}_N f = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{f}(\omega) e^{i\omega x} \quad (2.1.21)$$

the trigonometric interpolant of the gridfunction. We assume that

$$\lim_{N \rightarrow \infty} \|\text{Int}_N f - f\| = 0. \quad (2.1.22)$$

From Theorem 1.3.4, this convergence condition is satisfied if f is a smooth function.

We now want to prove the following theorem.

Theorem 2.1.1. Consider the difference approximation shown in Eq. (2.1.18) on a finite interval $0 \leq t \leq T$ for a sequence $h, k \rightarrow 0$. Assume that

1. The initial data satisfy Eqs. (2.1.20) and (2.1.22).
2. The approximation is stable, that is, there is a constant K_s such that for all k and h

$$\sup_{0 \leq t_n \leq T} |\hat{Q}^n| \leq K_s.$$

3. The approximation is consistent, that is, for every fixed ω

$$\lim_{k, h \rightarrow 0} \sup_{0 \leq t_n \leq T} |\hat{Q}^n(\xi) - e^{i\omega t_n}| = 0.$$

Then the trigonometric interpolant $\text{Int}_N v$ of the solution of the difference approximation converges to the solution of the differential equation,

$$\lim_{k,h \rightarrow 0} \sup_{0 \leq t_n \leq T} \|u(\cdot, t_n) - \text{Int}_N(v_j^n)\| = 0.$$

Proof. For every fixed t_n , we can represent the solution of the difference approximation by its trigonometric interpolant and, therefore, we can think of the solution as being represented in terms of simple waves. From Eq. (2.1.19), we obtain

$$\text{Int}_N(v_j^n) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{Q}^n(\xi) \tilde{f}(\omega) e^{i\omega x}.$$

Let $0 < M < N/2$ be a fixed integer. From Eq. (2.1.5) and Parseval's relation we obtain

$$\begin{aligned} \|u(\cdot, t_n) - \text{Int}_N(v_j^n)\|^2 &= \sum_{\omega=-N/2}^{N/2} |e^{i\omega t_n} \hat{f}(\omega) - \hat{Q}^n(\xi) \tilde{f}(\omega)|^2 \\ &\quad + \sum_{|\omega| > N/2} |\hat{f}(\omega)|^2 \leq I + II + III, \end{aligned}$$

where

$$\begin{aligned} I &= \sum_{\omega=-M}^M |\hat{Q}^n(\xi) \tilde{f}(\omega) - e^{i\omega t_n} \hat{f}(\omega)|^2, \\ II &= 2 \sum_{|\omega| > M} |\hat{f}(\omega)|^2, \\ III &= 2 \sum_{|\omega| > M} |\tilde{f}(\omega)|^2 |\hat{Q}^n(\xi)|^2. \end{aligned}$$

By Eq. (2.1.20),

$$\lim_{M \rightarrow \infty} II = 0.$$

From Eq. (2.1.22) and the second assumption,

$$\lim_{M \rightarrow \infty} III \leq 4K_s^2 \lim_{M \rightarrow \infty} \sum_{|\omega| > M} (|\tilde{f}(\omega) - \hat{f}(\omega)|^2 + |\hat{f}(\omega)|^2) = 0.$$

Finally, for every fixed M , the second and third assumptions together with Eq.

(2.1.22) imply that

$$\begin{aligned} \lim_{N \rightarrow \infty} I &\leq 2 \lim_{N \rightarrow \infty} \sum_{\omega=-M}^M (|\hat{Q}^n(\xi)(\tilde{f}(\omega) - \hat{f}(\omega))|^2 \\ &+ |(\hat{Q}^n(\xi) - e^{i\omega t_n})\hat{f}(\omega)|^2) = 0. \end{aligned}$$

Now convergence follows easily. Let $\epsilon > 0$ be a constant. Choose M so large that $II + III < \epsilon/2$. For sufficiently large N we also have $I \leq \epsilon/2$ and, therefore, convergence follows. This proves the theorem.

This theorem tells us that the solution of the difference approximation converges to the solution of the differential equation if the approximation is **stable and consistent**. In actual calculations one uses fixed values of k and h , and wants to know the size of the error for these particular values of k and h . We will discuss such estimates in Chapter 5.

EXERCISES

- 2.1.1. The convergence of the solutions in Figures 2.1.4 and 2.1.5 is rather slow. Explain why that is so and find which one of the terms I , II , or III is large for this example in the proof of Theorem 2.1.1.
- 2.1.2. Modify the scheme (2.1.11) such that it approximates $u_t = -u_x$. Prove that the conditions (2.1.14) and (2.1.15) are also necessary for stability in this case.
- 2.1.3. Choose σ in Eq. (2.1.11) such that Q uses only two gridpoints. What is the stability condition?

2.2. LEAP-FROG SCHEME

The difference approximations that we discussed in the previous section were all so-called **one-step** methods, that is, v_j^{n+1} could be expressed as a linear combination of neighboring values $v_{j-r}^n, \dots, v_{j+s}^n$ at the previous time level. The **leap-frog scheme**

$$v_j^{n+1} = v_j^{n-1} + \lambda(v_{j+1}^n - v_{j-1}^n), \quad \lambda = k/h, \quad (2.2.1)$$

is a **two-step** method because v_j^{n+1} is determined by values at two previous time levels. To start the calculation, we have to specify v_j^0 and v_j^1 . The initial data yields $v_j^0 = f_j$, while v_j^1 can be determined by a **one-step** method. It does not need to be stable, because we use it only once. The simplest one is Eq. (2.1.7), that is,

Image Point method

$$v_j^1 = (I + kD_0)v_j^0 = (I + kD_0)f_j. \quad (2.2.2)$$

We again seek simple wave solutions

$$v_j^n = \frac{1}{\sqrt{2\pi}} e^{i\omega x_j} \hat{v}^n(\omega),$$

and obtain

$$\hat{v}^{n+1}(\omega) = \hat{v}^{n-1}(\omega) + 2i\lambda (\sin \xi) \hat{v}^n(\omega). \quad (2.2.3)$$

To solve Eq. (2.2.3) we make the ansatz

$$\hat{v}^n(\omega) = z^n, \quad (2.2.4)$$

where z is a complex number.

Substituting Eq. (2.2.4) into Eq. (2.2.3) gives us

$$z^{n+1} = z^{n-1} + 2i\lambda (\sin \xi) z^n,$$

and, therefore, Eq. (2.2.4) is a solution of Eq. (2.2.3) if, and only if, z satisfies the so called **characteristic equation**

$$z^2 = 1 + 2i\lambda z \sin \xi. \quad (2.2.5)$$

For $0 < \lambda < 1$, Eq. (2.2.5) has two distinct solutions with

$$|z_j| = 1,$$

given by

$$z_{1,2} = i\lambda \sin \xi \pm \sqrt{1 - \lambda^2 \sin^2 \xi}. \quad (2.2.6)$$

The general solution of Eq. (2.2.3) is

$$\hat{v}^n = \sigma_1 z_1^n + \sigma_2 z_2^n. \quad (2.2.7)$$

The parameters σ_1 and σ_2 are determined by the initial data. If $\hat{v}^0(\omega) = \hat{f}(\omega)$, then by Eq. (2.2.2)

$$\hat{v}^1(\omega) = (1 + i\lambda \sin \xi) \hat{f}(\omega),$$

and we obtain the linear system of equations

$$\boxed{\begin{aligned}\sigma_1 + \sigma_2 &= \hat{f}(\omega), \\ \sigma_1 z_1 + \sigma_2 z_2 &= (1 + i\lambda \sin \xi) \hat{f}(\omega).\end{aligned}} \quad (2.2.8)$$

As in the one-step case, we consider the low frequencies with $|\omega h| \ll 1$. Then, if $\lambda = k/h = \text{constant}$,

$$\begin{aligned}z_1 &= 1 + i\omega k - \frac{1}{2}\omega^2 k^2 + \mathcal{O}(\omega^3 k^3) = e^{i\omega k(1 + \mathcal{O}(\omega^2 k^2))}, \\ z_2 &= -(1 - i\omega k - \frac{1}{2}\omega^2 k^2 + \mathcal{O}(\omega^3 k^3)) = -e^{-i\omega k(1 + \mathcal{O}(\omega^2 k^2))}.\end{aligned}$$

After a simple calculation, Eq. (2.2.8) gives us

$$\sigma_1 = \hat{f}(\omega)(1 + \mathcal{O}(\omega^2 k^2)), \quad \sigma_2 = \mathcal{O}(\omega^2 k^2) \hat{f}(\omega),$$

and, therefore,

$$\begin{aligned}\hat{v}^n(\omega) &= \hat{f}(\omega)(1 + \mathcal{O}(\omega^2 k^2)) e^{i\omega t_n(1 + \mathcal{O}(\omega^2 k^2))} \\ &\quad + \mathcal{O}(\omega^2 k^2) \hat{f}(\omega)(-1)^n e^{-i\omega t_n(1 + \mathcal{O}(\omega^2 k^2))}.\end{aligned}$$

Thus, the solution consists of two parts. The first part approximates the corresponding solution $\hat{u}(\omega, t_n) = \hat{f}(\omega) e^{i\omega t_n}$, and the error is $\mathcal{O}(\omega^3 k^2 t_n)$. The second part oscillates rapidly and is independent of the differential equation. It is often called a *parasitic solution*. Luckily, the amplitude is small for $\omega^2 k^2 \ll 1$ and does not grow with time.

Since the leap-frog scheme uses three time levels, Theorem 2.1.1 does not apply as formulated. However, using the form of $\hat{v}^n(\omega)$ derived above, we can again use trigonometric interpolation to prove convergence to the solution of the differential equation. Smoothness of the initial data and the *stability condition*

$$\lambda = k/h \leq 1 - \delta, \quad \delta > 0 \quad (2.2.9)$$

for any sequence $k, h \rightarrow 0$ are required for convergence (see Exercise 2.2.1).

We now discuss a property of the leap-frog scheme that can cause practical difficulties. Let $a > 0$ be a constant, and consider the differential equation

$$u_t = u_x - au.$$

The simple wave solutions are now of the form

$$u = \frac{1}{\sqrt{2\pi}} e^{i\omega(x+t)} e^{-at} \hat{f}(\omega),$$

which clearly decay exponentially with time. We use the approximation

$$v_j^{n+1} = v_j^{n-1} + 2kD_0 v_j^n - 2ka v_j^n. \quad (2.2.10)$$

The simple wave solutions again have the form

$$v_j^n = (\sigma_1 z_1^n + \sigma_2 z_2^n) e^{i\omega x_j},$$

where now $z_{1,2}$ are the solutions of

$$z^2 = 1 + (2i\lambda \sin \xi - 2ka)z,$$

that is,

$$z_{1,2} = i\lambda \sin \xi - ka \pm \sqrt{1 + (i\lambda \sin \xi - ka)^2}.$$

Consider the special case $\omega = 0$. For $ka \ll 1$, we have

$$\begin{aligned} z_1 &= 1 - ka + \frac{k^2 a^2}{2} + \mathcal{O}(k^3 a^3), \\ &= e^{-ka + \mathcal{O}(k^3 a^3)}, \\ z_2 &= -e^{ka + \mathcal{O}(k^3 a^3)}, \end{aligned}$$

and, as before,

$$\begin{aligned} \hat{v}^n(0) &= \hat{f}(0)(1 + \mathcal{O}(k^2 a^2)) e^{-at_n(1 + \mathcal{O}(k^2 a^2))} \\ &\quad + \mathcal{O}(k^2 a^2) \hat{f}(\omega) (-1)^n e^{at_n(1 + \mathcal{O}(k^2 a^2))}. \end{aligned}$$

Now the parasitic solution grows exponentially and can obliterate the exponentially decaying solution. Therefore, the (unmodified) leap-frog scheme cannot be used for long time intervals. It is easy to modify the scheme and suppress this behavior. Instead of Eq. (2.2.10), we use

$$(1 + ka)v_j^{n+1} = (1 - ka)v_j^{n-1} + 2kD_0 v_j^n. \quad (2.2.11)$$

Now $z_{1,2}$ are the solutions of

$$(1 + ka)z^2 = 1 - ka + 2i\lambda z \sin \xi,$$

and

$$z_{1,2} = \frac{i\lambda \sin \xi}{1 + ka} \pm \sqrt{\frac{1 - \lambda^2 \sin^2 \xi - k^2 a^2}{(1 + ka)^2}}.$$

Therefore,

$$|z_{1,2}| = \frac{(1 - k^2 a^2)^{1/2}}{1 + ka} \approx e^{-ka} \quad \text{for } \lambda^2 < 1 - k^2 a^2,$$

and both $z_{1,2}^n$ decay like e^{-at_n} ; that is, the solutions have the same decay rates as the solution of the differential equation.

We close this section by noting that the condition $k \leq h$, found necessary for the explicit schemes of Eqs. (2.1.11) and (2.2.1), is very natural. Recall that the solution $u(x, t)$ of Eq. (2.1.1) at any point (\tilde{x}, \tilde{t}) is determined by the value of $f(x)$ at the point $\tilde{x} + \tilde{t}$ on the x axis, because $u(x, t)$ is constant along the characteristic $x + t = \tilde{x} + \tilde{t}$ going through (\tilde{x}, \tilde{t}) and $(\tilde{x} + \tilde{t}, 0)$. Now assume that (\tilde{x}, \tilde{t}) is a gridpoint. Then the solution of the difference approximation at (\tilde{x}, \tilde{t}) depends on the initial data in the interval $\tilde{x} - \tilde{t}/\lambda \leq x \leq \tilde{x} + \tilde{t}/\lambda$ (see Figure 2.2.1). If $\tilde{x} + \tilde{t}$ does not belong to this interval, that is, if $\lambda > 1$, then we cannot hope to obtain an accurate approximation. The condition that the domain of dependence of the difference approximation include the domain of dependence of the differential equation is known as the *Courant–Friedrichs–Lowy condition* (CFL-condition).

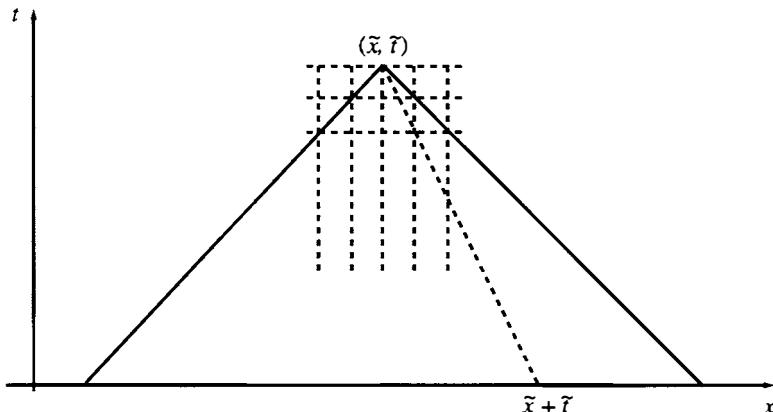


Figure 2.2.1. Domain of dependence for an explicit difference scheme.

The domain of dependence at a certain point (\tilde{x}, \tilde{t}) for a general hyperbolic differential equation is not a single point on the x axis but rather a set of points or a whole interval.

EXERCISES

- 2.2.1. Prove that the solution of the leap-frog scheme converges to the solution of the differential equation, if $\lambda \leq 1 - \delta, \delta > 0$.
- 2.2.2. Derive the explicit form of the leap-frog approximation (2.2.1) for $\lambda = 1$. Is the scheme suitable for computation?
- 2.2.3. Let $a = 10$. Estimate the time interval $[0, T]$, where the approximation (2.2.10) can be used. Does T depend on ω and/or k ?

2.3. IMPLICIT METHODS

There is another way to stabilize the approximation (2.1.7). If we replace the forward difference in time by a backward difference, we get the *backward Euler method*

$$(I - kD_0)v_j^{n+1} = v_j^n, \quad j = 0, 1, \dots, N. \quad (2.3.1)$$

If we introduce the vector $\mathbf{v} = (v_0, \dots, v_N)^T$, then we can write Eq. (2.3.1) in matrix form

$$\boxed{A\mathbf{v}^{n+1} = \mathbf{v}^n,}$$

$$A = \begin{bmatrix} 1 & -k/2h & 0 & \dots & 0 & k/2h \\ k/2h & 1 & -k/2h & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & k/2h & 1 & -k/2h \\ -k/2h & 0 & \dots & 0 & k/2h & 1 \end{bmatrix}. \quad (2.3.2)$$

This is called an *implicit* scheme, because it couples the solution values of all points at the new time level. A linear system of $N + 1$ equations must be solved to advance the scheme at each time step, and it, therefore, may seem to be an inefficient method. However, as we will see later, these schemes are often efficient and, in fact, the only realistic choice.

The now familiar way of introducing a Fourier component yields, for Eq. (2.3.1),

$$(1 - i\lambda \sin \xi)\hat{v}^{n+1}(\omega) = \hat{v}^n(\omega),$$

that is,

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \quad \hat{Q} = \frac{1}{1 - i\lambda \sin \xi}.$$

Obviously, $|\hat{Q}| \leq 1$, and, again, there is damping of all frequencies except for $\omega = 0, \pi/h$. Note the important difference between Eq. (2.3.1) and the explicit schemes in Eq. (2.1.11): The stability condition $|\hat{Q}| \leq 1$ is satisfied for *all* values of λ . In other words, the scheme is stable for an arbitrary time step. Such schemes are called *unconditionally stable*. This is typical for implicit schemes.

This approximation is only first-order accurate, because the time differencing is not centered. Instead, we can use the *trapezoidal rule* for time differencing, and obtain the *Crank–Nicholson method*

$$\left(I - \frac{k}{2} D_0 \right) v_j^{n+1} = \left(I + \frac{k}{2} D_0 \right) v_j^n, \quad j = 0, 1, \dots, N. \quad (2.3.3)$$

The amplification factor is

$$\hat{Q} = \frac{2 + i\lambda \sin \xi}{2 - i\lambda \sin \xi}. \quad (2.3.4)$$

Thus $|\hat{Q}| = 1$ for all values of λ . The scheme is also unconditionally stable and, as with the leap-frog scheme, there is no damping.

The explicit and implicit approximations can be combined into the so called *θ scheme*

$$(I - \theta k D_0) v_j^{n+1} = (I + (1 - \theta) k D_0) v_j^n, \quad j = 0, 1, \dots, N. \quad (2.3.5)$$

It is *unconditionally stable for $\theta \geq 1/2$* . The parameter θ is usually chosen to be in the interval $\frac{1}{2} \leq \theta \leq 1$. The reason for introducing such an approximation is that the damping can be controlled by adjusting θ .

The system (2.3.2) is most efficiently solved by a direct method. Let the nonzero elements of the matrix A be denoted by a_{ij} , where $i = 0, 1, \dots, N$ and $j = 0, 1, \dots, N$. We make a factorization $A = LR$, where L and R have the form

$$L = \begin{bmatrix} 1 & & & & \\ \times & 1 & & & 0 \\ & \times & 1 & & \\ & & \ddots & \ddots & \\ & 0 & & \ddots & \ddots \\ & & & & \times & 1 \\ \times & \times & \dots & \dots & \times & 1 \end{bmatrix},$$

$$R = \begin{bmatrix} \times & \times & & & \times \\ & \times & \times & 0 & \times \\ & & \times & \times & \times \\ & & & \ddots & \ddots & \vdots \\ & 0 & & & \ddots & \ddots & \vdots \\ & & & & & \times & \times \\ & & & & & & \times \end{bmatrix}.$$

The nonzero elements l_{ij} and r_{ij} of L and R , respectively, are given by the recursive formulas

$$r_{00} = a_{00},$$

$$r_{01} = a_{01},$$

$$\vdots$$

$$r_{0N} = a_{0N},$$

$$\left. \begin{aligned} l_{j,j-1} &= \frac{a_{j,j-1}}{r_{j-1,j-1}} \\ r_{jj} &= a_{jj} - l_{j,j-1} r_{j-1,j} \end{aligned} \right\}, \quad j = 1, \dots, N-1,$$

$$\left. \begin{aligned} r_{j,j+1} &= a_{j,j+1} \\ &\vdots \\ r_{jN} &= -l_{j,j-1} r_{j-1,N} \end{aligned} \right\}, \quad j = 1, \dots, N-2,$$

$$r_{N-1,N} = a_{N-1,N} - l_{N-1,N-2} r_{N-2,N},$$

$$\left. \begin{aligned} l_{N0} &= \frac{a_{00}}{r_{00}} \\ I_{Nj} &= -\frac{l_{N,j-1} r_{j-1,j}}{r_{jj}} \end{aligned} \right\}, \quad j = 1, \dots, N-2,$$

$$l_{N,N-1} = \frac{1}{r_{N-1,N-1}} (a_{N,N-1} - l_{N,N-2} r_{N-2,N-1}),$$

$$r_{NN} = a_{NN} - \sum_{j=0}^{N-1} l_{Nj} r_{jN}.$$

The system (2.3.2) is rewritten as

$$LR\mathbf{v}^{n+1} = \mathbf{v}^n. \quad (2.3.7)$$

The solution is obtained by backward and forward substitution

$$\begin{aligned} L\mathbf{w} &= \mathbf{v}^n, \\ R\mathbf{v}^{n+1} &= \mathbf{w}. \end{aligned} \quad (2.3.8)$$

The number of arithmetic operations for the procedure in Eqs. (2.3.6) to (2.3.8) is proportional to N . Hence, for problems in one space dimension, the work required for the implicit method is of the same order as that for an explicit method. Note, however, that on parallel or vector computers the simpler algorithmic structure of an explicit scheme may be an advantage.

The nonzero corner elements a_{0N} and a_{N0} in the matrix A are an effect of the periodicity conditions. For other types of boundary conditions, where A is tridiagonal without the corner elements, the formulas (2.3.6) still hold and we get

$$\begin{aligned} r_{iN} &= 0, \quad i = 0, 1, \dots, N-2, \\ l_{Nj} &= 0, \quad j = 0, 1, \dots, N-2. \end{aligned}$$

For methods with more than three points on time-level t_{n+1} coupled to each other, the bandwidth ν becomes larger. The same type of solution procedure can still be applied. The matrices L and R have the same number of nonzero subdiagonals and superdiagonals, respectively, as A has, and it can be shown that $\mathcal{O}(\nu^2 N)$ arithmetic operations are required for the solution.

For problems in two space dimensions on an $N \times N$ grid, the bandwidth is $\nu = \mathcal{O}(N)$, and a direct generalization of the method above leads to an operation count of the order of N^4 . In this case iterative methods can be considerably more efficient, and they are the only realistic methods in three space dimensions.

EXERCISES

- 2.3.1. Prove that Eq. (2.3.5) is unconditionally stable for $\theta \geq \frac{1}{2}$.
- 2.3.2. Calculate the exact number of arithmetic operations required to advance by one step the implicit scheme (2.3.3). Compare it with the work required to advance by one step the explicit scheme (2.1.11).

2.3.3. Derive the direct solution algorithm for a system $A\mathbf{v} = \mathbf{b}$, where A has ν nonzero diagonals. Prove that the operation count is $\mathcal{O}(\nu^2 N)$.

2.4. TRUNCATION ERROR

In the previous sections, we have derived several difference schemes to calculate the solution u of Eq. (2.1.1). In every case, we could write their solutions v in closed form and, therefore, we could calculate the error $u - v$ explicitly. In this section, we discuss the truncation error, which is a measure of the accuracy of a given scheme. Instead of estimating the error $u - v$ we calculate how well u satisfies the difference approximation. In Chapter 6, we use the truncation error to estimate $u - v$. The advantage of this procedure is that it can be used when u and v are not known explicitly. It can also be used for equations with variable coefficients.

Let u be a smooth function. Using a Taylor series expansion around any point (x, t) we obtain

$$\begin{aligned} D_0 u(x, t) &= \frac{u(x + h, t) - u(x - h, t)}{2h} \\ &= u_x(x, t) + \frac{h^2}{3!} u_{xxx}(x, t) + \frac{h^4}{5!} \varphi_0(x, t), \\ |\varphi_0(x, t)| &\leq \max_{x-h \leq \xi \leq x+h} \left| \frac{\partial^5 u(\xi, t)}{\partial x^5} \right|, \end{aligned} \quad (2.4.1)$$

$$\begin{aligned} D_+ D_- u(x, t) &= \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2} \\ &= u_{xx}(x, t) + \frac{2h^2}{4!} u_{xxxx}(x, t) + \frac{2h^4}{6!} \varphi_1(x, t), \\ |\varphi_1(x, t)| &\leq \max_{x-h \leq \xi \leq x+h} \left| \frac{\partial^6 u(\xi, t)}{\partial x^6} \right|, \end{aligned} \quad (2.4.2)$$

$$\begin{aligned} \frac{u(x, t+k) - u(x, t)}{k} &= u_t(x, t) + \frac{k}{2} u_{tt}(x, t) + \frac{k^2}{3!} \psi_0(x, t), \\ |\psi_0(x, t)| &\leq \max_{t \leq \xi \leq t+k} \left| \frac{\partial^3 u(x, \xi)}{\partial t^3} \right|, \end{aligned} \quad (2.4.3)$$

$$\begin{aligned} \frac{u(x, t+k) - u(x, t-k)}{2k} &= u_t(x, t) + \frac{k^2}{3!} u_{ttt}(x, t) + \frac{k^4}{5!} \psi_1(x, t), \\ |\psi_1(x, t)| &\leq \max_{t-k \leq \xi \leq t+k} \left| \frac{\partial^5 u(x, \xi)}{\partial t^5} \right|, \end{aligned} \quad (2.4.4)$$

$$\begin{aligned}
& \frac{u(x, t+k) - 2u(x, t) + u(x, t-k)}{k^2} \\
&= u_{tt}(x, t) + \frac{2k^2}{4!} u_{ttt}(x, t) + \frac{2k^4}{6!} \psi_2(x, t), \\
|\psi_2(x, t)| &\leq \max_{t-k \leq \xi \leq t+k} \left| \frac{\partial^6 u(x, \xi)}{\partial t^6} \right|. \tag{2.4.5}
\end{aligned}$$

Now assume that u is a smooth solution of Eq. (2.1.1) and substitute it into the difference scheme (2.1.11). Then we obtain from Eqs. (2.4.1) to (2.4.4) and $u_t = u_x, u_{tt} = u_{xx} = u_{xx}$,

$$\begin{aligned}
\frac{u_j^{n+1} - u_j^n}{k} - D_0 u_j^n - \sigma h D_+ D_- u_j^n &= u_t(x_j, t_n) - u_x(x_j, t_n) \\
&+ \frac{k}{2} u_{tt}(x_j, t_n) - \sigma h u_{xx}(x_j, t_n) + \mathcal{O}(h^2 + k^2) \\
&= \left(\frac{k}{2} - \sigma h \right) u_{xx}(x_j, t_n) + \mathcal{O}(h^2 + k^2) =: \tau_j^n. \tag{2.4.6}
\end{aligned}$$

We call τ_j^n the *truncation error* and say that the method is **accurate of order (p, q)** if $\tau = \mathcal{O}(h^p + k^q)$. For $\sigma \neq k/(2h)$ the method above is accurate of order $(1, 1)$. For $\sigma = k/(2h)$, the Lax–Wendroff method, the order of accuracy is $(2, 2)$.

Equation (2.4.6) implies that u satisfies

$$\begin{aligned}
u_j^{n+1} &= (I + kD_0)u_j^n + \sigma kh D_+ D_- u_j^n + k\tau_j^n, \\
u_j^0 &= f_j. \tag{2.4.7}
\end{aligned}$$

Subtracting Eq. (2.1.11) from Eq. (2.4.7), we obtain, for the error $e = u - v$,

$$\boxed{
\begin{aligned}
e_j^{n+1} &= (I + kD_0)e_j^n + \sigma kh D_+ D_- e_j^n + k\tau_j^n, \\
e_j^0 &= 0.
\end{aligned}}$$

We will show that e is of the same order as τ if the approximation is stable.

One can also easily derive expressions for τ for the other methods. The leap-frog and the Crank–Nicholson method are accurate of order $(2, 2)$, whereas the backward Euler method is accurate of order $(2, 1)$. Thus, we expect that the error in time will dominate when using the backward Euler method unless the solution varies much slower in time than in space (in the truncation error the time step k is multiplied by time derivatives).

EXERCISES

- 2.4.1.** When deriving the order of accuracy, Taylor expansion around some point (x_*, t_*) is used. Prove that (x_*, t_*) can be chosen arbitrarily and, in particular, that it does not have to be a gridpoint.
- 2.4.2.** Prove that the leap-frog scheme (2.2.1) and the Crank–Nicholson scheme (2.3.3) are accurate of order (2,2). Despite the same order of accuracy, one can expect that one scheme is more accurate than the other. Why is that so?

2.5. HEAT EQUATION

In this section we consider the simplest **parabolic** model problem for heat conduction,

$$\begin{aligned} u_t &= u_{xx}, & -\infty < x < \infty, & 0 \leq t, \\ u(x, 0) &= f(x), & -\infty < x < \infty, \end{aligned} \quad (2.5.1)$$

with 2π -periodic initial data. We again use the Fourier technique to obtain the solution. The differential operator $\partial^2/\partial x^2$ corresponds to the multiplication operator $-\omega^2$ in Fourier space, and we obtain

$$\begin{aligned} \frac{\partial \hat{u}(\omega, t)}{\partial t} &= -\omega^2 \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega). \end{aligned} \quad (2.5.2)$$

The solution of Eq. (2.5.2) is

$$\hat{u}(\omega, t) = e^{-\omega^2 t} \hat{f}(\omega), \quad (2.5.3)$$

which yields

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{-\omega^2 t} e^{i\omega x} \hat{f}(\omega), \quad f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{f}(\omega). \quad (2.5.4)$$

From Parseval's relation we obtain,

$$\|u(\cdot, t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{-\omega^2 t} \hat{f}(\omega)|^2 \leq \|f(\cdot)\|^2. \quad (2.5.5)$$

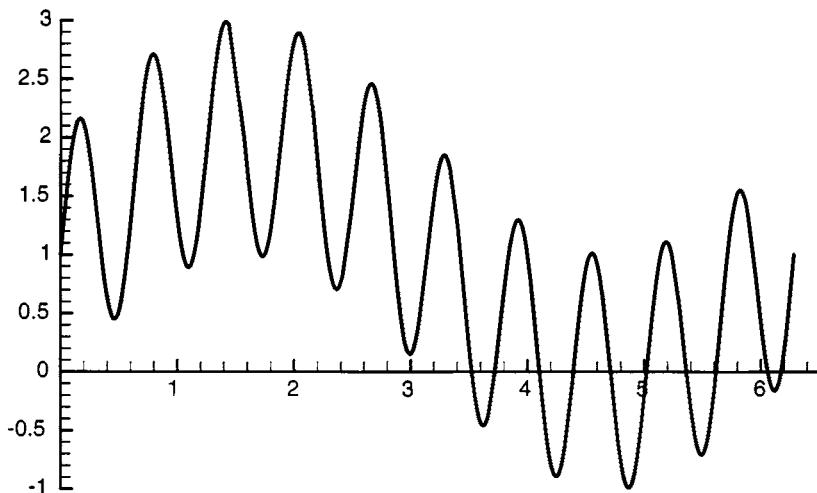


Figure 2.5.1.

Equation (2.5.3) illustrates typical parabolic behavior; each Fourier component is damped with time, and the damping is very strong for high frequencies. Even if the initial data are very rough, the solution is an analytic function for $t > 0$, that is, the Fourier coefficients decay exponentially. In Figures 2.5.1 to 2.5.3 we have plotted the solution of Eq. (2.5.1) with initial data $f(x) = 1 + \sin x + \sin 10x$ for $t = 0, 10^{-2}, 1$.

One can also show that, unlike the hyperbolic case, the speed of propagation

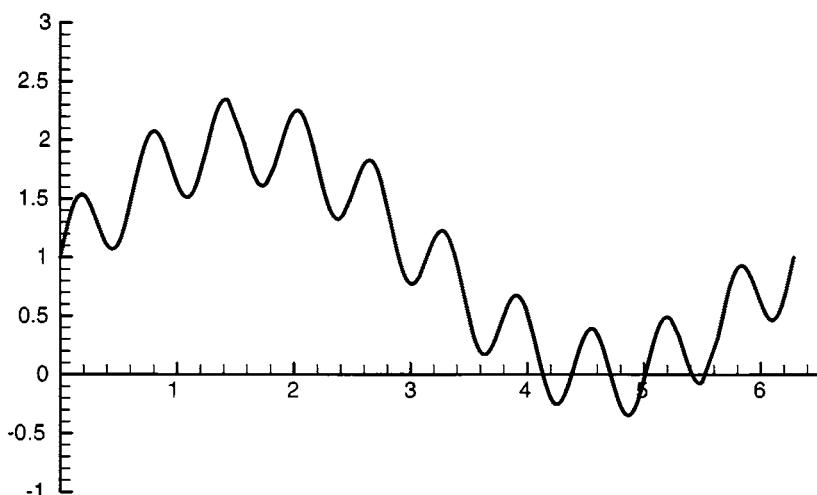


Figure 2.5.2.

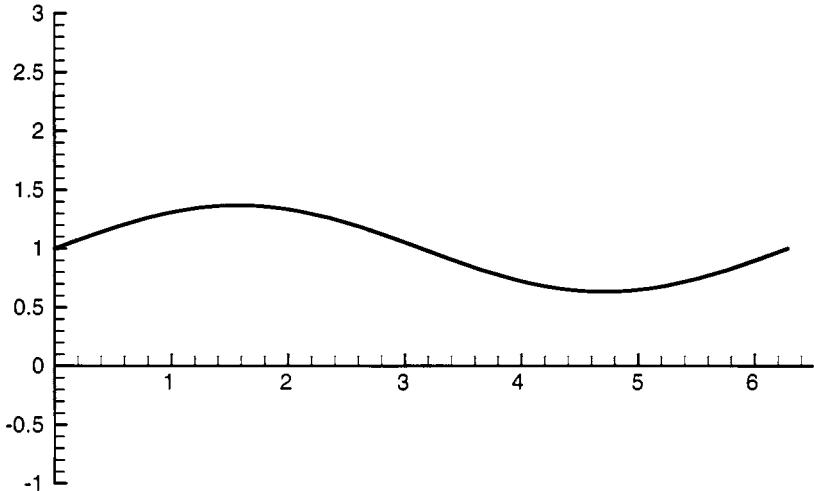


Figure 2.5.3.

is infinite. We now consider simple difference approximations of Eq. (2.5.1) and begin with

$$u_j^{n+1} = (I + kD_+D_-)u_j^n, \quad j = 0, 1, \dots, N. \quad (2.5.6)$$

The scheme is based on forward differencing in time and is often called the **Euler method**. We recall that the corresponding approximation (2.1.7) for u_t and u_x was useless, because it was unstable for any sequence $k, h \rightarrow 0$ with $k/h \geq c > 0$.

To compute the symbol \hat{Q} , we use the basic trigonometric formulas of Section 1.2. From Eq. (1.2.5)

$$kD_+D_- e^{i\omega x_j} = -4\sigma \sin^2 \frac{\xi}{2} e^{i\omega x_j}, \quad (2.5.7)$$

where $\sigma = k/h^2$, $\xi = \omega h$.

The transformed difference scheme is then

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \quad \hat{Q} = 1 - 4\sigma \sin^2 \frac{\xi}{2}. \quad (2.5.8)$$

The condition $|\hat{Q}| \leq 1$ is equivalent to

$$\sigma \leq \frac{1}{2}. \quad (2.5.9)$$

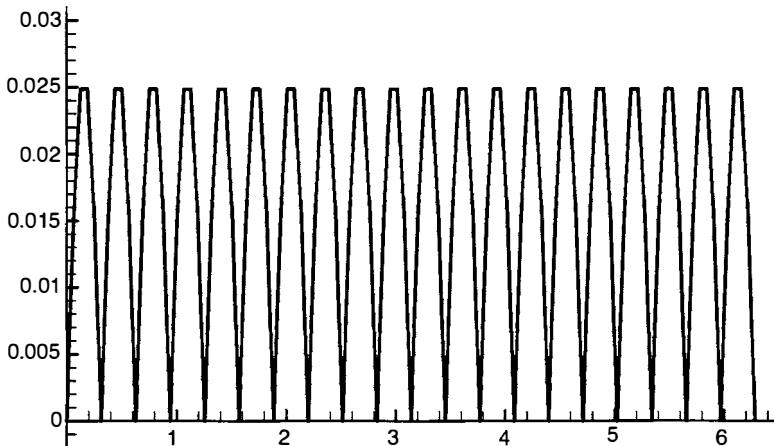


Figure 2.5.4.

We have calculated approximations of the same solution we plotted in Figures 2.5.1 to 2.5.3 using Eq. (2.5.6) with $\sigma = \frac{1}{2}$, $N = 100$. In Figures 2.5.4 and 2.5.5, we have plotted the error $u - v$, for $t = 10^{-2}, 1$.

The condition given in Eq. (2.5.9) implies that the time step k must be chosen proportional to h^2 . This is often too restrictive. On the other hand, it is natural for an explicit scheme. As noted above, there is no finite speed of propagation for parabolic problems. This means that the domain of dependence of the difference scheme must cover the whole interval in the limit $k \rightarrow 0, h \rightarrow 0$,

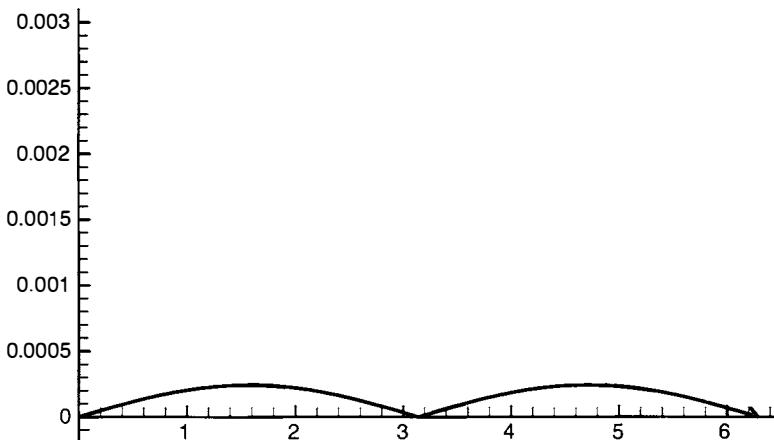


Figure 2.5.5.

even for points (\tilde{x}, \tilde{t}) arbitrarily close to the x axis, otherwise the approximation cannot converge to the true solution. Figure 2.5.6 shows the expanding domain of dependence for fixed t , decreasing h and $k = \sigma h^2$.

The leap-frog scheme approximating Eq. (2.5.1) is

$$v_j^{n+1} = 2k D_+ D_- v_j^n + v_j^{n-1}, \quad j = 0, 1, \dots, N. \quad (2.5.10)$$

The solution in Fourier space is of the form shown in Eq. (2.2.7), where z_1, z_2 are the roots of the characteristic equation

$$z^2 + 8\sigma \left(\sin^2 \frac{\xi}{2} \right) z - 1 = 0, \quad (2.5.11)$$

that is,

$$z_{1,2} = -4\sigma \sin^2 \frac{\xi}{2} \pm \sqrt{1 + \left(4\sigma \sin^2 \frac{\xi}{2} \right)^2}.$$

For $\xi \neq 0$, one of $z_{1,2}$ is larger than one in magnitude for all values of $\sigma > 0$, the scheme is useless since it is unstable for any sequence $k, h \rightarrow 0$ with $k/h^2 \geq c > 0$.

A small modification can be made to stabilize the scheme. Equation (2.5.10)

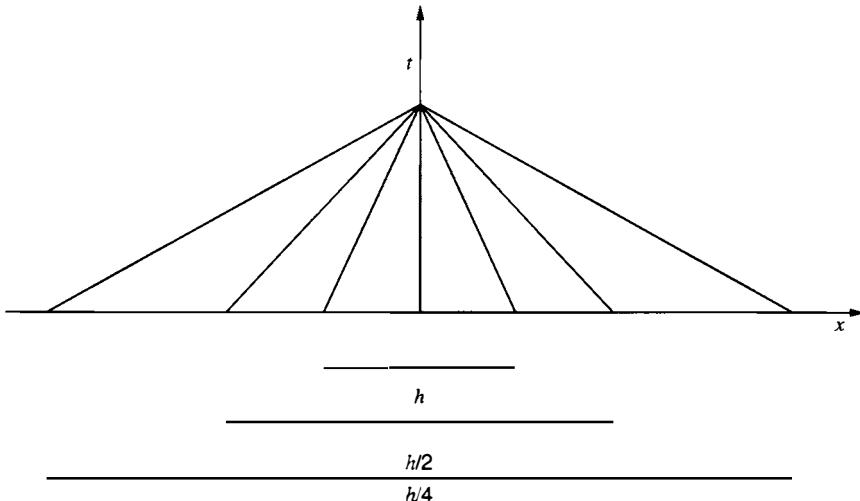


Figure 2.5.6.

can be written

$$v_j^{n+1} = 2\sigma(v_{j+1}^n - 2v_j^n + v_{j-1}^n) + v_j^{n-1},$$

and if we replace v_j^n by $(v_j^{n+1} + v_j^{n-1})/2$, then we obtain

$$\boxed{v_j^{n+1} = 2\sigma(v_{j+1}^n - v_j^{n+1} - v_j^{n-1} + v_{j-1}^n) + v_j^{n-1}, \quad j = 0, 1, \dots, N.} \quad (2.5.12)$$

This is known as the *DuFort–Frankel method*. It is still explicit, since we can solve for v_j^{n+1} and write it as

$$v_j^{n+1} = \frac{1}{1+2\sigma}(2\sigma(v_{j+1}^n + v_{j-1}^n) + (1 - 2\sigma)v_j^{n-1}).$$

Now the characteristic equation is

$$z^2 - \frac{4\sigma}{1+2\sigma}(\cos \xi)z - \frac{1-2\sigma}{1+2\sigma} = 0,$$

that is,

$$z_{1,2} = \frac{2\sigma}{1+2\sigma} \cos \xi \pm \frac{1}{1+2\sigma} \sqrt{A}, \quad (2.5.13)$$

where $A = 4\sigma^2 \cos^2 \xi + 1 - 4\sigma^2$. If $A \geq 0$, then $A \leq 1$, and

$$|z_{1,2}| \leq \frac{2\sigma}{1+2\sigma} + \frac{1}{1+2\sigma} = 1.$$

If $A < 0$, then we write

$$z_{1,2} = \frac{1}{1+2\sigma} (2\sigma \cos \xi \pm i \sqrt{4\sigma^2(1 - \cos^2 \xi) - 1}), \quad (2.5.14)$$

and get

$$|z_{1,2}|^2 = \frac{4\sigma^2 - 1}{(1+2\sigma)^2} = \frac{2\sigma - 1}{2\sigma + 1} < 1.$$

The DuFort–Frankel approximation is unconditionally stable, which is somewhat surprising, since it is explicit. The time step can be chosen independent of the space step. This seems to contradict the conclusion above that the domain of dependence must expand as h decreases. However, this apparently contradictory behavior is an illustration of the fact that **stability is only a necessary condition for convergence**. It does not guarantee that solutions are accurate approximations. It only guarantees that solutions remain bounded.

To investigate the order of accuracy we calculate the truncation error. From Eqs. (2.4.1) to (2.4.6),

$$\begin{aligned}\tau &= \frac{u_j^{n+1} - u_j^{n-1}}{2k} - D_+ D_- u_j^n + \frac{k^2}{h^2} \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{k^2} \\ &= u_t - u_{xx} + \frac{k^2}{h^2} u_{tt} + \mathcal{O}\left(k^2 + h^2 + \frac{k^4}{h^2}\right) \\ &= \frac{k^2}{h^2} u_{tt} + \mathcal{O}\left(k^2 + h^2 + \frac{k^4}{h^2}\right).\end{aligned}\quad (2.5.15)$$

Thus, $\lim_{k,h \rightarrow 0} \tau = 0$ only if $\lim_{k,h \rightarrow 0} k/h = 0$. Typically, one chooses

$$k = ch^{1+\delta}, \quad \delta > 0. \quad (2.5.16)$$

Then the **truncation error is $\mathcal{O}(h^{2\delta})$** , and the method is only accurate of order (2,2) if $\delta = 1$, i.e., $k = O(h^2)$, essentially the same restriction as that required for the explicit scheme.

We now examine analogues of the implicit schemes introduced in the previous section. The **backward Euler** approximation is

$$(I - kD_+ D_-)v_j^{n+1} = v_j^n, \quad j = 0, 1, \dots, N, \quad (2.5.17)$$

with the amplification factor

$$\hat{Q} = \frac{1}{1 + 4\sigma \sin^2 \frac{\xi}{2}}, \quad \sigma = \frac{k}{h^2}. \quad (2.5.18)$$

The magnitude of \hat{Q} is never greater than one independent of σ , and all nonzero frequencies are damped. Note that, as for the differential equation, the damping is stronger for larger ω .

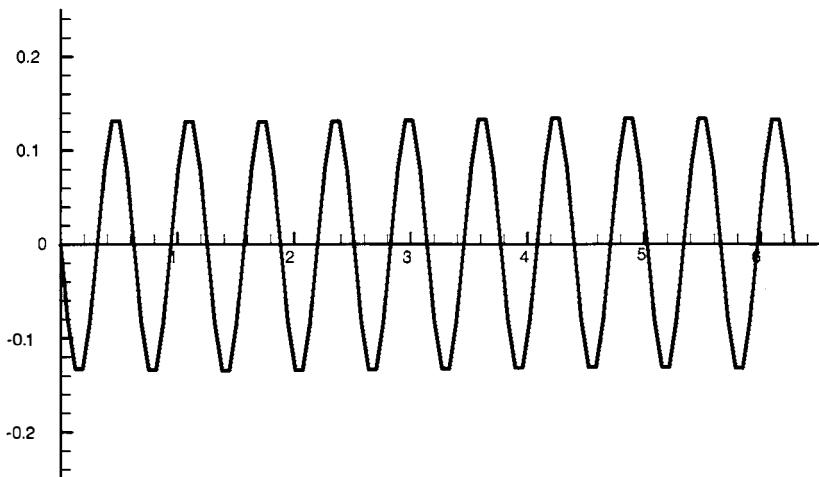


Figure 2.5.7.

In Figures 2.5.7 and 2.5.8, we show the error of backward Euler calculations with $k = h$ and $N = 100$ for $t = 10^{-2}, 1$, respectively. The initial data are the same as in Figure 2.5.1.

The Crank–Nicholson scheme

$$\left(I - \frac{k}{2} D_+ D_- \right) v_j^{n+1} = \left(I + \frac{k}{2} D_+ D_- \right) v_j^n, \quad j = 0, 1, \dots, N \quad (2.5.19)$$

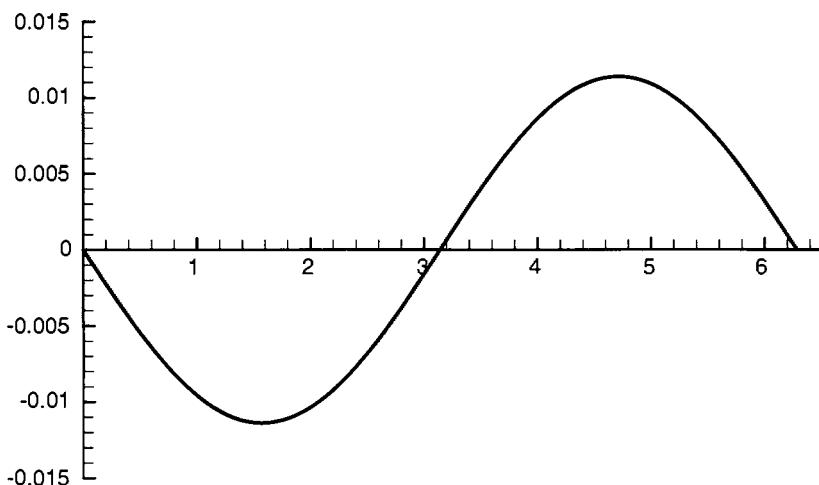


Figure 2.5.8.

has the amplification factor

$$\hat{Q} = \frac{1 - 2\sigma \sin^2 \frac{\xi}{2}}{1 + 2\sigma \sin^2 \frac{\xi}{2}}, \quad \sigma = \frac{k}{h^2}, \quad (2.5.20)$$

and, like the backward Euler method, it is unconditionally stable. However, when σ is large, \hat{Q} is near -1 for $\xi \neq 0$, and there is very little damping. This is a serious drawback because one would like to use time steps of the same order as the space step. With that choice, we get $\sigma = \mathcal{O}(1/h)$, and $\hat{Q} \rightarrow -1$ as $h \rightarrow 0$ for every fixed ξ (i.e., as $\omega \rightarrow \infty$).

We have calculated the approximate solution of Eq. (2.5.1) with the same initial data as before using the Crank–Nicholson method with $k = h$ and $N = 100$ for $t = 10^{-2}, 1$. The error is plotted in Figures 2.5.9 and 2.5.10. There is now an oscillating error (see Exercise 2.5.3).

We can also combine the two implicit schemes for parabolic equations obtaining the θ scheme

$$(I - \theta k D_+ D_-) v_j^{n+1} = (I + (1 - \theta) k D_+ D_-) v_j^n, \quad j = 0, 1, \dots, N, \quad 0 \leq \theta \leq 1, \quad (2.5.21)$$

which is unconditionally stable for $\theta \geq \frac{1}{2}$ (see Exercise 2.5.2). As in the hyperbolic case, the damping increases with θ up to $\theta = 1$ (backward Euler), but the accuracy decreases.

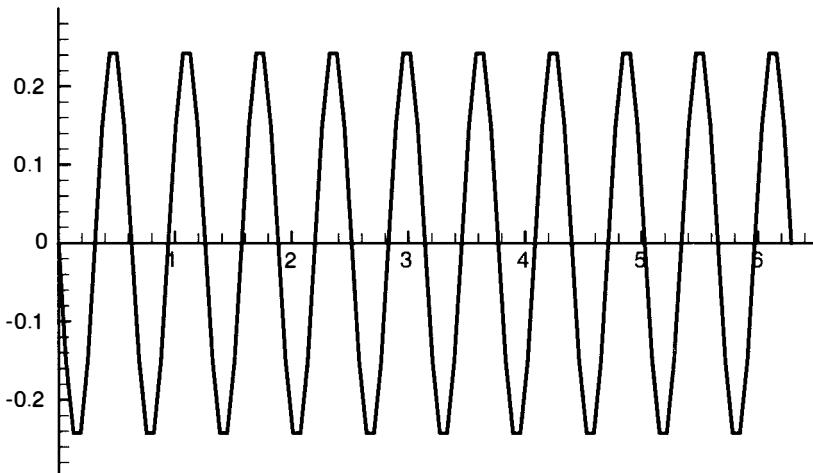


Figure 2.5.9.

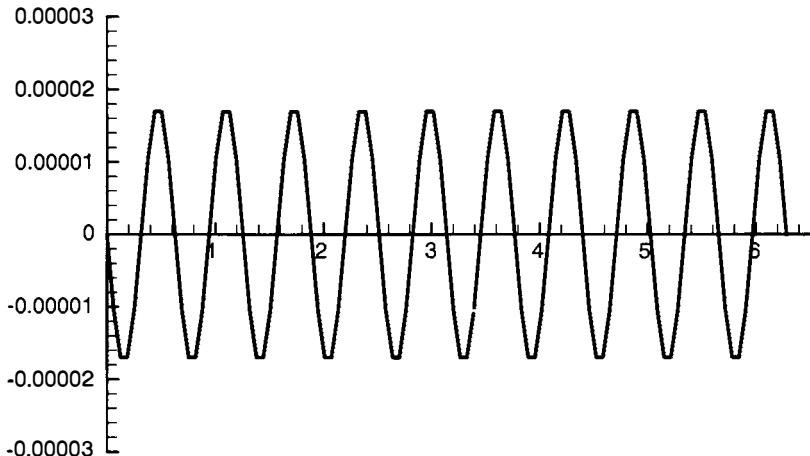


Figure 2.5.10.

EXERCISES

- 2.5.1.** Assume that the initial data for Eq. (2.5.1) is a simple wave $f(x) = e^{i\omega x}$. Determine the time t_1 , where $\|u(\cdot, t_1)\| = 10^{-6}$. Apply the Euler method from Eq. (2.5.6), and calculate the corresponding time t_2 . Determine the optimal time step for a given h .
- 2.5.2.** Prove that the θ scheme from Eq. (2.5.21), is unconditionally stable for $\theta \geq \frac{1}{2}$.
- 2.5.3.** Derive the truncation error for the backward Euler and the Crank–Nicholson methods applied to $u_t = u_{xx}$. Prove that it is $\mathcal{O}(h^2 + k)$ and $\mathcal{O}(h^2 + k^2)$, respectively. Despite this fact, at certain times the backward Euler method is more accurate for the example computed in this section. Explain this paradox.

2.6. CONVECTION–DIFFUSION EQUATION

In many applications, the differential equations have both first- and second-order derivatives in space. We now consider the model problem for convection–diffusion

$$\begin{aligned} u_t + au_x &= \eta u_{xx}, & -\infty < x < \infty, \quad 0 \leq t, \quad y = \text{constant} > 0, \\ u(x, 0) &= f(x), & -\infty < x < \infty, \end{aligned} \tag{2.6.1}$$

with 2π -periodic initial data. In Fourier space, the corresponding problem is

$$\begin{aligned}\frac{\partial \hat{u}(\omega, t)}{\partial t} + i\omega \hat{u}(\omega, t) &= -\eta \omega^2 \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega),\end{aligned}\quad (2.6.2)$$

with solutions

$$\hat{u}(\omega, t) = e^{-(i\omega + \eta \omega^2)t} \hat{f}(\omega). \quad (2.6.3)$$

Consider the difference approximation

$$v_j^{n+1} = v_j^n + k(\eta D_+ D_- - a D_0) v_j^n, \quad j = 0, 1, \dots, N. \quad (2.6.4)$$

The amplification factor is

$$\hat{Q} = 1 - 2\alpha \sin^2 \frac{\xi}{2} - i\lambda \sin \xi, \quad \alpha = \frac{2\eta k}{h^2}, \quad \lambda = \frac{ak}{h}. \quad (2.6.5)$$

The “parabolic” stability condition for the case $a = 0$ is

$$\frac{k\eta}{h^2} \leq \frac{1}{2}, \quad \text{that is, } \alpha \leq 1. \quad (2.6.6)$$

For $a \neq 0$ we have

$$\begin{aligned}|\hat{Q}|^2 &= 1 - 4\alpha \sin^2 \frac{\xi}{2} + 4\alpha^2 \sin^4 \frac{\xi}{2} + 4\lambda^2 \sin^2 \frac{\xi}{2} \left(1 - \sin^2 \frac{\xi}{2}\right), \\ &= 1 - 4(\lambda^2 - \alpha^2)s^2 + 4(\lambda^2 - \alpha)s,\end{aligned}\quad (2.6.7)$$

where $s = \sin^2 \xi/2$. Thus, $|\hat{Q}| \leq 1$ for all ξ if, and only if,

$$\phi(s) := -(\lambda^2 - \alpha^2)s + \lambda^2 - \alpha \leq 0, \quad 0 \leq s \leq 1. \quad (2.6.8)$$

$\phi(s)$ is a linear function of s and, therefore, Eq. (2.6.8) holds if and only if

$$\phi(0) = \lambda^2 - \alpha \leq 0, \quad \phi(1) = \alpha^2 - \alpha \leq 0,$$

that is,

$$\lambda^2 \leq \alpha \leq 1 \quad \text{or} \quad a^2 k \leq 2\eta \leq h^2/k. \quad (2.6.9)$$

The conditions in Eq. (2.6.9) can be interpreted in this way: The parabolic term makes it possible to stabilize the approximation of the hyperbolic part. However, the coefficient η must be large enough compared to a (or k small enough) in order to provide enough damping. Furthermore, the damping of the method is always less than that of the differential equation. The true parabolic decay rate for Eq. (2.6.1) is not preserved by the approximation. Part of it is required to stabilize the hyperbolic part. As η becomes small, this becomes more severe.

In the previous section, it was noted that, for parabolic problems, the stability restriction on the time step for explicit schemes (except the DuFort–Frankel method) is often too severe, and implicit approximations should be used. Implicit methods can also be used when first-order derivatives are present. For example, the **Crank–Nicholson method**

$$\begin{aligned} & \left(I + \frac{k}{2} (aD_0 - \eta D_+ D_-) \right) v_j^{n+1} \\ &= \left(I - \frac{k}{2} (aD_0 - \eta D_+ D_-) \right) v_j^n, \quad j = 0, 1, \dots, N \end{aligned} \quad (2.6.10)$$

is unconditionally stable. In applications, however, the hyperbolic part is often nonlinear, that is, $a = a(u)$, and a nonlinear system of equations must be solved at each step. In this case, it is convenient to use a so called **semi-implicit** method. The simplest approximation of this kind for our problem is

$$(I - k\eta D_+ D_-) v_j^{n+1} = -2kaD_0 v_j^n + (I + k\eta D_+ D_-) v_j^{n-1}, \quad j = 0, 1, \dots, N, \quad (2.6.11)$$

which is a combination of the **leap-frog** approximation and the **Crank–Nicholson approximation**. v_j^1 must be computed by some other one-step method. In Fourier space, Eq. (2.6.11) yields

$$\begin{aligned} & \left(1 + 4\sigma^2 \sin^2 \frac{\xi}{2} \right) \hat{v}^{n+1}(\omega) \\ &= -i2\lambda (\sin \xi) \hat{v}^n(\omega) + \left(1 - 4\sigma \sin^2 \frac{\xi}{2} \right) \hat{v}^{n-1}(\omega), \\ & \sigma = \frac{k\eta}{h^2}, \quad \lambda = \frac{ak}{h}. \end{aligned} \quad (2.6.12)$$

The corresponding characteristic equation is

$$z^2 + \frac{i2\lambda \sin \xi}{1 + \beta} z - \frac{1 - \beta}{1 + \beta} = 0, \quad \beta = 4\sigma \sin^2 \frac{\xi}{2}, \quad (2.6.13)$$

with solutions

$$z_{1,2} = \frac{-i\lambda \sin \xi \pm \sqrt{1 - \beta^2 - \lambda^2 \sin^2 \xi}}{1 + \beta}. \quad (2.6.14)$$

First, assume that the square root is real, that is,

$$\beta^2 + \lambda^2 \sin^2 \xi \leq 1. \quad (2.6.15)$$

Then,

$$|z_{1,2}|^2 = \frac{1 - \beta^2}{(1 + \beta)^2} = \frac{1 - \beta}{1 + \beta} \leq 1.$$

Next, assume that

$$\beta^2 + \lambda^2 \sin^2 \xi > 1. \quad (2.6.16)$$

Then the roots are purely imaginary, and we have, for $|\lambda| < 1$,

$$|z_{1,2}| = \left| \frac{\lambda \sin \xi \pm \sqrt{\beta^2 + \lambda^2 \sin^2 \xi - 1}}{1 + \beta} \right| \leq \frac{|\lambda| + \beta}{1 + \beta} < 1. \quad (2.6.17)$$

Thus $|z_{1,2}| \leq 1$ for $|\lambda| \leq 1$, which is the same stability condition we obtained for the leap-frog approximation in Eq. (2.2.1) of $u_t = u_x$. Note, however, that $z_1 = z_2$ for $\beta^2 + \lambda^2 \sin^2 \xi = 1$. Then the representation $\hat{v}^n(\omega) = \sigma_1 z_1^n + \sigma_2 z_2^n$ becomes $\hat{v}^n(\omega) = (\sigma_1 + \sigma_2 n)z_1^n$. Because $|z_1| \leq |\lambda| < 1$ in this case, we have $n|z_1|^n \leq \text{constant independent of } n$. Thus, $|\hat{v}^n(\omega)|$ is bounded independent of ω, n , and it is stable. We shall consider the approximation in Eq. (2.6.11) in a more general setting in Chapter 5.

The time step can be chosen to be of the same order as the space step with the semi-implicit scheme (2.6.11), which is a substantial gain in efficiency compared to an explicit scheme. This was achieved without involving the whole difference operator at the new time level.

EXERCISES

- 2.6.1.** Write a program that computes the solutions to Eq. (2.6.4) for $N = 10, 20, 40, \dots$. Choose the time step such that

1. α , defined in (2.6.5), is a constant with $\alpha \leq 1$,
2. λ , defined in (2.6.5), is a constant with $|\lambda| \leq 1$.

Compare the solutions and explain the difference in their behavior.

2.6.2. Newton's method for a nonlinear system $\mathbf{F}(\mathbf{v}) = 0$ is defined by

$$\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)} - \mathbf{F}'(\mathbf{v}^{(n)})^{-1} \mathbf{F}(\mathbf{v}^{(n)})$$

where \mathbf{F}' is the Jacobian matrix of \mathbf{F} with respect to \mathbf{v} . Assume that the coefficients a and η in Eq. (2.6.1) depend on u . Prove that Newton's method applied to each step of the Crank–Nicholson scheme (2.6.10) leads to linear systems of the same structure as discussed in Section 2.3.

2.7. HIGHER ORDER EQUATIONS

In this section, we briefly discuss differential equations of the form

$$\frac{\partial u}{\partial t} = a \frac{\partial^p u}{\partial x^p}, \quad p \geq 1, \quad -\infty < x < \infty, \quad t \geq 0, \quad (2.7.1a)$$

$$u(x, 0) = f(x), \quad -\infty < x < \infty, \quad (2.7.1b)$$

where a is a complex number and f is 2π -periodic. In Fourier space, Eq. (2.7.1) becomes

$$\frac{\partial \hat{u}(\omega, t)}{\partial t} = a(i\omega)^p \hat{u}(\omega, t), \quad (2.7.2a)$$

$$\hat{u}(\omega, 0) = \hat{f}(\omega), \quad (2.7.2b)$$

that is,

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

with

$$|\hat{u}(\omega, t)| = |e^{\operatorname{Re}[a(i\omega)^p] t} \hat{f}(\omega)|. \quad (2.7.3)$$

For the problem to be well posed, it is sufficient that the condition

$$\operatorname{Re}[a(i\omega)^p] \leq 0 \quad (2.7.4)$$

be fulfilled for all ω . This ensures that the solution will satisfy the estimate

$$\|u(\cdot, t)\| \leq \|f(\cdot)\|. \quad (2.7.5)$$

Since ω is real and can be positive or negative, we obtain the condition

$$\text{sign}(\operatorname{Re} a) = (-1)^{p/2+1}, \quad \begin{aligned} &\text{if } \operatorname{Re} a \neq 0 \text{ and } p \text{ is even,} \\ &\text{Im } a = 0, \quad \text{if } p \text{ is odd.} \end{aligned} \quad (2.7.6a)$$

$$(2.7.6b)$$

The most natural centered difference approximation is given by

$$\frac{\partial^p}{\partial x^p} \rightarrow Q_p = \begin{cases} (D_+ D_-)^{p/2}, & p \text{ even} \\ D_0 (D_+ D_-)^{(p-1)/2}, & p \text{ odd,} \end{cases} \quad (2.7.7)$$

which, in Fourier space, yields

$$(i\omega)^p \rightarrow \hat{Q}_p = \begin{cases} \left(-\frac{4}{h^2} \sin^2 \frac{\omega h}{2} \right)^{p/2}, & p \text{ even,} \\ \frac{i}{h} \sin(\omega h) \left(-\frac{4}{h^2} \sin^2 \frac{\omega h}{2} \right)^{(p-1)/2}, & p \text{ odd.} \end{cases} \quad (2.7.8)$$

If a is real the Euler method can always be used if p is even. The leap-frog scheme can always be used if p is odd. This follows directly from the calculations made in Sections 2.2 and 2.5. For the Euler method, we have

$$\hat{Q} = 1 + ka\hat{Q}_p, \quad (2.7.9)$$

where \hat{Q}_p is defined in Eq. (2.7.8). If a is real, stability requires that the condition

$$(-1)^{p/2-1} \cdot \frac{4^{p/2} ak}{h^p} \leq 2, \quad p \text{ even,} \quad (2.7.10)$$

be satisfied. For $p \geq 4$, the time step restriction is so severe that the method cannot be used in any realistic computation. Similarly, for the leap-frog scheme, we obtain a condition of the form

$$\frac{k}{h^p} \leq \text{constant}, \quad p \text{ odd.} \quad (2.7.11)$$

[This is easily seen if we follow the calculations leading to Eq. (2.2.9).] $p = 1$ is the practical limit in several space dimensions, and we conclude that implicit

methods are necessary for higher order equations. (In one space dimension, one could possibly use explicit methods for $p = 2, 3$).

EXERCISES

2.7.1. What explicit method could be used for the Schrödinger type equation

$$u_t = iu_{xx} ? \quad (2.7.12)$$

Derive the stability condition.

2.7.2. Define the Crank–Nicholson approximation for the Korteweg de Vries type equation

$$u_t = u_{xxx} + au_x. \quad (2.7.13)$$

Prove unconditional stability.

2.7.3. Define a semi-implicit approximation suitable for the efficient solution of Eq. (2.7.13) with $a = a(u)$. Derive the stability condition (for $a = \text{constant}$).

2.8. GENERALIZATION TO SEVERAL SPACE DIMENSIONS

In two space dimensions the hyperbolic model problem becomes

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2}, \quad -\infty < x_1, x_2 < \infty, \quad t \geq 0, \quad (2.8.1)$$

with initial data

$$u(x, 0) = f(x), \quad x = (x_1, x_2), \quad -\infty < x_1, x_2 < \infty,$$

which is 2π -periodic in x_1 and x_2 . If

$$f(x) = \frac{1}{2\pi} e^{i\langle \omega, x \rangle} \hat{f}(\omega), \quad \omega = (\omega_1, \omega_2),$$

we make the ansatz

$$u = \frac{1}{2\pi} e^{i\langle \omega, x \rangle} \hat{u}(\omega, t), \quad \hat{u}(\omega, 0) = \hat{f}(\omega)$$

and obtain

$$\hat{u}_t(\omega, t) = i(\omega_1 + \omega_2)\hat{u}(\omega, t).$$

Thus,

$$u(x, t) = \frac{1}{2\pi} e^{i(\omega_1 + \omega_2)t} e^{i\langle \omega, x \rangle} \hat{f}(\omega)$$

is the solution of our problem. For general f , we obtain, by the principle of superposition,

$$u = \frac{1}{2\pi} \sum_{\omega} e^{i(\omega_1 + \omega_2)t} e^{i\langle \omega, x \rangle} \hat{f}(\omega) = f(x_1 + t, x_2 + t). \quad (2.8.2)$$

Thus, we can solve the problem as we did in the one-dimensional case. Also, the solution is constant along the characteristics, which are the lines $x_1 + t = \text{constant}$ and $x_2 + t = \text{constant}$.

We now discuss difference approximations. We introduce a time step, $k > 0$, and a two-dimensional grid by

$$x_j = (j_1 h, j_2 h), \quad j_\nu = 0, \pm 1, \pm 2, \dots, h = 2\pi/(N + 1),$$

and gridfunctions by

$$v_j^n = v(x_j, t_n), \quad t_n = nk.$$

Corresponding to Eq. (2.1.7), we now have

$$v_j^{n+1} = (I + k(D_{0x_1} + D_{0x_2}))v_j^n. \quad (2.8.3)$$

Its Fourier transform is

$$\hat{v}^{n+1}(\omega) = (1 + i\lambda(\sin \xi_1 + \sin \xi_2))\hat{v}^n(\omega). \quad (2.8.4)$$

It is of the same form as Eq. (2.1.9). Therefore, the approximation is not useful. We add artificial viscosity, that is, we consider

$$v_j^{n+1} = (I + k(D_{0x_1} + D_{0x_2}) + \sigma kh(D_{+x_1}D_{-x_1} + D_{+x_2}D_{-x_2}))v_j^n. \quad (2.8.5)$$

As before, we can choose $\lambda = k/h$, $\sigma > 0$ such that $|\hat{Q}| \leq 1$; that is, the approximation is stable.

The leap-frog and the Crank–Nicholson approximations are also easily generalized. They are

$$v_j^{n+1} = v_j^{n-1} + 2k(D_{0x_1} + D_{0x_2})v_j^n, \quad (2.8.6a)$$

$$\left(I - \frac{k}{2} (D_{0x_1} + D_{0x_2}) \right) v_j^{n+1} = \left(I + \frac{k}{2} (D_{0x_1} + D_{0x_2}) \right) v_j^n, \quad (2.8.6b)$$

respectively. Equation (2.8.6a) is stable for $k/h < 1/2$, whereas Eq. (2.8.6b) is stable for all values of $\lambda = k/h$. Both methods are second-order accurate.

The parabolic model equation is

$$\begin{aligned} u_t &= u_{x_1 x_1} + u_{x_2 x_2}, \\ u(x, 0) &= f(x). \end{aligned}$$

Like the one-dimensional problem, its solution

$$u = \frac{1}{2\pi} \sum_{\omega} e^{-|\omega|^2 t} e^{i\langle \omega, x \rangle} \hat{f}(\omega)$$

becomes “smoother” with time because the highly oscillatory waves ($|\omega| \gg 1$) are rapidly damped. We can easily construct difference approximations analogous to those used for the one-dimensional problem. We need only replace $D_+ D_-$ in Section 2.5 by $D_{+x_1} D_{-x_1} + D_{+x_2} D_{-x_2}$. The analysis proceeds as before. The explicit Euler method in Eq. (2.5.6) is stable for $\sigma = k/h^2 \leq \frac{1}{4}$, whereas the implicit Euler, Crank–Nicholson, and DuFort–Frankel methods are unconditionally stable.

EXERCISES

- 2.8.1.** Derive the stability condition for the leap-frog approximation to $u_t = u_x + u_y + u_z$, where the step sizes Δx , Δy , Δz may be different.
- 2.8.2.** Derive the stability condition for the Euler approximation to $u_t = u_{xx} + u_{yy} + u_{zz}$. Prove that the DuFort–Frankel method is unconditionally stable for the same equation.

BIBLIOGRAPHIC NOTES

Most of the difference schemes introduced in this chapter were developed very early, in several cases before the electronic computer was invented. The leap-frog scheme was discussed in a classical paper by *Courant-Friedrichs-Levy* (1928). In the same paper, the so called *CFL-condition* was introduced, that is, the domain of dependence of the difference scheme must include the domain of

dependence of the differential equation. Today one often uses the term “CFL-number” which, for the model equation $u_t = au_x$, means $\lambda = ka/h$.

The Lax–Friedrichs scheme was introduced for conservation laws $u_t = F(u)_x$ by Lax (1954), and the Lax–Wendroff method was presented in its original form by Lax and Wendroff (1960). Various versions have been presented later, but for the simple model equations we have been considering so far, they are identical. Any approximation of a hyperbolic equation that is a one-step explicit scheme with a centered second order accurate approximation complemented with a damping term of second order, is usually called a Lax–Wendroff type approximation.

The Crank–Nicholson approximation was initially constructed for parabolic heat conduction problems by Crank and Nicholson (1947). The same name has later been used for other types of equations, where centered difference operators are used in space and the trapezoidal rule is used for discretization in time. The DuFort–Frankel method for parabolic problems was introduced by DuFort and Frankel (1953).

Stability analysis based on Fourier modes as presented here goes back to von Neumann, who used it at Los Alamos National Laboratory during World War II. It was first published by Crank and Nicholson (1947) and later by Charney, Fjortoft, and von Neumann (1950), von Neumann and Richtmyer (1950) and O’Brien, Hyman, and Kaplan (1951).

In this book we use Fourier series representations of periodic functions, but one could as well use Fourier integral representations of general L_2 functions:

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x} d\omega,$$

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\omega x} dx,$$

see, for example, Richtmyer and Morton (1967). The gridfunctions can be extended such that they are defined everywhere if the initial function is defined everywhere. In that way, the Fourier integrals are also defined for the solutions to the difference approximations. The Fourier transformed equations are exactly the same as they are for Fourier series and, accordingly, the stability conditions derived in Fourier space will be identical.

The stability condition (2.1.10) is the one that corresponds to the general stability definition, which will be given in Chapter 5. We notice that the condition $|\hat{Q}| \leq 1$ used in all our examples is stronger, since $|\hat{Q}| \leq 1 + \mathcal{O}(k)$ is sufficient for Eq. (2.1.10) to hold. However, if k/h is kept constant, our approximations are not explicitly dependent on k for the hyperbolic model equation. The same conclusion holds for the parabolic model problem if k/h^2 is constant. Therefore, $|\hat{Q}| \leq 1$ is the only possibility for a stable scheme. On the other hand, for the equation $u_t = u_x + u$, we must necessarily have $|\hat{Q}| = 1 + \mathcal{O}(k)$, and this motivates the more general stability condition.

3

HIGHER ORDER ACCURACY

3.1. EFFICIENCY OF HIGHER ORDER ACCURATE DIFFERENCE APPROXIMATIONS

In this section, we develop higher order accurate difference approximations and compare the efficiency of different methods. To illustrate the basic ideas, we first consider discretizations in space only. Later the results for fully discretized methods will be summarized.

We begin again with the model problem

$$u_t + au_x = 0, \quad u(x, 0) = e^{i\omega x}, \quad (3.1.1)$$

For simplicity we assume that $a > 0$. As in Section 2.1, the solution is easily computed, and we have

$$u(x, t) = e^{i\omega(x - at)}, \quad (3.1.2)$$

that is, the wave propagates with speed a to the right. The simplest centered difference approximation is

$$\frac{dv_j(t)}{dt} + aD_0v_j(t) = 0, \quad j = 0, 1, \dots, N, \\ v_j(0) = e^{i\omega x_j}. \quad (3.1.3)$$

The method of discretizing only the spatial variables is often called the **method of lines**. It yields a set of ordinary differential equations (ODEs) for the grid-values. In general, one has to solve the ODEs numerically. In the method of lines, usually a standard ODE solver is used.

The solution of Eq. (3.1.3) is

$$v_j(t) = e^{i\omega(x_j - a_2 t)}, \quad a_2 = \frac{a \sin \xi}{\xi}. \quad (3.1.4)$$

(The subscript 2 is used here to indicate that the difference operator is second-order accurate.) Therefore, the error $e_2 = \|u - v\|_\infty := \max_{0 \leq j \leq N} |u(x_j) - v_j|$ satisfies

$$\begin{aligned} e_2 &= \|e^{i\omega(x_j - at)} - e^{i\omega(x_j - a_2 t)}\|_\infty = \|e^{-i\omega at} - e^{-i\omega a_2 t}\|_\infty, \\ &= \omega t a \left(1 - \frac{\sin \xi}{\xi}\right) + \mathcal{O}(\xi^4) = \frac{\omega t a \xi^2}{6} + \mathcal{O}(\xi^4), \quad \xi = \omega h, \end{aligned} \quad (3.1.5)$$

where, for convenience, we have assumed that $\omega \geq 0$. Using results from Section 1.3, we can represent a general gridfunction by its interpolating polynomial. The largest wave number that can be represented on the grid is $\omega = N/2$. Therefore we consider simple waves with $0 < \omega \leq N/2$. If ω is much less than $N/2$ in magnitude, then there are many gridpoints per wavelength, and the solution is well represented by the gridfunction $v_j(t)$. In this case, ξ is a small number, and we can neglect $\mathcal{O}(\xi^4)$ and obtain

$$e_2 \approx \frac{\omega t a \xi^2}{6}. \quad (3.1.6)$$

If ξ is large, corresponding to fewer gridpoints per wavelength in the solution, then the difference operator D_0 will not yield a good approximation of $\partial/\partial x$ and approximations Q_p of order $p > 2$ are required.

It is convenient to represent Q_p as a perturbation of D_0 . We make the ansatz

$$Q_p = D_0 \sum_{\nu=0}^{p/2-1} (-1)^\nu \alpha_\nu (h^2 D_+ D_-)^\nu. \quad (3.1.7)$$

To determine the coefficients, we apply the operator to smooth functions $u(x)$. It is sufficient to apply Q_p to functions $e^{i\omega x}$, because general 2π -periodic functions can be written as a linear combination of these functions. We obtain

$$Q_p e^{i\omega x} = \frac{i}{h} \sin \xi \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} \left(\sin \frac{\xi}{2}\right)^{2\nu} e^{i\omega x}, \quad \xi = \omega h.$$

If Eq. (3.1.7) is accurate of order p , then

$$i\omega + \mathcal{O}(\omega^{p+1} h^p) = i\omega \frac{\sin \xi}{\xi} \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} \left(\sin \frac{\xi}{2}\right)^{2\nu}.$$

[cf. Eqs. (1.2.3) and (1.2.4)].

The substitution $\theta = \sin(\xi/2)$ yields

$$\frac{\arcsin \theta}{\sqrt{1 - \theta^2}} = \theta \sum_{\nu=0}^{p/2-1} \alpha_\nu 2^{2\nu} \theta^{2\nu} + \mathcal{O}(\theta^{p+1}).$$

By expanding the left-hand side in a Taylor series, the α_ν are uniquely determined (see Exercise 3.1.5). Clearly Q_p is accurate for order p when the α_ν are determined this way. We formulate the result in Lemma 3.1.1.

Lemma 3.1.1. *The centered difference operator Q_p , which approximates $\partial/\partial x$ with accuracy of order p , has the form (3.1.7) with*

$$\begin{aligned}\alpha_0 &= 1, \\ \alpha_\nu &= \frac{\nu}{4\nu+2} \alpha_{\nu-1}, \quad \nu = 1, 2, \dots, p/2-1.\end{aligned}$$

The fourth- and sixth-order accurate operators are

$$Q_4 = D_0 \left(I - \frac{h^2}{6} D_+ D_- \right), \quad (3.1.8)$$

$$Q_6 = D_0 \left(I - \frac{h^2}{6} D_+ D_- + \frac{h^4}{30} D_+^2 D_-^2 \right). \quad (3.1.9)$$

The form of these operators could, of course, be obtained by using Taylor series expansions directly in physical space. For example,

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

Thus, if the leading error term is eliminated by using a second-order accurate approximation of u_{xxx} , then a fourth-order approximation is obtained. The natural centered approximation is

$$\frac{\partial^3}{\partial x^3} \rightarrow D_0 D_+ D_- \quad (3.1.10)$$

and a Taylor series expansion shows that

$$D_0 D_+ D_- u(x, t) = u_{xxx}(x, t) + \mathcal{O}(h^2). \quad (3.1.11)$$

This procedure leads to the operator Q_4 in Eq. (3.1.8).

We now give precise comparisons of the operators Q_2 , Q_4 , and Q_6 . The error for Q_2 was given in Eq. (3.1.5). For Q_4 and Q_6 , we obtain, correspondingly,

$$e_4 = \omega t a \left[1 - \frac{\sin \xi}{\xi} \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} \right) \right] + \mathcal{O}(\xi^6), \quad (3.1.12a)$$

$$\begin{aligned} e_6 = \omega t a & \left[1 - \frac{\sin \xi}{\xi} \left(1 + \frac{2}{3} \sin^2 \frac{\xi}{2} + \frac{8}{15} \sin^4 \frac{\xi}{2} \right) \right] \\ & + \mathcal{O}(\xi^8). \end{aligned} \quad (3.1.12b)$$

We assume that $|\xi| \ll 1$ and neglect the $\mathcal{O}(\xi^{p+2})$ term.

The time required for a particular feature of the solution to traverse the interval $[0, 2\pi]$ is $2\pi/a$. Because we are considering wave numbers ω larger than one in magnitude, the same feature occurs ω times during the same period in time, and therefore $2\pi/\omega a$ is the time for one period to pass a given point x . [This is seen directly from the form of the solution (3.1.2).] We consider computing the solution over q periods in time, where $q \geq 1$, that is, $t = 2\pi q/(\omega a)$. We want to ensure that the error is smaller than a given tolerance ϵ ; that is, we allow

$$e_p(\omega) = \epsilon, \quad p = 2, 4, 6,$$

after q periods. The question now is: How many gridpoints N_p are required to achieve this level of accuracy for the p th order accurate method? (For convenience we use the notation N_p here instead of $N_p + 1$.) We introduce the number of points per wavelength, M_p , which is defined by

$$M_p = N_p/\omega = 2\pi/\xi, \quad p = 2, 4, 6. \quad (3.1.13)$$

This is a natural measure, because we can first determine the maximum wave number ω that must be accurately computed and then find out what resolution is required for that wave. All smaller wave numbers will be approximated at least as well. From Eqs. (3.1.5) and (3.1.13) we obtain

$$e_2(\omega) \approx 2\pi q \left(1 - \frac{\sin \xi}{\xi} \right) \approx \frac{\pi q \xi^2}{3} = \frac{4\pi^3 q}{3M_2^2} = \epsilon, \quad t = 2\pi q/(\omega a),$$

and M_2 can be obtained from the last equality.

A similar computation can be carried out for the fourth- and sixth-order operators, and the results are

$$\begin{aligned} M_2 &\approx 2\pi \left(\frac{\pi}{3} \right)^{1/2} \left(\frac{q}{\epsilon} \right)^{1/2}, \\ M_4 &\approx 2\pi \left(\frac{\pi}{15} \right)^{1/4} \left(\frac{q}{\epsilon} \right)^{1/4}, \\ M_6 &\approx 2\pi \left(\frac{\pi}{70} \right)^{1/6} \left(\frac{q}{\epsilon} \right)^{1/6}. \end{aligned} \quad (3.1.14)$$

For any even order of accuracy, the formula has the form

$$M_p = C_p \left(\frac{q}{\epsilon} \right)^{1/p}. \quad (3.1.15)$$

The relationships among M_2 , M_4 , and M_6 are

$$\begin{aligned} M_2 &\approx \frac{\sqrt{5}}{2\pi} M_4^2 \approx 0.36 M_4^2, \\ M_4 &\approx \left(\frac{14}{3} \right)^{1/4} \frac{1}{\sqrt{2\pi}} M_6^{3/2} \approx 0.58 M_6^{3/2}, \end{aligned} \quad (3.1.16)$$

and we note that they are independent of ϵ and q . Table 3.1.1 shows M_p as a function of q for two realistic values of ϵ .

How can we use the above information? Assume that we know for a given problem how many Fourier modes are needed to adequately describe the solution. Then we choose the number of points such that the wave with the highest frequency is well resolved according to our analysis above; that is, ϵ is sufficiently small.

If more detailed information is known about the spectral distribution of the solution $\hat{v}(\omega, t)$, then this can be used to obtain sharper comparisons of efficiency by weighting the error function appropriately.

Next we consider the fully discretized case. If leap-frog time differencing is used and the semidiscrete approximation is

$$\frac{dv}{dt} = Qv, \quad (3.1.17)$$

then the full approximation is given by

TABLE 3.1.1. M_p as a function of q

	M_2	M_4	M_6
$\epsilon = 0.1$	$20q^{1/2}$	$7q^{1/4}$	$5q^{1/6}$
$\epsilon = 0.01$	$64q^{1/2}$	$13q^{1/4}$	$8q^{1/6}$

$$\begin{aligned} v^{n+1} &= 2kQv^n + v^{n-1}, \\ v^0 &= e^{i\omega x}, \quad v^1 = (I + kQ)v^0. \end{aligned} \quad (3.1.18)$$

(Here v is considered as a vector of the gridvalues.) We choose the time step k so small that the error due to the time differencing is small compared with ϵ . One can show that this is the case if $k^2\omega^3/6 \ll \epsilon$.

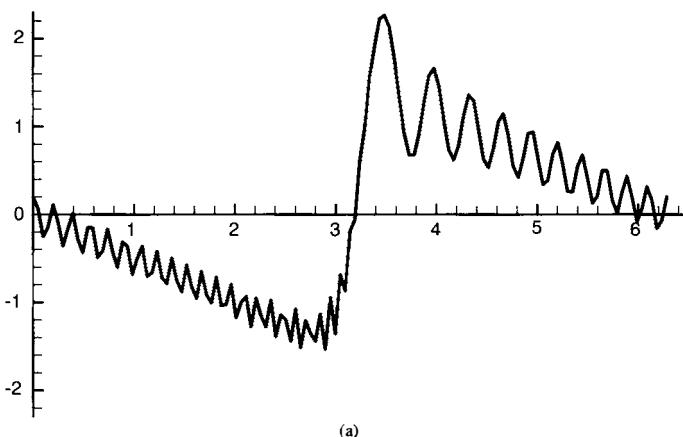
Table 3.1.1 tells us that the number of points per wavelength decreases by a factor three if $\epsilon = 10^{-1}$, $q = 1$, and we replace Q_2 by Q_4 . The amount of work decreases by a factor $3/2$, because it is twice as much work to calculate Q_4v as Q_2v . However, in applications we want to solve systems

$$u_t = A(x, t, u)u_x + F(x, t, u),$$

and often the evaluation of A and F is much more expensive than the calculation of the derivatives. In this case, we gain almost a factor of 3 in the work of these evaluations. The gain is greater in more space dimensions. In three dimensions, the number of points is reduced by a factor of 27, and the work is decreased by a factor of between 27/8 and 27, depending on the cost of evaluating the coefficients. The gain is even greater for the error level $\epsilon = 10^{-2}$. However, for general initial data, experience indicates that $\epsilon = 10^{-1}$ is appropriate to obtain an overall error of 10^{-2} . The reason is that most of the energy is contained in the small wave numbers, and for those we have enough points per wavelength. The energy in the large wave numbers is small and therefore $\epsilon = 10^{-1}$ is tolerable.

In the discussion here, it was assumed that only one period in time was computed. Table 3.1.1 shows that one gains even more by going to higher order accuracy if the computation is carried out over longer time intervals.

In Figure 3.1.1, we have calculated the solution of



(a)

Figure 3.1.1 (a) $p = 2$. (b) $p = 4$. (c) $p = 6$.

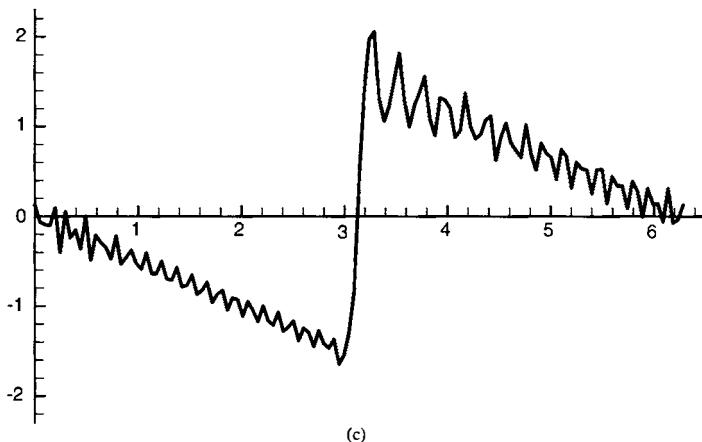
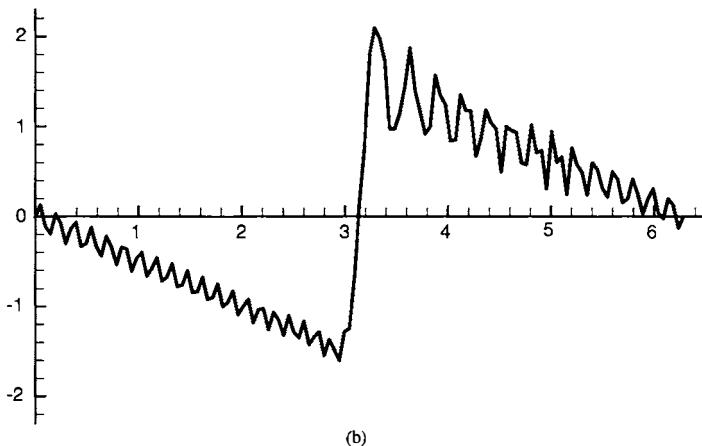


Figure 3.1.1 (Continued)

$$u_t + u_x = 0, \quad t \geq 0, \quad u(x + 2\pi, t) = u(x, t),$$

$$u(x, 0) = \frac{1}{2} (\pi - x) = \sum_{\omega=1}^{\infty} \frac{\sin \omega x}{\omega} = \text{the sawtooth function},$$

using Eq. (3.1.18) with $Q = Q_p$, $p = 2, 4, 6$, $N = 128$, and $k = 10^{-3}$ to $t = \pi$. The results are not satisfactory. If we use Q_6 , Table 3.1.1 shows that we can only compute the first $128/(5q^{1/6})$ waves in the Fourier expansion of the solution with an error $\leq 10^{-1}$. The large wave numbers are not approximated well enough. In Figure 3.1.2 we have replaced the initial data by

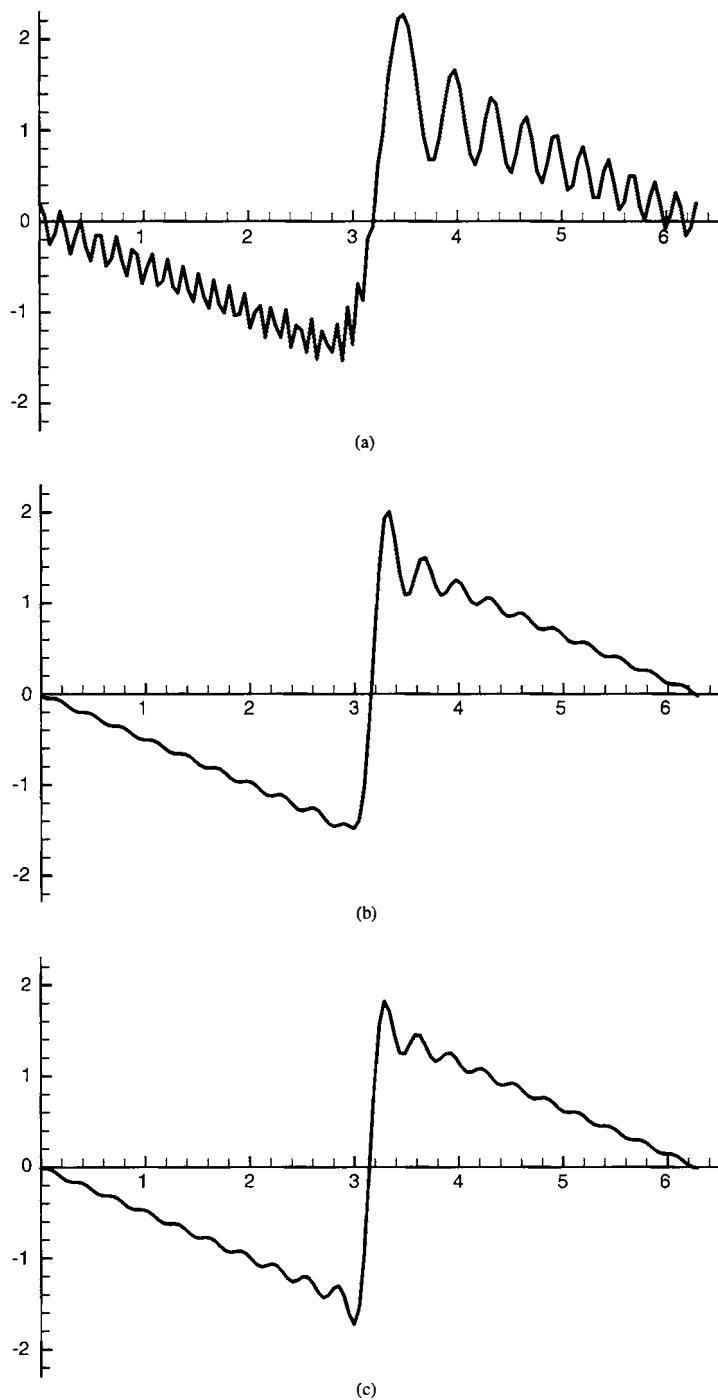


Figure 3.1.2 (a) $p = 2$. (b) $p = 4$. (c) $p = 6$.

$$u(x, 0) = \sum_{\omega=1}^{20} \frac{\sin \omega x}{\omega}.$$

Now the results for the fourth- and sixth-order methods are better.

We see again the Gibbs' phenomena as explained in Chapter 1. To suppress it, we replace the initial data in Figure 3.1.3 by

$$u(x, 0) = \sum_{\omega=1}^{30} \left(1 - \left(\frac{\omega}{30} \right)^2 \right) \frac{\sin \omega x}{\omega},$$

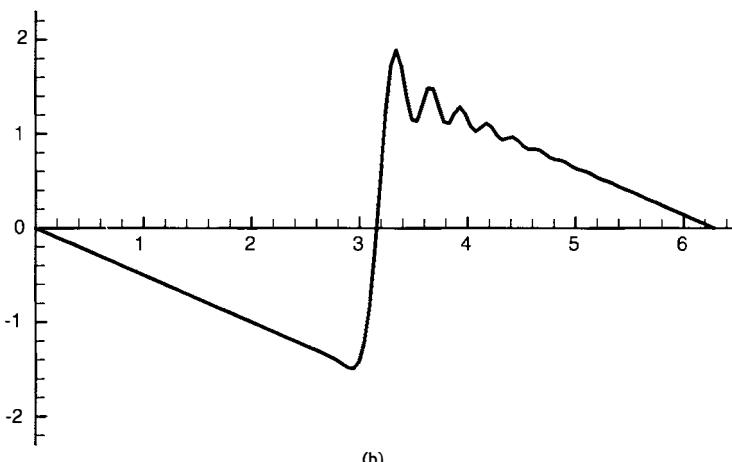
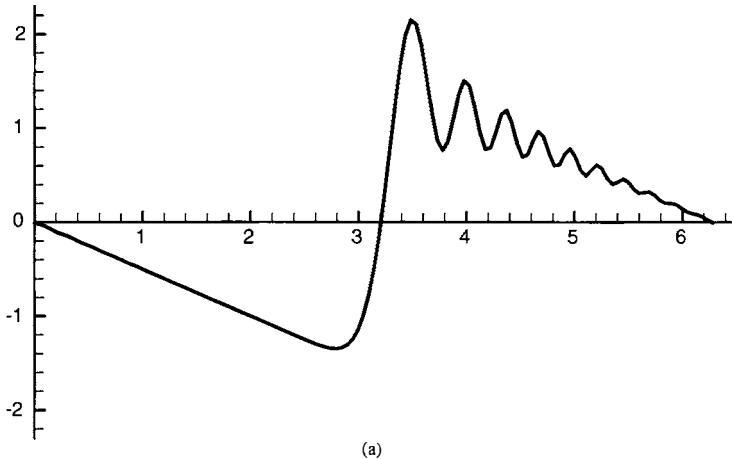


Figure 3.1.3 (a) $p = 2$. (b) $p = 4$. (c) $p = 6$.

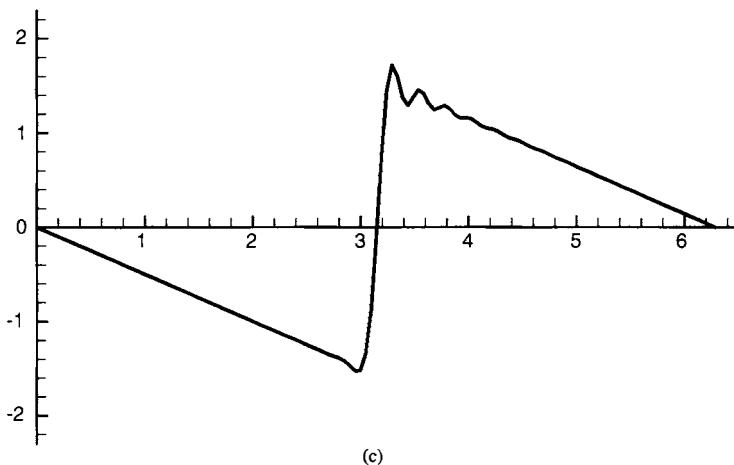
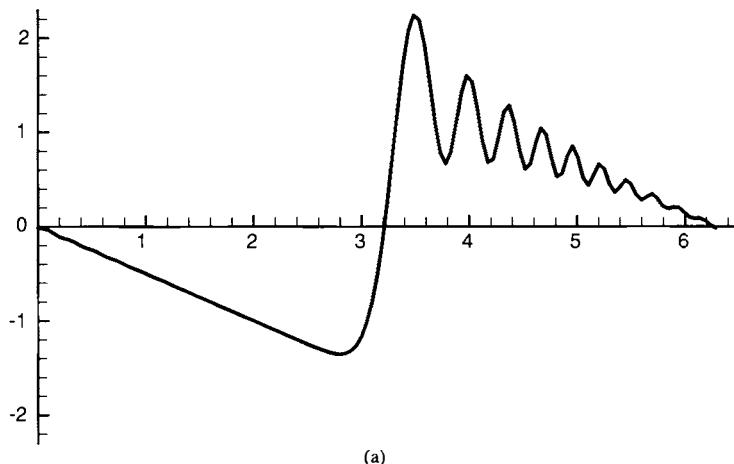


Figure 3.1.3 (Continued)

and in Figure 3.1.4 by

$$u(x, 0) = \sum_{\omega=1}^{30} \left(1 - \left(\frac{\omega}{30} \right)^4 \right) \frac{\sin \omega x}{\omega}.$$

These series are constructed according to the principle discussed for Eq. (1.1.5): The coefficients are reduced to zero in a smooth way as ω increases. For these smoother solutions the results for the three different methods show the expected behavior.

Figure 3.1.4 (a) $p = 2$. (b) $p = 4$. (c) $p = 6$.

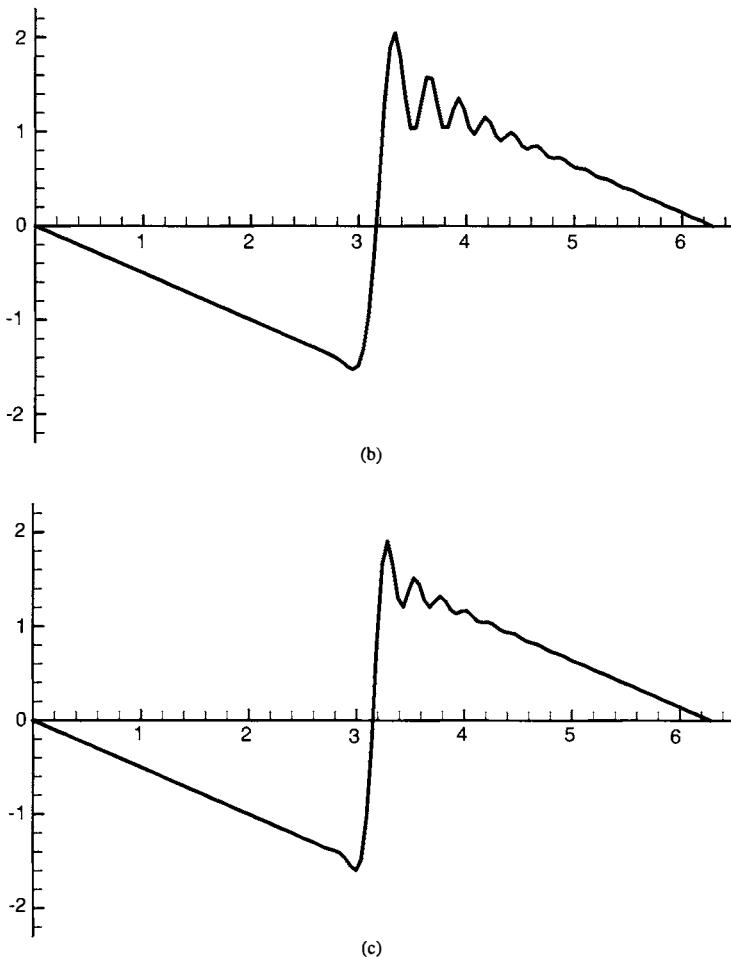


Figure 3.1.4 (Continued)

The leap-frog time differencing used here is only second-order accurate, and it seems reasonable to use higher order methods for the time discretization. However, the operators Q_p , derived for space discretization, cannot be used in the time direction because the resulting schemes are unstable.

There are many time differencing methods developed for ordinary differential equations $dv/dt = Qv$. One class of these methods is that of one-step multistage methods, where Qv is evaluated at different stages in order to go from v^n to v^{n+1} . The explicit Runge-Kutta methods are typical representatives for this class, and they will be discussed in Section 6.7. Hyman's method is another popular method of this class; it is also discussed in Section 6.7.

Another class of methods is that of linear multistep methods, which have the

form

$$\sum_{\nu=0}^q \alpha_\nu v^{n+1-\nu} = k \sum_{\nu=0}^q \beta_\nu Q v^{n+1-\nu}. \quad (3.1.19)$$

(Note, that “linear” refers to the fact that Q occurs linearly in the formula, Q itself does not have to be linear.) Explicit linear multistep methods of high accuracy can be constructed, but they require storage of the solution at many time levels, which is inconvenient for large partial differential equation problems. If the solution of implicit systems at every time step is acceptable, that is, $\beta_q \neq 0$ in Eq. (3.1.19), then it is possible to construct more compact linear multistep schemes that do not require so much storage. One such method is *Simpson's rule*, which we derive next.

Consider the semidiscrete problem in Eq. (3.1.17) again and the leap-frog scheme in Eq. (3.1.18). For a smooth solution $v(t)$ of Eq. (3.1.17), a Taylor expansion around $t = t_n$ yields

$$\begin{aligned} \frac{v^{n+1} - v^{n-1}}{2k} &= v_t^n + \frac{k^2}{6} v_{ttt}^n + \mathcal{O}(k^4) \\ &= Qv^n + \frac{k^2}{6} v_{ttt}^n + \mathcal{O}(k^4). \end{aligned}$$

We want to eliminate the k^2 term and note that

$$\begin{aligned} \frac{k^2}{2} v_{ttt}^n &= \frac{1}{2} (v_t^{n+1} + v_t^{n-1}) - v_t^n + \mathcal{O}(k^4) \\ &= \frac{1}{2} (Qv^{n+1} + Qv^{n-1}) - Qv^n + \mathcal{O}(k^4), \end{aligned}$$

that is,

$$\begin{aligned} Qv^n = v_t^n &= \frac{1}{2} (v_t^{n+1} + v_t^{n-1}) - \frac{k^2}{2} v_{ttt}^n + \mathcal{O}(k^4) \\ &= \frac{1}{2} (Qv^{n+1} + Qv^{n-1}) - \frac{k^2}{2} v_{ttt}^n + \mathcal{O}(k^4). \end{aligned}$$

Therefore, if we replace Qv^n by $\frac{2}{3}Qv^n + \frac{1}{6}(Qv^{n+1} + Qv^{n-1})$, then we obtain the fourth-order method

$$\frac{v^{n+1} - v^{n-1}}{2k} = \frac{1}{6} (Qv^{n+1} + 4Qv^n + Qv^{n-1}). \quad (3.1.20)$$

This implicit method is only conditionally stable (see Exercise 3.1.2). As for all implicit methods, the question of how to solve the resulting system of equations arises. Direct methods are prohibitively expensive for realistic problems in several space dimensions, and iterative methods are more attractive. No matter what iterative method is used, it is important to have a good initial guess $v_{[0]}^{n+1}$ for v^{n+1} . The value v^n on the previous time level is a natural choice. However, a better approximation can be obtained if $v_{[0]}^{n+1}$ is computed using an explicit approximation of the differential equation. This explicit method is called a *predictor*. If the approximate values are substituted into the right-hand side, we obtain the *corrector* formula. This can be used as an iteration for v^{n+1} . For Eq. (3.1.20), we obtain

$$v_{[\nu+1]}^{n+1} = v^{n-1} + \frac{k}{3} (Qv_{[\nu]}^{n+1} + 4Qv^n + Qv^{n-1}), \quad \nu = 0, 1, \dots \quad (3.1.21)$$

The combined procedure is called a *predictor-corrector* method and, in practice, only a few iterations are used.

Clearly, the stability for the combined method does not follow from the stability of the original implicit method. This would only be the case if, at each time level, the corrector is iterated to convergence. If a fixed number of iterations is used, then the combined scheme is effectively explicit, and it must be analyzed to ensure stability. One possible predictor is

$$v_{[0]}^{n+1} + 4v^n - 5v^{n-1} = 2k(2Qv^n + Qv^{n-1}) \quad (3.1.22)$$

followed by one step of Eq. (3.1.21). It can be shown to be stable with $Q = Q_p$ as defined by Eq. (3.1.7). (The stability limit for k/h depends on p .)

We next turn to the parabolic model problem

$$\begin{aligned} u_t &= au_{xx}, \quad a > 0, \\ u(x, 0) &= e^{i\omega x}, \end{aligned} \quad (3.1.23)$$

which has the solution

$$u(x, t) = e^{-a\omega^2 t} e^{i\omega x}. \quad (3.1.24)$$

This is a so-called standing wave, whose amplitude decreases with time. The semidiscrete second-order approximation is

$$\begin{aligned} \frac{dv_j(t)}{dt} &= aD_+D_-v_j(t), \\ v_j(0) &= e^{i\omega x_j}, \quad j = 0, 1, \dots, N, \end{aligned} \quad (3.1.25)$$

which has the solution

$$v_j(t) = e^{-(4a/h^2)t \sin^2 \xi/2} e^{i\omega x_j}. \quad (3.1.26)$$

This approximation has no phase error, but it has an error in amplitude. It is convenient to study the error of the exponent, which is

$$d_2(\omega) = -\left(\frac{4}{h^2} \sin^2 \frac{\xi}{2} - \omega^2 \right) at. \quad (3.1.27)$$

If we use a Taylor expansion about $\xi = 0$, we get

$$d_2(\omega) \approx \frac{a}{12} \omega^4 h^2 t, \quad (3.1.28)$$

which shows that the approximation has a smaller decay rate than the solution of the differential equation.

For the error, we have a first approximation,

$$\begin{aligned} |u(x_j, t) - v_j(t)| &= e^{-a\omega^2 t} |1 - e^{d_2(\omega)}| \\ &\approx \frac{1}{12} \omega^2 h^2 \cdot (a\omega^2 t) e^{-a\omega^2 t} \leq \frac{\xi^2}{12e}. \end{aligned}$$

Thus,

$$\max |u(x_j, t) - v_j(t)| \approx \frac{\xi^2}{12e} \quad \text{for } t \approx \frac{1}{a\omega^2};$$

that is, if the time interval $[0, T]$ for the computation is not very small, the maximum error does not depend on T . This is due to the dissipative character of parabolic equations. We again introduce the error level ϵ and the number of points per wavelength $M_p = 2\pi/\xi$, and we obtain

$$M_2 \approx \frac{\pi}{\sqrt{3e}} \epsilon^{-1/2} = 1.1 \cdot \epsilon^{-1/2}.$$

A corresponding calculation yields, for the fourth-order approximation,

TABLE 3.1.2. M_p for parabolic equations

	M_2	M_4
$\epsilon = 0.1$	3.5	2.5
$\epsilon = 0.01$	11	4.4

$$\frac{dv_j(t)}{dt} = aD_+D_- \left(I - \frac{h^2}{12} D_+D_- \right) v_j(t),$$

$$M_4 \approx \frac{2\pi}{(90e)^{1/4}} \epsilon^{-1/4} \approx 1.6\epsilon^{-1/4},$$

which gives us Table 3.1.2.

If the error level $\epsilon = 10^{-1}$ is tolerable, then there is no reason to use a higher order method, and even for $\epsilon = 10^{-2}$ it is questionable. Thus, for diffusion-convection problems, where the diffusion dominates, second-order methods are often adequate.

Instead of the high order accurate difference operators Q_p used above, one can use compact implicit operators of Padé type. The approximation w_j of $u_x(x_j)$ is obtained by solving a system

$$Pw_j = Qv_j \quad (3.1.29)$$

where v_j approximates $u(x_j)$. Here P and Q are difference operators using a few grid points. For example, a fourth order approximation of u_x is obtained with

$$P = \frac{1}{6} (E^{-1} + 4I + E),$$

$$Q = \frac{1}{2h} (E - E^{-1}). \quad (3.1.30)$$

A linear system of equations must be solved in each time-step. This extra work may well pay off, since the error constant is very small (see Exercise 3.1.6).

EXERCISES

- 3.1.1.** Derive a sixth-order approximation of $\partial^2/\partial x^2$ and an estimate of the number of gridpoints M_6 in terms of ϵ . Add the third column for M_6 to Table 3.1.2.
- 3.1.2.** Derive the stability condition for Simpson's rule (3.1.20) when used for the equation $u_t = u_x$ with Q_4 .

- 3.1.3.** Derive the order of accuracy in time for the predictor-corrector method [Eqs. (3.1.21) and (3.1.22)] when only one iteration is used in the corrector.
- 3.1.4.** Verify Table 3.1.1 by carrying out numerical experiments for $u_t + a(x)u_x = F(x, t)$ with the leap-frog scheme and a small time step. Hint: One can always find solutions of a partial differential equations in the following way. Let U be any function. Calculate

$$U_t - P(\partial/\partial x)U = F(x, t)$$

then U is the solution of

$$\begin{aligned} u_t - P(\partial/\partial x)u &= F \\ u(x, 0) &= U(x, 0). \end{aligned}$$

- 3.1.5.** Derive the coefficients of Lemma 3.1.1 with the help of the relationship

$$\frac{d}{d\theta} \left(\frac{\arcsin \theta}{\sqrt{1-\theta^2}} \right) = \frac{1}{1-\theta^2} \left(1 + \theta \frac{\arcsin \theta}{\sqrt{1-\theta^2}} \right).$$

- 3.1.6.** Derive the leading error term ch^4 in the approximation

$$u_x \approx P^{-1}Qu,$$

where P and Q are defined in Eq. (3.1.30). Compare this to the leading error term for Q_4 defined in Eq. (3.1.8).

- 3.1.7.** Derive a Padé type fourth order approximation of u_{xx} , and derive the leading error term.

3.2. FOURIER METHOD

Again, consider the hyperbolic problem

$$\begin{aligned} u_t + au_x &= 0, \\ u(0) = f(x) &= \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}, \end{aligned} \tag{3.2.1}$$

with its solution given by

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega(x- at)} \hat{f}(\omega). \quad (3.2.2)$$

In Section 1.4, it was shown how trigonometric interpolation can be used to derive high-order accurate approximations of derivatives using the S operator defined in Eq. (1.4.2). The spacial coordinate x is discretized and the differential operator $\partial/\partial x$ is replaced by S . We use the truncated Fourier series as initial data for the approximation. (In practice, the natural choice is the interpolating polynomial, which is not the same.)

Let \mathbf{v} be the vector with gridvalues $v_j, j = 0, 1, \dots, N$, as its elements. Then the Fourier method is defined by

$$\begin{aligned} \frac{d\mathbf{v}}{dt} + aS\mathbf{v} &= 0, \\ \mathbf{v}(0) &= \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} \hat{f}(\omega) \mathbf{e}_\omega. \end{aligned} \quad (3.2.3)$$

Here, the vector \mathbf{e}_ω consists of the Fourier component $e^{i\omega x}$ evaluated in each of the gridpoints, as defined in Eq. (1.4.3). It was shown that these vectors form an eigensystem for S . $\mathbf{v}(t)$ can be expressed as a linear combination [see Eq. (1.4.4)] of these vectors. If this is substituted into Eq. (3.2.3), we obtain

$$\begin{aligned} \sum_{|\omega| \leq N/2} \left[\frac{\partial \tilde{v}(\omega, t)}{\partial t} + i\omega \tilde{v}(\omega, t) \right] \mathbf{e}_\omega &= 0, \\ \sum_{|\omega| \leq N/2} [\tilde{v}(\omega, 0) - \hat{f}(\omega)] \mathbf{e}_\omega &= 0. \end{aligned} \quad (3.2.4)$$

Since the vectors \mathbf{e}_ω are linearly independent, the expressions in the brackets must vanish for each ω . [The process of isolating each component $\tilde{v}(\omega, t)$ can of course be considered as a diagonalization of the matrix S in Eq. (3.2.3).] We obtain

$$\tilde{v}(\omega, t) = e^{-i\omega at} \hat{f}(\omega), \quad |\omega| \leq N/2, \quad (3.2.5)$$

and the approximate solution is

$$v_j(t) = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} e^{i\omega(x_j - at)} \hat{f}(\omega), \quad j = 0, 1, \dots, N. \quad (3.2.6)$$

This can be compared to the true solution shown in Eq. (3.2.2). For smooth solutions, $|\hat{f}(\omega)|$ is very small for large values of $|\omega|$, and the approximate solution shown in Eq. (3.2.6) is very accurate. In particular, if $\hat{f}(\omega) = 0$ for $|\omega| > N/2$, then the exact solution is obtained (even if the interpolation polynomial is used, as demonstrated in Lemma 1.3.2). One way to define the accuracy requirements is to determine a maximum wave number, ω_{\max} , for which the solution must be determined within a certain error, as described in the previous section for difference methods. With Fourier methods, we obtain the exact solution (disregarding time discretization) for these first ω_{\max} wave numbers.

Now we compare the efficiency of difference methods. If one successively increases the order of accuracy p using centered difference operators, as was done in Section 3.1, and use periodicity one can show that the Fourier method represented by the S operator is obtained in the limit as $p \rightarrow \infty$. The work per point using the centered difference operators is proportional to p and, therefore, the work per time step is $\mathcal{O}(pN)$. For large values of p , the amount of work required to implement the high-order methods straightforwardly is excessive, and the methods are inefficient. When $p \approx \log N$ or larger, it is more efficient to use the Fourier method to implement the p th order approximation using trigonometric interpolation with an appropriate modified operator S . In this case, the operation count is reduced to $\mathcal{O}(N \log N)$ per time step.

It was noted above that the first N Fourier components are computed without error using the Fourier method. Disregarding the constant term $\tilde{v}(0, t)/\sqrt{2\pi}$, it takes N points to determine the corresponding N Fourier coefficients. ($N/2$ sine-coefficients and $N/2$ cosine-coefficients in the real case.) Thus, exactly two points per wavelength are necessary, which is the lower limit for M_p .

The work per wavelength for difference methods is defined as

$$W_p = pM_p, \quad p = 2, 4, 6, \quad (3.2.7)$$

where M_p , which is defined in Eq. (3.1.13), is a function of q/ϵ . For the spectral method, W_p is independent of q and ϵ , and it is only a function of the number of gridpoints, $(N + 1)$. Figure 3.2.1 is a plot of W_p for the different methods.

Again, we must remember that our analysis is for a very simple problem. Therefore, the conclusion can only be used as a rough guide. From Figure 3.2.1, we can expect that the Fourier method will perform well over long time intervals, when higher order accuracy is required, that is, when the maximum wave number to be accurately represented is large.

The success of the Fourier method to approximate solutions of Eq. (3.2.1) is due to the fact that the functions $e^{i\omega x}$ are the eigenfunctions of that problem. The Fourier method computes the first N terms of the expansion of the solution of Eq. (3.2.1) in its eigenfunctions. For problems where the eigenfunctions are not close to Fourier modes, the advantage of the Fourier method is not that pronounced.

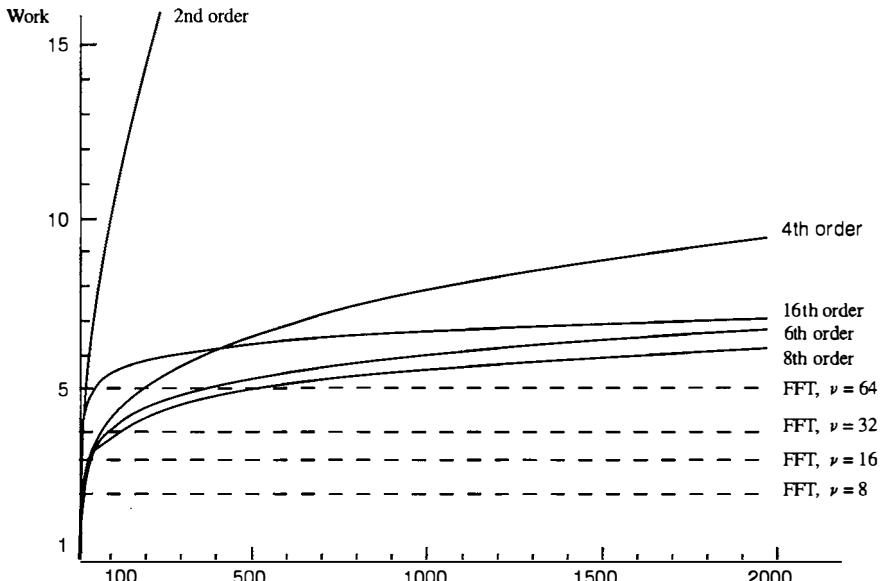


Figure 3.2.1. W_p as a function of q/ϵ . [Reprinted from Fornberg *SIAM Journal of Numerical Analysis* 12, 521 (1975), with permission. Note that $\nu = N$ in the figure.]

The Fourier method cannot be used directly for problems with nonperiodic solutions. A similar method based upon expansions in Chebyshev polynomials can be used in this case, but it will not be discussed in this book.

For time discretization, the leap-frog scheme is often used:

$$\mathbf{v}^{n+1} = -2kaS\mathbf{v}^n + \mathbf{v}^{n-1}. \quad (3.2.8)$$

In Fourier space, the scheme has the form

$$\tilde{v}(\omega)^{n+1} = -2kai\omega\tilde{v}(\omega)^n + \tilde{v}(\omega)^{n-1}, \quad |\omega| \leq N/2, \quad (3.2.9)$$

with the characteristic equation

$$z^2 + 2kai\omega z - 1 = 0. \quad (3.2.10)$$

The solutions are given by

$$z_{1,2} = -kai\omega \pm \sqrt{1 - (kai\omega)^2}, \quad |\omega| \leq N/2, \quad (3.2.11)$$

showing that the condition for stability is $|kaN/2| < 1$, or

$$\frac{k|a|}{h} \left(\pi - \frac{h}{2} \right) < 1. \quad (3.2.12)$$

This compares with $k|a|/h < 1$ for the second order leap-frog difference scheme. The dominating part of the work when using Eq. (3.2.8) is the two FFTs at each time step. The operation count is therefore $\mathcal{O}(N \log N)$ per step.

Another way of differencing in time is the trapezoidal rule (usually called the Crank–Nicholson method, when applied to centered difference approximations):

$$\left(I + \frac{k}{2} aS \right) \mathbf{v}^{n+1} = \left(I - \frac{k}{2} aS \right) \mathbf{v}^n. \quad (3.2.13)$$

The amplification factor is

$$\hat{Q} = \frac{1 - ik\omega/2}{1 + ik\omega/2}, \quad (3.2.14)$$

so it is unconditionally stable. The implementation of Eq. (3.2.13) can be carried out in Fourier space as follows:

If the FFT operation is denoted by the operator T , and the multiplications by $i\omega$ are carried out with the diagonal operator D , then S can be factored as

$$S = T^{-1}DT, \quad (3.2.15)$$

and Eq. (3.2.13) can be rewritten in the form

$$\left(I + \frac{k}{2} aT^{-1}DT \right) \mathbf{v}^{n+1} = \left(I - \frac{k}{2} aT^{-1}DT \right) \mathbf{v}^n. \quad (3.2.16)$$

A multiplication of Eq. (3.2.16) from the left by T gives, with $\tilde{\mathbf{v}}^n = T\mathbf{v}^n$,

$$\left(I + \frac{k}{2} aD \right) \tilde{\mathbf{v}}^{n+1} = \left(I - \frac{k}{2} aD \right) \tilde{\mathbf{v}}^n. \quad (3.2.17)$$

$\tilde{\mathbf{v}}^{n+1}$ is obtained from

$$\tilde{\mathbf{v}}^{n+1} = \left(I + \frac{k}{2} aD \right)^{-1} \left(I - \frac{k}{2} aD \right) \tilde{\mathbf{v}}^n, \quad (3.2.18)$$

and finally \mathbf{v}^{n+1} from $\mathbf{v}^{n+1} = T^{-1}\tilde{\mathbf{v}}^{n+1}$. Hence, the algorithm, for each step, is:

1. Compute $\tilde{\mathbf{v}}^n = T\mathbf{v}^n$ using FFT.
2. Compute $\tilde{\mathbf{v}}^{n+1}$ from Eq. (3.2.18) (D is diagonal so it takes only $\mathcal{O}(N)$ operations).
3. Compute $\mathbf{v}^{n+1} = T^{-1}\tilde{\mathbf{v}}^{n+1}$ using the inverse FFT.

Therefore, the operation count is approximately the same as for the explicit leap-frog scheme. If the coefficient a depends on x , then the algorithm is more complicated.

Only hyperbolic problems have been treated here, but the definition of the Fourier method is obvious for higher order equations. For example, the operator S^2 approximates the differential operator $\partial^2/\partial x^2$, and it takes only two FFTs to compute it. With the notation above, we have

$$S^2 = (T^{-1}DT)(T^{-1}DT) = T^{-1}D^2T, \quad (3.2.19)$$

where the matrix D^2 is diagonal with elements $-\omega^2$, where $|\omega| \leq N/2$. Similarly, $\partial^2/\partial x \partial y$ can be approximated by $S_x S_y$, where S_x and S_y are the one-dimensional operators corresponding to each space dimension.

The Fourier method has been presented here as a way of approximating derivatives with high accuracy, and it can be obtained as the limit of difference approximations on a given grid. Another type of Fourier technique can be derived using *Galerkin's method*. This method can be defined in a more abstract form without specifying the type of basis functions to be used. With piecewise polynomials as basis functions, it can be used to derive the *finite element method*. Here, we briefly describe how it can be used with trigonometric basis functions.

Consider the general problem

$$\frac{\partial u}{\partial t} = P\left(x, t, \frac{\partial}{\partial x}\right) u, \quad (3.2.20a)$$

$$u(x, 0) = f(x), \quad (3.2.20b)$$

where

$$P\left(x, t, \frac{\partial}{\partial x}\right) = \sum_{\nu=0}^p A_\nu(x, t) \frac{\partial^\nu}{\partial x^\nu}.$$

The differential equation is multiplied by a test function $\varphi(x)$, which belongs to an appropriate function space \mathcal{L} . The equation is then integrated over the interval $[0, 2\pi]$ to obtain

$$\begin{aligned}(u_t, \varphi) &= (Pu, \varphi), \\ (u(\cdot, 0), \varphi) &= (f, \varphi)\end{aligned}\quad (3.2.21)$$

Equation (3.2.20) is now reformulated as follows: Find a function $u(x, t)$ that belongs to an appropriate function space \mathcal{M} for each fixed t such that Eq. (3.2.21) is fulfilled for each $\varphi(x) \in \mathcal{L}$.

This is called the weak form of Eq. (3.2.20a). The spaces \mathcal{M} and \mathcal{L} are chosen so that the integrals in Eq. (3.2.21) exist. For example, if $P(\partial/\partial x) = a(x)\partial/\partial x$, then \mathcal{L} can be taken to be the set of all functions in $C^\infty(0, 2\pi)$ with 2π -periodic derivatives and \mathcal{M} as all periodic functions with derivatives in $L_2(0, 2\pi)$.

An approximation of Eq. (3.2.21) is obtained by replacing the spaces \mathcal{M} and \mathcal{L} by finite dimensional subspaces \mathcal{M}_N and \mathcal{L}_N . By choosing these spaces appropriately, we obtain the formulation of an approximation of the problem, which leads to a numerical algorithm. If $\mathcal{M}_N = \mathcal{L}_N$, this is called the *Galerkin method*.

Choose \mathcal{L}_N as the class of trigonometric polynomials and let $\tilde{f} \in \mathcal{L}_N$ be an approximation of f . The Galerkin approximation is then defined as follows:

Definition 3.2.1. The Galerkin Approximation. *Find a function*

$$v(x, t) = 1/\sqrt{2\pi} \sum_{\omega=-N/2}^{N/2} \hat{v}(\omega, t) e^{i\omega x}$$

such that

$$\begin{aligned}(v_t, \varphi) &= (Pv, \varphi), \\ (v(\cdot, 0), \varphi) &= (f, \varphi)\end{aligned}\quad (3.2.22)$$

for all $\varphi \in \mathcal{L}_N$.

Because the functions $\{e^{i\omega x}\}_{\omega=-N/2}^{N/2}$ span the space \mathcal{L}_N , Eq. (3.2.22) is fulfilled for all $\varphi \in \mathcal{L}_N$ if and only if it is fulfilled for each basis function $e^{i\omega x}$.

By introducing the Fourier series for v and multiplying by $\sqrt{2\pi}$ we obtain

$$\left(\sum_{|\omega| \leq N/2} \left(\frac{\partial}{\partial t} - P\left(x, t, \frac{\partial}{\partial x}\right) \right) \hat{v}(\omega, t) e^{i\omega x}, e^{i\nu x} \right) = 0, \quad |\nu| \leq N/2. \quad (3.2.23)$$

By using Lemma 1.1.1, this leads to

$$2\pi \frac{d\hat{v}(\nu, t)}{dt} = \sum_{|\omega| \leq N/2} \left(P\left(x, t, \frac{\partial}{\partial x}\right) e^{i\omega x}, e^{i\nu x} \right) \hat{v}(\omega, t), \quad |\nu| \leq N/2. \quad (3.2.24)$$

The coefficients of $\hat{v}(\omega, t)$ on the right-hand side are integrals of known functions and can be computed. The coefficients in the differential operator are periodic functions of x and can be expanded into a Fourier series with coefficients $\hat{a}(\mu, t)$. Therefore, the coefficients of $\hat{v}(\omega, t)$ can be expressed in terms of $\hat{a}(\omega, t)$. Equation (3.2.24) is a system of linear ODEs, which can be solved by numerical methods, usually difference methods. The procedure described here was originally known as the *spectral method*.

We next consider the case where P has constant coefficients. We take, as an example, $P(\partial/\partial x) = a\partial/\partial x$ and get, from Eq. (3.2.24), using Lemma 1.1.1,

$$\frac{d\hat{v}(\omega, t)}{dt} = i\omega a \hat{v}(\omega, t), \quad |\omega| \leq N/2. \quad (3.2.25)$$

Equation (3.2.25) also holds for the coefficients $\tilde{v}(\omega, t)$ of the Fourier method. If the initial data are chosen to be the same, then the solutions are identical. Because the coefficients $\hat{v}(\omega, t) = \tilde{v}(\omega, t)$ uniquely determine the trigonometric polynomial, the Galerkin and Fourier methods are identical in this case.

For problems with variable coefficients, the methods are not identical. The Fourier method is often called the *pseudospectral method*, because the time integration is carried out in physical space, not in Fourier space, as it is in the spectral method.

Let us consider another way of interpreting the Fourier method. Let $\delta_j(x)$ be a Dirac delta function defined such that

$$\int_0^{2\pi} f(x) \delta_j(x) dx = f(x_j), \quad j = 0, 1, \dots, N.$$

If φ in Eq. (3.2.22) is chosen as $\delta_j(x)$, we obtain

$$v_t(x_j, t) = P\left(x, t, \frac{\partial}{\partial x}\right)v(x, t)|_{x=x_j}, \quad j = 0, 1, \dots, N. \quad (3.2.26)$$

In other words, we are requiring that the differential equation be satisfied, but only at the gridpoints x_j . By assuming that v belong to a function space of dimension $N+1$ for each t , it is possible to satisfy the equalities in Eq. (3.2.26). This is known as the *collocation method*, and the points $\{x_j\}$ are called the *collocation points*.

If we choose v to be a trigonometric polynomial, we obtain

$$\sum_{|\omega| \leq N/2} \left(\left(\frac{d}{dt} - P\left(x, t, \frac{\partial}{\partial x}\right) \right) \hat{v}(\omega, t) e^{i\omega x} \right)_{x=x_j} = 0, \quad j = 0, 1, \dots, N, \quad (3.2.27)$$

or, equivalently,

$$\frac{d}{dt} v(x_j, t) = \sum_{|\omega| \leq N/2} \frac{d\hat{v}(\omega, t)}{dt} e^{i\omega x_j} = \sum_{|\omega| \leq N/2} \left(P\left(x, t, \frac{\partial}{\partial x}\right) \hat{v}(\omega, t) e^{i\omega x} \right)_{x=x_j}, \quad j = 0, 1, \dots, N. \quad (3.2.28)$$

This is exactly the Fourier method as defined earlier. With $P = a(x, t)\partial/\partial x$, we obtain

$$\frac{d}{dt} v(x_j, t) = a(x_j, t) \sum_{|\omega| \leq N/2} i\omega \hat{v}(\omega, t) e^{i\omega x_j}, \quad j = 0, 1, \dots, N. \quad (3.2.29)$$

Let $\mathbf{v}(t)$ denote the vector with components defined by

$$v_j(t) = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N/2} \hat{v}(\omega, t) e^{i\omega x_j}, \quad j = 0, 1, \dots, N. \quad (3.2.30)$$

We compute $\hat{v}(\omega, t)$ from $\mathbf{v}(t)$ using the FFT. Let

$$A(t) = \text{diag}(a(x_0, t), a(x_1, t), \dots, a(x_N, t)),$$

then Eq. (3.2.29) can be written as

$$\frac{d\mathbf{v}(t)}{dt} = A(t)S\mathbf{v}(t), \quad (3.2.31)$$

which is the Fourier method.

To summarize, Fourier and pseudospectral are different names for the same procedure. They are equivalent to the collocation method when the class of trigonometric polynomials is chosen to be the solution subspace. The Galerkin spectral method gives the same solution if the differential equation has constant coefficients.

EXERCISES

- 3.2.1.** Write a program for solving $u_t + u_x = 0$ with the Fourier method. Compute an approximation for each of the four initial functions presented in Section 3.1. Compare the results to Figures 3.1.1 to 3.1.4.

BIBLIOGRAPHIC NOTES

The analysis of the semidiscrete problem leading to the estimates of the necessary number of points per wavelength was introduced by Kreiss and Oliger (1972) (they also discussed the Fourier method). Earlier work on these ideas was done by Økland (1958) and Thompson (1961). In practice, the time discretization also plays an important role. An analysis of fully discretized approximations was presented by Swartz and Wendroff (1974). Their results indicate that higher order approximations, in space and time, are more efficient in most cases for hyperbolic problems. However, one should be aware that, so far, we are only dealing with periodic boundary conditions. When other types of boundaries are introduced, the requirement of stability creates an additional difficulty, which is more problematic with higher order accuracy. We discuss this in Part II of this book.

A comprehensive treatment of Fourier and spectral methods can be found in Gottlieb and Orszag (1977) and Canuto, et. al. (1988).

Our discussion of the relation between eigenfunction expansion and the Fourier method is based on Fornberg (1975). The accuracy of the Fourier method has been discussed by Tadmor (1986), where he shows that the convergence rate as $h \rightarrow 0$ is faster than any power h^p . In Abarbanel, Gottlieb, and Tadmor (1986), it is shown that the same “spectral accuracy” is also obtained for discontinuous solutions. Further references concerning the Fourier method are found in the Notes for Chapter 6.

Higher order discretization methods in time are discussed further in Section 6.7. In particular, stability results are presented for Runge–Kutta, for Adams–Bashford, and for Adams–Moulton methods. The time discretization for Fourier methods has been discussed by Gottlieb and Turkel (1980).

The predictor (3.1.22) for the Simpson rule corrector (3.1.21) was suggested by Stetter (1965).

4

WELL-POSED PROBLEMS

In Chapter 2, we considered several model equations and discussed their solutions and properties. In this chapter, we formalize the results and develop concepts that apply to general classes of problems. The concept of well-posedness was introduced by Hadamard and, simply stated, it means that a well-posed problem should have a solution, that this solution should be unique and that it should depend continuously on the problem's data. The first two requirements are certainly obvious minimal requirements for a reasonable problem, and the last ensures that perturbations, such as errors in measurement, should not unduly affect the solution. We refine and quantify this statement and conclude this chapter with a discussion of nonlinear behavior in this context.

4.1. WELL-POSEDNESS

In Chapter 2, we noted that the L_2 norm of the solutions of all our model equations could be estimated for all time in terms of the L_2 norm of the initial data; that is,

$$\|u(\cdot, t)\| \leq \|u(\cdot, 0)\|. \quad (4.1.1)$$

[See Eqs. (2.1.6), (2.5.5), and (2.7.5).] If we change the initial data

$$u(x, 0) = f(x)$$

to

$$\tilde{u}(x, 0) = f(x) + \delta g(x),$$

where $0 < \delta \ll 1$ is a small constant and $\|g(\cdot)\| = 1$, we obtain another solution $\tilde{u}(x, t)$. The difference $w(x, t) = \tilde{u}(x, t) - u(x, t)$ is also a solution of the same differential equation with initial data

$$w(x, 0) = \tilde{u}(x, 0) - u(x, 0) = \delta g(x).$$

The estimate (4.1.1) gives us

$$\|\tilde{u}(\cdot, t) - u(\cdot, t)\| \leq \delta \|g(\cdot)\| = \delta. \quad (4.1.2)$$

Thus, Eq. (4.1.1) guarantees that small changes of the initial data result in small changes in the solution, and we say that the solution depends continuously on the initial data.

The requirement that the estimate (4.1.1) hold for all time is too restrictive. In applications, the differential equations often contain lower order terms. Consider, for example, the equation

$$u_t = u_x + \alpha u, \quad \alpha = \text{constant},$$

with 2π -periodic initial data

$$u(x, 0) = f(x).$$

If we introduce a new dependent variable,

$$v = e^{-\alpha t} u,$$

we obtain the model equation (2.1.1)

$$v_t = v_x, \quad v(x, 0) = f(x).$$

Thus,

$$\|v(\cdot, t)\| = \|v(\cdot, 0)\|,$$

that is,

$$\|u(\cdot, t)\| = e^{\alpha t} \|u(\cdot, 0)\|.$$

In this case, Eq. (4.1.2) becomes

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

In any finite time interval $0 \leq t \leq T$, the difference is still of order δ , although for $\alpha > 0$ and large αt the estimate is useless. However, the solution still depends continuously on the initial data.

Now consider the system

$$u_t = Au_x, \quad A = \begin{bmatrix} 0 & d \\ d & 0 \end{bmatrix}, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \quad d = \text{constant}, \quad (4.1.3)$$

with 2π -periodic initial data

$$u(x, 0) = f.$$

The unitary transformation

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$

transforms A to diagonal form, that is,

$$U^*AU = \begin{bmatrix} -d & 0 \\ 0 & d \end{bmatrix}.$$

Substituting

$$v = U^*u$$

as a new variable into the system (4.1.3) gives us

$$v_t = U^*AUv_x;$$

that is,

$$v_t^{(j)} = \lambda_j v_x^{(j)}, \quad \lambda_1 = -d, \quad \lambda_2 = d. \quad (4.1.4)$$

Thus, we obtain two scalar model equations whose solutions satisfy the estimates

$$\|v^{(j)}(\cdot, t)\|^2 = \|v^{(j)}(\cdot, 0)\|^2, \quad j = 1, 2.$$

Therefore,

$$\begin{aligned} \|v(\cdot, t)\|^2 &= \|v^{(1)}(\cdot, t)\|^2 + \|v^{(2)}(\cdot, t)\|^2, \\ &= \|v^{(1)}(\cdot, 0)\|^2 + \|v^{(2)}(\cdot, 0)\|^2 = \|v(\cdot, 0)\|^2, \end{aligned}$$

and we obtain the energy estimate

$$\begin{aligned}\|u(\cdot, t)\|^2 &= \|Uv(\cdot, t)\|^2 = \|v(\cdot, t)\|^2 = \|v(\cdot, 0)\|^2 \\ &= \|U^*u(\cdot, 0)\|^2 = \|u(\cdot, 0)\|^2.\end{aligned}\quad (4.1.5)$$

The system (4.1.3) is **symmetric and hyperbolic**. We shall see that such systems always have estimates of the form Eq. (4.1.5). In many applications, the systems are **hyperbolic, but not symmetric**. In that case, we still obtain an estimate. A typical example is the system

$$u_t = \begin{bmatrix} 0 & 1 \\ d^2 & 0 \end{bmatrix} u_x, \quad d^2 = \text{constant} > 0.$$

We change the dependent variables by the so called **diagonal scaling**

$$u = D\tilde{u} := \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix} \tilde{u}, \quad d > 0,$$

and obtain the system (4.1.3),

$$\tilde{u}_t = \begin{bmatrix} 1 & 0 \\ 0 & d^{-1} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ d^2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix} \tilde{u}_x = d \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tilde{u}_x. \quad (4.1.6)$$

Therefore, the estimate

$$\|\tilde{u}(\cdot, t)\| = \|\tilde{u}(\cdot, 0)\|$$

holds and we obtain, in the original variables,

$$\begin{aligned}\|u(\cdot, t)\| &= \|D\tilde{u}(\cdot, t)\| \leq |D| \|\tilde{u}(\cdot, t)\|, \\ &= |D| \|\tilde{u}(\cdot, 0)\| = |D| \|D^{-1}u(\cdot, 0)\|, \\ &\leq |D| |D^{-1}| \|u(\cdot, 0)\|.\end{aligned}$$

Thus, we have the estimate

$$\|u(\cdot, t)\| \leq K \|u(\cdot, 0)\|,$$

with

$$K = \max(d, d^{-1}).$$

In this case Eq. (4.1.2) becomes

$$\|\tilde{u}(\cdot, t) - u(\cdot, t)\| \leq K\delta,$$

and the difference is still of order δ , although the estimate is poor for very large or very small d .

Instead of starting at $t = 0$ we can also solve the problem with initial data given at any time t_0 . In that case, our estimates have the form

$$\|u(\cdot, t)\| \leq Ke^{\alpha(t-t_0)}\|u(\cdot, t_0)\|. \quad (4.1.7)$$

We now introduce the concept of well-posedness for general systems. Consider a system of partial differential equations

$$u_t = P\left(x, t, \frac{\partial}{\partial x}\right)u, \quad t \geq t_0, \quad (4.1.8a)$$

with initial data

$$u(x, t_0) = f(x). \quad (4.1.8b)$$

Here $u = (u^{(1)}, \dots, u^{(m)})^T$ is a vector with m components, $x = (x^{(1)}, \dots, x^{(d)})$ and $P(x, t, \partial/\partial x)$ is a general differential operator

$$P\left(x, t, \frac{\partial}{\partial x}\right) = \sum_{|\nu| \leq p} A_\nu(x, t) \left(\frac{\partial}{\partial x^{(1)}} \right)^{\nu_1} \cdots \left(\frac{\partial}{\partial x^{(d)}} \right)^{\nu_d} \quad (4.1.9)$$

of order p . $\nu = (\nu_1, \dots, \nu_d)$ is a multiindex with nonnegative integer elements, and $|\nu| = \sum \nu_i$. The coefficients $A_\nu = A_{\nu_1, \dots, \nu_d}$ are $m \times m$ matrix functions.

For simplicity, we assume that $A_\nu(x, t) \in C^\infty(x, t)$. We also assume that the coefficients and data are 2π -periodic in every space dimension.

We can now make the fundamental definition.

Definition 4.1.1. The problem (4.1.8) is well posed if, for every t_0 and every $f \in C^\infty(x)$:

1. There exists a unique solution $u(x, t) \in C^\infty(x, t)$, which is 2π -periodic in every space dimension and
2. There are constants α and K , independent of f and t_0 , such that Eq. (4.1.7) holds; that is,

$$\|u(\cdot, t)\| \leq Ke^{\alpha(t-t_0)}\|f(\cdot)\|. \quad (4.1.10)$$

This is not the only way well-posedness can be defined. For example, different norms might be used and the functional form of growth allowed might be different—or suppressed entirely. As we will see, Definition 4.1.1 is natural for a large class of problems and allows us to simplify the analysis in a natural way. The exponential growth must be tolerated in order to treat equations with variable coefficients and to be able to ignore lower order terms.

It follows from the discussion in Chapter 2 that all of the periodic boundary problems presented are well posed. It is important to make the distinction that it is the problem that is well posed and not the differential equation itself. The problem setting is as important as the equation. Not all problems are well posed, as our next example will illustrate. If a problem is not well posed, then we say that it is *ill posed*.

Consider the periodic boundary problem for the backward heat equation

$$u_t = -u_{xx} \quad (4.1.11a)$$

for $0 \leq t$ and $0 \leq x \leq 2\pi$ and initial data

$$u(x, 0) = e^{i\omega x} \hat{f}(\omega) \quad (4.1.11b)$$

for $0 \leq x \leq 2\pi$. The solution is given by

$$u(x, t) = e^{i\omega x + \omega^2 t} \hat{f}(\omega), \quad (4.1.12)$$

which grows rapidly in time if ω is large. For example, if $\omega = 10$ and $\hat{f}(\omega) = 10^{-10}$, then $u(0, 1) = 2.7 \times 10^{43}$. We cannot find an α for which Eq. (4.1.10) holds independent of ω . On the other hand, if $\omega = 1$, then

$$u(x, t) = e^{ix + t} \hat{f}(1),$$

and the solution is well behaved. If one can choose the initial data so that only small wave numbers are present, then a reasonably well-behaved solution will exist, and it may not seem important that the problem is ill posed. However, in applications, large wave numbers are always present, for example, due to errors in measurement, and these severely contaminate the solution. One might think that this problem can be solved by simply filtering the initial data so that only small wave numbers are present. Theoretically, this is possible, but, as we have discussed in Section 2.1, in numerical computations, rounding errors always introduce large wave numbers and ruin the solution. In addition, because most application problems have variable coefficients or are nonlinear and high frequencies are then generated spontaneously in the solution at later times, these frequencies will destroy the reasonable behavior of their solutions.

There are difficulties with our definition of well-posedness. Consider the periodic boundary problem for the equation

$$u_t = u_{xx} + 100u \quad (4.1.13)$$

with initial data

$$u(x, 0) = e^{i\omega x} \hat{f}(\omega).$$

The solution of Eq. (4.1.13) is given by

$$u(x, t) = e^{i\omega x + (100 - \omega^2)t} \hat{f}(\omega).$$

Thus, for general data

$$u(x, 0) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{f}(\omega),$$

the solution is

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x + (100 - \omega^2)t} \hat{f}(\omega). \quad (4.1.14)$$

Using Parseval's relation, we obtain

$$\|u(\cdot, t)\|^2 = \sum_{\omega=-\infty}^{\infty} e^{2(100 - \omega^2)t} |\hat{f}(\omega)|^2 \leq e^{200t} \|u(\cdot, 0)\|^2,$$

which shows that the problem is well posed. Equation (4.1.14) shows that large wave numbers are damped but small wave numbers grow rapidly, and one might be tempted to call the problem ill posed. However, there is a fundamental difference between the behavior of the solutions of Eq. (4.1.11) and Eq. (4.1.13). Solutions of Eq. (4.1.11) have no growth limit; arbitrarily rapid growth can be obtained by increasing ω , whereas those of Eq. (4.1.13) do have a growth limit. This shows us that it is not sufficient to only know that the problem is well posed—we also need to have estimates for α and K . This is especially true for numerical calculations because these constants control the

growth rate of rounding and other errors. This will be discussed in the later chapters.

EXERCISES

- 4.1.1.** Suppose that one wants to solve the problem (4.1.13) for $0 \leq t \leq 2$ and will allow 1% relative error in the solution. Give a bound for the permissible rounding errors.

4.2. SCALAR DIFFERENTIAL EQUATIONS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

We consider the scalar equation

$$u_t = au_{xx} + bu_x + cu, \quad t \geq t_0, \quad (4.2.1a)$$

with 2π -periodic initial data

$$u(x, t_0) = f(x). \quad (4.2.1b)$$

The coefficients a, b , and c are complex numbers. We want to investigate what conditions a, b , and c must satisfy if the problem is to be well posed.

For equations with constant coefficients we can assume that $t_0 = 0$, because the coefficients are invariant under the transformation $t' = t - t_0$. We want to prove the following theorem.

Theorem 4.2.1. *The problem (4.2.1) is well posed if, and only if, there is a real constant α , such that, for all real ω ,*

$$\operatorname{Re} \kappa \leq \alpha, \quad \kappa := -a\omega^2 + i\omega b + c. \quad (4.2.2)$$

Proof. We proceed as we did with the model equations. Let

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{f}(\omega)$$

be a simple wave. We construct a simple wave solution

$$u(x, t) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{u}(\omega, t), \quad \hat{u}(\omega, 0) = \hat{f}(\omega). \quad (4.2.3)$$

Substituting Eq. (4.2.3) into Eq. (4.2.1a) gives us the ordinary differential equation

$$\frac{\partial}{\partial t} \hat{u}(\omega, t) = \kappa \hat{u}(\omega, t), \quad (4.2.4)$$

that is,

$$\hat{u}(\omega, t) = e^{\kappa t} \hat{f}(\omega).$$

In this case

$$\|u(\cdot, t)\|^2 = e^{2\operatorname{Re} \kappa t} |\hat{f}(\omega)|^2 = e^{2\operatorname{Re} \kappa t} \|f(\cdot)\|^2.$$

Therefore, the inequality (4.1.10) is satisfied if and only if Eq. (4.2.2) holds.

Now consider general initial data

$$u(x, 0) = f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}$$

and assume that Eq. (4.2.2) holds. The solution of our problem exists and is given by

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{\kappa(\omega)t + i\omega x} \hat{f}(\omega). \quad (4.2.5)$$

By Parseval's relation

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

Therefore, inequality (4.1.10) also holds for general data. This proves the theorem.

We now discuss condition (4.2.2)

1. The constant c , which is the coefficient of the undifferentiated term, has no influence on whether the problem is well posed or not, because we can always replace α by $\alpha + c$ and Eq. (4.2.2) becomes

$$\operatorname{Re}(\kappa - c) = \operatorname{Re}(-a\omega^2 + ib\omega) \leq \alpha.$$

As we shall see, this is typical for general systems. We shall, therefore, assume that $c = 0$.

2. The equation is called **parabolic** if $a_r = \operatorname{Re} a > 0$. In this case

$$\operatorname{Re} \kappa \leq -a_r \omega^2 + |b| |\omega| \leq \frac{|b|^2}{4a_r}.$$

Thus, the problem is **well posed for all values of b** . This is typical for general parabolic systems: the highest derivative term guarantees, by itself, that the problem is well posed.

3. **$\operatorname{Re} a = 0$** . Now

$$\operatorname{Re} \kappa = -\omega \operatorname{Im} b.$$

If $\operatorname{Im} b \neq 0$, then the problem is not well posed, because we can choose the sign of ω such that $\operatorname{Re} \kappa$ becomes arbitrarily large. Thus, well-posed problems have the form

$$u_t = ia_i u_{xx} + b_r u_x, \quad a_i, b_r \text{ real.}$$

If $a_i \neq 0$, the equation is called a **Schrödinger type equation**. If $a_i = 0$, we again have our hyperbolic model equation.

4. **$\operatorname{Re} a < 0$** . Now

$$\operatorname{Re} \kappa \geq |a_r| \omega^2 - |b| |\omega|,$$

there is no upper bound for $\operatorname{Re} \kappa$, and the problem is **not well-posed for any b** .

EXERCISES

4.2.1. Consider the differential equation

$$\frac{\partial u}{\partial t} = \sum_{j=0}^4 a_j \frac{\partial^j u}{\partial x^j}.$$

Derive the condition for well-posedness corresponding to Eq. (4.2.2). Is it true that the problem is always well posed if $\operatorname{Re} a_4 < 0$?

4.3 FIRST-ORDER SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

Let

$$A = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ a_{21} & \dots & a_{2m} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mm} \end{bmatrix}, \quad u = \begin{bmatrix} u^{(1)}(x, t) \\ \vdots \\ u^{(m)}(x, t) \end{bmatrix}$$

be an $m \times m$ matrix and a vector function with m components, respectively. We consider the initial value problem

$$\boxed{\begin{aligned} u_t &= Au_x, \\ u(x, 0) &= f(x). \end{aligned}} \quad (4.3.1)$$

We then want to prove Theorem 4.3.1.

Theorem 4.3.1. *The initial value problem for the system (4.3.1) is well-posed if, and only if, the eigenvalues λ of A are real and there is a complete system of eigenvectors.*

Proof. To begin, let us prove that the eigenvalues of A must be real for the problem to be well posed. Let λ be an eigenvalue and ϕ the corresponding eigenvector. Then

$$u = e^{i\omega(\lambda t + x)} \phi$$

is a solution of Eq. (4.3.1) with initial data

$$u(x, 0) = e^{i\omega x} \phi.$$

Thus,

$$\|u(\cdot, t)\| = e^{\operatorname{Re}(i\omega\lambda t)} \|f(\cdot)\|.$$

If the problem is well posed, then

$$\operatorname{Re}(i\omega\lambda) \leq \alpha, \quad \alpha = \text{constant}$$

for all ω . This is only possible if λ is real.

Now assume that the eigenvalues are real and that there is a complete set of eigenvectors $\phi^{(1)}, \dots, \phi^{(m)}$. Let

$$S = (\phi^{(1)}, \dots, \phi^{(m)}).$$

Then S transforms A to diagonal form

$$S^{-1}AS = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}. \quad (4.3.2)$$

Introduce a new variable by

$$u = S\tilde{u}.$$

Then, Eq. (4.3.1) becomes

$$\tilde{u}_t = \Lambda \tilde{u}_x, \quad (4.3.3)$$

that is, we obtain m scalar equations

$$\tilde{u}_t^{(j)} = \lambda_j \tilde{u}_x^{(j)}.$$

From our previous results,

$$\|\tilde{u}^{(j)}(\cdot, t)\| = \|\tilde{u}^{(j)}(\cdot, 0)\|.$$

Therefore,

$$\|\tilde{u}(\cdot, t)\| = \|\tilde{u}(\cdot, 0)\|$$

and

$$\begin{aligned} \|u(\cdot, t)\| &= \|S\tilde{u}(\cdot, t)\| \leq |S| \|\tilde{u}(\cdot, t)\| = |S| \|\tilde{u}(\cdot, 0)\| \\ &= |S| \|S^{-1}\tilde{u}(\cdot, 0)\| \leq |S| |S^{-1}| \|u(\cdot, 0)\|. \end{aligned}$$

Thus, the problem is well posed if the eigenvalues are real, and there is a complete set of eigenvectors. In particular, **if A is Hermitian, it has real eigenvalues and a complete set of eigenvectors**; therefore, Eq. (4.3.1) is then well posed.

Now assume that the eigenvalues are real, but that there is not a complete set of eigenvectors.

We start with a typical case

$$u_t = \begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} u_x =: (\lambda I + J)u_x; \quad (4.3.4)$$

that is, A consists of one **Jordan block**. Let

$$u(x, 0) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{u}(\omega, 0).$$

Then the solution of our problem is given by

$$u(x, t) = \frac{1}{\sqrt{2\pi}} e^{i\omega(\lambda I + J)t} e^{i\omega x} \hat{u}(\omega, 0),$$

and, therefore,

$$\|u(\cdot, t)\| \leq |e^{i\omega(\lambda I + J)t}| \|u(\cdot, 0)\|.$$

Thus, the problem is well posed if we can find constants K, α such that, for all ω ,

$$|e^{i\omega(\lambda I + J)t}| \leq K e^{\alpha t}. \quad (4.3.5)$$

However, because the matrices λI and J commute,

$$\begin{aligned} |e^{i\omega(\lambda I + J)t}| &= |e^{i\omega\lambda It} e^{i\omega Jt}| = |e^{i\omega\lambda It}| |e^{i\omega Jt}| \\ &= |e^{i\omega Jt}| = \left| \sum_{j=0}^{m-1} \frac{(i\omega)^j J^j t^j}{j!} \right|. \quad J^m = 0 \end{aligned}$$

(Observe that $J^j \equiv 0$ for $j \geq m$.)

The last expression grows like $|\omega|^{m-1}$. Therefore, we cannot find K, α such that Eq. (4.3.5) holds and the initial value problem is not well posed for Eq. (4.3.4).

Now consider the general case. There is a transformation S such that

$$S^{-1}AS = \begin{bmatrix} \lambda_1 I + J_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r I + J_r \end{bmatrix}$$

has block form and every block is a Jordan block. If all blocks are of dimension one, then the above matrix is diagonal and we have a complete set of eigenvectors. Otherwise, there is at least one Jordan block of dimension two, and the problem cannot be well posed. This proves the theorem.

Systems like Eq. (4.3.1), where the eigenvalues of A are real, are called **hyperbolic**. In particular, we have the following classification.

Definition 4.3.1. *The system in Eq. (4.3.1) is called symmetric hyperbolic if A is a Hermitian matrix. It is called strictly hyperbolic if the eigenvalues are real and distinct, it is called strongly hyperbolic if the eigenvalues are real and there exists a complete system of eigenvectors, and, finally, it is called weakly hyperbolic if the eigenvalues are real.*

From our previous results, the initial value problem is not well posed for weakly hyperbolic systems, which are not strongly hyperbolic. It is well posed for strongly hyperbolic systems. Also, strictly hyperbolic and symmetric hyperbolic systems are subclasses of strongly hyperbolic systems, see Figure 4.3.1.

We now prove that lower order terms do not destroy the well-posedness of the initial value problem for strongly hyperbolic systems.

Lemma 4.3.1. *Let $y \in C^1$ satisfy the differential inequality*

$$\frac{dy}{dt} \leq \alpha y, \quad \text{for } t \geq 0.$$

Then

$$y(t) \leq e^{\alpha t} y(0).$$

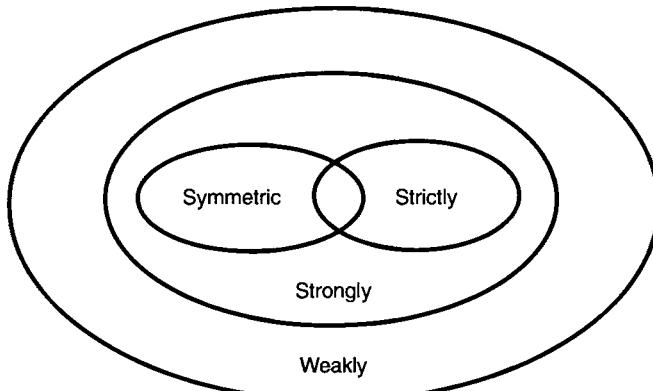


Figure 4.3.1.

Proof. $\tilde{y} = e^{-\alpha t} y$ satisfies

$$\frac{d\tilde{y}}{dt} \leq 0, \quad \text{that is, } \tilde{y}(t) \leq y(0)$$

and the desired estimate follows.

Now we can prove Theorem 4.3.2.

Theorem 4.3.2. Consider a strongly hyperbolic system (4.3.1) and perturb it with an undifferentiated term

$$\boxed{\begin{aligned} u_t &= Au_x + Bu, \\ u(x, 0) &= f(x), \end{aligned}} \quad (4.3.6)$$

where B is a constant $m \times m$ matrix. The problem (4.3.6) is well posed.

Proof. Let S be the transformation (4.3.2). Let

$$u = S\tilde{u}$$

and substitute into Eq. (4.3.6) to get

$$\begin{aligned} \tilde{u}_t &= \Lambda\tilde{u}_x + \tilde{B}\tilde{u}, & \tilde{B} &= S^{-1}BS, \\ \tilde{u}(x, 0) &= \tilde{f}(x), & \tilde{f}(x) &= S^{-1}f(x). \end{aligned} \quad (4.3.7)$$

Let $\hat{f} = e^{i\omega x}\hat{f}(\omega)$. We construct a simple wave solution. Substituting $\tilde{u}(x, t) = e^{i\omega x}\hat{u}(\omega, t)$ into Eq. (4.3.7) gives us

$$\begin{aligned} \hat{u}_t &= (i\omega\Lambda + \tilde{B})\hat{u}, \\ \hat{u}(\omega, 0) &= \hat{f}(\omega). \end{aligned}$$

Therefore,

$$\hat{u}(\omega, t) = e^{(i\omega\Lambda + \tilde{B})t}\hat{f}(\omega). \quad (4.3.8)$$

Also,

$$\begin{aligned}
 \frac{\partial}{\partial t} |\hat{u}|^2 &= \langle \hat{u}_t, \hat{u} \rangle + \langle \hat{u}, \hat{u}_t \rangle, \\
 &= \boxed{\langle i\omega\Lambda\hat{u}, \hat{u} \rangle + \langle \hat{u}, i\omega\Lambda\hat{u} \rangle} + \langle \tilde{B}\hat{u}, \hat{u} \rangle + \langle \hat{u}, \tilde{B}\hat{u} \rangle, \\
 &= \langle \tilde{B}\hat{u}, \hat{u} \rangle + \langle \hat{u}, \tilde{B}\hat{u} \rangle, \\
 &\leq \boxed{2\alpha|\hat{u}|^2}, \quad \alpha = |\tilde{B}|, \quad \text{Cauchy-Swartz}
 \end{aligned}$$

Therefore,

Integrate both side

$$|\hat{u}(\omega, t)|^2 = |e^{(i\omega\Lambda + \tilde{B})t} \hat{f}(\omega)|^2 \leq e^{2\alpha t} |\hat{f}(\omega)|^2,$$

that is,

$$\boxed{|e^{(i\omega\Lambda + \tilde{B})t}| \leq e^{\alpha t}}. \quad (4.3.9)$$

Now consider general smooth initial data. They can be expanded into a rapidly convergent Fourier series

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

By Eq. (4.3.8), the formal solution is

$$\tilde{u}(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} e^{(i\omega\Lambda + B)t} \hat{f}(\omega). \quad (4.3.10)$$

Using Eq. (4.3.9), we see that the series (4.3.10) converges rapidly for $t > 0$ and is a genuine solution of our problem. By Parseval's relation

$$\begin{aligned}
 \|\tilde{u}(\cdot, t)\|^2 &= \sum_{\omega=-\infty}^{\infty} |e^{(i\omega\Lambda + B)t} \hat{f}(\omega)|^2 \\
 &\leq e^{2\alpha t} \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = e^{2\alpha t} \|\tilde{f}(\cdot)\|^2.
 \end{aligned}$$

Thus, the problem is well posed in the transformed variables. For the original variables we obtain

$$\begin{aligned}\|u(\cdot, t)\| &= \|S\tilde{u}(\cdot, t)\| \leq |S| \|\tilde{u}(\cdot, t)\| \\ &\leq |S| e^{\alpha t} \|\tilde{f}(\cdot)\| = |S| e^{\alpha t} \|S^{-1}f(\cdot)\| \\ &\leq K e^{\alpha t} \|f(\cdot)\|, \quad K = |S| |S^{-1}|.\end{aligned}$$

Therefore, the problem is also well posed in the original variables.

EXERCISES

4.3.1. For which matrices A, B is the system

$$u_t = Au_x + Bu$$

energy conserving [i.e., $\|u(\cdot, t)\| = \|u(\cdot, 0)\|$]?

4.4. PARABOLIC SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

Let us consider second order systems

$$\begin{aligned}u_t &= Au_{xx} + Bu_x + Cu =: Pu, \\ u(x, 0) &= f(x).\end{aligned}\tag{4.4.1}$$

Definition 4.4.1. The system is called *parabolic* if the eigenvalues λ of A satisfy

$$\operatorname{Re} \lambda \geq \delta \quad \delta = \text{constant} > 0.$$

We find the following notation useful. Suppose $A = A^*$ is Hermitian, then we say that $A \geq 0$ if $\langle Av, v \rangle \geq 0$ for all vectors v . We also say $A \geq B$ if $A - B \geq 0$ when A and B are Hermitian. We now want to prove the following theorem.

Theorem 4.4.1. The initial value problem is well posed for parabolic differential equations.

Proof. Let $f(x) = (1/\sqrt{2\pi}) e^{i\omega x} \hat{f}(\omega)$. We construct a simple wave solution

$$u(x, t) = \frac{1}{\sqrt{2\pi}} e^{i\omega x} \hat{u}(\omega, t).\tag{4.4.2}$$

Substituting Eq. (4.4.2) into Eq. (4.4.1) gives us

$$\begin{aligned}\hat{u}_t &= (-\omega^2 A + i\omega B + C)\hat{u} =: \hat{P}(i\omega)\hat{u}, \\ \hat{u}(\omega, 0) &= \hat{f}(\omega);\end{aligned}\tag{4.4.3}$$

that is,

$$\hat{u}(\omega, t) = e^{\hat{P}(i\omega)t} \hat{f}(\omega).\tag{4.4.4}$$

Now assume that

$$A + A^* \geq \delta I, \quad \delta > 0.\tag{4.4.5}$$

Then

$$\begin{aligned}\hat{P}(i\omega) + \hat{P}^*(i\omega) &\leq (-\omega^2 \delta + 2|B|\omega + 2|C|)I, \\ &\leq \left(\frac{|B|^2}{\delta} + 2|C| \right) I =: 2\alpha I.\end{aligned}\tag{4.4.6}$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial t} \langle \hat{u}, \hat{u} \rangle &= \langle \hat{u}_t, \hat{u} \rangle + \langle \hat{u}, \hat{u}_t \rangle = \langle \hat{P}(i\omega)\hat{u}, \hat{u} \rangle + \langle \hat{u}, \hat{P}(i\omega)\hat{u} \rangle, \\ &= \langle \hat{u}, (\hat{P}^*(i\omega) + \hat{P}(i\omega))\hat{u} \rangle \leq 2\alpha \langle \hat{u}, \hat{u} \rangle,\end{aligned}$$

that is, by Lemma 4.3.1,

$$|\hat{u}(\omega, t)|^2 \leq e^{2\alpha t} |\hat{u}(\omega, 0)|^2 = e^{2\alpha t} |\hat{f}(\omega)|^2.$$

Therefore,

$$|e^{\hat{P}(i\omega)t}| \leq e^{\alpha t}.\tag{4.4.7}$$

Now consider general smooth initial data. They can be expanded into a rapidly convergent Fourier series

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

By Eq. (4.4.4), the formal solution is

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} e^{\hat{P}(i\omega)t} \hat{f}(\omega). \quad (4.4.8)$$

By Eq. (4.4.7), series (4.4.8) also converges rapidly for $t > 0$ and is, therefore, a genuine solution of our problem. By Parseval's relation

$$\begin{aligned} \|u(\cdot, t)\|^2 &= \sum_{\omega=-\infty}^{\infty} |e^{\hat{P}(i\omega)t} \hat{f}(\omega)|^2, \\ &\leq e^{2\alpha t} \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = e^{2\alpha t} \|f(\cdot)\|^2. \end{aligned}$$

We have proved that there exists a smooth solution which satisfies the desired estimate. We now show that such a solution is unique. Let u be any smooth solution of Eq. (4.4.1). It can be expanded into a rapidly converging Fourier series

$$\begin{aligned} u(x, t) &= \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{u}(\omega, t), \\ \hat{u}(\omega, t) &= \frac{1}{\sqrt{2\pi}} (e^{i\omega x}, u(x, t)), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega). \end{aligned}$$

The differential equation gives us

$$\frac{\partial}{\partial t} (e^{i\omega x}, u) = A(e^{i\omega x}, u_{xx}) + B(e^{i\omega x}, u_x) + C(e^{i\omega x}, u).$$

Using integration by parts, we obtain

$$(e^{i\omega x}, u_x) = \int_0^{2\pi} e^{-i\omega x} u_x dx = i\omega \int_0^{2\pi} e^{-i\omega x} u dx = i\omega (e^{i\omega x}, u).$$

Correspondingly,

$$(e^{i\omega x}, u_{xx}) = -\omega^2 (e^{i\omega x}, u).$$

Therefore,

$$\frac{\partial}{\partial t} \hat{u}(\omega, t) = \hat{P}(i\omega) \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) = \hat{f}(\omega);$$

that is, we obtain Eq. (4.4.4) again. This proves uniqueness.

We now show that inequality (4.4.5) can always be obtained for parabolic systems by a change of the dependent variables. By Schur's Lemma (see Appendix A1, Lemma A.1.3), there is a unitary transformation U such that

$$U^* A U = \begin{bmatrix} \lambda_1 & \tilde{a}_{12} & \cdots & \cdots & \tilde{a}_{1m} \\ & \lambda_2 & \tilde{a}_{23} & \cdots & \tilde{a}_{2m} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \tilde{a}_{m-1,m} \\ 0 & & & & \lambda_m \end{bmatrix}$$

has upper triangular form. Let

$$D = \begin{bmatrix} 1 & & & 0 \\ & d & & \\ & & \ddots & \\ 0 & & & d^{m-1} \end{bmatrix}, \quad d > 0,$$

be a diagonal matrix. Then

$$\tilde{A} := D^{-1} U^* A U D = \Lambda + G,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_m \end{bmatrix}, \\ G = \begin{bmatrix} 0 & d\tilde{a}_{12} & \cdots & \cdots & d^{m-1}\tilde{a}_{1m} \\ 0 & 0 & d\tilde{a}_{23} & \cdots & d^{m-2}\tilde{a}_{2m} \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & d\tilde{a}_{m-1,m} \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

For sufficiently small d ,

$$\tilde{A} + \tilde{A}^* = \Lambda + \Lambda^* + G + G^* \geq 2\delta I + G + G^* \geq \delta I.$$

Introduce a new variable by

$$u = UD\tilde{u}$$

and substitute into Eq. (4.4.1). The new system has the form

$$\tilde{u}_t = \tilde{A}\tilde{u}_{xx} + \tilde{B}\tilde{u}_x + \tilde{C}\tilde{u}$$

and satisfies Eq. (4.4.5). As for hyperbolic systems, the change of variable does not affect their well-posedness. The theorem is proved.

As in the scalar case for parabolic differential equations, the lower order terms do not have any influence on well-posedness. Also, by Eq. (4.4.6), for large ω

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) \lesssim -\delta\omega^2 I$$

and, therefore,

$$|e^{\hat{P}(i\omega)t}| \leq e^{-\delta\omega^2 t}$$

implies that the high frequency part of the solution is rapidly damped.

EXERCISES

4.4.1. Prove that there are positive constants δ, K such that the solutions to a parabolic system $u_t = Au_{xx}$ satisfy

$$\|u(\cdot, t)\|^2 + \delta \int_0^t \|u_x(\cdot, \xi)\|^2 d\xi \leq K\|u(\cdot, 0)\|^2. \quad (4.4.9)$$

4.4.2. Is it true that Eq. (4.4.9) holds with the same constants δ, K , if the system is changed to $u_t = Au_{xx} + Bu_x + Cu$, where B is Hermitian and C is skew-Hermitian?

4.5. GENERAL SYSTEMS WITH CONSTANT COEFFICIENTS

We now consider general systems of the type of Eq. (4.1.8)

$$\boxed{u_t = P\left(\frac{\partial}{\partial x}\right) u, \quad u(x, 0) = f(x),} \quad (4.5.1)$$

with constant coefficients. Let $\omega = (\omega_1, \dots, \omega_d)$ denote the real wave number vector and assume that

$$f(x) = (2\pi)^{-(d/2)} e^{i\langle\omega, x\rangle} \hat{f}(\omega), \quad \langle\omega, x\rangle = \sum_{j=1}^d \omega_j x^{(j)}.$$

As before, we construct simple wave solutions.

$$\boxed{u(x, t) = (2\pi)^{-(d/2)} e^{i\langle\omega, x\rangle} \hat{u}(\omega, t).} \quad (4.5.2)$$

Substituting Eq. (4.5.2) into Eq. (4.5.1) gives us

$$\begin{aligned} \frac{\partial}{\partial t} \hat{u}(\omega, t) &= \hat{P}(i\omega) \hat{u}(\omega, t) \\ \hat{u}(\omega, 0) &= \hat{f}(\omega). \end{aligned} \quad (4.5.3)$$

$\hat{P}(i\omega)$ is called the symbol, or Fourier transform, of the differential operator $P(\partial/\partial x)$ and is obtained by replacing $\partial/\partial x^{(j)}$ by $i\omega_j$. Thus, $\hat{P}(i\omega)$ is an $m \times m$ matrix whose coefficients are polynomials in $i\omega_j$. Equations (4.5.3) are a system of ordinary differential equations with constant coefficients, and the solution is given by

$$\hat{u}(\omega, t) = e^{\hat{P}(i\omega)t} \hat{f}(\omega). \quad (4.5.4)$$

Theorem 4.5.1. *The initial value problem (4.5.1) is well posed if and only if there are constants K, α such that, for all ω ,*

$$\boxed{|e^{\hat{P}(i\omega)t}| \leq K e^{\alpha t}.} \quad (4.5.5)$$

Proof. If the problem is well posed, then there are constants K, α such that, for all solutions,

$$\|u(\cdot, t)\| \leq Ke^{\alpha t} \|u(\cdot, 0)\|. \quad (4.5.6)$$

Therefore, by Eq. (4.5.4), inequality (4.5.5) is a necessary condition.

Now assume that Eq. (4.5.5) is valid and consider Eq. (4.5.1) with general smooth initial data. The data can be expanded into a rapidly convergent Fourier series

$$f(x) = (2\pi)^{-(d/2)} \sum_{\omega} e^{i\langle \omega, x \rangle} \hat{f}(\omega).$$

By Eq. (4.5.4), the formal solution of our problem is

$$u(x, t) = (2\pi)^{-(d/2)} \sum_{\omega} e^{i\langle \omega, x \rangle} e^{\hat{P}(i\omega)t} \hat{f}(\omega). \quad (4.5.7)$$

By Eq. (4.5.5), the series (4.5.7) also converges rapidly for $t > 0$ and is, therefore, a genuine solution. By Parseval's relation,

$$\begin{aligned} \|u(\cdot, t)\|^2 &= \sum_{\omega} |\hat{u}(\omega, t)|^2 = \sum_{\omega} |e^{\hat{P}(i\omega)t} \hat{f}(\omega)|^2, \\ &\leq Ke^{\alpha t} \sum_{\omega} |\hat{f}(\omega)|^2 = Ke^{\alpha t} \|f(\cdot)\|^2. \end{aligned}$$

We have constructed a smooth solution that satisfies the desired estimate. Because uniqueness can be proved in the same way as in the previous section, the theorem follows.

We now discuss the algebraic conditions that guarantee that Eq. (4.5.5) holds.

Theorem 4.5.2. The Petrovskii condition. A necessary condition for well-posedness is that, for all ω , the eigenvalues λ of $\hat{P}(i\omega)$ satisfy the inequality

$$\operatorname{Re} \lambda \leq \alpha. \quad (4.5.8)$$

Proof. Let λ be an eigenvalue and ϕ be the corresponding eigenvector. Then

$$e^{\hat{P}(i\omega)t} \phi = e^{\lambda t} \phi,$$

Use the arbitrariness of omega
 $u_t=Au_x: \operatorname{Im} \lambda(A)=0$
 $u_t=Au_{xx}: \operatorname{Re} \lambda(A)>=0$

and the theorem follows.

Theorem 4.5.3. Assume that the *Petrovskii condition* is satisfied and that there is a constant K and a transformation $S(\omega)$ with

$$|S(\omega)| |S^{-1}(\omega)| \leq K \quad (4.5.9)$$

for every ω and $S^{-1}(\omega)\hat{P}(i\omega)S(\omega)$ has diagonal form. Then the initial value problem (4.5.1) is well posed.

Proof.

$$\begin{aligned} |e^{\hat{P}(i\omega)t}| &= |S S^{-1} e^{\hat{P}(i\omega)t} S S^{-1}| \leq |S| |S^{-1}| |e^{\Lambda t}|, \\ &\leq K e^{\alpha t}. \end{aligned}$$

This proves the theorem.

Theorem 4.5.4. The initial value problem is *well-posed* if there is a constant α such that, for all ω ,

$$\hat{P}(i\omega) + \hat{P}^*(i\omega) \leq 2\alpha I. \quad (4.5.10)$$

Proof. By Eq. (4.5.3),

$$\begin{aligned} \frac{d}{dt} \langle \hat{u}, \hat{u} \rangle &= \langle \hat{P}\hat{u}, \hat{u} \rangle + \langle \hat{u}, \hat{P}\hat{u} \rangle, \\ &= \langle \hat{u}, (\hat{P}^* + \hat{P})\hat{u} \rangle \leq 2\alpha \langle \hat{u}, \hat{u} \rangle. \end{aligned}$$

Therefore, by Lemma 4.3.1,

$$|\hat{u}(\omega, t)|^2 \leq e^{2\alpha t} |\hat{u}(\omega, 0)|^2 = e^{2\alpha t} |\hat{f}(\omega)|^2,$$

that is, by Eq. (4.5.4),

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

This proves the theorem.

We can also express the above result in another way.

Definition 4.5.1. The differential operator $P(\partial/\partial x)$ is called *semibounded* if there is a constant α such that, for all smooth functions $w(x)$,

$$(w, Pw) + (Pw, w) \leq 2\alpha(w, w). \quad (4.5.11)$$

REMARK. This definition might be called semibounded from above. Note that it does not imply that the operator is bounded.

Next, we have the following theorem.

Theorem 4.5.5. *P is semibounded if, and only if, Eq. (4.5.10) holds.*

Proof. Expand w in a Fourier series

$$w = (2\pi)^{-(d/2)} \sum_{\omega} e^{i(\omega, x)} \hat{w}(\omega).$$

Then,

$$Pw = (2\pi)^{-(d/2)} \sum_{\omega} e^{i(\omega, x)} \hat{P}(i\omega) \hat{w}(\omega).$$

Therefore, Parseval's relation gives us

$$(w, Pw) + (Pw, w) = \sum_{\omega} \langle \hat{w}(\omega), (\hat{P}^*(i\omega) + \hat{P}(i\omega)) \hat{w}(\omega) \rangle. \quad (4.5.12)$$

If Eq. (4.5.10) holds, then Eq. (4.5.11) follows. The converse is also true. If Eq. (4.5.11) is satisfied, then it must hold for every simple wave $w = (2\pi)^{-(d/2)} e^{i(\omega, x)} \hat{w}(\omega)$, that is,

$$\langle \hat{w}(\omega), (\hat{P}^*(i\omega) + \hat{P}(i\omega)) \hat{w}(\omega) \rangle \leq 2\alpha |\hat{w}(\omega)|^2$$

for all vectors \hat{w} . Therefore, Eq. (4.5.10) follows, and we have proved the theorem.

If one knows that an operator is semibounded, then one can prove the energy estimate in Eq. (4.1.7) immediately.

Theorem 4.5.6. *If P is semibounded, then the solutions of Eq. (4.5.1) satisfy the estimate*

$$\|u(\cdot, t)\| \leq e^{\alpha t} \|f(\cdot)\|. \quad (4.5.13)$$

Proof.

$$\begin{aligned} \frac{d}{dt} (u, u) &= (u_t, u) + (u, u_t) = (Pu, u) + (u, Pu) \\ &\leq 2\alpha(u, u), \end{aligned}$$

and Eq. (4.5.13) follows.

EXAMPLE. As in Section 4.3, we call a first-order system

$$\frac{\partial u}{\partial t} = \sum_{j=1}^d A_j \frac{\partial u}{\partial x^{(j)}}$$

symmetric hyperbolic, if the matrices A_j are Hermitian, that is, if $A_j = A_j^*$, $j = 1, 2, \dots, d$. Now

$$P(i\omega) = i \sum_{j=1}^d A_j \omega_j$$

and

$$P(i\omega) + P^*(i\omega) = 0.$$

Thus, the initial-value problem is well posed and $P(\partial/\partial x)$ is semibounded.

The above result can be generalized.

Theorem 4.5.7. Assume that there are constants $\alpha, K > 0$ and, for every ω , a positive Hermitian matrix $\hat{H}(\omega) = \hat{H}^*(\omega)$ with

$$K^{-1}I \leq \hat{H}(\omega) \leq KI. \quad (4.5.14)$$

such that

$$\hat{H}(\omega)\hat{P}(i\omega) + \hat{P}^*(i\omega)\hat{H}(\omega) \leq 2\alpha\hat{H}(\omega). \quad (4.5.15)$$

Then the initial value problem is well posed.

Proof. By Eq. (4.5.3)

$$\begin{aligned}\frac{d}{dt} \langle \hat{u}, \hat{H} \hat{u} \rangle &= \langle \hat{P} \hat{u}, \hat{H} \hat{u} \rangle + \langle \hat{u}, \hat{H} \hat{P} \hat{u} \rangle, \\ &= \langle \hat{u}, (\hat{P}^* \hat{H} + \hat{H} \hat{P}) \hat{u} \rangle \leq 2\alpha \langle \hat{u}, \hat{H} \hat{u} \rangle.\end{aligned}\tag{4.5.15}$$

Therefore, by Lemma 4.3.1,

$$\langle \hat{u}(\omega, t), \hat{H}(\omega) \hat{u}(\omega, t) \rangle \leq e^{2\alpha t} \langle \hat{u}(\omega, 0), \hat{H}(\omega) \hat{u}(\omega, 0) \rangle;$$

that is, by Eq. (4.5.14),

$$|\hat{u}(\omega, t)|^2 \leq K \langle \hat{u}(\omega, t), \hat{H} \hat{u}(\omega, t) \rangle \leq K^2 e^{2\alpha t} |\hat{u}(\omega, 0)|^2.$$

Thus,

$$|e^{P(i\omega)t}| \leq K e^{\alpha t},$$

and the initial value problem is well posed. The theorem now follows from Theorem 4.5.1.

One can prove that the conditions of Theorem 4.5.7 characterize well-posed problems. Without proof we state the following theorem.

Theorem 4.5.8. *The initial-value problem (4.5.1) is well posed if and only if we can construct Hermitian matrices such that conditions (4.5.14) and (4.5.15) hold.*

In applications, $\hat{H}(\omega)$ often has a very simple structure. Either $\hat{H}(\omega) \equiv I$ or \hat{H} is a diagonal matrix which defines a scaling of the dependent variables. We can again express our results another way. Let $\hat{H}(\omega)$ satisfy Eq. (4.5.14) and let

$$v = (2\pi)^{-(d/2)} \sum_{\omega} e^{i\langle \omega, x \rangle} \hat{v}(\omega), \quad w = (2\pi)^{-(d/2)} \sum_{\omega} e^{i\langle \omega, x \rangle} \hat{w}(\omega)$$

be two functions. We define a scalar product by

$$(v, w)_H = \sum_{\omega} \langle \hat{v}(\omega), \hat{H}(\omega) \hat{w}(\omega) \rangle.$$

The scalar product $(v, w)_H$ defines the H norm, $(v, v)_H^{1/2} = \|v\|_H$, which is equiv-

alent to the L_2 norm $\|v\|^2$, because, by Eq. (4.5.14),

$$K^{-1}(v, v) \leq (v, v)_H \leq K(v, v). \quad (4.5.16)$$

Now assume that Eq. (4.5.15) also holds. Observing that

$$Pv = (2\pi)^{-(d/2)} \sum_{\omega} e^{i\langle \omega, x \rangle} \hat{P}(i\omega) \hat{v}(\omega, t),$$

we obtain, from Parseval's relation and Eq. (4.5.15),

$$\begin{aligned} (Pv, v)_H + (v, Pv)_H &= \sum_{\omega} (\langle \hat{P}(i\omega) \hat{v}(\omega), \hat{H}(\omega) \hat{v}(\omega) \rangle \\ &\quad + \langle \hat{v}(\omega), \hat{H}(\omega) \hat{P}(i\omega) \hat{v}(\omega) \rangle), \\ &\leq \langle \hat{v}(\omega), (\hat{P}^*(i\omega) \hat{H}(\omega) + \hat{H}(\omega) \hat{P}(i\omega)) \hat{v}(\omega) \rangle, \\ &\leq 2\alpha \sum_{\omega} \langle \hat{v}(\omega), \hat{H}(\omega) \hat{v}(\omega) \rangle = 2\alpha \|v\|_H^2. \end{aligned}$$

Thus, the expression above is bounded from above. We generalize Definition 4.5.1 to the following definition.

Definition 4.5.2. *The operator P is called semibounded in the H norm if there is a constant α such that*

$$(Pv, v)_H + (v, Pv)_H \leq 2\alpha \|v\|_H^2 \quad (4.5.17)$$

for all smooth functions v .

The following theorem corresponds to Theorem 4.5.6.

Theorem 4.5.9. *If P is semibounded in the H norm, then the solutions of Eq. (4.5.1) satisfy the estimate*

$$K^{-1} \|u(\cdot, t)\|^2 \leq \|u(\cdot, t)\|_H^2 \leq e^{2\alpha t} \|f(\cdot)\|_H^2 \leq K e^{2\alpha t} \|f(\cdot)\|^2.$$

Proof. The theorem follows from

$$\frac{d}{dt} (u, u)_H = (Pu, u)_H + (u, Pu)_H \leq 2\alpha (u, u)_H.$$

EXERCISES

4.5.1. Consider the first order system $u_t = Au_x$. Is it possible that the Petrovskii condition (4.5.8) is satisfied for some constant $\alpha > 0$ but not for $\alpha = 0$?

4.5.2. Derive a matrix $\hat{H}(\omega)$ satisfying Eq. (4.5.14) and (4.5.15) for the system

$$u_t = \begin{bmatrix} 1 & 10 \\ 0 & 2 \end{bmatrix} u_x.$$

4.6. SEMIBOUNDED OPERATORS WITH VARIABLE COEFFICIENTS

In the previous section, we discussed semibounded operators with constant coefficients. Here we generalize the concept to differential equations

$$u_t = P\left(x, t, \frac{\partial}{\partial x}\right) u, \quad t \geq t_0,$$

$$u(x, t_0) = f(x),$$

(4.6.1)

with variable coefficients.

Definition 4.6.1. The differential operator $P(x, t, \partial/\partial x)$ is **semibounded** if, for any interval $t_p \leq t \leq T$, there is a constant α such that, for all sufficiently smooth functions w ,

$$2 \operatorname{Re}(w, Pw) = \left(w, P\left(\cdot, t, \frac{\partial}{\partial x}\right) w \right) + \left(P\left(\cdot, t, \frac{\partial}{\partial x}\right) w, w \right) \leq 2\alpha \|w\|^2.$$

(4.6.2)

As for systems with constant coefficients, we have the following theorem.

Theorem 4.6.1. If the operator $P(x, t, \partial/\partial x)$ is **semibounded**, then the solutions of Eq. (4.6.1) satisfy the estimate

$$\|u(\cdot, t)\| \leq e^{\alpha(t - t_0)} \|f(\cdot)\|.$$

(4.6.3)

We shall now give a number of examples. First, we need the following lemma.

Lemma 4.6.1. If u, v are periodic smooth vector functions, then

$$\left(u, \frac{\partial v}{\partial x^{(j)}} \right) = - \left(\frac{\partial u}{\partial x^{(j)}}, v \right), \quad j = 1, 2, \dots, d.$$

Proof. Let $x_- = (x^{(1)}, \dots, x^{(i-1)}, x^{(i+1)}, \dots, x^{(d)})$. Then

$$\left(u, \frac{\partial v}{\partial x^{(j)}} \right) = \int_0^{2\pi} \cdots \int_0^{2\pi} \left(\int_0^{2\pi} \left\langle u, \frac{\partial v}{\partial x^{(j)}} \right\rangle dx^{(j)} \right) dx_-.$$

Integration by parts gives us

$$\begin{aligned} \int_0^{2\pi} \left\langle u, \frac{\partial v}{\partial x^{(j)}} \right\rangle dx^{(j)} &= \langle u, v \rangle \Big|_{x^{(j)}=0}^{2\pi} - \int_0^{2\pi} \left\langle \frac{\partial u}{\partial x^{(j)}}, v \right\rangle dx^{(j)}, \\ &= - \int_0^{2\pi} \left\langle \frac{\partial u}{\partial x^{(j)}}, v \right\rangle dx^{(j)}, \end{aligned}$$

and the lemma follows.

4.6.1. Symmetric Hyperbolic First Order Systems

We consider

$$\frac{\partial u}{\partial t} = P \left(x, t, \frac{\partial}{\partial t} \right) u := \sum_{j=1}^d B_j(x, t) \frac{\partial u}{\partial x^{(j)}} + C(x, t)u.$$

Here C is a general smooth periodic matrix function and $B_j = B_j^*$ are smooth periodic Hermitian matrices. P is semibounded because

$$\begin{aligned} \left(u, B_j \frac{\partial u}{\partial x^{(j)}} \right) &= \left(B_j u, \frac{\partial u}{\partial x^{(j)}} \right) \\ &= - \left(\frac{\partial (B_j u)}{\partial x^{(j)}}, u \right) \\ &= - \left(B_j \frac{\partial u}{\partial x^{(j)}}, u \right) - \left(\frac{\partial B_j}{\partial x^{(j)}} u, u \right). \end{aligned}$$

Therefore,

$$(u, Pu) + (Pu, u) = - \sum_{j=1}^d \left(\frac{\partial B_j}{\partial x^{(j)}} \underline{u}, \underline{u} \right) + (Cu, u) + (u, Cu) \\ \leq \left(\sum_{j=1}^d \left| \frac{\partial B_j}{\partial x^{(j)}} \right|_\infty + |C + C^*|_\infty \right) \|u\|^2$$

shows that Eq. (4.6.2) is satisfied.

Throughout this book, we often use differential equations from fluid dynamics to illustrate various concepts. The velocity of the flow has the components u, v, w in the x, y, z directions, respectively. The density is denoted by ρ and the pressure by p . If viscous and heat conduction effects are neglected, the flow is described by the momentum equations

$$\begin{aligned} u_t + uu_x + vu_y + wu_z + \rho^{-1}p_x &= 0 \\ v_t + uv_x + vv_y + wv_z + \rho^{-1}p_y &= 0 \\ w_t + uw_x + vw_y + ww_z + \rho^{-1}p_z &= 0 \end{aligned} \quad (4.6.4a)$$

and the continuity equation

$$\rho_t + (u\rho)_x + (v\rho)_y + (w\rho)_z = 0.$$

This system is called the *Euler equations*. There are five dependent variables; so we need another equation to close the system. Under certain conditions one can assume that p is only a function of ρ and that we have an equation of state

$$p = G(\rho), \quad a^2 := \frac{dG}{d\rho} \geq \delta > 0. \quad (4.6.4b)$$

For reasons that will become clear later, a is called the *speed of sound*. The system is nonlinear, and all the theory we have presented up to now only deals with linear equations. As we shall see later, the so called linearized equations play a fundamental role for well-posedness. These equations are obtained in the following way. Assume that there is a smooth solution of the original nonlinear problem. We denote this solution by (U, V, W, R) . Assume that a small perturbation is added to the initial data, so that the solution is perturbed by $\epsilon(u', v', w', p')$ where ϵ is small. We want to derive a simplified linear system for this perturbation. Consider the first equation of the Euler equations (4.6.4), and substitute the perturbed solution into it:

$$(U + \epsilon u')_t + (U + \epsilon u')(U + \epsilon u')_x + (V + \epsilon v')(U + \epsilon u')_y \\ + (W + \epsilon w')(U + \epsilon u')_z + (R + \epsilon \rho')^{-1}(G(R + \epsilon \rho'))_x = 0.$$

By assumption (U, V, W, R) is a solution, so we get

$$\epsilon(u'_t + Uu'_x + Vu'_y + Wu'_z + R^{-1}a^2(R)\rho'_x \\ + U_xu' + U_yv' + U_zw' - R^{-2}(G(R))_x\rho') + \mathcal{O}(\epsilon^2) = 0.$$

By neglecting the nonlinear quadratic terms, we obtain the linearized equation. For convenience, we consider uniform flow in the z direction; that is, $w = u_z = v_z = \rho_z = p_z = 0$ in Eqs. (4.6.4). Then the full linearized system is

$$\begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_t + \begin{bmatrix} U & 0 & \frac{a^2(R)}{R} \\ 0 & U & 0 \\ R & 0 & U \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_x \\ + \begin{bmatrix} V & 0 & 0 \\ 0 & V & \frac{a^2(R)}{R} \\ 0 & R & V \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix}_y + C \begin{bmatrix} u' \\ v' \\ \rho' \end{bmatrix} = 0,$$

where C is a matrix depending on U, V, R . This system is not symmetric hyperbolic. However, we can make it symmetric by introducing

$$\tilde{\rho} = \frac{a(R)}{R} \rho'$$

as a new function.

The new system is

$$\begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_t + \begin{bmatrix} U & 0 & a(R) \\ 0 & U & 0 \\ a(R) & 0 & U \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_x \\ + \begin{bmatrix} V & 0 & 0 \\ 0 & V & a(R) \\ 0 & a(R) & V \end{bmatrix} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix}_y + \tilde{C} \begin{bmatrix} u' \\ v' \\ \tilde{\rho} \end{bmatrix} = 0. \quad (4.6.5)$$

4.6.2. Parabolic Systems

We call the system

$$\frac{\partial u}{\partial t} = P_2 \left(x, t, \frac{\partial}{\partial x} \right) u := \sum_{i,j=1}^d \frac{\partial}{\partial x^{(j)}} \left(A_{ij} \frac{\partial}{\partial x^{(i)}} u \right)$$

strongly parabolic if, for all smooth w and all t , there is a constant $\delta > 0$ such that

$$\sum_{i,j=1}^d \left(\left(\frac{\partial w}{\partial x^{(j)}}, A_{ij} \frac{\partial w}{\partial x^{(i)}} \right) + \left(A_{ij} \frac{\partial w}{\partial x^{(i)}}, \frac{\partial w}{\partial x^{(j)}} \right) \right) \geq \delta \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2. \quad (4.6.6)$$

Clearly, if Eq. (4.6.6) holds, then P_2 is semibounded because

$$(w, P_2 w) + (P_2 w, w) \leq -\delta \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2.$$

If $A_{ij} \equiv 0$ for $i \neq j$, then Eq. (4.6.6) becomes

$$\sum_{j=1}^d \left(\frac{\partial w}{\partial x^{(j)}}, (A_{jj} + A_{jj}^*) \frac{\partial w}{\partial x^{(j)}} \right) \geq \delta \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2. \quad (4.6.7)$$

Equation (4.6.7) holds if, and only if, the eigenvalues κ of $A_{jj} + A_{jj}^*$ satisfy $\kappa \geq \delta$. A simple example is the heat equation

$$\frac{\partial u}{\partial t} = \sum_{j=1}^d \frac{\partial}{\partial x^{(j)}} \left(\alpha \frac{\partial}{\partial x^{(j)}} u \right).$$

Let

$$P_1 = \sum_{j=1}^d B_j(x, t) \frac{\partial}{\partial x^{(j)}} + C(x, t)$$

be any first-order operator with smooth coefficients and let P_2 satisfy Eq.

(4.6.6). The operator $P_2 + P_1$ is then semibounded. We have

$$\begin{aligned}
 & (w, P_2 w) + (P_2 w, w) + (w, P_1 w) + (P_1 w, w) \\
 & \leq -\delta \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2 + 2\|w\| \|P_1 w\|, \\
 & \leq -\delta \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2 + 2K_1 \|w\|^2 + \sum_{j=1}^d \frac{2K_2}{\sqrt{\delta/2}} \|w\| \cdot \sqrt{\frac{\delta}{2}} \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2, \\
 & \leq \left(2K_1 + 2K_2^2 \frac{\delta}{\delta} \right) \|w\|^2 - \frac{\delta}{2} \sum_{j=1}^d \left\| \frac{\partial w}{\partial x^{(j)}} \right\|^2. \tag{4.6.8}
 \end{aligned}$$

K_1 and K_2 depend only on bounds for B_j and C .

The examples above show that, for symmetric hyperbolic systems and for strongly parabolic systems, zero-order and first-order terms, respectively, do not affect the semiboundedness of the operator.

4.6.3. Mixed Hyperbolic-Parabolic Systems

Consider a **strongly parabolic** system

$$u_t = P_2 u, \tag{4.6.9}$$

and a **symmetric-hyperbolic** system

$$v_t = P_1 v. \tag{4.6.10}$$

Here the vectors u and v do not necessarily have the same number of components.

Let $Q^{lm} = \sum_{j=1}^d B_j^{lm} \partial/\partial x^{(j)} + C^{lm}$, $l = 1, 2; m = 1, 2$, denote general first-order operators. Then the coupled system

$$\begin{bmatrix} u \\ v \end{bmatrix}_t = \begin{bmatrix} P_2 + Q^{11} & Q^{12} \\ Q^{21} & P_1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \tag{4.6.11}$$

has a spatial differential operator which is semibounded. Let $w^{(1)}$ and $w^{(2)}$ be vector functions with the same number of components as u and v , respectively. We have, using integration by parts,

$$\begin{aligned}
& (w^{(1)}, Q^{12} w^{(2)}) + (Q^{12} w^{(2)}, w^{(1)}) \\
& \leq \text{constant} \left(\sum_{j=1}^d \left\| \frac{\partial w^{(1)}}{\partial x^{(j)}} \right\| \|w^{(2)}\| + \|w^{(1)}\| \|w^{(2)}\| \right) \\
& \leq \delta_1 \sum_{j=1}^d \left\| \frac{\partial w^{(1)}}{\partial x^{(j)}} \right\|^2 + K_1 \|w^{(1)}\|^2 + K_2(\delta_1) \|w^{(2)}\|^2. \quad (4.6.12)
\end{aligned}$$

The constant δ_1 can be chosen arbitrarily small. The same type of inequality follows immediately for

$$(w^{(2)}, Q^{21} w^{(1)}) + (Q^{21} w^{(1)}, w^{(2)}).$$

For $\operatorname{Re}(w^{(1)}, (P_2 + Q^{11})w^{(1)})$ we use an inequality of the form (4.6.8), i.e., we have a negative term containing the derivatives of $w^{(1)}$ which cancels the corresponding term in (4.6.12) if δ_1 is chosen sufficiently small. We know that P_1 is semibounded, thus the semiboundedness follows for the whole system (4.6.11).

As an application, we consider the ***Navier-Stokes equations***. These equations describe viscous flow and are obtained by adding extra terms to the momentum equations in (4.6.4). With $\mu, \mu > 0, \mu', \mu' \geq 0$ denoting constant viscosity coefficients we have, in two space dimensions,

$$\begin{aligned}
\rho(u_t + uu_x + vu_y) + p_x &= \mu \Delta u + \mu' \frac{\partial}{\partial x} (u_x + v_y), \\
\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \\
\rho(v_t + uv_x + vv_y) + p_y &= \mu \Delta v + \mu' \frac{\partial}{\partial y} (u_x + v_y), \\
\rho_t + u\rho_x + v\rho_y + \rho(u_x + v_y) &= 0, \\
p &= G(\rho).
\end{aligned}$$

The linearized equations are obtained in the same way as they were above for the Euler equations. Except for zero-order terms, they have the form

$$\begin{aligned}
u_t + U u_x + V u_y + \frac{a^2(R)}{R} \rho_x &= \frac{\mu}{R} \Delta u + \frac{\mu'}{R} \frac{\partial}{\partial x} (u_x + v_y), \\
v_t + U v_x + V v_y + \frac{a^2(R)}{R} \rho_y &= \frac{\mu}{R} \Delta v + \frac{\mu'}{R} \frac{\partial}{\partial y} (u_x + v_y), \\
\rho_t + U \rho_x + V \rho_y + R(u_x + v_y) &= 0. \quad (4.6.13)
\end{aligned}$$

We obtain the decoupled system by neglecting all first-order terms in the first two equations and $R(u_x + v_y)$ in the third equation. We can write

$$u_t = \frac{\mu}{R} \Delta u + \frac{\mu'}{R} \frac{\partial}{\partial x} (u_x + v_y), \quad (4.6.14a)$$

$$v_t = \frac{\mu}{R} \Delta v + \frac{\mu'}{R} \frac{\partial}{\partial y} (u_x + v_y), \quad (4.6.14b)$$

$$\rho_t + U\rho_x + V\rho_y = 0. \quad (4.6.14c)$$

Integration by parts gives us

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} (\|u\|^2 + \|v\|^2) &= - \left\{ \left(\mathbf{u}_x, \frac{\mu}{R} \mathbf{u}_x \right) + \left(\mathbf{u}_y, \frac{\mu}{R} \mathbf{u}_y \right) \right. \\ &\quad \left. + \left(u_x + v_y, \frac{\mu'}{R} (u_x + v_y) \right) \right\} \\ &\leq -\delta (\|u_x\|^2 + \|u_y\|^2 + \|v_x\|^2 + \|v_y\|^2), \end{aligned}$$

where we have used the notation $\mathbf{u} = (u, v)^T$. Thus, Eqs. (4.6.14a) and (4.6.14b) are strongly parabolic, and Eq. (4.6.14c) is scalar hyperbolic. Therefore, the differential operator in space for the linearized Navier–Stokes equations is semi-bounded.

We have shown that for large classes of initial-value problems the energy estimate (4.1.10) holds. One can show that these problems also have unique solutions, and we obtain Theorem 4.6.2.

Theorem 4.6.2. *The initial-value problem for symmetric hyperbolic, strongly parabolic, and mixed symmetric hyperbolic—strongly parabolic systems is well posed.*

EXERCISES

4.6.1. Let $B(x, t)$ be a Hermitian matrix and let $C(x, t)$ be a skew-Hermitian matrix. Prove that the system

$$u_t = (Bu)_x + Bu_x + Cu$$

is energy conserving, that is,

$$\|u(\cdot, t)\| = \|u(\cdot, 0)\|.$$

4.6.2. Derive the exact form of the matrix \tilde{C} in the linearized Euler equations (4.6.5).

- 4.6.3.** Consider the linearized one-dimensional Euler equations, where $U = 0$ and R is a constant. Prove that the system represents two “sound-waves” moving with the velocities $\pm a(R)$.

4.7. THE SOLUTION OPERATOR AND DUHAMEL’S PRINCIPLE

In applications, the differential equations often contain a forcing function $F(x, t) \in C^\infty$, which is 2π -periodic in every space dimension. In this case, the differential equations have the form

$$\begin{aligned} u_t &= P\left(x, t, \frac{\partial}{\partial x}\right) u + F(x, t), \quad t \geq t_0, \\ u(x, t_0) &= f(x). \end{aligned} \tag{4.7.1}$$

The appropriate generalization of Definition 4.1.1 is as follows.

Definition 4.7.1. *The problem (4.7.1) is well posed if, for every t_0 and every $f \in C^\infty(x)$, $F \in C^\infty(x, t)$,*

1. *There exists a unique solution $u \in C^\infty(x, t)$ which is 2π -periodic in every space dimension,*
2. *There are constants α and K independent of f , F , and t_0 such that*

$$\|u(\cdot, t)\| \leq K e^{\alpha(t - t_0)} (\|u(\cdot, t_0)\| + \max_{t_0 \leq \tau \leq t} \|F(\cdot, \tau)\|).$$

Thus, we require that we can solve Eq. (4.7.1) for every smooth F and that the solutions also depend continuously on F ; that is, if $\|F\|$ is small, then its effect on the solution is small. In this section, we want to show that Definitions 4.7.1 and 4.1.1 are equivalent. If we can solve the homogeneous equations (4.1.8) and the estimate (4.1.10) holds, then we can also solve the inhomogeneous system (4.7.1) and our estimate is valid. We begin by proving a generalization of Lemma 4.3.1 that will be used frequently throughout the rest of the book.

Lemma 4.7.1. *Let α be a constant, $\beta(t)$ a bounded function, and $u(t) \in C^1(t)$ a function satisfying*

$$\frac{du}{dt} \leq \alpha u + \beta(t), \quad t \geq t_0.$$

Then

$$\begin{aligned} |u(t)| &\leq e^{\alpha(t-t_0)}|u(t_0)| + \int_{t_0}^t e^{\alpha(t-\tau)}|\beta(\tau)|d\tau \\ &\leq e^{\alpha(t-t_0)}|u(t_0)| + \varphi^*(\alpha, t - t_0) \max_{t_0 \leq \tau \leq t} |\beta(\tau)|, \end{aligned}$$

where

$$\varphi^*(\alpha, t) = \begin{cases} \frac{1}{\alpha} (e^{\alpha t} - 1), & \text{if } \alpha \neq 0, \\ t, & \text{if } \alpha = 0. \end{cases} \quad (4.7.2)$$

Proof. By introducing the new variable $v = e^{-\alpha(t-t_0)}u$, we obtain

$$e^{\alpha(t-t_0)} \left(\frac{dv}{dt} + \alpha v \right) = \frac{du}{dt} \leq \alpha e^{\alpha(t-t_0)}v + \beta.$$

Thus,

$$\frac{dv}{dt} \leq e^{-\alpha(t-t_0)}\beta;$$

that is,

$$|v(t)| \leq |v(t_0)| + \int_{t_0}^t e^{-\alpha(\tau-t_0)}|\beta(\tau)|d\tau,$$

and the lemma follows.

We first consider the following system of ordinary differential equations

$$\begin{aligned} \frac{du}{dt} &= A(t)u + F(t), \quad t \geq t_0, \\ u(t_0) &= u_0, \end{aligned} \quad (4.7.3)$$

where $A(t)$ and $F(t)$ are continuous matrix and vector functions, respectively. We now construct the solution of Eqs. (4.7.3) using solutions of the corresponding homogeneous problem

$$\begin{aligned}\frac{dv}{dt} &= A(t)v, \quad t \geq t_0, \\ v(t_0) &= v_0.\end{aligned}\tag{4.7.4}$$

Let $t \geq t_0$ be fixed. The problem (4.7.4) has a unique continuously differentiable solution $v(t)$ for every v_0 . This gives us a mapping $v_0 \mapsto v(t)$ of the initial vectors onto the solution vectors $v(t)$. Because problem (4.7.4) is a linear differential equation, it easily follows that this mapping is linear. Thus, there exists an operator $S(t, t_0)$, which we call the *solution operator*, such that

$$v(t) = S(t, t_0)v(t_0).\tag{4.7.5}$$

Lemma 4.7.2. *The solution operator has the following properties*

$$\begin{aligned}S(t_0, t_0) &= I, \quad \text{the identity,} \\ S(t_2, t_0) &= S(t_2, t_1)S(t_1, t_0), \quad t_0 \leq t_1 \leq t_2,\end{aligned}\tag{4.7.6}$$

and there are constants K and α such that

$$|S(t, t_0)| \leq Ke^{\alpha(t-t_0)}.\tag{4.7.7}$$

Proof. The equalities in Eqs. (4.7.6) follow immediately from the definition of S . If we consider the solution of Eqs. (4.7.4) at initial time t_0 , it is unchanged; that is, $S(t_0, t_0) = I$. If we consider solving Eqs. (4.7.4) at time t_2 , then we can alternatively think of beginning at t_0 and solving to t_2 , or of beginning at t_0 and solving for $v(t_1)$ and using this value to solve the interval from t_1 to t_2 . Because Eqs. (4.7.4) have a unique solution, we must obtain the same value, that is, $S(t_2, t_0) = S(t_2, t_1)S(t_1, t_0)$. Now consider

$$\begin{aligned}\frac{d}{dt} |v|^2 &= \left\langle \frac{dv}{dt}, v \right\rangle + \left\langle v, \frac{dv}{dt} \right\rangle, \\ &= \langle Av, v \rangle + \langle v, Av \rangle, \\ &= \langle v, (A + A^*)v \rangle, \\ &\leq 2\alpha|v|^2,\end{aligned}$$

where 2α is an upper bound for $|A + A^*|$, that is, $|(A + A^*)v| \leq 2\alpha|v|$ for all v . Therefore, by Lemma 4.7.1 ,

$$|v(t)|^2 \leq e^{2\alpha(t-t_0)}|v(t_0)|^2$$

and Eq. (4.7.7) follows.

In this case, we obtain $K = 1$ in Eq. (4.7.7), but α may be large. By allowing $K > 1$ and using a different technique, we can often get better estimates. As an example, consider

$$A = \begin{bmatrix} 0 & 100 \\ 1 & 0 \end{bmatrix},$$

in which case

$$A + A^* = \begin{bmatrix} 0 & 101 \\ 101 & 0 \end{bmatrix}, \quad \alpha = 50.5;$$

that is,

$$|S(t, t_0)| \leq e^{50.5(t-t_0)}. \quad (4.7.8)$$

With

$$T = \begin{bmatrix} 10 & -10 \\ 1 & 1 \end{bmatrix}, \quad T^{-1} = \frac{1}{20} \begin{bmatrix} 1 & 10 \\ -1 & 10 \end{bmatrix},$$

we have

$$\begin{aligned} T^{-1}AT &= \begin{bmatrix} 10 & 0 \\ 0 & -10 \end{bmatrix} =: \Lambda, \\ |T| \cdot |T^{-1}| &= 10. \end{aligned}$$

For $w = T^{-1}v$ we have

$$\frac{dw}{dt} = \Lambda w$$

and

$$\frac{d}{dt}|w|^2 = 2\langle w, \Lambda w \rangle \leq 20|w|^2.$$

Thus,

$$|v(t)| \leq |T| \cdot |w(t)| \leq |T|e^{10(t-t_0)}|w(t_0)| \leq |T| \cdot |T^{-1}|e^{10(t-t_0)}|v(t_0)|,$$

which yields

$$|S(t, t_0)| \leq 10e^{10(t-t_0)},$$

which is to be compared with Eq. (4.7.8).

Another important property of S is proved in the following lemma.

Lemma 4.7.3. *Let u_0 be a constant vector. Then $S(t_2, t_1)u_0$ is a C^1 function of both t_2 and t_1 . Also,*

$$\frac{\partial}{\partial t} S(t, t_0) = A(t)S(t, t_0), \quad (4.7.9)$$

$$\frac{\partial}{\partial t_0} S(t, t_0) = -S(t, t_0)A(t_0). \quad (4.7.10)$$

Proof. We recall that the solutions $v(t)$ of Eqs. (4.7.4) are C^1 . S is continuous in its first argument because

$$\lim_{\Delta t \rightarrow 0} |(S(t + \Delta t, t_0) - S(t, t_0))v(t_0)| = \lim_{\Delta t \rightarrow 0} |v(t + \Delta t) - v(t)| = 0.$$

Continuity in the second argument follows from Eqs. (4.7.7). We have

$$\begin{aligned} & \lim_{\Delta t_0 \rightarrow 0} |S(t, t_0 + \Delta t_0) - S(t, t_0)| \\ &= \lim_{\Delta t_0 \rightarrow 0} |S(t, t_0 + \Delta t_0)(I - S(t_0 + \Delta t_0, t_0))| \\ &\leq \lim_{\Delta t_0 \rightarrow 0} Ke^{\alpha(t - (t_0 + \Delta t_0))} |I - S(t_0 + \Delta t_0, t_0)| = 0. \end{aligned}$$

Let $\Delta t > 0$, then

$$\frac{S(t + \Delta t, t_0) - S(t, t_0)}{\Delta t} v(t_0) = \frac{v(t + \Delta t, t) - v(t)}{\Delta t},$$

so

$$\lim_{\Delta t \rightarrow 0} \frac{S(t + \Delta t, t_0) - S(t, t_0)}{\Delta t} v(t_0) = \frac{dv(t)}{dt} = A(t)v(t) = A(t)S(t, t_0)v(t_0)$$

exists, establishes Eq. (4.7.9), and shows that S is C^1 in its first argument. Now let $\Delta t_1 > 0$ and consider

$$\frac{S(t_2, t_1 + \Delta t_1) - S(t_2, t_1)}{\Delta t_1} v(t_1) = S(t_2, t_1 + \Delta t_1) \frac{I - S(t_1 + \Delta t_1, t_1)}{\Delta t_1} v(t_1).$$

An argument analogous to the preceding one yields

$$\lim_{\Delta t_1 \rightarrow 0} \frac{I - S(t_1 + \Delta t_1, t_1)}{\Delta t_1} v(t_1) = - \frac{dv(t_1)}{dt}$$

and, by continuity,

$$\lim_{\Delta t_1 \rightarrow 0} S(t_2, t_1 + \Delta t_1) = S(t_2, t_1).$$

It follows that

$$\begin{aligned} \lim_{\Delta t_1 \rightarrow 0} & \frac{S(t_2, t_1 + \Delta t_1) - S(t_2, t_1)}{\Delta t_1} v(t_1) \\ &= -S(t_2, t_1) \frac{dv(t_1)}{dt} = -S(t_2, t_1)A(t_1)v(t_1). \end{aligned}$$

Thus, S is also C^1 in its second argument, Eq. (4.7.10) holds, and the lemma is proved.

We can now prove the following theorem.

Theorem 4.7.1. Duhamel's Principle. *Let S be the solution operator of the homogeneous problem (4.7.4) introduced in Eq. (4.7.5). The solution of the inhomogeneous problem (4.7.3) can be expressed as*

$$u(t) = S(t, t_0)u_0 + \int_{t_0}^t S(t, \tau)F(\tau) d\tau, \quad (4.7.11)$$

and Eq. (4.7.7) yields

$$|u(t)| \leq K(e^{\alpha(t-t_0)}|u_0| + \varphi^*(\alpha, t - t_0) \max_{t_0 \leq \tau \leq t} |F(\tau)|), \quad (4.7.12)$$

where $\varphi^*(\alpha, t)$ is defined in Eq. (4.7.2).

Proof. $S(t, \tau)$ is a continuous function of τ and is continuously differentiable in t , so Eq. (4.7.11) is well defined, and $u(t) \in C^1(t)$. Using Lemma 4.7.3, we have

$$\begin{aligned}\frac{du}{dt} &= A(t)S(t, t_0)u_0 + \int_{t_0}^t A(t)S(t, \tau)F(\tau)d\tau + F(t) \\ &= A(t)u(t) + F(t),\end{aligned}$$

and

$$u(0) = S(0, 0)u_0 = u_0.$$

The inequality (4.7.12) then follows immediately from Eqs. (4.7.7) and (4.7.11), which concludes the proof.

We now generalize this result to partial differential equations. Consider the system (4.7.1) and suppose that the corresponding homogeneous problem

$$\begin{aligned}v_t &= P\left(x, t, \frac{\partial}{\partial x}\right)v, \quad t \geq t_0, \\ v(x, t_0) &= f(x)\end{aligned}\tag{4.7.13}$$

is well posed according to Definition 4.1.1. We can define a solution operator for Eqs. (4.7.13) in the same way as we did for the ordinary differential equation. Let $t \geq t_0$ be fixed. For every periodic initial function $f(x) \in C^\infty(x)$ there is a unique solution $v(x, t) \in C^\infty(x)$ so we obtain a mapping $f(x) \mapsto v(x, t)$. This mapping is linear because the differential equation is linear, and it is bounded because we assumed that Eq. (4.1.10) holds. We can, therefore, define a solution operator $S(t, t_0)$ such that

$$\begin{aligned}v(x, t) &= S(t, t_0)f(x), \\ \|S(t, t_0)\| &\leq Ke^{\alpha(t-t_0)}.\end{aligned}\tag{4.7.14}$$

We can use the same arguments that we used before for ordinary differential equations to show that the equalities (4.7.6) also hold for this solution operator. We now obtain

$$\begin{aligned}\frac{\partial}{\partial t} S(t, t_0)f(x) &= P\left(x, t, \frac{\partial}{\partial x}\right)S(t, t_0)f(x), \\ \frac{\partial}{\partial t_0} S(t, t_0)f(x) &= -S(t, t_0)P\left(x, t_0, \frac{\partial}{\partial x}\right)f(x),\end{aligned}\tag{4.7.15}$$

which corresponds to Lemma 4.7.3. By repeating the differentiation, we may conclude that $S(t, t_0)f(x) \in C^\infty(x, t, t_0)$.

As before, these results allow us to represent the solution of the inhomogeneous problem in terms of the solution operator for the homogeneous problem,

$$u(x, t) = S(t, t_0)f(x) + \int_{t_0}^t S(t, \tau)F(x, \tau) d\tau. \quad (4.7.16)$$

Therefore, $u(x, t) \in C^\infty(x, t)$ and we obtain the estimate

$$\begin{aligned} \|u(\cdot, t)\| &\leq \|S(t, t_0)f(\cdot)\| + (\max_{t_0 \leq \tau \leq t} \|F(\cdot, \tau)\|) \int_{t_0}^t \|S(t, \tau)\| d\tau \\ &\leq K(e^{\alpha(t-t_0)} \|f(\cdot)\| + \varphi^*(\alpha, t - t_0) \max_{t_0 \leq \tau \leq t} \|F(\cdot, \tau)\|), \end{aligned} \quad (4.7.17)$$

where φ^* is defined in Eq. (4.7.2).

In summary, we have proved this theorem.

Theorem 4.7.2. Duhamel's principle for PDE. *If the solutions of the homogeneous system (4.7.13) satisfy the conditions of Definition 4.1.1, that is, if it is a well-posed problem, then the inhomogeneous problem is also well-posed, its solution can be represented in terms of the solution operator for the homogeneous problem using Eq. (4.7.16), and its solution satisfies the estimate (4.7.17).*

EXERCISES

- 4.7.1. Carry out each step in proving that Eq. (4.7.15) implies $S(t, t_0)f(x) \in C^\infty(x, t, t_0)$.
- 4.7.2. Derive the explicit form of the solution operator for $u_t = au_x$. Use this form to write down the solution of $u_t = au_x + F(x, t)$.

4.8. GENERALIZED SOLUTIONS

Until now we have assumed that the data $f(x) \in C^\infty(x)$ and $F(x, t) \in C^\infty(x, t)$. In this case, we say that there is a classical solution. However, we can relax this condition by introducing generalized solutions.

We again begin with the homogeneous problem (4.1.8) and assume that it is well posed. Let $g(x)$ be a 2π -periodic function that belongs only to L_2 . The function $g(x)$ may not be smooth, but it can be approximated by a sequence of smooth functions $f_\nu(x) \in C^\infty(x)$ such that

$$\lim_{\nu \rightarrow \infty} \|f_\nu(\cdot) - g(\cdot)\| = 0.$$

We can then solve Eq. (4.1.8) for the initial functions $f_\nu(x)$ and obtain a sequence of solutions

$$v_\nu(x, t) = S(t, t_0)f_\nu(x).$$

This sequence converges to a limit function $v(x, t)$, because

$$\|v_\nu(\cdot, t) - v_\mu(\cdot, t)\| = \|S(t, t_0)(f_\nu(\cdot) - f_\mu(\cdot))\| \leq Ke^{\alpha(t-t_0)}\|f_\nu(\cdot) - f_\mu(\cdot)\|.$$

Also, $v(x, t)$ is independent of the sequence $\{f_\nu\}$, that is, if $\{\tilde{f}_\nu\}$ is another sequence such that $\lim \tilde{f}_\nu = g$ and $\tilde{v}_\nu(x, t)$ are the corresponding solutions of Eq. (4.1.8), then

$$\begin{aligned} \|\tilde{v}_\nu(\cdot, t) - v_\nu(\cdot, t)\| &= \|S(t, t_0)(\tilde{f}_\nu(\cdot) - f_\nu(\cdot))\| \\ &\leq Ke^{\alpha(t-t_0)}(\|\tilde{f}_\nu(\cdot) - g(\cdot)\| + \|f_\nu(\cdot) - g(\cdot)\|), \end{aligned}$$

and we see that $\lim_{\nu \rightarrow \infty} \tilde{v}_\nu = \lim_{\nu \rightarrow \infty} v_\nu$. We call $v(x, t)$ the *generalized solution* of Eq. (4.1.8). This process is well known in functional analysis. $S(t, t_0)$ is a bounded linear operator in L_2 , which is densely defined. Therefore, it can be uniquely extended to all of L_2 . We have just described this process.

We look at an example of this process now. Consider the differential equation

$$v_t + v_x = 0, \quad 0 \leq t, \tag{4.8.1}$$

with piecewise constant periodic initial data

$$g(x) = \begin{cases} 0, & \text{for } 0 \leq x \leq \frac{2}{3}\pi, \\ 1, & \text{for } \frac{2}{3}\pi < x < \frac{4}{3}\pi, \\ 0, & \text{for } \frac{4}{3}\pi \leq x \leq 2\pi. \end{cases} \tag{4.8.2}$$

We want to show that its generalized solution is

$$v(x, t) = g(x - t),$$

that is, it is a square wave, which travels with speed 1 to the right.

We approximate $g(x)$ by $f_\nu(x) \in C^\infty(x)$, which are slightly rounded at the corners and converge to $g(x)$ (see Figure 4.8.1). Then

$$v_\nu(x, t) = f_\nu(x - t),$$

and it is clear that

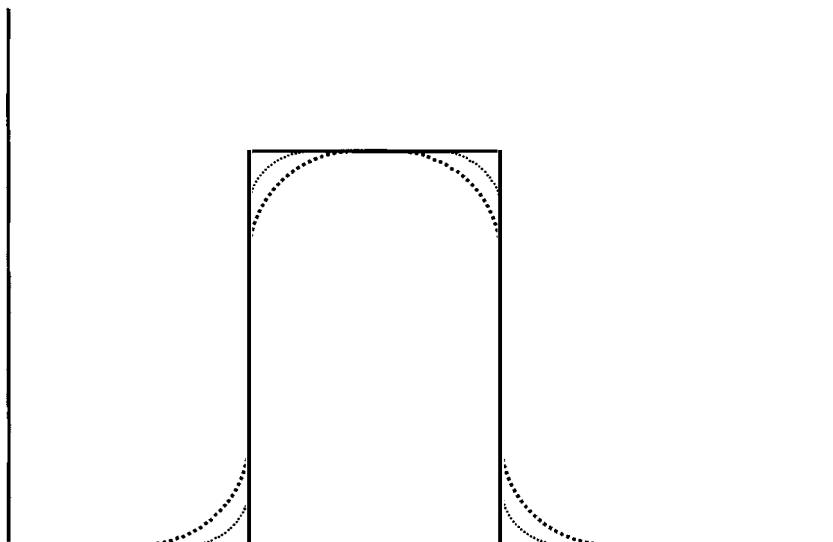


Figure 4.8.1.

$$v(x, t) := \lim_{\nu \rightarrow \infty} v_\nu(x, t) = g(x - t).$$

Using Duhamel's principle, the same process can be applied to the inhomogeneous problem. The only assumption we need to make regarding the forcing function $G(x, t)$ is that we can find a sequence $\{F_\nu(x, t)\}$, where for each ν , $F_\nu(x, t) \in C^\infty(x, t)$ such that

$$\lim_{\nu \rightarrow \infty} \sup_{0 \leq t \leq T} \|F_\nu(x, t) - G(x, t)\| = 0$$

in every finite time interval $0 \leq t \leq T$.

This extension process is very powerful, because we only need to prove existence theorems for smooth solutions. Then one applies the closure process above to extend the results.

EXERCISES

- 4.8.1.** Construct explicit functions $f_\nu(x)$ such that $\lim_{\nu \rightarrow \infty} f_\nu(x) = g(x)$, where $g(x)$ is defined in Eq. (4.8.2).
- 4.8.2.** Carry out each step in the definition of the generalized solution for

$$\begin{aligned} u_t &= P\left(\frac{\partial}{\partial x}\right) u + F, \\ u(x, 0) &= f(x), \end{aligned}$$

where F and f are not smooth functions.

4.9. WELL-POSEDNESS OF NONLINEAR PROBLEMS

We discuss nonlinear problems of the form

$$\begin{aligned} u_t &= P\left(x, t, u, \frac{\partial}{\partial x}\right) u + F \\ &:= \sum_{i,j=1}^d \frac{\partial}{\partial x^{(j)}} \left(A_{ij}(x, t, u) \frac{\partial}{\partial x^{(i)}} u \right) + \sum_{i=1}^d B_i(x, t, u) \frac{\partial}{\partial x^{(i)}} u \\ &\quad + C(x, t, u)u + F(x, t), \\ u(x, 0) &= f(x); \end{aligned} \tag{4.9.1}$$

that is, the coefficients are functions of u but not of the derivatives of u . This is no real restriction, however. Consider, for example,

$$u_t = u_x^2 u_{xx}. \tag{4.9.2}$$

Let $v = u_x$, $w = u_{xx}$. Differentiating Eq. (4.9.2) gives us

$$\begin{aligned} \begin{bmatrix} u \\ v \\ w \end{bmatrix}_t &= \begin{bmatrix} v^2 & 0 & 0 \\ 0 & v^2 & 0 \\ 0 & 0 & v^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}_{xx} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 6vw \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}_x \\ &\quad + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2w^2 & 0 \\ 0 & 0 & 2w^2 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \end{aligned}$$

which is of the form found in Eqs. (4.9.1). In fact, any second-order nonlinear system with smooth coefficients can be transformed into a system of the form found in Eqs. (4.9.1). We also assume that the coefficients and data are real and that we are only interested in real solutions.

The following definition corresponds to Section 4.3.

Definition 4.9.1. *The system (4.9.1) is called symmetric hyperbolic if all $A_{ij} = 0$ and if, for all x, t , and u , the matrices $B_i(x, t, u)$, $i = 1, 2, \dots, d$, are Hermitian.*

If the $A_{ij} = A_{ij}(x, t)$ do not depend on u , then we can define strongly parabolic systems in the same way as we did in Section 4.6; that is, the inequality (4.6.6) holds. If the A_{ij} depend on u , then it is generally impossible to define parabolicity without reference to a particular solution. Consider, for example,

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

If $f(x) = -1 + \varepsilon g$, $0 < \varepsilon \ll 1$, then the solutions will behave badly because we are close to the backward heat equation. On the other hand, if $f = 1 + \varepsilon g$, then the solution will stay close to 1, and the equation is strongly parabolic at the solution.

Definition 4.9.2. Let $u(x, t)$ be a solution of Eqs. (4.9.1). We call the system *strongly parabolic at the solution* if the inequality (4.6.6) holds. Here $A_{ij} = A_{ij}(x, t, u)$.

We can no longer expect that solutions will exist for all time. This difficulty occurs already for ordinary differential equations. The solution of

$$u_t = u^2, \quad u(0) = u_0,$$

is given by

$$u(t) = \frac{u_0}{1 - u_0 t}.$$

For $u_0 > 0$, we have $\lim_{t \rightarrow 1/u_0} u(t) = \infty$. If $u_0 < 0$, then the solution exists for all time and converges to zero as $t \rightarrow \infty$. In Chapter 8, we consider

$$u_t + uu_x = 0,$$

and show that its solutions stay bounded but that u_x may become arbitrarily large when t approaches some time t_0 .

In general, only local existence results are known; that is, for given smooth initial data $f(x)$, there exists a finite time interval $0 \leq t \leq T$ such that Eqs. (4.9.1) have a smooth solution. Here T depends on f . These existence results can be obtained by the iteration

$$\begin{aligned} u_t^{(n+1)} &= P\left(x, t, u^{(n)}(x, t), \frac{\partial}{\partial x}\right) u^{(n+1)} + F, \quad n = 0, 1, \dots, \\ u^{(n+1)}(x, 0) &= f(x), \\ u^{(0)}(x, t) &= f(x). \end{aligned} \tag{4.9.3}$$

Thus, one solves a sequence of linear problems and proves that the sequence of solutions $u^{(n)}(x, t)$ converges to a limit function $u(x, t)$ which is the solution of the nonlinear problem (4.9.1). Bounds of $u^{(n)}$ and its derivatives that do not depend on n are crucial in this process. Therefore, the linear problems must be well posed.

Theorem 4.9.1. *If all the linear problems above are symmetric hyperbolic or strongly parabolic or mixed symmetric hyperbolic–strongly parabolic, then $u^{(n)}$ and its derivatives can be bounded independently of n in some time interval $0 \leq t \leq T$ and the nonlinear problem has a unique smooth solution.*

For symmetric hyperbolic systems, the last result is satisfactory. We do not need to know $u^{(n)}(x, t)$ to be able to decide whether the system is symmetric hyperbolic. This follows from the algebraic structure of the equations. For strongly parabolic systems the same is true if the second-order terms are linear, that is, if the coefficients A_{ij} do not depend on u . If the A_{ij} depend on u , then we need to know $u^{(n)}(x, t)$ to decide whether the systems (4.9.3) satisfy the required condition. However, the next theorem tells us that we need only know that the condition is satisfied at $t = 0$.

Theorem 4.9.2. *Assume that the system*

$$w_t = P\left(x, t, f(x), \frac{\partial}{\partial x}\right) w$$

is strongly parabolic. Then there is a time interval $0 \leq t \leq T, T > 0$, such that the systems (4.9.3) have the same property, and the nonlinear system (4.9.1) is strongly parabolic at the solution.

One often solves partial differential equations numerically, although no analytic existence results are known. In this case, one would like to use the numerical results to infer that the analytic solution exists and that the numerical solution is close to that solution. We can proceed in the following way. We interpolate the numerical solution w and show that the interpolant \tilde{u} satisfies a perturbed differential equation

$$\begin{aligned}\tilde{u}_t &= P\left(x, t, \tilde{u}, \frac{\partial}{\partial x}\right) \tilde{u} + \tilde{F}, \\ \tilde{u}(x, 0) &= \tilde{f}.\end{aligned}\quad (4.9.4)$$

Estimates for $F - \tilde{F}$ and $f - \tilde{f}$ can be obtained from the numerical solution and, hopefully, they are small. This depends on the accuracy of the method and how smooth the numerical solution is, that is, on bounds of w and its divided differences. Thus, by numerical calculation we have established the existence of a solution of Eq. (4.9.4). From this, we want to infer the existence of a solution of Eq. (4.9.1) and show that $\sup_t \|u - \tilde{u}\|_N$ is small when $\sup_t \|F - \tilde{F}\|_N$ and $\sup_t \|f - \tilde{f}\|_N$ are small. Here $\|\cdot\|_N$ denotes a suitable norm, which, in general, depends on derivatives; that is,

$$\|f - \tilde{f}\|_N^2 = \sum_{|j| \leq p} \|D^{|j|}(f - \tilde{f})\|^2.$$

Typically $p = [\frac{d}{2}d] + 1$, where d is the number of space dimensions. We now can make the following definition.

Definition 4.9.3. Let \tilde{u} be a smooth solution of Eqs. (4.9.1). We call the nonlinear initial value problem well posed for a time interval $0 \leq t \leq T$ at the solution \tilde{u} , if

1. there is a neighborhood \mathcal{N} of \tilde{F}, \tilde{f} defined by a suitable norm $\|\cdot\|_N$:

$$\sup_t \|\tilde{F} - F\|_N < \eta, \quad \|\tilde{f} - f\|_N < \eta,$$

where $\eta > 0$ is a sufficiently small constant, such that Eqs. (4.9.1) have a smooth solution for all $(F, f) \in \mathcal{N}$, and

2. there is a constant K such that

$$\sup_{0 \leq t \leq T} \|\tilde{u} - u\|_N \leq K\eta.$$

One can therefore prove the following.

Theorem 4.9.3. If the system (4.9.4) is symmetric hyperbolic or strongly parabolic at \tilde{u} , or mixed symmetric hyperbolic-strongly parabolic, then it is well posed at \tilde{u} .

The error $w = u - \tilde{u}$ satisfies, to first approximation, a linear system

$$\begin{aligned} w_t &= P_1 \left(x, t, \tilde{u}, \frac{\partial}{\partial x} \right) w + F - \tilde{F}, \\ w(x, 0) &= f - \tilde{f}, \end{aligned} \quad (4.9.5)$$

which we obtain by linearizing Eqs. (4.9.1) at \tilde{u} . Thus, the bounds on w depend on the growth behavior of the solution operator of Eqs. (4.9.5).

EXERCISES

- 4.9.1.** Consider the nonlinear Euler equations (4.6.4). Prove that the corresponding linearized system is not hyperbolic with the equation of state $p = G(\rho)$ if the initial data are such that

$$\frac{dG}{d\rho} < 0, \quad t = 0.$$

BIBLIOGRAPHIC NOTES

Well-posedness of the Cauchy problem for linear partial differential equations was defined by Hadamard (1921). He used a weaker form than the one used here. Petrovskii (1937) gave a general analysis of PDEs with constant coefficients that are well posed in Hadamard's sense. Well posedness in our sense is discussed in Kreiss (1964) and in Kreiss and Lorenz (1989).

We have concentrated upon the estimates of the solutions in terms of the initial data. In the linear case this immediately leads to uniqueness. The existence of a solution is of course the most fundamental part of well-posedness, and we have indicated how this can be assured by using difference approximations for which the existence of solutions is trivial. In fact, it is often sufficient to discretize the differential equation in space. A system of ODEs of the type (4.7.4) is then obtained, and the classical theory for ODEs can be used to guarantee existence of a solution; see, for example, Coddington and Levinson (1955.)

The goal of our discussion has been to present the underlying theory for time-dependent PDEs, which is important for their numerical solution. A much more comprehensive treatment can be found in Kreiss and Lorenz (1989).

We have formulated general PDEs as $u_t = P(\partial/\partial x)u$. Many problems give rise to equations with higher order time derivatives. These equations can always be rewritten in first-order form by introducing $v = u_t, w = u_{tt}, \dots$ as new dependent variables and then be treated by the theory above. The theory can also be developed for the original higher order form. An example of this is in the paper by Friedrichs (1954), which deals with symmetric hyperbolic PDEs of second order.

5

STABILITY AND CONVERGENCE FOR NUMERICAL APPROXIMATIONS OF LINEAR AND NONLINEAR PROBLEMS

In this chapter, we define the basic concepts used for the analysis of approximations of general initial value problems. We focus upon difference methods, but the theory developed can be applied to other methods as well.

5.1. STABILITY AND CONVERGENCE

Consider the **periodic** initial value problem for a general linear system of partial differential equations

$$\boxed{u_t = P\left(x, t, \frac{\partial}{\partial x}\right) u, \\ u(x, 0) = f(x).} \quad (5.1.1)$$

Here x and u are vectors with d and m components, respectively. We assume that f and the coefficients are 2π -periodic, that is, 2π -periodic in all space directions $x^{(j)}$, and we are interested in 2π -periodic solutions.

As discussed earlier, we introduce gridpoints by

$$x_j = (j_1 h, \dots, j_d h), \quad j_\nu = 0, \pm 1, \pm 2, \dots, \\ h = 2\pi/(N + 1), \quad N \text{ a natural number},$$

and a time step $k > 0$. For convenience, we always assume that $k = \lambda h^p$, where λ is a constant greater than 0 and p is the order of the differential operator in space.

A general difference approximation is of the form

$$\boxed{Q_{-1} v^{n+1} = \sum_{\sigma=0}^q Q_\sigma v^{n-\sigma}, \quad n = q, \quad q + 1, \dots, \\ v^\sigma = f^{(\sigma)}, \quad \sigma = 0, 1, \dots, q.} \quad (5.1.2)$$

For simplicity, we write the gridfunction v^n without a subscript. It is to be understood that Eqs. (5.1.2) hold in every gridpoint. The Q_σ are difference operators with 2π -periodic matrix coefficients, which may depend on x, t, k, h . The initial data are given 2π -periodic functions, and we are interested in 2π -periodic solutions. We assume that the operator Q_{-1} is uniformly bounded and has a uniformly bounded inverse Q_{-1}^{-1} as $h, k \rightarrow 0$, so that we can advance the solution step by step in time.

For theoretical purposes, it is convenient to rewrite Eqs. (5.1.2) as a one-step scheme using the companion matrix. By introducing the vectors

$$\boxed{\mathbf{v}^n = (v^{n+q}, v^{n+q-1}, \dots, v^n)^T, \quad n = 0, 1, \dots, \\ \mathbf{f} = (f^{(q)}, f^{(q-1)}, \dots, f^{(0)})^T,} \quad (5.1.3)$$

Eqs. (5.1.2) take the form

$$\boxed{\mathbf{v}^{n+1} = Q(t_n) \mathbf{v}^n, \\ \mathbf{v}^0 = \mathbf{f},} \quad (5.1.4)$$

where

$$\boxed{Q(t_n) = Q(x_j, t_n) = \begin{bmatrix} Q_{-1}^{-1} Q_0 & Q_{-1}^{-1} Q_1 & \cdots & Q_{-1}^{-1} Q_q \\ I & & & \\ & I & & \\ & & I & 0 \end{bmatrix}} \quad (5.1.5)$$

is the companion matrix. For example, the leap-frog scheme in Eq. (2.2.1) can be written as

$$\boxed{\mathbf{v}^{n+1} = \begin{bmatrix} 2kD_0 & I \\ I & 0 \end{bmatrix} \mathbf{v}^n.} \quad (5.1.6)$$

Because it corresponds to the solution operator $S(t, t_0)$ for differential equations, we define the **discrete solution operator** S_h by

$$\boxed{\mathbf{v}^n = S_h(t_n, t_\nu) \mathbf{v}^\nu.} \quad (5.1.7)$$

The arguments k, h will be omitted.

S_h can be expressed explicitly by

$$\boxed{S_h(t_n, t_\nu) = \prod_{\mu=1}^{n-\nu} Q(t_{n-\mu}), \quad S_h(t_n, t_n) = I.}$$

If Q is independent of t , then we have

$$S_h(t_n, t_\nu) = Q^{n-\nu}. \quad (5.1.8)$$

As a norm we use

$$\boxed{\|\mathbf{v}^n\|_h = \left(\sum_{\sigma=0}^q \|v^{n+\sigma}\|_h^2 \right)^{1/2},} \quad (5.1.9)$$

and denote by $\|S_h\|$ the corresponding operator norm.

The concept of stability is the discrete analogue of well-posedness. The existence of solutions is easy to verify. We must obtain the needed estimates.

Definition 5.1.1. *The difference approximation (5.1.4) is called **stable** for $0 < h \leq h_0$, if there are constants α_S, C, K_S such that for all h*

$$\boxed{\|Q_{-1}^{-1}\|_h \leq C, \quad \|S_h(t_n, t_\nu)\|_h \leq K_S e^{\alpha_S(t_n - t_\nu)}} \quad (5.1.10)$$

This stability definition requires that the estimate

$$\boxed{\|\mathbf{v}^n\|_h \leq K(t_n) \|\mathbf{f}\|_h, \quad K(t_n) = K_S e^{\alpha_S t_n}} \quad (5.1.11)$$

hold for *all* initial functions \mathbf{f} . The definition includes an exponential factor, because we must be able to treat approximations to differential equations with exponentially growing solutions such as, for example, $u_t = u_x + u$. On any finite

time interval $[0, T]$, $K(t_n)$ is bounded by $K(T)$, which is bounded independently of h , k , \mathbf{f} . Stability guarantees that the solutions are bounded as $h \rightarrow 0$. Note the difference between the growth factors $e^{\alpha t_n}$ and $e^{\alpha n}$. The first is accepted, the second is not. For a given time step k_0 , one can always write

$$e^{\alpha n} = e^{\beta t_n}, \quad \beta = \alpha/k_0.$$

However, when the mesh is refined to $k = k_0/2$, it takes $2n$ steps to reach the same time value, and the growth is worse. This is the worst type of instability normally encountered. Another typical form of instability is

$$\|v^n\|_h \sim C/h^q, \quad q > 0, \quad (5.1.12)$$

which is a weaker instability, but still prohibited by our definition.

Next consider the case where the difference approximation includes a forcing function

$$Q_{-1}v^{n+1} = \sum_{\sigma=0}^q Q_\sigma v^{n-\sigma} + k\mathbf{F}^n, \quad n = q, q+1, \dots,$$

$$v^\sigma = f^{(\sigma)}, \quad \sigma = 0, 1, \dots, q.$$

(5.1.13)

We again write Eq. (5.1.13) as a one-step method

$$\mathbf{v}^{n+1} = Q(t_n)\mathbf{v}^n + k\mathbf{F}^n, \quad \mathbf{F} = (Q_{-1}^{-1}\mathbf{F}^n, 0, \dots, 0)^T,$$

$$\mathbf{v}^0 = \mathbf{f}.$$

(5.1.14)

The discrete version of Duhamel's principle is given in Lemma 5.1.1.

Lemma 5.1.1. *The solution of Eqs. (5.1.14) can be written in the form*

$$\mathbf{v}^n = S_h(t_n, 0)\mathbf{f} + k \sum_{\nu=0}^{n-1} S_h(t_n, t_{\nu+1})\mathbf{F}^\nu.$$

(5.1.15)

Proof. Let \mathbf{v} be defined by Eq. (5.1.15). Then

$$\mathbf{v}^{n+1} = S_h(t_{n+1}, 0)\mathbf{v}^0 + k \sum_{\nu=0}^n S_h(t_{n+1}, t_{\nu+1})\mathbf{F}^\nu.$$

Observing that

$$S_h(t_{n+1}, t_{n+1}) = I, \quad S_h(t_{n+1}, t_\mu) = Q(t_n)S_h(t_n, t_\mu), \quad \mu < n + 1,$$

we obtain

$$\begin{aligned} \mathbf{v}^{n+1} &= Q(t_n) \left(S_h(t_n, 0) \mathbf{v}^0 + k \sum_{\nu=0}^{n-1} S_h(t_n, t_{\nu+1}) \mathbf{F}^\nu \right) + k \mathbf{F}^n \\ &= Q(t_n) \mathbf{v}^n + k \mathbf{F}^n. \end{aligned}$$

Because $\mathbf{v}^0 = \mathbf{f}$, \mathbf{v}^n is the solution of Eqs. (5.1.14). This proves the lemma.

We can now estimate the solution of Eqs. (5.1.14) in terms of \mathbf{f} and \mathbf{F} .

Theorem 5.1.1. *Assume, that the difference approximation is stable. Then the solution of Eq. (5.1.14) satisfies the estimate*

$$\|\mathbf{v}^n\|_h \leq K_S \left(e^{\alpha_S t_n} \|\mathbf{f}\|_h + \varphi_h^*(\alpha_S, t_n) \max_{0 \leq \nu \leq n-1} \|\mathbf{F}^\nu\|_h \right),$$

where

$$\varphi_h^*(\alpha_S, t_n) = \sum_{\nu=0}^{n-1} e^{\alpha_S (t_n - t_{\nu+1}) k} \sim \int_0^t e^{\alpha_S (t-\xi)} d\xi = \varphi^*(\alpha_S, t_n).$$

Proof. Equation (5.1.15) gives us

$$\|\mathbf{v}^n\|_h \leq \|S_h(t_n, 0)\|_h \|\mathbf{f}\|_h + \max_{0 \leq \nu \leq n-1} \|\mathbf{F}^\nu\|_h \sum_{\nu=0}^{n-1} \|S_h(t_n, t_{\nu+1})\|_h k$$

and the lemma follows from Eq. (5.1.10).

This theorem shows that it is sufficient to consider homogeneous approximations. The proper estimates for inhomogeneous equations follow from stability. Note, that the factor k multiplying \mathbf{F}^ν is lost in the estimate. There is a step-by-step accumulation of \mathbf{F}^ν values; so the amplification is of order $n \sim k^{-1}$.

We can now discuss the influence of rounding errors, which are committed in each step. They can be interpreted in terms of a slight perturbation of the

difference equation; we are actually computing the solution of

$$Q_{-1}\tilde{v}^{n+1} = \sum_{\sigma=0}^q Q_\sigma \tilde{v}^{n-\sigma} + \varepsilon^n, \quad \|\varepsilon^n\|_h \leq \varepsilon. \quad (5.1.16)$$

instead of Eqs. (5.1.2). ε is the order of machine precision. Subtracting Eqs. (5.1.2) from Eq. (5.1.16) we get, for $w^n = \tilde{v}^n - v^n$,

$$Q_{-1}w^{n+1} = \sum_{\sigma=0}^q Q_\sigma w^{n-\sigma} + \varepsilon^n. \quad (5.1.17)$$

Therefore, by Theorem 5.1.1,

$$\|w^n\|_h \leq C(t_n) \frac{\varepsilon}{k}. \quad (5.1.18)$$

If the error due to truncation is of the order δ , then we require $\varepsilon/k \ll \delta$, otherwise there is danger that the rounding error will be dominant. For modern computers with 64-bit words, rounding errors are not often a problem.

We will next study the effect of perturbing each operator Q_σ by an operator of order $\mathcal{O}(k)$. It is essential to understand the effect of these perturbations. For example, if $v^{n+1} = Q_0 v^n$ approximates $u_t = u_x$, then $v^{n+1} = (Q_0 + kI)v^n$ approximates $u_t = u_x + u$. It is convenient to know that such perturbations of order k do not cause instability, because we can often simplify the analysis by neglecting terms of order k . Let $\{R_\sigma\}$ be operators with

$$\|R_\sigma\|_h \leq K_1, \quad \sigma = -1, 0, \dots, q. \quad (5.1.19)$$

Instead of Eqs. (5.1.2), we consider

$$(Q_{-1} + kR_{-1})v^{n+1} = \sum_{\sigma=0}^q (Q_\sigma + kR_\sigma)v^n. \quad (5.1.20)$$

We now prove that Eq. (5.1.20) is stable if the unperturbed approximation (5.1.2) is stable. For the proof, we need the following lemma.

Lemma 5.1.2. *Let Q be a bounded operator. For $\|kQ\|_h < 1$*

$$\|(I + kQ)^{-1}\|_h \leq \frac{1}{1 - \|kQ\|_h}$$

therefore $(I + kQ)^{-1} = I - k(I + kQ)^{-1}Q := I + kQ_1$, where

$$\|Q_1\|_h \leq \frac{\|Q\|_h}{1 - \|kQ\|_h}.$$

Proof. We have

$$\|(I + kQ)^{-1}\|_h = \sup_{\|v\|_h=1} \|(I + kQ^{-1})v\|_h.$$

Let

$$w = (I + kQ)^{-1}v, \quad \|v\|_h = 1,$$

then

$$1 = \|v\|_h = \|(I + kQ)w\|_h \geq \|w\|_h - k\|Qw\|_h \geq (1 - k\|Q\|_h)\|w\|_h.$$

Therefore the desired estimate and the representation follows.

Theorem 5.1.2. Assume that the approximation (5.1.2) is stable and that Eq. (5.1.19) holds. Then the perturbed approximation (5.1.20) is stable.

Proof. Using the last lemma, we have

$$(Q_{-1} + kR_{-1})^{-1} = Q_{-1}^{-1}(I + kR_{-1}Q_{-1}^{-1})^{-1}$$

Therefore, using the last lemma, we can write Eq. (5.1.20) in the form

$$\tilde{\mathbf{v}}^{n+1} = (Q + kR)\tilde{\mathbf{v}}^n, \quad (5.1.21)$$

where R is uniformly bounded. Let $\mathbf{w}^n = e^{-\beta t_n} \tilde{\mathbf{v}}^n$, $\beta > 0$. Then Eq. (5.1.21) becomes

$$\mathbf{w}^{n+1} = e^{-\beta k} Q \mathbf{w}^n + k e^{-\beta k} \tilde{\mathbf{F}}^n, \quad (5.1.22)$$

where

$$\tilde{\mathbf{F}}^n = R \mathbf{w}^n.$$

Duhamel's principle can be applied to Eq. (5.1.22). The solution operator $\tilde{S}_h(t_n, t_\nu)$, corresponding to $e^{-\beta k}Q$, is clearly $e^{-\beta k(n-\nu)} S_h(t_n, t_\nu)$, where S_h is the solution operator for Eq. (5.1.4). Thus,

$$\|e^{-\beta k(n-\nu)} S_h(t_n, t_\nu)\|_h \leq K_S e^{(\alpha_S - \beta)(t_n - t_\nu)}.$$

Now consider Eq. (5.1.22) for $0 \leq \nu \leq n$, and let

$$\|\mathbf{w}^\mu\|_h = \max_{0 \leq \nu \leq n} \|\mathbf{w}^\nu\|_h.$$

By Theorem 5.1.1,

$$\|\mathbf{w}^\mu\|_h \leq K_S(e^{(\alpha_S - \beta)t_\mu} \|\mathbf{w}^0\|_h + \text{constant } \varphi_h^*(\alpha_S - \beta, t_\mu) \|\mathbf{w}^\mu\|_h).$$

The function $\varphi^*(\alpha, t)$, defined in Theorem 5.1.1, decreases as α decreases. Hence, by choosing β large enough, the factor multiplying $\|\mathbf{w}^\mu\|_h$ can be made less than $\frac{1}{2}$. Therefore,

$$\|\mathbf{w}^\mu\|_h \leq \|\mathbf{w}^\mu\|_h \leq 2K_S e^{(\alpha_S - \beta)t_\mu} \|\mathbf{w}^0\|_h, \quad \text{for } \beta \text{ sufficiently large;}$$

that is,

$$\|\tilde{\mathbf{v}}^n\|_h \leq 2K_S e^{\beta t_n + (\alpha_S - \beta)t_\mu} \|\tilde{\mathbf{v}}^0\|_h \leq 2K_S e^{\beta t_n} \|\tilde{\mathbf{v}}^0\|_h.$$

This proves the theorem.

This theorem shows that we can neglect terms of order k in the stability analysis. However, as we have seen in Section 2.2 when discussing the leap-frog scheme, terms of order k can play an important role. For practical purposes, a more refined stability definition is useful.

Definition 5.1.2. Assume that the solution operator $S(t, t_0)$ of the continuous problem satisfies the estimate

$$\|S(t, t_0)\| \leq K e^{\alpha(t-t_0)}.$$

We call the difference approximation strictly stable for $0 < h \leq h_0$, if

$$\|S_h(t, t_0)\|_h \leq K_S e^{\alpha_S(t-t_0)}, \quad \alpha_S \leq \alpha + \mathcal{O}(h).$$

As an example we consider

$$\begin{aligned} u_t &= u_x - u, \\ u(x, 0) &= f(x). \end{aligned}$$

From Section 2.2,

$$\|S(t, t_0)\| = e^{-(t-t_0)}.$$

However, the leap-frog approximation

$$v^{n+1} = v^{n-1} + 2k(D_0 v^n - v^n)$$

generates spurious solutions such that

$$\|S_h(t, t_0)\|_h \approx e^{t-t_0}.$$

The approximation is stable but not strictly stable. The modified scheme

$$(I + k)v^{n+1} = 2kD_0 v^n + (I - k)v^{n-1}$$

is strictly stable.

Next we will define the *order of accuracy* which is a measure of how well the difference scheme (5.1.2) approximates the differential equation (5.1.1).

Definition 5.1.3. Let $u(x, t)$ be a smooth solution of Eq. (5.1.1). Then the local truncation error is defined by

$$kr_j^n = Q_{-1}u(x_j, t_{n+1}) - \sum_{\sigma=0}^q Q_\sigma u(x_j, t_{n-\sigma}). \quad (5.1.23)$$

The difference approximation (5.1.2) is accurate of order (p_1, p_2) if, for all sufficiently smooth solutions $u(x, t)$, there is a function $L(t_n)$ such that, for $h \leq h_0$,

$$\|\tau^n\|_h \leq L(t_n) \cdot (h^{p_1} + k^{p_2}), \quad (5.1.24)$$

where $L(t_n)$ is bounded on every finite time interval. If $p_1 > 0, p_2 > 0$, then Eq. (5.1.2) is called consistent.

Since we have assumed a relationship between k and h , the right-hand side of Eq. (5.1.24) is only a function of h . If $k = \lambda h^\rho$, then we say that Eq. (5.1.2) has order of accuracy $\min(p_1, pp_2)$.

Several calculations carried out in Chapter 2 illustrate how one determines the order of accuracy. One always uses Taylor series expansions. For example, for the leap-frog scheme applied to $u_t = u_x$, we obtain

$$\begin{aligned} u(x_j, t_{n+1}) - 2kD_0u(x_j, t_n) - u(x_j, t_{n-1}) &= k\tau_j^n, \\ \tau_j^n &= -\frac{h^2}{3} u_{xxx}(x_j, t_n) + \frac{k^2}{3} u_{ttt}(x_j, t_n) + \mathcal{O}(h^4, k^4), \end{aligned} \quad (5.1.25)$$

that is, the approximation is accurate of order (2,2).

The DuFort–Frankel approximation (2.5.12) for $u_t = u_{xx}$ has its truncation error (divided by 2) defined in Eq. (2.5.15). For consistency, the condition (2.5.16), $k = ch^{1+\delta}$, $\delta > 0$, is required. Then the truncation error has the form

$$\frac{\tau}{2} = c^2 h^{2\delta} u_{tt} - \frac{h^2}{12} u_{xxxx} + \mathcal{O}(h^{2+4\delta}). \quad (5.1.26)$$

If $\delta > 0$, the scheme is consistent, and the order of accuracy is $\min(2\delta, 2)$. (According to our standard assumption, we would choose $\delta = 1$, since it is a parabolic equation.)

For some approximations, it is more convenient to carry out the Taylor expansions about a point that is not in the grid. The Crank–Nicholson scheme (2.5.19), approximating $u_t = u_{xx}$, has its center point at

$$(x_*, t_*) = (x_j, (t_{n+1} + t_n)/2).$$

Any differentiable function $\varphi(t)$ satisfies

$$\begin{aligned} \frac{\varphi(t_{n+1}) + \varphi(t_n)}{2} &= \varphi(t_*) + \frac{k^2}{4} \varphi_{tt}(t_*) + \mathcal{O}(k^4), \\ \frac{\varphi(t_{n+1}) - \varphi(t_n)}{k} &= \varphi_t(t_*) + \frac{k^2}{24} \varphi_{ttt}(t_*) + \mathcal{O}(k^4). \end{aligned}$$

Using Eqs. (2.4.1) to (2.4.6), we get

$$\begin{aligned} \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{k} &- \frac{1}{2} D_+ D_- [u(x_j, t_{n+1}) + u(x_j, t_n)] \\ &= u_t(x_*, t_*) - u_{xx}(x_*, t_*) + \frac{k^2}{24} u_{ttt}(x_*, t_*) - \frac{k^2}{4} u_{xxtt}(x_*, t_*) \\ &\quad - \frac{h^2}{12} u_{xxxx}(x_*, t_*) + \mathcal{O}(h^4 + k^4). \end{aligned}$$

Therefore, the order of accuracy is (2,2).

We also need that the initial data of the difference approximation are consistent with the solution of the differential equation. For one-step schemes, the correct data, $f^{(0)}$, can be used; but, for multistep schemes, the first q steps must

be computed using some other method. One-step schemes are usually used to compute these values.

Definition 5.1.4. *The initial data have the order of accuracy (p_1, p_2) if, for all sufficiently smooth solutions $u(x, t)$ of Eq. (5.1.1),*

$$\|v^\sigma - u(\cdot, \sigma k)\|_h \leq L_I(h^{p_1} + k^{p_2}), \quad \sigma = 0, 1, \dots, q. \quad (5.1.27)$$

The initial data are consistent if $p_1 > 0, p_2 > 0$.

As an example, we use again the leap-frog approximation (2.2.1) for $u_t = u_x$. By Eq. (5.1.25), the order of accuracy is (2,2). The scheme is stable for $k/h < 1$. The only remaining question is how to compute v^1 . By Taylor expansion

$$\begin{aligned} u(x_j, k) &= u(x_j, 0) + ku_t(x_j, 0) + \mathcal{O}(k^2) \\ &= u(x_j, 0) + ku_x(x_j, 0) + \mathcal{O}(k^2) \\ &= (I + kD_0)u(x_j, 0) + \mathcal{O}(kh^2 + k^2). \end{aligned}$$

Therefore, if we use

$$\begin{aligned} v_j^0 &= u(x_j, 0), \\ v_j^1 &= (I + kD_0)v_j^0, \end{aligned}$$

then the initial conditions for the difference approximation are accurate of order (2,2).

Note that the Euler scheme, used here for the first step, is only first-order accurate in time and actually unstable, but it has a *local truncation error* of order k^2 . This is sufficient, because the scheme is applied for only one step. For a multistep scheme, this holds in general: the initial data can be generated by a difference scheme with one order less accuracy.

Consistency does not guarantee that the discrete solutions of the approximation will converge to the solutions of the differential equation as the mesh size tends to zero. It was shown in Section 2.1 that there are consistent schemes which have solutions that grow arbitrarily fast, although the differential equation has a bounded solution. The approximation must also be stable.

Theorem 5.1.3. *Assume that the solution $u(x, t)$ of Eq. (5.1.1) is smooth and that the approximation (5.1.2) is stable. Assume that the approximation and its initial data are accurate of order (p_1, p_2) . Then, on any finite interval $[0, T]$, the error satisfies*

$$\begin{aligned} \|v^n - u(\cdot, t_n)\|_h &\leq K_S(e^{\alpha S t_n} \|v^0 - u(\cdot, 0)\|_h + \|Q_{-1}^{-1}\|_h \varphi_h^*(\alpha, t_n) \max_{0 \leq j \leq n-1} \|\tau^j\|_h) \\ &= \mathcal{O}(h^{p_1} + k^{p_2}), \end{aligned} \quad (5.1.28)$$

that is, the solutions of the difference approximation converge as $h \rightarrow 0$ to the solution of the differential equation.

Proof. If the solution $u(x, t)$ is substituted into the difference scheme (5.1.2), then we obtain

$$Q_{-1}u(x_j, t_{n+1}) = \sum_{\sigma=0}^q Q_\sigma u(x_j, t_{n-\sigma}) + k\tau_j^n. \quad (5.1.29)$$

Let $w_j^n = u(x_j, t_n) - v_j^n$. Subtracting Eq. (5.1.2) from Eq. (5.1.29) yields

$$\begin{aligned} Q_{-1}w_j^{n+1} &= \sum_{\sigma=0}^q Q_\sigma w_j^{n-\sigma} + k\tau_j^n, \\ w_j^\sigma &= u(x_j, \sigma k) - v_j^\sigma, \quad \sigma = 0, 1, \dots, q, \end{aligned}$$

and the estimate follows from Theorem 5.1.1.

REMARK. There is a fundamental difference between the case $\alpha_S < 0$ and $\alpha_S > 0$. If $\alpha_S < 0$, the effect of the error in the initial data disappears, and the total error is of order $\mathcal{O}(h^{p_1} + k^{p_2})$ uniformly in time. Thus, we can solve the equations on long time intervals without deterioration of the error bound. If $\alpha_S > 0$, then the error grows exponentially, and if the solution does not grow as fast, then the error will dominate. Therefore, it is important that the approximation is strictly stable.

We end this section with a discussion of the connection between stability and convergence for generalized solutions. Let $g \in L_2$. As in Section 4.8, we consider a sequence of continuous problems

$$\frac{\partial}{\partial t} u_{[\nu]}(t) = Pu_{[\nu]}(t), \quad u_{[\nu]}(0) = f_{[\nu]}, \quad \text{for } \nu = 1, 2, \dots, \quad (5.1.30)$$

in a fixed time interval $0 \leq t \leq T$. Here $\{f_{[\nu]}\}$ is a sequence of smooth functions with $\lim_{\nu \rightarrow \infty} f_{[\nu]} = g$. Then $u(t) = \lim_{\nu \rightarrow \infty} u_{[\nu]}(t)$ is the generalized solution of Eqs. (5.1.30) with initial data $u(0) = g$. Let

$$\mathbf{v}^{n+1} = Q\mathbf{v}^n$$

be a consistent and stable difference approximation of type (5.1.2). We consider the corresponding sequence

$$\mathbf{v}_{[\nu]}^{n+1} = Q\mathbf{v}_{[\nu]}^n, \quad \mathbf{v}^0 = \mathbf{f}_{[\nu]}. \quad (5.1.31)$$

Let

$$\mathbf{u}_{[\nu]} = (u_{[\nu]}(t_{n+q}), \dots, u_{[\nu]}(t_n))^T, \quad \mathbf{v}_{[\nu]} = \mathbf{v}_{[\nu]}^n.$$

For every fixed ν the solution $u_{[\nu]}$ of Eqs. (5.1.30) is a smooth function. Therefore, for any $\varepsilon > 0$ and all t_n with $0 \leq t_n \leq T$,

$$\|\mathbf{v}_{[\nu]} - \mathbf{u}_{[\nu]}\|_h \leq \varepsilon,$$

provided $h = h(\varepsilon, \nu)$ is sufficiently small. We now use Fourier interpolation to define $\mathbf{v}_{[\nu]}$ everywhere. As in Section 2.1, we denote the Fourier interpolant by $\text{Int}_N \mathbf{v}_{[\nu]}$. Also, $\text{Int}_N \mathbf{u}_{[\nu]}$ denotes the Fourier interpolant of $\mathbf{u}_{[\nu]}$'s restriction to the grid. Then, by Theorem 1.3.3, we obtain, for every fixed ν ,

$$\begin{aligned} \|\text{Int}_N \mathbf{v}_{[\nu]} - \mathbf{u}_{[\nu]}\| &\leq \|\text{Int}_N \mathbf{v}_{[\nu]} - \text{Int}_N \mathbf{u}_{[\nu]}\| + \|\text{Int}_N \mathbf{u}_{[\nu]} - \mathbf{u}_{[\nu]}\|, \\ &= \|\mathbf{v}_{[\nu]} - \mathbf{u}_{[\nu]}\|_h + \|\text{Int}_N \mathbf{u}_{[\nu]} - \mathbf{u}_{[\nu]}\| < 2\varepsilon, \end{aligned}$$

provided $h = h(\varepsilon, \nu)$ is sufficiently small. Therefore, with $\mathbf{u} = (u(t_{n+q}), \dots, u(t_n))^T$,

$$\|\mathbf{u} - \text{Int}_N \mathbf{v}_{[\nu]}\| \leq \|\mathbf{u} - \mathbf{u}_{[\nu]}\| + \|\text{Int}_N \mathbf{v}_{[\nu]} - \mathbf{u}_{[\nu]}\| \leq 3\varepsilon,$$

provided ν is sufficiently large and h is sufficiently small. This proves convergence.

To prove that convergence implies stability, we assume that the approximation is not stable. Then there are sequences

$$h_\nu \rightarrow 0, \quad \nu \rightarrow \infty$$

$$k_\nu \rightarrow 0, \quad \nu \rightarrow \infty,$$

$$\mathbf{f}_{[\nu]}, \|\mathbf{f}_{[\nu]}\| = 1, \quad \nu \rightarrow \infty,$$

$$K_\nu \rightarrow \infty, \quad \nu \rightarrow \infty,$$

such that the solution of Eq. (5.1.31) with initial data $\mathbf{f}_{[\nu]}$ fulfills the inequality

$$\|\mathbf{v}_{[\nu]}^{t/k_\nu}\| > K_\nu \|\mathbf{f}_{[\nu]}\|, \quad \nu > \nu_0$$

for some t in the interval $[0, T]$.

Consider now the solution $\mathbf{w}_{[\nu]}$ of Eq. (5.1.31) with initial data

$$\mathbf{g}_{[\nu]} := \mathbf{f}_{[\nu]}/K_\nu.$$

By linearity it satisfies the inequality

$$\|\mathbf{w}_{[\nu]}^{t/k_\nu}\| > K_\nu \|\mathbf{f}_{[\nu]}/K_\nu\| = \|\mathbf{f}_{[\nu]}\| = 1, \quad \nu > \nu_0.$$

Because $\mathbf{g}_{[\nu]} \rightarrow \mathbf{0}$ as $\nu \rightarrow \infty$, the solution of the underlying well posed continuous problem is $w \equiv 0$. This contradicts convergence.

Thus we have proved the following theorem.

Theorem 5.1.4. *If the difference approximation (5.1.2) is stable and consistent, then we obtain convergence even if the underlying continuous problem only has a generalized solution. If the approximation is convergent, then it is stable.*

Theorem 5.1.4 is the classical Lax–Richtmyer equivalence theorem, which states that convergence is equivalent to consistency and stability. We have proved that stability plays a very important role for numerical methods. In the next sections, we derive algebraic conditions that allow us to decide whether a given method is stable.

EXERCISES

5.1.1. Derive the order of accuracy for the following difference approximations of $u_t = Au_x - u$, where A is a matrix.

1. $v^{n+1} = (I - kI + kAD_+)v^n$.

2. $(I + kI - kAD_+)v^{n+1} = v^n$.

3. $(I + kI)v^{n+1} = 2kAD_0v^n + (I - kI)v^{n-1}$.

5.1.2. Find a suitable method to compute v^1 when the DuFort–Frankel method (2.5.12) is used to approximate Eq. (2.5.1). Derive an error estimate for the solutions. (You may assume that stability holds for all values of k/h^2 .)

5.1.3. Derive the order of accuracy of the approximation

$$(I - kD_0D_+D_-)v^{n+1} = 2kaD_0v^n + (I + kD_0D_+D_-)v^{n-1}$$

for $u_t = u_{xxx} + au_x$.

5.2. STABILITY FOR APPROXIMATIONS WITH CONSTANT COEFFICIENTS

For approximations with constant coefficients, one can use Fourier analysis to discuss their stability properties. To simplify the notation, we only treat the one-dimensional case, and assume that the grid has $N + 1$ points in $[0, 2\pi]$. We assume that the operators Q_σ in Eqs. (5.1.2) have the form

$$Q_\sigma = \sum_{\nu=-r}^p A_{\nu\sigma} E^\nu, \quad (5.2.1)$$

where the matrices $A_{\nu\sigma}$ are smooth functions of h but do not depend on x_j or t_n . Let v_j^n be a solution of Eqs. (5.1.2). We can represent it by its interpolating polynomial, that is, by

$$v_j^n = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} e^{i\omega x_j} \tilde{v}^n(\omega), \quad j = 0, 1, \dots, N,$$

in the grid points. Substituting this expression into Eqs. (5.1.2) gives us

$$\sum_{\omega=-N/2}^{N/2} (Q_{-1} e^{i\omega x_j}) \tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^q (Q_\sigma e^{i\omega x_j}) \tilde{v}^{n-\sigma}(\omega), \quad j = 0, 1, \dots, N. \quad (5.2.2)$$

As in Section 2.1,

$$Q_\sigma e^{i\omega x_j} = e^{i\omega x_j} \hat{Q}_\sigma(\xi), \quad \hat{Q}_\sigma(\xi) = \sum_{\nu=-r}^p A_{\nu\sigma} e^{i\nu\xi}, \quad \xi = \omega h, \quad (5.2.3)$$

denotes the so-called *symbol*, or *Fourier transform*, of Q_σ . Equation (5.2.2) shows that the $\tilde{v}^{n+1}(\omega)$ are determined by

$$\hat{Q}_{-1}(\xi) \tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^q \hat{Q}_\sigma(\xi) \tilde{v}^{n-\sigma}(\omega), \quad (5.2.4)$$

because the vectors $(1, e^{i\omega h}, \dots, e^{i\omega Nh})^T$, $\omega = -N/2, \dots, N/2$, are linearly independent. Now we can prove the following lemma.

Lemma 5.2.1. *The inverse Q_{-1}^{-1} exists and*

$$\|Q_{-1}^{-1}\|_h \leq C, \quad \text{for } 0 < h \leq h_0, \quad h = 2\pi/(N + 1),$$

if and only if $\hat{Q}_{-1}^{-1}(\xi)$ exists and

$$|\hat{Q}_{-1}^{-1}(\xi)| \leq C, \quad \text{for } 0 < h \leq h_0, \quad \omega = 0, \pm 1, \dots, \pm N/2. \quad (5.2.5)$$

Proof. \hat{Q}_{-1}^{-1} exists if and only if the equation

$$\hat{Q}_{-1}w_j = g_j, \quad j = 0, 1, \dots, N, \quad (5.2.6)$$

has a unique solution. As above, we can represent w_j, g_j by their Fourier interpolants, and Eq. (5.2.6) is equivalent to

$$\hat{Q}_{-1}(\xi)\tilde{w}(\omega) = \tilde{g}(\omega). \quad (5.2.7)$$

Then Eq. (5.2.5) follows from the discrete version of Parseval's relation

$$\|w\|_h^2 = \sum_{\omega=-N/2}^{N/2} |\tilde{w}(\omega)|^2 = \sum_{\omega=-N/2}^{N/2} |\hat{Q}_{-1}^{-1}(\xi)\tilde{g}(\omega)|^2 = \|\hat{Q}_{-1}^{-1}g\|_h^2.$$

We look at some examples and begin with the Crank–Nicholson and Euler backward approximations of the hyperbolic system $u_t = Au_x$ [see Eqs. (2.3.3) and (2.3.1)]. The operator \hat{Q}_{-1} has the form

$$\hat{Q}_{-1} = I - \theta k A D_0, \quad \theta = 1/2, 1.$$

Here A is a matrix with real eigenvalues a_ν . Then

$$\hat{Q}_{-1} = I - \theta \lambda i A \sin \xi, \quad \lambda = k/h,$$

and the eigenvalues z_ν of \hat{Q}_{-1} are

$$z_\nu = 1 - \theta \lambda a_\nu i \sin \xi.$$

Because $|z_\nu| \geq 1$ for all ξ, h, k , the condition (5.2.5) is fulfilled.

As a second example, we consider the *box scheme* for the same equation

$$v_{j+1}^{n+1} + v_j^{n+1} - v_{j+1}^n - v_j^n = \lambda A(v_{j+1}^{n+1} - v_j^{n+1} + v_{j+1}^n - v_j^n). \quad (5.2.8)$$

\hat{Q}_{-1} is given by

$$\hat{Q}_{-1} = I + \lambda A + (I - \lambda A)E,$$

hence

$$\hat{Q}_{-1} = I + \lambda A + (I - \lambda A)e^{i\xi}.$$

If A has an eigenvalue $a_\nu = 0$, then \hat{Q}_{-1} has an eigenvalue

$$z_\nu = 1 + e^{i\xi}.$$

Therefore, $z_\nu = 0$ for $\xi = \pi$, and the condition (5.2.5) is not fulfilled. This shows that one has to be careful when using the box scheme. For many other types of boundary conditions this difficulty does not arise.

Now assume that Eq. (5.2.5) holds and let

$$\tilde{\mathbf{v}}^n(\omega) = (\tilde{v}^{n+q}(\omega), \tilde{v}^{n+q-1}(\omega), \dots, \tilde{v}^n(\omega))^T.$$

Then we can write Eq. (5.2.4) as a one-step method

$$\tilde{\mathbf{v}}^{n+1}(\omega) = \hat{Q}(\xi)\tilde{\mathbf{v}}^n(\omega), \quad |\omega| \leq N/2, \quad (5.2.9)$$

where

$$\hat{Q} = \begin{bmatrix} \hat{Q}_{-1}^{-1}\hat{Q}_0 & \hat{Q}_{-1}^{-1}\hat{Q}_1 & \cdots & \hat{Q}_{-1}^{-1}\hat{Q}_q \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix}. \quad (5.2.10)$$

We can now prove the next theorem.

Theorem 5.2.1. *The approximation (5.1.4) is stable if, and only if, Eq. (5.2.5) holds and*

$$|\hat{Q}^n(\xi)| \leq K_S e^{\alpha s t_n}, \quad (5.2.11)$$

for all h with $h = 2\pi/(N+1) \leq h_0$ and all $|\omega| \leq N/2$.

Proof. The theorem follows directly from Parseval's relation

$$\|\mathbf{v}^n\|_h^2 = \sum_{\omega} |\tilde{\mathbf{v}}^n(\omega)|^2 = \sum_{\omega} |\hat{Q}^n(\xi)\tilde{\mathbf{v}}^0(\omega)|^2.$$

For one-step approximations of scalar differential equations, \hat{Q} is a scalar and is easy to check. When \hat{Q} is a matrix, the condition is generally difficult to verify. It is convenient to replace it by conditions on the eigenvalues of \hat{Q} .

Theorem 5.2.2. *A necessary condition for stability is that the eigenvalues z_ν of \hat{Q} satisfy*

$$|z_\nu| \leq e^{\alpha s k}, \quad |\xi| \leq \pi \quad (5.2.12)$$

for all h with $h \leq h_0$ (the von Neumann condition).

Proof. z_j^n is an eigenvalue of \hat{Q}^n . Therefore, if the method is stable, we obtain for any n

$$|z_j^n| \leq |\hat{Q}^n| \leq K_S e^{\alpha s n k},$$

or

$$|z_j| \leq K_S^{1/n} e^{\alpha s k},$$

and Eq. (5.2.12) follows because n can be arbitrarily large.

The von Neumann condition is often thought to be a sufficient condition for stability. However, the following example shows that this is not true. Consider the trivial system of differential equations

$$u_t = 0, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \quad (5.2.13)$$

and the approximation

$$v^{n+1} = \left(I - \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} h^2 D_0^2 \right) v^n, \quad k = h, \quad (5.2.14)$$

which has order of accuracy (1,1). The symbol is

$$\hat{Q} = \begin{bmatrix} 1 & \sin^2 \xi \\ 0 & 1 \end{bmatrix}, \quad (5.2.15)$$

which satisfies the von Neumann condition. The powers of \hat{Q} are easily computed

$$\hat{Q}^n = \begin{bmatrix} 1 & n \sin^2 \xi \\ 0 & 1 \end{bmatrix}, \quad (5.2.16)$$

and the norm is of the order n , which cannot be bounded by $K e^{\alpha t_n}$.

The condition (5.2.11) can be expressed in a slightly different form: Multiplying it by $e^{-\alpha t_n}$, the condition becomes

$$|(e^{-\alpha s k} \hat{Q})^n| \leq K_S, \quad (5.2.17)$$

and the problem is reduced to finding conditions, which guarantee that a family of matrices are *power bounded*. If the parameters ξ, h are fixed, then it is well known what conditions the eigenvalues must satisfy to guarantee Eq. (5.2.17). They must be less than or equal to one in magnitude, and those on the unit circle must be distinct. The difficulty lies in the fact that the power-boundedness must be uniform in ξ, h .

In the special case that \hat{Q} can be uniformly diagonalized, the von Neumann condition is sufficient.

Theorem 5.2.3. *Assume that there is a matrix $T = T(\xi, h)$ with $|T| \cdot |T^{-1}| \leq C$, for C independent of ξ and h , such that*

$$T^{-1} \hat{Q} T = \text{diag}(z_1, z_2, \dots, z_{m(q+1)}). \quad (5.2.18)$$

Then the von Neumann condition (5.2.12) is sufficient for stability.

Proof. We have, with $\rho(A)$ denoting the spectral radius of a matrix A ,

$$\begin{aligned} |\hat{Q}^n| &= |TT^{-1}\hat{Q}TT^{-1}\hat{Q}T \dots T^{-1}\hat{Q}TT^{-1}| \\ &\leq |T| \cdot |\text{diag}(z_1, \dots, z_{m(q+1)})|^n \cdot |T^{-1}| \leq C\rho(\hat{Q}^n) \\ &\leq Ce^{\alpha S t_n}. \end{aligned}$$

Normal matrices are diagonalized by orthogonal matrices T with $|T| = |T^{-1}| = 1$. Therefore, we have the following corollary.

Corollary 5.2.1. *If \hat{Q} is a normal matrix, that is, if $\hat{Q}^*\hat{Q} = \hat{Q}\hat{Q}^*$, then the von Neumann condition is sufficient for stability. In particular, this is true for Hermitian and skew-Hermitian matrices \hat{Q} .*

As noted earlier, this means, in particular, that the von Neumann condition is sufficient for all one-step approximations of scalar differential equations.

To find the eigenvalues for multistep schemes, it is easier to work with the original multistep form (5.1.2) and the corresponding formula in Fourier space,

$$\hat{Q}_{-1} \tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^q \hat{Q}_\sigma \tilde{v}^{n-\sigma}(\omega). \quad (5.2.19)$$

We need the next lemma.

Lemma 5.2.2. *The eigenvalues z of the matrix \hat{Q} , defined by Eq. (5.2.10), are solutions of*

$$\det \left(\hat{Q}_{-1} z^{q+1} - \sum_{\sigma=0}^q \hat{Q}_\sigma z^{q-\sigma} \right) = 0. \quad (5.2.20)$$

This equation is usually called the characteristic equation for Eq. (5.2.19), and is formally obtained by the substitution $v^n \rightarrow z^n$.

Proof. The eigenvalue problem for \hat{Q} is

$$\hat{Q}\tilde{\mathbf{v}} = z\tilde{\mathbf{v}}, \quad \tilde{\mathbf{v}} = (\tilde{v}^q, \tilde{v}^{q-1}, \dots, \tilde{v}^0)^T.$$

An eigenvalue z with eigenvector $\tilde{\mathbf{v}}$ must satisfy

$$\begin{aligned} \hat{Q}_{-1}^{-1} \hat{Q}_0 \tilde{v}^q + \hat{Q}_{-1}^{-1} \hat{Q}_1 \tilde{v}^{q-1} + \cdots + \hat{Q}_{-1}^{-1} \hat{Q}_q \tilde{v}^0 &= z\tilde{v}^q, \\ \tilde{v}^q &= z\tilde{v}^{q-1}, \\ \tilde{v}^{q-1} &= z\tilde{v}^{q-2}, \\ &\vdots \\ \tilde{v}^1 &= z\tilde{v}^0. \end{aligned}$$

All vectors \tilde{v}^ν can be expressed in terms of \tilde{v}^0 , and we get, from the first equation,

$$\left(\hat{Q}_{-1}^{-1} \sum_{\sigma=0}^q \hat{Q}_\sigma z^{q-\sigma} - z^{q+1} I \right) \tilde{v}^0 = 0.$$

After multiplying by \hat{Q}_{-1} the determinant condition (5.2.20) follows.

EXAMPLES. We begin with the leap-frog scheme (2.2.1) for $u_t = u_x$. The determinant condition for the eigenvalues z just derived gives us Eq. (2.2.5) for the solutions z_1, z_2 . If $|\lambda \sin \xi| = 1$, there is a double eigenvalue, $z = i$ or $z = -i$. In the first case, the amplification matrix \hat{Q} is

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

Let T_1 be the matrix which transforms \hat{Q} to Jordan canonical form:

$$T_1^{-1} \hat{Q} T_1 = \begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}.$$

Then

$$\hat{Q}^n = T_1 \begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}^n T_1^{-1}$$

is unbounded. Again, the von Neumann condition is not sufficient. If, on the other hand,

$$\lambda \leq \lambda_0 < 1, \quad (5.2.21)$$

then $|z_1 - z_2| \geq \delta > 0$, where δ is independent of ξ, h . Therefore, \hat{Q} can be uniformly diagonalized and the von Neumann condition is sufficient.

Now consider a system $u_t = Au_x$, where A is a diagonalizable matrix such that $T^{-1}AT = D = \text{diag}(a_1, \dots, a_m)$. Substituting new variables $w = T^{-1}v$ into the leap-frog scheme

$$v^{n+1} = 2kAD_0v^n + v^{n-1} \quad (5.2.22)$$

gives us

$$w^{n+1} = 2kDD_0w^n + w^{n-1}, \quad (5.2.23)$$

which is a set of m scalar equations. We have

$$\left| \frac{ka_\nu}{h} \right| \leq \lambda_0 < 1, \quad \nu = 1, \dots, m, \quad (5.2.24)$$

which corresponds to the condition (5.2.21), and a necessary and sufficient stability condition for Eq. (5.2.22) is

$$\frac{k}{h} \rho(A) \leq \lambda_0 < 1. \quad (5.2.25)$$

General stability conditions are complicated. Without proof we state the *Kreiss matrix theorem*:

Theorem 5.2.4. *Let F be a family of matrices A of fixed order. The following four statements are equivalent. (The constants $C_1, C_2, C_{31}, C_{32}, C_4$ are fixed for a given family F .)*

1. The matrices in F are uniformly power bounded, that is,

$$|A^n| \leq C_1, \quad \text{for all integers } n \geq 0. \quad (5.2.26)$$

2. For all $A \in F$, the resolvent matrix $(A - zI)^{-1}$ exists for all complex numbers z , $|z| > 1$, and

$$|(A - zI)^{-1}| \leq \frac{C_2}{|z| - 1} \quad (\text{the resolvent condition}). \quad (5.2.27)$$

3. For each $A \in F$, there is a matrix T with $|T| \leq C_{31}$ and $|T^{-1}| \leq C_{31}$, such that

$$T^{-1}AT = \begin{bmatrix} z_1 & b_{12} & & \cdots & b_{1m} \\ & z_2 & b_{23} & \cdots & b_{2m} \\ & & \ddots & & \vdots \\ & & & & z_m \end{bmatrix},$$

where

$$|z_m| \leq |z_{m-1}| \leq \cdots \leq |z_1| \leq 1, \quad (5.2.28)$$

and

$$\begin{aligned} |b_{\nu\mu}| &\leq C_{32}(1 - |z_\nu|), \\ \nu &= 1, 2, \dots, m - 1, \\ \mu &= \nu + 1, \nu + 2, \dots, m. \end{aligned} \quad (5.2.29)$$

4. For each $A \in F$ there is a positive definite matrix H such that

$$\begin{aligned} C_4^{-1}I &\leq H \leq C_4I, \\ A^*HA &\leq (1 - \delta)H, \quad \delta = \frac{1}{2}(1 - \max_{1 \leq \nu \leq m} |z_\nu|), \end{aligned} \quad (5.2.30)$$

where $\{z_\nu\}$ are the eigenvalues of A .

This theorem is, in general, not easy to apply. The condition 3 is probably the most straightforward, because of its resemblance to Schur's normal form. However, it is difficult to find a matrix T with a bounded inverse that allows one to satisfy the inequalities (5.2.29) for the off-diagonal elements.

To obtain simpler stability conditions, one can require additional properties that are naturally built into the approximation. *Dissipativity* is such a property.

Recall from Sections 2.2 and 2.3 that some of the methods presented there damp the amplitudes of most frequencies. The word dissipation is sometimes used to describe any kind of decrease in norm. We define it more precisely here.

Definition 5.2.1. *The approximation (5.1.4) is dissipative of order $2r$ if all the eigenvalues z_ν of \hat{Q} satisfy*

$$|z_\nu| \leq (1 - \delta|\xi|^{2r})e^{\alpha sk}, \quad |\xi| \leq \pi, \quad (5.2.31)$$

where $\delta > 0$ is a positive constant independent of ξ, h .

The two important properties enforced by Eq. (5.2.31) can be expressed in the following way:

1. There is a damping of order $1 - \delta$ for all large wave numbers.
2. The damping factor approaches 1 in magnitude as a polynomial of degree $2r$ for small wave numbers.

We analyze the dissipative behavior of a few simple methods. As a first example, consider the parabolic equation $u_t = u_{xx}$ and the Euler approximation (2.5.6). There is only one eigenvalue

$$z = \hat{Q} = 1 - 4\sigma \sin^2 \frac{\xi}{2}, \quad \sigma = k/h^2. \quad (5.2.32)$$

This method is stable if $\sigma \leq \frac{1}{2}$, but for $\sigma = \frac{1}{2}$ the scheme is not dissipative, because $z = -1$ for $\xi = \pi$. If $\sigma < \frac{1}{2}$, then $|z| < 1$ for $0 < |\xi| \leq \pi$ and

$$z = 1 - \sigma(\xi^2 + \mathcal{O}(\xi^4)) \quad (5.2.33)$$

in a neighborhood of $\xi = 0$. Therefore the scheme is dissipative of order 2.

The dissipative property is very natural for parabolic problems, because the differential equation itself is dissipative [see Eq. (2.5.2)]. Actually, the relation (5.2.33), which is restricted to a neighborhood of $\xi = 0$, is a consequence of consistency.

As we will see in Section 6.5, the dissipative property also plays an important role for hyperbolic problems. Consider the Lax–Friedrichs method (2.1.16) for $u_t = u_x$. Again, there is only one eigenvalue z , and

$$|\hat{Q}|^2 = |z|^2 = |\cos \xi + i\lambda \sin \xi|^2 = 1 - (1 - \lambda^2) \sin^2 \xi. \quad (5.2.34)$$

The scheme exhibits the correct behavior (corresponding to dissipativity of order 2) near $\xi = 0$, if $\lambda < 1$. However, $|z| = 1$ for $\xi = \pi$, so the scheme is

not dissipative. (Some authors define approximations of this type to be dissipative and they call those schemes satisfying Definition 5.2.1 *strictly dissipative*.)

The backward Euler method (2.3.1) suffers from the same deficiency as the Lax–Friedrichs method: the inequality (5.2.31) fails at $\xi = \pi$. The Crank–Nicholson method (2.3.3) has no damping at all, and z is on the unit circle for all ξ .

To construct a one-step dissipative method for $u_t = u_x$, consider the approximation

$$v^{n+1} = (I + kD_0 + k^2 c D_+ D_-) v^n, \quad (5.2.35)$$

where c is constant, treated in Chapter 2. The amplification factor is

$$\hat{Q} = 1 + \lambda i \sin \xi - 4\lambda^2 c \sin^2 \frac{\xi}{2} \quad (5.2.36)$$

with

$$|\hat{Q}|^2 = 1 - 4\lambda^2(2c - 1) \sin^2 \frac{\xi}{2} - 4\lambda^2(1 - 4\lambda^2 c^2) \sin^4 \frac{\xi}{2},$$

where $\lambda = k/h$. In Section 2.1 we found that $|\hat{Q}| \leq 1$ for

$$\lambda \leq 1, \quad \frac{1}{2} \leq c. \quad (5.2.37)$$

With strict inequalities in Eq. (5.2.37), the scheme is dissipative of order 2. If $c = \frac{1}{2}$ and $\lambda < 1$, the order of dissipativity is 4. This is the *Lax–Wendroff method* (2.1.17), which is accurate of order (2,2).

We end this section by showing that conditions for consistency and order of accuracy can be given in Fourier space. For convenience, we limit ourselves to first-order systems.

Theorem 5.2.5. *We make the following assumptions:*

1. *The differential equation $u_t = Pu$ has constant coefficients.*
2. *The difference operators in the approximation (5.1.2) have the form (5.2.1), where the coefficients $A_{\nu\alpha}$ are independent of x, t and h , ($k = \lambda h$, λ constant).*

Then Eq. (5.1.2) has order of accuracy p if, for some constant $\xi_0 > 0$,

$$\begin{aligned} & \left| \hat{Q}_{-1}(\xi) e^{\hat{P}(i\omega)k} - \sum_{\sigma=0}^q \hat{Q}_\sigma(\xi) e^{-\hat{P}(i\omega)\sigma k} \right| \\ & \leq \text{constant } |\xi|^{p+1} \quad \text{for } |\xi| = |\omega h| \leq \xi_0. \end{aligned} \quad (5.2.38)$$

3. For explicit one-step methods, $Q_0(\xi)$ must agree with the Taylor expansion of $\exp(\hat{P}(i\omega)k)$ through terms of order $|\xi|^{p+1}$.

Proof. The solution to Eq. (5.1.1) is written in the form

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

where

$$\hat{u}(\omega, t) = e^{\hat{P}(i\omega)t} \hat{u}(\omega, 0).$$

By Definition 5.1.2, the truncation error is

$$k\tau(\cdot, t_n) = \sum_{\omega=-\infty}^{\infty} \left(\hat{Q}_{-1}(\xi) e^{\hat{P}(i\omega)k} - \sum_{\sigma=0}^q \hat{Q}_\sigma e^{-\hat{P}(i\omega)\sigma k} \right) e^{i\omega x} \hat{u}(\omega, t_n).$$

If Eq. (5.2.38) holds and the solution $u(x, t)$ is smooth, then

$$\begin{aligned} \|k\tau(\cdot, t_n)\|_h^2 & \leq \text{constant} \sum_{\omega=-\infty}^{\infty} |\xi|^{2(p+1)} |\hat{u}(\omega, t_n)|^2, \\ & = \text{constant } h^{2(p+1)} \sum_{\omega=-\infty}^{\infty} |\omega|^{2(p+1)} |\hat{u}(\omega, t_n)|^2, \\ & \leq \text{constant } h^{2(p+1)}, \end{aligned}$$

which shows that the order of accuracy is p .

To prove the theorem in the other direction, we let $\hat{u}(\omega, t_n) = 0$ for $\omega \neq \omega_1 \neq 0$, and $|\hat{u}(\omega_1, t_n)| = 1$. Then

$$\begin{aligned} \text{constant } ch^{p+1} & \geq \|k\tau(\cdot, t_n)\|_h \\ & = \left\| \left(\hat{Q}_{-1}(\xi) e^{\hat{P}(i\omega_1)k} - \sum_{\sigma=0}^q \hat{Q}_\sigma e^{-\hat{P}(i\omega_1)\sigma k} \right) e^{i\omega_1 x} \hat{u}(\omega_1, t_n) \right\|_h. \end{aligned}$$

Because $\hat{u}(\omega_1, t_n)$ is arbitrary, except for normalization, we get the necessary inequality

$$\left| Q_{-1}(\xi) e^{\hat{P}(i\omega_1)k} - \sum_{\sigma=0}^q \hat{Q}_\sigma e^{-\hat{P}(i\omega_1)\sigma k} \right| \leq \text{constant } h^{p+1} \leq \text{constant } |\omega_1 h|^{p+1}.$$

But ω_1 is arbitrary, so this proves Eq. (5.2.38). For explicit one-step schemes, Eq. (5.2.38) becomes

$$|e^{P(i\omega)k} - Q_0(\xi)| \leq \text{constant } |\xi|^{p+1}.$$

This proves the theorem.

The theory in this section has been developed for one space dimension, but it can be extended to several space dimensions without difficulty. In fact, all of the definitions, lemmas, and theorems hold as formulated, the only modification that needs to be made is that ω is a multiindex.

EXERCISES

- 5.2.1.** Formulate and prove Theorem 5.2.1 for approximations in d space dimensions.
- 5.2.2.** Define the Lax–Wendroff approximation for the system $u_t = Au_x$. Is it dissipative if

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

5.3. APPROXIMATIONS WITH VARIABLE COEFFICIENTS: THE ENERGY METHOD

In this section, we consider approximations whose coefficients depend on x_j and/or t_n . If there is only t dependence, the Fourier technique used in the previous chapter can still be used. However, the solution operator in Fourier space is a product of operators $\hat{Q}(t_\nu)$ in this case. We do not discuss this further.

If the coefficients depend on x , then the situation is different. The separation of variables technique, using the ansatz $v_j^n = (1/\sqrt{2\pi}) \sum_\omega \tilde{v}^n(\omega) e^{i\omega x_j}$, does not work. It is not possible to get a relation of the form

$$\hat{Q}_{-1}\tilde{v}^{n+1}(\omega) = \sum_{\sigma=0}^q \hat{Q}_\sigma \tilde{v}^{n-\sigma}(\omega)$$

for the Fourier coefficients.

The Fourier technique can still be used, but it must be embedded in more involved machinery. Stability conditions are derived for related problems with **frozen coefficients**, that is, the approximation is examined with $x = x_*$ and $t = t_*$ held fixed in the coefficients for every x_* and t_* in the domain. With certain extra conditions, it is possible to prove stability for the variable coefficient problem if all the “frozen” problems are stable. This has been done for approximations of hyperbolic and parabolic equations, and the results for hyperbolic equations will be given in Sections 6.5 and 6.6.

A more direct way of proving stability is the **energy method**, and this method will be developed in this section. In this method, no transformations are used, and the analysis is carried out directly in physical space.

The idea is simple (but the algebra involved may be difficult): Construct a norm $\|\cdot\|_h^*$ such that **the growth in each step is at most $e^{\alpha k}$** , that is

$$\|v^{n+1}\|_h^* \leq e^{\alpha k} \|v^n\|_h^*. \quad (5.3.1)$$

If this norm is **equivalent to the usual discrete l_2 -norm $\|\cdot\|_h$** , then

$$\|v^n\|_h \leq C_1 \|v^n\|_h^* \leq C_1 e^{\alpha t_n} \|v^0\|_h^* \leq C_2 e^{\alpha t_n} \|v^0\|_h, \quad (5.3.2)$$

which is the type of estimate that we want.

We first illustrate the idea using a simple example. The Crank–Nicholson method for $u_t = u_x$ can be written in the form

$$v_j^{n+1} - v_j^n = \frac{k}{2} Q(v_j^{n+1} + v_j^n), \quad j = 0, 1, \dots, N, \quad (5.3.3)$$

where $Q = D_0$. Assuming that v is real, we multiply both sides by $v_j^{n+1} + v_j^n$ and sum over all j . The result is

$$\|v^{n+1}\|_h^2 - \|v^n\|_h^2 = \left(v^{n+1} + v^n, \frac{k}{2} Q(v^{n+1} + v^n) \right)_h; \quad (5.3.4)$$

so $\|v^{n+1}\|_h \leq \|v^n\|_h$ if, for all gridfunctions v , the condition

$$(v, Qv)_h \leq 0 \quad (5.3.5)$$

is fulfilled. By periodicity,

$$(v, Qv)_h = \frac{1}{2} \sum_{j=0}^N v_j (v_{j+1} - v_{j-1}) = 0,$$

and stability is proved for all k . This was already shown in Section 2.3 using the Fourier technique. However, the energy method can be easily generalized to variable coefficients. Consider $u_t = a(x)u_x$, and the same approximation (5.3.3) where $Q = Q_j = a_j D_0$. We now have

$$(v, Qv)_h = \frac{1}{2} \sum_{j=0}^N v_j a_j (v_{j+1} - v_{j-1}) = \frac{1}{2} \sum_{j=0}^N (a_j - a_{j+1}) v_j v_{j+1}.$$

Assume that $a(x)$ is Lipschitz continuous, that is, $|D_+ a_j| \leq C$. Then

$$|(v, Qv)_h| \leq \frac{C}{2} \sum_{j=0}^N \frac{1}{2} (v_j^2 + v_{j+1}^2) h = \frac{C}{2} \|v\|_h^2. \quad (5.3.6)$$

From Eq. (5.3.3) we get

$$\|v^{n+1}\|_h^2 - \|v^n\|_h^2 \leq \frac{Ck}{4} \|v^{n+1} + v^n\|_h^2 \leq \frac{Ck}{2} (\|v^{n+1}\|_h^2 + \|v^n\|_h^2),$$

so

$$\|v^{n+1}\|_h^2 \leq \frac{1 + Ck/2}{1 - Ck/2} \|v^n\|_h^2 \leq (1 + \alpha_1 k) \|v^n\|_h^2, \quad (5.3.7)$$

where α_1 is a constant depending on C . Thus, Eq. (5.3.1) is satisfied with $\alpha = \alpha_1/2$ and $\|\cdot\|_h^* = \|\cdot\|_h$.

The inequality (5.3.5) is the key to the final stability estimate. We now treat general operators $Q = Q(x_j, t_n)$. As for differential operators, we define semi-boundedness:

Definition 5.3.1. *The discrete operator Q is semibounded if, for all periodic gridfunctions v , the inequality*

$$\operatorname{Re}(v, Qv)_h \leq \alpha \|v\|_h^2 \quad (5.3.8)$$

holds, where α is independent of h , x , t , and v .

For semidiscrete approximations,

$$\begin{aligned}\frac{dv_j}{dt} &= Qv_j, \quad j = 0, 1, \dots, N, \\ v_j(0) &= f_j,\end{aligned}\tag{5.3.9}$$

the so-called *method of lines*, we obtain the following theorem.

Theorem 5.3.1. *The solutions of Eq. (5.3.9) satisfy the estimate*

$$\|v(t)\|_h \leq e^{\alpha t} \|f\|_h \tag{5.3.10}$$

if Q is *semibounded*.

Proof.

$$\begin{aligned}\frac{d}{dt} \|v\|_h^2 &= 2 \operatorname{Re} \left(v, \frac{dv}{dt} \right)_h = 2 \operatorname{Re} (v, Qv)_h \\ &\leq 2\alpha \|v\|_h^2.\end{aligned}$$

Therefore, Eq. (5.3.10) follows.

We discretize time using the trapezoidal rule and obtain Theorem 5.3.2.

Theorem 5.3.2. *The approximation (5.3.3) is *unconditionally stable* if Q is *semibounded*. The solution satisfies the estimate*

$$\|v^n\|_h \leq e^{\beta(1+\tilde{C}(k))t_n} \|f\|_h, \tag{5.3.11}$$

where $\beta = \max(0, \alpha)$ and α is defined in Eq. (5.3.8).

Proof. Taking the scalar product of Eq. (5.3.3) with $v^{n+1} + v^n$ yields

$$\begin{aligned}&\operatorname{Re} (v^{n+1} + v^n, v^{n+1} - v^n)_h \\ &= \frac{k}{2} \operatorname{Re} (v^{n+1} + v^n, Q(v^{n+1} + v^n))_h, \\ &\leq \begin{cases} \frac{k\alpha}{2} \|v^{n+1} + v^n\|_h^2 \leq k\alpha(\|v^{n+1}\|_h^2 + \|v^n\|_h^2), & \text{if } \alpha > 0, \\ 0, & \text{if } \alpha \leq 0. \end{cases}\end{aligned}$$

Therefore,

$$\|v^{n+1}\|_h^2 - \|v^n\|_h^2 \leq 0, \quad \text{if } \alpha \leq 0,$$

otherwise,

$$(1 - k\alpha)\|v^{n+1}\|_h^2 \leq (1 + k\alpha)\|v^n\|_h^2,$$

and Eq. (5.3.11) follows.

The estimate (5.3.10) shows that the constant α determines the growth rate of the solution of the semidiscrete approximation. The fully discretized approximation preserves this growth rate with $\mathcal{O}(k)$ error only if $\alpha \geq 0$. We remark that Eq. (5.3.11) can be strengthened for $\alpha < 0$ with additional hypotheses on Q .

The backwards Euler method

$$\begin{aligned} (I - kQ)v^{n+1} &= v^n, \\ v^0 &= f, \end{aligned} \tag{5.3.12}$$

preserves the correct growth rate in general.

Theorem 5.3.3. *The backwards Euler method (5.3.12) is unconditionally stable if Q is semibounded. The solution satisfies the estimate*

$$\|v^n\|_h \leq e^{\alpha(1+\mathcal{O}(k))t_n} \|f\|_h. \tag{5.3.13}$$

Proof. Taking the scalar product of Eq. (5.3.12) with v^{n+1} gives us

$$\|v^{n+1}\|_h^2 - k \operatorname{Re}(v^{n+1}, Qv^{n+1})_h = \operatorname{Re}(v^{n+1}, v^n)_h \leq \|v^{n+1}\|_h \|v^n\|_h,$$

that is,

$$(1 - \alpha k)\|v^{n+1}\|_h^2 \leq \|v^{n+1}\|_h \|v^n\|_h,$$

and Eq. (5.3.13) follows.

We need the difference version of Lemma 4.6.1.

Lemma 5.3.1. *Let u, v be complex valued periodic vector gridfunctions, then*

$$(u, Ev)_h = (E^{-1}u, v)_h, \tag{5.3.14a}$$

$$(u, D_0 v)_h = -(D_0 u, v)_h, \quad (5.3.14b)$$

$$(u, D_+ v)_h = -(D_- u, v)_h. \quad (5.3.14c)$$

Proof. By definition

$$\begin{aligned} (u, E v)_h &= \sum_{j=0}^N \langle u_j, v_{j+1} \rangle h = \sum_{j=1}^{N+1} \langle u_{j-1}, v_j \rangle h, \\ &= \sum_{j=0}^N \langle u_{j-1}, v_j \rangle h + (\langle u_N, v_{N+1} \rangle - \langle u_{-1}, v_0 \rangle) h, \\ &= (E^{-1} u, v)_h. \end{aligned}$$

We can express $2hD_0 = E + E^{-1}$, $hD_+ = E - I$, and $hD_- = I - E^{-1}$ in terms of E , and the rest of the lemma follows.

A consequence is

$$(u, D_0 u)_h = -(D_0 u, u)_h = -\overline{(u, D_0 u)_h},$$

that is,

$$\text{Re}(u, D_0 u)_h = 0.$$

We also need the following lemma.

Lemma 5.3.2. Let u and v be complex vector gridfunctions, and let $A = A(x_j)$ be a Lipschitz continuous matrix. If D is any of the difference operators D_+, D_-, D_0 , then

$$(u, D A v)_h = (u, A D v)_h + R, \quad (5.3.15)$$

where

$$|R| \leq \text{constant} \|u\|_h \cdot \|v\|_h.$$

Proof.

$$\begin{aligned} E^\nu(A_j v_j) &= A_{\nu+j} v_{\nu+j} = A_j E^\nu v_j + h B_j E^\nu v_j, \\ B_j &= (A_{\nu+j} - A_j)/h, \end{aligned}$$

that is,

$$\|E^\nu(A_j v_j) - A_j E^\nu v_j\|_h \leq \text{constant } h \|v_j\|_h.$$

By expressing the difference operators in terms of E , the lemma follows.

For convenience, we now only consider examples with real solutions. Consider a system $u_t = Au_x$, where $A = A(x, t)$ is a symmetric matrix. The semi-boundedness condition (5.3.8) for the centered difference operator AD_0 follows from Lemma 5.3.2. Define R as in Eq. (5.3.15). Then

$$(u, AD_0 u)_h = (u, D_0 A u)_h - R = -(D_0 u, A u)_h - R = -(AD_0 u, u)_h - R,$$

and it follows that

$$(u, AD_0 u)_h \leq \text{constant } \|u\|_h^2. \quad (5.3.16)$$

This shows that both the Crank-Nicholson and backward Euler methods are stable with $Q = Q_j = A(x_j)D_0$. Note that the symmetry of A is essential for the energy method to succeed. If A is not symmetric, it must first be symmetrized. Consider, for simplicity, a constant matrix A , and let T be a symmetrizer; that is, $T^{-1}AT = \hat{A}$ is symmetric (\hat{A} could be diagonal). The Crank–Nicholson method is

$$\left(I - \frac{k}{2} AD_0 \right) v^{n+1} = \left(I + \frac{k}{2} AD_0 \right) v^n,$$

or, equivalently, with $w = T^{-1}v$,

$$\left(I - \frac{k}{2} \hat{A} D_0 \right) w^{n+1} = \left(I + \frac{k}{2} \hat{A} D_0 \right) w^n.$$

Theorem 5.3.2 gives us

$$\|w^n\|_h \leq \|w^0\|_h, \quad (\beta = 0 \text{ here}). \quad (5.3.17)$$

Actually, there is equality, since

$$(w, \hat{A} D_0 w) = (\hat{A} w, D_0 w) = -(D_0 \hat{A} w, w) = -(w, \hat{A} D_0 w).$$

The final estimate is obtained from

$$\begin{aligned}\|v^n\|_h &= \|Tw^n\|_h \leq |T| \cdot \|w^n\|_h \leq |T| \cdot \|w^0\|_h, \\ &\leq |T| \cdot |T^{-1}| \cdot \|v^0\|_h \leq \text{constant} \|v^0\|_h.\end{aligned}\quad (5.3.18)$$

Indeed, the transformation $T^{-1}v$ can be considered as the definition of a new norm for v^n ,

$$\|v^n\|_h^* = \|T^{-1}v^n\|_h = (v^n, (T^{-1})^* T^{-1} v^n)_h^{1/2}. \quad (5.3.19)$$

The norm $\|\cdot\|_h^*$ is used to obtain the step-by-step estimate $\|v^{n+1}\|_h^* \leq \|v^n\|_h^*$, and the final estimate (5.3.18) then follows from the equivalence of the norms. This is a special case of the basic principle of the energy method: the construction of a special norm such that Eq. (5.3.1) is satisfied.

Now consider the fourth-order accurate difference operator

$$Q = AD_0 \left(I - \frac{h^2}{6} D_+ D_- \right), \quad (5.3.20)$$

derived in Section 3.1. Assuming that A is symmetric, we get, by using the identities (5.3.14),

$$\begin{aligned}(v, Qv)_h &= (v, AD_0 v)_h - \frac{h^2}{6} (v, AD_0 D_+ D_- v)_h, \\ &= 0 + \frac{h^2}{6} (D_- v, AD_0 D_+ v)_h = 0.\end{aligned}$$

This shows that also the operator (5.3.20) is skew-symmetric and the Crank–Nicholson method is again stable.

Scalar parabolic problems are often given in self-adjoint form

$$u_t = (au_x)_x, \quad (5.3.21)$$

where $a = a(x) \geq \delta > 0$. We consider the centered operator

$$Qv_j = D_+ a_{j-1/2} D_- v_j, \quad (5.3.22)$$

where $a_{j-1/2} = a(x_{j-1/2})$.

For smooth functions $a(x), u(x)$ we have

$$\begin{aligned}
 Qu(x_j) &= \frac{1}{h} [a(x_{j+1/2})D_+u(x_j) - a(x_{j-1/2})D_-u(x_j)] \\
 &= \frac{1}{h} \left(a(x_{j+1/2})[u_x(x_{j+1/2}) + \frac{h^2}{24} u_{xxx}(x_{j+1/2}) + \mathcal{O}(h^4)] \right. \\
 &\quad \left. - a(x_{j-1/2})[u_x(x_{j-1/2}) + \frac{h^2}{24} u_{xxx}(x_{j-1/2}) + \mathcal{O}(h^4)] \right).
 \end{aligned}$$

Furthermore,

$$a(x_{j\pm 1/2})u_{xxx}(x_{j\pm 1/2}) = a(x_j)u_{xxx}(x_j) + \mathcal{O}(h),$$

so

$$\begin{aligned}
 Qu(x_j) &= D_+a(x_{j-1/2})u_x(x_{j-1/2}) + \mathcal{O}(h^2) \\
 &= [(au_x)_x]_{x=x_j} + \mathcal{O}(h^2),
 \end{aligned}$$

and we see that Eq. (5.3.22) is a second-order accurate approximation in space.

We also have

$$(u, Qu)_h = (u, D_+a_{-1/2}D_-u)_h = -(D_-u, a_{-1/2}D_-u)_h \leq -\delta \|D_-u\|_h^2 \leq 0 \quad (5.3.23)$$

so Q is semibounded.

We now discuss a combination of the leap-frog scheme and the Crank–Nicholson method.

Theorem 5.3.4. Consider

$$(I - kQ_1(x_j, t_{n+1}))v_j^{n+1} = 2kQ_0(x_j, t_n)v_j^n + (I + kQ_1(x_j, t_{n-1}))v_j^{n-1}. \quad (5.3.24)$$

Assume that for all gridfunctions w

$$\operatorname{Re}(w, Q_1 w)_h \leq \alpha_1 \|w\|_h^2, \quad (5.3.25a)$$

$$\operatorname{Re}(w, Q_0 w)_h = 0, \quad (5.3.25b)$$

$$k\|Q_0\|_h \leq 1 - \delta, \quad \delta > 0. \quad (5.3.25c)$$

Then the method is **stable**.

Proof. We assume that Q_0, Q_1 are only functions of x . The proof of the general case is left as an exercise. We write Eq. (5.3.24) in the form

$$v_j^{n+1} - v_j^{n-1} = 2kQ_0v_j^n + kQ_1(v_j^{n+1} + v_j^{n-1}).$$

By Eq. (5.3.25a)

$$\begin{aligned} \|v^{n+1}\|_h^2 - \|v^{n-1}\|_h^2 &= 2k \operatorname{Re}(v^{n+1} + v^{n-1}, Q_0 v^n)_h \\ &\quad + k \operatorname{Re}(v^{n+1} + v^{n-1}, Q_1(v^{n+1} + v^{n-1}))_h \\ &\leq 2k \operatorname{Re}(v^{n+1}, Q_0 v^n)_h + 2k \operatorname{Re}(v^{n-1}, Q_0 v^n)_h \\ &\quad + \alpha_1 k \|v^{n+1} + v^{n-1}\|_h^2. \end{aligned} \quad (5.3.26)$$

For any gridfunctions u, v , Eq. (5.3.25b) implies

$$0 = \operatorname{Re}(u + v, Q_0(u + v))_h = \operatorname{Re}((v, Q_0 u)_h + (u, Q_0 v)_h).$$

Also,

$$\alpha_1 k \|v^{n+1} + v^{n-1}\|_h^2 \leq 2\tilde{\alpha}k(\|v^{n+1}\|_h + \|v^{n-1}\|_h), \quad \tilde{\alpha} = \frac{1}{2}(\alpha_1 + |\alpha_1|).$$

Therefore, we can write Eq. (5.3.26) in the form

$$L^{n+1} - 2\tilde{\alpha}k \|v^{n+1}\|_h^2 \leq L^n + 2\tilde{\alpha}k \|v^{n-1}\|_h^2, \quad (5.3.27)$$

where

$$L^n = \|v^n\|_h^2 + \|v^{n-1}\|_h^2 - 2k \operatorname{Re}(v^n, Q_0 v^{n-1})_h.$$

By Eq. (5.3.25c),

$$\begin{aligned} |2k(v^{n+1}, Q_0 v^n)_h| &\leq 2k \|Q_0\|_h \|v^{n+1}\|_h \|v^n\|_h, \\ &\leq (1 - \delta)(\|v^n\|_h^2 + \|v^{n+1}\|_h^2). \end{aligned}$$

Therefore,

$$\delta(\|v^n\|_h^2 + \|v^{n+1}\|_h^2) \leq L^{n+1} \leq 2(\|v^n\|_h^2 + \|v^{n+1}\|_h^2), \quad (5.3.28)$$

showing that $(L^{n+1})^{\frac{1}{2}}$ is equivalent to the norm $(\|v^{n+1}\|_h^2 + \|v^n\|_h^2)^{\frac{1}{2}}$. If $\alpha_1 \leq 0$, then $\tilde{\alpha} = 0$, the norm L^n does not increase, and we have

$$\|v^{n+1}\|_h^2 + \|v^n\|_h^2 \leq \frac{1}{\delta} L^{n+1} \leq \frac{1}{\delta} L^1 \leq \frac{2}{\delta} (\|v^1\|_h^2 + \|v^0\|_h^2).$$

Thus, the method is stable. If $\alpha_1 > 0$, then $\tilde{\alpha} = \alpha_1$, and Eqs. (5.3.27) and (5.3.28) imply that

$$\left(1 - 2\frac{\alpha_1}{\delta} k\right) L^{n+1} \leq \left(1 + \frac{2\alpha_1 k}{\delta}\right) L^n;$$

that is, for $\alpha_2 = \alpha_1/\delta$,

$$L^{n+1} \leq e^{4\alpha_2 k} L^n \leq e^{4\alpha_2 t_n} L^1. \quad (5.3.29)$$

This proves stability.

Let

$$u_t = Pu \quad (5.3.30)$$

be a system of differential equations and assume that P is semibounded, that is,

$$\operatorname{Re}(w, Pw) \leq \alpha_1 \|w\|^2. \quad (5.3.31)$$

Now assume that we have succeeded in constructing a difference operator such that

$$\operatorname{Re}(w, Qw)_h \leq \alpha_1 \|w\|_h^2.$$

Let Q^* be the adjoint operator, so that

$$(u, Qv)_h = (Q^* u, v)_h$$

for all gridfunctions u and v . By Eq. (5.3.14), Q^* is also a difference operator, and it approximates the adjoint differential operator P^* . We write

$$Q = Q_0 + Q_1, \quad Q_0 = \frac{1}{2}(Q - Q^*), \quad Q_1 = \frac{1}{2}(Q + Q^*).$$

Then

$$\operatorname{Re}(w, Q_0 w)_h = 0, \quad \operatorname{Re}(w, Q_1 w)_h = \operatorname{Re}(w, Qw)_h \leq \alpha_1 \|w\|_h^2,$$

and we can use the approximation (5.3.24).

The inequality (5.3.31) implies that the solutions of the differential equation

satisfy

$$\frac{d}{dt} \|u\|^2 \leq 2\alpha_1 \|u\|^2;$$

that is,

$$\|u(\cdot, t)\|^2 \leq e^{2\alpha_1 t} \|u(\cdot, 0)\|^2.$$

Therefore, the estimate (5.3.29) might not be satisfactory. However, we can introduce new variables $\tilde{u} = e^{-\alpha_1 t} u$ and obtain

$$\tilde{u}_t = (P - \alpha_1 I)\tilde{u} =: \tilde{P}\tilde{u},$$

and apply the method to the new problem. After we have written down the method, we can revert to the original variables.

As an example, we consider the convection-diffusion equation

$$u_t = \epsilon u_{xx} + a(x, t)u_x + c(x, t)u =: Pu,$$

$$\epsilon = \text{constant} > 0, \quad a, c \text{ real.}$$

(5.3.32)

We first determine the adjoint operator P^* . Integration by parts gives us

$$(v, Pu) = (\epsilon v_{xx} - (av)_x + cv, u),$$

that is,

$$\begin{aligned} P^*u &= \epsilon u_{xx} - (au)_x + cu, \\ \frac{1}{2}(P + P^*)u &= \epsilon u_{xx} + \frac{1}{2}(au_x - (au)_x) + cu \\ &= \epsilon u_{xx} + (c - \frac{1}{2}a_x)u. \\ \frac{1}{2}(P - P^*)u &= \frac{1}{2}((au)_x + au_x). \end{aligned}$$

As a difference approximation we choose

$$Q_0 v = \frac{1}{2} (D_0(av) + aD_0 v),$$

$$Q_1 v = \epsilon D_+ D_- v + (c - \frac{1}{2}a_x)v.$$

By Eq. (5.3.14),

$$\operatorname{Re}(v, Q_0 v)_h = 0,$$

$$\operatorname{Re}(v, Q_1 v)_h \leq -\varepsilon \|D_- v\|_h^2 + \max_x (c - \frac{1}{2} a_x) \|v\|_h^2.$$

Also, since $\|D_0\|_h = 1/h$, we have

$$\begin{aligned} k \|Q_0 v\|_h &\leq \frac{k}{2} \|D_0(av)\|_h + \frac{k}{2} \|aD_0 v\|_h \\ &\leq \frac{k}{2} \left(\frac{1}{h} \|av\|_h + \|a\|_\infty \|D_0 v\|_h \right) \leq \frac{k}{h} \|a\|_\infty \|v\|_h. \end{aligned}$$

By Theorem 5.3.4, the method is stable if $\frac{k}{h} \|a\|_\infty < 1 - \delta$.

REMARK. The theorem is also valid if we replace Eq. (5.3.25b) by

$$\operatorname{Re}(w, Q_0 w)_h = R(w), \quad (5.3.33)$$

where

$$|R(w)| \leq \text{constant } \|w\|_h^2.$$

In Section 2.2, we have applied the leap-frog scheme to

$$u_t = u_x - au, \quad a = \text{constant};$$

that is, in the framework of Theorem 5.3.4, we have used

$$Q_0 v = D_0 v - av, \quad Q_1 = 0.$$

In this case

$$\operatorname{Re}(w, Q_0 w)_h = -a \|w\|_h^2,$$

and Eq. (5.3.33) is satisfied. The method is stable, but a parasitic solution develops and grows exponentially, although the solution of the differential equation decays. Therefore, one must be careful if one replaces Eq. (5.3.25b) by Eq. (5.3.33).

EXERCISES

5.3.1. Prove Theorem 5.3.4 when Q_0, Q_1 depend on t .

5.3.2. Prove Theorem 5.3.4 with the condition (5.3.25b) replaced by Eq. (5.3.33).

5.3.3. Use the energy method to prove stability for the difference scheme

$$v_j^{n+1} = (I + ka_j D_+) v_j^n, \quad a_j > 0.$$

5.4. SPLITTING METHODS

Splitting methods are commonly used for time-dependent partial differential equations. They are often used to reduce problems in several space dimensions to a sequence of problems in one space dimension—this can significantly reduce the work required for implicit methods.

Consider, for example, the simplest form of the two-dimensional heat equation

$$u_t = u_{xx} + u_{yy}, \quad (5.4.1)$$

with periodic boundary conditions, and approximate it by the standard second-order Crank–Nicholson method

$$\left(I - \frac{k}{2} (D_{+x} D_{-x} + D_{+y} D_{-y}) \right) v^{n+1} = \left(I + \frac{k}{2} (D_{+x} D_{-x} + D_{+y} D_{-y}) \right) v^n. \quad (5.4.2)$$

To advance v^n one time step, we have to solve a linear system of equations in $\mathcal{O}(h^{-2})$ unknowns. Now replace Eq. (5.4.2) by

$$\begin{aligned} & \left(I - \frac{k}{2} D_{+x} D_{-x} \right) \left(I - \frac{k}{2} D_{+y} D_{-y} \right) v^{n+1} \\ &= \left(I + \frac{k}{2} D_{+x} D_{-x} \right) \left(I + \frac{k}{2} D_{+y} D_{-y} \right) v^n. \end{aligned} \quad (5.4.3)$$

Equation (5.4.3) is still second-order accurate because it differs from Eq. (5.4.2) by the third-order term

$$\frac{k^3}{4} D_{+x} D_{-x} D_{+y} D_{-y} \frac{v^{n+1} - v^n}{k}.$$

Now assume that v^n is known. We write Eq. (5.4.3) in the form

$$\begin{aligned} \left(I - \frac{k}{2} D_{+x} D_{-x} \right) z &= F, \\ F &:= \left(I + \frac{k}{2} D_{+x} D_{-x} \right) \left(I + \frac{k}{2} D_{+y} D_{-y} \right) v^n, \\ \left(I - \frac{k}{2} D_{+y} D_{-y} \right) v^{n+1} &= z. \end{aligned} \quad (5.4.4)$$

The first step is to solve for z . Because the equation contains only difference operators in the x direction, we can solve it for every fixed $y = y_v$. This is particularly simple, because the resulting linear system is essentially tridiagonal. Direct solution methods for this type of system were discussed in Section 2.3. It was shown that the solution is obtained in $\mathcal{O}(h^{-1})$ arithmetic operations.

Once one has determined z , one can determine v^{n+1} on every line $x = x_j$. Thus, instead of solving a linear system in $\mathcal{O}(h^{-2})$ unknowns, we can determine v^{n+1} by solving $\mathcal{O}(h^{-1})$ systems in $\mathcal{O}(h^{-1})$ unknowns. This procedure requires $\mathcal{O}(h^{-2})$ arithmetic operations, which is generally cheaper. The gain is more pronounced for more general equations with variable coefficients where specially designed methods for the constant coefficient system (5.4.2) do not apply.

We next consider general types of splittings for one-step methods. Assume that the differential equation has the form

$$u_t = (P_1 + P_2) u, \quad (5.4.5)$$

where P_1, P_2 are linear differential operators in space. Let Q_1, Q_2 be approximate solvers for each part, that is,

$$v^{n+1} = Q_1 v^n \quad (5.4.6)$$

is an approximation of

$$v_t = P_1 v, \quad (5.4.7)$$

and

$$w^{n+1} = Q_2 w^n \quad (5.4.8)$$

is an approximation of

$$w_t = P_2 w. \quad (5.4.9)$$

We assume that Q_1, Q_2 are simple in the sense that both Eq. (5.4.6) and Eq. (5.4.8) together are much easier to compute (or to construct) than any direct solver of (5.4.5). One typical case is when Q_1 and Q_2 are one-dimensional but operate in different coordinate directions. If each of them is at least first-order accurate in time, then the approximation

$$u^{n+1} = Q_2 Q_1 u^n \quad (5.4.10)$$

is also first-order accurate. This follows from the fact that, for smooth functions u ,

$$Q_j u = (I + kP_j)u + \mathcal{O}(k^2), \quad j = 1, 2,$$

and, therefore,

$$Q_2 Q_1 u = (I + kP_1 + kP_2)u + \mathcal{O}(k^2).$$

If Q_1 and Q_2 satisfy $\|Q_j\|_h \leq 1 + \mathcal{O}(k)$, $j = 1, 2$, then obviously $\|Q_2 Q_1\|_h \leq 1 + \mathcal{O}(k)$. Under the general stability conditions on Q_1 and Q_2 in Definition 5.1.1, the stability of Eq. (5.4.10) does not necessarily follow.

We can also construct second-order accurate splittings. Assume that Q_1 and Q_2 are accurate of order $(p, 2)$ when applied to smooth solutions of Eqs. (5.4.7) and (5.4.9), respectively. Then

$$Q_j = Q_j(k, t) = I + k \frac{\partial}{\partial t} + \frac{k^2}{2} \frac{\partial^2}{\partial t^2} + \mathcal{O}(k^3 + kh^p), \quad j = 1, 2.$$

Since $v_{tt} = (P_1 v)_t = P_1 v_t + P_{1,t} v = (P_1^2 + P_{1,t})v$, and similarly for w_{tt} , we get

$$Q_j(k, t) = I + kP_j + \frac{k^2}{2} (P_j^2 + P_{j,t}) + \mathcal{O}(k^3 + kh^p), \quad j = 1, 2. \quad (5.4.11)$$

[If P_j has the form $A(x, t)\partial/\partial x$, then $P_{j,t}$ denotes the operator $(\partial A/\partial t)(\partial/\partial x)$.]

The second-order splitting is

$$u^{n+1} = Q u^n := Q_1 \left(\frac{k}{2}, t_{n+1/2} \right) Q_2(k, t_n) Q_1 \left(\frac{k}{2}, t_n \right) u^n. \quad (5.4.12)$$

By using the relations (5.4.11), we get

$$\begin{aligned}
Q = & I + k \left[\frac{1}{2} P_1(t_n) + \frac{1}{2} P_1(t_{n+1/2}) + P_2(t_n) \right] \\
& + \frac{k^2}{2} \left[\frac{1}{4} P_1^2(t_n) + \frac{1}{4} P_1^2(t_{n+1/2}) \right. \\
& + \frac{1}{2} P_1(t_{n+1/2})P_1(t_n) + P_2(t_n)P_1(t_n) \\
& + P_1(t_{n+1/2})P_2(t_n) + P_2^2(t_n) + \frac{1}{4} P_{1,t}(t_n) + \frac{1}{4} P_{1,t}(t_{n+1/2}) + P_{2,t}(t_n) \Big] \\
& + \mathcal{O}(k^3 + kh^p).
\end{aligned}$$

If we expand P_1 and $P_{1,t}$ around $t = t_n$ using Taylor series, we get

$$\begin{aligned}
Q = & I + k(P_1 + P_2) + \frac{k^2}{2} (P_1^2 + P_1P_2 + P_2P_1 + P_2^2 + P_{1,t} + P_{2,t}) \\
& + \mathcal{O}(k^3 + kh^p).
\end{aligned}$$

But this is the unique form of a $(p, 2)$ -order accurate one-step operator applied to a smooth solution of Eq. (5.4.5).

We summarize the result in the following theorem.

Theorem 5.4.1. Assume that Eqs. (5.4.6) and (5.4.8) are approximations of order (p, q) , $q \geq 1$, to Eqs. (5.4.7) and (5.4.9), respectively. Then Eq. (5.4.10) is an approximation of order $(p, 1)$. If $q \geq 2$, then Eq. (5.4.12) is an approximation of order $(p, 2)$.

In the second-order case, stability follows if $\|Q_1(k/2, t)\| \leq 1 + \mathcal{O}(k)$, $\|Q_2(k, t)\| \leq 1 + \mathcal{O}(k)$ for all t . If these relations do not hold, the stability of Eq. (5.4.12) must be verified directly.

When the operator Q in Eq. (5.4.12) is applied repeatedly, $Q_1(k/2, t_n) \cdot Q_1(k/2, t_{n-1/2})$ occurs in each step. By using Taylor expansions of P and Eq. (5.4.11), we get

$$Q_1\left(\frac{k}{2}, t_n\right) Q_1\left(\frac{k}{2}, t_{n-1/2}\right) = Q_1(k, t_{n-1/2}) + \mathcal{O}(k^3 + kh^p),$$

showing that the method

$$u^n = Q_1\left(\frac{k}{2}, t_{n-1/2}\right) Q_2(k, t_{n-1}) Q_1(k, t_{n-3/2}) \cdots Q_2(k, 0) Q_1\left(\frac{k}{2}, 0\right) u^0$$

is an approximation of order $(p, 2)$. Note, however, that each time a printout is required, a half-step with the operator Q_1 must be taken.

The splitting procedure described here can also be applied to nonlinear problems, and the arguments leading to second-order accuracy hold. This is useful, for example, when solving systems of conservation laws

$$u_t = F_x(u) + G_y(u),$$

where the operators Q_1 and Q_2 represent one-dimensional solvers.

The splitting method described above can be generalized to problems in more than two space dimensions. Assume that the problem

$$u_t = \sum_{j=1}^d P_j u$$

has time-independent coefficients and that

$$v^{n+1} = Q_j v^n$$

solves

$$u_t = P_j u, \quad j = 1, 2, \dots, d,$$

with at least first-order accuracy in time. Then the approximation

$$v^{n+1} = Q_d Q_{d-1} \cdots Q_1 v^n$$

is first-order accurate in time.

The second-order version does not generalize in a straightforward way for $d > 2$.

We emphasize that, as always when discussing the formal order of accuracy, it is assumed that the solutions are sufficiently smooth. The accuracy of splitting methods used for problems with discontinuous solutions is not well understood.

In the explicit case, the splitting methods are usually as expensive as the original ones when counting the number of arithmetic operations. Still, there may be a gain in the simplification of the stability analysis. For example, it is easy to find the stability limits on k for the one-dimensional Lax–Wendroff operators such that

$$\|I + kD_{0x} + \frac{k^2}{2} D_{+x}D_{-x}\|_h \leq 1,$$

$$\|I + kD_{0y} + \frac{k^2}{2} D_{+y}D_{-y}\|_h \leq 1.$$

It is more difficult to find the stability limit for the two-dimensional Lax–Wendroff type approximation for $u_t = u_x + u_y$

$$v^{n+1} = \left(I + k(D_{0x} + D_{0y}) + \frac{k^2}{2} (D_{+x}D_{-x} + 2D_{0x}D_{0y} + D_{+y}D_{-y}) \right) v^n. \quad (5.4.13)$$

The stability of Lax–Wendroff type methods will be discussed in Section 6.5. Furthermore, the implementation of factored schemes is easier for large-scale problems. For example, if each factor is one-dimensional, then, in each step, we operate on the data in one space dimension only.

With

$$Q_1^{(1)}(k) = I + kD_{+y}D_{-y},$$

$$Q_2(k) = \left(I - \frac{k}{2} D_{+x}D_{-x} \right)^{-1} \left(I + \frac{k}{2} D_{+x}D_{-x} \right),$$

$$Q_1^{(2)}(k) = I - kD_{+y}D_{-y},$$

the approximation (5.4.3) is

$$v^{n+1} = Q_1^{(2)}\left(\frac{k}{2}\right) Q_2(k) Q_1^{(1)}\left(\frac{k}{2}\right) v^n.$$

This is called an ***Alternating Direction Implicit (ADI)*** method, which is a special factored form different from Eq. (5.4.12). It is also second-order accurate as shown above, and it is, furthermore, unconditionally stable (Exercise 5.4.2).

EXERCISES

- 5.4.1. Find the stability limit on $\lambda = k/h$ for the approximation (5.4.13) with equal step size $h_x = h_y = h$.
- 5.4.2. Prove that the ADI-scheme (5.4.3) is unconditionally stable.

5.4.3. Construct a second-order accurate ADI approximation of type (5.4.3) of

$$u_t = (a(x, y)u_x)_x + (b(x, y)u_y)_y.$$

Use the energy method to prove that it is unconditionally stable.

5.5. STABILITY FOR NONLINEAR PROBLEMS

The stability theory presented so far holds for linear problems. Because most problems in applications are nonlinear, one might think that the linear theory is of no use for real problems. However, if the solution u of the nonlinear differential equation is sufficiently smooth, then to first approximation the error equation can be linearized about the solution u , and convergence follows if the linearized error equation is stable.

We start with a simple example. Consider Burger's equation

$$u_t + uu_x = \nu u_{xx}, \quad \nu = \text{constant} > 0, \quad (5.5.1)$$

with 2π -periodic initial data

$$u(x, 0) = f(x), \quad f(x) \in C^\infty. \quad (5.5.2)$$

One can then show the following.

Theorem 5.5.1. *Equations (5.5.1) and (5.5.2) have a solution that belongs to C^∞ for all t . Also, one can estimate the derivatives: for every i, j there is a constant C_{ij} such that*

$$\max_{x, t} \left| \frac{\partial^{i+j} u}{\partial x^i \partial t^j} \right| \leq C_{ij} \nu^{-(i+j)}. \quad (5.5.3)$$

This estimate cannot be improved. We approximate the above problem by the Crank–Nicholson method

$$\begin{aligned} \left(I - \frac{k}{2} Q(v^{n+1}) \right) v^{n+1} &= \left(I + \frac{k}{2} Q(v^n) \right) v^n, \\ Q(v^n) &= -v^n D_0 + \nu D_+ D_-, \quad v^0 = f, \end{aligned} \quad (5.5.4)$$

and assume that $k = \lambda h$, $\lambda > 0$ constant.

The error

$$h^2 e = v - u$$

satisfies

$$\begin{aligned} \left(I - \frac{k}{2} (Q_0(u^{n+1}) - h^2 e^{n+1} D_0) \right) e^{n+1} &= \left(I + \frac{k}{2} (Q_0(u^n) - h^2 e^n D_0) \right) e^n \\ &\quad + k R^{n+1/2}(u), \\ e^0 &= 0, \end{aligned} \tag{5.5.5}$$

where

$$Q_0(u^n) = Q(u^n) - (D_0 u^n),$$

and

$$\begin{aligned} kh^2 R^{n+1/2}(u) &= - \left(I - \frac{k}{2} Q(u^{n+1}) \right) u^{n+1} + \left(I + \frac{k}{2} Q(u^n) \right) u^n, \\ &= kh^2(R_2^{n+1/2} + h^2 R_4^{n+1/2} + h^4 R_6^{n+1/2} + \dots), \end{aligned} \tag{5.5.6}$$

is the truncation error. Thus, the R_{2j} are expressions in u and its derivatives of order up to $2j + 2$. If we neglect the nonlinear terms $kh^2 e^{n+1} D_0 e^{n+1}$ and $kh^2 e^n D_0 e^n$, then Eq. (5.5.5) is a stable linear difference scheme. Therefore, we obtain, in any finite time interval $0 \leq t \leq T$,

$$\|u^n - v^n\|_h = h^2 \|e^n\|_h \leq \text{constant } h^2 \max_{0 \leq j \leq n-1} \|R^{j+1/2}\|_h. \tag{5.5.7}$$

Observe, however, that this estimate is only useful if $h^2 \|R^{j+1/2}\|_h \ll 1$. If $\nu \ll 1$, then by Eq. (5.5.3) we can only guarantee this if $h \ll \nu$.

We can improve the estimate. Formally, Eq. (5.5.5) converges as $h \rightarrow 0$, $k/h = \lambda = \text{constant} > 0$, to the differential equation

$$\begin{aligned} w_{1t} + uw_{1x} + u_x w_1 &= \nu w_{1xx} + R(u), \\ w_1(x, 0) &= 0. \end{aligned} \tag{5.5.8}$$

Thus,

$$\begin{aligned} \left(I - \frac{k}{2} Q_0(u^{n+1}) \right) w_1^{n+1} &= \left(I + \frac{k}{2} Q_0(u^n) \right) w_1^n + kR^{n+1/2} + kh^2 S^{n+1/2}, \\ w_1^0 &= 0, \end{aligned} \quad (5.5.9)$$

where

$$kh^2 S^{n+1/2} = kh^2(S_2^{n+1/2} + h^2 S_4^{n+1/2} + h^4 S_6^{n+1/2} + \dots)$$

is the truncation error. The $S_{2j}^{n+1/2}$ are expressions in w_1 and its derivatives of order up to $2j+2$.

We expect that $e - w_1 = \mathcal{O}(h^2)$. Therefore, we make the ansatz

$$e = w_1 + h^2 e_1. \quad (5.5.10)$$

Substituting Eq. (5.5.10) into Eq. (5.5.5) and using Eq. (5.5.9) gives us

$$\begin{aligned} \left(I - \frac{k}{2} (Q_1(u^{n+1}) + h^4 e_1^{n+1} D_0) \right) e_1^{n+1} &= \left(I + \frac{k}{2} (Q_1(u^n) + h^4 e_1^n D_0) \right) e_1^n \\ &\quad + k(R^{(1)})^{n+1/2}, \\ e_1^0 &= 0, \end{aligned} \quad (5.5.11)$$

where

$$Q_1(u^n) = Q_0(u^n) - h^2 w_1^n D_0,$$

and

$$\begin{aligned} (R^{(1)})^{n+1/2} &= -S^{n+1/2} - w_1^{n+1} (D_0 w_1^{n+1}) + w_1^n (D_0 w_1^n), \\ &= \bar{R}_2^{n+1/2} + h^2 \bar{R}_4^{n+1/2} + h^4 \bar{R}_6^{n+1/2} + \dots. \end{aligned}$$

If we neglect the terms $kh^4 e_1^{n+1} D_0 e_1^{n+1}$ and $kh^4 e_1^n D_0 e_1^n$, then Eq. (5.5.11) is a stable linear difference scheme, and, in every finite time interval $0 \leq t \leq T$, we obtain a bound

$$\|e_1^n\|_h \leq \text{constant} \max_{0 \leq j \leq n-1} \|(R^{(1)})^{j+1/2}\|_h.$$

Thus,

$$v = u + h^2 e = u + h^2 w_1 + h^4 e_1,$$

provided we could neglect the nonlinear terms; that is,

$$u = v - h^2 w_1 - h^4 e_1 = v - h^2 w_1 + \mathcal{O}(h^4). \quad (5.5.12)$$

This process can be continued. After p steps, e_p is the solution of

$$\begin{aligned} & \left(I - \frac{k}{2} (Q_p(u^{n+1}) + h^{2p+2} e_p^{n+1} D_0) \right) e_p^{n+1} \\ &= \left(I + \frac{k}{2} (Q_p(u^n) + h^{2p+2} e_p^n D_0) \right) e_p^n \\ & \quad + k(R^{(p)})^{n+1/2}, \\ e_p^0 &= 0. \end{aligned} \quad (5.5.13)$$

Here

$$Q_p(u^n) = Q_{p-1}(u^n) + h^{2p} w_p^n D_0,$$

where the w_p are solutions of linear differential equations of type (5.5.8). Also,

$$u = v - \sum_{j=1}^p h^{2j} w_j - h^{2p+2} e_p, \quad (5.5.14)$$

and e_p is bounded, provided we can neglect the nonlinear terms in Eq. (5.5.13).

We shall now use Appendix A.2 to bound the solution of the nonlinear equation (5.5.13). We consider a fixed time interval $0 \leq t \leq T = Nk$ and write Eq. (5.5.13) in the form

$$A\mathbf{e} + \varepsilon \mathbf{f}(\mathbf{e}) = \mathbf{b}, \quad \varepsilon = h^{2p+2}.$$

Here

$$\begin{aligned}\mathbf{e} &= (e_p^N, e_p^{N-1}, \dots, e_p^1), \\ \mathbf{b} &= ((R^{(p)})^{N-1/2}, (R^{(p)})^{N-3/2}, \dots, (R^{(p)})^{1/2}), \\ \mathbf{f}(\mathbf{e}) &= -\frac{1}{2} (e_p^N D_0 e_p^N, e_p^{N-1} D_0 e_p^{N-1}, \dots, e_p^1 D_0 e_p^1) \\ &\quad + \frac{1}{2} (e_p^{N-1} D_0 e_p^{N-1}, e_p^{N-2} D_0 e_p^{N-2}, \dots, e_p^0 D_0 e_p^0),\end{aligned}$$

and A denotes the linear part of the equation. We use the norm

$$|\mathbf{e}|^2 = \max_n \|e^n\|_h^2.$$

Let

$$|\mathbf{e}|_\infty = \max_{n,j} |e^n(x_j)|.$$

Then

$$\|e^n\|_h^2 = \sum_j |e^n(x_j)|^2 h$$

implies that

$$|\mathbf{e}|_\infty \leq h^{-1/2} |\mathbf{e}|.$$

Stability of the linear part of the equation tells us that

$$|A^{-1}| \leq K.$$

For the nonlinear term we obtain

$$\begin{aligned}
|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})|^2 &\leq \max_n \|u^n D_0 u^n - v^n D_0 v^n\|_h^2 \\
&= \max_n \sum_j |u^n(x_j) D_0 u^n(x_j) - v^n(x_j) D_0 v^n(x_j)|^2 h \\
&= \max_n \sum_j |(u^n(x_j) - v^n(x_j)) D_0 u^n(x_j) + v^n(x_j) D_0(u^n(x_j) \\
&\quad - v^n(x_j))|^2 h \\
&\leq \text{constant } h^{-2} (|\mathbf{u}|_\infty^2 + |\mathbf{v}|_\infty^2) \max_n \|u^n - v^n\|_h^2 \\
&\leq \text{constant } h^{-3} (|\mathbf{u}|^2 + |\mathbf{v}|^2) |\mathbf{u} - \mathbf{v}|^2,
\end{aligned}$$

or

$$|\mathbf{f}(\mathbf{u}) - \mathbf{f}(\mathbf{v})| \leq \text{constant } h^{-3/2} \sqrt{|\mathbf{u}|^2 + |\mathbf{v}|^2} |\mathbf{u} - \mathbf{v}|. \quad (5.5.15)$$

Thus, referring to Theorem A.2.2 of Appendix A.2, \mathbf{f} is locally Lipschitz continuous, and the Lipschitz constant satisfies

$$L(\mathbf{e}^0, 1) \leq \text{constant } h^{-3/2},$$

that is,

$$\tau := \varepsilon |A^{-1}| L \leq \text{constant } h^{2p+1/2}.$$

Also, since $\mathbf{f}(0) = 0$ by Eq. (5.5.15), we get

$$|\mathbf{f}(\mathbf{e}^0)| \leq \text{constant } h^{-3/2} |\mathbf{e}^0|^2,$$

showing that the assumptions of Theorem A.2.2 are fulfilled for h sufficiently small. Thus, Eq. (5.5.13) has a solution e_p^n with

$$\max_n \|e_p^n - (e_p^n)^0\|_h \leq \text{constant } h^{2p+1/2}.$$

Clearly, the solution of Eq. (5.5.13) is locally unique. The initial guess $(e_p^n)^0$ is the solution of the stable linear problem, and it can be estimated in terms of $\max_n \|(R^{(p)})^{n+1/2}\|_h$. Therefore, the solution e_p^n of the nonlinear equation is bounded if h is sufficiently small.

One can use the asymptotic expansion Eq. (5.5.14) in two ways to improve the error bound.

1. *Richardson extrapolation.* Let us calculate the solution of Eq. (5.5.4) for two stepsizes h and $2h$ and denote the solutions v_h and v_{2h} . By Eq. (5.5.14), we have

$$\begin{aligned} u &= v_h + \sum_{j=1}^p h^{2j} w_j + \mathcal{O}(h^{2p+2}), \\ u &= v_{2h} + \sum_{j=1}^p (2h)^{2j} w_j + \mathcal{O}(h^{2p+2}). \end{aligned}$$

Observe that the w_j are solutions of the linear differential equations and do not depend on h . Therefore,

$$\begin{aligned} 3u &= 4v_h - v_{2h} + \sum_{j=2}^p (4 - 2^{2j}) h^{2j} w_j + \mathcal{O}(h^{2p+2}), \\ &= 4v_h - v_{2h} + \mathcal{O}(h^4), \end{aligned}$$

and we obtain a fourth-order accurate approximation $(4v_h - v_{2h})/3$ of u . Of course, one can use the above process to calculate even higher order approximations. For example, if we calculate v_h, v_{2h}, v_{3h} , then we can form a linear combination that gives us a sixth-order accurate result. The above procedure can also be used to calculate an approximation of the error. We have as a first approximation

$$u - v_h = h^2 w_1 + \mathcal{O}(h^4) = \frac{1}{3} (v_h - v_{2h}) + \mathcal{O}(h^4).$$

2. *Deferred correction.* In this case we first calculate v from Eq. (5.5.4). To compute an approximation of the error $h^2 e$, we replace u by v in Eq. (5.5.5) and neglect the nonlinear terms:

$$\begin{aligned} \left(I - \frac{k}{2} Q_0(v^{n+1}) \right) \tilde{e}^{n+1} &= \left(I + \frac{k}{2} Q_0(v^n) \right) \tilde{e}^n + k R^{n+1/2}(v), \\ \tilde{e}^0 &= 0. \end{aligned}$$

Using Eq. (5.5.14), we commit an error of order $\mathcal{O}(h^2)$ in this process, and we get

$$u = v + h^2 \tilde{e} + \mathcal{O}(h^4).$$

Again, one can generalize this procedure to obtain higher order approximations.

REMARK. Richardson extrapolation and deferred correction are general procedures that can be used to decrease the error for all kinds of problems whenever an asymptotic expansion exists.

We have shown that, to first approximation, the error is given by

$$u^n - v^n = h^2 w_1^n,$$

where w_1 is the solution of Eq. (5.5.8). This estimate is only useful if $h^2 \|w_1^n\|_h \ll 1$. Unfortunately, this often poses a problem. R is the truncation error and it is composed of u and its derivatives. The leading terms are proportional to u_{ttt} , u_{xxx} , νu_{xxxx} . The best estimate we can obtain using Eq. (5.5.3) is

$$|R| \leq \text{constant } \nu^{-3}.$$

Therefore, we need $h^2 \nu^{-3} \ll 1$, which can be a severe restriction on the stepsize.

If this condition is not satisfied, not only is the accuracy poor, but the approximation may blow up faster than exponentially. This is often referred to as nonlinear instability; however, it occurs in linear problems with nonsmooth coefficients as well.

The other difficulty is associated with the solution operator $S(t, t_0)$ of the differential equation (5.5.8). If $\|S(t, t_0)\|$ is exponentially increasing, then one has to make h exponentially small if one wants to solve the problem over long time intervals. In fact, the problem can be worse if the solution operator $S_h(t, t_0)$ of the linearized difference approximation grows faster than $S(t, t_0)$.

It is not difficult to generalize these results. Consider a quasilinear system

$$\begin{aligned} u_t &= P \left(x, t, u, \frac{\partial}{\partial x} \right) u, \\ u(x, 0) &= f, \end{aligned} \tag{5.5.16}$$

of order p and approximate it by a one-step method

$$\begin{aligned} v^{n+1} &= (I + kQ(x, t_n, v^n))v^n, \\ v^0 &= f. \end{aligned} \tag{5.5.17}$$

Assume that $k/h^p = \lambda$ is a constant and that the approximation is accurate of order q . Assume also that Eq. (5.5.16) has a smooth solution U . Substituting U into Eq. (5.5.17) gives us

$$U^{n+1} = (I + kQ(x, t_n, U^n))U^n + kh^q R^n. \tag{5.5.18}$$

We expect that $U - v = \mathcal{O}(h^q)$, and therefore make the ansatz

$$v = U + h^q e. \quad (5.5.19)$$

Substituting Eq. (5.5.19) into Eq. (5.5.17) gives us

$$\begin{aligned} e^{n+1} &= (I + kQ_{11}(x, t_n, U^n))e^n + kh^q Q_{12}(x, t_n, U^n, e^n)e^n + kR^n, \\ e^0 &= 0. \end{aligned} \quad (5.5.20)$$

Q_{11} is the difference operator which we obtain by linearizing $Q(x, t_n, v^n)v^n$ about U . Q_{12} is the remaining nonlinear part. If we neglect the nonlinear terms, then Eq. (5.5.20) converges formally to a linear differential equation

$$\begin{aligned} w_{1t} &= P_1(x, t, U)w_1 + R, \\ w_1(x, 0) &= 0. \end{aligned} \quad (5.5.21)$$

Here P_1 is the linearization of $P(x, t, u)u$ about U . Assume that Eq. (5.5.21) has a smooth solution [i.e., Eq. (5.5.21) is a well-posed problem], and that

$$\begin{aligned} w_1^{n+1} &= (I + kQ_{11})w_1^n + kR + kh^{q_1} S^n, \\ w_1^0 &= 0, \end{aligned}$$

is accurate of order q_1 . Then we define e_1 by

$$h^{q_1} e_1 = e - w_1.$$

For e_1 , we obtain the equation

$$\begin{aligned} e_1^{n+1} &= (I + Q_{21}(x, t_n, U^n, h^q w_1^n))e_1^n \\ &\quad + kh^{q+q_1} Q_{22}(x, t_n, U^n, h^q w_1^n, e_1^n)e_1^n + k(R^{(1)})^n, \\ e_1^0 &= 0. \end{aligned}$$

Now we can repeat this process. Thus, we can reduce the nonlinearity to higher and higher order in h . We can apply Theorem A.2.2 of Appendix A.2, provided the linearized difference approximation is stable. In particular,

$$v - U = h^q w_1 + \mathcal{O}(h^{q+q_1}).$$

Here w_1 is the solution of Eq. (5.5.21).

EXERCISES

5.5.1. Prove that the L_2 norm of the solutions of Eq. (5.5.1) is a nonincreasing function of t .

5.5.2. Consider the approximation (5.5.4) with $Q(v)v$ replaced by

$$Q(v)v := -\frac{1}{3} (vD_0v + D_0v^2) + \nu D_+ D_-.$$

Use the energy method to prove that the solutions fulfill

$$\|v^n\|_h \leq \|v^0\|_h.$$

5.5.3. Determine the coefficients α_ν in the general formula for Richardson extrapolation

$$u = \sum_{\nu=0}^p \alpha_\nu v_{2^\nu h} + \mathcal{O}(h^{2(p+1)}).$$

BIBLIOGRAPHIC NOTES

The Kreiss Matrix Theorem was given by Kreiss (1962); the proof is also found in Richtmyer and Morton (1967). There are several constants occurring in the theorem, and it may be useful to know the relations between them. In particular, if the constant C_2 in Eq. (5.2.27) is known, one would like to know the constant C_1 in Eq. (5.2.26). It has been proved by Spijker (1991) [see also LeVeque and Trefethen (1984)] that the best possible value is $C_1 = e \cdot m \cdot C_2$ and that C_1 and C_2 grow at most linearly in the dimension of the matrices [see Tadmor (1981)]. For further results concerning these constants, see McCarthy and Schwartz (1965), McCarthy (1994), and van Dorsselaer, Kraaijevanger, and Spijker (1993).

The general splitting procedure described in Section 5.4 is due to Strang (1968).

Alternating direction implicit methods of type (5.4.3) were originally developed by Peaceman and Rachford (1955) and Douglas (1955). Later generalizations have been given by Douglas (1962) and Douglas and Gunn (1964). See also Beam and Warming (1978).

The type of convergence proof technique used for nonlinear problems in Section 5.5 was first used by Strang (1964), who treated quasilinear hyperbolic terms.

Richardson extrapolation was introduced by Richardson (1927). Richardson extrapolation and deferred correction are discussed in Fox (1957) and Pereyra (1968).

Stability difficulties resulting from nonsmooth coefficients and with nonlinear problems have been discussed by Phillips (1959), Kreiss and Oliger (1972), and Fornberg (1975).

6

HYPERBOLIC EQUATIONS AND NUMERICAL METHODS

6.1. SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

We have already discussed first-order systems

$$u_t = Au_x \quad (6.1.1)$$

in Section 4.3. Here we only add some complementary results and an example.

In Section 4.5, we have proved that for general systems with constant coefficients we could obtain an energy estimate

$$\frac{d}{dt} (u, u)_H \leq 2\alpha(u, u)_H$$

if and only if the initial value problem is well posed. We now explicitly construct the H norm.

Lemma 6.1.1. *Let A be a matrix with real eigenvalues and a complete set of eigenvectors that are the columns of a matrix T . Let D be a real positive diagonal matrix. Then*

$$H = (T^{-1})^*DT^{-1}$$

is positive definite, and

$$B = HA$$

is Hermitian. H is called a symmetrizer of A .

Proof. The matrix H is Hermitian, and it is positive definite. By assumption, $T^{-1}AT = T^*A^*(T^{-1})^*$ is a real diagonal matrix and so is $DT^{-1}AT$. Therefore,

$$B - B^* = HA - A^*H = (T^{-1})^*(DT^{-1}AT - T^*A^*(T^{-1})^*D)T^{-1} = 0,$$

which proves the lemma.

Now consider a strongly hyperbolic system such as Eq. (6.1.1).

Theorem 6.1.1. *Let H be defined as in Lemma 6.1.1. Then the solutions of the strongly hyperbolic system (6.1.1) satisfy*

$$(u(\cdot, t), Hu(\cdot, t)) = (u(\cdot, 0), Hu(\cdot, 0)). \quad (6.1.2)$$

Proof. HA is Hermitian and, therefore, by Lemma 4.6.1,

$$\begin{aligned} \frac{d}{dt} (u, Hu) &= (u_t, Hu) + (u, Hu_t), \\ &= (Au_x, Hu) + (u, HAU_x), \\ &= -(u, A^*Hu_x) + (u, HAU_x), \\ &= (u, (HA - A^*H^*)u_x), \\ &= 0. \end{aligned}$$

The theorem shows how the symmetrizer H is used to construct a new norm such that the solution is nonincreasing in that norm.

As an example, we consider the Euler equations (4.6.4a) and (4.6.4b). If the linearization is made around a constant state, the lower order term in the linearized system vanishes. Considering perturbations that are independent of y and z , we arrive at a one-dimensional system with constant coefficients:

$$\begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & 0 & 0 & a^2/R \\ 0 & U & 0 & 0 \\ 0 & 0 & U & 0 \\ R & 0 & 0 & U \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_x = 0, \quad (6.1.3)$$

where a is the speed of sound. (We have dropped the prime sign here.) Thus, the system reduces to a 2×2 system and two scalar equations

$$\begin{bmatrix} u \\ \rho \end{bmatrix}_t + A \begin{bmatrix} u \\ \rho \end{bmatrix}_x = 0, \quad A = \begin{bmatrix} U & a^2/R \\ R & U \end{bmatrix}, \quad (6.1.4a)$$

$$v_t + Uv_x = 0, \quad (6.1.4b)$$

$$w_t + Uw_x = 0. \quad (6.1.4c)$$

Obviously, Eqs. (6.1.4b) and (6.1.4c) are hyperbolic. The eigenvalues of A are

$$\kappa = U \pm a,$$

showing that the system (6.1.4a) is strictly hyperbolic under the natural assumption $a > 0$. [The system (6.1.3) is not strictly hyperbolic, because the coefficient matrix has two eigenvalues U . However, this has no real significance, because the system decouples.]

The eigenvectors of A are $(1, \pm R/a)^T$ corresponding to the eigenvalues $U \pm a$. The matrix T^{-1} in Lemma 6.1.1 is, therefore,

$$T^{-1} = \frac{a}{2R} \begin{bmatrix} R/a & 1 \\ R/a & -1 \end{bmatrix}.$$

For any diagonal matrix $D = \text{diag}(d_1, d_2)$ we have

$$H := (T^{-1})^* D T^{-1} = \frac{1}{4} \begin{bmatrix} d_1 + d_2 & \frac{a}{R} (d_1 - d_2) \\ \frac{a}{R} (d_1 - d_2) & \frac{a^2}{R^2} (d_1 + d_2) \end{bmatrix},$$

and, with $d_1 = d_2 = 2$, we get the symmetrizer

$$H = \begin{bmatrix} 1 & 0 \\ 0 & a^2/R^2 \end{bmatrix}.$$

With $\tilde{\mathbf{u}} = (u, a\rho/R)^T$ we get, using Theorem 6.1.1,

$$\|\tilde{\mathbf{u}}(\cdot, t)\|^2 = \|\tilde{\mathbf{u}}(\cdot, 0)\|^2; \quad (6.1.5)$$

that is, the introduction of the new norm $(\mathbf{u}, H\mathbf{u})$ can be interpreted as a single rescaling of the dependent variables $\mathbf{u} = (u, \rho)^T$.

This transformation can of course be applied directly to the system (6.1.4a). Let $H^{1/2} = \text{diag}(1, a/R)$. Then

$$\tilde{\mathbf{u}} = H^{1/2}\mathbf{u},$$

and Eq. (6.1.4a) takes the form

$$\tilde{\mathbf{u}}_t + H^{1/2}A(H^{1/2})^{-1}\tilde{\mathbf{u}}_x = 0,$$

where

$$H^{1/2}A(H^{1/2})^{-1} = \begin{bmatrix} U & a \\ a & U \end{bmatrix}.$$

This is a symmetric system and Eq. (6.1.5) follows immediately using the energy method.

Finally, we note that if $U = 0$ and $\tilde{\rho} = a\rho/R$, then

$$\begin{aligned} u_{tt} &= -a\tilde{\rho}_{xt} = a^2 u_{xx}, \\ \tilde{\rho}_{tt} &= -au_{xt} = a^2 \tilde{\rho}_{xx}; \end{aligned} \quad (6.1.6)$$

that is, both u and $\tilde{\rho}$ satisfy the wave equation.

EXERCISES

- 6.1.1.** Derive the symmetrizer H for the system (6.1.3) without using the decoupled form.
- 6.1.2.** Derive the symmetrizer H of

$$A = \begin{bmatrix} 1 & a & 0 \\ b & 1 & a \\ 0 & b & 1 \end{bmatrix}.$$

What is the condition on a, b for H to have the desired properties?

6.2. SYSTEMS WITH VARIABLE COEFFICIENTS IN ONE SPACE DIMENSION

We now consider systems

$$\begin{aligned} u_t &= A(x, t)u_x + B(x, t)u + F(x, t), \\ u(x, 0) &= f(x). \end{aligned} \quad (6.2.1)$$

The generalization of Definition 4.3.1 is as follows.

Definition 6.2.1. *The variable coefficient system (6.2.1) is called strictly symmetric, and strongly or weakly hyperbolic if the matrix $A(x, t)$ satisfies the corresponding characterizations in Definition 4.3.1 at every fixed point $x = x_0, t = t_0$.*

We now assume that $A(x, t), B(x, t), F(x, t)$, and the initial values $f(x)$ are 2π -periodic in x . We start with a uniqueness theorem.

Theorem 6.2.1. *Assume that $A \in C^1(x, t)$ is Hermitian and that $B \in C(x, t)$. Then Eq. (6.2.1) has at most one 2π -periodic solution $u(x, t) \in C^1(x, t)$.*

Proof. Let $u(x, t), v(x, t)$ be two solutions of Eq. (6.2.1) belonging to $C^1(x, t)$. Then $w(x, t) = u(x, t) - v(x, t)$ satisfies the homogeneous equation

$$w_t = Aw_x + Bw$$

with homogeneous initial condition

$$w(x, 0) = 0.$$

As in Section 4.6, Lemma 4.6.1 gives us

$$\begin{aligned} \frac{d}{dt} \|w\|^2 &= (w, w_t) + (w_t, w) = (w, Aw_x) + (Aw_x, w) \\ &\quad + (w, Bw) + (Bw, w) \\ &= -(w, A_x w) + (w, (B + B^*)w) \leq 2\alpha \|w\|^2, \end{aligned}$$

where

$$2\alpha = \max_{x, t} (|A_x| + |B + B^*|).$$

Therefore,

$$\|w(\cdot, t)\|^2 \leq e^{2\alpha t} \|w(\cdot, 0)\|^2 = 0,$$

and the theorem follows.

Now assume that the system (6.2.1) is only strongly hyperbolic. By Lemma 6.1.1, we can find a positive definite Hermitian matrix $H = H(x, t)$ such that HA is Hermitian. If $H \in C^1(x, t)$, then we can proceed as in Theorem 6.1.1 and obtain the following theorem.

Theorem 6.2.2. *Replace the condition “ A is Hermitian” in Theorem 6.2.1 by “there is a positive definite matrix $H(x, t) \in C^1(x, t)$ such that HA is Hermitian.” Then the system (6.2.1) has at most one 2π -periodic solution $u(x, t) \in C^1(x, t)$.*

If the system is strictly hyperbolic, then one can choose the eigenvectors of A to be as smooth as the coefficients of A . Thus, if $A \in C^1(x, t)$, then the same

is true for the symmetrizer $H = (T^{-1})^* T^{-1}$. Without proof we now state an existence result.

Theorem 6.2.3. *Assume that the system (6.2.1) is strongly hyperbolic, that A, B, F , and f , are 2π -periodic functions of x and that they belong to $C^\infty(x, t)$. If there is a 2π -periodic symmetrizer $H \in C^\infty(x, t)$, then the initial value problem (6.2.1) has a solution $\in C^\infty(x, t)$. In particular, if the system is strictly hyperbolic, then there is a symmetrizer $H \in C^\infty(x, t)$.*

If the system is only weakly hyperbolic, then we cannot expect that the problem is well posed. This was shown in Section 4.3 for systems with constant coefficients. We show that the introduction of variable coefficients makes the situation worse in the sense that more rapid growth may occur.

Consider the system

$$\frac{\partial u}{\partial t} = U^*(t) \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} U(t) \frac{\partial u}{\partial x}, \quad (6.2.2)$$

where

$$U(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

The coefficient matrix in Eq. (6.2.2) has real eigenvalues but it cannot be diagonalized, so the system is only weakly hyperbolic. Introduce a new variable $v = U^* v$. Then v is the solution of

$$\begin{aligned} \frac{\partial v}{\partial t} &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \frac{\partial v}{\partial x} + Cv, \quad C = -UU_t^* = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \\ v(x, 0) &= Uu(x, 0). \end{aligned} \quad (6.2.3)$$

The system (6.2.3) has constant coefficients and we can solve it explicitly. Let the initial data consist of a simple wave

$$v(x, 0) = e^{i\omega x} \hat{q}(\omega), \quad \omega \text{ real.}$$

Then, as in Section 4.3, the solution is of the form $v(x, t) = \hat{v}(\omega, t)e^{i\omega x}$, where $\hat{v}(\omega, t)$ satisfies

$$\begin{aligned} d\hat{v}/dt &= \hat{A}\hat{v}, \quad \hat{v}(\omega, 0) = \hat{q}(\omega), \\ \hat{A} &= i\omega \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} + C = \begin{bmatrix} i\omega & i\omega + 1 \\ -1 & i\omega \end{bmatrix}. \end{aligned} \quad (6.2.4)$$

The eigenvalues of \hat{A} are given by

$$\kappa_1 = i\omega + \sqrt{-(i\omega + 1)}, \quad \kappa_2 = i\omega - \sqrt{-(i\omega + 1)}. \quad (6.2.5)$$

Thus, the solution of Eq. (6.2.4) has the form

$$\hat{v}(\omega, t) = \sigma_1 e^{\kappa_1 t} g_1 + \sigma_2 e^{\kappa_2 t} g_2,$$

where g_l denotes the eigenvector corresponding to the eigenvalue κ_l , and σ_1 and σ_2 are determined by the initial values, that is, by

$$\sigma_1 g_1 + \sigma_2 g_2 = \hat{q}(\omega).$$

For large values of $|\omega|$, one of the eigenvalues κ_l always satisfies

$$\operatorname{Re} \kappa = \left| \frac{\omega}{2} \right|^{1/2} + \mathcal{O}\left(\frac{1}{|\omega|^{1/2}} \right). \quad (6.2.6)$$

Thus, $\hat{v}(\omega, t)$ grows, in general, like $\exp(|\omega/2|^{1/2}t)$ and the same is true for $v(x, t)$ and $u(x, t)$.

To solve the system (6.2.3) numerically is rather difficult. Even if the initial data for the transformed system (6.2.4) only consists of a linear combination

$$v(x, 0) = \sum_{\omega=-M}^M e^{i\omega x} \hat{q}(\omega)$$

of low frequency components, rounding errors introduce high frequencies that can be greatly amplified. For example, if $\omega = 100, t = 10$, then $\exp(|\omega/2|^{1/2}t) = 5.12 \times 10^{30}$. Thus, if $\hat{v}_{100}(0) = 10^{-10}$ at $t = 0$, then $\hat{v}_{100}(t)$ is, in general, of the order 10^{20} at $t = 10$.

If the Euler equations are linearized around a solution that depends on x, y, z and t , there is a zero-order term left in the system as shown in Section 4.6. In the one-dimensional case, where all y and z derivatives vanish, the system does not decouple as it does for constant coefficients. However, if the zero-order term is disregarded, then we get a system with the same structure

as Eq. (6.1.3). Thus, the symmetrization can be done as for constant coefficients.

EXERCISES

- 6.2.1. Derive the explicit form of the zero-order term in the linearized Euler equations and prove that the system does not decouple in the one-dimensional case.
- 6.2.2. In Section 6.1, the second-order wave equation was derived from the linearized Euler equations with constant coefficients. Carry out the corresponding derivation for variable coefficients.

6.3 SYSTEMS WITH CONSTANT COEFFICIENTS IN SEVERAL SPACE DIMENSIONS

In this section, we consider first-order systems

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^d A_\nu \frac{\partial u}{\partial x^{(\nu)}} \quad (6.3.1)$$

with constant coefficients. Here $u = (u^{(1)}, \dots, u^{(m)})^T$ is a vector function with m components depending on $x = (x^{(1)}, \dots, x^{(d)})^T$ and t . The A_ν are constant $m \times m$ matrices. We are interested in 2π -periodic solutions, that is,

$$u(x + 2\pi e_\nu, t) = u(x, t), \quad \nu = 1, 2, \dots, d, \quad t \geq 0, \quad (6.3.2)$$

where e_ν denotes the unit vector in the $x^{(\nu)}$ direction. Also, at $t = 0$, $u(x, t)$ must satisfy 2π -periodic initial conditions

$$u(x, 0) = f(x). \quad (6.3.3)$$

We now generalize Definition 4.3.1 and define what we mean by hyperbolic.

Definition 6.3.1. Consider all linear combinations

$$\hat{P}(\omega) = \sum_{\nu=1}^d A_\nu \omega_\nu, \quad \omega_\nu \text{ real}, \quad |\omega| = \left(\sum_{\nu=1}^d \omega_\nu^2 \right)^{1/2} = 1.$$

The system (6.3.1) is called

1. strictly hyperbolic if, for every real vector $\omega = (\omega_1, \dots, \omega_d)$, the eigenvalues of $\hat{P}(\omega)$ are distinct and real,
2. symmetric hyperbolic if all matrices $A_v, v = 1, 2, \dots, d$, are Hermitian,
3. strongly hyperbolic if there is a constant $K > 0$ and, for every real vector ω , a nonsingular transformation $T(\omega)$ exists with

$$\sup_{|\omega|=1} (|T(\omega)| + |T^{-1}(\omega)|) \leq K$$

such that

$$T^{-1}(\omega)\hat{P}(\omega)T(\omega) = \Lambda, \\ \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda_j \text{ real},$$

4. weakly hyperbolic if the eigenvalues of $\hat{P}(\omega)$ are real.

We, therefore, have the following theorem.

Theorem 6.3.1. *If the system (6.3.1) is strictly or symmetric hyperbolic, then it is also strongly hyperbolic. If the system is strongly hyperbolic, then it is also weakly hyperbolic. Thus, the relations as expressed in Figure 4.3.1 are also valid in several space dimensions.*

Proof. The only nontrivial part is to prove that a strictly hyperbolic system is also strongly hyperbolic. It follows from Lemma A.1.9 in Appendix A.1 that the eigenvectors can be chosen as smooth functions of ω . Observing that $|\omega| = 1$ is a bounded set, it follows that $|T(\omega)| + |T^{-1}(\omega)|$ is uniformly bounded.

In view of the example in the last section, weakly hyperbolic systems are seldom considered in applications. In most cases the systems are either symmetric (or can be transformed to symmetric form) or strictly hyperbolic.

As in Section 4.5, we can solve the initial value problem using Fourier expansions. In particular, Theorem 4.5.3 gives us this theorem.

Theorem 6.3.2. *The initial value problem (6.3.1) is well posed for strongly hyperbolic systems.*

The generalization of the linearized Euler equations (6.1.3) to three space dimensions is

$$\begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & 0 & 0 & a^2/R \\ 0 & U & 0 & 0 \\ 0 & 0 & U & 0 \\ R & 0 & 0 & U \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_x + \begin{bmatrix} V & 0 & 0 & 0 \\ 0 & V & 0 & a^2/R \\ 0 & 0 & V & 0 \\ 0 & R & 0 & V \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_y \\ + \begin{bmatrix} W & 0 & 0 & 0 \\ 0 & W & 0 & 0 \\ 0 & 0 & W & a^2/R \\ 0 & 0 & R & W \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ \rho \end{bmatrix}_z = 0, \quad (6.3.4)$$

where we again assume that $a > 0$. The system is neither strictly nor symmetric hyperbolic. However, as in the one-dimensional case, it can be symmetrized by introducing $\tilde{\rho} = a\rho/R$ as a new variable. Consequently, the system is strongly hyperbolic.

The symbol is given by

$$\hat{P}(\omega) = \begin{bmatrix} \alpha & 0 & 0 & (a^2/R)\omega_1 \\ 0 & \alpha & 0 & (a^2/R)\omega_2 \\ 0 & 0 & \alpha & (a^2/R)\omega_3 \\ R\omega_1 & R\omega_2 & R\omega_3 & \alpha \end{bmatrix}, \quad \alpha = U\omega_1 + V\omega_2 + W\omega_3. \quad (6.3.5)$$

Its eigenvalues κ are

$$\kappa_1 = \kappa_2 = \alpha, \quad \kappa_3 = \alpha + a|\omega|, \quad \kappa_4 = \alpha - a|\omega|.$$

EXERCISES

6.3.1. Using the notation

$$\mathbf{u}_t + A_1 \mathbf{u}_x + A_2 \mathbf{u}_y + A_3 \mathbf{u}_z = 0.$$

for the system (6.3.4), find the matrix T such that $T^{-1}A_jT$, $j = 1, 2, 3$, are symmetric. Prove that there is no matrix S such that $S^{-1}A_jS$, $j = 1, 2, 3$, are diagonal.

6.3.2. Find a matrix $T(\omega)$ such that $T^{-1}(\omega)\hat{P}(\omega)T(\omega)$ is diagonal and

$$|T^{-1}(\omega)| \cdot |T(\omega)| \leq \text{constant for } |\omega| = 1,$$

where $\hat{P}(\omega)$ is defined in Eq. (6.3.5). Can one choose $T(\omega)$ as a smooth function of ω ?

6.4. SYSTEMS WITH VARIABLE COEFFICIENTS IN SEVERAL SPACE DIMENSIONS

In this section, we consider the initial value problem for

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{\nu=1}^d A_\nu(x, t) \frac{\partial u}{\partial x^{(\nu)}} + B(x, t)u + F(x, t), \\ u(x, 0) &= f(x).\end{aligned}\tag{6.4.1}$$

Here u, f , and F are vector functions with m components and A , and B are $m \times m$ matrices. We assume that A, B, F , and f are smooth [i.e., they belong to $C^\infty(x, t)$], 2π -periodic functions, and we want to find a 2π -periodic solution satisfying Eq. (6.3.2).

Hyperbolic is again defined pointwise.

Definition 6.4.1. *The system (6.4.1) is called strictly, symmetric, strongly, or weakly hyperbolic if the matrix*

$$\hat{P}(x, t, \omega) = \sum_{\nu=1}^d A_\nu(x, t)\omega_\nu, \quad \omega_\nu \text{ real},\tag{6.4.2}$$

satisfies the corresponding characterizations in Definition 6.3.1 at every point $x = x_0, t = t_0$. Using Theorem 6.2.3, one can prove the following theorem.

Theorem 6.4.1. *If the problem (6.4.1) is symmetric hyperbolic, then it has a unique smooth solution.*

Except for the zero-order term, the linearized Euler equations have the form (6.3.4), where U, V, W, a^2 , and R now depend on x, y, z , and t . Hence, the system can be symmetrized and the theorem can be applied.

If the system is only strongly hyperbolic, then the existence proof is more complicated and is beyond the scope of this book. In fact, a general existence theorem is not known, and one has to make more assumptions. We present the idea of the proof. Consider the symbol $i\hat{P}(\omega)$, defined by Eq. (6.4.2), and the diagonalizing matrix $T(x, t, \omega)$. By Lemma 6.1.1, we can construct a symmetrizer

$$\hat{H}(x, t, \omega) = (T^{-1}(x, t, \omega))^* D(x, t, \omega) T^{-1}(x, t, \omega)\tag{6.4.3}$$

for every fixed x, t, ω . Now assume that $\hat{H}(x, t, \omega)$ is a smooth function of all variables. We can then construct a positive definite bounded Hermitian operator H as in Section 4.5 such that we can obtain an energy estimate for (u, Hu) . We state the result as follows:

Theorem 6.4.2. Assume that the symmetrizer (6.4.3) can be chosen as a smooth function of all variables. Then the initial value problem (6.4.1) is well posed. If the system (6.4.1) is strictly hyperbolic, then there is a smooth symmetrizer.

EXERCISES

- 6.4.1.** Consider the linearized Euler equations (6.3.4) with variable coefficients obtained by linearizing around a nonconstant flow U, V, W, R, a^2 . Is there any flow such that the system is strictly hyperbolic at some point x_0, y_0, z_0, t_0 ?

6.5. APPROXIMATIONS WITH CONSTANT COEFFICIENTS

Throughout the previous chapters, hyperbolic model problems have often been used to demonstrate various features for different types of approximations. In this section, we shall give a more unified treatment for approximations of linear hyperbolic problems with constant coefficients.

In Section 5.2, it was shown that necessary and sufficient conditions for stability may be difficult to verify for problems with constant coefficients. Furthermore, for variable coefficient problems, the Fourier transform technique used in Section 5.2 does not apply directly. However, if the differential equation is hyperbolic, it is possible to give simple conditions such that stability follows from the stability of the constant coefficient problem. The key to this is the concept of dissipativity, which was discussed at the end of Section 5.2. [The definition of dissipativity was given for the one-dimensional case there. The only required change in the condition (5.2.31) is that $|\xi| \leq \pi$ must be replaced by $|\xi_j| \leq \pi, j = 1, 2, \dots, d$.]

First, we consider the hyperbolic system (6.3.1) and the explicit one-step approximation

$$\begin{aligned} v^{n+1} &= Qv^n, \\ Q &= \sum_l B_l E^l, \quad E^l = E_1^{l_1} \cdots E_d^{l_d}, \end{aligned} \tag{6.5.1}$$

where $l = (l_1, l_2, \dots, l_d)$ is a multiindex. E_ν is the translation operator in the coordinate $x^{(\nu)}$. It is assumed that the coefficient matrices B_l do not depend on (x_j, t_n) and that the space and time step are related by $k = \lambda h$, where λ is a constant.

All the results in Chapter 5 are given for general problems, and, consequently, they apply here. However, by using the structure of the differential equation, it is possible to derive sufficient stability conditions that are easier to verify. For example, we have the following theorem.

Theorem 6.5.1. Assume that Eq. (6.5.1) is an approximation of the strongly hyperbolic system (6.3.1). Then, it is stable if it is dissipative of order $2r$ and the order of accuracy is at least $2r - 1$.

Proof. The main idea is to use the information about the eigenvalues of $\hat{Q}(\xi)$ to bound the norm. Consistency is used to connect the properties of the approximation to those of the differential equation.

We first consider small values of $|\xi|$. It was shown in Theorem 5.2.5 that the order of accuracy can be defined in Fourier space, and, from the accuracy assumption, we have

$$\hat{Q} = e^{\hat{P}(i\omega)^k} + R = e^{\lambda i \sum_{\nu} A_{\nu} \xi_{\nu}} + R, \quad |R| \leq C|\xi|^{2r}, \quad |\xi| \leq \xi_0. \quad (6.5.2)$$

Since the system (6.3.1) is strongly hyperbolic, there is a matrix T such that

$$T^{-1}(\xi) \sum_{\nu} A_{\nu} \xi_{\nu} T(\xi) = D(\xi),$$

where D is real and diagonal. We now use the matrix T to transform \hat{Q} to “almost diagonal form” so that the magnitude of the norm is only slightly larger than the magnitude of the eigenvalues. Then the deviation from diagonal form can be compensated for by using dissipativity.

Let $S(\xi)$ be a unitary matrix that transforms

$$\hat{Q}_1 = T^{-1} \hat{Q} T \quad (6.5.3)$$

to upper triangular form. We get

$$S^* \hat{Q}_1 S = R_1 + R_2, \quad (6.5.4)$$

where

$$R_1 = S^* e^{i\lambda D} S, \quad R_2 = S^* T^{-1} R T S, \quad |R_2| = \mathcal{O}(|\xi|^{2r}).$$

Let

$$R_j = L_j + D_j + U_j, \quad j = 1, 2, \quad (6.5.5)$$

where L_j , D_j , and U_j denote the strictly lower triangular, diagonal, and strictly upper triangular parts, respectively. By construction $L_1 + L_2 = 0$, hence $|L_1| = \mathcal{O}(|\xi|^{2r})$. The matrix R_1 is unitary, and thus normal. Parlett (1966) has shown

that such matrices satisfy the inequality $|U_1| \leq \sqrt{m(m-1)}|L_1|$, and we obtain

$$|U_1| = \mathcal{O}(|\xi|^{2r}). \quad (6.5.6)$$

Because $|R_2| = \mathcal{O}(|\xi|^{2r})$, it follows that $|U_2| = \mathcal{O}(|\xi|^{2r})$ and, therefore,

$$|U_1 + U_2| \leq \alpha_0 |\xi|^{2r} \quad (6.5.7)$$

for some constant α_0 .

The next step is to transform $S^* \hat{Q}_1 S$ so that α_0 is replaced by a sufficiently small constant. In this way, the influence of the off-diagonal elements can be annihilated by the dissipative part of the eigenvalues that form the diagonal matrix $D_1 + D_2$. Let the diagonal matrix B be defined by

$$B = \text{diag}(b, b^2, \dots, b^m), \quad b > 1. \quad (6.5.8)$$

We have

$$\begin{aligned} |BS^* \hat{Q}_1 SB^{-1}| &= |B(R_1 + R_2)B^{-1}|, \\ &= |B(D_1 + D_2)B^{-1} + B(L_1 + L_2 + U_1 + U_2)B^{-1}|, \\ &\leq |D_1 + D_2| + |B(U_1 + U_2)B^{-1}|. \end{aligned}$$

An estimate of the first term is immediately obtained from the dissipativity condition. For the second term, we note that if U is any strictly upper triangular matrix, a direct calculation shows that the transformation BUB^{-1} decreases the norm by a factor b . Hence, from Eq. (6.5.7), we get

$$|BS^* \hat{Q}_1 SB^{-1}| \leq 1 - \left(\delta - \frac{\alpha_0}{b} \right) |\xi|^{2r} \leq 1 - \delta_1 |\xi|^{2r}, \quad \delta_1 > 0, \quad (6.5.9)$$

if b is chosen sufficiently large.

The inequality (6.5.9) can be interpreted as the construction of a new norm for \hat{Q} , which is bounded by one in the neighborhood of $|\xi| = 0$. We have

$$\begin{aligned} (S^* \hat{Q}_1 S)^* B^2 S^* \hat{Q}_1 S &= B [BS^* \hat{Q}_1 SB^{-1}]^* [BS^* \hat{Q}_1 SB^{-1}] B, \\ &\leq (1 - \delta_1 |\xi|^{2r}) B^2. \end{aligned}$$

Recalling that S is unitary, we get, by Eq. (6.5.3),

$$\hat{Q}^* (T^{-1})^* S B^2 S^* T^{-1} \hat{Q} \leq (1 - \delta_1 |\xi|^{2r}) (T^{-1})^* S B^2 S^* T^{-1};$$

that is, with

$$\hat{H} = (T^{-1})^* S B^2 S^* T^{-1}, \quad (6.5.10)$$

it follows that

$$\hat{Q}^* \hat{H} \hat{Q} \leq (1 - \delta_1 |\xi|^{2r})^2 \hat{H}. \quad (6.5.11)$$

\hat{H} clearly defines a norm $\langle \hat{v}, \hat{H} \hat{v} \rangle^{1/2}$ in the m -dimensional vector space, because it follows from Eq. (6.5.10) that \hat{H} is Hermitian and positive definite with

$$C_1 b^2 I \leq \hat{H} \leq C_2 b^{2m} I \quad (6.5.12)$$

for $C_1 > 0$ and $C_2 > 0$.

So far, we have only considered a neighborhood of $|\xi| = 0$. However, for $|\xi| \geq \xi_0 > 0$, we have $\rho(\hat{Q}) \leq 1 - \delta$, $\delta > 0$ and, therefore, by Lemma A.1.5 of Appendix A.1, we can find a uniformly positive definite $\hat{H}(\xi)$ such that

$$\hat{Q}^* \hat{H} \hat{Q} \leq \left(1 - \frac{\delta}{2} \right) \hat{H}. \quad (6.5.13)$$

Without restriction, we can assume that \hat{H} satisfies Eq. (6.5.12) for all ξ . In the new norm $\langle \hat{v}, \hat{H} \hat{v} \rangle$, the difference approximation is a so-called *contraction*, that is,

$$\begin{aligned} |\hat{v}^n|_{\hat{H}}^2 &= \langle \hat{v}^n, \hat{H} \hat{v}^n \rangle = \langle \hat{Q} \hat{v}^{n-1}, \hat{H} \hat{Q} \hat{v}^{n-1} \rangle = \langle \hat{v}^{n-1}, \hat{Q}^* \hat{H} \hat{Q} \hat{v}^{n-1} \rangle, \\ &\leq \langle \hat{v}^{n-1}, \hat{H} \hat{v}^{n-1} \rangle = |\hat{v}^{n-1}|_{\hat{H}}^2. \end{aligned}$$

The norm $|\cdot|_{\hat{H}}$ can be transformed into a corresponding norm in physical space.

Any vector gridfunction u_j can be written as

$$v_j = (2\pi)^{-d/2} \sum_{\omega} \hat{v}_{\omega} e^{i\langle \omega, x_j \rangle}. \quad (6.5.14)$$

The operator H is defined by

$$H v_j = (2\pi)^{-d/2} \sum_{\omega} \hat{H}(\omega h) \hat{v}_{\omega} e^{i\langle \omega, x_j \rangle}, \quad (6.5.15)$$

and we obtain, from Eq. (6.5.12),

$$b^2 \|v\|_h^2 \leq (v, Hv)_h \leq b^{2m} \|v\|_h^2. \quad (6.5.16)$$

Therefore, Parseval's relation and Eq. (6.5.11) give us

$$\begin{aligned} \|v^n\|_h^2 &\leq b^{-2}(v^n, Hv^n)_h = b^{-2} \sum_{\omega} \langle \hat{v}^n, \hat{H}\hat{v}^n \rangle, \\ &\leq b^{-2} \sum_{\omega} \langle \hat{v}^{n-1}, \hat{H}\hat{v}^{n-1} \rangle, \\ &\leq \dots \leq b^{-2} \sum_{\omega} \langle \hat{v}^0, \hat{H}\hat{v}^0 \rangle = b^{-2}(v^0, Hv^0)_h, \\ &\leq b^{2(m-1)} \|v^0\|_h^2. \end{aligned} \quad (6.5.17)$$

Thus, the method is stable.

This way of constructing a new norm is essential for approximations with variable coefficients. The technique used is a generalized form of the energy method introduced in Section 5.3. It is very similar to the method of proving well-posedness for nonsymmetric continuous problems.

The assumption that the order of accuracy must be at least $2r-1$ is restrictive. For example, the Lax-Wendroff method (2.1.17) is dissipative of order 4 and accurate of order 2. Therefore, Theorem 6.5.1 does not apply. There is a way to handle this difficulty. For general approximations, some extra condition must be used. For example, if we require strict hyperbolicity, the accuracy restriction can be removed.

Theorem 6.5.2. *Assume that Eq. (6.3.1) is strictly hyperbolic and has constant coefficients. Then the approximation (6.5.1) is stable if it is consistent and dissipative.*

Proof. We first consider a neighborhood of $\xi = 0$, $|\xi| \leq \xi_0$. By consistency, it follows from Theorem 5.2.5 that the symbol has the form

$$\hat{Q} = e^{\hat{P}(i\omega)k} + \mathcal{O}(|\xi|^2) = I + \lambda i |\xi| \left(\sum_{\nu} A_{\nu} \xi'_{\nu} + \mathcal{O}(|\xi|) \right), \quad \xi' = \frac{\xi}{|\xi|}.$$

By assumption, the eigenvalues of $\sum_{\nu} A_{\nu} \xi'_{\nu}$ are distinct. Thus, there is a matrix $T(\xi')$ with $|T(\xi')| \leq C$, $|T^{-1}(\xi')| \leq C$, such that $T^{-1} \sum_{\nu} A_{\nu} \xi'_{\nu} T$ is diagonal, and, for $|\xi|$ sufficiently small, we have

$$\hat{Q}_1 = T^{-1} \hat{Q} T = \text{diag}(z_1, \dots, z_m), \quad (6.5.18)$$

that is, by dissipativity,

$$|\hat{Q}_1| \leq 1 - \delta |\xi|^{2r}, \quad |\xi| \leq \xi_0. \quad (6.5.19)$$

The remaining part of the proof is almost identical to that of Theorem 6.5.1. For $|\xi| > \xi_0$, all of the eigenvalues are bounded away from the unit circle. We construct a new norm such that $|\hat{Q}|_{\hat{H}} < 1$ and obtain the result.

Dissipativity by itself is not sufficient to guarantee stability. Consider the approximation

$$v^{n+1} = v^n + \alpha \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} h^2 D_+ D_- v^n,$$

which is consistent with $u_t = 0$, $u = (u^{(1)}, u^{(2)})^T$. Note that this is not a strictly hyperbolic system. The amplification matrix is

$$\hat{Q} = I + 4\alpha \sin^2 \frac{\xi}{2} \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix},$$

showing that, for $\alpha < 1/2$, the approximation is dissipative of order 2. However, for small ξ , we have

$$\begin{aligned} |\hat{Q}^n| &= \left| \left(\begin{bmatrix} 1 - \alpha\xi^2 & \alpha\xi^2 \\ 0 & 1 - \alpha\xi^2 \end{bmatrix} + \mathcal{O}(\xi^4) \right)^n \right|, \\ &= \left| \begin{bmatrix} (1 - \alpha\xi^2)^n & \alpha n \xi^2 \\ 0 & (1 - \alpha\xi^2)^n \end{bmatrix} + \mathcal{O}(n\xi^4) \right|, \end{aligned}$$

and obviously the stability condition (5.2.11) is not satisfied.

One can also prove the following theorem.

Theorem 6.5.3. *Assume that Eq. (6.3.1) is symmetric hyperbolic and has constant coefficients. Then the approximation (6.5.1) is stable if*

1. *the coefficient matrices B_l are Hermitian*
2. *it is dissipative of order $2r$*
3. *the order of accuracy is at least $2r - 2$.*

The Lax–Wendroff method is based on the Taylor expansion

$$u(t_{n+1}) = u(t_n) + k u_t(t_n) + \frac{k^2}{2} u_{tt}(t_n) + \mathcal{O}(k^3). \quad (6.5.20)$$

For the differential equation,

$$u_t = A u_x + B u_y, \quad (6.5.21)$$

we have

$$u_{tt} = A u_{xt} + B u_{yt} = A^2 u_{xx} + (AB + BA) u_{xy} + B^2 u_{yy}, \quad (6.5.22)$$

which yields the approximation

$$\begin{aligned} v^{n+1} &= [I + k(AD_{0x} + BD_{0y}) + \frac{k^2}{2} (A^2 D_{+x} D_{-x} \\ &\quad + (AB + BA) D_{0x} D_{0y} + B^2 D_{+y} D_{-y})]v^n. \end{aligned} \quad (6.5.23)$$

We want to prove that Eq. (6.5.23) is stable if A and B are Hermitian. We can also write the scheme in the form

$$v^{n+1} = \left(I + kQ_0 + \frac{k^2}{2} Q_0^2 + \frac{k^2}{2} Q_1 \right) v^n, \quad (6.5.24)$$

where

$$\begin{aligned} Q_0 &= AD_{0x} + BD_{0y}, \\ Q_1 &= A^2(D_{+x}D_{-x} - D_{0x}^2) + B^2(D_{+y}D_{-y} - D_{0y}^2), \\ &= -\frac{h^2}{4} (A^2 D_{+x}^2 D_{-x}^2 + B^2 D_{+y}^2 D_{-y}^2). \end{aligned}$$

The amplification factor is

$$\begin{aligned} \hat{Q} &= I + \lambda \hat{Q}_0 + \frac{\lambda^2}{2} \hat{Q}_0^2 + \frac{\lambda^2}{2} \hat{Q}_1, \\ \hat{Q}_0 &= i(A \sin \xi_1 + B \sin \xi_2), \\ \hat{Q}_1 &= -4 \left(A^2 \sin^4 \frac{\xi_1}{2} + B^2 \sin^4 \frac{\xi_2}{2} \right). \end{aligned}$$

By Lemma A.1.4 in Appendix A.1, we can estimate the eigenvalues z of \hat{Q} with the help of the field of values. We have

$$\begin{aligned}
|z|^2 &\leq \max_{|v|=1} |\langle v, \hat{Q}v \rangle|^2 =: |\langle w, \hat{Q}w \rangle|^2, \\
&= \left(1 - \frac{\lambda^2}{2} |\hat{Q}_0 w|^2 + \frac{\lambda^2}{2} \langle w, \hat{Q}_1 w \rangle \right)^2 + \lambda^2 |\langle w, \hat{Q}_0 w \rangle|^2, \\
&= 1 - \lambda^2 |\hat{Q}_0 w|^2 + \lambda^2 \langle w, \hat{Q}_1 w \rangle - \frac{\lambda^4}{2} |\hat{Q}_0 w|^2 \langle w, \hat{Q}_1 w \rangle, \\
&\quad + \frac{\lambda^4}{4} |\hat{Q}_0 w|^4 - \frac{\lambda^4}{4} (\langle w, \hat{Q}_1 w \rangle)^2 + \lambda^2 |\hat{Q}_0 w|^2, \\
&= 1 - \lambda^2 \left(1 - \frac{\lambda^2}{2} |\hat{Q}_0 w|^2 - \frac{\lambda^2}{4} \langle w, \hat{Q}_1 w \rangle \right) \langle w, \hat{Q}_1 w \rangle \\
&\quad + \frac{\lambda^4}{4} |\hat{Q}_0 w|^4.
\end{aligned}$$

Also,

$$\begin{aligned}
\langle w, \hat{Q}_1 w \rangle &= 4 \left[\left| A \left(\sin^2 \frac{\xi_1}{2} \right) w \right|^2 + \left| B \left(\sin^2 \frac{\xi_2}{2} \right) w \right|^2 \right] \\
&\leq 8 \max(|A|^2, |B|^2), \\
|\hat{Q}_0 w|^4 &\leq [|A(\sin \xi_1)w| + |B(\sin \xi_2)w|]^4, \\
&\leq 8[|A(\sin \xi_1)w|^4 + |B(\sin \xi_2)w|^4], \\
&\leq 8 \cdot 16 \max(|A|^2, |B|^2) \left[\left| A \left(\sin^2 \frac{\xi_1}{2} \right) w \right|^2 \right. \\
&\quad \left. + \left| B \left(\sin^2 \frac{\xi_2}{2} \right) w \right|^2 \right], \\
&= 32 \max(|A|^2, |B|^2) \langle w, \hat{Q}_1 w \rangle, \\
|\hat{Q}_0 w|^2 &\leq 4 \max(|A|^2, |B|^2).
\end{aligned}$$

Therefore, we obtain

$$|z|^2 \leq 1 - \lambda^2 (1 - 12 \max(|A|^2, |B|^2) \lambda^2) \langle w, \hat{Q}_1 w \rangle.$$

Observing that $|A|^2 = \rho^2(A)$, it follows that $|z|^2 \leq 1$ for

$$\lambda \leq \lambda_0 := \frac{1}{2\sqrt{3}} \min \left(\frac{1}{\rho(A)}, \frac{1}{\rho(B)} \right). \quad (6.5.25)$$

If A and B are nonsingular, then $\langle w, \hat{Q}_1 w \rangle \geq \delta(|\xi_1|^4 + |\xi_2|^4)$ and, consequently, the scheme is dissipative of order four for $\lambda < \lambda_0$. Therefore, by Theorem 6.5.3, it is stable. If A or B is singular, we cannot use the theorem because the approximation is not dissipative. To prove stability we use the Kreiss Matrix Theorem 5.2.4. We have proved that

$$\max_{|v|=1} |\langle v, \hat{Q}v \rangle| \leq 1, \quad \text{for } \lambda \leq \lambda_0.$$

Therefore, the solutions of the resolvent equation

$$(\hat{Q} - zI)w = g$$

can be estimated. The equality

$$-\langle w, \hat{Q}w \rangle + z|w|^2 = -\langle w, g \rangle$$

implies, for $|z| > 1$, that

$$(|z| - 1)|w|^2 \leq |w||g|;$$

that is,

$$|(\hat{Q} - zI)^{-1}| \leq \frac{1}{|z| - 1}.$$

Therefore, the approximation is stable. We have proved Theorem 6.5.4.

Theorem 6.5.4. *The Lax–Wendroff method (6.5.23) is stable if $\lambda \leq \lambda_0$, where λ_0 is defined in Eq. (6.5.25). It is dissipative of order four if A and B are non-singular and $\lambda < \lambda_0$.*

If A or B is singular, we can modify Eq. (6.5.24) so that it will be dissipative by adding a fourth-order dissipation term; in other words, we use

$$v^{n+1} = \left(I + kQ_0 + \frac{k^2}{2} Q_0^2 + \frac{k^2}{2} Q_2 \right) v^n, \quad (6.5.26)$$

where

$$Q_2 = Q_1 - \sigma h^2 (D_{+x}^2 D_{-x}^2 + D_{+y}^2 D_{-y}^2), \quad \sigma = \text{constant} > 0.$$

As above, we can show that the approximation is dissipative of order four and, therefore, stable for all sufficiently small λ .

Next we consider linear multistep approximations.

Theorem 6.5.5. *Assume that Eq. (6.3.1) is strictly hyperbolic with constant coefficients and that Q is a difference operator that is consistent with $\sum_\nu A_\nu \partial/\partial x^{(\nu)}$. Then the approximation*

$$\sum_{\sigma=-1}^q \alpha_\sigma v^{n-\sigma} = k \sum_{\sigma=-1}^q \beta_\sigma Q v^{n-\sigma}$$

is stable if it is dissipative and if the only root on the unit circle of

$$\sum_{\sigma=-1}^q \alpha_\sigma z^{-\sigma} = 0 \quad (6.5.27)$$

is the simple root $z = 1$. (For convenience it is assumed that the coefficients α_σ and β_σ are independent of h .)

Proof. By Lemma 5.2.2, the eigenvalues of the symbol corresponding to the one-step form are given by

$$\det \left(\sum_{\sigma=-1}^q (\alpha_\sigma I - \beta_\sigma k \hat{Q}) z^{-\sigma} \right) = 0, \quad (6.5.28)$$

where \hat{Q} is the symbol of Q . By consistency

$$\begin{aligned} k \hat{Q} &= k \left(i \sum_\nu A_\nu \omega_\nu + \mathcal{O}(|\omega|^2 h) \right) = i \lambda \sum_\nu A_\nu \xi_\nu + \mathcal{O}(|\xi|^2), \\ &= i \lambda |\xi| \sum_\nu A_\nu \xi'_\nu + \mathcal{O}(|\xi|^2), \quad |\xi'| = 1. \end{aligned}$$

Strict hyperbolicity implies that there is a matrix $T = T(\xi')$ such that for $|\xi|$ sufficiently small

$$T^{-1} \hat{Q} T = D(\xi')$$

is diagonal with distinct eigenvalues. Here $|T^{-1}| \leq \text{constant}$, $|T| \leq \text{constant}$. Thus, Eq. (6.5.28) is equivalent to

$$\det \left(\sum_{\sigma=-1}^q (\alpha_\sigma I - \beta_\sigma |\xi| i \lambda D(\xi') z^{-\sigma}) \right) = 0. \quad (6.5.29)$$

By assumption, for $\xi = 0$, this equation has exactly m roots z_j with $z_j = 1$.

Because $D(\xi')$ is diagonal, we can separate Eq. (6.5.29) into m different equations

$$\sum_{\sigma=-1}^q (\alpha_\sigma - \beta_\sigma |\xi| i \lambda d_j) z^{-\sigma} = 0 \quad j = 1, 2, \dots, m.$$

Each one of these correspond to the one-step form

$$(\hat{w}^{(j)})^{n+1} = \tilde{Q}^{(j)}(\hat{w}^{(j)})^n$$

for a scalar equation in Fourier space (cf. Lemma 5.2.2). Here, the $\tilde{Q}^{(j)}$ are $(q+1) \times (q+1)$ matrices.

Thus, with $\hat{w} = (\hat{w}^{(1)}, \hat{w}^{(2)}, \dots, \hat{w}^{(m)})^T$, we have transformed the problem in Fourier space to the block-diagonal one-step form

$$\hat{w}^{n+1} = \tilde{\mathbf{Q}} \hat{w}^n,$$

where

$$\tilde{\mathbf{Q}} = \text{diag}(\tilde{Q}^{(1)}, \tilde{Q}^{(2)}, \dots, \tilde{Q}^{(m)}).$$

By assumption, for $\xi = 0$, there is only one eigenvalue of $\tilde{Q}^{(j)}$ with $z = 1$. Hence, by the dissipativity assumption, the eigenvalues satisfy

$$\begin{aligned} |z_1| &\leq 1 - \delta_1 |\xi|^{2r}, & \delta_1 > 0, \\ |z_\nu| &\leq 1 - \delta_2, & \delta_2 > 0, & \nu = 2, 3, \dots, q + 1. \end{aligned}$$

Thus, by Lemma A.1.6 in Appendix A.1, $\tilde{Q}^{(j)}$ can be transformed to the form

$$S^{-1} \tilde{Q}^{(j)} S = \begin{bmatrix} z_1 & 0 \\ 0 & \tilde{Q}_2^{(j)} \end{bmatrix},$$

where $\rho(\tilde{Q}_2^{(j)}) \leq 1 - \delta_2$, $|\xi| \leq \xi_0$, and $|S| + |S^{-1}| \leq \text{constant}$. Recalling the technique in the proof of Theorem 6.5.1, there is another norm $|\cdot|_{\hat{H}_2}$ such that

$$|\tilde{Q}_2^{(j)}|_{\hat{H}_2} \leq 1 - \delta_3, \quad \delta_3 > 0, \quad |\xi| \leq \xi_0.$$

With

$$\hat{H} = \begin{bmatrix} 1 & 0 \\ 0 & \hat{H}_2 \end{bmatrix},$$

we now have

$$|S^{-1}\tilde{Q}^{(j)}S|_{\hat{H}} \leq 1, \quad |\xi| \leq \xi_0.$$

This concludes the case $|\xi| \leq \xi_0$.

For $|\xi| > \xi_0$, all eigenvalues z_j are bounded away from the unit circle. Therefore, we can complete the proof as in Theorem 6.5.1.

Up to this point, we have assumed that there are no lower order terms in the differential equation. However, we recall from Section 5.1 that lower order perturbations and forcing functions do not affect stability. This means that we can treat general systems of the form

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^d A_\nu \frac{\partial u}{\partial x^{(\nu)}} + B(x, t)u + F(x, t), \quad (6.5.30)$$

where the matrices A_ν are constant.

For most approximations, the additional terms can be approximated in an obvious way. For example, one version of the leap-frog scheme is

$$v^{n+1} = 2k \sum_{\nu=1}^d A_\nu D_{0\nu} v^n + 2kB^n v^n + v^{n-1} + 2kF^n. \quad (6.5.31)$$

However, the example given in Section 2.2 shows that it is often better to substitute

$$B^n v^n \rightarrow \frac{1}{2}B^n(v^{n+1} + v^{n-1}).$$

In both cases, it is clear that the resulting scheme has the form of Eq. (5.1.20). Therefore, the scheme is stable if the principal part is stable. The effect of the forcing function is described in Theorem 5.1.1.

A straightforward derivation of a scheme of Lax-Wendroff type is obtained

by differentiating Eq. (6.5.30). For the one-dimensional case, we have

$$\begin{aligned} u_{tt} &= Au_{xt} + B(x, t)u_t + B_t(x, t)u + F_t(x, t), \\ &= A^2u_{xx} + (AB + BA)u_x + (AB_x + B^2 + B_t)u + AF_x + BF + F_t. \end{aligned}$$

Assume that the derivatives of B and F are available analytically. Then the scheme is

$$\begin{aligned} v^{n+1} &= \left(I + kAD_0 + \frac{k^2}{2} A^2 D_+ D_- \right) v^n \\ &\quad + k \left(\frac{k}{2} (AB(t_n) + B(t_n)A)D_0 + \frac{k}{2} (AB_x(t_n) \right. \\ &\quad \left. + B(t_n)^2 + B_t(t_n)) + B(t_n) \right) v^n \\ &\quad + k \left(F^n + \frac{k}{2} (AF_x^n + B(t_n)F^n + F_t^n) \right), \end{aligned} \quad (6.5.32)$$

and consists of three parts of different orders in h . The first term is the principal part, and we assume that $(k\rho(A) \leq h)$ so that it is stable. The operator kD_0 occurring in the second part is bounded because

$$\|kD_0v\|_h = \lambda h \|D_0v\|_h = \lambda \|v\|_h.$$

If B_x and B_t are bounded, then stability follows from Theorem 5.1.2 and an estimate in terms of F , F_x , and F_t is obtained from Theorem 5.1.1.

Actually, the requirements on B and F can be weakened. If B is bounded, but B_x and B_t are not, we substitute

$$\begin{aligned} kB_x(x_j, t_n) &\rightarrow \frac{\lambda}{2} (B(x_{j+1}, t_n) - B(x_{j-1}, t_n)), \\ kB_t(x_j, t_n) &\rightarrow \frac{1}{2} (B(x_j, t_{n+1}) - B(x_j, t_{n-1})), \end{aligned}$$

which are bounded operators. Since these terms are multiplied by an extra factor k , stability follows. Similarly, we can substitute differences for derivatives of F and obtain an estimate in terms of F as for the differential equation.

EXERCISES

- 6.5.1.** Prove stability for the approximation (6.5.26) for the case that A or B in Eq. (6.5.21) is singular.

6.5.2. Let

$$Q = AD_{0x} + BD_{0y} - \delta h^3((D_{+x}D_{-x})^2 + (D_{+y}D_{-y})^2)$$

and use “backward differentiation”

$$\frac{3}{2}v^{n+1} - 2v^n + \frac{1}{2}v^{n-1} = kQv^{n+1}$$

for an approximation of $u_t = Au_x + Bu_y$. Derive the stability condition for k/h .

6.6. APPROXIMATIONS WITH VARIABLE COEFFICIENTS

In Section 5.3, general approximations with variable coefficients were treated, and the energy method was used as the main tool for stability analysis. For dissipative difference schemes, it is also possible to generalize the theorems discussed in the previous section, which were based on Fourier analysis. The symbol \hat{Q} and dissipativity are defined pointwise. Note, however, that \hat{Q} cannot be used to obtain an analytic representation of the solution in general.

Consider the symmetric hyperbolic system (6.4.1) with variable coefficients and a one-step explicit difference scheme

$$v_j^{n+1} = Q(x_j, t_n, h)v_j^n, \quad Q = \sum_l A_l(x, t, h)E^l. \quad (6.6.1)$$

The symbol is given by

$$\hat{Q}(x, t, h, \xi) = \sum_l A_l(x, t, h)\hat{E}^l,$$

where

$$\hat{E}^l = e^{-i\langle \omega, x \rangle} E^l e^{i\langle \omega, x \rangle} = e^{-i\langle \omega, x \rangle} E_1^{l_1} \cdots E_d^{l_d} e^{i\langle \omega, x \rangle} = e^{i\langle l, \xi \rangle}.$$

Dissipativity is defined as follows.

Definition 6.6.1. *The approximation (6.6.1) is dissipative of order $2r$ if all the eigenvalues z_μ of $\hat{Q}(x, t, 0, \xi)$ satisfy*

$$|z_\mu(x, t, 0, \xi)| \leq (1 - \delta|\xi|^{2r}), \\ |\xi_\nu| \leq \pi, \quad \nu = 1, 2, \dots, d, \quad \mu = 1, 2, \dots, m,$$

for all x, t , where $\delta > 0$ is a constant independent of x, t , and ξ .

We now generalize Theorem 6.5.3.

Theorem 6.6.1. Assume that the hyperbolic system (6.4.1) and the approximation (6.6.1) have Hermitian coefficient matrices that are Lipschitz continuous in x and t . Then the approximation is stable if it is accurate of at least order $2r - 2$ and dissipative of order $2r$.

We shall not give the proof here. The technique used is similar to that used to prove Theorem 6.5.1. A new norm is constructed via the matrix \hat{H} defined in Eq. (6.5.10) (where T is now unitary since we have Hermitian coefficients).

For strictly hyperbolic systems we have the following theorem.

Theorem 6.6.2. Assume that Eq. (6.4.1) is strictly hyperbolic and that the approximation (6.6.1) is consistent, dissipative, and has coefficients that are Lipschitz continuous in x and t . Then the approximation is stable.

As an example, consider the Lax–Wendroff method for the system

$$u_t = A(x)u_x,$$

where $A(x)$ is uniformly nonsingular. Using

$$u_{tt} = A(x)u_{xt} = A(x)[A(x)u_x]_x$$

we obtain the second-order accurate approximation

$$v_j^{n+1} = v_j^n + kA_j D_0 v_j^n + \frac{k^2}{2} A_j D_+(A_{j-1/2} D_- v_j^n). \quad (6.6.2)$$

Here A_j can be used in place of $A_{j-1/2}$ without losing second-order accuracy. Let $k/h = \lambda = \text{constant} > 0$. If $A(x)$ is at least Lipschitz continuous we can write Eq. (6.6.2) in the form

$$v_j^{n+1} = (I + kA_j D_0 + \frac{k^2}{2} A_j^2 D_+ D_-) v_j^n + kR v_j^n, \quad (6.6.3)$$

where the operator R is uniformly bounded. Therefore, the last term can be

disregarded and we obtain the symbol

$$\hat{Q}(x, \xi) = I + \lambda A_i \sin \xi - 2\lambda^2 A^2 \sin^2 \frac{\xi}{2}.$$

The approximation is consistent and dissipative of order 4 if

$$k/h \max_x \varrho(A(x)) < 1.$$

Hence, by Theorem 6.6.2, it is stable if $A(x)$ has distinct eigenvalues.

Nonlinear systems can often be written in so called conservation form

$$\begin{aligned} u_t &= \frac{\partial}{\partial x} F(x, t, u), \\ u(x, 0) &= f(x). \end{aligned} \quad (6.6.4)$$

Consider the second-order accurate two-step procedure

$$v_{j+1/2}^{n+1/2} = \frac{1}{2}(v_{j+1}^n + v_j^n) + \frac{k}{2} D_+ F_j^n, \quad (6.6.5a)$$

$$v_j^{n+1} = v_j^n + \frac{k}{h}(F_{j+1/2}^{n+1/2} - F_{j-1/2}^{n+1/2}), \quad (6.6.5b)$$

where we have used the notation $F_j^n = F(x_j, t_n, v_j^n)$. Note that the halfstep subscripts and superscripts are just notation, no extra gridpoints are generated. Equation (6.6.5a) just defines provisional values for Eq. (6.6.5b) and the approximation is a one-step scheme in the sense that no extra initial data are needed.

Assume that Eq. (6.6.4) has a smooth solution U . As in Section 5.5, the stability behavior of the scheme is determined by the linearized error equation. We make the ansatz

$$v = U + h^2 e.$$

Then

$$F_j^n = F(x_j, t_n, v_j^n) = F(x_j, t_n, U_j^n) + h^2 (Ae)_j^n + \mathcal{O}(h^4 |e|^2),$$

where

$$A(x, t) = \frac{\partial}{\partial u} F(x, t, U),$$

gives us

$$\begin{aligned} e_j^{n+1/2} &= \frac{1}{2} (e_{j+1}^n + e_j^n) + \frac{k}{2} D_+(Ae)_j^n + kG_{j+1/2}^n, \\ e_j^{n+1} &= e_j^n + \frac{k}{h} ((Ae)_{j+1/2}^{n+1/2} - (Ae)_{j-1/2}^{n+1/2}) + kH_j^n, \\ e_j^0 &= 0. \end{aligned} \quad (6.6.6)$$

Here the forcing functions G and H are functions of U and e . The coefficients of the nonlinear terms in e are of the order h^4 . By referring back to Section 5.5, we find that e is bounded if the linear part of Eq. (6.6.6) is stable. For the discussion of the stability, we can assume that A is constant and disregard G and H . Then we can easily eliminate the halfstep and obtain

$$\begin{aligned} e_j^{n+1} &= e_j^n + \frac{k}{h} A(e_{j+1/2}^{n+1/2} - e_{j-1/2}^{n+1/2}), \\ &= e_j^n + \frac{k}{h} A \left[\frac{1}{2} (e_{j+1}^n + e_j^n) + \frac{k}{2} AD_+ e_j^n, \right. \\ &\quad \left. - \frac{1}{2} (e_j^n + e_{j-1}^n) - \frac{k}{2} AD_- e_{j-1}^n \right], \\ &= (I + kAD_0 + \frac{k^2}{2} A^2 D_- D_+) e_j^n, \end{aligned}$$

which is equivalent with the original Lax–Wendroff approximation (6.6.3) with $R = 0$. (All methods that reduce to this scheme for constant coefficients are said to be of the Lax–Wendroff type.) Hence, for $k\rho(A)/h < 1$, e is bounded and the error $v - U = \mathcal{O}(h^2)$.

EXERCISES

- 6.6.1.** Prove that the second-order accuracy of Eq. (6.6.2) is retained if $A_{j-1/2}$ is replaced by A_j .
- 6.6.2.** Consider the Lax–Wendroff approximation (6.6.2) of the linearized Euler equations (6.1.3), where U , R , and a^2 depend on x and t . Theorem 6.6.1 or 6.6.2 cannot be used to prove stability if the flow has a *stagnation point* $U(x_0, t_0) = 0$ or a *sonic point* $U(x_1, t_1) = a(x_1, t_1)$. Why is that so, and how can the approximation be modified to be stable?

6.7. THE METHOD OF LINES

If one discretizes in space and leaves time continuous, one obtains a system of ordinary differential equations. This system is then solved by a standard method

like a Runge–Kutta method or a multistep method. Most methods in use can be constructed in this way, for example, the leap-frog and the Crank–Nicholson methods. In the latter case, the trapezoidal rule is used for time discretization. Exceptions are the Lax–Wendroff method and the combination of the leap-frog and Crank–Nicholson methods.

Consider a strongly hyperbolic system

$$\frac{\partial u}{\partial t} = P \left(\frac{\partial}{\partial x} \right) u = \sum_{\nu} A_{\nu} \frac{\partial u}{\partial x^{(\nu)}} \quad (6.7.1)$$

with constant coefficients. We can solve it by Fourier transform; that is, we solve the system of ordinary differential equations

$$\frac{d\hat{u}}{dt} = \hat{P}(i\omega)\hat{u}, \quad \hat{P} = i \sum_{\nu} A_{\nu} \omega_{\nu}. \quad (6.7.2)$$

We now construct difference approximations. To begin with, we discretize only space and not time; that is, we approximate $P(\partial/\partial x)$ by a difference operator

$$Q_1 = \frac{1}{h} \sum_l B_l E^l, \quad (6.7.3)$$

where the matrices B_l do not depend on h . We obtain a system of ordinary differential equations

$$\frac{dw}{dt} = Q_1 w, \quad (6.7.4)$$

where w is the vector containing the gridvalues. We assume that the approximation is accurate of order $2r - 1$, $r \geq 1$; thus, for smooth solutions $u(x, t)$, we have

$$\|Pu - Q_1 u\|_h \leq \text{constant } h^{2r-1}. \quad (6.7.5)$$

In general, Eq. (6.7.4) is not useful because it is not stable [which means that Eq. (6.7.4) has solutions that grow like $\exp(at/h)$, $a > 0$]. We modify the system and consider

$$\frac{dw}{dt} = Qw, \quad Q = Q_1 + \sigma Q_2, \quad \sigma = \text{constant} > 0, \quad (6.7.6)$$

where

$$Q_2 = (-1)^{r-1} h^{2r-1} \sum_{\nu} D_{+x^{(\nu)}}^r D_{-x^{(\nu)}}^r.$$

Observe that the order of accuracy is not changed. We want to show that, for sufficiently large σ , the solutions of Eq. (6.7.6) decay exponentially. Fourier transforming Eq. (6.7.6) gives us

$$\frac{d\hat{w}}{dt} = \hat{Q}(i\xi)\hat{w} = \left(\hat{Q}_1(i\xi) + \frac{\sigma}{h} \hat{Q}_2(i\xi) \right) \hat{w}, \quad |\xi_{\nu}| \leq \pi, \quad \xi = \omega h,$$

where

$$\begin{aligned} h\hat{Q}_1(i\xi) &= \hat{P}(i\xi) + \hat{R}_{2r}(i\xi), \quad |\hat{R}_{2r}(i\xi)| \leq \text{const. } |\xi|^{2r}, \\ \hat{Q}_2 &= -4^r \left(\sum_{\nu} \sin^{2r} \frac{\xi_{\nu}}{2} \right) I. \end{aligned}$$

The form of \hat{Q}_1 is derived using the same technique as in Theorem 5.2.5. By assumption, Eq. (6.7.1) is strongly hyperbolic. Therefore, there is a transformation S such that $S^{-1}\hat{P}S = i\Lambda$ where Λ is a real diagonal matrix. Introducing $S^{-1}\hat{w} = \tilde{w}$ as a new variable gives us

$$\frac{d\tilde{w}}{dt} = h^{-1}(i\Lambda + S^{-1}\hat{R}_{2r}S + \sigma\hat{Q}_2)\tilde{w} =: \tilde{Q}\tilde{w}. \quad (6.7.7a)$$

For sufficiently large σ ,

$$\tilde{Q} + \tilde{Q}^* \leq \frac{\sigma}{2h} \hat{Q}_2, \quad (6.7.7b)$$

and we have proved that the solutions of Eq. (6.7.6) decay exponentially.

In analogy to Definition 5.2.1, we define dissipativity for the method of lines.

Definition 6.7.1. *The approximation (6.7.6) is dissipative of order $2r$ if all the eigenvalues s_{ν} of \hat{Q} satisfy*

$$\operatorname{Re} s_{\nu} \leq \alpha_s - \delta h^{-1} |\xi|^{2r}, \quad |\xi_{\nu}| \leq \pi, \quad (6.7.8)$$

where α_s is the exponential growth factor.

Now assume that Q_1 is accurate of only order $2r - 2$. Then

$$h\hat{Q}_1(i\xi) = \hat{P}(i\xi) + \hat{R}_{2r-1}(i\xi) + \hat{R}_{2r}(i\xi),$$

where

$$|\hat{R}_{2r-1}| \leq \text{constant } |\xi|^{2r-1}.$$

Thus,

$$hS^{-1}\hat{Q}_1S = i\Lambda + S^{-1}\hat{R}_{2r-1}S + S^{-1}\hat{R}_{2r}S,$$

and Eq. (6.7.7b) holds if $S^{-1}\hat{R}_{2r-1}S$ is anti-Hermitian. If Eq. (6.7.1) is a symmetric hyperbolic system, then S is a unitary matrix, and we need anti-Hermitian coefficients in \hat{R}_{2r-1} . This condition is fulfilled for all centered approximations of the type Eq. (3.1.7). Thus we have proved the following theorem.

Theorem 6.7.1. *Approximate P by any difference operator Q_1 that is accurate of order $2r - 1$. We can add dissipation terms of order $2r$ that do not change the order of accuracy so that the semidiscrete approximation (6.7.6) is stable and dissipative of order $2r$.*

If Q_1 is only accurate of order $2r - 2$, then the procedure leads to a stable approximation if the system of differential equations is symmetric hyperbolic and if Q_1 is a centered approximation of the type of Eq. (3.1.7).

We now discretize time and discuss some of the standard methods used to solve ordinary differential equations numerically. We start with the Runge–Kutta methods.

6.7.1. Runge–Kutta Methods

Let $y' = f(y, t)$ be a system of ordinary differential equations. The classical fourth-order accurate Runge–Kutta method is given by

$$v^{n+1} = v^n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

where $k > 0$ denotes the time step, and

$$\begin{aligned}k_1 &= kf(v^n, t_n), \\k_2 &= kf\left(v^n + \frac{k_1}{2}, t_n + \frac{k}{2}\right), \\k_3 &= kf\left(v^n + \frac{k_2}{2}, t_n + \frac{k}{2}\right), \\k_4 &= kf(v^n + k_3, t_n + k).\end{aligned}$$

For the system (6.7.6), we obtain

$$\begin{aligned}k_1 &= kQv^n, \\k_2 &= kQ\left(I + \frac{k}{2}Q\right)v^n, \\k_3 &= kQ\left(I + \frac{k}{2}Q\left(I + \frac{k}{2}Q\right)\right)v^n, \\k_4 &= kQ\left(I + kQ\left(I + \frac{k}{2}Q\left(I + \frac{k}{2}Q\right)\right)\right)v^n.\end{aligned}$$

Thus,

$$v^{n+1} = \left(\sum_{j=0}^4 \frac{(kQ)^j}{j!} \right) v^n. \quad (6.7.9)$$

General p th-order accurate Runge–Kutta methods for Eq. (6.7.6) are of the form

$$v^{n+1} = \left(\sum_{j=0}^p \frac{(kQ)^j}{j!} + \sum_{j=p+1}^m \alpha_j \frac{(kQ)^j}{j!} \right) v^n. \quad (6.7.10)$$

Let Q be the modified difference operator in Eq. (6.7.6). We Fourier transform, change variables as in Eq. (6.7.7a) and obtain

$$\tilde{v}^{n+1} = \left(\sum_{j=0}^p \frac{(k\tilde{Q})^j}{j!} + \sum_{j=p+1}^m \alpha_j \frac{(k\tilde{Q})^j}{j!} \right) \tilde{v}^n, \quad (6.7.11a)$$

where

$$k\tilde{Q} = i\lambda\Lambda + \lambda\tilde{R}_{2r}(i\xi) + \sigma\lambda\hat{Q}_2(i\xi) \quad \text{if } p = 2r - 1, \quad (6.7.11b)$$

or

$$k\tilde{Q} = i\lambda\Lambda + \lambda\tilde{R}_{2r-1}(i\xi) + \lambda\tilde{R}_{2r}(i\xi) + \sigma\lambda\hat{Q}_2(i\xi), \quad \text{if } p = 2r - 2. \quad (6.7.11c)$$

In the second case, we assume that

$$\tilde{R}_{2r-1}^* = -\tilde{R}_{2r-1}. \quad (6.7.12)$$

This is a natural assumption for symmetric hyperbolic systems and centered difference operators.

We need the following lemma.

Lemma 6.7.1. *Let A be a matrix and $\{\alpha_j\}_{j=p+1}^m$ given scalars, where p is an integer with $1 \leq p \leq m$. For sufficiently small $|A|$ we can find a matrix B that commutes with A such that*

$$e^{A+B} = \sum_{j=0}^p \frac{A^j}{j!} + \sum_{j=p+1}^m \alpha_j \frac{A^j}{j!}. \quad (6.7.13)$$

The matrix B has a series expansion

$$B = \sum_{j=p+1}^{\infty} \beta_j A^j, \quad (6.7.14a)$$

where

$$\begin{aligned} \beta_{p+1} &= \frac{\alpha_{p+1} - 1}{(p+1)!}, \\ \beta_{p+2} &= \frac{\alpha_{p+2} - 1}{(p+1)!} - \frac{\alpha_{p+1} - 1}{(p+1)!}. \end{aligned} \quad (6.7.14b)$$

Proof. We make the ansatz (6.7.14a) and write Eq. (6.7.13) as

$$e^{A+B} = e^A + S,$$

where

$$S = \sum_{j=p+1}^m \frac{\alpha_j - 1}{j!} A^j - \sum_{j=m+1}^{\infty} \frac{A^j}{j!}.$$

After multiplying by e^{-A} , we get

$$e^B = I + e^{-A}S,$$

and after taking the logarithm of both sides

$$B = \log(I + e^{-A}S) = e^{-A}S - \frac{1}{2}(e^{-A}S)^2 + \dots \quad (6.7.15)$$

Because $\log(I + e^{-A}S)$ is well defined, the matrix B exists, and the coefficients β_j in Eq. (6.7.14) are defined by the series expansion in Eq. (6.7.15). When calculating the coefficients of the leading terms, we obtain Eq. (6.7.14b). This proves the lemma.

When substituting $kQ \rightarrow A, kR \rightarrow B$ in the lemma, the solution of Eq. (6.7.11) can be written as

$$\tilde{v}^n = e^{(\tilde{Q} + \tilde{R})t_n} \tilde{v}^0,$$

where

$$k\tilde{R} = \beta_{p+1}(k\tilde{Q})^{p+1} + \beta_{p+2}(k\tilde{Q})^{p+2} + \mathcal{O}((k|\tilde{Q}|)^{p+3}), \quad (6.7.16)$$

with β_{p+1} and β_{p+2} defined in Eq. (6.7.14b).

Therefore, the approximation is stable if

$$\operatorname{Re} k(\tilde{Q} + \tilde{R}) = \frac{1}{2}(k(\tilde{Q} + \tilde{R}) + k(\tilde{Q} + \tilde{R})^*) \leq 0.$$

First, consider the case $p = 2r - 1$. Then, recalling that $\Lambda(\xi) = \mathcal{O}(|\xi|)$, we have

$$\operatorname{Re}(k\tilde{R}) = \mathcal{O}((k|\tilde{Q}|)^{2r}) = \mathcal{O}(|\lambda\xi|^{2r}).$$

Furthermore,

$$\begin{aligned} \operatorname{Re}(k\tilde{Q}) &= \operatorname{Re}(\lambda R_{2r}(i\xi) + \sigma\lambda\hat{Q}_2(i\xi)), \\ &= \mathcal{O}(\lambda|R_{2r}(i\xi)|) - 4^r\sigma\lambda \left(\sum_{\nu} \sin^{2r} \frac{\xi_{\nu}}{2} \right) I. \end{aligned} \quad (6.7.17)$$

Thus, for λ sufficiently small and σ sufficiently large,

$$\operatorname{Re} k(\tilde{Q} + \tilde{R}) \leq 0.$$

Next, consider the case $p = 2r - 2$, and assume that Eq. (6.7.12) is satisfied. Then Eq. (6.7.17) holds and, furthermore,

$$\begin{aligned}\operatorname{Re}(k\tilde{R}) &= \beta_{p+1} \operatorname{Re}(k\tilde{Q})^{2r-1} + \mathcal{O}((k|\tilde{Q}|)^{2r}), \\ &= \beta_{p+1} \operatorname{Re}(i\lambda\Lambda(\xi))^{2r-1} + \mathcal{O}(|\lambda\xi|^{2r}) = \mathcal{O}(|\lambda\xi|^{2r}).\end{aligned}$$

Thus, we have the same situation as for $p = 2r - 1$ and we obtain Theorem 6.7.2.

Theorem 6.7.2. *If the Runge-Kutta method is accurate of order $p = 2r - 1$, then approximation (6.7.10) is stable, provided σ is sufficiently large and λ is sufficiently small. If $p = 2r - 2$, this is also true if the system is symmetric and condition (6.7.12) holds.*

The last theorem tells us that we can stabilize any Runge-Kutta method by adding dissipative terms. We now construct stable methods by using nondissipative operators in space. We approximate

$$P\left(\frac{\partial}{\partial x}\right) = \sum_{\nu} A_{\nu} \frac{\partial}{\partial x^{(\nu)}}$$

by

$$Q = \sum_{\nu} A_{\nu} Q_{x^{(\nu)}}, \quad (6.7.18)$$

where

$$Q_{x^{(\nu)}} = D_{0x^{(\nu)}} \sum_{j=0}^{(q/2)-1} (-1)^j \alpha_j (h^2 D_{+x^{(\nu)}} D_{-x^{(\nu)}})^j,$$

are the centered operators (3.1.7). Then

$$\hat{Q} = i \sum_{\nu} A_{\nu} \tau_{\nu}, \quad h\tau_{\nu} = \sin \xi_{\nu} \sum_{j=0}^{(q/2)-1} 4^j \alpha_j \left(\sin \frac{\xi_{\nu}}{2} \right)^{2j};$$

that is, the symbol is of the same form as $P(i\omega)$. Therefore, there is a transfor-

mation that transforms \hat{Q} to diagonal form \tilde{Q} , and we can assume that

$$k\tilde{Q} = i\lambda\Lambda,$$

where Λ is a real diagonal matrix, that is,

$$\operatorname{Re}(k\tilde{Q}) = 0.$$

The derivation leading to the form shown in Eq. (6.7.16) for the matrix \tilde{R} also applies here, and, for sufficiently small λ ,

$$\operatorname{Re}(k\tilde{R}) = \begin{cases} \beta_{p+1}(i\lambda\Lambda)^{p+1} + \mathcal{O}((i\lambda\Lambda)^{p+3}), & \text{if } p \text{ is odd,} \\ \beta_{p+2}(i\lambda\Lambda)^{p+2} + \mathcal{O}((i\lambda\Lambda)^{p+4}), & \text{if } p \text{ is even.} \end{cases}$$

Therefore, we obtain the following theorem.

Theorem 6.7.3. *For sufficiently small λ , the Runge–Kutta methods shown in Eq. (6.7.10), with Q defined by Eq. (6.7.18), are stable if*

$$(-1)^{(p+1)/2}\beta_{p+1} < 0, \quad \text{for } p \text{ odd,}$$

or

$$(-1)^{(p+2)/2}\beta_{p+2} < 0, \quad \text{for } p \text{ even.}$$

Here β_{p+1} and β_{p+2} are defined as in Eq. (6.7.14b). If the above expressions are positive, then the methods are unstable.

For the classical Runge–Kutta methods, all $\alpha_\nu = 0$ in Eq. (6.7.10). In this case, the last theorem and Eq. (6.7.14b) tell us

Theorem 6.7.4. *If $\alpha_{p+1} = 0$ for p odd or $\alpha_{p+1} = \alpha_{p+2} = 0$ for p even, then the Runge–Kutta methods are stable for $p = 3, 7, 11, \dots$ or $p = 4, 8, 12, \dots$ and λ sufficiently small. For all other p , the methods are unstable.*

Thus, the classical fourth-order accurate Runge–Kutta method is stable. The same is true for the third-order accurate Runge–Kutta method (usually called the Runge–Kutta Heun method).

We now consider the model equation

$$u_t = u_x, \tag{6.7.19}$$

in more detail. We approximate $\partial/\partial x$ by the fourth-order accurate operator

$$D_0(h) \left(I - \frac{h^2}{6} D_+(h) D_-(h) \right) = \frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h)$$

and obtain the system of ordinary differential equations

$$w_t = (\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h)) w, \quad (6.7.20)$$

which, after Fourier transformation, becomes

$$\hat{w}_t = i\alpha \hat{w}, \quad h\alpha = \frac{4}{3} \sin \xi - \frac{1}{6} \sin 2\xi. \quad (6.7.21)$$

We discuss the classical Runge–Kutta methods of order $p = 1, 2, 3, 4$. In this case, all coefficients $\alpha_\nu = 0$ in Eq. (6.7.10). For $p = 1$, we obtain Euler's explicit method. In Section 2.1, we proved that it is unstable. For $p = 2$, the method is also unstable. This follows from Theorem 6.7.4. For $p = 3$, we obtain the Fourier-transformed Runge–Kutta Heun method:

$$v^{n+1} = \left(1 + iq - \frac{q^2}{2} - \frac{iq^3}{6} \right) v^n =: zv^n, \quad q = \lambda h\alpha, \quad \lambda = k/h.$$

The method is stable for $0 < |q| \leq \sqrt{3}$, because

$$|z|^2 = \left(1 - \frac{q^2}{2} \right)^2 + \left(q - \frac{q^3}{6} \right)^2 = 1 - \frac{q^4}{12} + \frac{q^6}{36}.$$

A simple calculation shows that

$$\max(|z|) \approx 1.372.$$

Therefore, we obtain stability for

$$\lambda \leq 1.26.$$

The classical fourth-order Runge–Kutta method becomes

$$v^{n+1} = \left(1 + iq - \frac{q^2}{2} - \frac{iq^3}{6} + \frac{q^4}{24} \right) v^n =: zv^n.$$

The method is stable for $0 < |q| \leq \sqrt{8}$, because

$$|z|^2 = \left(1 - \frac{q^2}{2} + \frac{q^4}{24}\right)^2 + \left(q - \frac{q^3}{6}\right)^2 = 1 - \frac{q^6}{72} + \frac{q^8}{(24)^2}.$$

Thus, we get the stability condition

$$\lambda \leq 2.06.$$

For any given approximation Q in space, the work required for the fourth-order Runge–Kutta method is $\frac{4}{3}$ times the work required for the third-order method. However, the increase in the stability limit more than offsets the increase in work. Therefore, the fourth-order method is preferable. It is also preferable to the Lax–Wendroff method. The amount of work doubles, but the increase in the stability limit again offsets it. Compared with the leap-frog scheme, the fourth-order Runge–Kutta method requires four times as much work. Therefore, the leap-frog scheme is more efficient if we can use the maximal time step without loss of accuracy. In general, this is impossible, because the method is only second-order accurate in time. However, in many problems with different time scales the accuracy is retained.

The stability condition for the leap-frog time differencing with the fourth-order approximation (6.7.20) is

$$\lambda < 0.72.$$

However, the storage requirement for the leap-frog scheme is twice that of the Runge–Kutta method. Also, the Runge–Kutta method has no spurious solutions that can grow exponentially. Thus, the fourth-order Runge–Kutta method may be preferable in this case.

There are also higher order Runge–Kutta methods. Here, not all $\alpha_j = 0$. Many of them, like the Runge–Kutta Fehlberg or the Nyström methods, are unstable for nondissipative approximations.

We again consider the model equation (6.7.19), but now we approximate it by the dissipative approximation

$$w_t = (\frac{4}{3}D_0(h) - \frac{1}{3}D_0(2h)w + \sigma(-1)^{r-1}h^{2r-1}D'_+D'_-w), \quad \sigma > 0. \quad (6.7.22)$$

Fourier transform gives us, instead of Eq. (6.7.21),

$$\hat{w}_t = \gamma\hat{w}, \quad h\gamma = ih\alpha - 4^r\sigma \sin^{2r} \frac{\xi}{2}. \quad (6.7.23)$$

For the time integration, we use the fourth-order Runge–Kutta method. Theorem

6.7.2 tells us that the approximation is stable for sufficiently small $\lambda = k/h$. In practical applications, one wants to know the largest possible λ . In the theory of difference approximations for ordinary differential equations, one calculates the region Ω of absolute stability. If $\lambda h \gamma \in \Omega$ for all ξ , then the method is stable in our sense. For the fourth-order Runge–Kutta method, Ω is the shaded region in Figure 6.7.1.

6.7.2. Multistep Methods

One can use multistep methods instead of Runge–Kutta methods. In this case, we only need to calculate Qv once at every time step. However, storage requirements increase with the order of accuracy. We have already discussed the leap-frog scheme that is the simplest one. Now, we discuss the explicit

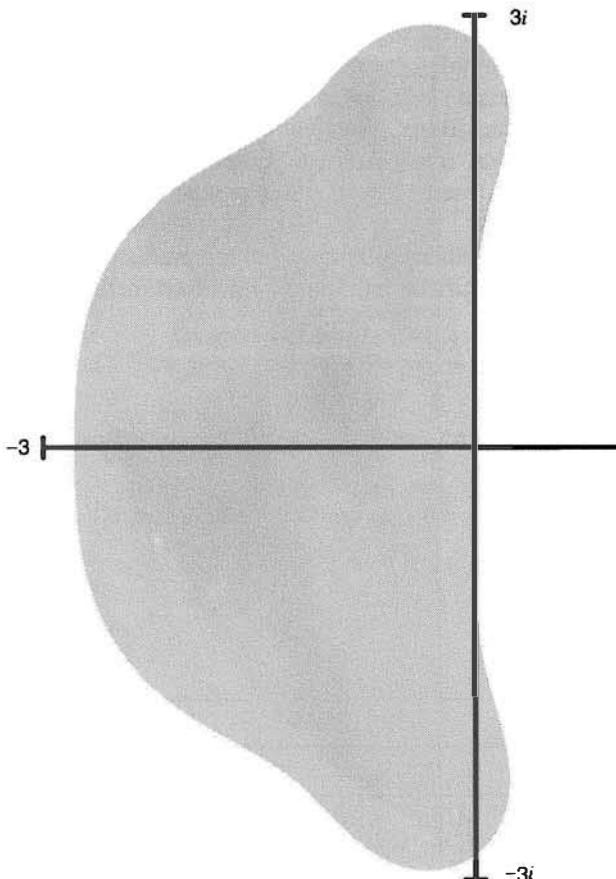


Figure 6.7.1. Stability region for the fourth-order Runge–Kutta method.

Adams–Bashford method

$$\begin{aligned} v^{n+1} &= v^n + kQ \sum_{m=0}^p \gamma_m (kD_{-t})^m v^n, \\ kD_{-t} v^n &= v^n - v^{n-1}, \quad \gamma_m = (-1)^m \int_0^1 \binom{-s}{m} ds, \end{aligned} \quad (6.7.24)$$

and the implicit Adams–Moulton method

$$\begin{aligned} v^{n+1} &= v^n + kQ \sum_{m=0}^p \gamma_m^* (kD_{-t})^m v^{n+1}, \\ \gamma_m^* &= (-1)^m \int_0^1 \binom{1-s}{m} ds. \end{aligned} \quad (6.7.25)$$

They are both accurate of order $p + 1$. Observe, however, that, in the second case, we need to know v only at $\max(p, 1)$ time levels to calculate v^{n+1} . Table 6.7.1 shows the first seven coefficients.

For $p = 0$, the Adams–Bashford method is just the explicit Euler method, and the Adams–Moulton method is the Euler backward method. For $p = 1$, the Adams–Moulton method is the trapezoidal rule, which gives us the Crank–Nicholson method.

Adams–Moulton methods are often used in the predictor–corrector mode. As an example, we use the third-order Adams–Bashford method

$$v_{[0]}^{n+1} = v^n + \frac{kQ}{12} (23v^n - 16v^{n-1} + 5v^{n-2}) \quad (6.7.26a)$$

as a predictor and the fourth-order Adams–Moulton method

$$v^{n+1} = v^n + \frac{kQ}{24} (9v_{[0]}^{n+1} + 19v^n - 5v^{n-1} + v^{n-2}) \quad (6.7.26b)$$

as a corrector.

TABLE 6.7.1.

m							
0	1	2	3	4	5	6	
γ_m	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$
γ_m^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$

For the dissipative operators Q in Eq. (6.7.6), one can prove the following theorem.

Theorem 6.7.5. *Theorem 6.7.2 is also valid for the Adams methods.*

If we use the nondissipative approximation (6.7.20), then we obtain the following theorem.

Theorem 6.7.6. *Use the nondissipative approximation (6.7.20). For sufficiently small λ , the Adams–Bashford methods are stable for $p = 3, 4, 7, 8, \dots$ and the Adams–Moulton methods are stable for $p = 1, 2, 5, 6, \dots$. Otherwise, they are unstable. The predictor–corrector method (6.7.26) is stable.*

The stability regions Ω for the fourth-order accurate Adams–Bashford and Adams–Moulton methods are shown in Figures 6.7.2 and 6.7.3, respectively.

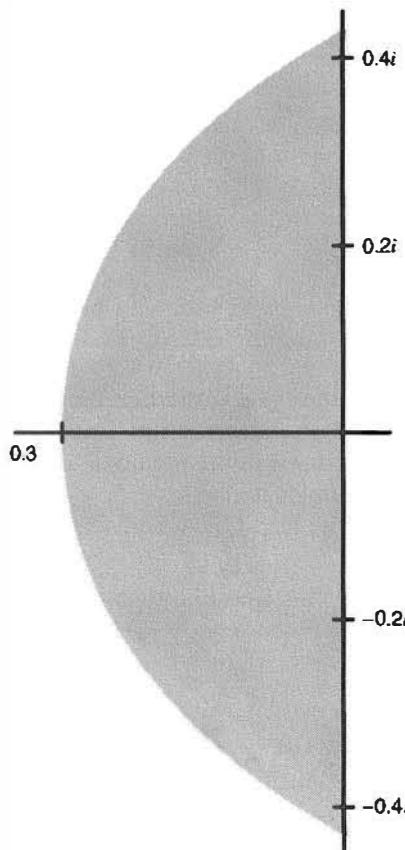


Figure 6.7.2. Stability region for the fourth-order Adams–Bashford method.

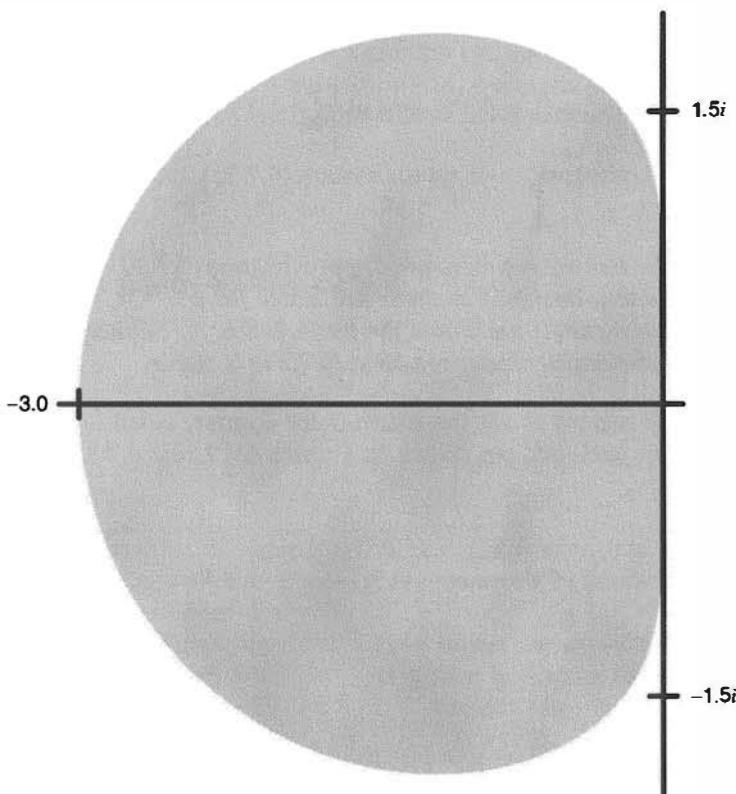


Figure 6.7.3. Stability region for the fourth-order Adams–Moulton method.

For the Adams–Bashford method, the restriction on the step size is so severe that it cannot compete with the fourth-order Runge–Kutta method. If the Adams–Moulton methods are used in the predictor–corrector mode, then they are competitive if one has enough storage.

The Hyman method uses the leap-frog scheme as predictor

$$v_{[0]}^{n+1} = v^{n-1} + 2kQv^n \quad (6.7.27a)$$

combined with the corrector

$$v^{n+1} = \frac{4}{5} v^n + \frac{1}{5} v^{n-1} + \frac{2k}{5} (Qv_{[0]}^{n+1} + 2Qv^n). \quad (6.7.27b)$$

The stability region in Fourier space of the predictor itself is limited to part of the imaginary axis; this region is extended by the corrector step so that the

combined method has a stability region that includes part of the left half-plane and a small part of the right half-plane (see Figure 6.7.4).

This indicates that Q may contain dissipative terms and also that the time differencing introduces dissipation itself, even if Q is one of the centered difference operators (6.7.18). The stability limit for the model equation (6.7.20) is

$$\lambda < 1.09.$$

The order of accuracy in time can be shown to be 3. Again the efficiency is comparable with the classical fourth-order Runge–Kutta method.

We have only discussed difference operators Q for the approximation in space in this section. The results can also be applied to Fourier operators $Q = S$. This was done in Section 3.2, where the leap-frog time differencing and the trapezoidal rule were discussed.

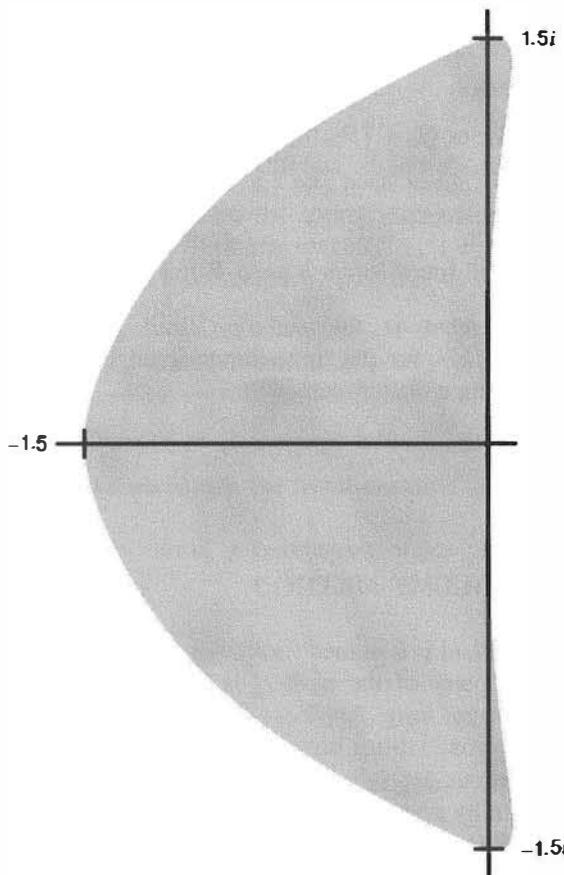


Figure 6.7.4. Stability region for the Hyman method.

EXERCISES

6.7.1. Prove the the Hyman predictor-corrector method (6.7.27) is third-order accurate in time. Derive the stability condition.

6.7.2. Consider a strongly hyperbolic system

$$u_t = \sum_{\nu=1}^d A_\nu u_{x^\nu}, \quad (6.7.28)$$

and the space approximation

$$Q = \sum_{\nu=1}^d A_\nu Q_\nu,$$

where $Q_\nu = D_{0\nu}$ or $Q_\nu = \frac{4}{3}D_{0\nu}(h) - \frac{1}{3}D_{0\nu}(2h)$. Assuming equal step sizes in all directions, derive the stability conditions for

1. the leap-frog scheme,
2. the third- and fourth-order Runge–Kutta methods.

In particular, what is the stability limit expressed in terms of $U, V, W, a, \lambda = k/h$ for the three-dimensional Euler equations and the fourth-order Runge–Kutta method?

6.7.3. Replace Q_ν by the Fourier operator S_ν in Exercise 6.7.2 and derive the stability conditions.

6.8 THE FINITE-VOLUME METHOD

The finite-volume method is designed for systems written in conservation form, and an attractive property of the method is that it automatically produces a conservative semidiscrete form. Another advantage is that it gives a convenient way of designing approximations on an irregular grid.

The main applications are in several space dimensions, but, to explain the main ideas, we begin by considering the one-dimensional case. The method is based on a conservation law

$$u_t + F_x(x, t, u) = 0, \quad (6.8.1)$$

which implies

$$\frac{d}{dt} \int_0^{2\pi} u(x, t) dx = 0. \quad (6.8.2)$$

In the linear case, $F(x, t, u) = A(x, t)u$. We consider nonuniform grids, since the method is strongly associated with such approximations. The x axis is divided into a sequence of finite volumes $\Delta_j = [x_{j-1/2}, x_{j+1/2}]$ with lengths h_j . The vector u_j represents u on Δ_j , and is often thought of as an average over Δ_j . We associate u_j with the value of u at some point x_j in Δ_j (see Figure 6.8.1). It is first assumed that the points x_j (and $x_{j+1/2}$) are generated by a smooth transformation $x_j = x(\xi_j)$, where the points ξ_j are uniformly distributed in the ξ space. This is most often the case in applications.

The conservation law is integrated over the finite volume yielding

$$\frac{d}{dt} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t) dx + F_{j+1/2}(t) - F_{j-1/2}(t) = 0,$$

where the notation $F_j(t) = F(x_j, t, u_j)$ has been used. [We also write F_j for $F_j(t)$.] Because we want to express the approximations in terms of u_j and F_j , the vector $F_{j+1/2}$ is replaced by the average $(F_{j+1} + F_j)/2$, and similarly for $F_{j-1/2}$. In the integral, u is represented by its value at $x = x_j$. In this way, we obtain the *centered finite volume method*

$$\frac{dv_j}{dt} + \frac{1}{2h_j} (F_{j+1} - F_{j-1}) = 0. \quad (6.8.3)$$

On a uniform mesh, we recognize the usual centered semidiscrete difference approximation.

Disregarding boundary terms, a summation over j yields the conservation property

$$\frac{d}{dt} \sum_j v_j h_j = 0,$$

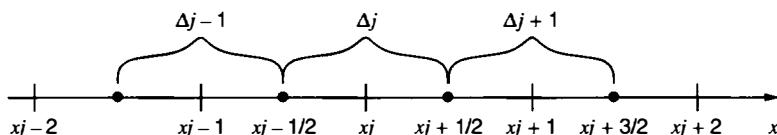


Figure 6.8.1. Finite volumes Δ_j and gridpoints x_j .

which is the discrete analogue of Eq. (6.8.2). Let h be a constant given step size, for example, the uniform spacing in ξ -space, and let $a_j = h_j/h$. Then Eq. (6.8.3) takes the form

$$\frac{dv_j}{dt} + \frac{1}{a_j} D_0 F_j = 0.$$

With $F = Au$, where A is a constant matrix, we get

$$\frac{dv_j}{dt} + \tilde{A}_j D_0 v_j = 0, \quad \tilde{A}_j = \frac{A}{a_j}, \quad (6.8.4)$$

which can be viewed as a linear approximation with variable coefficients. If the coefficient matrix \tilde{A}_j is Lipschitz continuous, that is,

$$|\tilde{A}_{j+1} - \tilde{A}_j| \leq Ch, \quad (6.8.5)$$

then the operator $\tilde{A}_j D_0$ is semibounded [cf. Eq. (5.3.16)], which implies that Eq. (6.8.4) is stable. The condition (6.8.5) is equivalent to

$$h|a_{j+1} - a_j| = |h_{j+1} - h_j| \leq C_1 h^2,$$

or

$$|x_{j+3/2} - 2x_{j+1/2} + x_{j-1/2}| \leq C_1 h^2. \quad (6.8.6)$$

But this inequality follows immediately from the assumption that the gridpoints $x_{j+1/2}$ are generated from a smooth transformation of the ξ space, which has a uniform grid.

The second-order accuracy of Eq. (6.8.3) also follows from this assumption. The differential equation (6.8.1) takes the form

$$\frac{\partial \tilde{u}}{\partial t} + \frac{d\xi}{dx} \frac{\partial F}{\partial \xi} = 0, \quad (6.8.7)$$

where $\tilde{u}(\xi, t) = u(x(\xi), t)$. A second-order approximation of Eq. (6.8.7) is obviously

$$\frac{d\tilde{v}_j}{dt} + b_j D_0 F_j = 0, \quad (6.8.8)$$

where $b(x) = d\xi/dx$, $\tilde{v}_j = \tilde{v}(\xi_j, t)$, $F_j = F(x(\xi_j), t, \tilde{v}(\xi_j, t))$. With $F = Av$, Eq.

(6.8.8) can as well be considered as the approximation (6.8.4) in physical space, with a_j replaced by $1/b_j$. Therefore, second-order accuracy is established for Eq. (6.8.4) if

$$a_j = \frac{1}{b_j} + \mathcal{O}(h^2) = \left(\frac{dx}{d\xi} \right)_{\xi=\xi_j} + \mathcal{O}(h^2).$$

This follows immediately from

$$ha_j = x_{j+1/2} - x_{j-1/2} = x(\xi_{j+1/2}) - x(\xi_{j-1/2}) = h \left(\frac{dx}{d\xi} \right)_{\xi=\xi_j} + \mathcal{O}(h^3).$$

For an actual implementation it is sufficient to store either the sequence $\{x_{j+1/2}\}_{j=1}^N$ or the sequence $\{x_j\}_{j=1}^N$. In the latter case, Eq. (6.8.3) is modified as

$$\frac{dv_j}{dt} + \frac{F_{j+1} - F_{j-1}}{x_{j+1} - x_{j-1}} = 0. \quad (6.8.9)$$

Since $x_{j+1} - x_{j-1} = 2h(dx/d\xi)_{\xi=\xi_j} + \mathcal{O}(h^3)$, this is still second-order accurate. Actually, the latter formulation is preferable, because contrary to Eq. (6.8.3), it is consistent if the grid $\{x_j\}$ is not generated by a smooth transformation from a uniform grid.

Also note that x_j is, in general, not the midpoint of the interval Δ_j . However, after a completed calculation with Eq. (6.8.3), where the grid $\{x_{j+1/2}\}$ has been used, the v_j values can be plotted at the centerpoints $(x_{j+1/2} + x_{j-1/2})/2$. Because

$$\frac{x_{j+1/2} + x_{j-1/2}}{2} = x_j + \mathcal{O}(h^2),$$

we have

$$u\left(\frac{x_{j+1/2} + x_{j-1/2}}{2}, t\right) = u(x_j, t) + \mathcal{O}(h^2),$$

showing that second-order accuracy is preserved with such a procedure. [Of course, this does not hold for a nonsmooth transformation $x(\xi)$.]

The method (6.8.3) is sometimes modified as

$$\frac{dv_j}{dt} + \frac{1}{h_j} \left(F\left(x_{j+1/2}, t, \frac{v_{j+1} + v_j}{2}\right) - F\left(x_{j-1/2}, t, \frac{v_j + v_{j-1}}{2}\right) \right) = 0,$$

which has essentially the same stability and accuracy properties.

Next, we consider the two-dimensional case for a system in conservation law form

$$u_t + F_x(x, y, t, u) + G_y(x, y, t, u) = 0. \quad (6.8.10)$$

The mesh consists of quadrilaterals, but these are not necessarily rectangles. It is assumed that the mesh has been obtained by a smooth transformation of a uniform rectangular grid such that neighboring quadrilaterals do not differ too much from each other in shape and size. In general, the new coordinate lines are not straight lines, but locally they are considered as such, which is consistent with the second-order accurate methods we consider. Figure 6.8.2 shows the mesh with the finite volume Δ in the middle with corners A, B, C, D . Green's formula yields

$$\iint_{\Delta} u_t dx dy + \int_{\partial\Delta} (F(x, y, t, u) dy - G(x, y, t, u) dx) = 0 \quad (6.8.11)$$

after integrating Eq. (6.8.10) over Δ and where $\partial\Delta$ is the boundary of Δ . The

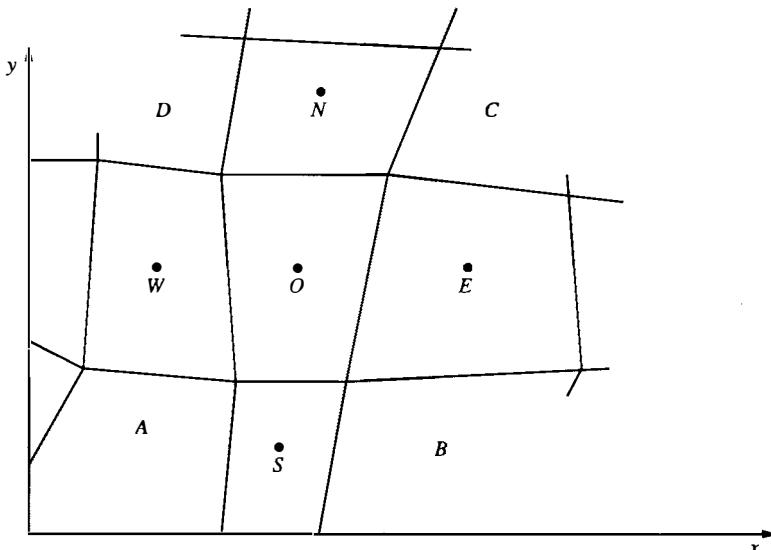


Figure 6.8.2. Finite volume mesh.

notation u_O can be used for the vector function value at the point O (the image point of the center of the rectangle in the corresponding uniform grid), or for the average of u over Δ . It makes no difference for the method to be discussed here, and we use the first definition. The coordinates at the point A are denoted by (x_A, y_A) , and so on, and we use the notation F_A for $F(x_A, y_A, t, u(x_A, y_A, t))$, and so on. The line integrals between A and B are now approximated by the average of the vector values at O and S multiplied by the proper distance, and the other ones analogously. We get

$$\begin{aligned} \int_{\partial\Delta} F dy &\approx (y_B - y_A) \frac{F_O + F_S}{2} + (y_C - y_B) \frac{F_O + F_E}{2} \\ &\quad + (y_D - y_C) \frac{F_O + F_N}{2} + (y_A - y_D) \frac{F_O + F_W}{2}, \\ \int_{\partial\Delta} G dx &\approx (x_B - x_A) \frac{G_O + G_S}{2} + (x_C - x_B) \frac{G_O + G_E}{2} \\ &\quad + (x_D - x_C) \frac{G_O + G_N}{2} + (x_A - x_D) \frac{G_O + G_W}{2}. \end{aligned}$$

With $\text{vol}(\Delta)$ denoting the area of Δ , the integral in Eq. (6.8.11) is represented by

$$\iint_{\Delta} u_t dx dy = \text{vol}(\Delta) \frac{du_0}{dt}.$$

In this way we get the *centered finite volume method*

$$\begin{aligned} 2\text{vol}(\Delta) \frac{du_O}{dt} &+ (y_B - y_A)(F_O + F_S) + (y_C - y_B)(F_O + F_E) \\ &+ (y_D - y_C)(F_O + F_N) + (y_A - y_D)(F_O + F_W) \\ &- (x_B - x_A)(G_O + G_S) - (x_C - x_B)(G_O + G_E) \\ &- (x_D - x_C)(G_O + G_N) - (x_A - x_D)(G_O + G_W) = 0. \end{aligned} \tag{6.8.12}$$

When summed over all volumes Δ , all of the F and G terms cancel except for boundary terms, establishing the conservative property in the discrete case. If the mesh is rectangular, we have

$$y_B - y_A = y_D - y_C = x_C - x_B = x_A - x_D = 0.$$

By the usual notation $(x_O, y_O) = (x_i, y_j)$,

$$\Delta x_i = x_B - x_A, \quad \Delta y_j = y_C - y_B,$$

and the finite volume method reduces to

$$\frac{dv_{ij}}{dt} + \frac{F_{i+1,j} - F_{i-1,j}}{2\Delta x_i} + \frac{G_{i,j+1} - G_{i,j-1}}{2\Delta y_j} = 0. \quad (6.8.13)$$

When designing finite volume methods with the v_{ij} representing point values, it seems more natural to integrate over the quadrilateral Δ with corners at the gridpoints according to Figure 6.8.3.

The line integrals are now approximated by the trapezoidal rule using the vector values at the corners, yielding

$$\begin{aligned} 2 \int_{\partial\Delta} F dy &= (y_B - y_A)(F_A + F_B) + (y_C - y_B)(F_B + F_C), \\ &\quad + (y_D - y_C)(F_C + F_D) + (y_A - y_D)(F_D + F_A), \\ &= (y_C - y_A)(F_B - F_D) - (y_B - y_D)(F_C - F_A), \end{aligned}$$

and analogously for $\int_{\partial\Delta} G dx$. The integral $\iint_{\Delta} u_t dx dy$ cannot be approximated with second-order accuracy by using only one gridpoint. For this reason, a second-order fully discrete method must be implicit. When using equal weights for all gridpoints, we get the *box finite volume method*

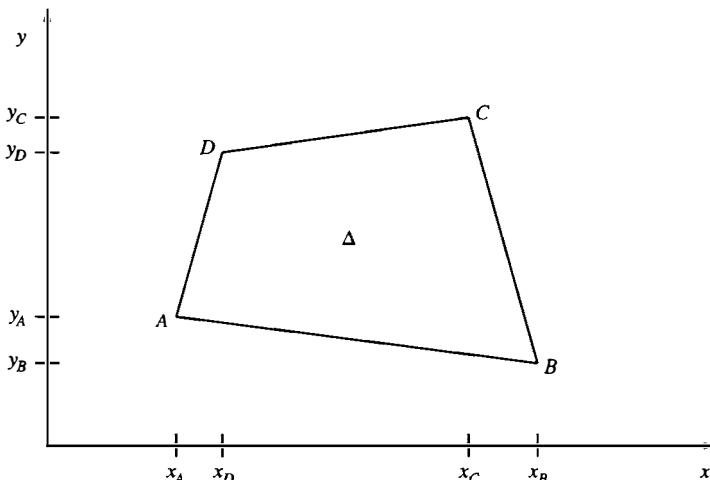


Figure 6.8.3. Finite volume Δ with corners at the gridpoints.

$$\frac{\text{vol}(\Delta)}{2} \frac{d}{dt} (v_A + v_B + v_C + v_D) \\ + (y_C - y_A)(F_B - F_D) + (y_D - y_B)(F_C - F_A) \\ + (x_A - x_C)(G_B - G_D) + (x_B - x_D)(G_C - G_A) = 0. \quad (6.8.14)$$

On a rectangular grid this method reduces to

$$\frac{1}{2} \frac{d}{dt} (I + E_x)(I + E_y)v + (I + E_y)D_{+x}F \\ + (I + E_x)D_{+y}G = 0, \quad (6.8.15)$$

which is the regular box difference scheme. Regardless of time discretization, Eq. (6.8.14) leads to an implicit scheme, hence it is not in common use. Furthermore, dissipation terms are often included in finite-volume methods, and in this case compactness is lost.

For discretization in time of the finite-volume methods we use the results of Section 6.7. The finite-volume method is based on a conservation law form of the differential equation, and, as we have demonstrated, nonlinearity is no hindrance. For application to the Euler equations, we need to rewrite Eq. (4.6.4). The conservation law form of the Euler equations is

$$\frac{\partial \phi}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z} = 0, \quad (6.8.16)$$

where

$$\phi = \begin{bmatrix} \rho u \\ \rho v \\ \rho w \\ \rho \end{bmatrix}, \quad F = \begin{bmatrix} p + \rho u^2 \\ \rho uv \\ \rho uw \\ \rho u \end{bmatrix}, \\ G = \begin{bmatrix} \rho uv \\ p + \rho v^2 \\ \rho vw \\ \rho v \end{bmatrix}, \quad H = \begin{bmatrix} \rho uw \\ \rho vw \\ p + \rho w^2 \\ \rho w \end{bmatrix}.$$

These equations are complemented by the equation of state $p = p(\rho)$. For many types of flow there is no such simple relation between pressure and density. Instead one introduces the internal energy e as a new dependent variable and use an equation of state of the type $p = p(\rho, e)$. This requires a new differential equation, which is

$$\frac{\partial E}{\partial t} + \frac{\partial}{\partial x} (uE + up) + \frac{\partial}{\partial y} (vE + vp) + \frac{\partial}{\partial z} (wE + wp) = 0, \quad (6.8.17)$$

where

$$E = \rho e + \frac{\rho}{2} (u^2 + v^2 + w^2).$$

With E as the fifth component of the vector ϕ we have an extended and more general conservation law of the type in Eq. (6.8.16).

EXERCISES

- 6.8.1.** Verify that the conservation law (6.8.16) is equivalent to the system (4.6.4).
- 6.8.2.** Derive the linearized form

$$\phi_t + A\phi_x + B\phi_y + C\phi_z = 0$$

of Eq. (6.8.16) by introducing a small perturbation around $\phi = \Phi$. Prove that the eigenvalues of A , B , and C are identical to those of the corresponding coefficient matrices for the original linear Euler equations.

- 6.8.3.** Use the results of Exercise 6.8.2 to derive the stability limit for the finite volume method (6.8.13) applied to the Euler equations with a fourth-order Runge–Kutta time discretization.

6.9. THE FOURIER METHOD

When applying the Fourier method to symmetric systems with constant coefficients there is no difficulty to verify stability (Exercise 6.7.3). This is because the S operator is skew-symmetric. However, in general, it cannot be expected that a straightforward implementation of the Fourier method will be stable for systems with variable coefficients. The energy method used for difference methods is based on the fact that the difference operator Q approximating $\partial/\partial x$ almost commutes with a variable coefficient $A(x)$; that is, $QA(x) = A(x)Q + \mathcal{O}(1)$. This is generally not true for the Fourier differentiation operator S , even if $A(x)$ is Lipschitz continuous, and we must modify the method to ensure that it is stable.

Corresponding to the concept of dissipativity introduced for difference methods, we introduce a smoothing operator H for the Fourier method. The basic idea is to modify the coefficients for higher wave numbers in each step so that the oscillatory components are damped. At the same time, we do not want to adversely affect the accuracy significantly, because high accuracy is one of the essential features of the Fourier method. We only consider problems in one space dimension here.

As usual, we work with $N + 1$ gridpoints in the interval $[0, 2\pi)$, N even, and for convenience we introduce the notation $M = N/2$. Let T_M denote the space of trigonometric polynomials with basis $\{e^{i\omega x}/\sqrt{2\pi}\}_{|\omega| \leq M}$. We divide $u \in T_M$ into two parts, u_1 and u_2 , corresponding to low and high wave numbers

$$\begin{aligned} u_1 &= \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq M_1} \hat{u}(\omega) e^{i\omega x}, \quad 0 < M_1 < M, \\ u_2 &= u - u_1. \end{aligned} \tag{6.9.1}$$

We want to construct a smoothing operator $H = H(l, M_1, d)$ such that u_1 is not affected, and such that u_2 is affected only if $\hat{u}(\omega)$ is large. We take

$$v = Hu = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq M} \hat{v}(\omega) e^{i\omega x}, \tag{6.9.2a}$$

where

$$\hat{v}(\omega) = \begin{cases} \hat{u}(\omega), & \text{if } |\omega| \leq M_1 \text{ or } |\hat{u}(\omega)| \leq \frac{d\|u_1\|}{|\omega|^l}, \\ \frac{d\|u_1\|}{|\omega|^l} \text{ sign}(\hat{u}(\omega)), & \text{otherwise.} \end{cases} \tag{6.9.2b}$$

Here l is a natural number and d is a constant. Obviously, we have $Hu_1 = u_1$ and $\|Hu\| \leq \|u\|$.

The following lemma shows that the smoothing operator does not affect a large class of functions.

Lemma 6.9.1. *Let $u \in T_M$ and define the smoothing operator $H(l, M_1, d)$ by Eq. (6.9.2). Then $H(l, M_1, d)u = u$ if*

$$\left\| \frac{d^l u}{d^l} \right\| \leq \gamma \|u\|, \tag{6.9.3}$$

where

$$\gamma \leq \frac{1}{\sqrt{2}} \min(d, M_1^l). \tag{6.9.4}$$

Proof. The function is split into two parts u_1 and u_2 as in Eq. (6.9.1), and we have

$$\|u\|^2 = \|u_1\|^2 + \|u_2\|^2.$$

The condition (6.9.3) implies

$$M_1^{2l} \|u_2\|^2 \leq \left\| \frac{d^l u_2}{du^l} \right\|^2 \leq \left\| \frac{d^l u}{du^l} \right\|^2 \leq \gamma^2 \|u\|^2 = \gamma^2 (\|u_1\|^2 + \|u_2\|^2),$$

and from Eq. (6.9.4) we obtain

$$\|u_2\|^2 \leq \frac{1}{M_1^{2l} - \gamma^2} \|u_1\|^2 \leq \|u_1\|^2.$$

Thus, if $\omega \neq 0$,

$$\omega^{2l} |\hat{u}(\omega)|^2 \leq \left\| \frac{d^l u}{du^l} \right\|^2 \leq \gamma^2 \|u\|^2 = \gamma^2 (\|u_1\|^2 + \|u_2\|^2) \leq 2\gamma^2 \|u_1\|^2,$$

which yields

$$|\hat{u}(\omega)| \leq \frac{\gamma\sqrt{2}}{|\omega|^l} \|u_1\| \leq \frac{d\|u_1\|}{|\omega|^l}.$$

Therefore by Eq. (6.9.2b), the coefficients $\hat{u}(\omega)$ are unaffected by H by definition, and the lemma is proved.

Now consider the differential equation

$$u_t = A(x, t)u_x, \quad (6.9.5)$$

where $A(x, t)$ is a Hermitian matrix. Integration by parts gives us

$$\frac{d}{dt} \|u\|^2 = -\operatorname{Re}(u, A_x u), \quad (6.9.6)$$

and therefore we obtain an energy estimate.

Denoting by \mathbf{w} the gridfunction written as a vector with $N+1$ components, we can write the semidiscrete modified Fourier method for Eq. (6.9.5) as

$$\mathbf{w}_t = \mathbf{A} H S \mathbf{w}. \quad (6.9.7)$$

Here S is the Fourier operator for differentiation defined in Section 1.4. \mathbf{A} is the $Nm \times Nm$ matrix with $A(x_j, t)$ as blocks on the diagonal. The implementation of the operation HS can be done using only two FFTs. The S operator is factorized as $S = T^{-1}DT$, where T represents the FFT and where D represents the multiplication by $i\omega$. If \hat{H} denotes the operation $\hat{v}(\omega) = \hat{H}\hat{u}(\omega)$, where $\hat{v}(\omega)$ is defined by Eq. (6.9.2), then Eq. (6.9.7) becomes

$$\mathbf{w}_t = \mathbf{AT}^{-1}\hat{H}DT\mathbf{w}. \quad (6.9.8)$$

The approximation (6.9.8) can be considered as an evolutionary equation in the space T_M . Let $w = Tw$ and $A_N = T\mathbf{A}$ denote the Fourier interpolant of w and \mathbf{A} respectively, and define A_{N^*} by

$$A_{N^*}u = T\mathbf{A}T^{-1}u.$$

Then the method (6.9.8) can be written in the form

$$w_t = A_N * Hw_x, \quad w \in T_M. \quad (6.9.9)$$

Let $\tilde{A}_{N,\omega}$ be the Fourier coefficient of A_N , and assume that

$$|\tilde{A}_{N,\omega}| \leq \frac{C(\beta)}{1 + |\omega|^\beta}, \quad \beta > 1, \quad (6.9.10)$$

where the constants C and β are independent of N . w in Eq. (6.9.9) is divided into two parts, w_1 and w_2 , as defined by Eq. (6.9.1). Without proof, we state the following theorem.

Theorem 6.9.1. *Solutions of the approximation (6.9.9) [or equivalently (6.9.8)] satisfy the estimate*

$$\frac{d}{dt} \|w\|^2 \leq -\operatorname{Re} \left(w, \frac{dA_N}{dx} * w \right) + (C_1(\beta)M^{2-\beta} + C_2(d)M_1^{2-l})\|w\|^2. \quad (6.9.11)$$

Here β is defined in Eq. (6.9.10) and l and d are defined in the definition (6.9.2) of the smoothing operator H . If $l > 2, \beta > 2$, then the estimate (6.9.11) converges to the estimate (6.9.6) for the differential equation as $M \rightarrow \infty$.

REMARK. The restriction of w to the grid gives us \mathbf{w} with norm $\|w\|_h$. Thus, Eq. (6.9.11) represents also a stability estimate for the solution of Eq. (6.9.8).

We will derive an error estimate for the method (6.9.8). As with difference methods, the error estimate is easily obtained for smooth solutions once stability is established. This is because the truncation error can be calculated in a straightforward manner, and then Duhamel's principle can be applied. We have the following theorem.

Theorem 6.9.2. *Assume that the coefficients and data of Eq. (6.9.5) are C^∞ -smooth. If M_1 is sufficiently large and $l \geq 2$ then, on every finite time interval $[0, T]$, there are constants C_p such that for any p*

$$\sup_{0 \leq t \leq T} \|u(t) - w(t)\|_h \leq C_p h^p, \quad h \leq h_0. \quad (6.9.12)$$

(This is often called spectral accuracy.)

Proof. For each t , the true solution $u(x, t)$ is divided into two parts

$$u(x, t) = u_N(x, t) + u_R(x, t), \quad (6.9.13)$$

where

$$u_N(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| \leq N} \hat{u}(\omega, t) e^{i\omega x}, \quad (6.9.14a)$$

$$u_R(x, t) = \frac{1}{\sqrt{2\pi}} \sum_{|\omega| > N} \hat{u}(\omega, t) e^{i\omega x}. \quad (6.9.14b)$$

We consider now the solution in a fixed time interval $0 \leq t \leq T$. Since $u(x, t)$ is C^∞ -smooth and we can derive estimates of u and its derivatives in terms of the coefficients and data it follows that

$$Hu_N = u_N$$

provided M_1 is sufficiently large. Also there are constants K_p such that

$$\sup_{0 \leq t \leq T} \|u_{Rt} - AHu_{Rx}\| \leq K_p N^{-p}, \quad \text{for } p = 1, 2, \dots \quad (6.9.15)$$

We shall now determine the truncation error

$$0 = u_t - Au_x = u_{Nt} - Au_{Nx} + u_{Rt} - Au_{Rx} = u_{Nt} - AHu_{Nx} + \varphi,$$

$$\varphi := u_{Rt} - Au_{Rx}.$$

Therefore we have on the grid

$$\mathbf{u}_{Nt} = \mathbf{AHS}\mathbf{u}_N + \varphi.$$

By Eq. (6.9.15) the error estimate follows from the stability estimate. Finally, we want to point out that the smoothing process is not needed for stability reasons if we write the differential equation in the form

$$\frac{\partial u}{\partial t} = \frac{1}{2} (Au_x + (Au)_x) - A_x u \quad (6.9.16)$$

and approximate it by

$$\mathbf{w}_t = S_1 \mathbf{w} - \mathbf{A}_x \mathbf{w}, \quad S_1 := \frac{1}{2} (\mathbf{AS} + \mathbf{SA}). \quad (6.9.17)$$

Since S is anti-Hermitian and \mathbf{A} is Hermitian, the operator S_1 is also anti-Hermitian, and we obtain an energy estimate.

EXERCISES

- 6.9.1.** Write a program for the Fourier method (6.9.7) for the scalar equation $u_t = a(x, t)u_x$. Carry out numerical experiments, and compare the accuracy by varying the parameters l, M_1 and d in the smoothing operator H .
- 6.9.2.** Use the form (6.9.16) with $A = a(x, t)$, and write a program for the Fourier method of type (6.9.17). Compare the performance with the method in Exercise 6.9.1.

BIBLIOGRAPHIC NOTES

Proofs for Theorems 6.2.3, 6.4.1, and 6.4.2 can be found in Nirenberg (1972). For approximations of the Lax–Wendroff type, separate stability investigations have been performed that do not use the theorems given in Section 6.5. Lax and Wendroff (1964) proved the weaker stability limit

$$\lambda \leq \frac{1}{2\sqrt{2}} \min\left(\frac{1}{\rho(A)}, \frac{1}{\rho(B)}\right) \quad (6.N.1)$$

for the approximation (6.5.23). The modified Lax–Wendroff method

$$v^{n+1} = \left[I + k(AD_{0x} + BD_{0y}) + \frac{k^2}{2} (A^2 D_{+x} D_{-x} + (AB + BA) D_{0x} D_{0y} + B^2 D_{+y} D_{-y}) + \frac{h^4}{8} (A^2 + B^2) D_{+x} D_{-x} D_{+y} D_{-y} \right] v^n, \quad (6.N.2)$$

was given by Lax and Wendroff (1962). The addition of the last term may well pay off, because the stability condition is relaxed. For Hermitian matrices A and B the stability condition is

$$\lambda^2(A^2 + B^2) \leq \frac{1}{2} I. \quad (6.N.3)$$

This scheme, as well as (6.5.23), is dissipative if the matrices A and B are nonsingular, but the latter condition is not required for stability of (6.N.2).

Another method of the Lax–Wendroff type is the MacCormack (1969) method. Several versions of this method have been developed, but the most popular one is the time-split method, (MacCormack and Pauly, 1972). For the conservation law $u_t = F_x + G_y$, we define the one-dimensional operators

$$\begin{aligned} u^* &= u^n + kD_{-x}F^n, \\ u^{**} &= u^* + kD_{+x}F^*, \\ Q_1(k)u^n &= \frac{1}{2}(u^n + u^{**}), \\ u^+ &= u^n + kD_{-y}G^n, \\ u^{++} &= u^+ + kD_{+y}G^+, \\ Q_2(k)u^n &= \frac{1}{2}(u^n + u^{++}). \end{aligned}$$

The Strang type splitting for two-dimensional problems yields

$$u^{n+2} = Q_1(k)Q_2(k)Q_2(k)Q_1(k)u^n. \quad (6.N.4)$$

Stability conditions have been given for the Euler equations. These conditions reduce to the one-dimensional ones:

$$k(|U| + a) \leq h_1, \quad k(|V| + a) \leq h_2, \quad (6.N.5)$$

where h_1 and h_2 are the stepsizes in the x and y direction, respectively.

One-step schemes have many advantages for obvious reasons, but it is more complicated to achieve higher order accuracy. Jeltsch and Smit (1985) obtained

general conditions on the number of points required for general one-step methods. In particular, they proved that the order of accuracy q for any stable one-step “upwind” scheme

$$v_j^{n+1} = \sum_{l=0}^r \alpha_l E^l v_j^n$$

is limited by

$$q \leq \min(r, 2). \quad (6.N.6)$$

This shows that any linear upwind method with order of accuracy three or higher is unstable.

Theorems 6.5.1, 6.5.2, 6.6.1, and 6.6.2 were proved by Kreiss (1964) for t -independent coefficients. The proof of Theorem 6.5.1 presented here was given by Parlett (1966). The proofs can also be found in Richtmyer and Morton (1967).

In Section 6.7, several methods for time discretization were discussed. These methods and several others are found in the literature on ordinary differential equations, for example, in Hairer, Nørsett, and Wanner (1980) or Gear (1971).

A thorough discussion of Runge–Kutta methods, including a software package for selecting optimal parameters for application to PDEs, was given by Otto (1988). The Hyman method was presented by Hyman (1979). The proof of Theorem 6.9.1 was given by Kreiss and Oliger (1979). Further stability results for the Fourier method are given in Tadmor (1986).

7

PARABOLIC EQUATIONS AND NUMERICAL METHODS

7.1 GENERAL PARABOLIC SYSTEMS

We have already treated a few parabolic model examples in Chapter 2. In Chapter 4, we defined strongly parabolic systems of second order and showed that they give rise to semibounded operators that lead to well-posed problems. In this section, we give a brief summary of the parabolicity definitions and main results for general systems.

First, consider one-dimensional second order systems

$$u_t = (Au_x)_x + Bu_x + Cu. \quad (7.1.1)$$

The general definition of parabolicity is now given.

Definition 7.1.1. *The system (7.1.1) is called **parabolic** if, for every fixed x_0 and t_0 , the eigenvalues λ_j of $A(x_0, t_0)$ satisfy*

$$\operatorname{Re} \lambda_j \geq \delta > 0. \quad (7.1.2)$$

We then have the following theorem.

Theorem 7.1.1. *If the system (7.1.1) is **parabolic**, then the initial value problem is **well posed**.*

Proof. By Lemma A.1.10 there is a matrix T such that

$$T^{-1}AT + (T^{-1}AT)^* \geq \delta I. \quad (7.1.3)$$

Introduce a new variable by $v = T^{-1}u$. Then the coefficient matrix of the prin-

cipal part of the resulting system is $T^{-1}AT$, which satisfies Eq. (7.1.3). The theorem then follows as shown in Section 4.6.

We now consider higher order parabolic equations. We begin with a scalar equation

$$\begin{aligned} u_t &= -(au_{xx})_{xx} + (bu_x)_{xx} + cu_{xx} \\ &\quad + du_x + eu + F. \end{aligned} \tag{7.1.4}$$

Equation (7.1.4) is called parabolic if $a = a(x, t)$ satisfies

$$\operatorname{Re} a \geq \delta > 0. \tag{7.1.5}$$

The initial-value problem is well-posed. To show this we need the following inequality.

Lemma 7.1.1. *For any constant $\tau > 0$,*

$$\|u_x\|^2 \leq \tau \|u_{xx}\|^2 + \frac{1}{4\tau} \|u\|^2. \tag{7.1.6}$$

Proof. Integration by parts gives us

$$\begin{aligned} \|u_x\|^2 &= (u_x, u_x) = |(u, u_{xx})| \\ &\leq \tau \|u_{xx}\|^2 + \frac{1}{4\tau} \|u\|^2. \end{aligned}$$

Let u be a smooth solution of Eq. (7.1.4). Then, we obtain

$$\begin{aligned} \frac{d}{dt} \|u\|^2 &= 2 \operatorname{Re}(-(u, (au_{xx})_{xx}) + (u, (bu_x)_{xx}) \\ &\quad + (u, cu_{xx}) + (u, du_x) + (u, eu) + (u, F)). \end{aligned}$$

Let $\tau > 0$ and $\mu > 0$ be constants, $b_0 = \max_{x,t} |b(x, t)|$, and define c_0 , d_0 , and e_0 correspondingly. Integration by parts gives us

$$\begin{aligned}
\operatorname{Re}(-(u, (au_{xx})_{xx})) &= \operatorname{Re}(-(u_{xx}, au_{xx})) \leq -\delta \|u_{xx}\|^2, \\
|(u, (bu_x)_{xx})| &= |(bu_{xx}, u_x)| \leq \frac{1}{4\mu} \|u_x\|^2 + b_0^2 \mu \|u_{xx}\|^2 \\
&\leq \frac{1}{16\mu\tau} \|u\|^2 + \left(\frac{\tau}{4\mu} + b_0^2 \mu \right) \|u_{xx}\|^2, \\
|(u, cu_{xx})| &\leq c_0 \left(\mu \|u_{xx}\|^2 + \frac{1}{4\mu} \|u\|^2 \right), \\
|(u, du_x)| &\leq \tau d_0^2 \|u_x\|^2 + \frac{1}{4\tau} \|u\|^2 \leq \tau d_0^2 \|u_{xx}\|^2 \\
&\quad + \left(\frac{1}{4\tau} + \frac{\tau d_0^2}{4} \right) \|u\|^2, \\
|(u, eu)| &\leq e_0 \|u\|^2, \\
|(u, F)| &\leq \frac{1}{2} \|u\|^2 + \frac{1}{2} \|F\|^2.
\end{aligned}$$

Thus, we obtain

$$\frac{d}{dt} \|u\|^2 \leq \alpha \|u_{xx}\|^2 + \beta \|u\|^2 + \|F\|^2,$$

where $\alpha = 2(-\delta + (\tau/4\mu) + b_0^2 \mu + c_0 \mu + \tau d_0^2)$ and β is a constant which does not depend on u . With $\tau = 2\mu\delta$ we obtain

$$\alpha = -\delta + (b_0^2 + c_0 + 2d_0^2 \delta) \mu,$$

and, by choosing μ small enough, α becomes negative. Then Lemma 4.3.1 gives us the estimate.

We have here shown that the desired estimate holds. To prove existence, we can construct a difference approximation satisfying the corresponding estimates. The limit process as $h \rightarrow 0$ then gives us existence.

Systems of the form

$$u_t = -(Au_{xx})_{xx} + P_3(x, t, \partial/\partial x)u,$$

where P_3 is a general third-order differential operator, are called parabolic if the eigenvalues of A satisfy Eq. (7.1.2). The proof that the initial value problem is well posed follows as before.

Now we consider systems in more than one space dimension. The simplest parabolic problems are those where no mixed derivative terms appear. In Section 4.6, it was shown that the initial value problem is well posed for systems

$$u_t = (Au_x)_x + (Bu_y)_y + P_1(x, t, \partial/\partial x)u + F, \quad (7.1.7)$$

provided that

$$A + A^* \geq \delta I, \quad B + B^* \geq \delta I.$$

Here P_1 is a general first-order operator. If mixed derivative terms appear, the estimates can become more technically difficult. We start with an equation with real constant coefficients

$$u_t = au_{xx} + bu_{xy} + cu_{yy}. \quad (7.1.8)$$

It is called parabolic if there is a constant $\delta > 0$ such that, for all real ω_1 and ω_2 ,

$$a\omega_1^2 + b\omega_1\omega_2 + c\omega_2^2 \geq \delta(\omega_1^2 + \omega_2^2). \quad (7.1.9)$$

In this case, the amplitudes of the large wave numbers are damped.

Now consider an equation with variable coefficients

$$u_t = a(x, y, t)u_{xx} + b(x, y, t)u_{xy} + c(x, y, t)u_{yy} + P_1u + F, \quad (7.1.10)$$

where P_1 is a general first order operator. We have the following definition.

Definition 7.1.2. *Equation (7.1.10) is called parabolic if Eq. (7.1.9) holds for every x_0, y_0 , and t_0 .*

For systems

$$u_t = A(x, y, t)u_{xx} + B(x, y, t)u_{xy} + C(x, y, t)u_{yy} + P_1u + F \quad (7.1.11)$$

we have this definition.

Definition 7.1.3. *Equation (7.1.11) is called parabolic if, for all real ω_1 and ω_2 and all x_0, y_0 , and t_0 , there is a constant $\delta > 0$ such that the eigenvalues λ_j of*

$$A(x_0, y_0, t_0)\omega_1^2 + B(x_0, y_0, t_0)\omega_1\omega_2 + C(x_0, y_0, t_0)\omega_2^2$$

satisfy the inequality

$$\operatorname{Re} \lambda_j \geq \delta(\omega_1^2 + \omega_2^2). \quad (7.1.12)$$

Finally, we define a general parabolic system of order $2r$ in d space dimensions.

Definition 7.1.4. Consider a system of order $2r$ given by

$$\frac{\partial u}{\partial t} = \sum_{j=0}^{2r} P_j(x, t, \partial x)u + F, \quad (7.1.13)$$

where

$$P_j(x, t, \partial x) = \sum_{\nu_1 + \dots + \nu_d = j} A_{\nu_1 \dots \nu_d}(x, t) \frac{\partial^j}{\partial x^{(1)\nu_1} \dots \partial x^{(d)\nu_d}}$$

are general homogeneous differential operators of order j with smooth 2π -periodic matrix coefficients. The system (7.1.13) is called parabolic if, for all real $\omega_1, \dots, \omega_d$ and all x_0, t_0 , there is a constant $\delta > 0$ such that the eigenvalues λ_j of

$$P_{2r}(x_0, t_0, \omega) = \sum_{\nu_1 + \dots + \nu_d = 2r} A_{\nu_1 \dots \nu_d}(x_0, t_0) \omega_1^{\nu_1} \dots \omega_d^{\nu_d}$$

satisfy the inequality

$$\operatorname{Re} \lambda_j \geq \delta(\omega_1^{2r} + \dots + \omega_d^{2r}). \quad (7.1.14)$$

One can prove the following theorem.

Theorem 7.1.2. If the system (7.1.13) is parabolic, then the initial value problem is well posed.

EXERCISES

7.1.1. The “viscous part” of the Navier–Stokes equations for the fluid velocity field (u, v, w) is

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nu \Delta u + (\nu + \nu') \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial x \partial z} \right), \\ \frac{\partial v}{\partial t} &= \nu \Delta v + (\nu + \nu') \left(\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial y \partial z} \right), \\ \frac{\partial w}{\partial t} &= \nu \Delta w + (\nu + \nu') \left(\frac{\partial^2 w}{\partial x \partial z} + \frac{\partial^2 w}{\partial y \partial z} + \frac{\partial^2 w}{\partial z^2} \right), \end{aligned}$$

where

$$\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Prove that the system is parabolic if the viscosity coefficients fulfill $\nu > 0$, $\nu' \geq 0$.

7.2. STABILITY FOR DIFFERENCE AND FOURIER METHODS

As indicated in the previous section, parabolic problems are easier to solve numerically than hyperbolic problems and other types of problems as well. The damping property built into the differential equation carries over to the approximation in a natural way. The stability analysis is also simplified, since the condition of dissipativity, which plays an important role for hyperbolic problems, is almost automatically satisfied if the method is consistent.

Most parabolic equations contain derivatives of different orders in space. The first question to resolve is whether or not it is sufficient to study the principal part; that is, can lower order terms be ignored? The general stability theorem to be given later shows that this is the case, but it does not follow from the general perturbation theorem (Theorem 5.1.2). To illustrate this, we look at the simple equation

$$u_t = u_{xx} + u_x,$$

and the approximation

$$v^{n+1} = (I + kD_+D_- + kD_0)v^n. \quad (7.2.1)$$

This scheme was studied in Section 2.5, and it was shown that the condition

$$\lambda = \frac{k}{h^2} \leq \frac{1}{2} \quad (7.2.2)$$

is necessary for stability. If λ is a constant, the last term of Eq. (7.2.1) is

$$kD_0v_j^n = \frac{\sqrt{\lambda k}}{2} (v_{j+1}^n - v_{j-1}^n),$$

which is not of the order k , as required by Theorem 5.1.2. However, the symbol of the difference operator on the right-hand side of Eq. (7.2.1) is

$$\hat{Q} = 1 - 4\lambda \sin^2 \frac{\xi}{2} + \sqrt{\lambda k} i \sin \xi$$

with

$$|\hat{Q}| = \left(\left(1 - 4\lambda \sin^2 \frac{\xi}{2} \right)^2 + \lambda k \sin^2 \xi \right)^{1/2} \leq 1 + \mathcal{O}(k),$$

if $\lambda \leq \frac{1}{2}$. The perturbation itself is of the order \sqrt{k} , but its effect on the magnitude of the symbol of the operator is of the order k .

For general parabolic equations and general approximations, this is not true. For example, consider the differential equation

$$u_t = -u_{xxxx} + \alpha u_{xx}, \quad (7.2.3)$$

where α is a constant, and the (admittedly somewhat strange) approximation

$$v^{n+1} = (I - kD_0^4 + k\alpha D_+ D_-)v^n. \quad (7.2.4)$$

The symbol is

$$\hat{Q} = 1 - \lambda \sin^4 \xi - 4\alpha \sqrt{\lambda k} \sin^2 \frac{\xi}{2}, \quad \lambda = \frac{k}{h^4}. \quad (7.2.5)$$

The stability condition without the second-order term is $\lambda \leq 2$. The critical points for the principal part are $\xi = 0$ and $\{\xi = \pi/2, \lambda = 2\}$, that is, at those points where \hat{Q} touches the unit circle. For all other values of ξ and λ , the perturbed \hat{Q} is inside the unit circle if k is small enough.

For $\lambda = 2$, $\xi = \pi/2$ we have

$$\hat{Q} = -(1 + 2\alpha \sqrt{2k}),$$

which yields

$$|\hat{Q}|^n \sim e^{2\alpha \sqrt{2k} n} = e^{2\alpha \sqrt{2} t_n / \sqrt{k}}. \quad (7.2.6)$$

Thus, the method is unstable.

This example exhibits typical behavior. The approximation behaves well for low wave numbers, that is, for small values of ξ , because the properties are then essentially determined by the differential equation. For high wave numbers,

however, the approximation has its own character, and perturbation results for the differential equation cannot be carried over.

The remedy for this example is simple. If the limit point $\lambda = 2$ is eliminated, then the symbol for the principal part only touches the unit circle at $\xi = 0$, and the perturbation is not harmful. In this case, the approximation is dissipative according to Definition 5.2.1. The importance of this concept has been demonstrated for hyperbolic problems. It is also essential for parabolic problems, where it is closely related to the differential equation. This close connection makes it possible to give simple conditions so that the condition (5.2.31) is fulfilled. For a parabolic problem of order $2r$, it is natural to set $\lambda = k/h^{2r}$, where λ is a constant. Then the approximation can be written so that the coefficients only depend on h . The general approximation (5.1.2) can be rewritten in the form

$$\sum_{\sigma=-1}^q (\alpha_\sigma I + Q_\sigma) v^{n-\sigma} = k F^n, \quad (7.2.7a)$$

$$v^\sigma = f^\sigma, \quad \sigma = 0, 1, \dots, q, \quad (7.2.7b)$$

where α_σ are constants and $\alpha_{-1} = 1$. For convenience, it is assumed that the step size is h in all space directions. For $F^n \equiv 0$, $x = x_*$, and $t = t_*$ fixed, the Fourier transform of Eq. (7.2.7a) is

$$\sum_{\sigma=-1}^q (\alpha_\sigma I + \hat{Q}_\sigma(x_*, t_*, h, \xi)) \hat{v}^{n-\sigma} = 0, \quad (7.2.8)$$

where the matrices \hat{Q}_σ are polynomials in h . The principal part is given by

$$\sum_{\sigma=-1}^q (\alpha_\sigma I + \hat{Q}_\sigma(x_*, t_*, 0, \xi)) \hat{v}^{n-\sigma} = 0, \quad (7.2.9)$$

where

$$\hat{Q}_\sigma(x_*, t_*, 0, 0) = 0, \quad \sigma = -1, 0, \dots, q.$$

The symbol for the corresponding one-step form is

$$\hat{Q}(0, \xi) = \begin{bmatrix} -(I + \hat{Q}_{-1})^{-1}(\alpha_0 I + \hat{Q}_0) & \cdots & \cdots & -(I + \hat{Q}_{-1})^{-1}(\alpha_q I + \hat{Q}_q) \\ I & 0 & \cdots & 0 \\ & I & \cdots & 0 \\ & & & 0 \end{bmatrix} \quad (7.2.10)$$

where the arguments x_* , t_* have been left out everywhere and where $(0, \xi)$ have been left out in the right-hand side. The eigenvalues of $\hat{Q}(0, 0)$ are the eigenvalues of the matrix

$$A = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_q \\ 1 & 0 & \cdots & 0 \\ \ddots & \ddots & & \\ & 1 & 0 \end{bmatrix}. \quad (7.2.11)$$

The conditions for stability are given in terms of the eigenvalues of this matrix. Without proof we give sufficient conditions for dissipativity.

Theorem 7.2.1. *Assume that Eq. (7.2.7) is a consistent approximation of the parabolic equation (7.1.13). Then, it is dissipative of order $2r$ if all the eigenvalues of $\hat{Q}(0, \xi)$ are inside the unit circle for $\xi \neq 0$, $|\xi_\nu| \leq \pi$, $\nu = 1, 2, \dots, d$, and all the eigenvalues of A except one are inside the unit circle.*

REMARK. If the initial function is a constant, the solution of the equation with only the principal part is constant. The corresponding value of ξ is zero, hence, there must be one eigenvalue of A which is one.

The general stability estimate can be given in the maximum norm defined by

$$\|v^n\|_{h,\infty} = \sup_j \sum_{\nu=1}^m |v_j^{(\nu)n}|. \quad (7.2.12)$$

The divided differences can also be estimated, and we write

$$D^\tau = D_{+x^{(1)}}^{\tau_1} D_{+x^{(2)}}^{\tau_2} \cdots D_{+x^{(d)}}^{\tau_d}, \quad |\tau| = \sum_{\nu=1}^d \tau_\nu. \quad (7.2.13)$$

Theorem 7.2.2. *Assume that Eq. (7.2.7) is a consistent and dissipative approximation of the parabolic equation (7.1.13), and that the coefficient matrices of*

Eq. (7.2.7a) are Lipschitz continuous. Also, assume that $k = \lambda h^{2r}$, that the eigenvalues of A are inside or on the unit circle, and that the ones on the unit circle are simple. Then the approximation is stable and there is an estimate

$$\sup_n \|t_n^{|\tau|/2r} D^r u^n\|_{h,\infty} \leq K \left(\sum_{\sigma=0}^q \|f^\sigma\|_{h,\infty} + \sup_n \|F^n\|_{h,\infty} \right),$$

$$|\tau| = 0, 1, \dots, 2r - 1. \quad (7.2.14)$$

REMARK. Because the conditions on the eigenvalues of A are less restrictive than those of Theorem 7.2.1, the dissipativity condition is needed in this formulation.

This stability result is very general. The only essential limitation is that $\lambda = k/h^{2r}$ is a constant. This is natural for explicit schemes where the von Neumann condition usually has the form $k/h^{2r} \leq \text{constant}$. For implicit methods, however, it is usually not a natural condition. As an example, consider the Crank–Nicholson approximation of $u_t = -u_{xxxx}$

$$\left(I + \frac{k}{2} (D_+ D_-)^2 \right) v^{n+1} = \left(I - \frac{k}{2} (D_+ D_-)^2 \right) v^n. \quad (7.2.15)$$

The amplification factor is

$$\hat{Q} = \frac{1 - 8\lambda \sin^4 \xi/2}{1 + 8\lambda \sin^4 \xi/2},$$

which fulfills all the conditions of Theorem 7.2.2 for any value of λ . The scheme has order of accuracy (2, 2). If the time and space derivatives are of the same order, it is natural to choose k and h to be of the same order. For example, with $k = h = 0.01$, we get $\lambda = k/h^4 = 10^6$. Some of the constants involved in the stability proof depend on λ , and K in the final estimate may be very large if λ is very large.

In general, since it is desirable to choose k proportional to h^p , $p < 2r$, implicit methods must be used, and it is usually possible to apply the energy method. In many applications the differential operator in space is semibounded, and the difference operator can be constructed such that it is also semibounded. Then Theorems 5.3.2 and 5.3.3 can be applied, and stability is obtained with an arbitrary relation between k and h .

As an application of the stability theory, we study a generalized form of the DuFort-Frankel method. Equation (2.5.13) shows that one root is on the unit circle for $\xi = \pi$ and, accordingly, the method is not dissipative. For $u_t = au_{xx}$, a modified scheme is

$$u^{n+1} = u^{n-1} + 2kaD_+D_-u^n - 2\lambda\gamma a(u^{n+1} - 2u^n + u^{n-1}), \quad \lambda = k/h^2.$$
(7.2.16)

The original method is obtained with $\gamma = 1$. The characteristic equation for the Fourier-transformed scheme is

$$z^2 - 1 = -2\beta z - \alpha(z - 1)^2,$$
(7.2.17)

where

$$\alpha = 2\lambda\gamma a, \quad \beta = 4\lambda a \sin^2 \frac{\xi}{2}.$$

Now we assume

$$\gamma > 1,$$
(7.2.18)

which yields the inequalities

$$2\alpha > \beta \geq 0.$$
(7.2.19)

The properties of the roots of Eq. (7.2.17) are given by the following lemma.

Lemma 7.2.1. *Assume that the conditions (7.2.19) hold. Then the roots z_1, z_2 of Eq. (7.2.17) are inside the unit circle, except for $\beta = 0$ when $z_1 = 1$.*

Proof. The roots of Eq. (7.2.17) are

$$z_{1,2} = \frac{1}{1+\alpha} \left(\alpha - \beta \pm \sqrt{1 - \beta(2\alpha - \beta)} \right).$$
(7.2.20)

If $1 - \beta(2\alpha - \beta) \geq 0$, then, by Eq. (7.2.19),

$$\sqrt{1 - \beta(2\alpha - \beta)} \leq 1.$$

Therefore, because $|\alpha - \beta| \leq \alpha$ from Eq. (7.2.19), we get $|z_{1,2}| \leq 1$. Equality only holds if $\beta = 0$, and in that case

$$z_1 = 1,$$
(7.2.21a)

$$|z_2| = \left| \frac{\alpha - 1}{\alpha + 1} \right| < 1.$$
(7.2.21b)

If $1 - \beta(2\alpha - \beta) < 0$, then the roots are complex, and

$$|z_{1,2}| = \left| \frac{\alpha - 1}{\alpha + 1} \right| < 1. \quad (7.2.22)$$

This proves the lemma.

Because the scheme is consistent for any value of the constant λ , it follows from this lemma and Theorem 7.2.1 that it is also dissipative under the condition (7.2.18). Because only one root is on the unit circle for $\xi = 0$, all the conditions of Theorem 7.2.2 are fulfilled for any value of λ .

Let us next consider the system $u_t = Au_{xx}$, where A has positive eigenvalues. A straightforward generalization of Eq. (7.2.16) is obtained by replacing a by A at both places where it occurs. However, because the last term is only present for stability reasons, a more convenient scheme can be obtained by substituting $\rho(A)I$ for A giving

$$v^{n+1} = v^{n-1} + 2kAD_+D_-v^n - 2\lambda\gamma\rho(A)(v^{n+1} - 2v^n + v^{n-1}). \quad (7.2.23)$$

Denote an eigenvalue of A by a . Then the eigenvalues z of the amplification matrix are given by Eq. (7.2.17), where

$$\alpha = 2\lambda\gamma\rho(A) > 0, \quad \beta = 4\lambda a \sin^2 \frac{\xi}{2} > 0.$$

The calculation above for the scalar case also goes through in this case, and Eq. (7.2.18) is still the condition for stability.

The DuFort–Frankel method can also be used for time discretization with the Fourier method. For systems $u_t = Au_{xx}$, we get

$$v^{n+1} = v^{n-1} + 2kAS^2v^n - 2\lambda\gamma\rho(A)(v^{n+1} - 2v^n + v^{n-1}) \quad (7.2.24)$$

with its Fourier transform

$$\hat{v}^{n+1} = \hat{v}^{n-1} - 2k\omega^2A\hat{v}^n - 2\lambda\gamma\rho(A)(\hat{v}^{n+1} - 2\hat{v}^n + \hat{v}^{n-1}). \quad (7.2.25)$$

Again, the characteristic equation is of the form (7.2.17) with

$$\alpha = 2\lambda\gamma\rho(A), \quad \beta = k\omega^2a.$$

The condition $2\alpha > \beta$ used in Lemma 7.2.1 implies $4\gamma\rho(A) > a(h\omega)^2$, $|\omega h| \leq \pi$. The approximation will be unconditionally stable if we choose

$$\gamma > \frac{\pi^2}{4}. \quad (7.2.26)$$

EXERCISES

- 7.2.1.** Prove that the eigenvalues of the amplification matrix for the scheme (7.2.23) are given by Eq. (7.2.17).
- 7.2.2.** If a time-marching procedure is used for computing a steady-state solution, it is desirable that the error be independent of k , particularly if large time steps are used. Does the DuFort–Frankel method have this property?

7.3. DIFFERENCE APPROXIMATIONS IN SEVERAL SPACE DIMENSIONS

We first consider equations of second order, because they are the only parabolic equations for which explicit methods can possibly be used in realistic applications. Even for these problems, there is a severe restriction on the time step, but because explicit methods are so much easier to implement, in particular, on vector and parallel computers, they are sometimes used.

First, consider a parabolic system written in the form

$$\frac{\partial \mathbf{u}}{\partial t} = \sum_{\nu=1}^d \left(\sum_{\mu=1}^d A_{\nu\mu}(x, t) \frac{\partial^2 u}{\partial x^{(\nu)} \partial x^{(\mu)}} + B_{\nu}(x, t) \frac{\partial u}{\partial x^{(\nu)}} \right) + C(x, t) \mathbf{u}. \quad (7.3.1)$$

The Euler method is

$$\mathbf{v}^{n+1} = (I + Q_2(t_n)) \mathbf{v}^n, \quad (7.3.2)$$

where

$$\begin{aligned} Q_2(t_n) &= k \sum_{\nu=1}^d \left[A_{\nu\nu}(t_n) D_{+x^{(\nu)}} D_{-x^{(\nu)}} \right. \\ &\quad \left. + \sum_{\mu=\nu+1}^d A_{\nu\mu}(t_n) D_{0x^{(\nu)}} D_{0x^{(\mu)}} + B_{\nu}(t_n) D_{0x^{(\nu)}} \right] + k C(t_n). \end{aligned} \quad (7.3.3)$$

The amplification factor for the principal part is $\hat{Q} = I - \hat{A}$, where

$$\hat{A} = \sum_{\nu=1}^d \left(4\lambda_{\nu} A_{\nu\nu} \sin^2 \frac{\xi_{\nu}}{2} + \sum_{\mu=\nu+1}^d \sqrt{\lambda_{\nu} \lambda_{\mu}} A_{\nu\mu} \sin \xi_{\nu} \sin \xi_{\mu} \right). \quad (7.3.4)$$

To derive a convenient stability condition, we assume that all the matrices $A_{\nu\mu}$ are Hermitian. We then have the following theorem.

Lemma 7.3.1. *Assume that the matrices $A_{\nu\mu}$ are Hermitian and that the system (7.3.1) is parabolic. Then the matrix \hat{A} defined by Eq. (7.3.4) is positive semi-definite, and*

$$\rho(\hat{A}) \leq 4 \sum_{\nu=1}^d \lambda_\nu \rho(A_{\nu\nu}). \quad (7.3.5)$$

Proof. The symbol (with reversed sign) of the principal part of Eq. (7.3.1) is

$$\hat{P} = \sum_{\nu=1}^d \left(A_{\nu\nu} \omega_\nu^2 + \sum_{\mu=\nu+1}^d A_{\nu\mu} \omega_\nu \omega_\mu \right).$$

By the parabolicity condition, we have

$$\operatorname{Re} \langle v, \hat{P}v \rangle \geq 0. \quad (7.3.6)$$

The discrete symbol corresponding to \hat{P} is

$$\begin{aligned} \hat{A} &= \sum_{\nu=1}^d \left(4 \lambda_\nu A_{\nu\nu} \sin^2 \frac{\xi_\nu}{2} + \sum_{\mu=\nu+1}^d \sqrt{\lambda_\nu \lambda_\mu} A_{\nu\mu} \sin \xi_\nu \sin \xi_\mu \right), \\ \lambda_\nu &= k/h_\nu^2. \end{aligned} \quad (7.3.7)$$

The condition (7.3.6) implies

$$\begin{aligned} &\left| \operatorname{Re} \left\langle v, \sum_{\nu=1}^d \sum_{\mu=\nu+1}^d \sqrt{\lambda_\nu \lambda_\mu} A_{\nu\mu} \sin \xi_\nu (\sin \xi_\mu) v \right\rangle \right| \\ &\leq \left| \operatorname{Re} \left\langle v, \sum_{\nu=1}^d \lambda_\nu A_{\nu\nu} (\sin^2 \xi_\nu) v \right\rangle \right|, \end{aligned}$$

for any vector v . The inequality

$$\operatorname{Re} \langle v, \hat{A}v \rangle \geq 0, \quad (7.3.8)$$

then follows from $|\sin \xi| \leq 2|\sin \xi/2|$; that is, \hat{A} is positive semi-definite.

Furthermore,

$$\begin{aligned}\rho(\hat{A}) &= \max_{|v|=1} \langle v, \hat{A}v \rangle \leq \max_{|v|=1} \left\langle v, \sum_{\nu=1}^d \lambda_\nu A_{\nu\nu} \left(4 \sin^2 \frac{\xi_\nu}{2} + \sin^2 \xi_\nu \right) v \right\rangle, \\ &\leq \sum_{\nu=1}^d \max_{|\xi_\nu| \leq \pi} \left(4 \sin^2 \frac{\xi_\nu}{2} + \sin^2 \xi_\nu \right) \lambda_\nu \max_{|v|=1} \langle v, A_{\nu\nu} v \rangle, \\ &= \alpha_0 \sum_{\nu=1}^d \lambda_\nu \rho(A_{\nu\nu}),\end{aligned}$$

where $\alpha_0 = 4$ in Eq. (7.3.5) is obtained by elementary calculus.

All the eigenvalues of the principal part of $\hat{Q} = I - \hat{A}$ are less than one if

$$\rho(\hat{A}) < 2.$$

Thus, a sufficient stability condition is obtained from Lemma 7.3.1 if

$$\sum_{\nu=1}^d \lambda_\nu \rho(A_{\nu\nu}) \leq \frac{1}{2}. \quad (7.3.9)$$

We remind the reader that a more restrictive condition on k may be required for practical computations when lower order terms are present, as shown for the model problem (2.6.1). In particular, this is true when the matrices $A_{\nu\mu}$ are small compared to B_ν .

The second-order accurate operators used in the Euler method (7.3.2) can be replaced by fourth-order accurate operators, for example,

$$\begin{aligned}D_{+x^{(\nu)}} D_{-x^{(\nu)}} &\rightarrow D_{+x^{(\nu)}} D_{-x^{(\nu)}} \left(I - \frac{h_\nu^2}{12} D_{+x^{(\nu)}} D_{-x^{(\nu)}} \right), \\ D_{0x^{(\nu)}} &\rightarrow D_{0x^{(\nu)}} \left(I - \frac{h_\nu^2}{6} D_{+x^{(\nu)}} D_{-x^{(\nu)}} \right).\end{aligned}$$

The corresponding stability condition is a little more restrictive but, for a given accuracy requirement, a coarser mesh in space can be used, and the scheme on a coarser mesh can be more efficient (cf. the discussion in Section 3.1).

The DuFort-Frankel method requires a small time step to be accurate. It is still sometimes used since it is convenient to use an unconditionally stable explicit method. In particular, this is true when the time-dependent equations

are used to compute a steady-state solution. In this case, the behavior of the solution as a function of time need not be computed accurately, and the time step can be chosen to optimize the convergence rate as $t_n \rightarrow \infty$.

Another advantage with the DuFort–Frankel method is that it can be conveniently combined with leap-frog differencing for the first-order terms. The resulting scheme is

$$\begin{aligned} v^{n+1} = & v^{n-1} + 2k \left(\sum_{\nu=1}^d \left(A_{\nu\nu} D_{+x^{(\nu)}} D_{-x^{(\nu)}} \right. \right. \\ & \left. \left. + \sum_{\mu=\nu+1}^d A_{\nu\mu} D_{0x^{(\nu)}} D_{0x^{(\mu)}} + B_\nu D_{0x^{(\nu)}} \right) + C \right) v^n \\ & - 2k\gamma \sum_{\nu=1}^d \frac{\rho(A_{\nu\nu})}{h_\nu^2} (v^{n+1} - 2v^n + v^{n-1}), \end{aligned} \quad (7.3.10)$$

where all the matrices are evaluated at $t = t_n$. It is unconditionally stable for $\gamma > 1$. The order of accuracy is 2 if the $\lambda_\nu = k/h_\nu^2$ are constants. The zero-order term is taken at the central time level in the version given in Eq. (7.3.10). As for the model equation (2.6.1), this may give rise to an increasing parasitic solution if k is not very small. The remedy for this is, again, to replace Cv^n by $C(v^{n+1} + v^{n-1})/2$, but there is a penalty. A system of equations with m unknowns must be solved at each gridpoint.

The severe restriction on the time step is not present when implicit methods are used. The θ scheme introduced in Section 2.3 for a hyperbolic model problem can be used for parabolic problems as well. With $Q_2(t_n)$ defined by Eq. (7.3.3), the approximation is

$$(I - \theta k Q_2(t_{n+1}))v^{n+1} = (I - (1 - \theta)k Q_2(t_n))v^n. \quad (7.3.11)$$

The Crank–Nicholson method is obtained for $\theta = 1/2$; this is the only case that is second-order accurate in time. Unconditional stability holds for all θ with $\theta \geq \frac{1}{2}$ (see Exercise 7.3.7).

The θ scheme (7.3.11) can also be used for general parabolic problems. For problems of the type

$$u_t = \sum_{\nu=1}^d A_\nu \frac{\partial^{2r} u}{\partial x^{(\nu)} 2r}, \quad (7.3.12)$$

we simply define

$$Q_2 = k \sum_{\nu=1}^d A_\nu (D_{+x}(\nu) D_{-x}(\nu))^r. \quad (7.3.13)$$

In several space dimensions, direct methods for solving the large algebraic systems at each step are expensive. However, iterative methods usually work well for parabolic problems. This is because the coefficient matrix is naturally positive definite and fulfills the conditions for fast convergence of most iterative methods.

It is usually more efficient to split the scheme into a sequence of one-dimensional steps. For two-dimensional problems and the Crank–Nicholson method ($\theta = \frac{1}{2}$), the splitting procedure resulting in the ADI-scheme was described in Section 5.4. With Q_x and Q_y defined by

$$\begin{aligned} Q_x(t) &= A_1(x, t) (D_{+x} D_{-x})^r, \\ Q_y(t) &= A_2(x, t) (D_{+y} D_{-y})^r, \end{aligned} \quad (7.3.14)$$

the scheme can be written in the form

$$\left(I - \frac{k}{2} Q_x(t_{n+1/2}) \right) v^{n+1/2} = \left(I + \frac{k}{2} Q_y(t_n) \right) v^n, \quad (7.3.15a)$$

$$\left(I - \frac{k}{2} Q_y(t_{n+1}) \right) v^{n+1} = \left(I + \frac{k}{2} Q_x(t_{n+1/2}) \right) v^{n+1/2}, \quad (7.3.15b)$$

and it is unconditionally stable if

$$\begin{aligned} \operatorname{Re}(u, Q_x u)_h &\leq 0, \\ \operatorname{Re}(u, Q_y u)_h &\leq 0, \end{aligned} \quad (7.3.16)$$

(see Exercise 5.4.2).

With a different ordering of the operators, we obtain the fractional step method

$$\left(I - \frac{k}{2} Q_x(t_{n+1/2}) \right) v^{n+1/2} = \left(I + \frac{k}{2} Q_x(t_n) \right) v^n, \quad (7.3.17a)$$

$$\left(I - \frac{k}{2} Q_y(t_{n+1}) \right) v^{n+1} = \left(I + \frac{k}{2} Q_y(t_{n+1/2}) \right) v^{n+1/2}. \quad (7.3.17b)$$

Stability follows directly from semiboundedness. Assuming real solutions and constant coefficients, we obtain from Eq. (7.3.16)

$$\|v^{n+1}\|_h^2 - \|v^{n+1/2}\|_h^2 \leq \frac{k}{2} (v^{n+1} + v^{n+1/2}, Q_y(v^{n+1} + v^{n+1/2}))_h \leq 0,$$

$$\|v^{n+1/2}\|_h^2 - \|v^n\|_h^2 \leq \frac{k}{2} (v^{n+1/2} + v^n, Q_x(v^{n+1/2} + v^n))_h \leq 0,$$

which gives us

$$\|v^{n+1}\|_h \leq \|v^{n+1/2}\|_h \leq \|v^n\|_h.$$

The method (7.3.17) is a splitting method, like those discussed in Section 5.4, because each half-time step is one-dimensional. To obtain second-order accuracy in time for general problems, Eq. (7.3.17) must be followed by two half-steps in reversed order. If Q_x and Q_y commute (which is a severe restriction), then Eq. (7.3.17) is, actually, equivalent to the ADI scheme and, consequently, second-order accurate. Applying the operator $I - (k/2)Q_x$ to Eq. (7.3.17b) we get

$$\begin{aligned} \left(I - \frac{k}{2} Q_x \right) \left(I - \frac{k}{2} Q_y \right) v^{n+1} &= \left(I - \frac{k}{2} Q_x \right) \left(I + \frac{k}{2} Q_y \right) v^{n+1/2}, \\ &= \left(I + \frac{k}{2} Q_y \right) \left(I - \frac{k}{2} Q_x \right) v^{n+1/2}, \\ &= \left(I + \frac{k}{2} Q_y \right) \left(I + \frac{k}{2} Q_x \right) v^n, \\ &= \left(I + \frac{k}{2} Q_x \right) \left(I + \frac{k}{2} Q_y \right) v^n. \end{aligned}$$

But this is the ADI scheme, which is seen by applying the operator $[I - (k/2)Q_x]$ to Eq. (7.3.15b) and using Eq. (7.3.15a).

Fractional step methods have an advantage over ADI methods. Because each half-step is strictly one-dimensional, only one line of data is needed to define each system to be solved, and this line can be directly overwritten by a new solution. This is convenient for large problems.

The fractional step method is based on one-dimensional Crank–Nicholson steps. Another method is obtained if the backward Euler method is used as a basis:

$$(I - kQ_x)v^{n+1/2} = v^n, \quad (7.3.18a)$$

$$(I - kQ_y)v^{n+1} = v^{n+1/2}. \quad (7.3.18b)$$

If the operators Q_x and Q_y satisfy Eq. (7.3.16), there is an energy estimate for each half-step and stability follows. This method is only first-order accurate in time, and, because each half-step is only first-order accurate for the one-dimen-

sional problem, the accuracy cannot be raised by the usual reversal procedure. However, the method (7.3.18) is sometimes preferred over second-order accurate ones when strong damping is required. The generalization of fractional step methods to more than two space dimensions is obvious.

Finally, we note that parabolic problems often are given in self-adjoint form, as discussed in Section 7.1. Consider, for example, second-order equations in the form

$$\frac{\partial u}{\partial t} = \sum_{\nu=1}^d \frac{\partial}{\partial x^{(\nu)}} \left(A_\nu(x, t), \frac{\partial u}{\partial x^{(\nu)}} \right), \quad (7.3.19)$$

where the matrices A_ν are positive definite. In such cases, the space operator is negative semidefinite, because

$$\left(u, \sum_{\nu=1}^d \frac{\partial}{\partial x^{(\nu)}} \left(A_\nu, \frac{\partial u}{\partial x^{(\nu)}} \right) \right) = - \sum_{\nu=1}^d \left(\frac{\partial u}{\partial x^{(\nu)}}, A_\nu \frac{\partial u}{\partial x^{(\nu)}} \right) \leq 0. \quad (7.3.20)$$

The same property carries over to the discrete case if the approximation is defined to be

$$\frac{dv_j}{dt} = \sum_{\nu=1}^d D_{+x^{(\nu)}}(A_\nu(x_{*\nu}, t)D_{-x^{(\nu)}})v_j, \quad (7.3.21)$$

where

$$x_{*\nu} = (x_{j_1}^{(1)}, \dots, x_{j_\nu-1/2}^{(\nu)}, \dots, x_{j_d}^{(d)})^T.$$

For further comments on this method, see the Bibliographic Notes.

EXERCISES

7.3.1. Is it true that the matrix \hat{A} given by Eq. (7.3.4) always has eigenvalues with non-negative real part, when the corresponding differential equation is parabolic?

7.3.2. Derive the estimate (7.3.5) by proving

$$\max_{|\xi| \leq \pi} \left(4 \sin^2 \frac{\xi}{2} + \sin^2 \xi \right) = 4.$$

- 7.3.3.** Derive the stability condition for the Euler method (7.3.2) when applied to the parabolic equation

$$u_t = (a + ib)u_{xx},$$

where a and b are real numbers.

- 7.3.4.** Find a parabolic differential equation for which the Euler method (7.3.2) is unstable for all values of $\lambda = k/h^2$ (equal step size in all space dimensions).
- 7.3.5.** Define the DuFort–Frankel method (7.2.23) with D_+, D_- replaced by the fourth-order accurate operator. Derive the condition on γ for unconditional stability.
- 7.3.6.** Assume that sufficient accuracy is obtained by using the Crank–Nicholson method on a grid with 100 points in each space direction and $k = h$ for a parabolic problem of order $2r$. When compared with the Euler method on the same grid in space, is there any breakpoint d_0 for the number of space dimensions and/or $2r_0$ for the order of the equation, when one method becomes more efficient than the other?
- 7.3.7.** Prove unconditional stability for the θ scheme (7.3.11) if $\theta \geq \frac{1}{2}$.

BIBLIOGRAPHIC NOTES

The treatment of general parabolic systems in Section 7.1 is very brief. A more thorough discussion is found in Kreiss and Lorenz (1989), where Theorem 7.1.2 is proved for second-order equations.

The general theory for difference approximations in Section 7.2 was originated by Widlund (1966). The proof of Theorem 7.2.1 is easy, but the proof of Theorem 7.2.2 is not. The basic technique is similar to the one used for hyperbolic problems in Sections 7.5 and 7.6.

From a stability point of view, the method (7.3.21) seems to be the best way of approximating the spacial differential operator in Eq. (7.3.19), because no lower order terms perturb the decay rate. However, with regard to accuracy, the situation is not that simple. Dykson and Rice (1985) investigated this for time-independent problems in two space dimensions. Their conclusion is that, unless the solution varies much more rapidly than A_1 and A_2 as functions of x , the form

$$\frac{dv_j}{dt} = \sum_{\nu=1}^d \left(A_\nu(x_j) D_{+x^{(\nu)}} D_{-x^{(\nu)}} + \frac{\partial A_\nu(x_j)}{\partial x^{(\nu)}} D_{0x^{(\nu)}} \right) v_j$$

is preferred.

8

PROBLEMS WITH DISCONTINUOUS SOLUTIONS

In this chapter, we consider the difficulties that arise when discontinuous solutions of hyperbolic equations are approximated. There is an extensive literature on this subject including several recent books. We do not survey the many special methods that have been developed. Instead we discuss the basic phenomena and the numerical techniques necessary to overcome the difficulties.

8.1. DIFFERENCE METHODS FOR LINEAR HYPERBOLIC EQUATIONS

So far we have concentrated on the approximation of smooth solutions. The notion of generalized solutions was introduced in Section 4.8, and these need not be smooth, or even continuous. We carried out some computations in Section 3.1 for our model hyperbolic equation using centered difference methods with the sawtooth function as initial data and observed that the results were no good. In Figure 3.1.1, we observed that the approximation was obliterated by high-frequency oscillations. This is typical behavior of difference methods when used to approximate a discontinuous solution. In this situation, these difficulties arise solely from the discontinuous initial data. We again return to our model hyperbolic equation

$$u_t = au_x, \quad 0 \leq x \leq 2\pi, \quad 0 \leq t \quad (8.1.1)$$

with the piecewise constant 2π -periodic initial data considered in Section 4.3,

$$u(x, 0) = f(x) = \begin{cases} 0, & \text{for } 0 \leq x < \frac{2}{3}\pi, \\ 1, & \text{for } \frac{2}{3}\pi \leq x \leq \frac{4}{3}\pi, \\ 0, & \text{for } \frac{4}{3}\pi < x \leq 2\pi, \end{cases} \quad (8.1.2)$$

which has a periodic solution

$$u(x, t) = u(x + 2\pi, t). \quad (8.1.3)$$

As we know, the solution of this problem is given by $u(x, t) = f(x + at)$ and is constant along the characteristics $x + at = \text{constant}$.

In Figure 8.1.1, we display approximations of the solution of the problem (8.1.1) to (8.1.3) with $a = -1, h = 2\pi/240, k = 2h/3$ at $t = 40k$ and $t = 360k = 2\pi$ obtained using the leap-frog, Lax–Wendroff, Lax–Friedrichs, and first-order upwind methods. Note that $u(x, 2\pi) = u(x, 0)$. The first-order upwind method is defined by

$$\begin{aligned} v_j^{n+1} &= v_j^n + \frac{ak}{h} (v_j^n - v_{j-1}^n), && \text{if } a < 0, \text{ and} \\ v_j^{n+1} &= v_j^n + \frac{ak}{h} (v_{j+1}^n - v_j^n), && \text{if } a \geq 0. \end{aligned} \quad (8.1.4)$$

The other methods have all been defined and used extensively in earlier chapters.

These computations are all unacceptable unless extremely fine grids are used. Our earlier linear convergence proofs do apply, as $h \rightarrow 0$ the solutions do converge. The leap-frog scheme has severe oscillations near the discontinuities that spread at a rate proportional to t . This is typical of all nondissipative methods. The Lax–Wendroff method has overshoots and undershoots near the discontinuities, but they do not spread as rapidly. Dissipation is essential when approximating discontinuous solutions. The Lax–Friedrichs and first-order upwind methods have no oscillations or overshoots and undershoots. However, they smear the discontinuity over a large region. They are examples of so-called monotone methods, which introduce no new maxima or minima for sufficiently small $\lambda = k/h$ when used for scalar equations. Unfortunately, they can be at most first-order accurate.

It has been shown that it is generally better to add a dissipative term to a nondissipative approximation than to use an inherently dissipative method such as the Lax–Wendroff method. Consider centered methods accurate of even order p and dissipative of order $2r$. It has been shown that the spread and amplitude of the oscillations slowly decrease as p increases. If $2r < p$, as is the case with the upstream method, the viscous effects dominate as t/h increases. If $2r > p$, as is the case with the Lax–Wendroff method, the viscous effects decay as t/h increases (see the Bibliographic Notes).

We next approximate Eqs. (8.1.1) and (8.1.2) using a fourth-order centered method with a fourth-order dissipative term

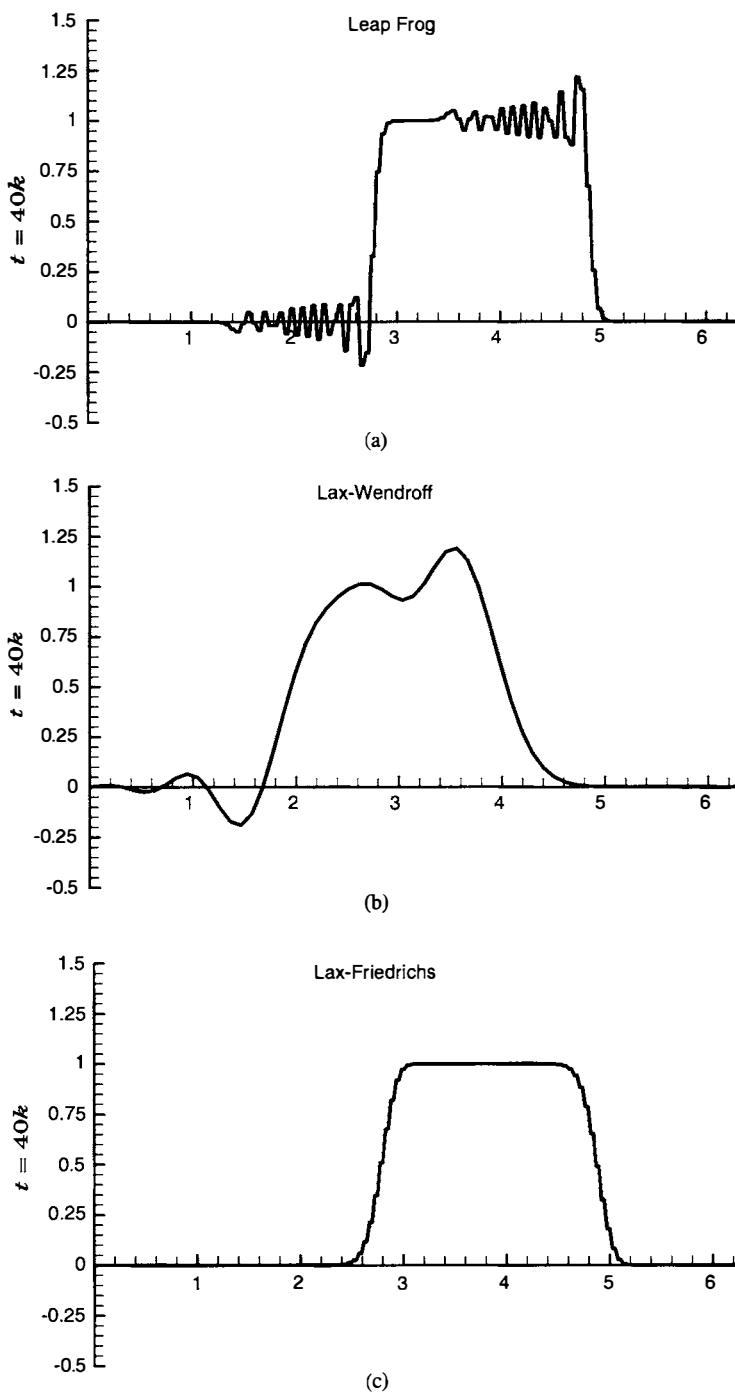


Figure 8.1.1. (a)–(c).

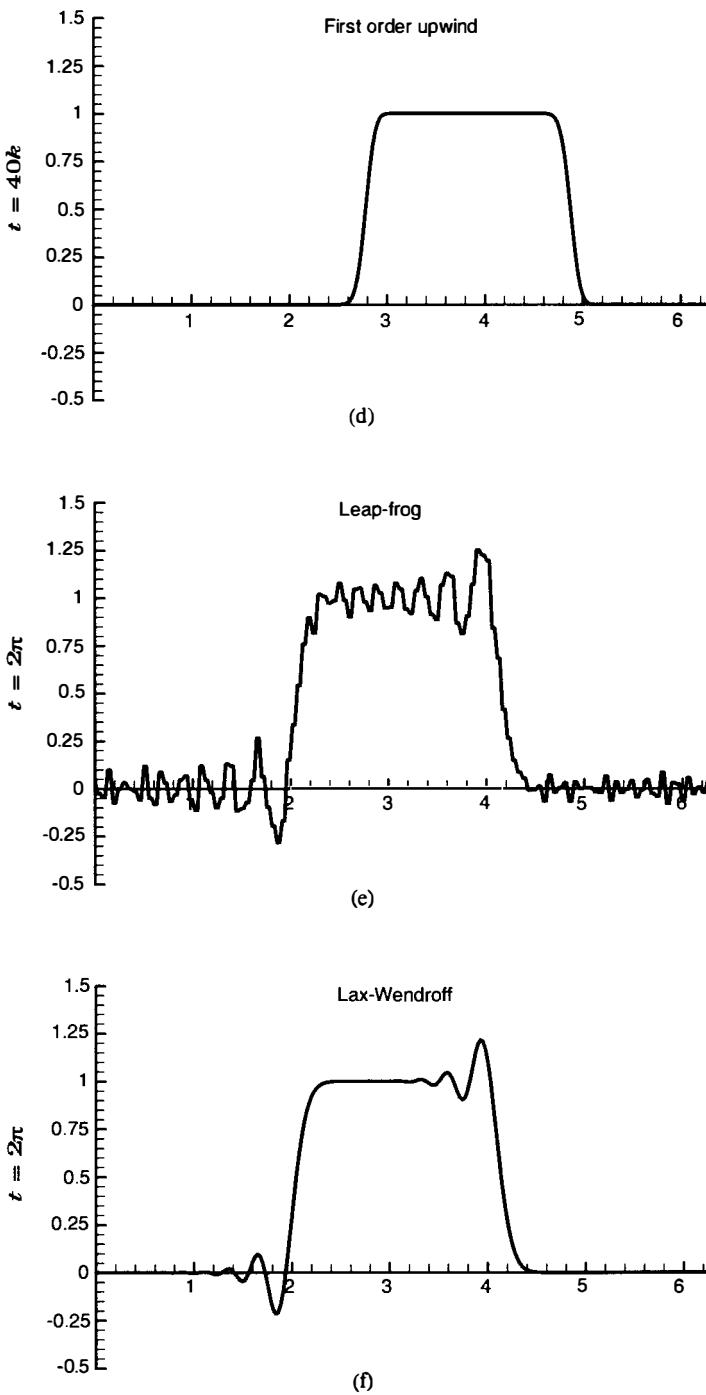


Figure 8.1.1. (d)–(f).

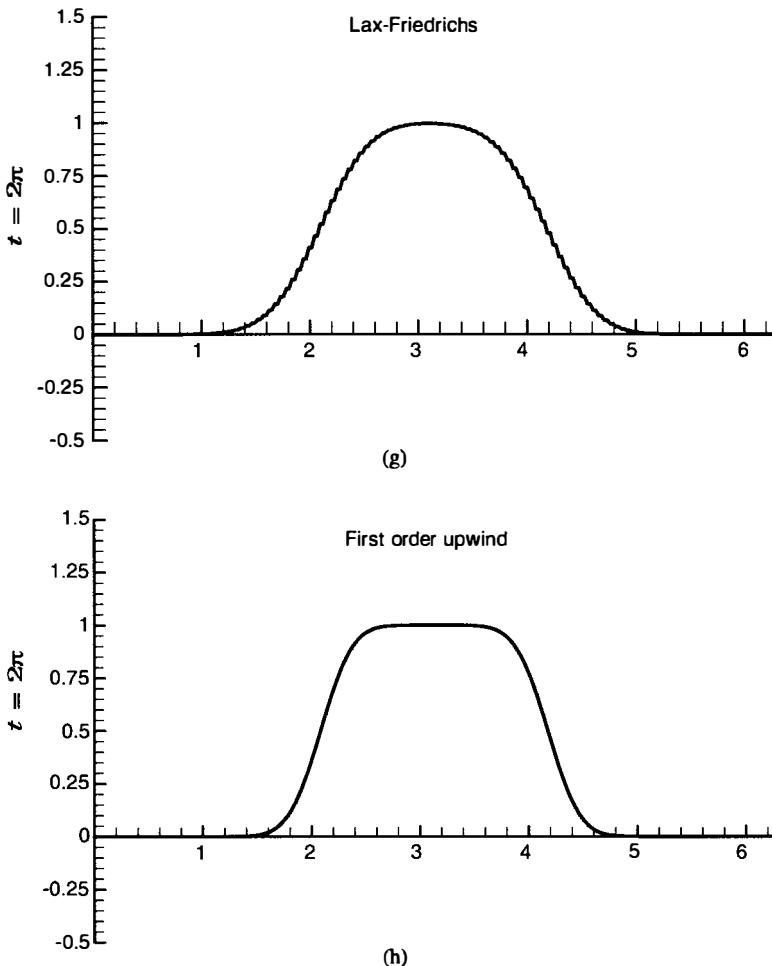


Figure 8.1.1. (g),(h).

$$v_t = aD_0 \left(I - \frac{h^2}{6} D_+ D_- \right) v - \epsilon h^3 (D_+ D_-)^2 v. \quad (8.1.5)$$

We use the fourth-order Runge–Kutta method in time with $h = 2\pi/240$ and $k = 2h/3$; $\epsilon = 0.1, 0.05$, and 0.01 . Dissipation introduced by the time discretization is of sixth-order, so the fully discretized method is dissipative of order 4. The results are shown in Figure 8.1.2 at $t = 40k$ and $t = 2\pi$. It is clear that $\epsilon = 0.01$ is too small to control the oscillations. The results with $\epsilon = 0.05$ and 0.1 are much better; they are also better than the results in Figure 8.1.1, but still leave a lot to be desired. A fine global, or adaptive, grid is still needed to get very accurate results.

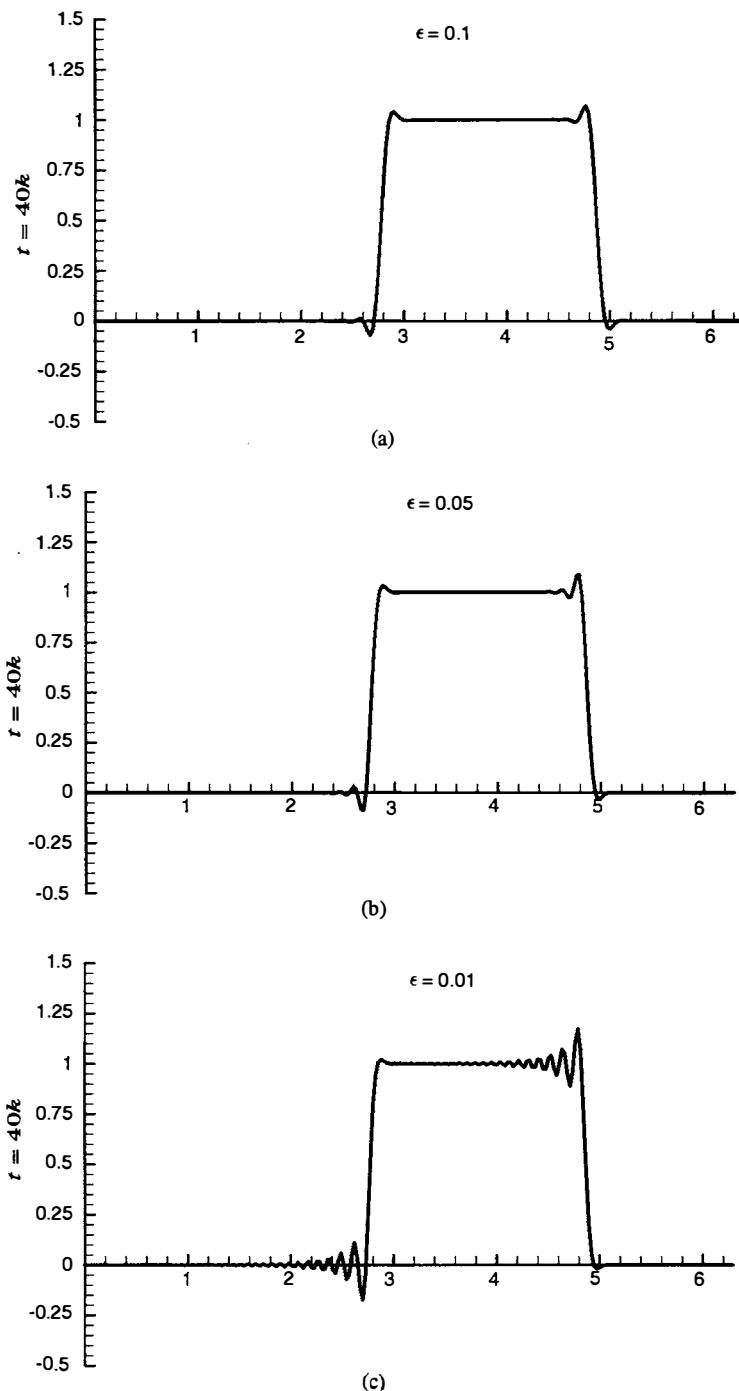
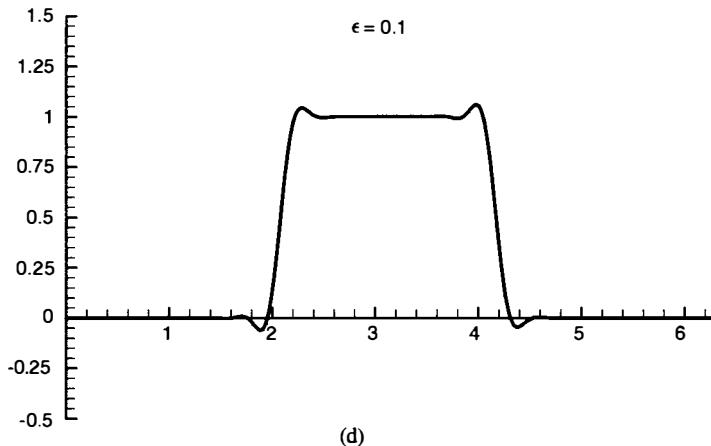
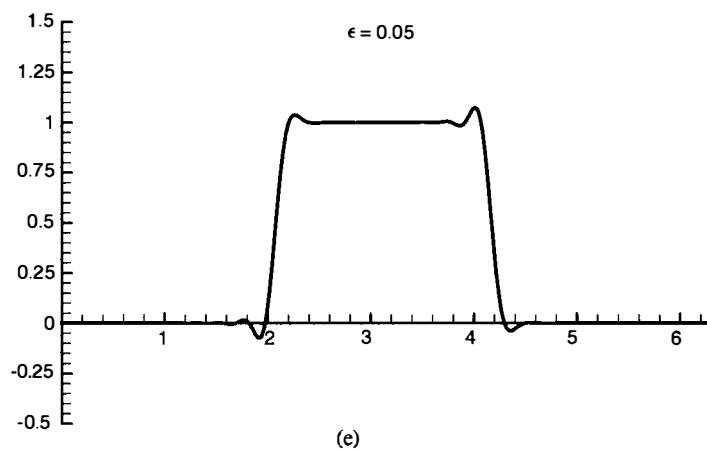


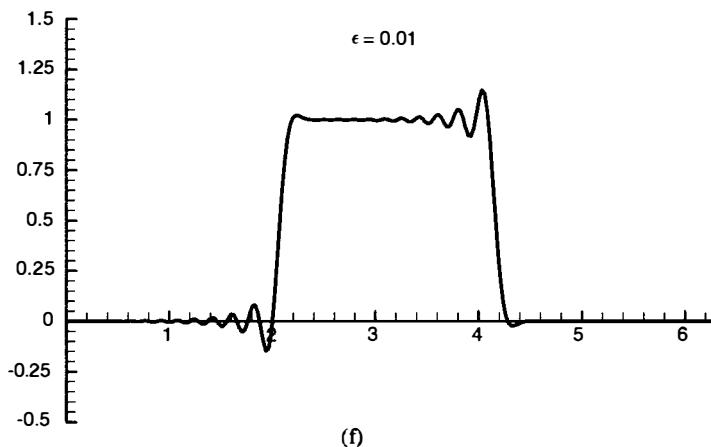
Figure 8.1.2. (a)–(c).



(d)



(e)



(f)

Figure 8.1.2. (d)–(f).

Special techniques for smoothing the initial data to increase the convergence rate and to recover piecewise smooth traveling waves from oscillatory approximations have been developed. Some of these ideas are noted in the Bibliographic Notes.

Of course, we could compute the exact solution of this problem using, for example, the upwind scheme with $ak/h = -1$. However, this is only possible for a problem with constant coefficients. In this case, the method degenerates to the method of characteristics that we consider next.

8.2. METHOD OF CHARACTERISTICS

First, we again consider the scalar initial value problem for Eq. (8.1.1) with periodic initial data $u(x, 0) = f(x)$. If a is constant, then we can write down the solution directly:

$$u(x, t) = f(x + at);$$

that is,

$$u(x, t) = f(x_0), \quad (8.2.1)$$

for all x, t with $x + at = x_0$. Thus, the solution does not change along the so called *characteristic lines* or *characteristics* (see Figure 8.2.1):

$$x + at = x_0. \quad (8.2.2)$$

If, for example, $f(x)$ consists of a pulse, then the pulse moves with the speed $-a$ without changing form (see Figure 8.2.2).

If we solve this problem with difference methods, as we have seen in Section 8.1, we have a lot of difficulty with spurious oscillations.

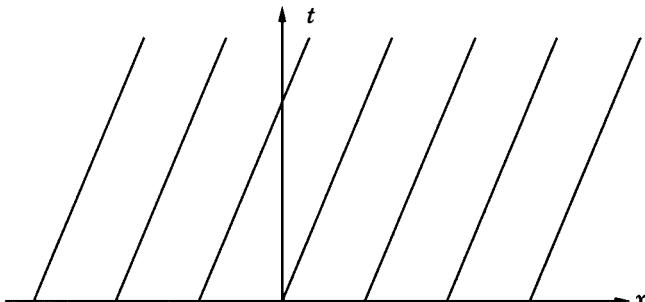


Figure 8.2.1. Characteristics for $a < 0$.

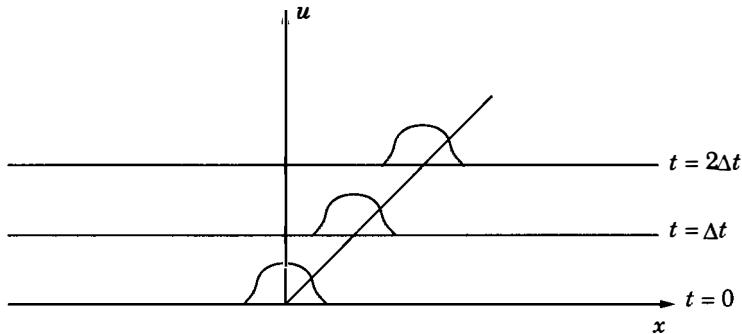


Figure 8.2.2. A moving pulse.

Now assume that $a = a(x, t)$ is a smooth real function of x, t . In this case the characteristics are defined as solutions of the ordinary differential equation

$$\frac{dx(t)}{dt} = -a(x(t), t), \quad (8.2.3a)$$

with initial data

$$x(0) = x_0. \quad (8.2.3b)$$

If a is constant, then we recover Eq. (8.2.2). The solution of Eq. (8.2.3) does not change along the characteristic lines, because

$$\frac{d}{dt} u(x(t), t) = u_x \frac{dx}{dt} + u_t = 0.$$

The problem (8.2.3) is an initial-value problem for an ordinary differential equation. Its solution exists for all time, because $a(x, t)$ is, by assumption, a bounded smooth function of x and t . Also, every point (x, t) lies on exactly one characteristic line, because we can solve Eq. (8.2.3a) backwards in time; that is, given a point (x, t) we can solve Eq. (8.2.3a) starting at (x, t) in the negative t direction. This process defines a characteristic as a unique relation between x and t for any given x_0 (see Figure 8.2.3), which we write in the form

$$\psi(x, t) = x_0. \quad (8.2.4)$$

From Eq. (8.2.3), the solution of Eq. (8.1.1) is given by

$$u(x, t) = f(x_0) = f(\psi(x, t)). \quad (8.2.5)$$

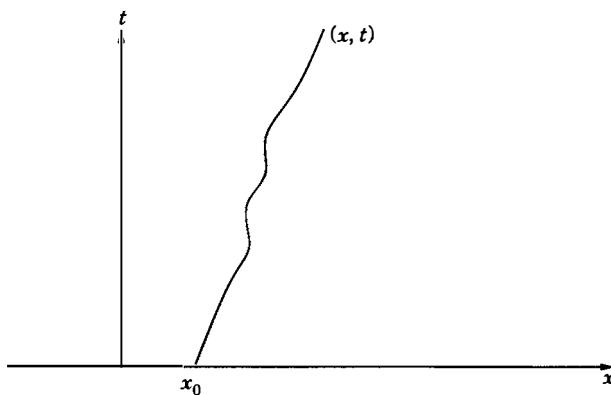


Figure 8.2.3. A characteristic line.

This is the method of characteristics.

EXAMPLE 8.2.1. If $a = -x$, then Eq. (8.2.3a) becomes

$$\frac{dx(t)}{dt} = x(t);$$

or

$$x(t) = e^t x_0.$$

Thus, the characteristics diverge (see Figure 8.2.4); if one solves the equation

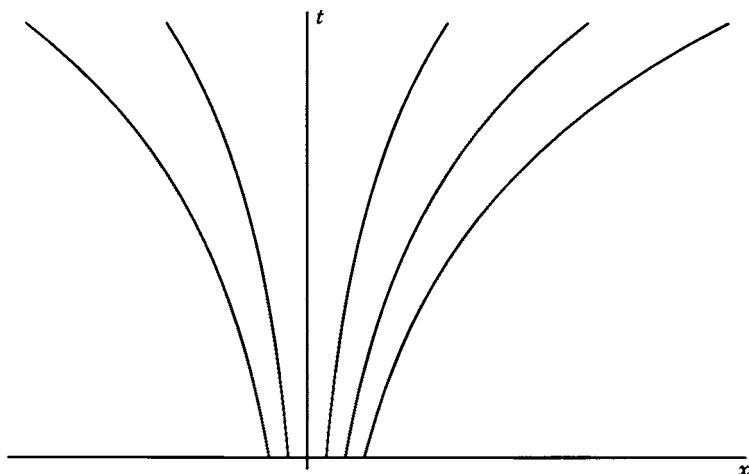


Figure 8.2.4. Diverging characteristics.

along a fixed set of characteristics, then the distance between the computed points increases as t increases. This is adequate because the solution

$$u(x, t) = f(xe^{-t})$$

becomes smoother with time since the derivative

$$u_x = f' e^{-t},$$

decays exponentially. Also, in every finite interval $|x| \leq a$,

$$\lim_{t \rightarrow \infty} u(x, t) = f(0).$$

Finite difference methods on a fixed grid will yield good results as t increases. In fact, one could decrease the number of gridpoints with time.

EXAMPLE 8.2.2. If $a = x$, then Eq. (8.2.3a) becomes

$$\frac{dx(t)}{dt} = -x(t),$$

or

$$x(t) = e^{-t} x_0.$$

Now the characteristics converge (see Figure 8.2.5) and the resulting solution

$$u(x, t) = f(xe^t)$$

becomes rougher with time, because the derivative

$$u_x = f' e^t$$

grows exponentially. Thus, finite difference methods will only provide accurate answers if one increases the number of points exponentially with time.

None of these problems occur if one uses the method of characteristics, that is, if one calculates the characteristics by an ODE solver and then uses the fact that u is constant along the characteristic. Also, the distance between computed points will decrease exponentially.

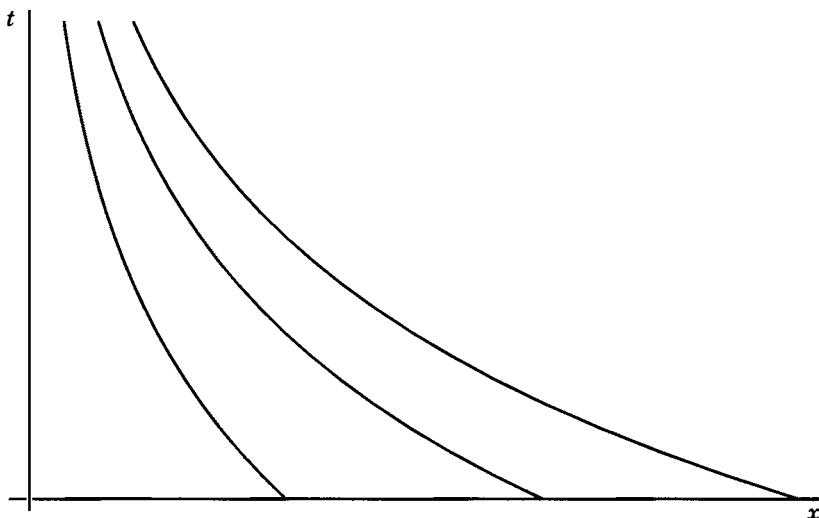


Figure 8.2.5. Converging characteristics.

We can also solve general scalar initial value problems

$$\begin{aligned} u_t &= a(x, t)u_x + b(x, t)u + F(x, t), \\ u(x, 0) &= f(x), \end{aligned} \quad (8.2.6)$$

by the same method. On a characteristic, Eq. (8.2.6) becomes a nonlinear system of ordinary differential equations

$$\begin{aligned} \frac{dx}{dt} &= -a(x, t), & x(0) &= x_0, \\ \frac{du}{dt} &= b(x, t)u + F(x, t), & u(0) &= f(x_0). \end{aligned} \quad (8.2.7)$$

This system can be solved by a Runge–Kutta or a multistep method.

The method of characteristics seems to be optimal for the homogeneous equation (8.1.1) with regard to accuracy. However, we may have difficulty with an inhomogeneous equation. Consider, for example,

$$u_t = -xu_x + \sin x, \quad (8.2.8)$$

which is equivalent to

$$\begin{aligned}\frac{dx}{dt} &= x, \\ \frac{du}{dt} &= \sin x,\end{aligned}$$

that is,

$$x(t) = e^t x_0, \quad \frac{du}{dt} = \sin(e^t x_0).$$

It is not difficult to obtain an accurate solution. However, the computational points diverge exponentially, and we may need to know the solution between computational points. In this case, the solution does not become smoother with time, and interpolation may be inaccurate. This difficulty can be overcome using the method of characteristics on a grid. We consider this later.

If we were to use the method of characteristics to solve Eq. (8.1.1) with the initial step function (8.1.2) we would find it easy to obtain any desired accuracy. The strength of the method of characteristics comes from the fact that we are integrating the solution along lines on which it is smooth. We are not differencing across discontinuities.

Next we consider systems

$$\begin{aligned}u_t &= A(x, t)u_x + B(x, t)u + F(x, t), \\ u(x, 0) &= f(x).\end{aligned}\tag{8.2.9}$$

We assume that the eigenvalues λ of A are real and that there is a smooth transformation $S = S(x, t)$ such that

$$S^{-1}AS = \Lambda, \quad \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix}.$$

Introducing $v = S^{-1}u$ as a new variable gives us

$$Sv_t + S_tv = ASv_x + AS_xv + BSv + F;$$

that is,

$$v_t = \Lambda v_x + \tilde{B}v + \tilde{F}, \quad \tilde{B} = S^{-1}(BS + AS_x - S_t), \quad \tilde{F} = S^{-1}F.\tag{8.2.10}$$

If $\tilde{B} \equiv 0$, then Eq. (8.2.10) reduces to a set of scalar equations

$$v_t^{(i)} = \lambda_i(x, t)v_x^{(i)} + \tilde{F}^{(i)}, \quad i = 1, 2, \dots, m, \quad (8.2.11)$$

which can be solved by the method of characteristics. The system (8.2.10) can then be solved using the iteration

$$v_t^{[j+1]} = \Lambda v_x^{[j+1]} + \tilde{\tilde{F}}^{[j]}, \quad \tilde{\tilde{F}}^{[j]} = \tilde{B}v^{[j]} + \tilde{F}^{[j]}. \quad (8.2.12)$$

Thus, the general case can also be solved using the method of characteristics. Observe that there are now m different sets of characteristics

$$\frac{dx^{(i)}}{dt} = -\lambda_i(x, t), \quad x^{(i)}(0) = x_0^{(i)}, \quad i = 1, 2, \dots, m, \quad (8.2.13)$$

and that these characteristics do not change from iteration to iteration.

As an example, we consider the system

$$\begin{aligned} u_t &= -xu_x + v, \\ v_t &= xv_x - u. \end{aligned} \quad (8.2.14)$$

We calculate u and v along the divergent u characteristics of Figure 8.2.4 and along the convergent v characteristics of Figure 8.2.5, respectively. In the first equation, we need to know v along the u characteristics. In practice, one can obtain those values by interpolation from the values on the v characteristics. This poses no difficulty because the v characteristics are convergent. However, for the second equation, the interpolation of u from values on the v characteristics can be a problem because those characteristics are divergent.

EXERCISES

- 8.2.1.** Give an estimate of $|\partial u / \partial x|$ for the solutions of Eq. (8.2.8), and compare it to the solutions of $u_t = -xu_x$.
- 8.2.2.** Write a program that solves Eq. (8.2.8) by the method of characteristics. Compute the solution $v_j(t_n)$ on a regular grid using linear interpolation, and show that the accuracy deteriorates as t increases.
- 8.2.3.** The solution of a scalar hyperbolic initial-value problem is uniquely defined by the concept of characteristics, even if the initial data is discontinuous. Prove that this solution is equivalent to the generalized solution defined in Section 4.8.
- 8.2.4.** Write a program that solves Eq. (8.2.14) by the method of characteristics as described above. Use initial data with compact support.

8.3. METHOD OF CHARACTERISTICS IN SEVERAL SPACE DIMENSIONS

Scalar equations

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{\nu=1}^d a_\nu(\mathbf{x}, t) \frac{\partial u}{\partial x^{(\nu)}} + b(\mathbf{x}, t)u + F(\mathbf{x}, t), \\ u(x, 0) &= f(x),\end{aligned}\tag{8.3.1}$$

can be solved in the same way as the one-dimensional problem. Using the notation $\mathbf{a} = (a_1, \dots, a_d)$, $\mathbf{x} = (x_1, \dots, x_d)$ the characteristics are the solution of the system

$$\frac{d\mathbf{x}}{dt} = -\mathbf{a}(\mathbf{x}, t), \quad \mathbf{x}(0) = \mathbf{x}_0, \tag{8.3.2}$$

and, on the characteristics, Eq. (8.3.1) becomes

$$\frac{du}{dt} = b(\mathbf{x}, t)u + F(\mathbf{x}, t), \quad u(0) = f(\mathbf{x}_0). \tag{8.3.3}$$

The behavior of the characteristics can be very complicated. For example, the characteristics of

$$u_t = -xu_x + yu_y$$

are

$$x(t) = e^t x_0, \quad y(t) = e^{-t} y_0.$$

They diverge in the x direction and converge in the y direction.

If the matrices $A_\nu(\mathbf{x}, t)$ in the hyperbolic system (6.4.1) are all diagonal (i.e., if the components of u are only coupled through lower order terms), then the characteristics are well-defined for every component of u , and we can proceed as in the one-dimensional case. General systems such as Eq. (6.4.1) cannot easily be solved by the method of characteristics, because one can not diagonalize all of the matrices A_ν with a single transformation $S(\mathbf{x}, t)$ in general.

However, sometimes one splits the differential operator so that partial steps can be taken using characteristics. As an example, we consider the Euler equations for incompressible flow, where the density ρ is assumed to be constant. Formally, we obtain the new reduced equations by letting $\rho \equiv 1$ in the original Euler equations (4.6.4). In two space dimensions, the new system is

$$u_t + uu_x + vu_y + p_x = 0, \quad (8.3.4a)$$

$$v_t + uv_x + vv_y + p_y = 0, \quad (8.3.4b)$$

$$u_x + v_y = 0. \quad (8.3.4c)$$

A more convenient system for computation is obtained by differentiating Eq. (8.3.4a) with respect to x and Eq. (8.3.4b) with respect to y and adding the two. The new equation replaces Eq. (8.3.4c), and we get

$$\begin{aligned} u_t + uu_x + vu_y + p_x &= 0, \\ v_t + uv_x + vv_y + p_y &= 0, \\ p_{xx} + p_{yy} &= -((u_x)^2 + 2u_yv_x + (v_y)^2). \end{aligned} \quad (8.3.5)$$

Locally we can solve Eq. (8.3.5) by the iteration

$$\begin{aligned} u_t^{[n+1]} + u^{[n]}u_x^{[n+1]} + v^{[n]}u_y^{[n+1]} + p_x^{[n]} &= 0, \\ v_t^{[n+1]} + u^{[n]}v_x^{[n+1]} + v^{[n]}v_y^{[n+1]} + p_y^{[n]} &= 0, \\ p_{xx}^{[n]} + p_{yy}^{[n]} &= -((u_x^{[n]})^2 + 2u_y^{[n]}v_x^{[n]} + (v_y^{[n]})^2). \end{aligned}$$

Thus, we consider p in the first two equations as a given function and, therefore, the first two equations are scalar equations with the same characteristics

$$\frac{d}{dt}x^{[n+1]} = u^{[n]}, \quad \frac{d}{dt}y^{[n+1]} = v^{[n]}.$$

EXERCISES

- 8.3.1.** Consider the solution of $u_t = -xu_x + yu_y$ at $t = t_0$ in the square $0 \leq x, y \leq 1$. What is the domain of dependence at $t = 0$? Draw a picture showing the distribution of the initial data required to define the solution on a uniform grid $\{x_i, y_j\}$ at $t = 10$.

8.4. METHOD OF CHARACTERISTICS ON A REGULAR GRID

In the previous sections, we have shown how a solution can be computed by using ODE methods along the characteristics. If this is done, the solution is, in general, obtained at points that have no regular distribution in the computational domain. Recall the examples in Section 8.2. For practical reasons, one often wants to have the solution represented on a regular grid. This can be achieved by interpolation after the computation is completed. The interpolation can also

be done as part of the method at each time step; we now discuss this technique. These methods are called semi-Lagrangian methods.

First consider the model equation $u_t = u_x$, and let v_j^n denote the computed value at the gridpoint (x_j, t_n) . To find a value v_j^{n+1} at the new time level, the intersection (x_*, t_n) of the characteristic with the previous time level is needed. In general, this is not a gridpoint, and an interpolation formula is used to approximate $v = v_*$. The simplest formula is obtained by using linear interpolation between neighboring points. The characteristic is moving a distance k in the t direction during one time step. We assume that it intersects the previous time level in the interval $(x_\nu, x_{\nu+1})$ at a distance h_* from x_ν [i.e., $(\nu - j)h + h_* = k$] (see Figure 8.4.1).

The interpolation formula is

$$v_*^n = \frac{h_* v_{\nu+1}^n + (h - h_*) v_\nu^n}{h},$$

and $v_j^{n+1} = v_*^n$ yields the simplest version of this modified method of characteristics

$$v_j^{n+1} = v_\nu^n + \frac{h_*}{h} (v_{\nu+1}^n - v_\nu^n). \quad (8.4.1)$$

This method is a special form of difference method; but, because x_* is not necessarily a neighboring point (or equal) to x_j , it is not of the usual form.

A stability analysis finds

$$\hat{Q} = e^{i(\nu - j)\xi} (1 + \alpha(e^{i\xi} - 1)), \quad \alpha = h_*/h,$$

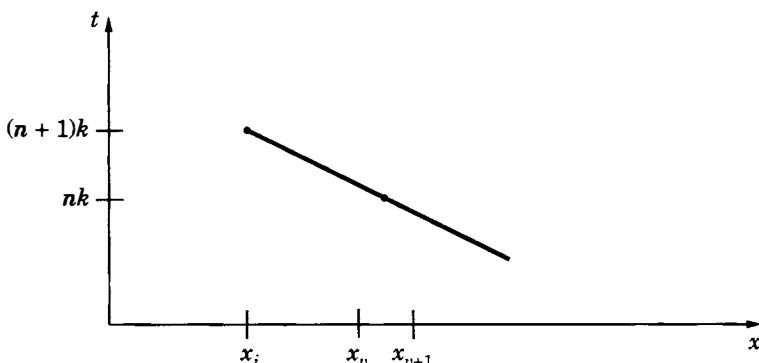


Figure 8.4.1. The method of characteristics ($\nu - j = 2$).

with

$$|\hat{Q}|^2 = 1 - 2\alpha(1 - \alpha)(1 - \cos \xi).$$

By definition, $0 < \alpha \leq 1$; therefore, $|\hat{Q}| \leq 1$, and the method is unconditionally stable.

Next we generalize to variable coefficients and consider

$$u_t + a(x, t)u_x + b(x, t)u = F(x, t), \quad a(x, t) < 0. \quad (8.4.2)$$

The characteristic $x(t)$ is defined by

$$\frac{dx}{dt} = a(x, t), \quad (8.4.3)$$

and with d/dt denoting differentiation along such a curve, Eq. (8.4.2) can be written in the form

$$\frac{du}{dt} + b(x, t)u = F(x, t). \quad (8.4.4)$$

Assume that the solution is known at $t = t_n$. Let $\Gamma(x, t)$ denote the characteristic passing through the point (x, t) , and let (x_*, t_n) be the intersection of $\Gamma(x_j, t_{n+1})$ with the line $t = t_n$. The trapezoidal rule for Eq. (8.4.3) is

$$x_j - x_* = \frac{k}{2} (a(x_j, t_{n+1}) + a(x_*, t_n)). \quad (8.4.5)$$

This is a nonlinear equation for the unknown x_* . Because a very good initial approximation is available,

$$x_*^{[0]} = x_j - ka(x_j, t_{n+1}), \quad (8.4.6)$$

Newton's method can be used to solve it efficiently.

When x_* is known, the trapezoidal rule can be used to approximate Eq. (8.4.4):

$$v_j^{n+1} - v_*^n = -\frac{k}{2} (b_*^n v_*^n + b_j^{n+1} v_j^{n+1}) + \frac{k}{2} (F_*^n + F_j^{n+1}). \quad (8.4.7)$$

The point x_* will not generally fall on a gridpoint. The functions b and F may be known for all x, t ; in that case only v_*^n need be computed. We use quadratic interpolation here. Let ν be the index such that $x_\nu < x_* \leq x_{\nu+1}$, with $x_* - x_\nu = h_*$.

Then the points $x_\nu, x_{\nu+1}, x_{\nu+2}$ are used for interpolation and the formula is

$$\begin{aligned} v_*^n &= \frac{1}{2} (1 - \alpha)(2 - \alpha)v_\nu^n + \alpha(2 - \alpha)v_{\nu+1}^n - \frac{\alpha}{2} (1 - \alpha)v_{\nu+2}^n, \\ \alpha &= \frac{h_*}{h} \leq 1. \end{aligned} \quad (8.4.8)$$

The complete algorithm for one time step is as follows:

1. Compute x_* from Eq. (8.4.5) using Newton's method with the initial approximation defined by Eq. (8.4.6).
2. For each j compute v_j^{n+1} using Eq. (8.4.7), where v_*^n is defined by Eq. (8.4.8).

To establish stability, we assume, without loss of generality, that a is a negative constant and $b = F = 0$. Then, it follows from Eq. (8.4.3) that $x_* > x_j$; that is, $\nu \geq j$ in the formula

$$\begin{aligned} v_j^{n+1} &= \frac{1}{2} (1 - \alpha)(2 - \alpha)v_\nu^n + \alpha(2 - \alpha)v_{\nu+1}^n - \frac{\alpha}{2} (1 - \alpha)v_{\nu+2}^n, \\ 0 < \alpha &\leq 1. \end{aligned} \quad (8.4.9)$$

In this case, we obtain the amplification factor

$$\hat{Q} = e^{i(\nu-j)\xi} \left(\frac{1}{2} (1 - \alpha)(2 - \alpha) + \alpha(2 - \alpha)e^{i\xi} - \frac{\alpha}{2} (1 - \alpha)e^{2i\xi} \right), \quad (8.4.10)$$

and a straightforward calculation shows that

$$|\hat{Q}| \leq 1,$$

for $0 < \alpha \leq 1$. We note that the method is unconditionally stable; that is, the time step k can be chosen arbitrarily. The method looks like an explicit difference method, and unconditional stability may seem surprising. However, it is not a contradiction of earlier results. There will be a growing number of gridpoints between x_j and x_ν as the mesh ratio k/h increases and the method cannot be classified as explicit, even if only three points are used at the previous time level. The essential fact is that these three points are chosen so that the domain of dependence (which is just a curve in the x, t plane) is always included in the expanding stencil.

If k is chosen such that $k|a| \leq h$, then the scheme is a regular explicit difference scheme

$$\begin{aligned} v_j^{n+1} &= \frac{1}{2} (1 - \alpha)(2 - \alpha)v_j^n + \alpha(2 - \alpha)v_{j+1}^n - \frac{\alpha}{2} (1 - \alpha)v_{j+2}^n, \\ \alpha &= k|a|/h. \end{aligned} \quad (8.4.11)$$

To calculate the order of accuracy, we rewrite the scheme as

$$\frac{v_j^{n+1} - v_j^n}{k} = |a| \left(D_+ - \frac{h}{2} D_+^2 + \frac{k|a|}{2} D_+^2 \right) v_j^n.$$

A Taylor expansion about (x_j, t_n) yields a truncation error $ku_{tt}/2$ on the left-hand side, which is canceled by the last term on the right hand side ($u_{tt} = a^2 u_{xx}$). Similarly, the one-sided operator D_+ yields a truncation error $hu_{xx}/2$ that is canceled by the second term. Accordingly, the scheme has only truncation error terms of order (h^2+k^2) , which shows that second-order accuracy is automatically attained with the method of characteristics if quadratic interpolation is used for v_* .

For the general case with arbitrary k and variable coefficients $a(x, t)$ and $b(x, t)$, the method is still second-order accurate. This can be proved by first considering Eq. (8.4.5) as a second-order approximation of Eq. (8.4.3), which determines the x_* points. The interpolation formula for v_*^n yields an $\mathcal{O}(h^3)$ error if the point x_* is exact. The perturbation of this point introduced by the numerical method computing x_* is locally $\mathcal{O}(h^3)$, and the total interpolation error is $\mathcal{O}(h^3)$ in each step. The method has a second-order global error.

It should be mentioned that the sign of the coefficient $a(x, t)$ may be different in different parts of the domain. The only modification of the algorithm required, if the solution x_* of Eq. (8.4.5) satisfies the inequality $x_{\nu-1} \leq x_* < x_\nu \leq x_j$ (corresponding to $a > 0$), is that a quadratic interpolation formula using the points $x_{\nu-2}, x_{\nu-1}, x_\nu, x_j$ is substituted for Eq. (8.4.8).

The method described here is based on an interpolation formula that uses only points on the same side of x_j as the incoming characteristic at (x_j, t_{n+1}) , even if $j = \nu$. This seems natural, because the whole method is based on tracing the domain of dependence. Furthermore, for nonlinear problems, when shocks may be present, it is advantageous to use information from only one side of the shock (see Section 8.8). However, from a stability point of view, there is nothing that prohibits use of points $x_{\nu-1}, x_\nu$, or $x_{\nu+1}$ (for $a < 0$). Actually, if a is constant and $k|a| < h$, this translation changes the scheme (8.4.11) into the Lax–Wendroff method.

We again discuss the generalization to systems in one space dimension and consider

$$u_t + A(x, t)u_x = 0. \quad (8.4.12)$$

The basic idea is to separate the various characteristics from each other and to

integrate the corresponding combinations of variables along these characteristic curves. The vector u has m components and, because the system is hyperbolic, we can find m left eigenvectors $\phi_i(x, t)$, such that

$$\phi_i^T(x, t)A(x, t) = \lambda_i(x, t)\phi_i^T(x, t), \quad i = 1, \dots, m, \quad (8.4.13)$$

where the λ_i are the eigenvalues of A . After multiplying Eq. (8.4.12) by ϕ_i^T , we obtain

$$\phi_i^T(u_t + \lambda_i(x, t)u_x) = 0, \quad i = 1, \dots, m. \quad (8.4.14)$$

The m families of characteristics are defined by

$$\frac{dx}{dt} = \lambda_i(x, t), \quad i = 1, \dots, m. \quad (8.4.15)$$

For each family, we choose that characteristic $\Gamma_i(x_j, t_{n+1})$ which passes through the point (x_j, t_{n+1}) and the trapezoidal rule becomes [as in Eq. (8.4.5)]

$$x_j - x_{*i} = \frac{k}{2}(\lambda_i(x_j, t_{n+1}) + \lambda_i(x_{*i}, t_n)), \quad (8.4.16)$$

where x_{*i} is the intersection of $\Gamma_i(x_j, t_{n+1})$ with $t = t_n$. When the m points x_{*i} are known, the corresponding vectors v_{*i}^n are computed from Eq. (8.4.8), where ν , h_* , and α now depend on i . The vectors v_j^{n+1} are then computed using the trapezoidal rule applied to Eq. (8.4.14) differentiated along each characteristic

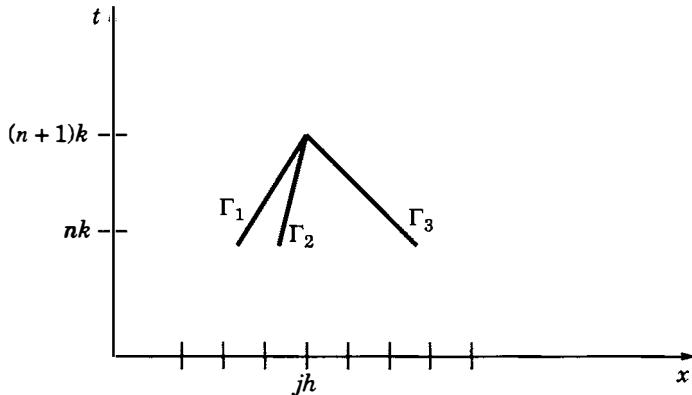
$$(\phi_i^T(x_{*i}, t_n) + \phi_i^T(x_j, t_{n+1}))(v_j^{n+1} - v_{*i}^n) = 0, \quad i = 1, \dots, m. \quad (8.4.17)$$

This is an $m \times m$ system for the unknowns $v_j^{(i)n+1}$, $i = 1, \dots, m$, which can be solved by a direct method. Figure 8.4.2 shows the situation for $m = 3$.

The generalization to systems with lower order terms can be done by iteration as discussed in Section 8.3. It can also be done by adding the extra terms (premultiplied by ϕ_i^T) to Eq. (8.4.14). The approximation (8.4.17) is modified, but the system (8.4.16) is unchanged.

If the system is nonlinear, then the eigenvalues λ_i also depend on u . Hence, the systems (8.4.16) and (8.4.17) become coupled, and they must be solved simultaneously.

The method of characteristics is generally much more difficult to apply to problems in several space dimensions. However, it is easily generalized for scalar problems as discussed in Section 8.3. The equation

Figure 8.4.2. Method of characteristics for a 3×3 -system.

$$u_t + a(x, y, t)u_x + b(x, y, t)u_y + c(x, y, t)u = F(x, y, t) \quad (8.4.18)$$

is rewritten in the form

$$\frac{du}{dt} + c(x, y, t)u = F(x, y, t), \quad (8.4.19)$$

along $\Gamma(t)$, where the coordinates $x(t)$, and $y(t)$ of the curve $\Gamma(t)$ are defined by

$$\frac{dx}{dt} = a(x, y, t), \quad \frac{dy}{dt} = b(x, y, t). \quad (8.4.20)$$

The system (8.4.20) is solved using the trapezoidal rule, and Newton's method is applied at each time step with (x_i, y_j) known at $t = t_{n+1}$ and with the point (x_*, y_*) as the unknown intersection of the characteristic curve with the plane $t = t_n$. To find the corresponding value v_{i*}^n , interpolation is required, and this can be done in many ways. As an example, consider negative values of a and b , and let (x_ν, y_μ) be the point such that

$$0 < h_{1*} = x_* - x_\nu \leq h_1,$$

$$0 < h_{2*} = y_* - y_\mu \leq h_2,$$

where h_1 and h_2 are the regular step sizes. Interpolated values on the line $y = y_*$ are computed using

$$v_{i*}^n = \frac{1}{2} (1 - \alpha_2)(2 - \alpha_2)v_{i\mu}^n + \alpha_2(2 - \alpha_2)v_{i,\mu+1}^n - \frac{\alpha_2}{2} (1 - \alpha_2)v_{i,\mu+2}^n,$$

$$\alpha_2 = h_{2*}/h_2, \quad i = \nu, \nu + 1, \nu + 2. \quad (8.4.21)$$

These values are then used to define v_*^n by applying the same formula in the x direction

$$\begin{aligned} v_*^n &= \frac{1}{2} (1 - \alpha_1)(2 - \alpha_1) v_{\nu,*}^n + \alpha_1(2 - \alpha_1) v_{\nu+1,*}^n - \frac{\alpha_1}{2} (1 - \alpha_1) v_{\nu+2,*}^n, \\ \alpha_1 &= h_{1*}/h_1. \end{aligned} \quad (8.4.22)$$

Figure 8.4.3 illustrates the procedure.

As mentioned above, the method of characteristics is not very convenient for general systems in several space dimensions. However, in many applications as, for example, in fluid dynamics, the system has the form

$$u_t + au_x + bu_y + Pu = 0, \quad (8.4.23)$$

where a and b are scalar functions and P is a differential operator with matrix coefficients. In Section 8.3, it was shown how the method of characteristics can be used with iteration. Another possibility is to use a Strang-type splitting described in Section 5.4. Let $Q_1(k)$ be the operator that solves the system

$$u_t + au_x + bu_y = 0 \quad (8.4.24)$$

using the method of characteristics (for m identical components) over one time step, and let

$$v^{n+1} = Q_2(k)v^n \quad (8.4.25)$$

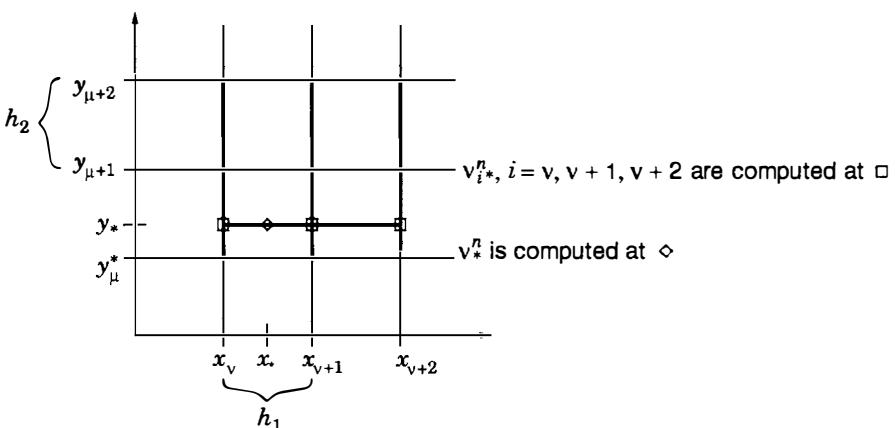


Figure 8.4.3. Interpolation scheme in two space dimensions.

be a difference approximation to $u_t + Pu = 0$. Then the system (8.4.23) can be approximated by

$$v^{n+1} = Q_1(k/2)Q_2(k)Q_1(k/2)v^n, \quad (8.4.26)$$

and the results in Section 5.4 concerning stability and accuracy are valid.

One case in which this type of splitting may be advantageous is where part (8.4.24) of the differential equation represents the fast waves in the system. The unconditional stability of $Q_1(k)$ then makes it possible to run the full scheme (8.4.26) with larger time steps than could be used with an explicit method.

EXERCISES

- 8.4.1. Prove that $|\hat{Q}| \leq 1$, with \hat{Q} as defined in Eq. (8.4.10).
- 8.4.2. Define the modified method of characteristics if Eq. (8.4.8) is replaced by quadratic interpolation using the points $x_{\nu-1}$, x_ν , and $x_{\nu+1}$. Prove unconditional stability if the differential equation is $u_t + au_x = 0$. Also prove that the method is equivalent to the Lax–Wendroff method in this case, if $k|a| < h$.

8.5. REGULARIZATION USING VISCOSITY

We have seen that variable coefficients can cause converging characteristics. As $t \rightarrow \infty$, discontinuities can occur. As an example, consider the problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sin x \frac{\partial u}{\partial x}, & -\pi \leq x \leq \pi, \quad t \geq 0, \\ u(x, 0) &= \sin x. \end{aligned} \quad (8.5.1)$$

The behavior of the characteristics near $x = 0$ is determined by the coefficient $\sin x$. Near $x = 0$, $\sin x$ behaves essentially like x . Figure 8.5.1 shows the solution at $t = 4\pi/3$ computed with a very fine mesh.

Thus, we have the same situation here that we discussed in Section 8.2. After some time, the solution develops a large gradient at $x = 0$, and we expect numerical difficulties. We approximate $\partial/\partial x$ by the fourth-order accurate operator $Q_4 = D_0(I - (h^2/6)D_+D_-)$ (see Section 3.1) and solve the resulting ordinary differential equations by the classical fourth-order Runge–Kutta method on the

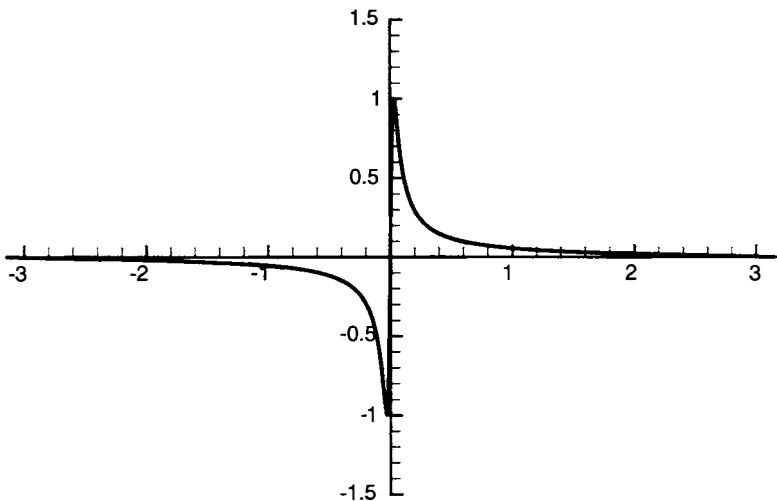


Figure 8.5.1.

interval $-\pi \leq x \leq \pi$. Figure 8.5.2 shows the numerical solution at $t = 4\pi/3$ with $h = 2\pi/120$. Difficulties near $x = 0$ are apparent.

We now alter Eq. (8.5.1) to

$$\begin{aligned} \frac{\partial w}{\partial t} &= \sin x \frac{\partial w}{\partial x} + \epsilon \frac{\partial^2 w}{\partial x^2}, \\ w(x, 0) &= \sin x. \end{aligned} \quad (8.5.2)$$

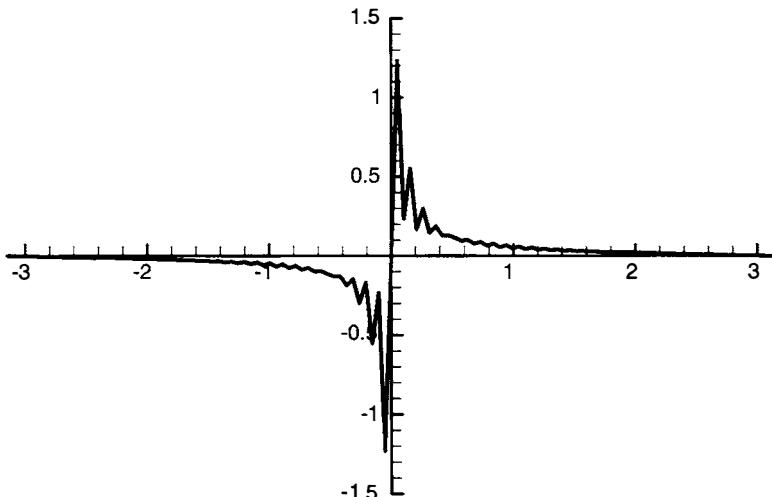


Figure 8.5.2.

Here $\epsilon > 0$ is a small constant. We expect that w will be close to the solution u of (8.5.1) in those regions where the derivatives of u are small compared to $1/\epsilon$.

One can prove that the derivatives of w satisfy an estimate

$$\|\partial^{p+q} w(\cdot, t)/\partial x^p \partial t^q\|_\infty \leq C_{p,q} \epsilon^{-(p+q)},$$

(recall Theorem 5.5.1). Here $C_{p,q}$ depends only on p, q and not on t . Thus, in contrast to Eq. (8.5.1), there are uniform bounds for the derivatives of the solution. Therefore, difference methods can be used to solve Eq. (8.5.2). The error is of the form $h^q D^{q+1} w$, where D denotes differentiation. Thus, if $h^q \epsilon^{-(q+1)} \ll 1$, then the error is small. We have also solved Eq. (8.5.2) by the method of lines using the fourth-order Runge–Kutta method in time. In space, we replaced

$$\frac{\partial}{\partial x} \rightarrow Q_4, \quad \frac{\partial^2}{\partial x^2} \rightarrow D_+ D_- \left(I - \frac{h^2}{12} D_+ D_- \right).$$

The truncation error is $\mathcal{O}(h^4 D^5 w + \epsilon h^4 D^6 w) = \mathcal{O}(h^4 \epsilon^{-5})$ in space. In Figure 8.5.3, we show w for $\epsilon = 0.0025$ and for the same h and t as shown in Figure 8.5.2. It is clear that, away from a small neighborhood of $x = 0$, w is close to u . There is an error near $x = 0$ partly because w is not equal to u , and partly because there are not enough gridpoints to approximate w well. An accurate approximation of w would require $h \approx 0.1 \epsilon^{5/4}$.

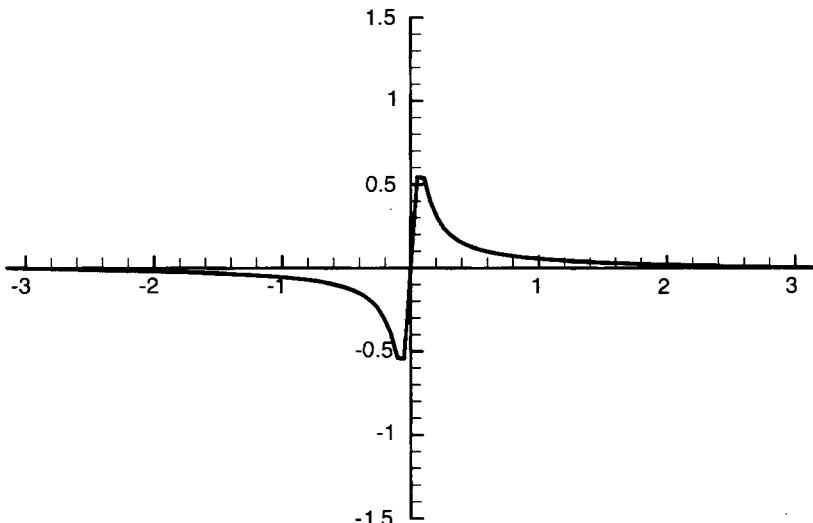


Figure 8.5.3. A regularized solution.

EXERCISES

- 8.5.1.** Solve Eq. (8.5.1) using the Lax–Wendroff scheme with decreasing step sizes h . Explain why the oscillations become more severe as h decreases.

8.6. THE INVISCID BURGERS' EQUATION

In this section, we consider nonlinear problems. In this case, discontinuous solutions can be generated spontaneously in finite time from smooth initial data. We consider Burgers' equation as an example. First, consider the Cauchy problem

$$u_t + uu_x = 0, \quad -\infty < x < \infty, \quad (8.6.1a)$$

$$u(x, 0) = f(x). \quad (8.6.1b)$$

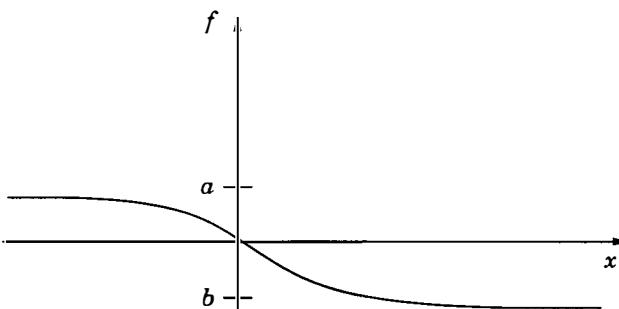
We assume that $f(x) \in C^\infty$ is a strictly monotonic function such that

$$\lim_{x \rightarrow -\infty} f(x) = a, \quad \lim_{x \rightarrow +\infty} f(x) = b, \quad (8.6.2)$$

where either $a > 0 > b$ or $a < 0 < b$ (see Figure 8.6.1).

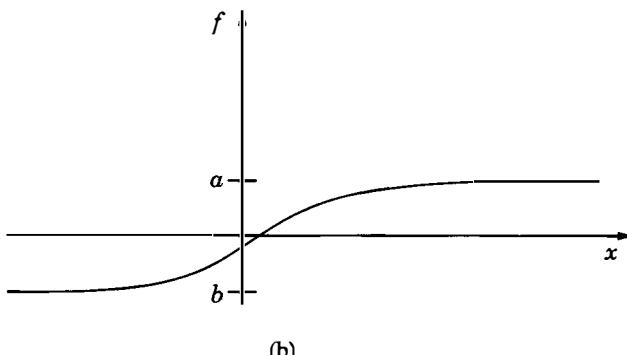
We think of the coefficient u in Eq. (8.6.1) as a given function, and, therefore, we can solve the problem by the method of characteristics. The characteristics are given by

$$\frac{dx}{dt} = u, \quad x(0) = x_0. \quad (8.6.3)$$



(a)

Figure 8.6.1a. $a > 0 > b$.

Figure 8.6.1b. $a < 0 < b$.

Along the characteristics

$$u(x, t) = f(x_0) = \text{constant} \quad (8.6.4)$$

Therefore, Eq. (8.6.3) becomes very simple. The characteristics are the straight lines

$$x(t) = f(x_0)t + x_0. \quad (8.6.5)$$

Now, assume that $a > 0 > b$. Then there is a point \bar{x} such that

$$f(x_0) > 0 \quad \text{for } x_0 < \bar{x}, \quad f(x_0) < 0 \quad \text{for } x_0 > \bar{x}.$$

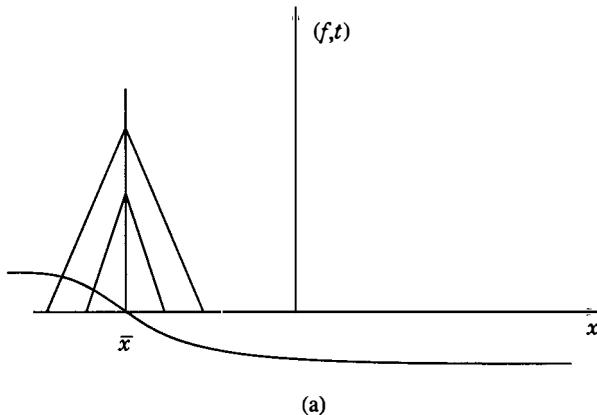
Therefore, the characteristics to the left of \bar{x} and the characteristics to the right of \bar{x} will intersect (see Figure 8.6.2a).

By Eq. (8.6.4), u is constant along the characteristic lines. If two characteristics intersect, then u will not, in general, be a unique function, and the solution will not exist. Also, just before the intersection the solution has a very large gradient, because the solution is converging to different values.

We can calculate the blow up time [i.e., the first time when two different characteristics arrive at the same point (x, t)]. In this case, there are x_0, \bar{x}_0 such that

$$x = f(x_0)t + x_0 = f(\bar{x}_0)t + \bar{x}_0;$$

that is,

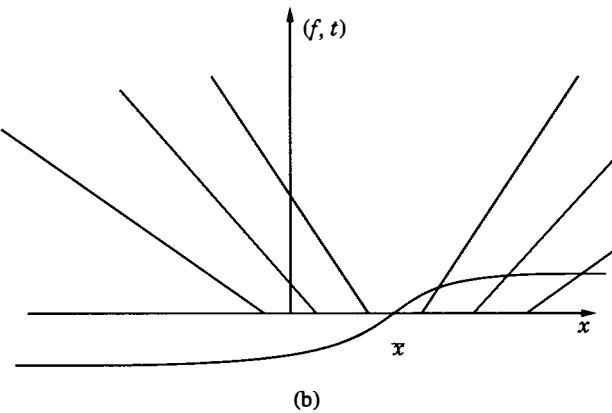
Figure 8.6.2a. $a > 0 > b$.

$$t = -\frac{\bar{x}_0 - x_0}{f(\bar{x}_0) - f(x_0)} = -\frac{1}{f'(\xi)},$$

where ξ lies between x_0 and \bar{x}_0 . Thus, the blow up occurs at

$$T = \min_{-\infty < x < \infty} \left(-\frac{1}{f'(x)} \right). \quad (8.6.6)$$

For $t > T$, the solution forms a shock wave that is defined as the limit of viscous solutions in the next section. Now assume that $a < 0 < b$. In this case, the characteristics diverge (see Figure 8.6.2b) and the solution becomes smoother with time.

Figure 8.6.2b. $a < 0 < b$.

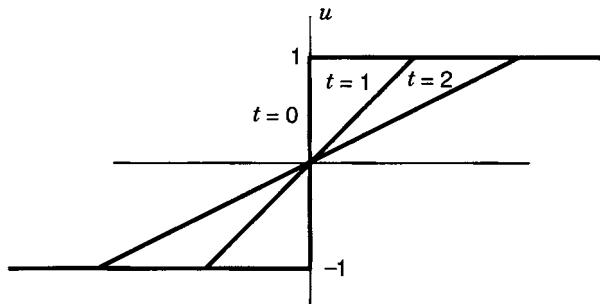


Figure 8.6.3.

Consider a sequence $f_\nu(x)$ of monotonically increasing initial data with

$$\lim_{\nu \rightarrow \infty} f_\nu(x) = \begin{cases} -1, & \text{for } x \leq 0, \\ 1, & \text{for } x > 0. \end{cases} \quad (8.6.7)$$

Using the characteristics, a simple calculation shows that the corresponding solutions converge to

$$u(x, t) = \begin{cases} 1, & \text{for } t \leq x, \\ x/t, & \text{for } -t \leq x \leq t, \\ -1, & \text{for } x \leq -t. \end{cases} \quad (8.6.8)$$

[Observe that x/t is a solution of Eq. (8.6.1a).] Thus, for every fixed t , the solution has the form shown in Figure 8.6.3. $u(x, t)$ is called a rarefaction wave.

We have already studied similar effects for linear equations in Section 8.2. The difference in the linear case is that the characteristics never intersect. They can only come arbitrarily close to each other. Therefore, the solution exists for all time although its derivatives can become arbitrarily large.

Now consider initial data with a finite number of maxima and minima (see Figure 8.6.4). As long as the solution exists, maxima and minima remain

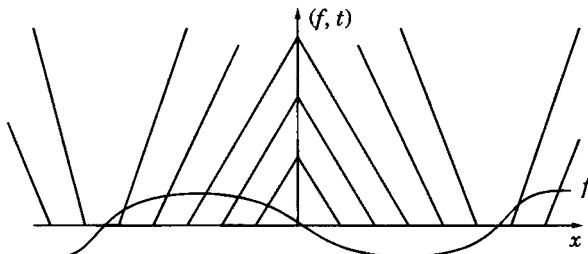


Figure 8.6.4.

maxima and minima. On intervals where u is monotonically increasing or monotonically decreasing, the characteristics diverge or converge, respectively. In the monotonically decreasing parts, the characteristics eventually intersect and the solution does not exist beyond the blow up.

8.7. THE VISCOUS BURGERS' EQUATION AND TRAVELING WAVES

Corresponding to Section 8.5, we regularize the inviscid equation (8.6.1) by adding dissipation, and consider the Cauchy problem

$$\begin{aligned} u_t + uu_x &= \epsilon u_{xx}, \quad -\infty < x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{8.7.1}$$

where $\epsilon > 0$ is a small constant. We also assume that $f(x)$ and its derivatives are uniformly bounded with respect to x .

Theorem 5.5.1 shows that for every i, j there exists a constant $C_{i,j}$ such that

$$\left\| \frac{\partial^{i+j} u}{\partial x^i \partial t^j} \right\|_\infty \leq C_{i,j} \epsilon^{-(i+j)}. \tag{8.7.2}$$

As we will see later, this is the best we can hope for with general initial data. However, for monotonically increasing data, we can do better. One can prove the following theorem.

Theorem 8.7.1. *Consider Eq. (8.7.1) with initial data as in Eq. (8.6.2), where $a < b$ (i.e., f is monotonically increasing). For every i, j , there exists a constant $C_{i,j}^*$, which does not depend on ϵ , such that for all t*

$$\left\| \frac{\partial^{i+j} u}{\partial x^i \partial t^j} \right\|_\infty \leq C_{i,j}^*.$$

Therefore, the solution of Eq. (8.7.1) converges to the solution of the inviscid equation (8.6.1) as $\epsilon \rightarrow 0$.

In Figure 8.7.1, we show the solution of Eq. (8.7.1) calculated with initial data as in Eq. (8.6.2) with $a = 1$ and $b = -1$. The solution is shown at $t = 200k$ with $\epsilon = 0.1$ using a fourth-order accurate centered approximation in space and the fourth-order Runge-Kutta method in time with $k = 0.1h$ and $h = 0.02$. The approximation is satisfactory when $k \ll \epsilon$ and $h \ll \epsilon$. In view of Eq. (8.7.2), this is consistent with truncation error analysis.

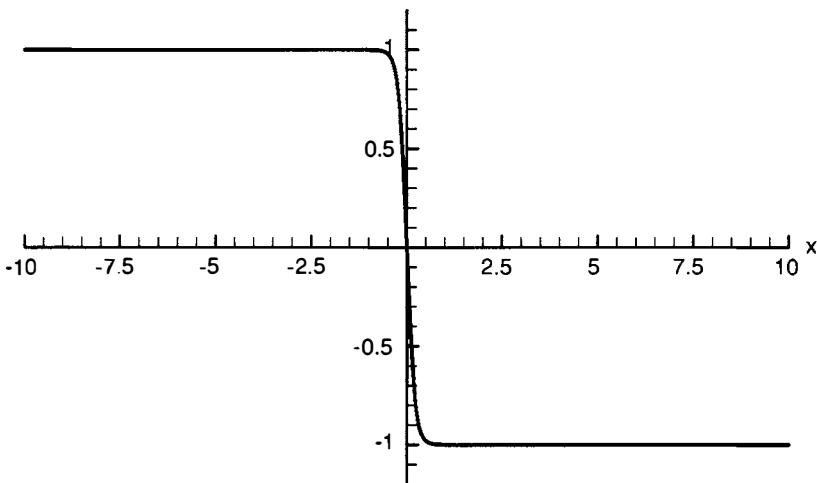


Figure 8.7.1.

In this case the solution converges to steady state. In general, if $a > b$, then the solution converges to a *traveling wave*, that is, there is a constant s such that

$$u(x, t) = \varphi(x - st), \quad \lim_{x \rightarrow -\infty} \varphi = a, \quad \lim_{x \rightarrow +\infty} \varphi = b, \quad a > b. \quad (8.7.3)$$

We shall prove the next theorem.

Theorem 8.7.2. *The viscous Burgers' equation has traveling wave solutions satisfying Eq. (8.7.3).*

Proof. We introduce a moving coordinate system

$$\begin{aligned} z &= x - st, \\ t' &= t. \end{aligned}$$

Neglecting the prime sign and using the same notation u for the dependent variable, Eq. (8.7.1) becomes

$$u_t - su_z + uu_z = \varepsilon u_{zz},$$

which can be written in the so called conservation form

$$u_t - \left(su - \frac{1}{2} u^2 \right)_z = \varepsilon u_{zz}. \quad (8.7.4)$$

In this new frame, a traveling wave of the form of Eq. (8.7.3) becomes stationary. Thus, φ must satisfy the ordinary differential equation

$$(-s\varphi + \frac{1}{2}\varphi^2)' = \epsilon\varphi'', \quad \lim_{z \rightarrow -\infty} \varphi = a, \quad \lim_{z \rightarrow +\infty} \varphi = b. \quad (8.7.5)$$

We can integrate Eq. (8.7.5) and obtain

$$\epsilon\varphi' = -s\varphi + \frac{\varphi^2}{2} + d, \quad d = \text{constant} \quad (8.7.6)$$

The constants s and d are defined by boundary conditions: $\lim_{z \rightarrow \pm\infty} \varphi' = 0$ implies

$$-sa + \frac{a^2}{2} + d = -sb + \frac{b^2}{2} + d = 0.$$

That is,

$$\begin{aligned} s &= \frac{b^2 - a^2}{2(b - a)} = \frac{b + a}{2}, \\ d &= \frac{(b + a)a}{2} - \frac{a^2}{2} = \frac{ab}{2}. \end{aligned}$$

With these values Eq. (8.7.6) becomes

$$\begin{aligned} \epsilon\varphi' &= -\frac{a+b}{2}\varphi + \frac{\varphi^2}{2} + \frac{ab}{2}, \\ &= \frac{1}{2} \left(\varphi - \frac{(a+b)}{2} \right)^2 - \frac{(a-b)^2}{8}. \end{aligned} \quad (8.7.7)$$

We now choose a value z_0 such that

$$\varphi(z_0) = \frac{a+b}{2}. \quad (8.7.8)$$

Such a point must exist because of the boundary conditions in Eq. (8.7.5). We then integrate Eq. (8.7.7) in the forward and backward direction of z with $z = z_0$ as a starting point. Clearly,

$$\varphi' < 0 \quad \text{for} \quad \left| \varphi - \frac{a+b}{2} \right| < \frac{|a-b|}{2}, \quad \text{and} \quad \varphi' = 0 \quad \text{for} \quad \varphi = a, b.$$

If $a > b$, φ decreases toward $\varphi = b$ when $z \rightarrow \infty$ and increases toward a as $z \rightarrow -\infty$. Therefore, Eqs. (8.7.7) and (8.7.8) give us the desired solution if $a > b$. If $a < b$, we cannot obtain a solution because the solutions of Eq. (8.7.7) are monotonically decreasing for $a < \varphi < b$. The traveling waves obtained for $a > b$ are called *viscous shock waves*.

The solution of Eq. (8.7.7) is not unique because we can choose z_0 arbitrarily. We can determine z_0 in the following way. Consider the initial-value problem (8.2.5). For $t \rightarrow \infty$, it will converge to a traveling wave φ of the above type. Let φ_0 be the traveling wave with $z_0 = 0$. For $s = (b+a)/2$, the differential equation gives us

$$\begin{aligned} \frac{\partial}{\partial t} \int_{-\infty}^{\infty} (u - \varphi_0) dz &= \int_{-\infty}^{\infty} u_t dz = \int_{-\infty}^{\infty} \left(\left(su - \frac{u^2}{2} \right)_z + \epsilon u_{zz} \right) dz \\ &= sb - \frac{b^2}{2} - sa + \frac{a^2}{2} = 0. \end{aligned}$$

Therefore,

$$\int_{-\infty}^{\infty} (u - \varphi_0) dz = \int_{-\infty}^{\infty} (f - \varphi_0) dz.$$

Because u rapidly converges to φ as $t \rightarrow \infty$, we obtain that

$$\int_{-\infty}^{\infty} (\varphi - \varphi_0) dz = \int_{-\infty}^{\infty} (f - \varphi_0) dz$$

determines the correct z_0 and $\varphi(z)$.

We can normalize φ by introducing

$$\psi = \frac{\varphi - \frac{1}{2}(a+b)}{\frac{1}{2}(a-b)}, \tag{8.7.9}$$

as a new variable. Equations (8.7.7) and (8.7.8) then become

$$\begin{aligned} \tilde{\epsilon} \psi' &= \psi^2 - 1, & \tilde{\epsilon} &= \frac{4\epsilon}{a-b}, \\ \psi(z_0) &= 0. \end{aligned} \tag{8.7.10}$$

Equation (8.7.10) can be solved explicitly. We obtain

$$\psi = \frac{e^{-z/\tilde{\epsilon}} - e^{z/\tilde{\epsilon}}}{e^{-z/\tilde{\epsilon}} + e^{z/\tilde{\epsilon}}}. \quad (8.7.11)$$

Equation (8.7.11) shows that the steep gradient of the traveling wave is confined to an interval of width $\tilde{\epsilon} \log \tilde{\epsilon}$ (i.e., there is an internal layer at $z = 0$). Except in this layer, a change of $\tilde{\epsilon}$ has very little effect on the solution. Also, in agreement with Theorem 8.7.1, differentiating Eq. (8.7.10) gives us

$$\|\psi\|_\infty = 1, \quad \left\| \frac{\partial \psi_z}{\partial z} \right\|_\infty = \frac{1}{\tilde{\epsilon}}, \quad \left\| \frac{\partial^j \psi_z}{\partial z^j} \right\|_\infty = \mathcal{O}(\tilde{\epsilon}^{-j}), \quad j = 0, 1, 2, \dots \quad (8.7.12)$$

The estimates (8.7.12) are sharper than those of Eq. (8.7.2), because they give us the correct dependence on the so called *shock strength* $a - b$. The stronger the shock, the thinner the internal layer and the larger the gradient.

We can now consider the limit process $\tilde{\epsilon} \rightarrow 0$. In this case, ψ converges to the step function

$$\psi = \begin{cases} +1, & \text{for } z < 0, \\ -1, & \text{for } z > 0. \end{cases}$$

The limits of viscous shock waves are called *inviscid shock waves*.

In Figure 8.7.2, we have calculated a viscous rarefaction wave, that is, we solve Eq. (8.7.1) with initial data from Eq. (8.6.7) and $\epsilon = 0.005$, $h = 0.02$, and $k = 0.1h$. We use the same method as in Figure 8.7.1. The result agrees well with Eq. (8.6.8). Due to the viscosity effect, the solution is C^∞ smooth for $t > 0$.

We can summarize our results. If the initial data are monotonically increasing, then the solution u converges to the solution of the inviscid equation as

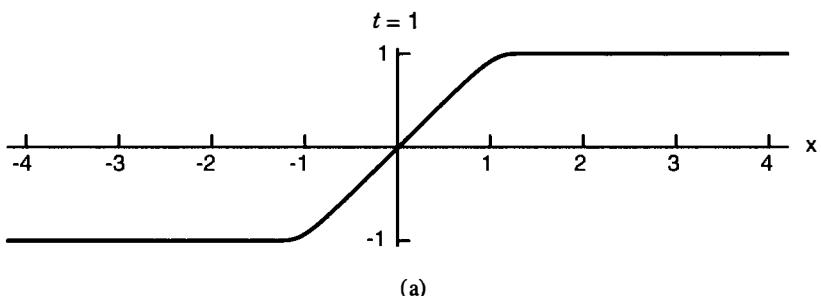
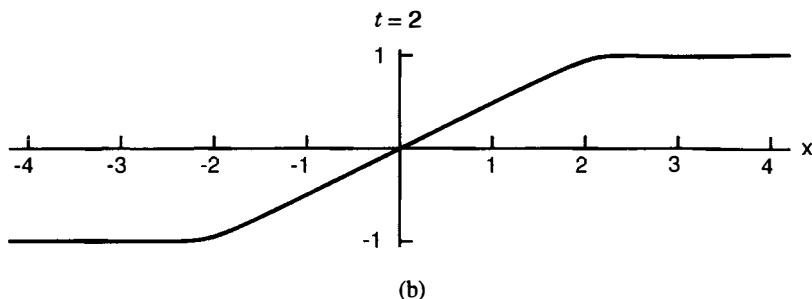


Figure 8.7.2. (a) $t = 1$.

Figure 8.7.2. (b) $t = 2$.

$\epsilon \rightarrow 0$. If $f(x)$ is of the form shown in Figure 8.6.1a, then the state $u_- = a$ will propagate to the right until it approaches the state $u_+ = b$ which has either been moving to the left ($b < 0$) or more slowly than u_- to the right. These states are then connected by a viscous layer and a traveling wave is formed. The traveling wave moves with speed $s = \frac{1}{2}(a + b)$. As $\epsilon \rightarrow 0$, this traveling wave converges to the step function

$$u = \begin{cases} a, & \text{for } x - st < z_0, \\ b, & \text{for } st > z_0. \end{cases}$$

For general initial data as in Figure 8.7.3, the monotonically increasing parts will remain smooth and the monotonically decreasing parts will be transformed into traveling waves that converge to “jump” functions as $\epsilon \rightarrow 0$. The traveling waves move with speed $\frac{1}{2}(u_+ + u_-)$, where u_- and u_+ are the values to the left and right of the viscous layer. We have computed the solution of Eq. (8.7.1) with the initial data displayed in Figure 8.7.3, $\epsilon = 0.005$, $h = 2\pi/250$, $k = 2h/3$. The result is shown in Figure 8.7.4 at $t = 1.5\pi$.

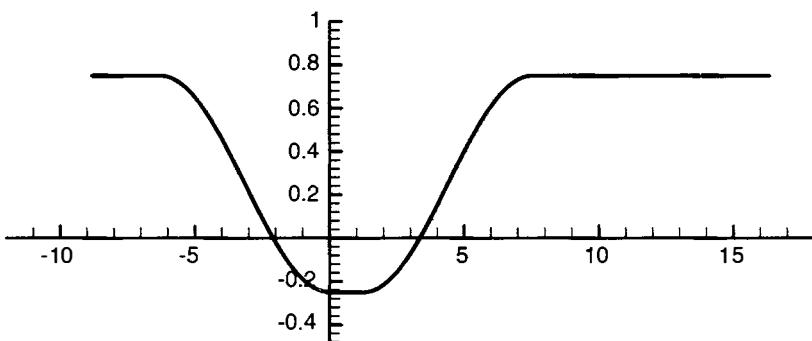


Figure 8.7.3. Initial data.

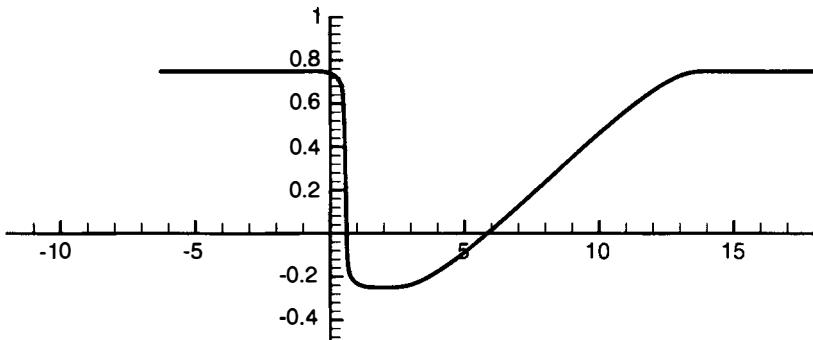


Figure 8.7.4. Solution at $t \equiv 1.5\pi$.

All the results above can be extended to more general equations

$$\begin{aligned} u_t + g(u)_x &= \epsilon u_{xx}, \quad -\infty < x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x). \end{aligned} \tag{8.7.13}$$

One can prove the following theorem.

Theorem 8.7.3. Consider Eq. (8.7.13) with initial data $f(x)$ satisfying Eq. (8.6.2). Assume that $g(u)$ is a convex function with

$$\frac{d^2g(u)}{du^2} > \delta > 0.$$

If $a > b$, then the solution of Eqs. (8.7.13) and (8.6.2) converges to a traveling wave with speed

$$s = \frac{g(a) - g(b)}{a - b}.$$

One does not need to calculate a traveling wave precisely to determine its speed. Assume that there is a traveling wave whose front at time t_0 is positioned at x_0 and at time $t_0 + \Delta t$ at $x_0 + s\Delta t$ (see Figure 8.7.5). Let $\delta > s\Delta t$ be a constant. We now integrate Eq. (8.7.13) over the interval $[x_0 - \delta, x_0 + \delta]$. During the time Δt , the integral increases by $s\Delta t(u_- - u_+)$. This is just an increase due to the state u_- moving into the interval less the state u_+ moving out of the interval. Thus, we obtain

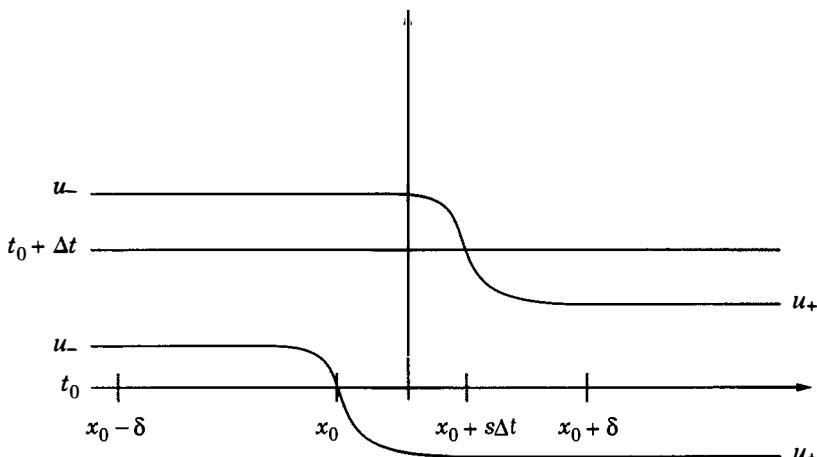


Figure 8.7.5.

$$s(u_- - u_+) = \frac{d}{dt} \int_{x_0 - \delta}^{x_0 + \delta} u \, dx = -(g(u_+) - g(u_-)) \\ + \epsilon(u_x(x_0 + \delta) - u_x(x_0 - \delta));$$

that is,

$$s = \frac{g(u_+) - g(u_-)}{u_+ - u_-} + \mathcal{O}(\epsilon).$$

REMARK. If $g(u) = \frac{1}{2}u^2$, then $s = (u_+ + u_-)/2 + \mathcal{O}(\epsilon)$, that is, we recover our previous result.

The interesting fact is that the speed does not depend on the detailed profile of the wave. This is because the derivative terms in Eq. (8.7.13) are in conservation form. If we have an equation of the form

$$u_t + a(x, t, u)u_x = \epsilon b(x, t, u, u_x)u_{xx},$$

we cannot proceed in the same way and the speed of a traveling wave may depend on the profile.

In the next section, we discuss numerical methods. A difference approximation is said to be in conservation form, or conservative, if it can be written as

$$\frac{dv_j}{dt} + \frac{H_{j+1} - H_j}{h} = \epsilon D_+ D_- v_j, \quad (8.7.14)$$

where $H_j = H(v_{j+p}, \dots, v_{j-q})$ is a gridfunction. For example, if $H_j = \frac{1}{4}(v_j^2 + v_{j-1}^2)$, we obtain Eq. (8.8.3). If the differential equation is in conservation form, consistent approximations need not be. The discussion above can be generalized to the discrete case. The conclusion is that the shock speed will not necessarily be obtained accurately with such approximations unless the shock profile is resolved using a very fine grid near the shock. This restriction is usually too severe. Thus, one should use conservative approximations.

The above construction can also be used for systems of conservation laws. Then u and g are vector functions with m components. We apply the construction above, component by component, obtaining m relations

$$s(u_+^{(\nu)} - u_-^{(\nu)}) = g^{(\nu)}(u_+) - g^{(\nu)}(u_-) + \mathcal{O}(\epsilon), \quad \nu = 1, 2, \dots, m, \quad (8.7.15)$$

which connect the state on both sides of the traveling wave to its speed. For $\epsilon = 0$, the conditions (8.7.15) are called the *Rankine–Hugoniot shock relations*.

As we have seen, as $\epsilon \rightarrow 0$, the solutions of our problem converge to limiting solutions with smooth sections separated by traveling jump functions. Such solutions are called *weak solutions* of the inviscid equation. There is a large amount of literature on weak solutions, which are usually defined directly using a so-called *weak formulation* and not as the limit of viscous solutions. We will not discuss that here and refer to the literature.

8.8. NUMERICAL METHODS FOR SCALAR EQUATIONS BASED ON REGULARIZATION

Equations for physical systems often contain dissipative terms to model physical processes such as diffusion or viscosity. If we think of Burgers' equation as a model of fluid flow, then it is natural to consider the viscous equation

$$u_t + \frac{1}{2}(u^2)_x = \nu u_{xx}. \quad (8.8.1)$$

(We have chosen to write ν instead of ϵ to emphasize that we are considering the naturally occurring viscosity; ϵ will denote artificial or numerical viscosity.) Unfortunately, ν is very small in many applications, for example, $\nu = 10^{-8}$. If we solve the problem using a difference approximation, truncation error analysis tells us that we must choose $h \ll \nu$. This is often impractical. So we might increase ν and solve

$$u_t + \frac{1}{2}(u^2)_x = \epsilon u_{xx} \quad (8.8.2)$$

instead with, say, $\epsilon \approx 10^{-2}$. In this case, the requirement $h \ll \epsilon$ is practically feasible. In the previous section, we have seen that the solutions of Eqs. (8.8.1) and (8.8.2) consist of smooth parts between traveling waves. In the smooth parts, the solutions differ by terms of order $\mathcal{O}(\epsilon)$. Also, the speed of the traveling waves is, to first approximation, independent of ϵ . The solutions differ in the thickness of the transition layers; $\mathcal{O}(\epsilon)$ instead of $\mathcal{O}(v)$. For these reasons, methods based on the concepts above often give reasonable and useful answers.

We discuss the procedure above in more detail, which will also explain why difficulties may arise. We approximate Eq. (8.8.2) by

$$\frac{dv_j}{dt} + \frac{1}{2} D_0 v_j^2 = \epsilon D_+ D_- v_j \quad (8.8.3)$$

and we want to calculate traveling waves. We only solve for stationary waves, that is, we want to determine the solution of

$$\epsilon D_+ D_- v_j = \frac{1}{2} D_0 v_j^2, \quad \lim_{j \rightarrow \infty} v_j = -a, \quad \lim_{j \rightarrow -\infty} v_j = a, \quad a > 0. \quad (8.8.4)$$

We can also write Eq. (8.8.4) in the form

$$\epsilon D_- (D_+ v_j) = \frac{1}{2} D_- (\frac{1}{2} (v_j^2 + v_{j+1}^2)).$$

Therefore,

$$\epsilon D_+ v_j - \frac{1}{4} (v_j^2 + v_{j+1}^2) = (\epsilon D_+ v_j - \frac{1}{4} (v_j^2 + v_{j+1}^2))_{j \rightarrow -\infty} = -\frac{a^2}{2}; \quad (8.8.5a)$$

that is,

$$v_{j+1} - \frac{h}{4\epsilon} v_{j+1}^2 = v_j + \frac{h}{4\epsilon} v_j^2 - \frac{h}{2\epsilon} a^2. \quad (8.8.5b)$$

We normalize Eq. (8.8.5b) and write it in the form

$$\begin{aligned} F(\tilde{v}_{j+1}) &:= \tilde{v}_{j+1} - \tau \tilde{v}_{j+1}^2 + \tau = \tilde{v}_j + \tau \tilde{v}_j^2 - \tau =: G(\tilde{v}_j), \\ \tilde{v} &= \frac{v}{a}, \quad \tau = \frac{ha}{4\epsilon}. \end{aligned} \quad (8.8.6)$$

The functions $F(\tilde{v}), G(\tilde{v})$ are parabolas with $F'' \equiv -\tau, G'' \equiv \tau$. They have their extreme values at $\tilde{v} = 1/(2\tau)$ and $\tilde{v} = -1/(2\tau)$, respectively, where $F = -G = \tau + 1/(4\tau)$. Also, $F(\pm 1) = G(\pm 1) = \pm 1$. There are two cases.

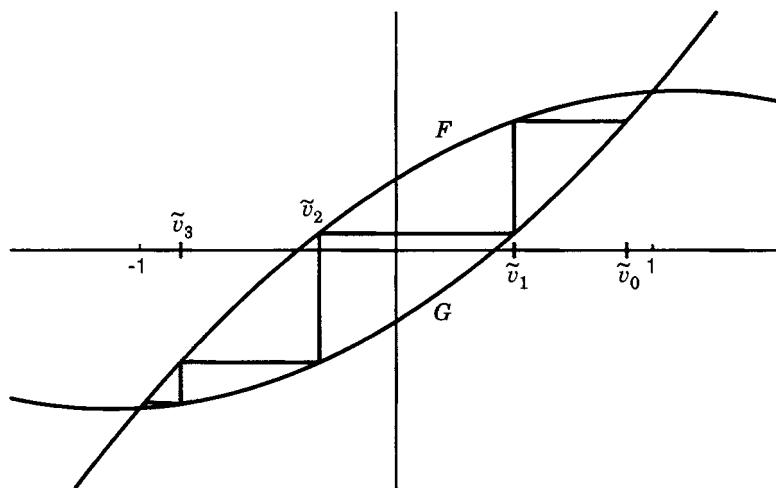
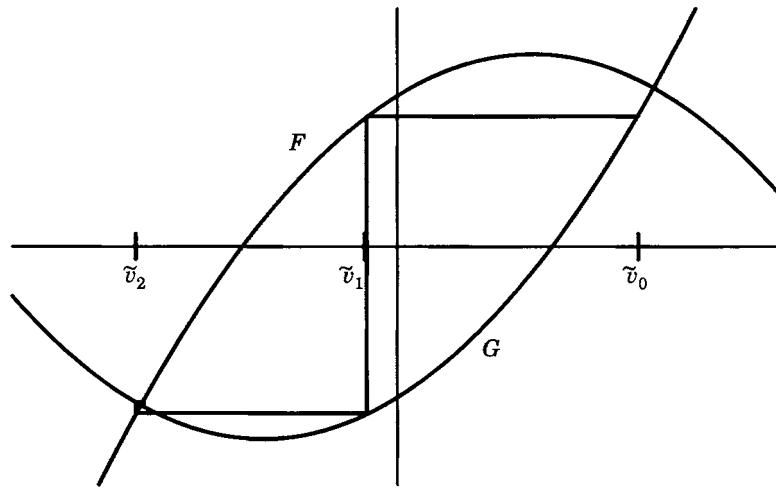


Figure 8.8.1.

CASE 1. $\tau \leq \frac{1}{2}$. Figure 8.8.1 clearly shows that, for any given \tilde{v}_0 with $-1 < \tilde{v}_0 < 1$, the solution \tilde{v}_j of Eq. (8.8.6) is monotonically decreasing toward -1 .

CASE 2. $\tau > \frac{1}{2}$. Figure 8.8.2b is a blowup of the area around the point of convergence. Now the convergence, in general, will not be monotone. The solution \tilde{v}_j will oscillate around -1 as j increases.



(a)

Figure 8.8.2a.

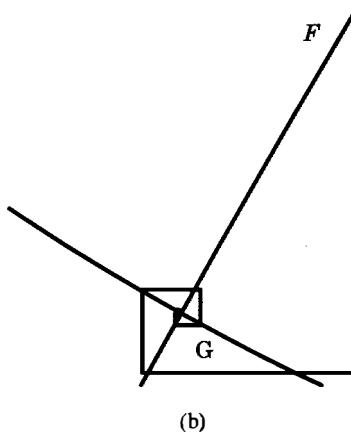


Figure 8.8.2b.

Starting with $\tilde{v}_0 = 0$ we have calculated $\tilde{v}_j, j = 0, 1, \dots$, for different values of τ (see Figure 8.8.3). Again, we see that, for $\tau < \frac{1}{2}$, the convergence is monotonic as j increases and, for $\tau > \frac{1}{2}$, it is oscillatory. This can also be seen by linearizing Eq. (8.8.6) around $\tilde{v} = -1$.

Let $\tilde{v} = -1 + v'$. If we neglect quadratic terms, we obtain

$$v'_{j+1} = \frac{1 - 2\tau}{1 + 2\tau} v'_j. \quad (8.8.7)$$

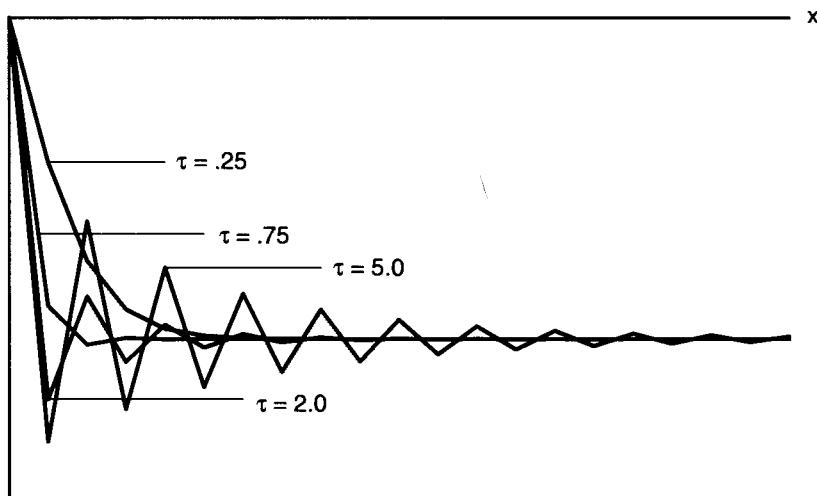


Figure 8.8.3.

Therefore, $\text{sign } v_{j+1} = -\text{sign } v_j$ if $\tau > \frac{1}{2}$. Furthermore, Eq. (8.8.7) also shows that the convergence is slow as j increases if $0 < \tau \ll 1$ or $\tau \gg 1$. If the convergence is slow, then the discontinuity will affect many gridpoints. We can also solve Eq. (8.8.4) for $j \leq 0$. The result is completely symmetric.

The calculation above and Eq. (8.8.7) show that the discontinuity is particularly sharp if we choose the dissipation so that

$$\tau = \frac{1}{2} \quad \text{or} \quad \epsilon = \frac{ha}{2}.$$

In this case, Eq. (8.8.3) becomes

$$\frac{dv_j}{dt} + \frac{1}{2} D_0 v_j^2 = \frac{ha}{2} D_+ D_- v_j. \quad (8.8.8)$$

Observing that

$$D_0 = D_+ - \frac{1}{2} h D_+ D_- = D_- + \frac{1}{2} h D_+ D_-,$$

we can write Eq. (8.8.8) in two equivalent forms:

$$\frac{dv_j}{dt} + \frac{1}{2} D_+ v_j^2 = \frac{h}{2} D_+ D_- \left(a v_j + \frac{v_j^2}{2} \right) \quad (8.8.9a)$$

or

$$\frac{dv_j}{dt} + \frac{1}{2} D_- v_j^2 = \frac{h}{2} D_+ D_- \left(a v_j - \frac{v_j^2}{2} \right). \quad (8.8.9b)$$

For large j , we have $v_j = -a + w_j$, $|w_j| \ll 1$, and, therefore,

$$\left| \frac{h}{2} D_+ D_- \left(a v_j + \frac{v_j^2}{2} \right) \right| = \left| \frac{h}{4} D_+ D_- w_j^2 \right| \ll 1, \quad j \gg 1.$$

Correspondingly,

$$\left| \frac{h}{2} D_+ D_- \left(a v_j - \frac{v_j^2}{2} \right) \right| \ll 1, \quad \text{for } j \ll -1.$$

Thus, Eqs. (8.8.9a) and (8.8.9b) are closely related to the so-called *upwind* difference schemes

$$\frac{dv_j}{dt} + \frac{1}{2} D_+ v_j^2 = 0 \quad \text{for } v_j < 0, \quad \frac{dv_j}{dt} + \frac{1}{2} D_- v_j^2 = 0 \quad \text{for } v_j > 0. \quad (8.8.10)$$

One can use Eq. (8.8.8) as an integration scheme. The choice of $\epsilon = ha/2$ was based on an analysis of the stationary solution, so we cannot expect the method to be useful if the shock speed is not small. In that case, one should locally introduce a moving coordinate system as discussed earlier.

Instead of using Eq. (8.8.10), where one uses a different method when v changes sign, one can use one formula. This leads to so-called *flux-splitting* methods. Introduce functions

$$g_+ = \begin{cases} v^2, & \text{if } v > 0, \\ 0, & \text{if } v \leq 0, \end{cases} \quad g_- = \begin{cases} 0, & \text{if } v \geq 0, \\ v^2, & \text{if } v < 0. \end{cases}$$

Then $v^2 = g_+ + g_-$ and, instead of the formulation (8.8.10), we can use

$$\frac{dv_j}{dt} + \frac{1}{2} (D_+(g_-)_j + D_-(g_+)_j) = 0. \quad (8.8.11)$$

There are some drawbacks with these methods. To obtain sharp waves, one must know the shock strength and use a regularization coefficient $\epsilon = \mathcal{O}(h)$. Thus, in the smooth part of the solution, the method is only first-order accurate.

One can avoid these problems by either using a “switch” or a more complicated dissipation term (or both). Instead of Eq. (8.8.3) we consider

$$\frac{dv_j}{dt} + \frac{1}{2} D_0 v_j^2 = \epsilon D_+(\varphi_j D_- v_j). \quad (8.8.12)$$

One way to build in the shock strength is to use

$$\varphi_j = \sum_{\nu=-p}^{p-1} h |D_+ v_{j+\nu}|. \quad (8.8.13)$$

Typically, one uses $p = 2$, because discontinuities are not smeared over more than three grid points when efficient methods are used. Then $\varphi_j \approx a - b$ near the discontinuity, and we can achieve the correctly scaled viscosity by choosing $\epsilon = h/4$. Where the solution is smooth, the dissipation term is of order $\mathcal{O}(h^2)$. Thus, the approximation is second-order accurate in the smooth regions. In Figure

8.8.4, we show a calculation obtained using Eq. (8.8.12) with $\epsilon = h/4$. It is free of oscillations.

We can also combine Eq. (8.8.12) with a switch. The idea is to only use dissipation near a discontinuity. To do this, one must construct a monitor M that can locate discontinuities. One possibility is

$$M_j = |v_{j+p} - v_{j-p}|, \quad p > 0.$$

Typically, $p = 2$. Then we can select a threshold \bar{M} and define

$$\varphi_j = \begin{cases} \sum_{\nu=-p}^{p-1} h|D_+ v_{j+\nu}|, & \text{if } M_j \geq \bar{M}, \\ 0, & \text{if } M_j < \bar{M}. \end{cases}$$

We now consider the more general equation (8.7.13) with initial data (8.6.2), where $a > b$. One can construct traveling waves as before. In the moving coordinate system, where $z \rightarrow x - st$, Eq. (8.7.13) becomes

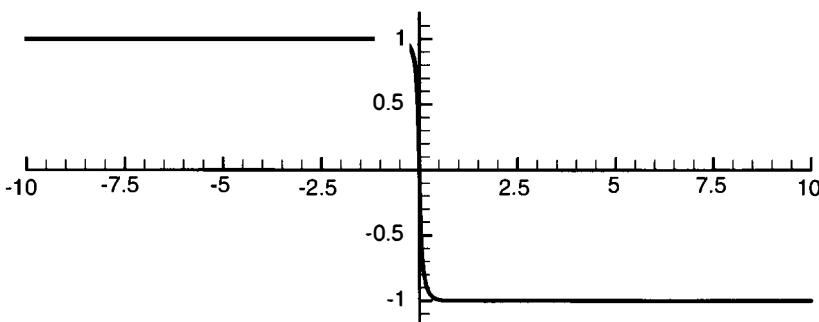
$$u_t + (g_1(u))_z = \epsilon u_{zz}, \quad (8.8.14)$$

where

$$g_1(u) := g(u) - su, \quad s = \frac{g(a) - g(b)}{a - b}. \quad (8.8.15)$$

For so called *weak shocks*, that is, $0 < a - b \ll 1$,

$$s \approx g'(u_0), \quad u_0 = \frac{a+b}{2}.$$



$t = 100k$

Figure 8.8.4.

Therefore, since $g'_1(u_0) \approx 0$,

$$\begin{aligned} (g_1(u))_z &\approx \left(g_1(u_0) + g'_1(u_0)(u - u_0) + g''_1(u_0) \frac{(u - u_0)^2}{2} \right)_z, \\ &\approx \frac{g''_1(u_0)}{2} ((u - u_0)^2)_z. \end{aligned}$$

By assumption, $g'' > 0$. Therefore, Eq. (8.8.14) behaves essentially like Burgers' equation. In particular, the optimal amount of dissipation is proportional to the shock strength $a - b$. If $a - b \gg 1$, then it is not possible in general to relate the optimal amount of dissipation to the shock strength. Corresponding to the relation (8.8.5a) for Burgers' equation, it follows that the profile of the discrete traveling wave is the solution of

$$\epsilon D_+ v_j - \frac{g_1(v_j) + g_1(v_{j+1})}{2} = -\bar{g}, \quad \bar{g} = g_1(a) = g_1(b) \quad (8.8.16)$$

using Eq. (8.8.15) as $j \rightarrow \infty$, $v_j \rightarrow b$. Introducing $v = b + w$ as a new variable and neglecting quadratic terms gives us the linearized equation

$$\epsilon D_+ w_j = g'_1(b) \left(\frac{w_j + w_{j+1}}{2} \right),$$

that is,

$$w_{j+1} = \frac{1 + \frac{h}{2\epsilon} g'_1(b)}{1 - \frac{h}{2\epsilon} g'_1(b)} w_j = \frac{1 - \frac{h}{2\epsilon} |g'_1(b)|}{1 + \frac{h}{2\epsilon} |g'_1(b)|} w_j. \quad (8.8.17)$$

[Observe that $g'' > 0$, and, therefore, for some number ξ , $b \leq \xi \leq a$, we have $g'_1(b) = g'(b) - s = g'(b) - g'(\xi) < 0$.] The optimal dissipation in front of the traveling wave is now

$$\epsilon = \frac{h}{2} |g'_1(b)|.$$

Correspondingly, we should choose

$$\epsilon = \frac{h}{2} |g'_1(a)|$$

behind the discontinuity. These two values can be vastly different and need not be related to the shock strength. We can incorporate these values into the difference approximation by using

$$\frac{dv_j}{dt} + D_0 g_1(v_j) = \frac{h}{2} D_- (|g'_1(v_j)| D_+ v_j). \quad (8.8.18)$$

REMARK. In practice, one replaces $|g'_1(v_j)|$ by $[1/(2p+1)] \sum_{\nu=-p}^p |g'_1(v_{j+\nu})|$. Again, recall that Eq. (8.8.18) is defined in the moving coordinate system. In the original coordinate system, the approximation is not useful if the shock speed is large.

EXERCISES

8.8.1. Prove that Eq. (8.8.16) has a unique monotone solution provided ϵ/h is sufficiently large.

8.9. REGULARIZATION FOR SYSTEMS OF EQUATIONS

Consider a system of conservation laws

$$u_t + g(u)_x = \epsilon u_{xx} \quad ((8.9.1))$$

where u and g are m vectors. Again we are interested in traveling waves

$$u(x, t) = \varphi(x - st), \quad \lim_{x \rightarrow \pm\infty} \varphi = u_{\pm}. \quad (8.9.2)$$

As in the scalar case, we introduce a moving coordinate system

$$z = x - st, \quad t' = t,$$

and we obtain, after dropping the prime sign,

$$u_t + (g(u) - su)_z = \epsilon u_{zz}. \quad (8.9.3)$$

Now φ is the solution of the stationary system

$$(g(\varphi) - s\varphi)_z = \epsilon \varphi_{zz}, \quad \lim_{z \rightarrow \pm\infty} \varphi(z) = u_{\pm}. \quad (8.9.4)$$

Integrating Eq. (8.9.4) gives us the Rankine–Hugoniot relation

$$g(u_+) - su_+ = g(u_-) - su_-. \quad (8.9.5)$$

Thus, the end states u_+ and u_- cannot be chosen arbitrarily. We now discuss the choice of u_+, u_- in more detail.

For many systems occurring in real applications, extensive analysis and numerical computations have been done. Based on those results, certain properties of the solutions are well understood. For example, there are solutions of the same form as in the scalar case; outside an internal layer of width $\mathcal{O}(\epsilon |\log \epsilon|)$, they converge rapidly to the end states u_\pm . Therefore, for large $|z|$, we can replace Eq. (8.9.3) by

$$u_t + (A(u_+) - sI)u_z = \epsilon u_{zz}, \quad \text{for } z \gg 1 \quad (8.9.6a)$$

or

$$u_t + (A(u_-) - sI)u_z = \epsilon u_{zz}, \quad \text{for } z \ll -1. \quad (8.9.6b)$$

Here $A = \partial g / \partial u$ is the Jacobian evaluated at u_+ and u_- , respectively. We now assume that the systems (8.9.6) are strictly hyperbolic for $\epsilon = 0$, that is, we can order the eigenvalues $\lambda^\pm - s$ of $A(u_\pm) - sI$ in ascending order

$$\lambda_1^\pm - s < \lambda_2^\pm - s < \dots < \lambda_m^\pm - s. \quad (8.9.7)$$

We first assume that $\lambda_j^\pm - s \neq 0$ for all j . The case $\lambda_j^\pm - s = 0$ will be discussed for the Euler equations at the end of this section. The corresponding eigenvectors are linearly independent, and, therefore, there are nonsingular matrices S_\pm such that

$$S_\pm(A(u_\pm) - sI)S_\pm^{-1} = \begin{pmatrix} \Lambda_1^\pm - sI & 0 \\ 0 & \Lambda_2^\pm - sI \end{pmatrix}. \quad (8.9.8)$$

Here $\Lambda_1^\pm - sI > 0$ and $\Lambda_2^\pm - sI < 0$ are positive and negative definite diagonal matrices. We introduce new variables in Eq. (8.9.6) by

$$v = Su, \quad S = S_+ \quad \text{for } z \gg 1 \quad \text{and} \quad S = S_- \quad \text{for } z \ll -1,$$

and obtain

$$\begin{aligned} v_t^I + (\Lambda_1^\pm - sI)v_z^I &= \epsilon v_{zz}^I, \\ v_t^{II} + (\Lambda_2^\pm - sI)v_z^{II} &= \epsilon v_{zz}^{II}, \end{aligned} \quad (8.9.9)$$

for $z \gg 1$ and $z \ll -1$, respectively. Observe that the dimension of Λ_1^\pm and Λ_2^\pm can depend on z where $z > 0$ or $z < 0$. Now consider the case $\epsilon = 0$. The components of v are the characteristic variables. They are also called the Riemann invariants. Let $z \gg 1$. Then, the components of v^I are constant along those characteristics that move to the right in the new coordinate system (z, t) . Therefore, the values v_+^I at $z = +\infty$ do not have any direct influence on the solution in the shock layer z . On the other hand, the components of v^{II} are constant along characteristics that move into the shock layer. Therefore, it is reasonable to describe

$$v_+^{II} = \lim_{z \rightarrow \infty} v^{II}. \quad (8.9.10a)$$

Correspondingly, for $z \ll -1$, we obtain that the components of v^I are constant along characteristics that move into the shock, and we describe

$$v_-^I = \lim_{z \rightarrow -\infty} v^I. \quad (8.9.10b)$$

Assume that the dimension of v_+^{II} is k and that of v_-^I is p . In the original variables, we obtain k linear relations between the components of u_+ and p linear relations between the components of u_- . Also, the m (nonlinear) relations of Eqs. (8.9.5) have to be satisfied. Thus, the $2m + 1$ variables u_+ and u_- and the shock speed s have to satisfy $m + k + p$ relations. Therefore, we require that

$$k + p = m + 1;$$

that is, the number of characteristics entering the shock shall be $m + 1$. This is called the entropy condition. Under reasonable assumptions, one can show that one can solve this system of equations.

Assume now that there is a traveling wave $\varphi(x - st)$. For large values of z , the function φ satisfies to first approximation

$$(A(u_+) - sI)\varphi_z = \epsilon\varphi_{zz}.$$

Introducing $\psi = S\varphi$ as new variables, we obtain for every component $\psi^{(\nu)}$

$$(\lambda_j^+ - s)\psi_z^{(\nu)} = \epsilon\psi_{zz}^{(\nu)}. \quad (8.9.11)$$

The general solution of Eq. (8.9.11) is given by

$$\psi^{(\nu)} = \sigma_1 + \sigma_2 e^{\frac{\lambda_j^+ - s}{\epsilon} z}.$$

We are interested in bounded solutions. Therefore, we have to distinguish between the two cases.

CASE 1. $\lambda_\nu^+ - s > 0$. Then necessarily $\sigma_2 = 0$ and

$$\psi^{(\nu)} = \sigma_1 = \psi_\infty^{(\nu)}.$$

CASE 2. $\lambda_\nu^+ - s < 0$. Then σ_2 is undetermined and

$$\psi^{(\nu)} = \psi_\infty^{(\nu)} + \sigma_2 e^{\frac{\lambda_\nu^+ - s}{\varepsilon} z}.$$

converges rapidly to $\psi_\infty^{(\nu)}$ as $z \rightarrow \infty$. Thus, in the original variables we obtain

$$u = u_+ + \mathcal{O}\left(\max_{\lambda_\nu^+ - s < 0} e^{\frac{\lambda_\nu^+ - s}{\varepsilon} z}\right). \quad (8.9.12)$$

The corresponding result holds for $z \ll -1$.

The conclusion from this analysis is the following: The Riemann invariants corresponding to characteristics coming out from the shock are constant, and they will cause no numerical difficulties. For the characteristics going into the shock, the corresponding Riemann invariants have a sharp layer. This is natural, since the value given at $z = \infty$ must suddenly adjust to the conditions at the shock after having traveled smoothly leftwards across the right half-plane. For the discrete approximation, the sharp layer causes an oscillatory solution if no extra dissipation term is added.

We now use a central difference method with an added artificial viscosity term. Our method of lines approximation in the moving coordinate system is

$$\frac{d}{dt} v_j + D_0(g(v_j) - sv_j) = D_+(M_j D_- v_j), \quad (8.9.13)$$

where M_j is an $m \times m$ positive definite matrix and both v_j and M_j are gridfunctions on the grid $x_j = jh$, $j = 0, \pm 1, \pm 2, \dots$

In the scalar case, our linear analysis showed that we could eliminate spurious oscillations if we choose the coefficient of the viscous term to be proportional to the local characteristic speed. The situation is more difficult here, because we generally have many different speeds. Suppose we choose $M = \varepsilon I$, where $\varepsilon \approx \max_{1 \leq \nu \leq m} |\lambda_\nu|$. Spurious oscillations will be suppressed, but components of the solution corresponding to small $|\lambda_\nu|$ will be smeared out excessively. This is a severe problem if the eigenvalues differ by orders of magnitude.

We now look at this situation in more detail. Suppose that we have a steady

solution with nearly constant states separated by a shock. We consider this solution outside the shock layer and linearize around the states on the left and right of the shock. After diagonalization of the resulting equations, the linearized steady problems are of the form

$$(J - sI)D_0 v_j = D_+ M_j D_- v_j, \quad (8.9.14)$$

where J is a constant matrix corresponding to $A = g'$, which is different on the left and right of the shock. Assuming that J has a complete set of eigenvectors, we can write

$$\Lambda = Q^{-1} J Q, \quad (8.9.15)$$

where Λ is diagonal and the columns of Q are the right eigenvectors of J . Set

$$v_j = Q w_j.$$

Then Eq. (8.9.14) can be written as

$$(\Lambda - sI)D_0 w_j = D_+(Q^{-1} M_j Q D_- w_j). \quad (8.9.16)$$

The components of w_j are the characteristic variables for this system. Because J is constant, Q and Q^{-1} are also constant on each side of the shock. If we determine M from

$$Q^{-1} M Q = \text{diag}(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m), \quad (8.9.17)$$

then Eq. (8.9.16) is an uncoupled system of m scalar equations approximating Eq. (8.9.11). Consider the right-hand side of the shock. Again, we have to distinguish between the two cases.

1. $\lambda_\nu^+ - s > 0$. As in the continuous case, the bounded solutions are constants and therefore no dissipation is needed.
2. $\lambda_\nu^+ - s < 0$. Then we obtain a scalar equation of the same form as in the previous section. We can suppress oscillations by choosing $\varepsilon_\nu \approx |\lambda_\nu^+ - s| h / 2$.

Thus, in the general case

$$\varepsilon_\nu = \begin{cases} \frac{1}{2} |\lambda_\nu - s| h, & \text{for ingoing characteristic variables,} \\ \bar{\varepsilon}, & \text{otherwise.} \end{cases}$$

Here $\bar{\epsilon}$ is a minimum level of dissipation, which one always should add to control noise.

We now consider the equations for gas dynamics as an example. In particular,

$$\begin{aligned}\rho_t + (\rho u)_x &= 0 \\ (\rho u)_t + (\rho u^2 + p)_x &= 0 \\ (\rho E)_t + ((\rho E + p)u)_x + \rho E &= 0\end{aligned}\tag{8.9.18}$$

for $-\infty < x < \infty$ and $t \geq 0$. The variables ρ , u , and E are the density, the velocity, and the total energy, respectively. The pressure p is determined from an equation of state

$$p = \left(\rho E - \frac{\rho u^2}{2} \right) (\gamma - 1),$$

where the constant γ is the ratio of specific heats.

The eigenvalues of the Jacobian of this system are $\lambda_1 = u$, $\lambda_2 = u + a$, and $\lambda_3 = u - a$, where $a = \sqrt{\gamma p / \rho}$ is the speed of sound. The so-called Riemann invariants that result from the diagonalization are

$$\begin{aligned}r_1 &= \frac{e}{\rho^{\gamma-1}} \\ r_2 &= u + \frac{2a}{\gamma-1} \\ r_3 &= u - \frac{2a}{\gamma-1}\end{aligned}$$

where $e = E - u^2/2$ is the internal energy. These quantities satisfy the differential equations

$$\frac{\partial}{\partial t} r_\nu + \lambda_\nu^\pm \frac{\partial}{\partial x} r_\nu = 0, \quad \nu = 1, 2, 3.$$

The values of the r_ν and λ_ν are determined by the states on either side of the shock.

Suppose we want to compute a steady-state solution that consists of a shock separating a supersonic region ($u > a$) on its left from a subsonic ($0 < u < a$) region on its right. Then

$$\lambda_2^- > \lambda_1^- > \lambda_3^- > 0$$

on the left and

$$\lambda_2^+ > \lambda_1^+ > 0 > \lambda_3^+$$

on the right. All of the characteristics go into the shock from the left and only one, λ_3 , goes into the shock from the right. The coefficients ϵ_ν can now be chosen according to our rule above, that is, $\epsilon_\nu = |\lambda_\nu| h/2$, $\nu = 1, 2, 3$.

The resulting matrix M will, in general, be nondiagonal, and we need to compute the eigensystem Q . However, if we have a strong shock with $u \geq a$, then $|u + a| \approx |u| \approx |u - a|$, and we can use $M = \varepsilon I$ with $\varepsilon \approx |u| h/2$ without computing Q . Again, the derivation is only valid for slowly moving shocks.

In addition to shocks, the Euler equations have solutions with contact discontinuities, where the density ρ is discontinuous. However, the variables u , p , and E are continuous across a contact discontinuity. The wave travels with speed $s = u$, and, because $\lambda_1 = u$ is continuous, the characteristics are parallel along the discontinuity. Thus, the behavior of the solution is similar to those of linear equations.

We have seen that discontinuous solutions of the linear model equation $u_t + u_x = 0$ are not computed very accurately, even if a high-order method is used. The reason is that the phase error for high wave numbers is large, which causes oscillatory solutions. By giving up some of the accuracy for low wave numbers, we can increase the accuracy for high wave numbers and thereby improve the situation.

Consider the approximation

$$dv_j/dt = Qv_j, \quad (8.9.19)$$

where

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

Instead of selecting the parameters $\alpha = 1$ and $\beta = 1$ such that Q is a sixth-order approximation of $\partial/\partial x$, we minimize the phase error in the least-square sense over a certain interval $[-\xi_0, \xi_0]$:

$$\min_{\alpha, \beta} \int_{-\xi_0}^{\xi_0} \left(1 - \frac{\sin \xi}{\xi} \left(1 + \alpha \frac{2}{3} \sin^2 \frac{\xi}{2} + \beta \frac{8}{15} \sin^4 \frac{\xi}{2} \right) \right)^2 d\xi$$

The following table shows α and β for three different values of ξ_0 .

Note that, for $\xi_0 = \pi/3$ and $\xi_0 = \pi/4$, Q is close to a fourth-order approximation. (In fact, the standard fourth-order approximation gives essentially the same numerical result.) For the very large wave numbers, we still need a damping term, and we substitute

$$Q \rightarrow \tilde{Q} = Q + \gamma h^2 D_+ D_-. \quad (8.9.20)$$

TABLE 8.9.1. Coefficients to Minimize over $[-\xi_0, \xi_0]$

ξ_0	α	β
$\pi/2$	0.88	1.90
$\pi/3$	0.98	1.35
$\pi/4$	0.99	1.84

Note that the extra term is of order h^2 , which is smaller than what is required for shocks. Furthermore, it affects only the amplitude and not the phase.

We have computed the solutions of the well-known shock-tube problem, also called the Riemann problem. The flow is governed by Eq. (8.9.18), and the initial data are constant on each side of the point $x_0 = 8.35$

$$(\rho, \rho u, \rho E)_{t=0} = \begin{cases} (0, 445, 0.311, 8.928), & \text{for } x < x_0, \\ (0.5, 0.0, 1.4275), & \text{for } x \geq x_0. \end{cases}$$

The solution develops a shock, a rarefaction wave, and a contact discontinuity, all of them originate at $x = x_0$.

For proper treatment of the relatively strong shock, we use the scalar viscosity coefficient $(h/2)D_- |u_j + a_j| D_+$ in all equations. However, it is activated only in the neighborhood of the shock and is cut off gradually by using the so-called van Leer limiter (see Section 8.10). In the implementation, the sign of $u + a - s$ is tested to find the location of the shock, where the shock speed s is known a priori, for this example.

The fourth-order Runge-Kutta method has been used for time discretization. In Figure 8.9.1, the variables $\rho + 3.3$, p , and $u - 1.3$ are shown for $h = 0.1$ at $t = 1.8$ (the shifting of ρ and u has been introduced for clarity of the picture). The expansion wave and the shock look fine, but the contact discontinuity in ρ is smeared out too much. The reason for this is that all these waves are located at the same point initially. Therefore, the first-order viscosity term for the shock acts also on the contact discontinuity in the beginning, and the damping is too strong to keep the sharp profile. Once it is smeared out, there is no mechanism for sharpening it, because the characteristics do not converge, as they do for the shock. The expansion wave has good accuracy in agreement with the experiment done in Figure 8.7.2.

The most straightforward and efficient procedure to overcome the difficulty with the smeared out contact discontinuity is to refine the grid. In this way the first-order viscosity term becomes smaller, and the profile becomes sharper. The refined grid is used only in the beginning. When the two discontinuities are well separated, the computation continues on the coarse grid. In our example, the change was made at $t = 0.6$, when there were five grid points between the discontinuities. The new and better result is shown in Figure 8.9.2.

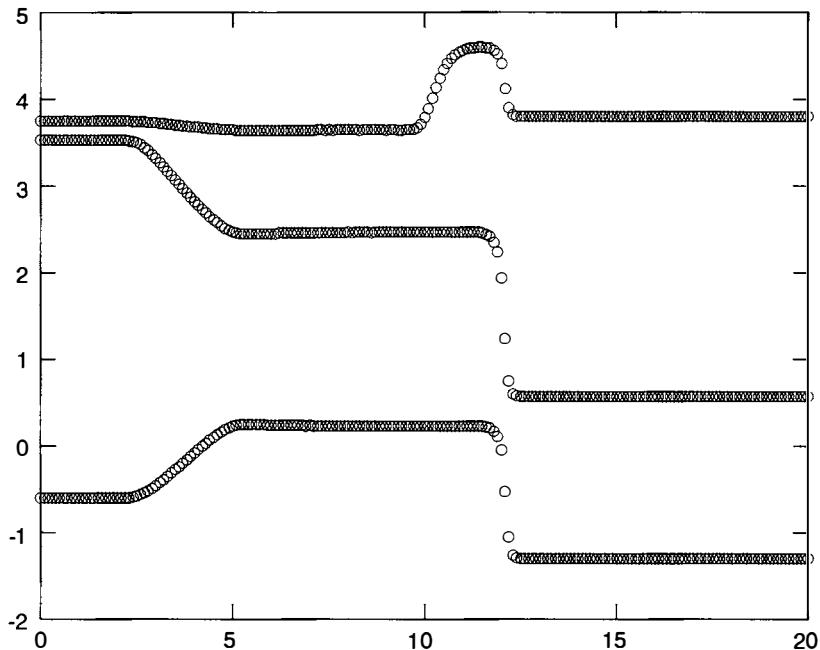


Figure 8.9.1. Shock-tube computation with one grid.

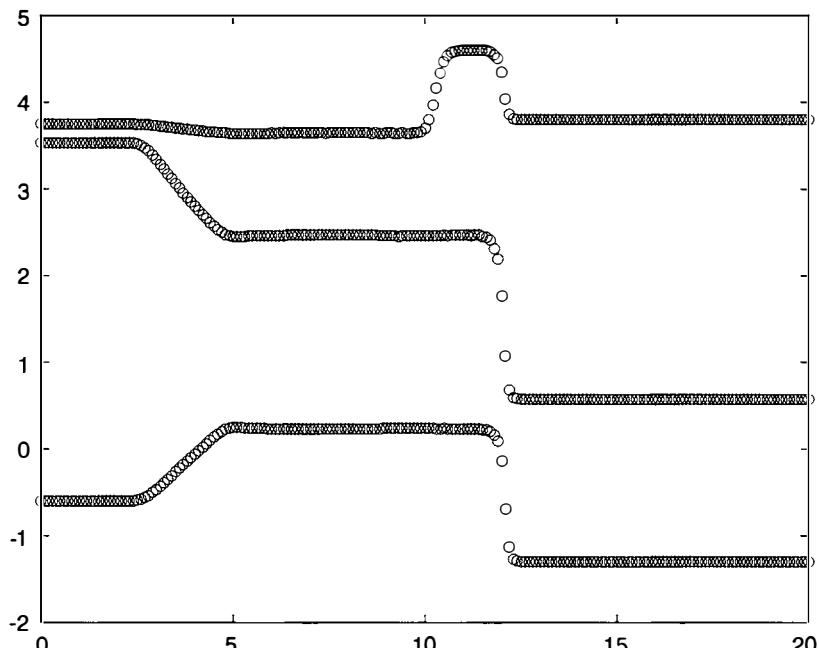


Figure 8.9.2. Shock-tube computation with a refined grid in the beginning.

In the analysis above, we have considered each side of the shock separately. The viscosity term has been designed so that the solution behaves well outside the shock layer and so that the shock speed is accurate. To obtain the correct detailed behavior of the solution inside the shock layer, further analysis is required.

8.10. HIGH-RESOLUTION METHODS

We have seen that it is necessary to use a viscosity term to control spurious oscillations and avoid too much smearing when discontinuous solutions are approximated. In this section, we discuss several methods of this type that have been used.

We consider approximations of the scalar equation

$$u_t + g(u)_x = 0, \quad -\infty < x < \infty, \quad t \geq 0 \quad (8.10.1)$$

with periodic initial data

$$u(x, 0) = f(x) = f(x + 1), \quad -\infty < x < \infty \quad (8.10.2)$$

and a periodic solution

$$u(x, t) = u(x + 1, t), \quad t \geq 0. \quad (8.10.3)$$

We first consider so-called flux-limiter methods in conservation form

$$v_j^{n+1} = v_j^n - \lambda(F(v_j^n) - F(v_{j-1}^n)) \quad (8.10.4)$$

where $\lambda = k/h$ and $F(v_j^n) = F(v_{j-\ell}^n, \dots, v_{j+r}^n)$ on a grid $x_j = jh$, $t_n = nk$. Consistency requires that $F(u, u, \dots, u) = g(u)$ (see Exercise 8.10.1).

For example, if we let $A(u) = g'(u)$ (for systems A is the Jacobian matrix), we can write the Lax–Wendroff method in conservation form

$$\begin{aligned} v_j^{n+1} = v_j^n - \frac{k}{2h} (g(v_{j+1}^n) - g(v_{j-1}^n)) + \frac{k^2}{2h^2} (A_{j+1/2}(g(v_{j+1}^n) - g(v_j^n)) \\ - A_{j-1/2}(g(v_j^n) - g(v_{j-1}^n))), \end{aligned} \quad (8.10.5)$$

where $A_{j\pm 1/2}$ is evaluated at $(v_j^n + v_{j\pm 1}^n)/2$.

Flux limiter methods are based on the idea of using a higher order flux F_2 , like the Lax–Wendroff flux, in regions when the solution is smooth and a lower order flux F_1 when the solution is not smooth. A single flux function is built of these basic flux functions. One can view the higher order flux as the lower

order flux plus a correction, that is,

$$F_2 = F_1 + (F_2 - F_1). \quad (8.10.6)$$

A flux limiter $\varphi(v_j^n)$ can be introduced to smoothly go from F_1 to F_2 and define a new flux function

$$F = F_1 + \varphi(F_2 - F_1),$$

which can be rewritten as

$$F = F_2 - (1 - \varphi)(F_2 - F_1). \quad (8.10.7)$$

If the data is smooth then φ should be near one; and φ should be near zero near a discontinuity.

To discuss several flux limiters in a simple setting and to compare them with the computational results of Section 8.1, we again return to the model equation (8.1.1) with periodic solutions. We consider combining the Lax–Wendroff flux with the lower order upwind method flux. If we assume $a > 0$ in $u_t + au_x = 0$, then we can write this method as

$$v_j^{n+1} = v_j^n - \frac{ak}{h} (v_j^n - v_{j-1}^n) - \frac{ak}{2h} \left(1 - \frac{ak}{h} \right) (v_{j+1}^n - 2v_j^n + v_{j-1}^n). \quad (8.10.8)$$

This flux function is

$$F(v_j^n) = av_j^n + \frac{a}{2} \left(1 - \frac{ak}{h} \right) (v_{j+1}^n - v_j^n). \quad (8.10.9)$$

Equation (8.10.9) is of the form of Eq. (8.10.6), with $F_1 = av_j^n$. To obtain a flux limiter, we modify Eq. (8.10.9) as

$$F(v_j^n) = av_j^n + \frac{1}{2} a \left(1 - \frac{ak}{h} \right) (v_{j+1}^n - v_j^n) \varphi(v_j^n). \quad (8.10.10)$$

The φ function needs to be a function of the smoothness of the data, so it is natural to consider $\varphi(\theta_j^n)$, where

$$\theta_j^n = \frac{v_j^n - v_{j-1}^n}{v_{j+1}^n - v_j^n} \quad (8.10.11)$$

is the ratio of neighboring gradients. If θ_j^n is near 1, then the solution is smooth and if it is very different from 1, then the gradient is changing rapidly. This measure will not be accurate near extreme points of the solution. In that case, it is possible that $\theta_j^n < 0$, even if v_j^n is smooth.

Beam and Warming have used the limiter

$$\varphi(\theta) = \theta. \quad (8.10.12)$$

Sweby suggested using

$$\varphi(\theta) = \begin{cases} 0, & \theta \leq 0, \\ \theta, & 0 \leq \theta \leq 1, \\ 1, & 1 < \theta, \end{cases} \quad (8.10.13)$$

and this yields the so-called second-order TVD scheme of Sweby. Roe has suggested his “superbee” limiter

$$\varphi(\theta) = \max(0, \min(1, 2\theta), \min(\theta, 2)). \quad (8.10.14)$$

Van Leer proposed the limiter defined by

$$\varphi(\theta) = \frac{|\theta| + \theta}{1 + |\theta|}. \quad (8.10.15)$$

We repeat the calculations shown in Figure 8.1.1 using these limiters. Recall that we used step function initial data given by Eq. (8.1.2), set $a = 1$, $h = 2\pi/240$, and $k = 2h/3$, and displayed the results at $t = 40k$ and $t = 2\pi$. In Figure 8.10.1 we repeat these calculations using the Beam–Warming, Sweby, Roe, and van Leer smoothers, respectively.

The flux-limiter methods were developed as nonlinear methods to obtain more accuracy than could be obtained with monotone schemes and still prevent oscillations. These methods were developed to have nonincreasing total variation. If we define the total variation $TV(v_j^n)$ as

$$TV(v_j^n) = \sum_j |v_j^n - v_{j-1}^n|. \quad (8.10.16)$$

Then TVD schemes are those that satisfy $TV(v_j^{n+1}) \leq TV(v_j^n)$. The methods we have described here, excepting the Beam–Warming method, were developed as TVD methods and have accuracy that is second order over most of the domain. However, it is known that TVD schemes must degenerate to first-order accuracy at extreme points.

In our discussion above, we assumed $a > 0$. It is clear that a similar method can be developed for the case $a < 0$. However, these two cases can be written using a single formula that is valid for any wave speed.

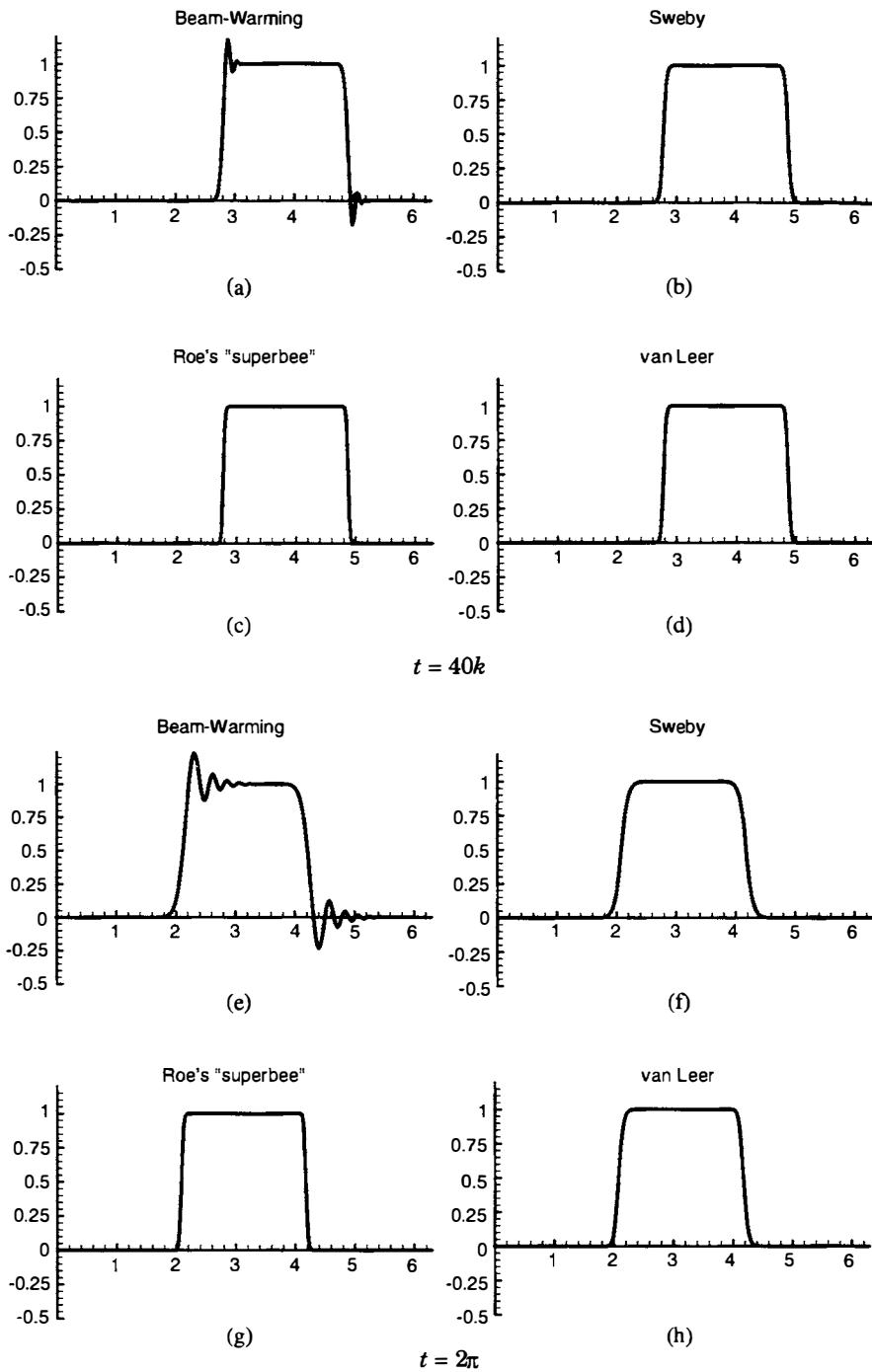


Figure 8.10.1. (a)-(h).

The upwind flux function can be written as

$$F_1(v_j^n) = \frac{a}{2} (v_j^n + v_{j+1}^n) - \frac{|a|}{2} (v_{j+1}^n - v_j^n) \quad (8.10.17)$$

and the Lax–Wendroff flux as

$$F_2(v_j^n) = \frac{a}{2} (v_j^n + v_{j+1}^n) - \frac{a^2 k}{2h} (v_{j+1}^n - v_j^n). \quad (8.10.18)$$

We can then introduce a limiter and write

$$F(v_j^n) = F_1(v_j^n) + \frac{\varphi(v_j^n)}{2} \left(\operatorname{sign}\left(\frac{ak}{h}\right) - \frac{ak}{h} \right) a(v_{j+1}^n - v_j^n) \quad (8.10.19)$$

since $|a| = \operatorname{sign}(a)a = \operatorname{sign}(ak/h)a$. We can now define φ as before, but need to define θ to be the ratio of slopes in the upwind direction. If we write $j_{\pm} = j - \operatorname{sign}(ak/h)$, then

$$\theta_j^n = \frac{v_{j\pm+1}^n - v_{j\pm}^n}{v_{j+1}^n - v_j^n}. \quad (8.10.20)$$

We now discuss slope-limiter methods. These methods are based on cell averages. If we use a grid of points $x_j = jh$, $h > 0$, then we can associate the cell average

$$V_j(t) = h^{-1} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x, t) dx \quad (8.10.21)$$

with the grid point x_j . The first method of this type was that of Godunov, which can be described as follows:

- Given data V_j^n at time t_n , construct a function $v^n(x)$ defined for all x . Godunov's method uses

$$v^n(x) = \begin{cases} V_j^n, & x_j \leq x < x_j + h/2, \\ V_{j+1}^n, & x_j + h/2 \leq x < x_{j+1}, \end{cases}$$

on $x_j \leq x \leq x_{j+1}$, and so on.

2. Solve the conservation law exactly on this subinterval to obtain $v^{n+1}(x)$ at t_{n+1} .
3. Define V_j^{n+1} using Eq. 8.10.21 with $t = t_{n+1}$.

This method has been generalized by using more accurate reconstructions in step 1. above. For instance, we could consider a piecewise linear reconstruction

$$v^n(x) = V_j^n + s_j^n(x - x_j), \quad x_{j-1/2} \leq x < x_{j+1/2}, \quad (8.10.22)$$

with slope s_j based on the data V_j^n . If one takes the obvious choice

$$s_j^n = \frac{V_{j+1}^n - V_j^n}{h},$$

one recovers the Lax–Wendroff method for Eq. (8.1.1). This shows that these methods can be second-order accurate, but may still exhibit oscillatory behavior. Several slope limiter methods have been constructed that yield TVD schemes. One simple such choice is the min mod limiter with

$$s_j^n = h^{-1} \min \text{mod}(V_{j+1}^n - V_j^n, V_j^n - V_{j-1}^n),$$

where

$$\min \text{mod}(a, b) = \frac{1}{2}(\text{sign}(a) + \text{sign}(b)) \min(|a|, |b|).$$

We now look at the ENO methods, which are the so-called essentially nonoscillatory methods. To achieve higher order accuracy, the TVD criteria is replaced by the condition that there exist a constant α such that

$$TV(v_j^{n+1}) \leq (1 + \alpha k)TV(v_j^n), \quad (8.10.23)$$

which guarantees that

$$TV(v_j^n) \leq (1 + \alpha k)^n TV(v_j^0) \leq e^{\alpha t_n} TV(v_j^0), \quad (8.10.24)$$

so that these methods will be total variation stable.

We describe a second-order ENO method for simplicity and then indicate how the ideas can be generalized. We again consider Eq. (8.10.1)

$$u_t + g(u)_x = 0, \quad t \geq 0$$

with

$$u(x, 0) = f(x).$$

If we integrate this equation over $[x_{j-1/2}, x_{j+1/2}] \times [t_n, t_{n+1}]$ we get

$$u_j^{n+1} = u_j^n - \frac{k}{h} (\bar{g}_{j+1/2} - \bar{g}_{j-1/2}), \quad (8.10.25)$$

where

$$\bar{g}_{j+1/2} = k^{-1} \int_0^k g(u(x_{j+1/2}, t_n + \tau)) d\tau \quad (8.10.26)$$

and

$$u_j^n = h^{-1} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_n) dt.$$

The function u in the integral in Eq. (8.10.26) must be reconstructed from the cell averages u_j^n to define an algorithm. The ENO method uses piecewise polynomials to reconstruct $u(x, t_n)$ and then Taylor expansion in time over $[t_n, t_{n+1}]$.

Assume that V_j^n approximates u_j^n . We then compute V_j^{n+1} from

$$V_j^{n+1} = V_j^n - \frac{k}{h} (\bar{g}_{j+1/2} - \bar{g}_{j-1/2}). \quad (8.10.27)$$

We define

$$\bar{g}_{j+1/2} = g(v_{j+1/2}^L, v_{j+1/2}^R), \quad (8.10.28)$$

where

$$\begin{aligned} v_{j+1/2}^L &= v_j^n + \frac{h}{2} \left(1 - \frac{k}{h} a_j^n \right) s_j^n, \\ v_{j+1/2}^R &= v_{j+1}^n - \frac{h}{2} \left(1 + \frac{k}{h} a_{j+1}^n \right) s_{j+1}^n, \end{aligned} \quad (8.10.29)$$

with $a_j^n = g'(v_j^n)$, and s_j^n is a piecewise smooth function of x around x_j . The

problem of solving the initial-value problem for a nonlinear conservation law with piecewise constant data is referred to as the Riemann problem.

In Figure 8.10.2, we show the result using a second-order ENO method obtained for the same linear coefficient problem used for Figure 8.10.1. For this constant scalar case with $g(u)_x = u_x$ and periodic boundary conditions, we use

$$\bar{g}_{j+1/2} = v_{j+1/2}^L = v^n + h \left(1 - \frac{k}{h} \right) s_j^n / 2$$

and

$$s_j^n = h^{-1} \min \text{mod}(2(V_{j+1}^n - V_j^n), \frac{1}{2}(V_{j+1}^n - V_{j-1}^n), 2(V_j^n - V_{j-1}^n)).$$

The min mod function is the same as defined previously extended to three variables.

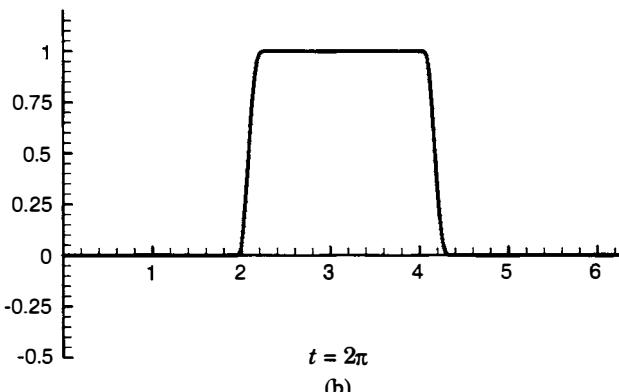
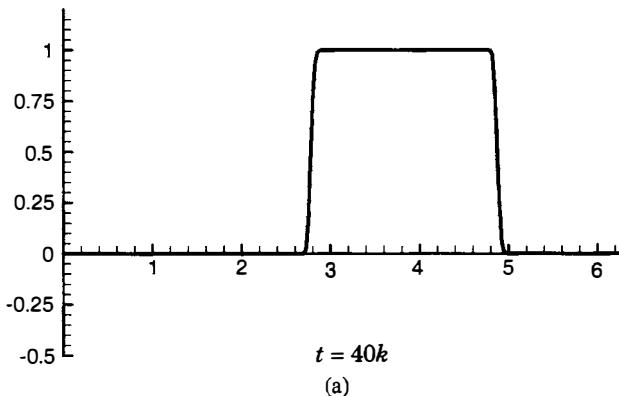


Figure 8.10.2.

The ENO results are not as accurate as the results of Roe's "superbee." The ENO methods smear out contact discontinuities; that is, a discontinuity that traverses a region where the characteristics are parallel. In this situation, there is no propagation of information into the discontinuity to sharpen it. To further compare the ENO scheme with the other four schemes, we applied each method to a new set of initial conditions. The equation we are approximating is the scalar equation defined in Eq. (8.10.1) with $g(u)_x = u_x$. The initial conditions are those taken from a paper by Harten (1989),

$$u_0(x + 0.5) = \begin{cases} -x \sin(\frac{3}{2}\pi x^2), & -1 < x < -\frac{1}{3}, \\ |\sin(2\pi x)|, & |x| < \frac{1}{3}, \\ 2x - 1 - \sin(3\pi x)/6, & \frac{1}{3} < x < 1. \end{cases} \quad (8.10.30)$$

There is a shift in the values of the right-hand side for display purposes. Figure 8.10.3 shows the exact solution as a solid line and the approximated solutions

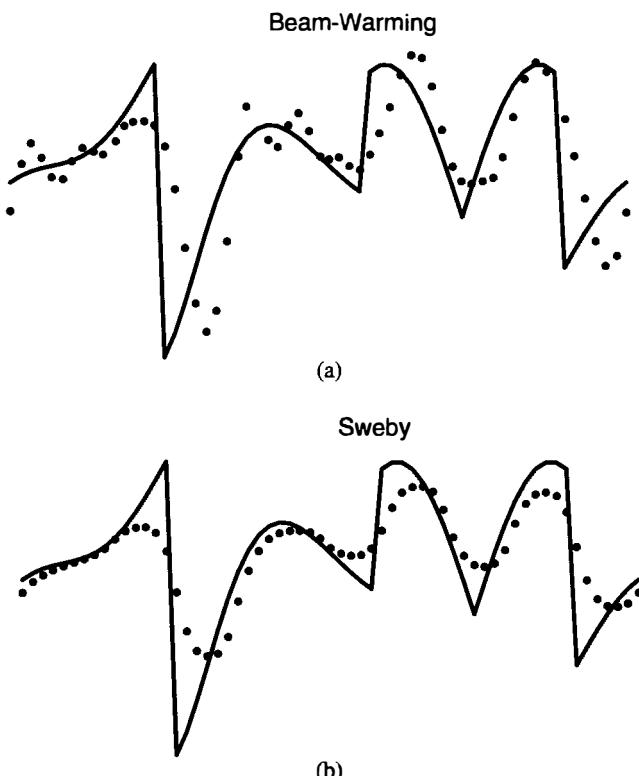


Figure 8.10.3. A comparison of methods.

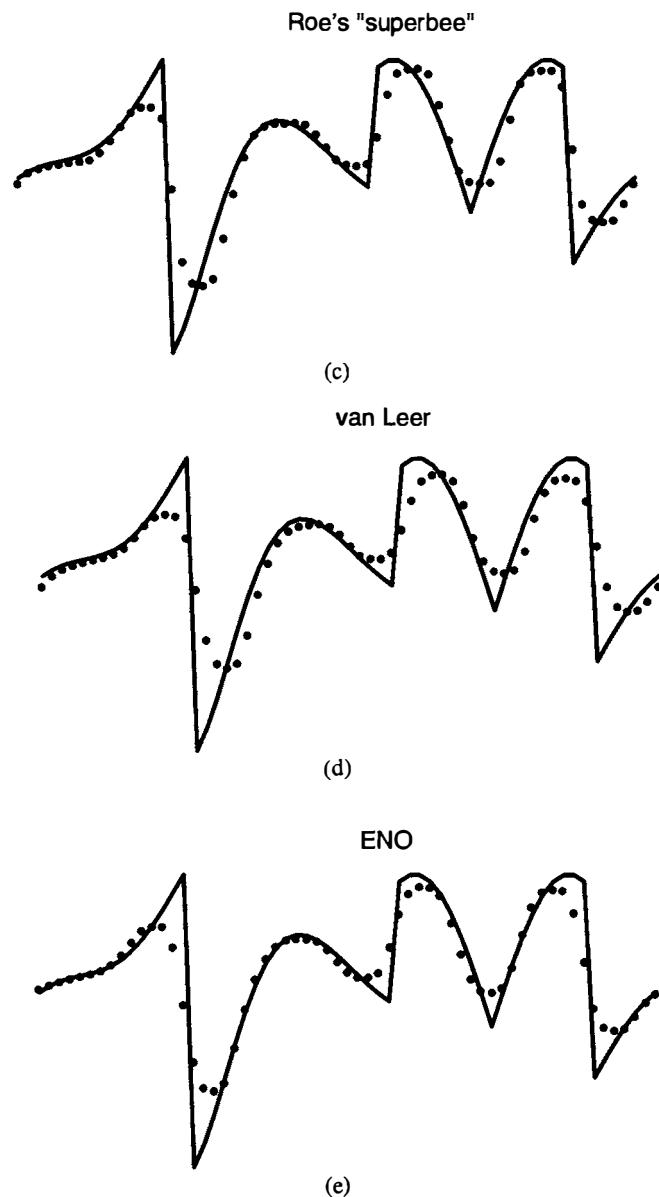


Figure 8.10.3. (Continued)

as dots at $t = 2$. These figures show that the second-order ENO scheme is more accurate for these complex initial conditions than the other second-order schemes described in this section.

EXERCISES

8.10.1. Prove that consistency requires $F(u, u, \dots, u) = g(u)$ in Eq. (8.10.4).

BIBLIOGRAPHIC NOTES

Much work has been done on the theory and development of numerical methods for nonlinear hyperbolic conservation laws and shock solutions during the last decade. The material in this book does not cover all that. Recent books devoted to this topic have been written by Godlewski and Raviart (1990) and LeVeque (1990). There is also the review paper by Osher and Tadmor (1988). Applications in fluid dynamics are detailed in the book by Hirsch (1990).

The results discussed on the behavior of the oscillations at a discontinuity for centered difference methods are from Hedstrom (1975) and Chin and Hedstrom (1978). Other results and references on the spreading error can be found in Brenner, Thomee, and Wahlbin (1975).

Linear monotone schemes are shown to be at most first order accurate in Harten, Hyman, and Lax (1976). The proof of Theorem 8.8.1 can be found in Kreiss and Lorenz (1989). The recovery of piecewise smooth solutions from oscillatory approximations has been discussed by Mock and Lax (1978) and by Gottlieb and Tadmor (1984). Smoothing of the initial data to increase the rate of convergence has been treated in Majda, McDonough, and Osher (1978). Our discussion of the required size of the dissipative coefficient is based on the work of Kreiss and Johansson (1993). Gunilla Johansson also did the shock-tube computation presented in Section 8.9.

Hybrid methods that combine difference methods with the method of characteristics have been considered (see e.g., Henshaw, 1985). Difference methods in conservation form were considered by Lax and Wendroff (1960).

Osher and Chakravarthy (1984) have shown that TVD methods are first order at extreme points. Tadmor (1984) has shown that all three point TVD schemes are at most first order accurate and Goodman and LeVeque (1985) have shown that, except for trivial cases, any TVD scheme in two space dimensions is at most first order accurate. Engquist, Lotstedt, and Sjögren (1989) have derived a different type of TVD method by adding a nonlinear filter to standard centered methods.

The flux limiters discussed were introduced by Sweby (1984), Roe (1985), and van Leer (1974). The ENO schemes were developed by Harten et. al. (1986, 1987), Harten (1989) and Shu and Osher (1988). The second-order example is based on Harten (1989). In Harten (1989), he developed ENO methods with “subcell resolution” to reduce the smearing of contact discontinuities. Tadmor (1984) has shown that the Lax-Friedrichs, Engquist and Osher (1980), Godunov and Roe methods all have numerical viscosity in decreasing order.

II

INITIAL-BOUNDARY-VALUE PROBLEMS

9

THE ENERGY METHOD FOR INITIAL-BOUNDARY-VALUE PROBLEMS

9.1. CHARACTERISTICS AND BOUNDARY CONDITIONS FOR HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION

We start with the scalar hyperbolic equation

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x}, \quad a = \text{constant}, \quad (9.1.1)$$

in the domain $0 \leq x \leq 1, t \geq 0$ (see Figure 9.1.1). At $t = 0$, we give initial data

$$u(x, 0) = f(x). \quad (9.1.2)$$

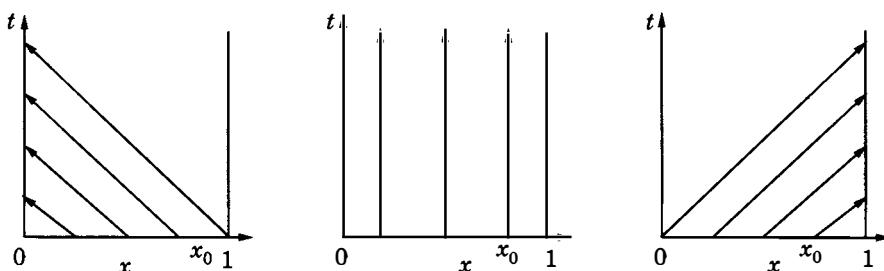


Figure 9.1.1. Characteristics of Eq. (9.1.1) for $a > 0$, $a = 0$, $a < 0$.

As we have seen in Chapter 8, the solutions are constant along the characteristic lines $x + at = \text{constant}$.

If $a > 0$, then the solution of our problem is uniquely determined for

$$x \geq 0, \quad t \geq 0, \quad x + at \leq 1.$$

To extend the solution for $x + at > 1$, we specify a boundary condition

$$u(1, t) = g_1(t), \quad x = 1. \quad (9.1.3)$$

For the solution to be smooth in the whole domain, it is necessary that $g_1(t)$ and $f(x)$ are smooth functions. It is also necessary that $g_1(t)$ and $f(x)$ be compatible or satisfy compatibility conditions. The most obvious necessary condition is that

$$g_1(0) = f(1). \quad (9.1.4)$$

Otherwise, the solution has a jump. (In that case, we only obtain a generalized solution.) If Eq. (9.1.4) is satisfied and $g_1 \in C^1(t)$ and $f(x) \in C^1(x)$, then $u(x, t)$ is Lipschitz continuous. To obtain solutions belonging to $C^1(x, t)$, we first note that $v = u_x$ satisfies

$$\begin{aligned} v_t &= av_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ v(x, 0) &= f'(x), \\ v(1, t) &= a^{-1}u_t(1, t) = a^{-1}g'_1(t). \end{aligned}$$

To ensure that v be continuous everywhere, the initial and boundary values must match each other at $(x = 1, t = 0)$. This leads to the condition

$$af'(1) = g'_1(0). \quad (9.1.5)$$

Also, $w = u_t$ satisfies

$$\begin{aligned} w_t &= aw_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ w(x, 0) &= au_x(x, 0) = af'(x), \\ w(1, t) &= g'_1(t), \end{aligned}$$

and the condition (9.1.5) also ensures that w is continuous everywhere. Thus, u is $C^1(x, t)$.

Higher order derivatives satisfy the same differential equation (9.1.1), and we get higher order regularity by adding more restrictions on higher order derivatives of f and g at $(x = 1, t = 0)$. The same technique can be applied to any

problem to ensure that we get smooth solutions. For systems of nonlinear equations, these compatibility conditions may become complicated. The easiest way to satisfy all of them is to require that all initial and boundary data (and forcing functions) vanish near the boundaries at $t = 0$.

We now consider the other cases for α . If $\alpha = 0$, we do not need any boundary conditions because $\partial u / \partial t \equiv 0$ implies

$$u(x, t) \equiv f(x), \quad \alpha = 0. \quad (9.1.6)$$

If $\alpha < 0$, then the solution is uniquely determined if we give the boundary condition

$$u(0, t) = g_0(t), \quad \alpha < 0. \quad (9.1.7)$$

Now we consider a strongly hyperbolic system with constant coefficients

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (9.1.8)$$

where u has m components. Let S be composed of the eigenvectors of A such that

$$S^{-1}AS = \Lambda = \begin{bmatrix} \Lambda^I & 0 & 0 \\ 0 & \Lambda^{II} & 0 \\ 0 & 0 & \Lambda^{III} \end{bmatrix}, \quad (9.1.9)$$

where

$$\Lambda^I = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda_r \end{bmatrix} > 0,$$

$$\Lambda^{II} = \begin{bmatrix} \lambda_{r+1} & 0 & \cdots & 0 \\ 0 & \lambda_{r+2} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda_{m-s} \end{bmatrix} < 0, \quad \Lambda^{III} \equiv 0$$

are diagonal matrices.

We introduce a new variable $v = S^{-1}u$. Then, we obtain the system

$$\frac{\partial v}{\partial t} = \Lambda \frac{\partial v}{\partial x}, \quad (9.1.10)$$

or

$$\frac{\partial}{\partial t} v^I = \Lambda^I \frac{\partial}{\partial x} v^I, \quad \frac{\partial}{\partial t} v^{II} = \Lambda^{II} \frac{\partial}{\partial x} v^{II}, \quad \frac{\partial}{\partial t} v^{III} = 0.$$

Using the previous argument, we obtain a unique solution if we specify the initial condition

$$v(x, 0) = f(x), \quad 0 \leq x \leq 1,$$

and the boundary conditions

$$v^I(1, t) = g^I(t), \quad v^{II}(0, t) = g^{II}(t).$$

With these conditions the problem decomposes into m scalar problems. We can couple the components by generalizing the boundary conditions to

$$\begin{aligned} v^I(1, t) &= R_1^{II} v^{II}(1, t) + R_1^{III} v^{III}(1, t) + g^I(t), \\ v^{II}(0, t) &= R_0^I v^I(0, t) + R_0^{III} v^{III}(0, t) + g^{II}(t). \end{aligned} \quad (9.1.11)$$

Here R_1^{II} , R_1^{III} , R_0^I , and R_0^{III} are rectangular matrices that may depend on t .

It is easy to describe these conditions in geometrical terms. $\Lambda^{III} = 0$ implies that $v^{III}(x, t) = f^{III}(x)$. Thus, we need only discuss the influence of the boundary conditions on v^I and v^{II} . We write Eq. (9.1.11) as

$$\begin{aligned} v^I(1, t) &= R^{II} v^{II}(1, t) + \tilde{g}^I(t), & v^{II}(0, t) &= R^I v^I(0, t) + \tilde{g}^{II}(t), \\ \tilde{g}^I &:= g^I + R_1^{III} f^{III}(1), & \tilde{g}^{II} &:= g^{II} + R_0^{III} f^{III}(0), \end{aligned} \quad (9.1.12)$$

where $R^I := R_0^I$ and $R^{II} := R_1^{II}$. Starting with $t = 0$ the initial values for v^I and v^{II} are transported along the characteristics to the boundaries $x = 0$ and $x = 1$, respectively. Using the boundary conditions, these values are transformed into values for $v^{II}(0, t)$ and $v^I(1, t)$, which are then transported along the characteristics to the boundaries $x = 1$ and $x = 0$, respectively. Here the process is repeated (see Figure 9.1.2). Because of these geometrical properties the components of v are called *characteristic variables*.

The number of boundary conditions for $x = 0$ is equal to the number of negative eigenvalues of Λ , or, equivalently, the number of characteristics entering the region. Correspondingly, at $x = 1$, the number of boundary conditions is equal to the number of positive eigenvalues of Λ . No boundary conditions are required, or may be given, for vanishing eigenvalues.

In most applications, the differential equations are given in the nondiagonal

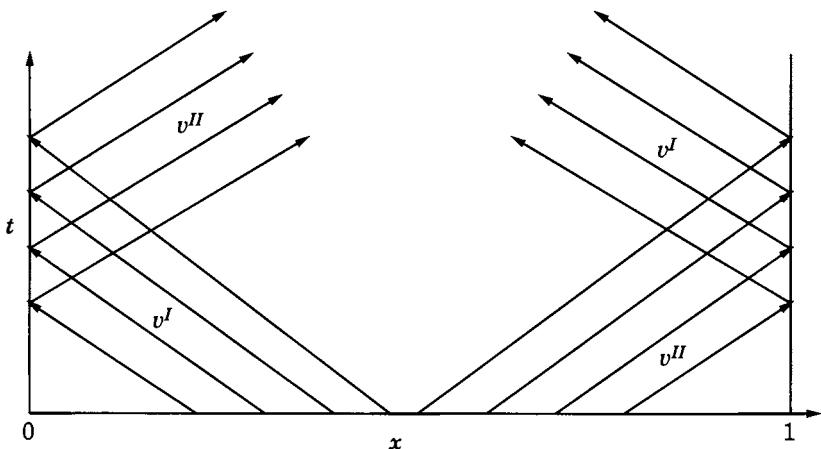


Figure 9.1.2.

form (9.1.8), and the boundary conditions are linear relations

$$L_0 u(0, t) = g_0(t), \quad L_1 u(1, t) = g_1(t). \quad (9.1.13)$$

Here

$$L_0 = \begin{bmatrix} l_{r+1,1} & \cdots & l_{r+1,m} \\ \vdots & \vdots & \vdots \\ l_{m-s,1} & \cdots & l_{m-s,m} \end{bmatrix}, \quad L_1 = \begin{bmatrix} l_{1,1} & \cdots & l_{1,m} \\ \vdots & \vdots & \vdots \\ l_{r,1} & \cdots & l_{r,m} \end{bmatrix}$$

are rectangular matrices whose rank is equal to the number of negative and positive eigenvalues of A , respectively (or better, the number of characteristics that enter the region at the boundary).

If we use the transformation (9.1.9), the differential equations are transformed into Eq. (9.1.10), and the boundary conditions become

$$L_0 S v(0, t) = g_0(t), \quad L_1 S v(1, t) = g_1(t). \quad (9.1.14)$$

That is, we again obtain linear relations for the characteristic variables. Our initial-boundary-value problem can be solved if Eq. (9.1.14) can be written in the form (9.1.11); then, we can solve the relations (9.1.14) for $v''(0, t)$ and $v'(1, t)$, respectively. We have now proved the following theorem.

Theorem 9.1.1. *Consider the system (9.1.8) for $0 \leq x \leq 1$, $t \geq 0$ with initial data at $t = 0$ and boundary conditions (9.1.13). This problem has a solution if the system is strongly hyperbolic, the number of boundary conditions is equal*

to the number of characteristics entering the region at the boundary, and we can write the boundary conditions so that the characteristic variables connected with the ingoing characteristics can be expressed in terms of the other variables.

As an example, we consider the normalized wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (9.1.15)$$

with the initial conditions

$$u(x, 0) = f_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = f_1(x), \quad (9.1.16)$$

and boundary conditions

$$u(0, t) = g_0(t), \quad u(1, t) = g_1(t). \quad (9.1.17)$$

Introducing a new function v by

$$v(x, t) = \int_0^t \frac{\partial u}{\partial x}(x, \tau) d\tau + \int_0^x f_1(\xi) d\xi,$$

or, in other words,

$$\frac{\partial v}{\partial t} = \frac{\partial u}{\partial x},$$

we can write Eq. (9.1.15) as a first-order system

$$\frac{\partial}{\partial t} \begin{bmatrix} u \\ v \end{bmatrix} = A \frac{\partial}{\partial x} \begin{bmatrix} u \\ v \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad (9.1.18a)$$

with initial conditions

$$u(x, 0) = f_0(x), \quad v(x, 0) = \int_0^x f_1(\xi) d\xi \quad (9.1.18b)$$

and boundary conditions (9.1.17). The matrix A has one positive and one negative eigenvalue, and, therefore, there is exactly one characteristic that enters the region on each side. Thus, the number of boundary conditions is correct. We

now transform A to diagonal form. The eigenvalues λ_j and the corresponding eigenvectors ϕ_j are

$$\lambda_1 = 1, \quad \phi_1 = \frac{1}{\sqrt{2}} (1, 1)^T, \quad \lambda_2 = -1, \quad \phi_2 = \frac{1}{\sqrt{2}} (1, -1)^T.$$

Thus,

$$S = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = S^{-1}.$$

Introducing

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = S^{-1} \begin{bmatrix} u \\ v \end{bmatrix}$$

as new variables gives us

$$\frac{\partial}{\partial t} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial}{\partial x} \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}$$

with boundary conditions

$$\begin{aligned} \tilde{u}(0, t) + \tilde{v}(0, t) &= \sqrt{2} u(0, t) = \sqrt{2} g_0(t), \\ \tilde{u}(1, t) - \tilde{v}(1, t) &= \sqrt{2} u(1, t) = \sqrt{2} g_1(t). \end{aligned}$$

Therefore, the conditions of Theorem 9.1.1 are satisfied, and we can solve the initial-boundary-value problem.

We now consider equations with variable coefficients. We start with the scalar equation

$$\frac{\partial u}{\partial t} = \lambda(x, t) \frac{\partial u}{\partial x}, \quad 0 \leq x \leq 1, \quad t \geq 0 \quad (9.1.19)$$

with initial values

$$u(x, 0) = f(x).$$

From Section 8.2, we know its solution is constant along the characteristic lines

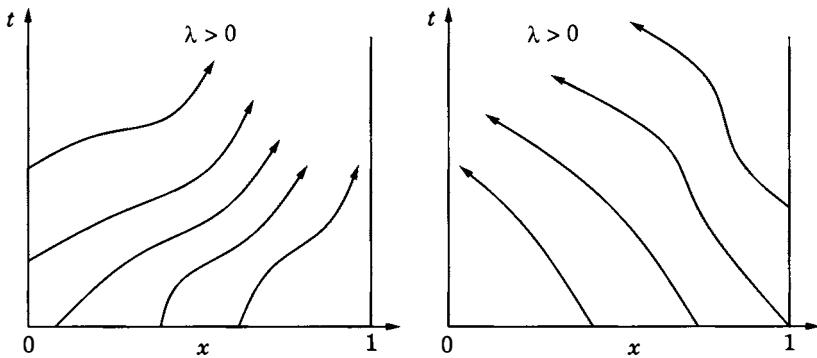


Figure 9.1.3. Characteristics of (9.1.19).

$$\frac{dx}{dt} = -\lambda(x, t), \quad x(0) = x_0.$$

If $\lambda(x, t) < 0$, then we have to give boundary conditions on the boundary $x = 0$,

$$u(0, t) = g_0(t)$$

, and, if $\lambda(x, t) > 0$, we give

$$u(1, t) = g_1(t)$$

(see Figure 9.1.3). Thus, we have the same situation as we had for equations with constant coefficients. However, $\lambda(x, t)$ can change sign in the interior of the region.

If $\lambda(0, t) > 0, \lambda(1, t) < 0$, no boundary conditions need to be specified anywhere, and if $\lambda(0, t) < 0, \lambda(1, t) > 0$, then we have to specify boundary conditions on both sides. As before, if $\lambda(0, t) \equiv 0$, no boundary conditions need to be given at $x = 0$ (see Figure 9.1.4). As in Section 8.2, we can also solve the

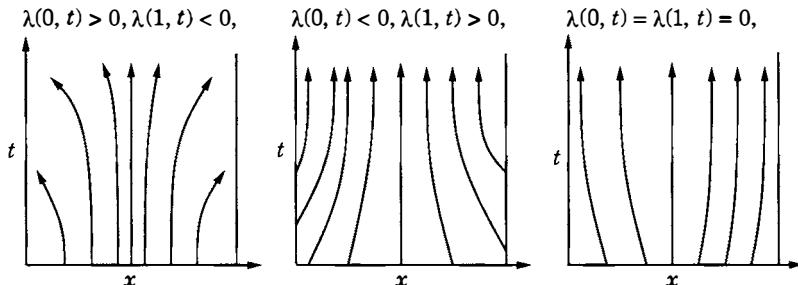


Figure 9.1.4. Characteristics of Eq. (9.1.19).

inhomogeneous equation

$$u_t = \lambda(x, t)u_x + F(x, t),$$

by the method of characteristics.

We now consider systems

$$\begin{aligned} \frac{\partial u}{\partial t} &= A(x, t) \frac{\partial u}{\partial x} + B(x, t)u + F(x, t), & 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), & 0 \leq x \leq 1. \end{aligned} \quad (9.1.20)$$

We assume that this is a strongly hyperbolic system, that is, that the eigenvalues of A are real and that A can be smoothly transformed into diagonal form. Therefore, we can, without restriction, assume that $A = \Lambda$ is diagonal. We solve the system by iteration,

$$\frac{\partial}{\partial t} u^{[n+1]} = \Lambda \frac{\partial}{\partial x} u^{[n+1]} + Bu^{[n]} + F;$$

that is, we have to solve m scalar equations at every step. It is clear that one should specify boundary conditions of the type of Eq. (9.1.11). We now have the following theorem.

Theorem 9.1.2. Consider the system (9.1.20), where A is diagonal. Also assume that the eigenvalues λ_j do not change sign at the boundaries; that is, one of the following relations hold at the boundaries for each j and, for all t : $\lambda_j > 0$, $\lambda_j \equiv 0$, $\lambda_j < 0$. If the boundary conditions are of the form in Eq. (9.1.11), then the initial-boundary-value problem has a unique solution.

EXERCISES

9.1.1. Determine all boundary conditions of the type of Eq. (9.1.13) such that the initial-boundary-value problem for the differential equation

$$u_t = \begin{bmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{bmatrix} u_x, \quad 0 \leq x \leq 1, \quad t \geq 0,$$

has a unique solution. Here a and b are real constants.

9.2. ENERGY ESTIMATES FOR HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION

In this section, we consider systems

$$\frac{\partial u}{\partial t} = \Lambda(x, t) \frac{\partial u}{\partial x} + B(x, t)u, \quad 0 \leq x \leq 1, \quad t \geq t_0, \quad (9.2.1a)$$

$$u(x, t_0) = f(x), \quad (9.2.1b)$$

where

$$\begin{aligned} \Lambda &= \begin{bmatrix} \Lambda' & 0 \\ 0 & \Lambda'' \end{bmatrix}, \\ \Lambda' &= \begin{bmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & \cdots & 0 & \lambda_r \end{bmatrix} > 0, \\ \Lambda'' &= \begin{bmatrix} \lambda_{r+1} & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_{r+2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & \cdots & 0 & \lambda_m \end{bmatrix} < 0, \end{aligned}$$

is a real diagonal matrix. For simplicity only, we assume that Λ is nonsingular. As we have seen in the last section, the problem above has a unique solution if we specify the boundary conditions

$$u''(0, t) = R^I(t)u'(0, t), \quad u'(1, t) = R''(t)u''(1, t). \quad (9.2.2)$$

The aim of this section is to derive energy estimates. We assume that the data are compatible, that is, that at $t = 0$ the initial data satisfy the boundary conditions.

We use the notation

$$G(x)|_0^1 = G(1) - G(0)$$

and prove the following lemma.

Lemma 9.2.1. *Let u and v be smooth vector functions and A a smooth matrix function. Then*

$$(u, Av_x) = -(u_x, Av) - (u, A_x v) + \langle u, Av \rangle|_0^1,$$

and

$$|(u, Av)| \leq \|A\|_\infty \|u\| \|v\|,$$

where $\|A\|_\infty = \sup_x |A|$.

Proof. The first relation follows by integration by parts. The second follows directly from the definition of the scalar product [cf. Eq. (1.2.18)].

Next, we want to prove the following theorem.

Theorem 9.2.1. *Let $u(x, t)$ be a smooth solution of the initial-boundary-value problem shown in Eqs. (9.2.1) and (9.2.2). There are constants K and α that do not depend on f such that*

$$\|u(\cdot, t)\| \leq Ke^{\alpha(t-t_0)}\|u(\cdot, t_0)\| = Ke^{\alpha(t-t_0)}\|f\|. \quad (9.2.3)$$

Proof. We have

$$\begin{aligned} \|u\|_t^2 &= (u_t, u) + (u, u_t), \\ &= (Bu, u) + (u, Bu) + (\Lambda u_x, u) + (u, \Lambda u_x), \\ &= (Bu, u) + (u, Bu) - (u, \Lambda_x u) + \langle u, \Lambda u \rangle |_0^1, \\ &\leq 2\alpha\|u\|^2 + \langle u, \Lambda u \rangle |_0^1, \end{aligned}$$

where

$$2\alpha = \max_{x,t} \frac{(u, (B + B^* - \Lambda_x)u)}{\|u\|^2}.$$

Using the boundary conditions, we obtain

$$\begin{aligned} &\langle u(1, t), \Lambda(1, t)u(1, t) \rangle \\ &= \langle u'(1, t), \Lambda'(1, t)u'(1, t) \rangle + \langle u''(1, t), \Lambda''(1, t)u''(1, t) \rangle, \\ &= \langle u''(1, t), C(1, t)u''(1, t) \rangle, \end{aligned}$$

where

$$C(1, t) = \Lambda''(1, t) + R^{II^*}(t)\Lambda'(1, t)R^{II}(t).$$

Now assume that $|R^{II}(t)|$ is small enough to guarantee

$$C(1, t) < \frac{1}{2} \Lambda^{II}(1, t).$$

Then

$$\langle u(1, t), \Lambda(1, t)u(1, t) \rangle \leq -\frac{1}{2}\lambda_0 |u^{II}(1, t)|^2, \quad \lambda_0 = \min_{1 \leq j \leq m} |\lambda_j|.$$

Correspondingly,

$$-\langle u(0, t), \Lambda(0, t)u(0, t) \rangle \leq -\frac{1}{2}\lambda_0 |u^I(0, t)|^2,$$

if $|R^I(t)|$ is also sufficiently small. In this case we have

$$\frac{d}{dt} \|u\|^2 + \frac{1}{2} \lambda_0 (|u^{II}(1, t)|^2 + |u^I(0, t)|^2) \leq 2\alpha \|u\|^2,$$

and the desired estimate (9.2.3) follows. Because $u^I(1, t)$ and $u^{II}(0, t)$ are linear combinations of $u^{II}(1, t)$ and $u^I(0, t)$ respectively, we get the sharper estimate

$$\|u(\cdot, t)\|^2 + \int_{t_0}^t (|u(0, \tau)|^2 + |u(1, \tau)|^2) d\tau \leq \text{constant } e^{2\alpha(t-t_0)} \|f(\cdot)\|^2.$$

If R^I, R^{II} are not sufficiently small, then we can proceed in the following way. Introduce a new variable w into Eq. (9.2.1),

$$u^I = w^I, \quad u^{II} = dw^{II},$$

where d is a function of x . Then w is the solution of

$$\begin{aligned} \frac{\partial w}{\partial t} &= \Lambda \frac{\partial w}{\partial x} + \tilde{B}w, \\ w^{II}(0, t) &= d^{-1}(0)R^I(t)w^I(0, t), \quad w^I(1, t) = d(1)R^{II}(t)w^{II}(1, t). \end{aligned}$$

Here \tilde{B} depends on B and d . Now choose d as a smooth function such that $|d^{-1}(0)R^I|$ and $|d(1)R^{II}|$ are small enough that the previous estimates are valid for w . Then we also obtain estimates for u . This proves the theorem.

As before for periodic problems, we define a solution operator as the mapping

$$u(x, t) = S(t, t_0)u(x, t_0),$$

where $u(x, t)$ is the solution of Eqs. (9.2.1a) and (9.2.2) with initial data $u(x, t_0)$ at $t = t_0$. By Theorem 9.2.1,

$$\|S(t, t_0)\| \leq Ke^{\alpha(t-t_0)}.$$

By Duhamel's principle, the function

$$u(x, t) = S(t, t_0)f(x) + \int_{t_0}^t S(t, \tau)F(x, \tau) d\tau$$

is the solution of the inhomogeneous problem

$$\begin{aligned} u_t &= \Lambda u_x + Bu + F, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, t_0) &= f(x), \\ u^{II}(0, t) &= R^I(t)u^I(0, t), \quad u^I(1, t) = R^{II}(t)u^{II}(1, t). \end{aligned} \quad (9.2.4)$$

The bound on the solution operator allows us to estimate the solution of problem (9.2.4), precisely as in Eq. (4.7.17) for the initial-value problem.

The boundary conditions are also often inhomogeneous:

$$u^{II}(0, t) = R^I(t)u^I(0, t) + g^{II}(t), \quad u^I(1, t) = R^{II}(t)u^{II}(1, t) + g^I(t).$$

In this case, we introduce two new variables

$$\tilde{u}^I(x, t) = u^I(x, t) - g^I(t)x, \quad \tilde{u}^{II}(x, t) = u^{II}(x, t) - g^{II}(t)(1 - x). \quad (9.2.5)$$

The new problem has homogeneous boundary conditions, but the initial function f and the forcing function F are modified.

To derive energy estimates, it is not necessary that the system be in diagonal form. Let us consider systems

$$u_t = B(x, t)u_x + C(x, t)u, \quad 0 \leq x \leq 1, \quad t \geq t_0, \quad (9.2.6)$$

with boundary conditions

$$L_0u(0, t) = 0, \quad L_1u(1, t) = 0. \quad (9.2.7)$$

Here $B = B^*$ is a Hermitian matrix with exactly $m - r$ negative eigenvalues at $x = 0$ and exactly r positive eigenvalues at $x = 1$. L_0 and L_1 are $(m - r) \times m$ and $r \times m$ matrices of maximal rank, respectively. Now, we want to prove the following theorem.

Theorem 9.2.2. *If*

$$(-1)^j \langle w_j, B(j, t)w_j \rangle \geq 0, \quad j = 0, 1, \quad (9.2.8)$$

for all vectors w_j satisfying

$$L_0 w_0 = 0, \quad L_1 w_1 = 0, \quad (9.2.9)$$

then any smooth solution of Eqs. (9.2.6) and (9.2.7) satisfies an energy estimate of the form of Eq. (9.2.3).

Proof. Integration by parts gives us

$$\begin{aligned} \frac{d}{dt} \|u\|^2 &= (u, Cu) + (Cu, u) + (u, Bu_x) + (Bu_x, u) \\ &= (u, Cu) + (Cu, u) - (u, B_x u) + \langle u, Bu \rangle |_0^1 \\ &\leq \text{constant } \|u\|^2. \end{aligned}$$

This proves the theorem.

As an example we consider the linearized Euler equations

$$\begin{bmatrix} u \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & a^2/R \\ R & U \end{bmatrix} \begin{bmatrix} u \\ \rho \end{bmatrix}_x = 0 \quad (9.2.10)$$

discussed in Section 6.1. Introducing the new variables $\tilde{u} = u$ and $\tilde{\rho} = a\rho/R$ gives us a symmetric system

$$\begin{bmatrix} \tilde{u} \\ \tilde{\rho} \end{bmatrix}_t = \tilde{B} \begin{bmatrix} \tilde{u} \\ \tilde{\rho} \end{bmatrix}_x, \quad \tilde{B} = - \begin{bmatrix} U & a \\ a & U \end{bmatrix}. \quad (9.2.11)$$

The eigenvalues of \tilde{B} are

$$\lambda = -(U \pm a), \quad a > 0.$$

To discuss the boundary conditions we have to distinguish between three cases:

1. *Supersonic Inflow*, $U(0, t) > a(0, t)$. There are two characteristics that enter the region, and we have to specify \tilde{u} and $\tilde{\rho}$. The homogeneous

boundary conditions are

$$\tilde{u} = \tilde{\rho} = 0.$$

Condition (9.2.8) is satisfied, because $\langle w_0, Bw_0 \rangle = 0$.

2. *Subsonic Flow*, $|U(0, t)| < a(0, t)$. There is only one characteristic that enters the region, and we use

$$\tilde{u} = -\alpha \tilde{\rho},$$

where α is real, as a boundary condition. For all $w_0 = (w_0^{(1)}, w_0^{(2)})$ with $w_0^{(1)} = -\alpha w_0^{(2)}$ we obtain

$$\begin{aligned} \langle w_0, \tilde{B}(0, t)w_0 \rangle &= (2a\alpha - U(\alpha^2 + 1))|w_0^{(2)}|^2, \\ &= (2(a - U)\alpha - U(1 - \alpha)^2)|w_0^{(2)}|^2, \\ &\geq 0, \end{aligned}$$

provided $|1 - \alpha|$ is sufficiently small. Note that $\alpha = 1$ corresponds to specifying the ingoing characteristic variable $\tilde{u} + \tilde{\rho}$.

3. *Supersonic Outflow*, $U(0, t) < -a(0, t)$. Both characteristics are leaving the region, and no boundary condition may be given. The second boundary at $x = 1$ is treated in the same way.

REMARK. In Case 2, we can also include sonic inflow $U(0, t) = a(0, t)$, and in Case 3, we can include sonic outflow $U(0, t) = -a(0, t)$.

Until now, we have assumed homogeneous boundary conditions when using the energy method. For some classes of problems it is also possible to obtain an estimate in a direct way with inhomogeneous boundary conditions. As an example, consider the problem

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \quad 0 \leq x \leq 1, \tag{9.2.12a}$$

$$u(x, 0) = f(x), \tag{9.2.12b}$$

$$u(0, t) = g(t). \tag{9.2.12c}$$

Integration by parts gives us, for any $\eta > 0$,

$$\begin{aligned} \frac{d}{dt} \|u\|^2 &= -(u, u_x) - (u_x, u) \\ &= |u(0, t)|^2 - |u(1, t)|^2 \end{aligned}$$

$v = e^{-\eta t} u$ satisfies

$$\begin{aligned}\frac{d}{dt} \|v\|^2 + |v(\cdot, t)|_{\Gamma}^2 &\leq 2|v(0, t)|^2 \\ &\leq 2|e^{-\eta t} g(t)|^2\end{aligned}$$

where $|v(\cdot, t)|_{\Gamma}^2 = |v(0, t)|^2 + |v(1, t)|^2$. Therefore,

$$\begin{aligned}\|v(\cdot, T)\|^2 + \int_0^T |v(\cdot, t)|_{\Gamma}^2 dt &\\ \leq \|v(\cdot, 0)\|^2 + 2 \int_0^T |e^{-\eta t} g(t)|^2 dt &\end{aligned}$$

or, for $\eta > \eta_0 \geq 0$,

$$\|e^{-\eta T} u(\cdot, T)\|^2 + \int_0^T |e^{-\eta t} u(\cdot, t)|_{\Gamma}^2 dt \leq C(\|f(\cdot)\|^2 + \int_0^T |e^{-\eta t} g(t)|^2 dt) \quad (9.2.13)$$

In this estimate, one can directly read off the dependence of u on the initial data f and the boundary data g . A forcing function F can also be included in Eq. (9.2.12a), and the estimate is obtained by using Duhamel's principle.

As demonstrated in Section 9.1, hyperbolic systems in one space dimension with constant coefficients can always be transformed to diagonal form, and, if the ingoing characteristic dependent variables are prescribed at the boundaries, we get the completely decoupled problem

$$\begin{aligned}\frac{\partial u}{\partial t} &= \Lambda \frac{\partial u}{\partial x} + F, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \\ u^{II}(0, t) &= g_0(t), \\ u^I(1, t) &= g_1(t).\end{aligned} \quad (9.2.14)$$

Here u^I and u^{II} correspond to the positive and negative eigenvalues of Λ , respectively. By applying the technique used for the scalar problem (9.2.12), we immediately obtain the estimate (9.2.13), where $|g(t)|^2$ is interpreted as $|g_0(t)|^2 + |g_1(t)|^2$. In other words, we can always find boundary conditions such that, with an arbitrary forcing function F , initial function f , and boundary functions g_0, g_1 , an estimate of the form of Eq. (9.2.13) holds.

Indeed, we can obtain such an estimate for general boundary conditions of the type of Eq. (9.2.2) in essentially the same direct way. However, when dealing with the discrete case in the next chapter, that will not generally be possible.

We have derived the energy estimates under the assumption that a smooth solution exists. This is no restriction, because we have already constructed such a solution in the previous section by the method of characteristics. However, one can also prove the existence of a solution in the following way. We construct difference approximations, which satisfy the corresponding discrete estimates. Then one can interpolate the discrete solution in such a way that the interpolant is smooth and converges as $h, k \rightarrow 0$ to the solution of our problem.

This is a general principle: Given an initial boundary value problem, one derives estimates under the assumption that a smooth solution exists. Then one constructs a difference approximation whose solutions satisfy corresponding discrete estimates. Suitable interpolants converge as $h, k \rightarrow 0$ to the solution of the problem.

EXERCISES

- 9.2.1.** Derive boundary conditions at $x = 1$ for the system (9.2.10) such that an estimate

$$\|\mathbf{u}(\cdot, t)\| \leq K\|\mathbf{u}(\cdot, 0)\|$$

holds for $\mathbf{u} = (u, \rho)^T$. What is the smallest possible value of K ?

- 9.2.2.** Consider the system (9.2.11) with the inhomogeneous versions of the boundary conditions discussed above. Derive an estimate of the form of Eq. (9.2.13). Prove that such an estimate holds in general.

9.3. ENERGY ESTIMATES FOR PARABOLIC DIFFERENTIAL EQUATIONS IN ONE SPACE DIMENSION

The simplest parabolic initial-boundary-value problem is the normalized heat equation

$$u_t = u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (9.3.1)$$

with initial conditions

$$u(x, 0) = f(x), \quad (9.3.2)$$

and the so called *Dirichlet boundary conditions*

$$u(0, t) = u(1, t) = 0. \quad (9.3.3)$$

(See Figure 9.3.1.)

Assume that Eqs. (9.3.1) to (9.3.3) have a smooth solution. We want to derive an energy estimate. Lemma 9.2.1 gives us

$$\begin{aligned} \frac{d}{dt} (u, u) &= (u, u_t) + (u_t, u) = (u, u_{xx}) + (u_{xx}, u), \\ &= -2\|u_x\|^2 + (\bar{u}u_x + \bar{u}_x u)|_0^1 = -2\|u_x\|^2 \leq 0, \end{aligned}$$

or

$$\|u(\cdot, t)\|^2 \leq \|u(\cdot, 0)\|^2 = \|f(\cdot)\|^2.$$

This estimate can be generalized to the equation

$$u_t = a(x, t)u_{xx} + b(x, t)u_x + c(x, t)u, \quad (9.3.4)$$

with the same initial and boundary conditions. Here a, b , and c are smooth functions with $a(x, t) \geq a_0 > 0$. Now we obtain

$$\frac{d}{dt} (u, u) = I + II + III,$$

where, by Lemma 9.2.1,

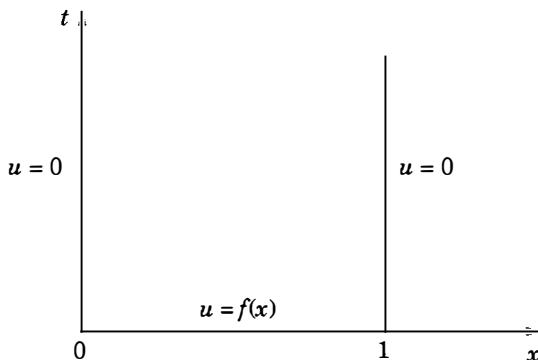


Figure 9.3.1.

$$\begin{aligned}
I &= (u, au_{xx}) + (au_{xx}, u), \\
&= -(u_x, au_x) - (u, a_x u_x) - (au_x, u_x) - (a_x u_x, u) + a(\bar{u}u_x + \bar{u}_x u)_0^1, \\
&\leq -2(u_x, au_x) + 2\|a_x\|_\infty \|u\| \|u_x\|, \\
&\leq -2a_0 \|u_x\|^2 + 2\|a_x\|_\infty \sqrt{\frac{2}{a_0}} \|u\| \sqrt{\frac{a_0}{2}} \|u_x\|, \\
&\leq -\frac{3}{2}a_0 \|u_x\|^2 + 2\left(\frac{\|a_x\|_\infty^2}{a_0}\right) \|u\|^2; \\
II &= (u, bu_x) + (bu_x, u) \leq 2\|b\|_\infty \|u\| \|u_x\|, \\
&\leq \frac{\|b\|_\infty^2}{a_0} \|u\|^2 + a_0 \|u_x\|^2; \\
III &= (u, cu) + (cu, u) \leq 2\|c\|_\infty \|u\|^2.
\end{aligned} \tag{9.3.5}$$

Thus,

$$\frac{d}{dt} \|u\|^2 \leq -\frac{a_0}{2} \|u_x\|^2 + \alpha \|u\|^2, \quad \alpha = \frac{2\|a_x\|_\infty^2 + \|b\|_\infty^2}{a_0} + 2\|c\|_\infty, \tag{9.3.6}$$

or

$$\|u(\cdot, t)\|^2 \leq e^{\alpha t} \|u(\cdot, 0)\|^2.$$

Instead of Dirichlet boundary conditions we can use

$$u_x(j, t) + r_j u(j, t) = 0, \quad j = 0, 1. \tag{9.3.7}$$

To obtain an energy estimate, we need a so called *Sobolev inequality*, as in Lemma 9.3.1.

Lemma 9.3.1. *Let $f \in C^1(0 \leq x \leq \ell)$. For every $\epsilon > 0$*

$$\|f\|_\infty^2 \leq \epsilon \|f_x\|^2 + (\epsilon^{-1} + \ell^{-1}) \|f\|^2.$$

Proof. Let x_1 and x_2 be points with

$$|f(x_1)| = \min_x |f(x)|, \quad |f(x_2)| = \max_x |f(x)| = \|f\|_\infty.$$

Without restriction, we can assume that $x_1 < x_2$. Then

$$\int_{x_1}^{x_2} \bar{f} f_x dx = |f|^2|_{x_1}^{x_2} - \int_{x_1}^{x_2} \bar{f}_x f dx ;$$

that is,

$$\begin{aligned} \|f\|_\infty^2 - |f(x_1)|^2 &\leq 2 \int_{x_1}^{x_2} |f| |f_x| dx \leq 2 \int_0^\ell |f| |f_x| dx, \\ &\leq 2\sqrt{\epsilon} \|f_x\| \frac{1}{\sqrt{\epsilon}} \|f\| \leq \epsilon \|f_x\|^2 + \epsilon^{-1} \|f\|^2. \end{aligned}$$

Observing that $\ell |f(x_1)|^2 \leq \|f\|^2$ the lemma follows.

REMARK. Lemma 9.3.1 holds for $\ell = \infty$ (Exercise 9.3.1).

Now consider Eq. (9.3.4) with the boundary conditions (9.3.7). By Eq. (9.3.5), the only new terms in the energy estimate, compared with the Dirichlet case, come from the boundary terms

$$E = a(\bar{u}u_x + \bar{u}_x u)|_0^1 = a(-\bar{u}ru - \bar{r}uu)|_0^1.$$

By Lemma 9.3.1, these can be estimated by

$$\begin{aligned} E &\leq 2\|a\|_\infty(|r_0| + |r_1|)\|u\|_\infty^2, \\ &\leq 2\|a\|_\infty\epsilon(|r_0| + |r_1|)\|u_x\|^2 + 2\|a\|_\infty(\epsilon^{-1} + 1)(|r_0| + |r_1|)\|u\|^2. \end{aligned}$$

Choosing $\epsilon = (a_0/8)(\|a\|_\infty(|r_0| + |r_1|))^{-1}$, we obtain instead of Eq. (9.3.6)

$$\begin{aligned} \frac{d}{dt} \|u\|^2 &\leq -\frac{a_0}{4} \|u_x\|^2 + \alpha^* \|u\|^2, \\ \alpha^* &= \alpha + 2\|a\|_\infty(\epsilon^{-1} + 1)(|r_0| + |r_1|). \end{aligned}$$

Thus, we have an energy estimate in this case.

We can also generalize the results to systems

$$\begin{aligned} \frac{\partial u}{\partial t} &= Au_{xx} + Bu_x + Cu, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{9.3.8}$$

where u is a vector-valued function and $A = A(x, t)$, $B = B(x, t)$, and $C = C(x, t)$ are $m \times m$ matrices that depend smoothly on x, t . We assume that

$$A + A^* \geq 2a_0 I, \quad a_0 > 0 \text{ constant.} \quad (9.3.9)$$

The boundary conditions consist of m linearly independent relations at each of the boundaries $x = 0$ and $x = 1$:

$$R_{1j}u_x(j, t) + R_{0j}u(j, t) = 0, \quad j = 0, 1. \quad (9.3.10)$$

To obtain an energy estimate we make the following assumption.

Assumption 9.3.1. *The boundary conditions are such that, for all vectors v_j and w_j with*

$$R_{1j}w_j + R_{0j}v_j = 0, \quad j = 0, 1,$$

the inequalities

$$(-1)^{j+1}(\langle v_j, Aw_j \rangle + \langle Aw_j, v_j \rangle) \leq c|v_j|^2, \quad c \geq 0 \text{ constant,} \quad (9.3.11)$$

hold.

REMARK. In the next chapter, we will use the Laplace transform technique to show that problems can be well-posed under less restrictive conditions than Eq. (9.3.11).

We can prove the following theorem.

Theorem 9.3.1. *If Eqs. (9.3.9) and (9.3.11) hold, then the smooth solutions of the initial-boundary-value problem shown in Eqs. (9.3.8) and (9.3.10) satisfy an energy estimate.*

Proof. As for the scalar equation (9.3.4), integration by parts gives us

$$\frac{d}{dt} \|u\|^2 = I + II + III,$$

where

$$\begin{aligned} I &= (u, Au_{xx}) + (Au_{xx}, u), \\ &= -(u_x, Au_x) - (u, A_x u_x) - (Au_x, u_x) - (A_x u_x, u) \\ &\quad + (\langle Au_x, u \rangle + \langle u, Au_x \rangle)|_0^1, \\ &\leq - (u_x, (A + A^*)u_x) + 2\|A_x\|_\infty \|u_x\| \|u\| + c(|u(0, t)|^2 + |u(1, t)|^2) \\ &\leq - 2a_0 \|u_x\|^2 + 2\|A_x\|_\infty \|u_x\| \|u\| + c(|u(0, t)|^2 + |u(1, t)|^2). \end{aligned}$$

By using Lemma 9.3.1 we obtain, as for the scalar case,

$$I \leq -c_1 \|u_x\|^2 + c_2 \|u\|^2, \quad c_1 > 0, \quad c_2 > 0.$$

The terms II and III have the same structure as the corresponding terms in Eq. (9.3.5) for the scalar case and, therefore, we obtain the estimates

$$\begin{aligned} II &\leq c_3 \|u\|^2 + \delta \|u_x\|^2, \quad \delta > 0 \\ III &\leq 2\|C\|_\infty \|u\|^2. \end{aligned}$$

By choosing $\delta < c_1$, we have an energy estimate.

We have shown that the smooth solutions of the above initial-boundary-value problem satisfy an energy estimate. This does not guarantee that such solutions exist. However, as for hyperbolic systems, one can prove existence using difference approximations provided that the initial and boundary conditions are compatible. The compatibility conditions are certainly satisfied if $f(x)$ has compact support in $0 \leq x \leq 1$, that is, if $f(x)$ vanishes in a neighborhood of $x = 0, 1$. If the compatibility conditions are not satisfied, then we approximate f by a sequence $\{f_\nu\}$ with compact support, and we define, in the same way as earlier, a unique generalized solution that still satisfies the energy estimate. One can show that the generalized solution is smooth except in the corners $x = 0, 1$; where $t = 0$.

Theorem 9.3.1 shows that the existence of energy estimates is independent of the form of the lower order terms (first- and zero-order terms). Therefore, B does not need to have real eigenvalues.

In many applications, systems are of the form

$$u_t = Bu_x + \nu u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (9.3.12)$$

where ν is a constant with $0 < \nu \ll 1$. For example, in fluid flow problems, ν represents a small viscosity. In this case, we still obtain an energy estimate, even if the eigenvalues of B are complex. However, the exponential growth constant α in the energy estimate is proportional to ν^{-1} . If B is symmetric, we can often do better. As an example, we consider Eq. (9.3.12) with

$$u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b \\ b & 0 \end{bmatrix}, \quad b_{11}, b \text{ real constants}, \quad (9.3.13a)$$

and boundary conditions

$$u^{(1)}(0, t) = u^{(1)}(1, t) = 0, \quad u_x^{(2)}(0, t) = u_x^{(2)}(1, t) = 0. \quad (9.3.13b)$$

Then

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

One can also show that, as $\nu \rightarrow 0$, the solutions converge to the solution of the “inviscid” problem

$$u_t = Bu_x \quad (9.3.14a)$$

with boundary conditions

$$u^{(1)}(0, t) = u^{(1)}(1, t) = 0 \quad (9.3.14b)$$

(see Exercise 9.3.3).

EXERCISES

9.3.1. Prove Lemma 9.3.1 for $\ell = \infty$.

9.3.2. Assume that the matrix A in Eq. (9.3.8) is positive definite and diagonal. What are the most general matrices R_{0j}, R_{1j} such that Eq. (9.3.11) is satisfied?

9.3.3. Prove that the solutions of Eqs. (9.3.12) and (9.3.13) converge to the solution of Eq. (9.3.14) as $\nu \rightarrow 0$.

9.4. WELL-POSED PROBLEMS

In the last two sections, we derived energy estimates for hyperbolic and parabolic systems and discussed possible boundary conditions that lead to well-posed problems. In this section, we define well-posed problems in general. We consider systems of partial differential equations

$$u_t = P \left(x, t, \frac{\partial}{\partial x} \right) u + F, \quad 0 \leq x \leq l, \quad t \geq t_0, \\ u(x, t_0) = f(x), \quad (9.4.1)$$

in the strip $0 \leq x \leq l, t \geq t_0$. (If $l = \infty$, we obtain a quarter space problem.) Here $u = (u^{(1)}, \dots, u^{(m)})^T$ is a vector function with m components and

$$P\left(x, t, \frac{\partial}{\partial x}\right) = \sum_{\nu=0}^p A_\nu(x, t) \frac{\partial^\nu}{\partial x^\nu}$$

is a differential operator of order p with smooth matrix coefficients. At $x = 0, l$ we give boundary conditions

$$L_0\left(t, \frac{\partial}{\partial x}\right)u(0, t) = g_0, \quad L_1\left(t, \frac{\partial}{\partial x}\right)u(l, t) = g_1. \quad (9.4.2)$$

Here L_0 and L_1 are differential operators of order r . In most applications $r \leq p - 1$. However, there are cases, where $r \geq p$. In analogy with the corresponding definition for the Cauchy problem, we define well-posedness for homogeneous boundary conditions.

Definition 9.4.1. Consider Eqs. (9.4.1) and (9.4.2) with $F = g_0 = g_1 = 0$. We call the problem well posed if, for every $f \in C^\infty$ that vanishes in a neighborhood of $x = 0, l$, it has a unique smooth solution that satisfies the estimate

$$\|u(\cdot, t)\| \leq K e^{\alpha(t-t_0)} \|u(\cdot, t_0)\|, \quad (9.4.3)$$

where K and α do not depend on f and t_0 .

REMARK. If the coefficients of P, L_0 , and L_1 do not depend on t , then we can replace Eq. (9.4.3) by

$$\|u(\cdot, t)\| \leq K e^{\alpha t} \|u(\cdot, 0)\|,$$

because the transformation $t' = t - t_0$ does not change the differential equation or boundary conditions.

As before, we can define a solution operator $S(t, t_0)$. If $F = g_0 = g_1 = 0$, then we can write the solution of the differential equation in the form

$$u(x, t) = S(t, t_0)u(x, t_0),$$

and Eq. (9.4.3) says that

$$\|S(t, t_0)\| \leq K e^{\alpha(t-t_0)}.$$

This again shows that we can extend the admissible initial data to all functions in L_2 .

Also, as before, we can use the solution operator to solve the inhomogeneous

differential equation with homogeneous boundary conditions ($g_0 = g_1 = 0$). If $F \in C^\infty(x, t)$ vanishes in a neighborhood of $x = 0, l$, then, by Duhamel's principle, the solution is

$$u(x, t) = S(t, t_0)f(x) + \int_{t_0}^t S(t, \tau)F(x, \tau)d\tau.$$

Also, $u \in C^\infty$. Again we can extend the admissible F to all functions $F \in L_2(x, t)$.

We can also solve problems with inhomogeneous boundary conditions provided we can find a smooth function φ that satisfies the boundary conditions, that is,

$$L_0\varphi(0, t) = g_0, \quad L_1\varphi(l, t) = g_1.$$

In this case $\tilde{u} = u - \varphi$ satisfies homogeneous boundary conditions. Also, \tilde{u} satisfies Eq. (9.4.1) with f and F replaced by $\tilde{f} = f - \varphi$ and $\tilde{F} = F - \varphi_t + P(x, t, \partial/\partial x)\varphi$, respectively. Thus, the estimate for \tilde{u} depends on the derivatives of φ . In general, this does not cause any difficulty with respect to x , because we can choose φ as a smooth function of x . However, φ_t can only be bounded by $\partial g_j/\partial t$. Thus, the boundary data must be differentiable. So it is desirable that the reduction process be avoided. Instead, one would like to estimate u directly in terms of F, f , and g . This leads to the following definition.

Definition 9.4.2. *The problem is strongly well posed if it is well posed and, instead of Eq. (9.4.3), the estimate*

$$\|u(\cdot, t)\|^2 \leq K(t, t_0) \left(\|u(\cdot, t_0)\|^2 + \int_{t_0}^t (\|F(\cdot, \tau)\|^2 + |g_0(\tau)|^2 + |g_1(\tau)|^2) d\tau \right) \quad (9.4.4)$$

holds. Here $K(t, t_0)$ is a function that is bounded in every finite time interval and does not depend on the data.

One can prove the following theorem.

Theorem 9.4.1. *For hyperbolic systems (9.2.1) with general inhomogeneous boundary conditions*

$$\begin{aligned} u^{II}(0, t) &= R^I(t)u^I(0, t) + g_0(t), \\ u^I(1, t) &= R^{II}(t)u^{II}(1, t) + g_1(t), \end{aligned}$$

and for parabolic systems (9.3.8) with Neumann boundary conditions

$$u_x(j, t) = R_j(t)u(j, t) + g_j(t), \quad j = 0, 1, \quad (9.4.5)$$

the initial boundary value problem is strongly well posed

Proof. For hyperbolic problems we have indicated a proof in Section 9.2. For parabolic problems, the proof follows by an easy modification of the estimates in Section 9.3.

The definition of strong well-posedness is given for general differential operators. As we demonstrated earlier, we obtain stronger estimates including the boundary values for hyperbolic equations (Exercise 9.4.1). Furthermore, if a second-order parabolic problem is strongly well posed, then we can estimate the derivative of the solution as well as the boundary values (Exercise 9.4.2).

It should be pointed out that strongly well posed is, in general, a more stringent requirement than well posed. Parabolic problems with other than Neumann boundary conditions are not generally strongly well posed. In more than one space dimension, hyperbolic problems can be well posed but not strongly well posed. The same is true for higher order systems. However, even for general systems such as (9.4.1), one can always find boundary conditions such that the initial-boundary-value problem is strongly well posed if P is a semibounded operator for the Cauchy problem.

EXERCISES

9.4.1. Carry out the proof of Theorem 9.4.1 in detail and prove that the estimate

$$\begin{aligned} \|u(\cdot, t)\|^2 &+ \int_{t_0}^t (|u(0, \tau)|^2 + |u(1, \tau)|^2) d\tau \\ &\leq K(t, t_0) \left(\|u(\cdot, t_0)\|^2 + \int_{t_0}^t (\|F(\cdot, \tau)\|^2 + |g_0(\tau)|^2 + |g_1(\tau)|^2) d\tau \right) \end{aligned} \quad (9.4.6)$$

holds for hyperbolic equations.

9.4.2. Prove Theorem 9.4.1 for second order parabolic systems with Neumann boundary conditions and derive the estimate

$$\begin{aligned} & \|u(\cdot, t)\|^2 + \int_{t_0}^t (\|u_x(\cdot, \tau)\|^2 + |u(0, \tau)|^2 + |u(1, \tau)|^2) d\tau \\ & \leq K(t, t_0) \left(\|u(\cdot, t_0)\|^2 + \int_{t_0}^t (\|F(\cdot, \tau)\|^2 + |g_0(\tau)|^2 + |g_1(\tau)|^2) d\tau \right). \end{aligned} \quad (9.4.7)$$

9.5. SEMIBOUNDED OPERATORS

In the previous sections, we considered differential equations

$$u_t = Pu, \quad 0 \leq x \leq l, \quad t \geq t_0,$$

with boundary conditions consisting of homogeneous linear combinations of u and its derivatives

$$L_0 u(0, t) = L_1 u(l, t) = 0.$$

If $l = \infty$, the boundary condition at $x = l$ is replaced by the condition $\|u(\cdot, t)\| < \infty$. For all practical purposes we can assume that $u(x, t)$ and all its derivatives converge to zero as $x \rightarrow \infty$.

In most cases, the solutions satisfied an energy estimate of the form

$$2\operatorname{Re}(Pu, u) = (Pu, u) + (u, Pu) \leq 2\alpha\|u\|^2.$$

We now formalize these results to some extent. For every fixed t , the differential operator P can be considered as an operator \mathcal{P} in the functional analysis sense, if we make its domain \mathcal{V} of definition precise. We define \mathcal{V} to be the set of functions

$$\mathcal{V} := \{w \in C^\infty, L_0 w(0) = L_1 w(l) = 0\}.$$

[If $l = \infty$, the condition $L_1 w(l) = 0$ is replaced by $\|w\| < \infty$.]

Definition 9.5.1. *We call \mathcal{P} semibounded if there exists a constant α such that, for all t and all $w \in \mathcal{V}$,*

$$2\operatorname{Re}(Pw, w) = (Pw, w) + (w, Pw) \leq 2\alpha\|w\|^2.$$

(Later in this book we will also use the notation P for \mathcal{P}).

If \mathcal{P} is semibounded and u is a solution of the differential equation with $u \in \mathcal{V}$ for every fixed t , then

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

and the basic energy estimate follows.

A theorem of the following type would be ideal. If \mathcal{P} is semibounded, then the corresponding initial-boundary-value problem is well posed. However, this is not true. If we change the domain \mathcal{V} to $\mathcal{V}_1 \subset \mathcal{V}$ by adding more boundary conditions, then the operator \mathcal{P}_1 is still semibounded, but the corresponding initial-boundary-value problem might not have a solution, because we may have overdetermined the solution at the boundary. Therefore, we define maximally semibounded as follows.

Definition 9.5.2. *\mathcal{P} is called maximally semibounded if the number q of linearly independent boundary conditions is minimal, that is, if there exist no boundary conditions such that \mathcal{P} is semibounded and their number is smaller than q .*

REMARK. One uses the word *maximal* because, in some sense, \mathcal{V} is as large as possible.

Let us apply the concept to

$$u_t = u_x, \quad 0 \leq x \leq 1, \quad t \geq 0.$$

We assume that at $x = 0, 1$ boundary conditions of the form

$$\sum_{j=0}^p a_j(x) \frac{\partial^j u(x, t)}{\partial x^j} = 0, \quad x = 0, 1,$$

are to be specified. We want to choose them so that \mathcal{P} will be maximally semibounded. For all $w \in C^\infty$ we have

$$(w_x, w) + (w, w_x) = |w|^2|_0^1.$$

Therefore, we must choose the boundary conditions so that

$$|w(1)|^2 - |w(0)|^2 \leq 2\alpha \|w\|^2.$$

Clearly, no boundary conditions are necessary for $x = 0$, and we need to bound

$$|w(1)|^2 \leq \alpha \|w\|^2.$$

This is only possible if

$$w(1) = 0.$$

Thus, the minimal number of boundary conditions is one.

If we consider systems

$$u_t = \Delta u_x,$$

then we arrive at our previous conditions, namely, one must express the ingoing characteristic variables in terms of those which are outgoing.

Let us apply the principle to the linearized Korteweg-de Vries equation

$$u_t = u_x + \delta u_{xxx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad \delta > 0.$$

For all $w \in C^\infty$, we have

$$(Pw, w) + (w, Pw) = (|w|^2 + \delta(\bar{w}w_{xx} + \bar{w}_{xx}w) - \delta|w_x|^2)|_0^1.$$

Thus, we must choose the boundary conditions to guarantee

$$(|w|^2 + \delta(\bar{w}w_{xx} + \bar{w}_{xx}w) - \delta|w_x|^2)|_0^1 \leq 2\alpha\|w\|^2.$$

This is only possible if

$$(|w|^2 + \delta(\bar{w}w_{xx} + \bar{w}_{xx}w) - \delta|w_x|^2)|_0^1 \leq 0.$$

For $x = 1$, we need one condition

$$w_{xx} = a_1 w, \quad \text{with} \quad 1 + 2\delta \operatorname{Re}(a_1) \leq 0,$$

or

$$w = 0.$$

For $x = 0$, we need two conditions, because $|w_x|^2$ appears with a positive multiplier. These conditions are

$$w_x = a_{01}w, \quad w_{xx} = a_{02}w, \quad \text{with} \quad 1 + \delta(2\operatorname{Re}(a_{02}) - |a_{01}|^2) \geq 0,$$

or

$$w = w_x = 0.$$

One can show that the resulting initial-boundary-value problem is well posed. In general, one can prove the following theorem.

Theorem 9.5.1. *Consider a general system*

$$u_t = \sum_{j=0}^p A_j(x, t) \frac{\partial^j u}{\partial x^j}, \quad 0 \leq x \leq \ell, \quad t \geq t_0,$$

with boundary conditions

$$\sum_{j=0}^r B_j(x, t) \frac{\partial^j u}{\partial x^j} = 0, \quad x = 0, \ell; \quad t \geq t_0.$$

Assume that A_p is nonsingular and that the coefficients are smooth. If the associated operator \mathcal{P} is maximally semibounded, then the initial-boundary-value problem is well posed.

We now discuss another general result. Consider a parabolic system

$$u_t^I = Au_{xx}^I, \quad A + A^* > \delta I > 0,$$

in n unknowns $u^I = (u^{(1)}, \dots, u^{(n)})^T$ and a symmetric hyperbolic system

$$u_t^{II} = Bu_x^{II},$$

in m unknowns $u^{II} = (u^{(n+1)}, \dots, u^{(n+m)})^T$. Assume that there are exactly r eigenvalues λ of B with $\lambda < 0$ at $x = 0$. Now we couple the systems to obtain

$$\begin{aligned} u_t &= Pu + F := \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} u_{xx} + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B \end{bmatrix} u_x + Cu + F, \\ u &= \begin{bmatrix} u^I \\ u^{II} \end{bmatrix}, \end{aligned} \tag{9.5.1}$$

and consider the quarter-space problem $0 \leq x < \infty, \quad t \geq 0$. At $x = 0$, we describe as boundary conditions the linear combinations

$$L_{10}u_x(0, t) + L_{00}u(0, t) = 0. \tag{9.5.2}$$

We are interested in solutions with $\|u(\cdot, t)\| < \infty$, and assume that $F(x, t), u(x, 0)$ are smooth functions with compact support. For this problem, one can prove the following theorem:

Theorem 9.5.2. *The operator \mathcal{P} associated with Eqs. (9.5.1) and (9.5.2) is maximally semibounded if the number of boundary conditions is equal to $n + r$ and*

$$\begin{aligned} & -\operatorname{Re}(\langle w^I(0), Aw_x^I(0) \rangle + \frac{1}{2}\langle w^{II}(0), Bw^{II}(0) \rangle + \langle w^I(0), B_{12}w^{II}(0) \rangle) \\ & \leq \text{constant } |w^I(0)|^2, \end{aligned} \quad (9.5.3)$$

for all t and all w that satisfy the boundary conditions. In this case, the quarter-space problem is well posed.

REMARK. We could have treated the problem in Theorem 9.5.2 on a bounded strip instead of on a quarter space. In that case, we would also need boundary conditions on the other boundary, these could have been derived by considering the corresponding left quarter-space problem. This technique of reducing problems to quarter space problems is an important tool.

As an example, we consider the one-dimensional version of the linearized Navier–Stokes equations (4.6.13) with constant coefficients. By introducing $\tilde{\rho} = ap/R$ as a new variable we get

$$\begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix} = - \begin{bmatrix} U & a \\ a & U \end{bmatrix} \begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix}_x + \nu \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u \\ \tilde{\rho} \end{bmatrix}_{xx}, \quad \nu = \frac{(\mu + \mu')}{R} > 0. \quad (9.5.4)$$

The uncoupled equations are

$$\begin{aligned} u_t &= \nu u_{xx}, \\ \tilde{\rho}_t &= -U\tilde{\rho}_x, \end{aligned}$$

and the expression (9.5.3) becomes

$$-\nu uu_x + \frac{U}{2} |\tilde{\rho}|^2 + au\tilde{\rho} \leq \text{constant } |u|^2, \quad x = 0. \quad (9.5.5)$$

Recalling that U is the velocity of the flow we have linearized around, we distinguish among three different cases.

1. *Inflow, $U > 0$.* In this case two boundary conditions have to be specified. The inequality (9.5.5) is satisfied if, for example,

$$u = \tilde{\rho} = 0, \quad \text{or} \quad u_x = 0, \quad \tilde{\rho} = 0.$$

2. *Rigid Wall*, $U = 0$. In this case we have to specify one boundary condition. The inequality (9.5.5) is satisfied if, for example,

$$u = 0,$$

which is the natural condition at a rigid wall.

3. *Outflow*, $U < 0$. One boundary condition is needed, and Eq. (9.5.5) is satisfied if, for example,

$$u = 0, \quad \text{or} \quad -\nu u_x + a\tilde{\rho} = 0.$$

In all three cases, the given conditions are stricter than necessary, since the constant in the right-hand side of Eq. (9.5.5) need not be zero.

EXERCISES

- 9.5.1.** If $\nu \rightarrow 0$ in Eq. (9.5.4), the limiting system is the linearized Euler equations. Derive boundary conditions that give well-posed problems for both systems. Is that possible for all values of U and a ?
- 9.5.2.** Derive boundary conditions such that the initial-boundary-value problem for

$$u_t = -u_{xxxx}, \quad 0 \leq x \leq 1, \quad t \geq 0.$$

is well posed.

9.6. QUARTER-SPACE PROBLEMS IN MORE THAN ONE SPACE DIMENSION

As an example, we consider hyperbolic systems

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x} + B \frac{\partial u}{\partial y} + Cu + F =: P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u + F. \quad (9.6.1)$$

All coefficients are smooth functions of $\mathbf{x} = (x, y, t)$, and $A = A(\mathbf{x}, t)$, $B = B(\mathbf{x}, t)$ are Hermitian matrices. $C = C(\mathbf{x}, t)$ can be a general matrix. In many applications, it is skew Hermitian, that is, $C = -C^*$.

9.6.1 Quarter-Space Problems

We consider Eq. (9.6.1) in the quarter space $0 \leq x < \infty$, $-\infty < y < \infty$, $t \geq 0$ as shown in Figure 9.6.1.

For $t = 0$, we give initial data

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \quad (9.6.2)$$

and at $x = 0$ we describe boundary conditions

$$L_0(y, t)u(0, y, t) = 0, \quad -\infty < y < \infty, \quad (9.6.3)$$

that are like Eq. (9.2.7), that is, they consist of linear relations between the components of u . We also assume that $A(0, y, t)$ is nonsingular.

REMARK. It is not necessary to assume that $A(0, y, t)$ be nonsingular. One can use the following condition. Let $\lambda_j(0, y, t)$ be an eigenvalue of $A(0, y, t)$. If $\lambda_j(0, y, t) = 0$ for some $y = y_0$, $t = t_0$, then either $\lambda_j \equiv 0$ for all \mathbf{x}, t in a neighborhood of $x = 0$, or $\lambda_j = x^p \tilde{\lambda}_j(\mathbf{x}, t)$, $\tilde{\lambda}_j \neq 0$, $p \geq 1$.

We now consider solutions that are 2π -periodic in y . Therefore, we assume that all coefficients and data have this property. Let

$$(u, v) = \int_0^{2\pi} \int_0^\infty \langle u, v \rangle dx dy, \quad \|u\|^2 = (u, u),$$

denote the L_2 scalar product and norm. We assume that $\|f\| < \infty$ and that we are interested in smooth solutions, for which

$$\|u(\cdot, t)\| < \infty, \quad (9.6.4)$$

for all t . We consider Eq. (9.6.4) as a boundary condition at $x = \infty$. One can again use difference approximations to prove that solutions exist.

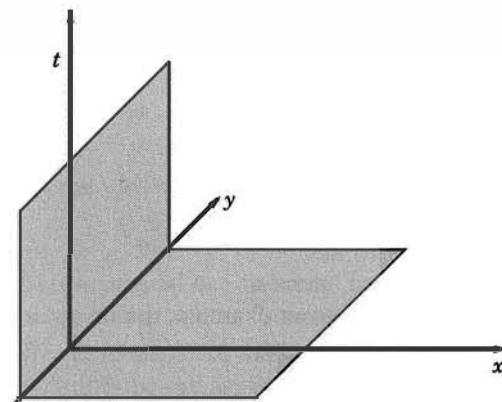


Figure 9.6.1.

Derive an energy estimate proceeding as before. Let u be a smooth solution to the homogeneous problem with $F = 0$. Then, we apply integration by parts to the right hand side of

$$\frac{d}{dt} \|u\|^2 = (u, Pu) + (Pu, u).$$

Observing that there are no boundary terms in the y direction and, because of Eq. (9.6.4), there are no boundary terms at $x = \infty$, we obtain, as in the one-dimensional case,

$$(u, Pu) + (Pu, u) = (u, (-A_x - B_y + C + C^*)u) \\ - \int_0^{2\pi} \langle u(0, y, t), A(0, y, t)u(0, y, t) \rangle dy.$$

Therefore, for every fixed y , the boundary term is of the same form as in the one-dimensional case, and we obtain an energy estimate

$$\frac{d}{dt} \|u\|^2 \leq 2\alpha \|u\|^2 + 2\|u\| \|F\|,$$

if the boundary conditions are such that

$$\langle u(0, y, t), A(0, y, t)u(0, y, t) \rangle \geq 0,$$

for all y, t . Definitions 9.4.1 and 9.4.2 are generalized to two space dimensions in an obvious way. One can prove the following theorem.

Theorem 9.6.1. *Assume that $A(0, y, t)$ is nonsingular and has exactly $m - r$ negative eigenvalues and that Eq. (9.6.3) consists of $m - r$ linearly independent relations. If, for all $w \in \mathcal{V} := \{L_0(y, t)w(0) = 0\}$,*

$$\langle w(0), A(0, y, t)w(0) \rangle \geq \delta |w(0)|^2, \quad \delta = \text{const.} \geq 0,$$

for all y, t , then the initial-boundary-value problem [Eqs. (9.6.1) to (9.6.4)] is strongly well posed if $\delta > 0$ and well posed if $\delta = 0$.

The concept of semibounded operators can be generalized to the two-dimensional case in an obvious way. In our example, the operator is semibounded if

$$\langle w(0), A(0, y, t)w(0) \rangle \geq 0, \quad w \in \mathcal{V} \tag{9.6.5}$$

for all y, t . If $A(0, y, t)$ has $m - r$ negative eigenvalues, one needs at least $m - r$

boundary conditions for Eq. (9.6.5) to hold. Thus, the operator is maximally semibounded if the number of boundary conditions is equal to the number of negative eigenvalues of A .

The results above show that boundary conditions must comply with the one-dimensional theory. This is also true for parabolic and mixed hyperbolic-parabolic systems. As an example, we consider the linearized and symmetrized Euler equations [see Eq. (4.6.5)]

$$\mathbf{u}_t = - \begin{bmatrix} U & 0 & a \\ 0 & U & 0 \\ a & 0 & U \end{bmatrix} \mathbf{u}_x - \begin{bmatrix} V & 0 & 0 \\ 0 & V & a \\ 0 & a & V \end{bmatrix} \mathbf{u}_y, \quad \mathbf{u} = \begin{bmatrix} u \\ v \\ \tilde{\rho} \end{bmatrix}. \quad (9.6.6)$$

The eigenvalues λ of A are

$$\begin{aligned} \lambda_1 &= -U, \\ \lambda_{2,3} &= -U \pm a, \end{aligned}$$

and we have

$$\langle \mathbf{u}, A\mathbf{u} \rangle = -U(|u|^2 + |v|^2 + |\tilde{\rho}|^2) - 2au\tilde{\rho}.$$

As in the one-dimensional case discussed in Section 9.2, the boundary conditions depend on U and a .

1. *Supersonic Inflow*, $U > a$. Three eigenvalues are negative, and all variables must be specified at the boundary:

$$u = v = \tilde{\rho} = 0.$$

2. *Subsonic Inflow*, $0 < U < a$. Two eigenvalues are negative, and we must specify two conditions. For example,

$$u = -\alpha\tilde{\rho}, \quad v = 0,$$

with

$$-U(\alpha^2 + 1) + 2\alpha a \geq 0,$$

will make $\langle \mathbf{u}, A\mathbf{u} \rangle \geq 0$.

3. *Rigid Wall*, $U = 0$. In this case the boundary is characteristic, A is singular and, according to the remark above, the function U must have special properties near the boundary. We need only specify one boundary condition

$$u = -\alpha \tilde{\rho}$$

with $\alpha \geq 0$. The most natural case is $\alpha = 0$ corresponding to flow that is parallel to the wall.

4. *Subsonic Outflow*, $-a < U < 0$. One eigenvalue is negative, and we need one boundary condition. For example,

$$u = -\alpha \tilde{\rho},$$

with

$$-U(\alpha^2 + 1) + 2\alpha a \geq 0,$$

or

$$\tilde{\rho} = 0$$

will make $\langle \mathbf{u}, A\mathbf{u} \rangle \geq 0$.

5. *Supersonic Outflow*, $U < -a$. All eigenvalues are positive and no boundary conditions may be given.

The energy method is very powerful, when it works, that is, when the boundary conditions are such that the boundary terms have the right sign. If this is not the case, then the method does not tell us anything, and we have to use Laplace transform methods instead. These are discussed in the next chapter.

9.6.2. Problems in General Domains

Next we consider the differential equation (9.6.1) in a general domain Ω in the x, y plane, bounded by a smooth curve $\partial\Omega$ (see Figure 9.6.2). We give initial

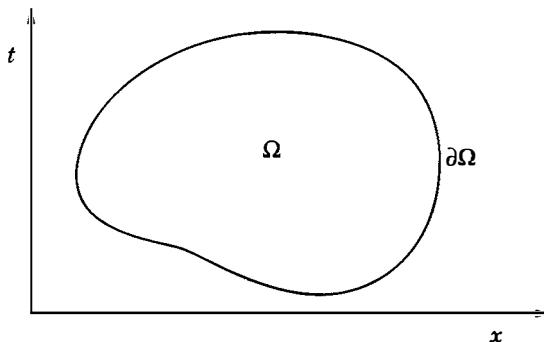


Figure 9.6.2.

data (9.6.2) for $\mathbf{x} \in \Omega$ and boundary conditions of the type of Eq. (9.6.3) on $\partial\Omega$. We want to show that this problem can be solved in terms of a quarter-space problem and a Cauchy problem. We assume that A, B , and C are defined in the whole \mathbf{x} plane and $F \equiv 0$.

Let $d > 0$ be a real number. At every point $\mathbf{x}_0 = (x_0, y_0)$ of $\partial\Omega$ we determine the inward normal and on it the point $\mathbf{x}_1 = (x_1, y_1)$ with $|\mathbf{x}_1 - \mathbf{x}_0| = d$. If d is sufficiently small (in relation to the curvature of $\partial\Omega$), then the process defines a subdomain Ω_1 , bounded by a smooth curve $\partial\Omega_1$, (see Figure 9.6.3). Now let $\varphi \in C^\infty$ be a function with $\varphi \equiv 1$ in the neighborhood of $\partial\Omega$ and $\varphi \equiv 0$ in Ω_1 and the neighborhood of $\partial\Omega_1$. Multiply Eq. (9.6.1) by φ and introduce new variables by $u_1 = \varphi u$, $u_2 = (1 - \varphi)u$, $u = u_1 + u_2$. Then we obtain the system

$$u_{1t} = Au_{1x} + Bu_{1y} + Cu_1 - (A\varphi_x + B\varphi_y)(u_1 + u_2). \quad (9.6.7a)$$

Correspondingly, multiplying Eq. (9.6.1) by $(1 - \varphi)$ gives us

$$u_{2t} = Au_{2x} + Bu_{2y} + Cu_2 + (A\varphi_x + B\varphi_y)(u_1 + u_2), \quad \mathbf{x} \in \Omega. \quad (9.6.7b)$$

By construction, $u_2 \equiv 0$ in the neighborhood of $\partial\Omega$. Therefore, we can extend the definition of u_2 to the whole \mathbf{x} plane. If $(A\varphi_x + B\varphi_y)u_1$ were a known function, then Eq. (9.6.7b) is a Cauchy problem for u_2 .

Now we construct a mapping

$$\tilde{\mathbf{x}} = \Psi(\mathbf{x}) \in C^\infty,$$

which transforms the region $\Omega - \Omega_1$ into the rectangle $\tilde{\Omega} := \{0 \leq \tilde{x} \leq 1, 0 \leq \tilde{y} \leq 2\pi\}$. Here the lines $\tilde{x} = 0$ and $\tilde{x} = 1$ correspond to the boundary curves $\partial\Omega$ and $\partial\Omega_1$, respectively. We assume also that the derivatives $\partial/\partial n, \partial/\partial s$ in the normal and tangential direction, respectively, of $\partial\Omega$ are transformed into $\partial/\partial \tilde{x}$ and $\partial/\partial \tilde{y}$, respectively.

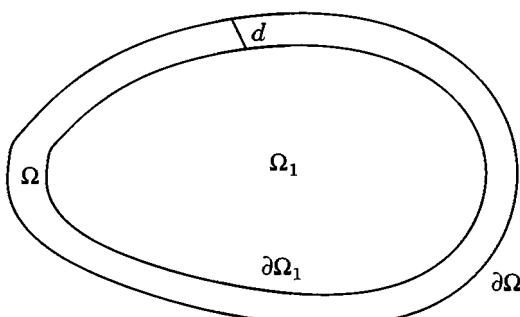


Figure 9.6.3. Partition of a domain with smooth boundary into two subdomains.

After the transformation, the system (9.6.7a) becomes

$$\tilde{u}_{1t} = \tilde{A}\tilde{u}_{1\tilde{x}} + \tilde{B}\tilde{u}_{1\tilde{y}} + C\tilde{u}_1 - (A\varphi_x + B\varphi_y)(\tilde{u}_1 + \tilde{u}_2), \quad \tilde{x} \in \tilde{\Omega}, \quad (9.6.7c)$$

where $\tilde{u}_j(\tilde{x}, t) = u_j(x, t)$ and $j = 1, 2$. For $\tilde{x} = 0$

$$\bar{A}(0, \tilde{y}, t) = A(0, \tilde{y}, t)\cos \alpha + B(0, \tilde{y}, t)\sin \alpha,$$

with α denoting the angle between the x axis and the inward normal derivative (see Figure 9.6.4).

The boundary conditions on $\partial\Omega$ are transformed into boundary conditions on $\tilde{x} = 0$. Also, \tilde{u}_1 and \tilde{u}_2 are 2π -periodic with respect to \tilde{y} . By construction, φ_x, φ_y , and u_1 vanish in a neighborhood of $\tilde{x}_1 = 1$. Therefore, we can extend the definition of u_1 to the whole quarter space $\tilde{x} \geq 0$. If we knew the term $(A\varphi_x + B\varphi_y)\tilde{u}_2$, then \tilde{u}_1 would be the solution of a quarter-space problem.

The systems (9.6.7b) and (9.6.7c) are only coupled by lower order terms. From our previous results, we know that these lower order terms have no influence upon whether or not a problem is well posed. Therefore, the boundary conditions on $\partial\Omega$ must be such that the quarter-space problem

$$\tilde{u}_{1t} = \tilde{A}\tilde{u}_{1\tilde{x}} + \tilde{B}\tilde{u}_{1\tilde{y}} \quad (9.6.8)$$

is well posed, and we obtain the following theorem for the original problem.

Theorem 9.6.2. Assume that $\bar{A} = A \cos \alpha + B \sin \alpha$ is nowhere singular on $\partial\Omega$ and has exactly $m - r$ negative eigenvalues. Then the initial-boundary-value problem is well posed, if

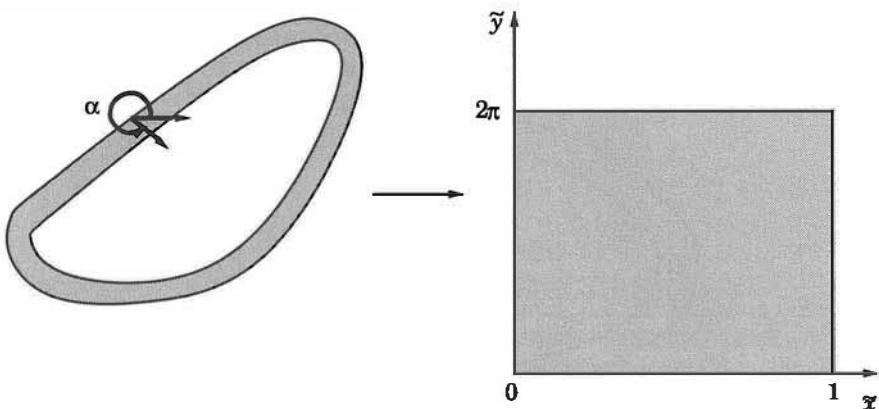


Figure 9.6.4. Mapping of the domain $\Omega - \Omega_1$ into a rectangle.

$$\langle u, \tilde{A}u \rangle \geq 0 \quad (9.6.9)$$

for all $\mathbf{x} \in \partial\Omega$ and all vectors u that satisfy the boundary conditions.

Proof. Using the construction above we must solve Eqs. (9.6.7b) and (9.6.7c). We solve these by iteration

$$\begin{aligned}\tilde{u}_{1t}^{[n+1]} &= \tilde{A}\tilde{u}_{1\tilde{x}}^{[n+1]} + \tilde{B}\tilde{u}_{1\tilde{y}}^{[n+1]} + C\tilde{u}_1^{[n+1]} \\ &\quad - (A\varphi_x + B\varphi_y)(\tilde{u}_1^{[n+1]} + \tilde{u}_2^{[n]}), \tilde{\mathbf{x}} \in \tilde{\Omega}, \\ u_{2t}^{[n+1]} &= Au_{2x}^{[n+1]} + Bu_{2y}^{[n+1]} + Cu_2^{[n+1]} \\ &\quad + (A\varphi_x + B\varphi_y)(u_1^{[n]} + u_2^{[n+1]}), \mathbf{x} \in \mathbb{R}^2.\end{aligned}$$

At every step of the iteration we solve a Cauchy problem and a quarter-space problem. By construction, both problems are well posed, and one can show that the iteration converges to a solution of the original problem.

This construction is easily extended to parabolic and mixed parabolic-hyperbolic systems.

EXERCISES

9.6.1. Consider the system

$$\begin{aligned}u_t &= \begin{bmatrix} 1 & 1 \\ a & 1 \end{bmatrix} u_x + \begin{bmatrix} 1 & b \\ 1 & 1 \end{bmatrix} u_y, \\ 0 \leq x < \infty, \quad -\infty < y < \infty, \quad t \geq 0.\end{aligned}$$

For which values of a and b can an energy estimate be obtained? Derive the most general well-posed boundary conditions for this system.

9.6.2. Consider the symmetric linearized Euler equations (9.6.6) with $U > a$, $V = 0$ in a circular disc Ω . Define well-posed boundary conditions on $\partial\Omega$.

BIBLIOGRAPHIC NOTES

The results in this chapter are classical. More details and references are given in *Kreiss and Lorenz* (1989).

10

THE LAPLACE TRANSFORM METHOD FOR INITIAL-BOUNDARY-VALUE PROBLEMS

10.1 SOLUTION OF HYPERBOLIC SYSTEMS

Initial-boundary-value problems for systems with constant coefficients can be solved using the Laplace transform. As we will see, the Laplace transform method is often the only way we can decide on the stability of a given finite difference scheme. We only use elementary properties of the Laplace transform (see Appendix A.2).

We consider the quarter-space problem for the system

$$\frac{\partial}{\partial t} \begin{bmatrix} u^I \\ u^{II} \end{bmatrix} = A \frac{\partial}{\partial x} \begin{bmatrix} u^I \\ u^{II} \end{bmatrix} + F, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (10.1.1a)$$

where A is diagonal and

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \quad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

We prescribe initial data

$$u(x, 0) = f(x), \quad (10.1.1b)$$

and boundary conditions

$$\begin{aligned} L''u''(0,t) + L' u'(0,t) &= g, \\ \|u(\cdot, t)\| &< \infty \text{ for every fixed } t. \end{aligned} \quad (10.1.1c)$$

where L' and L'' are constant matrices. We assume that F, f , and g are smooth functions with compact support.

We know already that the above problem is well posed if, and only if, L'' is nonsingular. However, we arrive at the same conclusion using the Laplace transform. We start with the following lemma.

Lemma 10.1.1. *Consider the initial-boundary-value problem (10.1.1) with $F \equiv g \equiv 0$. It is not well posed if we can find a complex number s with $\operatorname{Re} s > 0$ and initial values $f(x)$ with $0 < \|f\| < \infty$ such that*

$$w(x, t) = e^{st}f(x), \quad \operatorname{Re} s > 0 \quad (10.1.2)$$

is a solution.

Proof. Assume that there is a solution of the above type. Define the sequence $\{f_j(x)\}_{j=1}^{\infty}$ by

$$f_j(x) = \frac{f(jx)}{\|f(jx)\|}, \quad \text{i.e.,} \quad \|f_j\| = 1.$$

Then

$$w_j(x, t) = e^{jst}f_j(x),$$

are also solutions satisfying

$$\|w_j(\cdot, t)\| = e^{j(\operatorname{Re} s)t}, \quad j = 1, 2, \dots$$

Therefore, the problem cannot be well-posed because we can construct solutions that grow arbitrarily fast. This proves the lemma.

REMARK. If Eq. (10.1.1) had contained lower order terms, then solutions w with $\operatorname{Re} s \leq \eta_0$ are permissible.

We will now give conditions guaranteeing that solutions of the type of Eq. (10.1.2) exist.

Theorem 10.1.1. *A solution of the type of Eq. (10.1.2) exists; that is, the initial-boundary-value problem is not well posed, if the eigenvalue problem*

$$s\varphi = A \frac{d\varphi}{dx}, \quad 0 \leq x < \infty, \quad (10.1.3a)$$

with boundary conditions

$$L^{II}\varphi^{II}(0) + L^I\varphi^I(0) = 0, \quad \|\varphi\|^2 < \infty, \quad (10.1.3b)$$

has an eigenvalue s with $\operatorname{Re} s > 0$.

The eigenvalue problem has an eigenvalue s with $\operatorname{Re} s > 0$ if, and only if, L^{II} is singular. Therefore, by our previous result, the initial-boundary-value problem is well posed if, and only if, the eigenvalue problem (10.1.3) has no eigenvalue s with $\operatorname{Re} s > 0$.

Proof. If there is an eigenvalue s with $\operatorname{Re} s > 0$, then

$$w(x, t) = e^{st}\varphi(x)$$

is a solution of the type of Eq. (10.1.2). Therefore, the problem is not well posed.

Let us now derive algebraic conditions such that there are no eigenvalues with $\operatorname{Re} s > 0$. These conditions are necessary conditions for the problem to be well posed. The general solution of Eq. (10.1.3a) can be written in the form

$$\varphi^I(x) = e^{s(\Lambda^I)^{-1}x}\varphi^I(0), \quad \varphi^{II}(x) = e^{s(\Lambda^{II})^{-1}x}\varphi^{II}(0).$$

For $\|\varphi\| < \infty$, a necessary and sufficient condition is that

$$\varphi^I(0) = 0. \quad (10.1.4)$$

Then the relations (10.1.3b) are satisfied if, and only if,

$$L^{II}\varphi^{II}(0) = 0. \quad (10.1.5)$$

There are two possibilities.

1. L^{II} is Nonsingular: Then the only solution of Eq. (10.1.5) is $\varphi^{II}(0) = 0$, and there is no eigenvalue s with $\operatorname{Re} s > 0$. In this case, we know, from our previous results, that the problem is well posed.
2. L^{II} is Singular: Then Eq. (10.1.5) has a nontrivial solution, and, therefore,

fore, there is an eigenvalue s with $\operatorname{Re} s > 0$. In fact, all s with $\operatorname{Re} s > 0$ are eigenvalues.

This proves the theorem.

Thus, for a well-posed problem, L^{II} must be nonsingular, and, therefore, we can write the boundary conditions in the form

$$u^{II}(0, t) = R^I u^I(0, t) + g^{II}(t); \quad (10.1.6)$$

that is, they are necessarily of the form we have discussed earlier.

We can now solve the problem using the Laplace transform. Without restriction, we can assume that $f(x) \equiv 0$. Otherwise, we introduce a new variable $\tilde{u} = u - h(t)f(x)$, $h(0) = 1$, h smooth with compact support. We define the Laplace transform by

$$\begin{aligned} \hat{u}(x, s) &= \int_0^\infty e^{-st} u(x, t) dt, \\ s &= i\xi + \eta, \quad \xi, \eta \text{ real}, \quad \eta > 0. \end{aligned} \quad (10.1.7)$$

REMARK. From Section 9.2, we know that the solution of Eq. (10.1.1) satisfies an energy estimate. Therefore, the right-hand side of Eq. (10.1.7) is finite for every $\eta > 0$.

By Eq. (10.1.1a)

$$\int_0^\infty e^{-st} u_t dt = A \int_0^\infty e^{-st} u_x dt + \int_0^\infty e^{-st} F dt = A\hat{u}_x + \hat{F}.$$

Therefore,

$$\int_0^\infty e^{-st} u_t dt = e^{-st} u|_0^\infty + s \int_0^\infty e^{-st} u dt$$

implies (observe that $f \equiv 0$)

$$s\hat{u}^I = \Lambda^I \hat{u}_x^I + \hat{F}^I, \quad s\hat{u}^{II} = \Lambda^{II} \hat{u}_x^{II} + \hat{F}^{II}. \quad (10.1.8a)$$

The boundary conditions are

$$\hat{u}^{II}(0, s) = R^I \hat{u}^I(0, s) + \hat{g}^{II}(s), \quad \|\hat{u}(\cdot, s)\| < \infty. \quad (10.1.8b)$$

Because the eigenvalue problem (10.1.3) only has a trivial solution, the solution \hat{u} of the homogeneous problem (10.1.8) is identically zero. Thus, the inhomogeneous problem (10.1.8) has a unique solution. We can write down the solution explicitly:

$$\begin{aligned}\hat{u}^I(x, s) &= - \int_{-\infty}^x e^{s(\Lambda^I)^{-1}(x-y)} (\Lambda^I)^{-1} \hat{F}^I(y, s) dy, \\ \hat{u}^{II}(x, s) &= \int_0^x e^{s(\Lambda^{II})^{-1}(x-y)} (\Lambda^{II})^{-1} \hat{F}^{II}(y, s) dy + e^{s(\Lambda^{II})^{-1}x} \hat{u}^{II}(0, s),\end{aligned}\quad (10.1.9)$$

where $\hat{u}^{II}(0, s)$ is determined by Eq. (10.1.8b).

Using Eq. (10.1.9), we can estimate $\hat{u}(x, s)$ in terms of \hat{F} , \hat{g}^{II} . However, it is easy to use energy estimates directly. We take the scalar product of Eq. (10.1.8a) with \hat{u}^I and \hat{u}^{II} , respectively, and obtain

$$(\hat{u}^I, s\hat{u}^I) + (s\hat{u}^I, \hat{u}^I) = (\hat{u}^I, \Lambda^I \hat{u}_x^I) + (\Lambda^I \hat{u}_x^I, \hat{u}^I) + (\hat{u}^I, \hat{F}^I) + (\hat{F}^I, \hat{u}^I),$$

or

$$\eta \|\hat{u}^I\|^2 = \operatorname{Re}(\hat{u}^I, \Lambda^I \hat{u}_x^I) + \operatorname{Re}(\hat{u}^I, \hat{F}^I).$$

Integration by parts gives us

$$\operatorname{Re}(\hat{u}^I, \Lambda^I \hat{u}_x^I) = -\frac{1}{2} \langle \hat{u}^I(0, s), \Lambda^I \hat{u}^I(0, s) \rangle,$$

and, therefore,

$$\eta \|\hat{u}^I\|^2 + \frac{1}{2} \langle \hat{u}^I(0, s), \Lambda^I \hat{u}^I(0, s) \rangle \leq \|\hat{u}^I\| \|\hat{F}^I\|;$$

that is,

$$\|\hat{u}^I\| \leq \frac{1}{\eta} \|\hat{F}^I\|, \quad |\hat{u}^I(0, s)| \leq \frac{C}{\eta^{1/2}} \|\hat{F}^I\|.$$

For \hat{u}^{II} , we obtain, correspondingly,

$$\begin{aligned}\eta \|\hat{u}^{II}\|^2 &\leq \|\hat{u}^{II}\| \|\hat{F}^{II}\| - \frac{1}{2} \langle \hat{u}^{II}(0, s), \Lambda^{II} \hat{u}^{II}(0, s) \rangle, \\ &\leq \frac{\eta}{2} \|\hat{u}^{II}\|^2 + \frac{1}{2\eta} \|\hat{F}^{II}\|^2 - \frac{1}{2} \langle \hat{u}^{II}(0, s), \Lambda^{II} \hat{u}^{II}(0, s) \rangle,\end{aligned}$$

or

$$\eta \|\hat{u}^{II}\|^2 \leq \text{constant} \left(\frac{1}{\eta} \|\hat{F}^{II}\|^2 + |\hat{u}^{II}(0, s)|^2 \right).$$

Using the boundary condition we obtain

$$|\hat{u}^{II}(0, s)|^2 \leq \text{constant} (|\hat{g}^{II}|^2 + |\hat{u}^I(0, s)|^2) \leq \text{constant} \left(|\hat{g}^{II}|^2 + \frac{1}{\eta} \|\hat{F}^I\|^2 \right).$$

Therefore,

$$\eta \|\hat{u}\|^2 \leq \text{constant} \left(\frac{1}{\eta} \|\hat{F}\|^2 + |\hat{g}^{II}|^2 \right), \quad (10.1.10a)$$

$$|\hat{u}(0, s)|^2 \leq \text{constant} \left(|\hat{g}^{II}|^2 + \frac{1}{\eta} \|\hat{F}\|^2 \right). \quad (10.1.10b)$$

Inverting the Laplace transform gives us the solution of our problem

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

and, by Parseval's relation, Eq. (A.2.17), we obtain, for any $\eta > 0$ and $s = i\xi + \eta$,

$$\begin{aligned}\int_0^{\infty} e^{-2\eta t} \|u(\cdot, t)\|^2 dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\hat{u}(\cdot, s)\|^2 d\xi, \\ &\leq \text{constant} \int_{-\infty}^{\infty} \left(\frac{1}{\eta^2} \|\hat{F}(\cdot, s)\|^2 + \frac{1}{\eta} |\hat{g}^{II}(s)|^2 \right) d\xi, \\ &= \text{constant} \int_0^{\infty} e^{-2\eta t} \left(\frac{1}{\eta^2} \|F(\cdot, t)\|^2 + \frac{1}{\eta} |g(t)|^2 \right) dt.\end{aligned} \quad (10.1.11a)$$

Correspondingly, Eq. (10.1.10b) is equivalent to

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} |u(0, t)|^2 dt \\ & \leq \text{constant} \int_0^\infty e^{-2\eta t} \left(\frac{1}{\eta} \|F(\cdot, t)\|^2 + |g(t)|^2 \right) dt. \end{aligned} \quad (10.1.11b)$$

Thus, we can estimate the solution in terms of the data. In Section 10.3, we use this estimate to define another concept of well-posedness.

EXERCISES

- 10.1.1.** Assume that $f(x) \neq 0$ in Eq. (10.1.1b). Derive the estimate (10.1.11b) by applying the Laplace transform technique to $\tilde{u} = u - h(t)f(x)$ as described above.

10.2. SOLUTION OF PARABOLIC PROBLEMS

We start with an example. Consider the quarter-space problem for a parabolic system

$$\begin{aligned} u_t &= \Lambda u_{xx} + F, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \quad (10.2.1)$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad \lambda_j = \text{constant} > 0, \quad j = 1, 2, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix},$$

with boundary conditions

$$\begin{aligned} u^{(1)}(0, t) &= \alpha u^{(2)}(0, t) + g^{(1)}(t), \\ u_x^{(2)}(0, t) &= \beta u_x^{(1)}(0, t) + g^{(2)}(t), \\ \|u(\cdot, t)\| &< \infty. \end{aligned} \quad (10.2.2)$$

where α and β are constants. Let us first investigate under what conditions on α and β we can obtain an energy estimate. Here we assume that $g^{(1)} = g^{(2)} = 0$. Integration by parts gives us

$$\frac{d}{dt} \|u\|^2 = -2(u_x, \Lambda u_x) - 2B + (u, F) + (F, u),$$

where, by Eq. (10.2.2),

$$B = \operatorname{Re} \langle u(0, t), \Lambda u_x(0, t) \rangle = \operatorname{Re} ((\lambda_1 \alpha + \lambda_2 \beta) u^{(2)}(0, t) u_x^{(1)}(0, t)).$$

Therefore, to obtain an energy estimate, we need $\lambda_1 \alpha + \lambda_2 \beta = 0$.

We now investigate the case $\lambda_1 \alpha + \lambda_2 \beta \neq 0$. As in the previous section, we can prove that the problem is not well posed if the eigenvalue problem corresponding to Eqs. (10.2.1) and (10.2.2)

$$s\varphi = \Lambda\varphi_{xx}, \quad 0 \leq x < \infty, \quad (10.2.3)$$

$$\varphi^{(1)}(0) = \alpha\varphi^{(2)}(0), \quad \varphi_x^{(2)}(0) = \beta\varphi_x^{(1)}(0), \quad \|\varphi\| < \infty, \quad (10.2.4)$$

has eigenvalues in the right half of the complex plane. In that case, there are solutions that grow arbitrarily fast.

For our example, we have the following theorem:

Theorem 10.2.1. *The problem [Eqs. (10.2.3) and (10.2.4)] has an eigenvalue s with $\operatorname{Re} s > 0$ if, and only if,*

$$\lambda_2^{-1/2} - \lambda_1^{-1/2}\alpha\beta = 0, \quad \lambda_j^{-1/2} > 0 \quad j = 1, 2. \quad (10.2.5)$$

Proof. The general solution of Eq. (10.2.3) with $\|\varphi\| < \infty$ for $\operatorname{Re} s > 0$ is given by

$$\begin{aligned} \varphi^{(j)} &= e^{-\lambda_j^{-1/2}s^{1/2}x} y^{(j)}, \quad \operatorname{Re} \lambda_j^{-1/2}s^{1/2} > 0, \\ y^{(j)} &= \text{constant}, \quad j = 1, 2. \end{aligned}$$

The boundary conditions are satisfied if, and only if,

$$\begin{bmatrix} 1 & -\alpha \\ -\beta\lambda_1^{-1/2} & \lambda_2^{-1/2} \end{bmatrix} y = 0, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}. \quad (10.2.6)$$

Equation (10.2.6) has a nontrivial solution if, and only if, $\lambda_2^{-1/2} - \lambda_1^{-1/2}\alpha\beta = 0$. This proves the theorem.

We now show that if $\lambda_2^{-1/2} - \lambda_1^{-1/2}\alpha\beta \neq 0$, then the problem has a unique solution that can be estimated in terms of the data. We assume again that $f \equiv 0$. The Laplace transform \hat{u} is the solution of

$$\begin{aligned} s\hat{u} &= \Lambda\hat{u}_{xx} + \hat{F}, \\ \hat{u}^{(1)}(0, s) &= \alpha\hat{u}^{(2)}(0, s) + \hat{g}^{(1)}(s), \\ \hat{u}_x^{(2)}(0, s) &= \beta\hat{u}_x^{(1)}(0, s) + \hat{g}^{(2)}(s), \quad \|\hat{u}\| < \infty. \end{aligned} \quad (10.2.7)$$

Because there are no eigenvalues with $\eta = \operatorname{Re} s > 0$, Eq. (10.2.7) has a unique solution for $\eta > 0$. Inverting the Laplace transform gives us the desired solution of the initial-boundary-value problem. We now estimate the solution in terms of F and g . We introduce into Eq. (10.2.7) a new variable

$$\hat{w} = s^{-1/2} \Lambda^{1/2} \hat{u}_x, \quad -\frac{\pi}{4} < \arg s^{1/2} \leq \frac{\pi}{4},$$

and obtain the system

$$s^{1/2} \mathbf{y} = H \mathbf{y}_x + s^{-1/2} \hat{\mathbf{F}}, \quad L \mathbf{y}(0, s) = \mathbf{g},$$

where

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix} \hat{u} \\ \hat{w} \end{bmatrix}, & \hat{\mathbf{F}} &= \begin{bmatrix} \hat{F} \\ 0 \end{bmatrix}, & H &= \begin{bmatrix} 0 & \Lambda^{1/2} \\ \Lambda^{1/2} & 0 \end{bmatrix}, \\ L &= \begin{bmatrix} 1 & -\alpha & 0 & 0 \\ 0 & 0 & -\beta \lambda_1^{-1/2} & \lambda_2^{-1/2} \end{bmatrix}, & \mathbf{g} &= \begin{bmatrix} \hat{g}^{(1)} \\ s^{-1/2} \hat{g}^{(2)} \end{bmatrix}. \end{aligned}$$

The unitary transformation

$$T = \frac{1}{\sqrt{2}} \begin{bmatrix} I & I \\ -I & I \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

transforms H to diagonal form. Therefore, we introduce new variables by $\tilde{\mathbf{y}} = T^* \mathbf{y}$, $\tilde{\mathbf{F}} = T^* \hat{\mathbf{F}}$ and obtain

$$\begin{aligned} s^{1/2} \tilde{\mathbf{y}}^I &= -\Lambda^{1/2} \tilde{\mathbf{y}}_x^I + (2s)^{-1/2} \tilde{\mathbf{F}}^I, \\ s^{1/2} \tilde{\mathbf{y}}^{II} &= \Lambda^{1/2} \tilde{\mathbf{y}}_x^{II} + (2s)^{-1/2} \tilde{\mathbf{F}}^{II}, \\ D_1 \tilde{\mathbf{y}}^I(0, s) + D_2 \tilde{\mathbf{y}}^{II}(0, s) &= \tilde{g}(s), \quad \|\tilde{\mathbf{y}}\| < \infty, \end{aligned} \quad (10.2.8)$$

where $\|\tilde{\mathbf{F}}^I\| \leq \text{constant } \|\hat{\mathbf{F}}\|$, $\|\tilde{\mathbf{F}}^{II}\| \leq \text{constant } \|\hat{\mathbf{F}}\|$, and $|\tilde{g}| \leq \text{constant } (|\hat{g}^{(1)}| + |s^{-1/2}| |\hat{g}^{(2)}|)$.

We need the lemma below.

Lemma 10.2.1. *D_1 is nonsingular if, and only if, $\lambda_2^{-1/2} - \lambda_1^{-1/2} \alpha \beta \neq 0$.*

Proof. Consider the homogeneous equation (10.2.8). For $\operatorname{Re} s > 0$, the general solution belonging to L_2 is given by

$$\tilde{y}^I = e^{-s^{1/2}\Lambda^{-1/2}x}\tilde{y}^I(0), \quad \tilde{y}^{II} = 0.$$

Therefore, the homogeneous problem has a nontrivial solution if, and only if,

$$D_1\tilde{y}^I(0, s) = 0$$

has a nontrivial solution. Inverting the transformations above, any nontrivial solution generates a nontrivial solution $e^{s'x}\varphi(x)$ and vice versa. The lemma then follows from Theorem 10.2.1.

Assume that D_1 is nonsingular. We can proceed as in the previous section. We first estimate \tilde{y}^{II} . Integration by parts gives us

$$\begin{aligned} \operatorname{Re} s^{1/2} \|\tilde{y}^{II}\|^2 &= \operatorname{Re}(\tilde{y}^{II}, \Lambda^{1/2}\tilde{y}_x^{II}) + \operatorname{Re}\left(\frac{1}{(2s)^{1/2}} (\tilde{y}^{II}, \tilde{F}^{II})\right), \\ &= -\frac{1}{2} \langle \tilde{y}^{II}(0, s), \Lambda^{1/2}\tilde{y}^{II}(0, s) \rangle + \operatorname{Re}\left(\frac{1}{(2s)^{1/2}} (\tilde{y}^{II}, \tilde{F}^{II})\right). \end{aligned}$$

Therefore, observing that $\sqrt{2} \operatorname{Re} s^{1/2} \geq |s|^{1/2}$,

$$|s^{1/2}| \|\tilde{y}^{II}\|^2 + \frac{1}{2} |\Lambda^{-1/2}|^{-1} |\tilde{y}^{II}(0, s)|^2 \leq \frac{\text{constant}}{|s|^{3/2}} \|\tilde{F}^{II}\|^2,$$

that is,

$$\|\tilde{y}^{II}\|^2 \leq \frac{\text{constant}}{|s|^2} \|\tilde{F}^{II}\|^2, \quad |\tilde{y}^{II}(0, s)|^2 \leq \frac{\text{constant}}{|s|^{3/2}} \|\tilde{F}^{II}\|^2.$$

Because D_1 is nonsingular and D_1 and D_2 are independent of s , the boundary conditions give us

$$\begin{aligned} |\tilde{y}^I(0, s)|^2 &\leq \text{constant} (|\tilde{y}^{II}(0, s)|^2 + |\tilde{g}(s)|^2) \\ &\leq \text{constant} \left(\frac{1}{|s|^{3/2}} \|\tilde{F}^{II}\|^2 + |\tilde{g}(s)|^2 \right). \end{aligned}$$

Using integration by parts we obtain

$$\begin{aligned}\operatorname{Re} s^{1/2} \|\tilde{y}^I\|^2 &= \operatorname{Re}(-\tilde{y}^I, \Lambda^{1/2} \tilde{y}_x^I) + \operatorname{Re}\left(\frac{1}{(2s)^{1/2}} (\tilde{y}^I, \tilde{F}^I)\right) \\ &= \frac{1}{2} \langle \tilde{y}^I(0, s), \Lambda^{1/2} \tilde{y}^I(0, s) \rangle + \operatorname{Re}\left(\frac{1}{(2s)^{1/2}} (\tilde{y}^I, \tilde{F}^I)\right).\end{aligned}$$

Therefore,

$$|s^{1/2}| \|\tilde{y}^I\|^2 \leq \text{constant} \left(\frac{1}{|s|^{3/2}} \|\tilde{F}\|^2 + |\tilde{g}(s)|^2 \right).$$

Inverting all of the transformations we obtain

$$\begin{aligned}\|\hat{u}\|^2 &\leq \text{constant} \left(\frac{1}{|s|^2} \|\hat{F}\|^2 + \frac{1}{|s|^{1/2}} |\mathbf{g}|^2 \right), \\ \|\hat{u}_x\|^2 &\leq \text{constant} \left(\frac{1}{|s|} \|\hat{F}\|^2 + |s|^{1/2} |\mathbf{g}|^2 \right), \\ |\hat{u}(0, s)|^2 &\leq \text{constant} \left(\frac{1}{|s|^{3/2}} \|\hat{F}\|^2 + |\mathbf{g}|^2 \right). \quad (10.2.9)\end{aligned}$$

If $g \equiv 0$, then, from Parseval's relation, we obtain the estimate

$$\begin{aligned}\int_0^\infty e^{-2\eta t} (\|u(\cdot, t)\|^2 + \|u_x(\cdot, t)\|^2) dt \\ \leq \text{constant} \left(\frac{1}{\eta} + \frac{1}{\eta^2} \right) \int_0^\infty e^{-2\eta t} \|F(\cdot, t)\|^2 dt. \quad (10.2.10a)\end{aligned}$$

For $g \neq 0$, we can only estimate u and obtain

$$\begin{aligned}\int_0^\infty e^{-2\eta t} \|u(\cdot, t)\|^2 dt \\ \leq \text{constant} \int_0^\infty e^{-2\eta t} \left(\frac{1}{\eta^2} \|F(\cdot, t)\|^2 + \frac{1}{\eta^{1/2}} |g^{(1)}(t)|^2 \right. \\ \left. + \frac{1}{\eta^{3/2}} |g^{(2)}(t)|^2 \right) dt. \quad (10.2.10b)\end{aligned}$$

We can also estimate $u(0, t)$

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} |u(0, t)|^2 dt \\ & \leq \text{constant} \int_0^\infty e^{-2\eta t} \left(\frac{1}{\eta^{3/2}} \|F(\cdot, t)\|^2 \right. \\ & \quad \left. + |g^{(1)}(t)|^2 + \frac{1}{\eta} |g^{(2)}(t)|^2 \right) dt. \end{aligned} \quad (10.2.10c)$$

We use this estimate to define a new concept of well-posedness in the next section.

We have shown that good estimates are obtained for all α and β with

$$\alpha\beta \neq \left(\frac{\lambda_1}{\lambda_2} \right)^{1/2},$$

whereas an energy estimate requires

$$\frac{\beta}{\alpha} = -\frac{\lambda_1}{\lambda_2}.$$

Therefore, this example shows that the conditions for the existence of energy estimates can be too restrictive.

One can generalize the above results considerably. Consider the quarter-space problem for a parabolic system

$$u_t = Au_{xx} + Bu_x + Cu + F, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (10.2.11a)$$

$$u(x, 0) = 0, \quad (10.2.11b)$$

with m linearly independent boundary conditions

$$R_{11}u_x(0, t) + R_{10}u(0, t) = 0, \quad R_{00}u(0, t) = 0. \quad (10.2.11c)$$

The principle part of this problem is obtained by neglecting the lower order terms in the differential equation and in the boundary conditions. The corresponding eigenvalue problem is

$$\begin{aligned} s\varphi &= A\varphi_{xx}, \quad 0 \leq x < \infty, \\ R_{11}\varphi_x(0, s) &= 0, \\ R_{00}\varphi(0, s) &= 0, \\ \|\varphi\| &< \infty. \end{aligned} \quad (10.2.12)$$

One can prove the following theorem (Exercise 10.2.2).

Theorem 10.2.2. *The problem (10.2.11) has a unique solution satisfying an estimate of type (10.2.10), for $\eta > \eta_0$, if, and only if, Eq. (10.2.12) has no eigenvalue s with $\operatorname{Re} s > 0$.*

This theorem is valid for much more general boundary conditions

$$\sum_{j=0}^p B_{1j} \frac{\partial^{j+1} u}{\partial t^j \partial x} + B_{0j} \frac{\partial^j u}{\partial t^j} = 0, \quad x = 0. \quad (10.2.13)$$

EXERCISES

10.2.1. Formulate and prove the analogy to Lemma 10.1.1 for parabolic systems

$$u_t = Au_{xx}.$$

10.2.2. Prove Theorem 10.2.2.

10.3. GENERALIZED WELL-POSEDNESS

We again consider the system of partial differential equations

$$\frac{\partial u}{\partial t} = Pu + F, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (10.3.1a)$$

with initial data

$$u(x, 0) = f(x), \quad (10.3.1b)$$

and boundary conditions

$$L_0 u(0, t) = g_0(t), \quad L_1 u(1, t) = g_1(t). \quad (10.3.1c)$$

We assume that the system of differential equations is either symmetric hyperbolic, that is,

$$P = A \frac{\partial}{\partial x} + B, \quad A = A^*,$$

or second-order parabolic; that is,

$$P = A \frac{\partial^2}{\partial x^2} + B \frac{\partial}{\partial x} + C, \quad A + A^* \geq 2a_0 I > 0.$$

For simplicity, we assume that A , B , and C and the coefficients of L_0 and L_1 are smooth functions of x and t . We also assume that F , f , g_0 , and g_1 , are smooth bounded functions. To guarantee that the compatibility conditions are satisfied, we assume that F vanishes near the boundary and the initial line and that $f(x)$, $g_0(t)$, and $g_1(t)$ vanish near $x = 0, 1$ and $t = 0$, respectively. Of course, less stringent conditions need be met if only solutions with a fixed number of derivatives are of interest. As before, one can also extend the solution concept to include generalized solutions.

In Section 9.4, we have discussed two different definitions of well-posedness. They differ with respect to the estimate required. All bounds are natural in the context of energy estimates. Unfortunately, energy estimates are not available in many circumstances, and, therefore, other techniques must be used. A very powerful tool is the Laplace transform, which we used in the last two sections.

When using the Laplace transform, it is convenient to assume that $f(x) \equiv 0$. In Section 10.1, how to transform the problem so that this condition is satisfied was explained.

To begin, we assume that the system (10.3.1) has constant coefficients. As in previous sections, we introduce the Laplace transform

$$\hat{u}(x, s) = \int_0^\infty e^{-st} u(x, t) dt, \quad s = i\xi + \eta, \quad \xi, \eta \text{ real}, \quad \eta > 0.$$

It is the solution of the *resolvent equation*

$$(sI - P)\hat{u} = \hat{F}, \\ L_0\hat{u}(0, s) = \hat{g}_0, \quad L_1\hat{u}(1, s) = \hat{g}_1. \quad (10.3.2)$$

Typical estimates for the solutions of Eqs. (10.3.2) are listed below [see Eqs. (10.1.10) and (10.2.9)]:

1. Consider Eqs. (10.3.2) with homogeneous boundary conditions ($\hat{g}_j \equiv 0$). There is a constant η_0 and a function $K(\eta)$, with $\lim_{\eta \rightarrow \infty} K(\eta) = 0$, such that, for all \hat{F} and all s with $\operatorname{Re} s > \eta_0$,

$$\|\hat{u}(\cdot, s)\|^2 + \delta \|\hat{u}_x(\cdot, s)\|^2 \leq K(\eta) \|\hat{F}(\cdot, s)\|^2. \quad (10.3.3a)$$

Here $\delta = 0$ for hyperbolic systems and $\delta > 0$ for parabolic systems.

2. Instead of Eq. (10.3.3a), for inhomogeneous boundary conditions we have

$$\|\hat{u}(\cdot, s)\|^2 + \delta \|\hat{u}_x(\cdot, s)\|^2 \leq K(\eta) (\|\hat{F}(\cdot, s)\|^2 + |\hat{g}_0(s)|^2 + |\hat{g}_1(s)|^2). \quad (10.3.4a)$$

(This estimate is stronger than the one derived for the parabolic problem in Section 10.2.)

If the differential operator P is defined on the function space satisfying the homogeneous boundary conditions $L_0 v(0) = L_1 v(1) = 0$, the operator $(sI - P)^{-1}$ is called the *resolvent operator*, and the *resolvent condition* is usually formulated as

$$\|(sI - P)^{-1}\| \leq \frac{\text{constant}}{\operatorname{Re} s}.$$

Our conditions (10.3.3a) and (10.3.4a) are generalized forms of the resolvent condition.

By Parseval's relation, these inequalities imply the following estimates:

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} (\|u(\cdot, t)\|^2 + \delta \|u_x(\cdot, t)\|^2) dt \\ & \leq K(\eta) \int_0^\infty e^{-2\eta t} \|F(\cdot, t)\|^2 dt, \quad \eta > \eta_0, \quad \lim_{\eta \rightarrow \infty} K(\eta) = 0, \end{aligned} \quad (10.3.3b)$$

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} (\|u(\cdot, t)\|^2 + \delta \|u_x(\cdot, t)\|^2) dt \\ & \leq K(\eta) \int_0^\infty e^{-2\eta t} (\|F(\cdot, t)\|^2 + |g_0(t)|^2 + |g_1(t)|^2) dt, \\ & \eta > \eta_0, \quad \lim_{\eta \rightarrow \infty} K(\eta) = 0, \end{aligned} \quad (10.3.4b)$$

respectively.

For the examples treated above we have $\eta_0 = 0$. However, if the problem has variable coefficients, lower order terms, or two boundaries, then we have to choose $\eta_0 > 0$. The last two of these generalizations will be discussed later.

Because the integrals are taken over an infinite time interval, the constant $K(\eta)$ necessarily goes to infinity at some point $\eta = \eta_0$. For example, if $F = g_0 = g_1 = 0$ for $t \geq T$, then the integrals on the right-hand side of the estimates are finite for any value of y . However, the solution u is in general nonzero for

all t , and the integrals on the left-hand side of the estimates do not exist for $y \leq y_0$. Therefore, $K(\eta) \rightarrow \infty$ as $\eta \rightarrow y_0$.

We now use these estimates to introduce a new concept of well-posedness for the systems (10.3.1) with variable coefficients.

Definition 10.3.1. Consider the problem (10.3.1) with $f \equiv g_0 \equiv g_1 \equiv 0$. We call the problem well posed in the generalized sense if, for any smooth compatible F , there is a smooth solution that satisfies the estimate (10.3.3b) for all $\eta > \eta_0$. Here

$$\begin{aligned}\delta &= 0 && \text{for hyperbolic problems,} \\ \delta &> 0 && \text{for parabolic problems,}\end{aligned}$$

and η_0 , $K(\eta)$ are constants that do not depend on F . We call the problem strongly well posed in the generalized sense if the estimate (10.3.4b) holds.

REMARK. The initial and boundary conditions can be made homogeneous by subtracting a suitable function $\psi(x, t)$ that satisfies Eqs. (10.3.1b) and (10.3.1c). The function $v = u - \psi$ satisfies

$$\begin{aligned}\frac{\partial v}{\partial t} &= Pv + \tilde{F}, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ v(x, 0) &= 0, \\ L_0v(0, t) &= 0, \quad L_1v(1, t) = 0,\end{aligned}$$

where $\tilde{F} = P\psi - \partial\psi/\partial t + F$. Assuming that the problem is well posed in the generalized sense, we obtain the estimate (10.3.3b) as $u \rightarrow v$, $F \rightarrow \tilde{F}$. Hence, we get an estimate for $u = v + \psi$, but the bound depends on dg_0/dt , dg_1/dt , df/dx , and also on d^2f/dx^2 in the parabolic case.

There are many other ways to define well-posedness. Any definition of well-posedness must satisfy the following requirements.

1. For smooth compatible data, the problem has a smooth solution.
2. There is an estimate of the solution in terms of the data.
3. It should be stable against perturbations of lower order terms; that is, if we perturb the differential equations by changing the lower order terms, then the solutions of the perturbed problem should also satisfy an estimate of the same type.
4. It should hold for a large class of problems.
5. There should be equivalent algebraic conditions that are easy to verify.

The definitions of Section 9.4 fulfill these requirements. The necessary esti-

mates can be verified for symmetric first-order systems and for parabolic systems by the use of integration by parts. However, there are large classes of problems that cannot be investigated in this way. By using Definition 10.3.1, we can cover a much wider class of problems.

The concept of strong well-posedness in the generalized sense does not play the same role for parabolic equations as for hyperbolic equations. It holds for parabolic equations only if all boundary conditions are derivative conditions, that is, $\text{rank}(R_{11}) = m$ in Eq. (10.2.11c).

We now derive a number of fundamental properties for our new concepts.

Theorem 10.3.1. *Assume that the differential operator P is semibounded with the boundary conditions (10.3.1c) such that*

$$\operatorname{Re}(v, Pv) \leq -\delta \|v_x\|^2 + \alpha \|v\|^2,$$

where $\delta > 0$ if P is parabolic and $\delta = 0$ if P is hyperbolic. Then, the problem (10.3.1) is well posed in the generalized sense.

Proof. Introduce into Eq. (10.3.1) a new variable, $w = e^{-\eta t}u$. Then, we obtain

$$\frac{\partial w}{\partial t} = (P - \eta I)w + e^{-\eta t}F.$$

Therefore, for $\tilde{\eta} = \eta - \alpha \geq \tilde{\eta}_0 > 0$,

$$\begin{aligned} \frac{d}{dt} \|w\|^2 &\leq -2\delta \|w_x\|^2 - 2(\eta - \alpha)\|w\|^2 + 2|(w, e^{-\eta t}F)|, \\ &\leq -2\delta \|w_x\|^2 - \tilde{\eta}\|w\|^2 + \frac{1}{\tilde{\eta}} \|e^{-\eta t}F\|^2. \end{aligned}$$

Thus,

$$\|w(\cdot, t)\|^2 \leq \int_0^t e^{-\tilde{\eta}(t-\tau)}(G(\tau) - H(\tau))\tau,$$

where

$$G(\tau) = \frac{1}{\tilde{\eta}} \|e^{-\eta\tau}F(\cdot, \tau)\|^2, \quad H(\tau) = 2\delta \|w_x(\cdot, \tau)\|^2.$$

Define $\varphi(t)$ by

$$\varphi(t) = \begin{cases} e^{-\tilde{\eta}t}, & \text{for } t \geq 0, \\ 0, & \text{for } t < 0. \end{cases}$$

Then, we obtain

$$\begin{aligned} \int_0^\infty \|w(\cdot, t)\|^2 dt &\leq \int_0^\infty \left(\int_0^\infty \varphi(t - \tau) d\tau \right) G(\tau) d\tau \\ &\quad - \int_0^\infty \left(\int_0^\infty \varphi(t - \tau) d\tau \right) H(\tau) d\tau \\ &\leq \frac{1}{\tilde{\eta}^2} \int_0^\infty e^{-2\eta t} \|F(\cdot, t)\|^2 dt - \frac{2\delta}{\tilde{\eta}} \int_0^\infty \|w_x(\cdot, t)\|^2 dt; \end{aligned}$$

that is,

$$\tilde{\eta} \int_0^\infty \|w(\cdot, t)\|^2 dt + 2\delta \int_0^\infty \|w_x(\cdot, t)\|^2 dt \leq \frac{1}{\tilde{\eta}} \int_0^\infty e^{-2\eta t} \|F(\cdot, t)\|^2 dt.$$

Thus, Eq. (10.3.3b) is satisfied with

$$K(\eta) = \max \left(\frac{1}{(y - \alpha)^2}, \frac{1}{y - \alpha} \right)$$

and with δ replaced by $2\delta/\tilde{\eta}_0$.

The existence of a smooth solution can also be verified and, therefore, the theorem is proved.

Now we will prove that lower order terms do not affect generalized well-posedness.

Theorem 10.3.2. *Assume that the problem (10.3.1) is well posed in the generalized sense. Then the perturbed problem*

$$\begin{aligned} \partial w / \partial t &= (P + P_0)w + F, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ w(x, 0) &= 0, \\ L_0 w(0, t) &= 0, \quad L_1 w(1, t) = 0, \end{aligned}$$

has the same property. For hyperbolic equations $P_0 w$ is a zero order term, for parabolic equations it may also contain derivatives of first order.

Proof. Formally $P_0 w$ can be considered as a forcing function, and by assumption we have

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} (\|w(\cdot, t)\|^2 + \delta \|w_x(\cdot, t)\|^2) dt \\ & \leq K(\eta) \int_0^\infty e^{-2\eta t} (\|P_0 w(\cdot, t)\|^2 + \|F(\cdot, t)\|^2) dt. \end{aligned}$$

By choosing η sufficiently large and, therefore, $K(\eta)$ sufficiently small and recalling that $\delta > 0$ for parabolic problems, we can move the $P_0 w$ term to the left-hand side giving us

$$\begin{aligned} & \int_0^\infty e^{-2\eta t} (\|w(\cdot, t)\|^2 + \delta \|w_x(\cdot, t)\|^2) dt \\ & \leq K_1(\eta) \int_0^\infty e^{-2\eta t} \|F(\cdot, t)\|^2 dt, \quad \eta > \eta_1, \quad \lim_{\eta \rightarrow \infty} K_1(\eta) = 0. \end{aligned}$$

The existence of a solution can also be assured and the theorem is proved.

As we will see later, it is convenient to treat each boundary by itself. For that purpose, we define the two quarter-space problems

$$\begin{aligned} \partial u / \partial t &= P u + F, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= 0, \\ L_0 u(0, t) &= 0, \end{aligned} \tag{10.3.5}$$

$$\begin{aligned} \partial u / \partial t &= P u + F, \quad -\infty < x \leq 1, \quad t \geq 0, \\ u(x, 0) &= 0, \\ L_1 u(1, t) &= 0, \end{aligned} \tag{10.3.6}$$

and the Cauchy problem

$$\begin{aligned} \partial u / \partial t &= P u + F, \quad -\infty < x < \infty, \quad t \geq 0, \\ u(x, 0) &= 0. \end{aligned} \tag{10.3.7}$$

The coefficient matrices in P are extended in a smooth way to the whole x axis so that they are constant for large $|x|$. We assume that the functions F have compact support in all three cases.

Definition 10.3.1 of well-posedness in the generalized sense is now the same for each one of these three problems, but with the norms defined by

$$\begin{aligned}
 \|u(\cdot, t)\|^2 &= \|u(\cdot, t)\|_{0,\infty}^2 := \int_0^\infty |u(x, t)|^2 dx, && \text{for Eq. (10.3.5),} \\
 &= \|u(\cdot, t)\|_{-\infty, 1}^2 := \int_{-\infty}^1 |u(x, t)|^2 dx, && \text{for Eq. (10.3.6),} \\
 &= \|u(\cdot, t)\|_{-\infty, \infty}^2 := \int_{-\infty}^\infty |u(x, t)|^2 dx, && \text{for Eq. (10.3.7).}
 \end{aligned}$$

We can now prove the following theorem.

Theorem 10.3.3. *The problem (10.3.1) is well posed in the generalized sense if the quarter-space problems (10.3.5) and (10.3.6) and the Cauchy problem (10.3.7) are all well posed in the generalized sense.*

Proof. Let $\varphi_1(x) \in C^\infty(-\infty, \infty)$ be a monotone function with

$$\varphi_1(x) = \begin{cases} 1, & \text{for } x \leq 1/8, \\ 0, & \text{for } x \geq 1/4, \end{cases}$$

and define

$$\begin{aligned}
 \varphi_2(x) &= \varphi_1(1 - x), \\
 \varphi_3(x) &= 1 - \varphi_1(x) - \varphi_2(x).
 \end{aligned}$$

Let

$$\begin{aligned}
 u_j(x, t) &= \varphi_j(x)u(x, t), \quad j = 1, 2, 3, \\
 F_j(x, t) &= \varphi_j(x)F(x, t), \quad j = 1, 2, 3,
 \end{aligned}$$

and define $u_1 \equiv 0$ for $x \geq 1$, $u_2 \equiv 0$ for $x \leq 0$, and $u_3 = 0$ for $x \leq 0, x \geq 1$, and, correspondingly, define $F_j(x, t)$. Multiplying Eqs. (10.3.1) by φ_j where $j = 1, 2, 3$, the problem with homogeneous initial and boundary conditions can be written in the form

$$\begin{aligned}
 (u_1)_t &= Pu_1 + P_1u + F_1, \quad 0 \leq x < \infty, \quad t \geq 0, \\
 u_1(x, 0) &= 0, \\
 L_0u_1(0, t) &= 0,
 \end{aligned} \tag{10.3.8a}$$

$$\begin{aligned}
 (u_2)_t &= Pu_2 + P_2u + F_2, \quad -\infty < x \leq 1, \quad t \geq 0, \\
 u_2(x, 0) &= 0, \\
 L_1u_2(1, t) &= 0,
 \end{aligned} \tag{10.3.8b}$$

$$(u_3)_t = Pu_3 + P_3u + F_3, \quad -\infty < x < \infty, \quad t \geq 0, \\ u_3(x, 0) = 0. \quad (10.3.8c)$$

Here $\{P_j\}_1^3$ are bounded matrices in the hyperbolic case. For parabolic equations, $\{P_j u\}_1^3$ are linear combinations of u and its first derivatives.

Now assume that the two quarter-space problems and the Cauchy problem are well posed in the generalized sense. The solutions of Eq. (10.3.8) satisfy

$$\int_0^\infty e^{-2\eta t} (\|u_1\|_{0,\infty}^2 + \delta \|u_{1x}\|_{0,\infty}^2) dt \\ \leq K_1(\eta) \int_0^\infty e^{-2\eta t} (\|F_1\|_{0,1}^2 + \|u\|_{0,1}^2 + \delta \|u_x\|_{0,1}^2) dt, \\ \eta > \eta_1, \quad (10.3.9a)$$

$$\int_0^\infty e^{-2\eta t} (\|u_2\|_{-\infty,1}^2 + \delta \|u_{2x}\|_{-\infty,1}^2) dt \\ \leq K_2(\eta) \int_0^\infty e^{-2\eta t} (\|F_2\|_{0,1}^2 + \|u\|_{0,1}^2 + \delta \|u_x\|_{0,1}^2) dt, \\ \eta > \eta_2, \quad (10.3.9b)$$

$$\int_{-\infty}^\infty e^{-2\eta t} (\|u_3\|_{-\infty,\infty}^2 + \delta \|u_{3x}\|_{-\infty,\infty}^2) dt \\ \leq K_3(\eta) \int_{-\infty}^\infty e^{-2\eta t} (\|F_3\|_{0,1}^2 + \|u\|_{0,1}^2 + \delta \|u_x\|_{0,1}^2) dt, \\ \eta > \eta_3, \quad (10.3.9c)$$

Here we have used the fact that the functions u_j were smoothly extended and the coefficients of P_j vanish outside the intervals $1/8 \leq x \leq 1/4$, $3/4 \leq x \leq 7/8$. The inequalities are added and η is chosen large enough so that the $\|u\|_{0,1}^2$ and $\|u_x\|_{0,1}^2$ terms can be moved to the left-hand side. Observing that $u = u_1 + u_2 + u_3$ for $0 \leq x \leq 1$, we obtain

$$\int_0^\infty e^{-2\eta t} (\|u\|_{0,1}^2 + \delta \|u_x\|_{0,1}^2) dt \\ \leq \text{constant} \int_0^\infty e^{-2\eta t} \sum_{j=1}^3 (\|u_j\|_{0,1}^2 + \delta \|u_{jx}\|_{0,1}^2) dt \\ \leq K_4(\eta) \int_0^\infty e^{-2\eta t} \sum_{j=1}^3 \|F_j\|_{0,1}^2 dt \\ \leq K_5(\eta) \int_0^\infty e^{-2\eta t} \|F\|_{0,1}^2 dt, \quad \eta > \eta_4,$$

where $\lim_{\eta \rightarrow \infty} K_5(\eta) = 0$.

One can also use the above representation to prove the existence of solutions of the original problem. By construction, $u = u_1 + u_2 + u_3$ is a solution of the original problem for $x \in [0, 1]$. This proves the theorem.

There are no difficulties in generalizing Definition 10.3.1 to several space dimensions. As in Section 9.6, we consider the differential equation in some domain Ω bounded by a smooth curve $\partial\Omega$. The norms $\|\cdot\|$, $|\cdot|$ in Eq. (10.3.3) now represent the L_2 norm over Ω and $\partial\Omega$, respectively. As we show later in Section 10.6, we can split such a problem into a Cauchy problem and a quarter-space problem.

10.4. SYSTEMS WITH CONSTANT COEFFICIENTS IN ONE SPACE DIMENSION

In this section, we consider systems (10.3.1) with constant coefficients. We derive algebraic conditions such that the initial boundary value problem is well posed in the generalized sense.

As in Sections 10.1 and 10.2, we can derive a necessary condition for well-posedness.

Theorem 10.4.1. *The problem is not well posed if the eigenvalue problem*

$$(P - sI)\varphi = 0, \\ L_0 \varphi(0) = L_1 \varphi(1) = 0, \quad (10.4.1)$$

has a sequence of eigenvalues s_j , $j = 1, 2, \dots$, with

$$\lim_{j \rightarrow \infty} \operatorname{Re} s_j = \infty. \quad (10.4.2)$$

Proof. Assume that there is such a sequence. Denote the corresponding eigenfunctions by $\varphi_j(x)$ with $\|\varphi_j(\cdot)\| = 1$. Then

$$u_j(x, t) = e^{s_j t} \varphi_j(x) \quad (10.4.3)$$

are solutions of Eq. (10.3.1) where

$$u_j(x, 0) = \varphi_j(x), \quad g_0 \equiv g_1 \equiv F \equiv 0.$$

Corresponding to Lemma 10.1.1, the relation

$$\frac{\|u_j(\cdot, t)\|}{\|u_j(\cdot, 0)\|} = e^{\operatorname{Re} s_j t}$$

tells us that the problem cannot be well posed. This proves the theorem.

REMARK. In Lemma 10.1.1, we found that if there is a fixed eigenvalue s with $\operatorname{Re} s > 0$, then the problem (10.1.1) is not well posed. The reason for this weaker condition is that the problem (10.1.1) has no lower order terms in the differential equation or in the boundary conditions. Indeed, if such an eigenvalue exists, then it was demonstrated that there is a sequence of eigenvalues satisfying Eq. (10.4.2). If lower order terms are present, then there might be a fixed eigenvalue in the right half-plane, and yet the problem may be well posed.

The system (10.4.1) has constant coefficients, and, therefore, we can, at least in principle, solve the eigenvalue problem explicitly. However, we shall not do this because, by Section 10.3, we can simplify our task considerably. We can neglect all lower order terms and replace the strip problem by quarter-space problems. Because the substitution $x' = 1 - x$ transforms the left quarter-space problem into a right quarter-space problem, we now only consider the problem

$$\frac{\partial u}{\partial t} = Pu + F, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (10.4.4a)$$

$$u(x, 0) = f(x), \quad (10.4.4b)$$

$$L_0 u(0, t) = g_0, \quad \|u(\cdot, t)\| < \infty, \quad (10.4.4c)$$

where

$$P = A \frac{\partial}{\partial x} \quad \text{or} \quad P = A \frac{\partial^2}{\partial x^2}.$$

By Sections 10.1 and 10.2, the eigenvalue condition becomes particularly simple if the boundary conditions are of the form of Eq. (10.1.1c) or Eq. (10.2.11c). We now have the following theorem.

Theorem 10.4.2. *Assume that the boundary conditions have the form of Eq. (10.1.1c) if $P = A\partial/\partial x$ or Eq. (10.2.11c) if $P = A\partial^2/\partial x^2$. Then the quarter-space problem (10.4.4) is not well posed if the eigenvalue problems (10.1.3) or (10.2.12) for the principal part have an eigenvalue s_0 with $\operatorname{Re} s_0 > 0$.*

Now we shall derive necessary and sufficient conditions for well-posedness for the general problem (10.4.4). As in the previous sections, we assume that $f \equiv 0$, and we Laplace transform Eq. (10.4.4) and obtain

$$(sI - P)\hat{u} = \hat{F},$$

$$L_0 \hat{u}(0, s) = \hat{g}_0, \quad \|\hat{u}(\cdot, s)\| < \infty. \quad (10.4.5)$$

We now can prove Theorem 10.4.3.

Theorem 10.4.3. *The problem (10.4.4) is well posed in the generalized sense if, and only if, Eq. (10.4.5), with $\hat{g}_0 = 0$, has a unique solution for all \hat{F} and all $s = i\xi + \eta$, $\eta > \eta_0$ such that Eq. (10.3.3a) holds.*

Proof. If the problem is well posed, then the Laplace transform \hat{u} is well defined and is a solution of Eq. (10.4.5) with $\hat{g}_0 = 0$. By Parseval's relation, it follows from Eq. (10.3.3b) that

$$\begin{aligned} & \int_{-\infty}^{\infty} (\|\hat{u}(\cdot, i\xi + \eta)\|^2 + \delta \|\hat{u}_x(\cdot, i\xi + \eta)\|^2) d\xi \\ & \leq K(\eta) \int_{-\infty}^{\infty} \|\hat{F}(\cdot, i\xi + \eta)\|^2 d\xi. \end{aligned} \quad (10.4.6)$$

Next we prove that this inequality implies the corresponding "pointwise" inequality (10.3.3a). Assume the contrary; that is, there is some point $s = s_1 = i\xi_1 + \eta_1$, and a function $F_1(x)$ with $\|F_1(x)\| < \infty$ such that the solution of

$$\begin{aligned} (s_1 I - P)\hat{w} &= F_1, \\ L_0\hat{w} &= 0, \quad \|\hat{w}\| < \infty, \end{aligned}$$

satisfies

$$\|\hat{w}(\cdot, \eta_1 + i\xi_1)\|^2 + \delta \|\hat{w}_x(\cdot, \eta_1 + i\xi_1)\|^2 > K(\eta_1) \|F_1(\cdot)\|^2. \quad (10.4.7)$$

Let

$$F(x, t) = \begin{cases} \frac{1}{\sqrt{T}} e^{\eta_1 t + i\xi_1 t} F_1(x), & t \leq T, \\ 0, & t > T. \end{cases} \quad (10.4.8)$$

Then, for $s = \eta_1 + i\xi$,

$$\begin{aligned} \hat{F}(x, s) &= \int_0^T e^{-st} F(x, t) dt = \frac{1}{\sqrt{T}} F_1(x) \int_0^T e^{i(\xi_1 - \xi)t} dt, \\ &= \frac{1}{\sqrt{T}} F_1(x) \frac{e^{i(\xi_1 - \xi)T} - 1}{i(\xi_1 - \xi)} = F_1(x) \cdot \varphi((\xi_1 - \xi)T), \end{aligned}$$

where

$$\varphi((\xi_1 - \xi)T) = \sqrt{T} \cdot e^{(i/2)(\xi_1 - \xi)T} \cdot \frac{\sin(\frac{1}{2}(\xi_1 - \xi)T)}{\frac{1}{2}(\xi_1 - \xi)T}.$$

Thus

$$\begin{aligned}\int_{-\infty}^{\infty} \|\hat{F}(\cdot, s)\|^2 d\xi &= \|F_1(\cdot)\|^2 \cdot \int_{-\infty}^{\infty} |\varphi((\xi_1 - \xi)T)|^2 d\xi, \\ &= 2\|F_1(\cdot)\|^2 \cdot \int_{-\infty}^{\infty} \frac{\sin^2 \tau}{\tau^2} d\tau = 2\pi\|F_1(\cdot)\|^2.\end{aligned}$$

With F defined by Eq. (10.4.8), the system (10.4.5) becomes, for $s = \eta_1 + i\xi$,

$$\begin{aligned}((\eta_1 + i\xi)I - P)\hat{u} &= F_1(x)\varphi(\xi, T), \\ L_0\hat{u} &= 0, \quad \|\hat{u}\| < \infty.\end{aligned}$$

Because φ is independent of x , the function $\hat{v} = \hat{u}/\varphi((\xi_1 - \xi)T)$ satisfies

$$\begin{aligned}((\eta_1 + i\xi)I - P)\hat{v} &= F_1(x), \\ L_0\hat{v} &= 0, \quad \|\hat{v}\| < \infty.\end{aligned}$$

The solution \hat{v} is a continuous function of its coefficients, and, by Eq. (10.4.7), there is a constant ϵ independent of T such that

$$\|\hat{v}(\cdot, \eta_1 + i\xi)\|^2 + \delta\|\hat{v}_x(\cdot, \eta_1 + i\xi)\|^2 > K(\eta_1)\|F_1(\cdot)\|^2,$$

for $|\xi - \xi_1| \leq \epsilon$. Thus

$$\|\hat{u}(\cdot, \eta_1 + i\xi)\|^2 + \delta\|\hat{u}_x(\cdot, \eta_1 + i\xi)\|^2 > K(\eta_1)|\varphi(\xi, T)|^2 \cdot \|(F_1(\cdot))\|^2, \quad (10.4.9)$$

for $|\xi - \xi_1| \leq \epsilon$.

For large T , the magnitude of $\varphi(\xi, T)$ is small outside a small interval around $\xi = \xi_1$. Thus, for any $\epsilon_1 > 0$, there is a $T = T_0 = T_0(\epsilon_1)$ such that

$$\int_{\xi_1 - \epsilon}^{\xi_1 + \epsilon} |\varphi(\xi, T_0)|^2 d\xi \geq 2\pi - \epsilon_1.$$

By integrating Eq. (10.4.9), we obtain

\

$$\begin{aligned}
& \int_{-\infty}^{\infty} (\|\hat{u}(\cdot, \eta_1 + i\xi)\|^2 + \delta \|\hat{u}_x(\cdot, \eta_1 + i\xi)\|^2) d\xi, \\
& \geq \int_{\xi_1 - \varepsilon}^{\xi_1 + \varepsilon} (\|\hat{u}(\cdot, \eta_1 + i\xi)\|^2 + \delta \|\hat{u}_x(\cdot, \eta_1 + i\xi)\|^2) d\xi, \\
& > K(\eta_1) \int_{\xi_1 - \varepsilon}^{\xi_1 + \varepsilon} |\varphi((\xi_1 - \xi)T_0)|^2 \cdot \|F_1(\cdot)\|^2 d\xi \\
& \geq K(\eta_1) (2\pi - \varepsilon_1) \|F_1(\cdot)\|^2, \\
& = K(\eta_1) \int_{-\infty}^{\infty} \|\hat{F}(\cdot, \eta_1 + \xi)\|^2 d\xi - \varepsilon_1 K(\eta_1) \|F(\cdot)\|^2.
\end{aligned}$$

Because ε_1 is arbitrary, this inequality contradicts Eq. (10.4.6). Accordingly, Eq. (10.4.7) cannot hold.

To prove the theorem in the other direction, we consider (10.4.5), for $\hat{g}_0 = 0$, where \hat{F} is the Laplace transform of F in Eq. (10.4.4). We assume that there is a unique solution satisfying Eq. (10.3.3a). Formally, we can invert the Laplace transform and obtain

$$u(x, t) = \frac{1}{2\pi i} \int_{\mathcal{L}} e^{st} \hat{u}(x, s) ds. \quad (10.4.10)$$

Here \mathcal{L} denotes the line $s = i\xi + \eta$, $\eta = \text{constant} > \eta_0$, $-\infty < \xi < \infty$.

The theory of ordinary differential equations tells us that $\hat{u}(x, s)$ is an analytic function of s for $\operatorname{Re} s > \eta_0$. Also, the smoothness of F implies that there are constants C_p such that

$$\|\hat{F}(\cdot, s)\| \leq \frac{C_p}{|s|^p}, \quad \operatorname{Re} s > \eta_0,$$

and, by Eq. (10.3.3a)

$$\|\hat{u}(\cdot, s)\| \leq K(\eta) \frac{C_p}{|s|^p}, \quad \operatorname{Re} s > \eta_0. \quad (10.4.11)$$

Thus, $u(x, t)$ is a well-defined smooth function of x, t and we can choose any line \mathcal{L} in Eq. (10.4.10) with $\eta > \eta_0$. Substituting Eq. (10.4.10) into Eq. (10.4.4) shows that it solves the differential equation and satisfies the homogeneous boundary conditions. Furthermore, by Parseval's relation, Eq. (10.3.3b) follows.

It remains to be shown that the initial function is zero. We have, for any $\eta > \eta_0$,

$$u(x, 0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(x, i\xi + \eta) d\xi,$$

that is,

$$\begin{aligned} \|u(\cdot, 0)\|^2 &= \frac{1}{(2\pi)^2} \int_0^\infty \left| \int_{-\infty}^\infty \hat{u}(x, i\xi + \eta) d\xi \right|^2 dx, \\ &\leq \frac{1}{(2\pi)^2} \int_0^\infty \left(\int_{-\infty}^\infty \frac{1}{(|\xi| + 1)^2} d\xi \right. \\ &\quad \left. \cdot \int_{-\infty}^\infty (|\xi| + 1)^2 |\hat{u}(x, i\xi + \eta)|^2 d\xi \right) dx, \\ &\leq \text{constant} \int_{-\infty}^\infty (|\xi| + 1)^2 \|\hat{u}(\cdot, i\xi + \eta)\|^2 d\xi, \\ &\leq K(\eta) \cdot C_p^2 \text{constant} \int_{-\infty}^\infty \frac{(|\xi| + 1)^2}{(|s|^p + 1)^2} d\xi. \end{aligned}$$

This integral exists for $p \geq 2$ and every $\eta > \eta_0$. By letting $\eta \rightarrow \infty$, it follows that $u(x, 0) = 0$. This proves the theorem.

In Sections 10.1 and 10.2, we have shown that the eigenvalue condition is sufficient for well-posedness for hyperbolic and parabolic problems in one space dimension with standard boundary conditions. As we see in the next section, this is not true for hyperbolic problems in more than one space dimension. Also, in the discrete case, the corresponding eigenvalue condition is not sufficient for stability, even in one space dimension. This is the reason for introducing our new concept of well-posedness and the conditions (10.3.3a) and (10.3.4a). In fact, it is even needed for problems in one space dimension, if the boundary conditions are not of the standard type. Consider the system

$$\left. \begin{array}{l} u_t + u_x = F, \\ w_t - w_x = G, \end{array} \right\} \quad 0 \leq x < \infty, \quad t \geq 0, \quad (10.4.12a)$$

with boundary conditions

$$u_t(0, t) = w(0, t) \quad (10.4.12b)$$

or, alternatively,

$$u(0, t) = w_t(0, t). \quad (10.4.12c)$$

The transformed equations are

$$\left. \begin{aligned} s\hat{u} + \hat{u}_x &= \hat{F}, \\ s\hat{w} - \hat{w}_x &= \hat{G}, \end{aligned} \right\} \quad 0 \leq x < \infty, \quad (10.4.13a)$$

with boundary conditions

$$s\hat{u}(0, s) = \hat{w}(0, s), \quad (10.4.13b)$$

or, alternatively,

$$\hat{u}(0, s) = s\hat{w}(0, s). \quad (10.4.13c)$$

The eigenvalue problem is

$$\left. \begin{aligned} s\varphi + \varphi_x &= 0, & \|\varphi\| &< \infty, \\ s\psi - \psi_x &= 0, & \|\psi\| &< \infty, \end{aligned} \right\} \quad 0 \leq x < \infty,$$

$$s\varphi(0) = \psi(0) \quad \text{or} \quad \varphi(0) = s\psi(0).$$

For $\operatorname{Re} s > 0$, we have $\psi \equiv 0$, and, therefore, $\varphi \equiv 0$; that is, there are no eigenvalues with $\operatorname{Re} s > 0$.

We now estimate the solutions of Eq. (10.4.13). With $\eta = \operatorname{Re} s$, integration by parts gives us

$$\begin{aligned} \eta \|\hat{u}\|^2 - \frac{1}{2} |\hat{u}(0, s)|^2 &\leq \|\hat{u}\| \|\hat{F}\|, \\ \eta \|\hat{w}\|^2 + \frac{1}{2} |\hat{w}(0, s)|^2 &\leq \|\hat{w}\| \|\hat{G}\|, \end{aligned}$$

that is,

$$\begin{aligned} \frac{\eta}{2} \|\hat{u}\|^2 &\leq \frac{1}{2\eta} \|\hat{F}\|^2 + \frac{1}{2} |\hat{u}(0, s)|^2, \\ \frac{\eta}{2} \|\hat{w}\|^2 + \frac{1}{2} |\hat{w}(0, s)|^2 &\leq \frac{1}{2\eta} \|\hat{G}\|^2. \end{aligned}$$

The boundary condition (10.4.13b) implies

$$|\hat{u}(0, s)|^2 = \frac{1}{|s|^2} |\hat{w}(0, s)|^2 \leq \frac{1}{\eta |s|^2} \|\hat{G}\|^2,$$

and, therefore, the estimate (10.3.3a) [with \hat{F} replaced by $(\hat{F}, \hat{G})^T$] holds and the problem is well posed in the generalized sense.

On the other hand, the boundary condition (10.4.13c) gives us

$$|\hat{u}(0, s)| = |s|^2 |\hat{w}(0, s)| \leq \frac{|s|^2}{\eta} \|\hat{G}\|^2,$$

that is,

$$\frac{\eta}{2} \|\hat{u}\|^2 \leq \frac{1}{2\eta} \|\hat{F}\|^2 + \frac{|s|^2}{2\eta} \|\hat{G}\|^2. \quad (10.4.14)$$

One can prove that this estimate is sharp and, therefore, that the estimate (10.3.3a) does not hold because $\lim_{\eta \rightarrow \infty} K(\eta) \neq 0$. One might think that the definition of well-posedness could be changed such that the last case would be included. However, consider the strip problem

$$\begin{aligned} u_t + u_x &= F, & 0 \leq x \leq 1, \quad t \geq 0, \\ w_t - w_x &= G, \\ u(x, 0) &= w(x, 0) = 0, \end{aligned}$$

with boundary conditions

$$u(0, t) = w_t(0, t), \quad w(1, t) = u(1, t).$$

The corresponding eigenvalue problem is

$$\begin{aligned} s\varphi + \varphi_x &= 0, \\ s\psi - \psi_x &= 0, \\ \varphi(0) &= s\psi(0), \quad \psi(1) = \varphi(1), \end{aligned}$$

that is,

$$\varphi = e^{-sx} \varphi(0), \quad \psi = e^{s(x-1)} \psi(1),$$

where

$$\begin{bmatrix} 1 & -se^{-s} \\ e^{-s} & -1 \end{bmatrix} \begin{bmatrix} \varphi(0) \\ \psi(1) \end{bmatrix} = 0.$$

Then s is an eigenvalue if, and only if,

$$s = e^{2s}. \quad (10.4.15)$$

This equation has solutions with arbitrarily large $\operatorname{Re} s$ (see Exercise 10.4.2), and, therefore, the strip problem is not well posed in any computationally suitable meaning.

One can geometrically explain what happens. The characteristics that support w leave the strip at the boundary $x = 0$. To obtain $u(0, t)$, we have to differentiate w ; that is, we lose one derivative. The value u is transported to the other boundary, and there it is transferred to w through the boundary condition. This value is again transported to the boundary $x = 0$ and loses another derivative when its value is transferred to u . Thus, we lose more and more derivatives as time increases.

EXERCISES

- 10.4.1.** Prove that the estimate (10.4.14) is sharp, that is, that Eq. (10.3.3a) does not hold.
- 10.4.2.** Prove that Eq. (10.4.15) has solutions s with arbitrarily large $\operatorname{Re} s$.
- 10.4.3.** Prove by direct calculation that the eigenvalues s of

$$\left. \begin{aligned} s\varphi + \varphi_x &= 0, \\ s\psi - \psi_x &= 0, \end{aligned} \right\} \quad 0 \leq x \leq 1, \\ s\varphi(0) = \psi(0), \\ \psi(1) = \varphi(1),$$

satisfy $\operatorname{Re} s \leq \eta_0 = \text{constant}$ in agreement with the generalized well-posedness of Eqs. (10.4.12a) and (10.4.12b).

10.5. HYPERBOLIC SYSTEMS WITH CONSTANT COEFFICIENTS IN SEVERAL SPACE DIMENSIONS

In this section, we consider hyperbolic systems

$$u_t = Au_x + Bu_y + F =: P\left(\frac{\partial}{\partial x}\right) u + F \quad (10.5.1a)$$

in the quarter space $x \geq 0$, $-\infty < y < \infty$, $t \geq 0$. For $t = 0$, we give initial data

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \quad \mathbf{x} = (x, y) \quad (10.5.1b)$$

and, at $x = 0$, we prescribe boundary conditions

$$\begin{aligned} L_0 u(0, y, t) &= g(y, t), \quad -\infty < y < \infty, \\ \|u(\cdot, t)\| &< \infty, \end{aligned} \quad (10.5.1c)$$

which are of the same form as in Eq. (10.1.1c).

We assume that all the coefficients are real and A is nonsingular. We also assume that F and g are 2π -periodic in y , and we consider solutions that are 2π -periodic in y .

We want to derive algebraic conditions guaranteeing that the above problem is well posed or strongly well posed in the generalized sense. Corresponding to Eqs. (10.3.3) and (10.3.4), the estimates are now defined as integrals over the domain $0 \leq x < \infty$, $0 \leq y \leq 2\pi$. Corresponding to Lemma 10.1.1, we now have the following lemma.

Lemma 10.5.1. *Consider Eq. (10.5.1) with $F \equiv g \equiv 0$. The problem is not well posed if we can find a complex number s with $\operatorname{Re} s > 0$, an integer ω , and initial data*

$$u(\mathbf{x}, 0) = e^{i\omega y} \varphi(x), \quad \|\varphi(\cdot)\| < \infty,$$

such that

$$u(\mathbf{x}, t) = e^{st + i\omega y} \varphi(x) \quad (10.5.2)$$

is a solution of Eq. (10.5.1) (the Lopatinsky condition).

Proof. If Eq. (10.5.2) is a solution, so is

$$u_n(\mathbf{x}, t) = e^{snt + i\omega ny} \varphi(nx),$$

for any positive integer n . Therefore, we obtain solutions that grow arbitrarily fast and the problem is not well posed.

We now give conditions such that solutions of the form of Eq. (10.5.2) exist. Substituting Eq. (10.5.2) into Eq. (10.5.1) gives us

$$s\varphi = A\varphi_x + i\omega B\varphi, \quad 0 \leq x < \infty, \quad (10.5.3a)$$

$$L_0\varphi(0) = 0, \quad \|\varphi(\cdot)\| < \infty. \quad (10.5.3b)$$

As before, we have the lemma.

Lemma 10.5.2. *There is a solution of the form of Eq. (10.5.2) if, and only if, for some fixed ω , the eigenvalue problem (10.5.3) has an eigenvalue s with $\operatorname{Re} s > 0$.*

REMARK. ω need not be an integer, because, if we have a solution for s, ω , then we also have a solution for $s/|\omega|, \omega/|\omega| = \pm 1$.

By assumption, A is nonsingular, and we can write Eq. (10.5.3a) in the form

$$\varphi_x = M\varphi, \quad M = A^{-1}(sI - i\omega B).$$

We need the following lemma.

Lemma 10.5.3. *Assume that the system (10.5.1) is strongly hyperbolic. Then there is a constant $\delta > 0$ such that, for $\operatorname{Re} s > 0$, the eigenvalues κ of the matrix M satisfy the estimate*

$$|\operatorname{Re} \kappa| \geq \delta |\operatorname{Re} s|. \quad (10.5.4)$$

Proof. Let β be a real number, and consider

$$\begin{aligned} (M - i\beta I)^{-1} &= (A^{-1}(sI - i\omega B) - i\beta I)^{-1}, \\ &= (sI - i\omega B - i\beta A)^{-1}A. \end{aligned}$$

By assumption, the system is strongly hyperbolic, and, therefore, there is a transformation $T = T(\omega, \beta)$ with $\sup_{\omega, \beta}(|T| |T^{-1}|) < \infty$ such that

$$T^{-1}(\omega B + \beta A)T = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} =: \Lambda, \quad \lambda_j \text{ real.}$$

Thus,

$$(sI - i\omega B - i\beta A)^{-1}A = T(sI - i\Lambda)^{-1}T^{-1}A;$$

that is,

$$\begin{aligned} |(M - i\beta I)^{-1}| &\leq |T| |T^{-1}| |A| \cdot |(sI - i\Lambda)^{-1}|, \\ &\leq \delta |\operatorname{Re} s|^{-1}, \quad \delta = |A| |T^{-1}| |T|, \end{aligned} \quad (10.5.5)$$

which implies $|\kappa - i\beta| \geq \delta |\operatorname{Re} s|$. Because β is arbitrary, we choose $\beta = \operatorname{Im} \kappa$, and Eq. (10.5.4) follows.

The last lemma gives us the following lemma.

Lemma 10.5.4. *Assume that the system (10.5.1a) is strongly hyperbolic. For $\operatorname{Re} s > 0$, the matrix M has no eigenvalues κ with $\operatorname{Re} \kappa = 0$. If A has exactly $m - r$ negative eigenvalues, then M has exactly $m - r$ eigenvalues κ with $\operatorname{Re} \kappa < 0$, for all s with $\operatorname{Re} s > 0$ and all real ω .*

Proof. The first statement of the lemma is a weaker statement than Eq. (10.5.4). The eigenvalues κ of M are continuous functions of ω . Therefore, the number of κ with $\operatorname{Re} \kappa < 0$ does not depend on ω since $\operatorname{Re} \kappa$ cannot change sign if we vary ω . In particular, for $\omega = 0$, we obtain

$$M = sA^{-1},$$

and the second statement of the lemma follows.

Assume for a moment that the eigenvalues of M are distinct and denote by $\kappa_1, \dots, \kappa_{m-r}$ the eigenvalues with $\operatorname{Re} \kappa < 0$. Then, the general solution of Eq. (10.5.3a), belonging to L_2 , can be written in the form

$$\varphi = \sum_{j=1}^{m-r} \sigma_j y_j e^{\kappa_j x}.$$

Here the y_j are eigenvectors satisfying

$$My_j = \kappa_j y_j.$$

Substituting this expression into the boundary conditions gives us a linear system of equations for $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_{m-r})$, which we write in the form

$$C(s, \omega)\boldsymbol{\sigma} = 0. \quad (10.5.6)$$

There is a solution of the form of Eq. (10.5.2) if Eq. (10.5.6) has a nontrivial solution.

If the eigenvalues of M are not distinct, then we can still write the general solution, belonging to L_2 , in the form

$$\varphi = \sum_j \varphi_j(x) e^{\kappa_j x}, \quad (10.5.7)$$

where now $\varphi_j(x)$ are polynomials in x with vector coefficients containing altogether $m - r$ parameters σ_j . Therefore, we also obtain a linear system of type (10.5.6) in this case.

We have shown the following theorem to be true.

Theorem 10.5.1. *The initial-boundary-value problem (10.5.1) is not well posed if, for some s with $\operatorname{Re} s > 0$ and some ω ,*

$$\operatorname{Det}(C(s, \omega)) = 0.$$

Now assume that the eigenvalue problem (10.5.3) has no eigenvalue s with $\operatorname{Re} s > 0$. We want to show that we can solve the initial-boundary-value problem using Fourier and Laplace transforms. As before, we assume that the initial data are zero. By assumption, the data are 2π -periodic in y , that is, we can expand them into Fourier series with respect to y . For example,

$$F(\mathbf{x}, t) = \sum_{\omega=-\infty}^{\infty} \tilde{F}(x, \omega, t) e^{i\omega y}.$$

Therefore, we can also expand the solution into a Fourier series

$$u(\mathbf{x}, t) = \sum_{\omega=-\infty}^{\infty} \tilde{u}(x, \omega, t) e^{i\omega y}.$$

Substituting this expression into Eq. (10.5.1) gives us, for every frequency ω , a one-dimensional problem

$$\begin{aligned} \tilde{u}_t &= A\tilde{u}_x + i\omega B\tilde{u} + \tilde{F}, \\ \tilde{u}(x, \omega, 0) &= 0, \\ L_0\tilde{u}(0, \omega, t) &= \tilde{g}(\omega, t). \end{aligned} \tag{10.5.8}$$

We can solve Eq. (10.5.8) using the Laplace transform. The equation

$$\hat{u}(x, \omega, s) = \int_0^\infty e^{-st} \tilde{u}(x, \omega, t) dt$$

satisfies

$$\begin{aligned} s\hat{u} &= A\hat{u}_x + i\omega B\hat{u} + \hat{F}, \quad \|\hat{u}\| < \infty, \\ L_0\hat{u}(0, \omega, s) &= \hat{g}(\omega, s). \end{aligned} \tag{10.5.9}$$

By assumption, the eigenvalue problem (10.5.3) has no eigenvalue with $\operatorname{Re} s > 0$, and, therefore, we can solve Eq. (10.5.9) for $\operatorname{Re} s > 0$ and every ω . Inverting the Laplace and Fourier transforms gives us the desired solution.

By Parseval's relation, the estimates (10.3.3a) and (10.3.4a) now take the form

$$\|\hat{u}(\cdot, \omega, s)\|^2 \leq K(\eta) \|\hat{F}(\cdot, \omega, s)\|^2 \quad (10.5.10)$$

and

$$\|\hat{u}(\cdot, \omega, s)\|^2 \leq K(\eta) (\|\hat{F}(\cdot, \omega, s)\|^2 + |\hat{g}(\omega, s)|^2), \quad ((10.5.11))$$

respectively. Here $K(\eta)$ does not depend on ω .

We now consider the case where $\hat{F} \equiv 0$ and write the differential equation (10.5.9) in the form

$$\begin{aligned} \hat{u}_x &= \tau A^{-1}(s' - i\omega'B)\hat{u} =: \tau M\hat{u}, \\ L_0\hat{u}(0, \omega, s) &= \hat{g}(\omega, s), \end{aligned} \quad (10.5.12)$$

where

$$\tau = \sqrt{|s|^2 + \omega^2}, \quad s' = \frac{s}{\tau}, \quad \omega' = \frac{\omega}{\tau}.$$

By Lemma 10.5.4, for every s', ω' with $\operatorname{Re} s' > 0$, the eigenvalues κ of M split into two groups. By Schur's lemma, we can find a unitary transformation $U = U(\omega', s')$ such that

$$U^*(\omega', s')M(\omega', s')U(\omega', s') = \begin{bmatrix} M_{11} & M_{21} \\ 0 & M_{22} \end{bmatrix},$$

where the eigenvalues κ of M_{11} and M_{22} satisfy $\operatorname{Re} \kappa < 0$ and $\operatorname{Re} \kappa > 0$, respectively. Substituting a new variable $\hat{w} = U^*\hat{u}$ into Eq. (10.5.12), we obtain

$$\begin{aligned} \hat{w}_x^I &= \tau M_{11}\hat{w}^I + \tau M_{12}\hat{w}^{II}, \\ \hat{w}_x^{II} &= \tau M_{22}\hat{w}^{II}, \\ L_0U\hat{w} &= C^I(\omega', s')\hat{w}^I(0, \omega, s) + C^{II}(\omega', s')\hat{w}^{II}(0, \omega, s) = \hat{g}. \end{aligned} \quad (10.5.13)$$

Since we are interested in solutions with $\|w\| < \infty$ and the eigenvalues of M_{22} have a positive real part, it follows that

$$\begin{aligned}\hat{w}^I(x, \omega, s) &= e^{\tau M_{11}x} \hat{w}^I(0, \omega, s), & \hat{w}^{II} &\equiv 0, \\ C^I(\omega', s') \hat{w}^I(0, \omega, s) &= \hat{g}.\end{aligned}\tag{10.5.14}$$

There are two possibilities: Firstly, there exist s'_*, ω'_* with $\operatorname{Re}s'_* \geq 0$ and sequences s'_ν, ω'_ν with $\lim_{\nu \rightarrow \infty} s'_\nu = s'_*$, $\lim_{\nu \rightarrow \infty} \omega'_\nu = \omega'_*$ such that

$$\lim_{\nu \rightarrow \infty} |(C^I(\omega'_\nu, s'_\nu))^{-1}| = \infty. \tag{10.5.15}$$

One can prove that we can choose U such that it is continuous at ω'_*, s'_* . Therefore, Eq. (10.5.15) holds if, and only if,

$$\operatorname{Det}(C^I(\omega'_*, s'_*)) = 0.$$

If $\operatorname{Re}s'_* > 0$, then the homogeneous equations (10.5.14) have a nontrivial solution and, therefore, $s = \tau s'_*$ are eigenvalues of the eigenvalue problem (10.5.3) for $\omega = \tau \omega'_*$. Thus, the problem is not well posed in any sense. If $\operatorname{Re}s_* = 0$, then we obtain a solution of Eq. (10.5.3a) that satisfies $L_0\varphi = 0$ but might not belong to $L_2(0, \infty)$ because some of the eigenvalues of M_{11} might be purely imaginary [cf. Eq. (10.5.4)]. We make the following definition.

Definition 10.5.1. *If $\operatorname{Det}(C^I(\omega'_*, s'_*)) = 0$, where s'_* is purely imaginary, then s_* defined by $s_* = s' \sqrt{|s'_*|^2 + \omega'^2}$ is called a generalized eigenvalue of the eigenvalue problem (10.5.3) if $\|\varphi\| \notin L_2(0, \infty)$.*

REMARK. Even if s'_* is purely imaginary, the corresponding eigenfunction φ might belong to $L_2(0, \infty)$, that is, $\operatorname{Re}\kappa_\nu < 0$, $\nu = 1, \dots, m - r$. In such a case s_* is an eigenvalue.

The theory for the case with generalized eigenvalues, or eigenvalues on the imaginary axis is incomplete. In some cases the initial-boundary-value problem is well posed in the generalized sense, in other cases it is not. We discuss this in more detail for difference approximations.

Secondly, the alternative to Eq. (10.5.15) is that $(C^I(\omega', s'))^{-1}$ is uniformly bounded, or equivalently, that the *determinant condition* is fulfilled:

$$\operatorname{Det}(C^I(\omega', s')) \neq 0, \quad |\omega'| \leq 1, \quad |s'| \leq 1, \quad \operatorname{Re}s' \geq 0. \tag{10.5.16}$$

This is a strengthened version of the Lopatinsky condition given in Lemma 10.5.1. We now have the following lemma.

Lemma 10.5.5. *Consider the initial-boundary-value problem (10.5.1) with $f = F = 0$ and with g satisfying $\int_0^\infty \int_0^{2\pi} |g(y, t)|^2 dy dt < \infty$. Then there is a constant*

$K > 0$ such that its solutions satisfy

$$\int_0^\infty \int_0^{2\pi} |u(0, y, t)|^2 dy dt \leq K \int_0^\infty \int_0^{2\pi} |g(y, t)|^2 dy dt \quad (10.5.17)$$

if, and only if, Eq. (10.5.16) holds.

Proof. First assume that Eq. (10.5.16) holds. Then, because $(C')^{-1}$ is uniformly bounded, the solution \hat{w} of Eq. (10.5.13) satisfies

$$|\hat{w}^I(0, \omega, s)|^2 \leq K |\hat{g}(\omega, s)|^2, \quad \operatorname{Re} s > 0, \quad (10.5.18a)$$

where K is a constant independent of ω, s . The vector function $\hat{u} = U\hat{w}$ satisfies Eq. (10.5.9) with $\hat{F} = 0$, and, because U is a unitary matrix we have $|\hat{u}| = |\hat{w}|$. Therefore,

$$|\hat{u}(0, \omega, s)|^2 \leq K |\hat{g}(\omega, s)|^2, \quad \operatorname{Re} s > 0. \quad (10.5.18b)$$

By Parseval's relation, this inequality implies

$$\begin{aligned} \int_0^\infty \int_0^{2\pi} e^{-2\eta t} |u(0, y, t)|^2 dy dt &\leq K \int_0^\infty \int_0^{2\pi} e^{-2\eta t} |g(y, t)|^2 dy dt, \\ &\leq K \int_0^\infty \int_0^{2\pi} |g(y, t)|^2 dy dt, \quad \eta > 0. \end{aligned}$$

Because the right-hand side is independent of η , Eq. (10.5.17) follows.

Next assume that Eq. (10.5.17) holds. Then, by Parseval's relation, we get the corresponding integral inequality in the Fourier-Laplace space, and, as demonstrated above, this leads to the pointwise estimate (10.5.18b) for arbitrary g . From this estimate, we obtain Eq. (10.5.18a), which is equivalent to Eq. (10.5.16). This proves the lemma.

We now make the following definition.

Definition 10.5.2. Consider the system (10.5.9) for $\hat{F} = 0$. If its solutions satisfy Eq. (10.5.18b), we say that it satisfies the Kreiss condition.

Because the constant K is independent of ω', s' , one might think that the condition $\operatorname{Re} s > 0$ could be replaced by $\operatorname{Re} s \geq 0$. However, the reason for keeping the strict inequality is that it automatically selects the correct general solution \hat{u} through the condition $\|\hat{u}\| < \infty$, because the exponentially growing part is annihilated.

Using the arguments above, we get the following lemma.

Lemma 10.5.6. *The Kreiss condition is satisfied if, and only if, the eigenvalue problem (10.5.3) has no eigenvalue or generalized eigenvalue for $\operatorname{Re} s \geq 0$.*

The main result of the theory is presented in the following theorem.

Theorem 10.5.2. *Assume that Eq. (10.5.1a) is a strictly hyperbolic system. If the Kreiss condition is satisfied, then the initial boundary value problem is strongly well posed in the generalized sense.*

We will not give a proof here. In applications it is not necessary to go through the transformation process leading to the formulation (10.5.13). The general solution \hat{u} of Eq. (10.5.9) with $\|\hat{u}\| < \infty$ for $\hat{F} = 0$ and $\operatorname{Re} s > 0$ is obtained just as for the eigenvalue problem, and we arrive at a system

$$C(s, \omega)\sigma = \tilde{g}, \quad \operatorname{Re} s > 0, \quad (10.5.19)$$

where $C(s, \omega)$ is the matrix occurring in Eq. (10.5.6). With the proper normalization, the Kreiss condition is equivalent to

$$\operatorname{Det}(C(s, \omega)) \neq 0, \quad \operatorname{Re} s \geq 0. \quad (10.5.20)$$

Note that $C(s, \omega)$ must always be defined for $\operatorname{Re} s = 0$ as a limit when s is approaching the imaginary axis from the right. We demonstrate the procedure in an example at the end of this section.

If a hyperbolic problem is well posed, then we have proved that it is also well posed in the generalized sense. One might conjecture that the converse is also true, but no general results are known. However, for strictly hyperbolic equations, we have the following theorem.

Theorem 10.5.3. *Assume that Eq. (10.5.1a) is strictly hyperbolic or symmetric hyperbolic. If the Kreiss condition is satisfied, then the initial-boundary-value problem is strongly well posed.*

Proof. We will only prove this result for symmetric hyperbolic systems. Without restriction, we can assume that

$$A = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix},$$

$$\Lambda_1 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{bmatrix} > 0, \quad \Lambda_2 = \begin{bmatrix} \lambda_{r+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{bmatrix} < 0,$$

is diagonal. We first solve an auxiliary problem

$$\begin{aligned} v_t &= Av_x + Bv_y, \\ v(\mathbf{x}, 0) &= f(\mathbf{x}), \\ v^{II}(0, y, t) &= 0, \quad v^{II} = (v^{(r+1)}, \dots, v^{(m)})^T, \quad \|v(\cdot, t)\| < \infty. \end{aligned} \quad (10.5.21)$$

For its solution, we have the energy estimate

$$\frac{d}{dt} \|v\|^2 + \int_0^{2\pi} \langle v(0, y, t), Av(0, y, t) \rangle dy = 0;$$

that is,

$$\begin{aligned} \|v(\cdot, t)\|^2 &\leq \|v(\cdot, 0)\|^2 = \|f(\cdot)\|^2, \\ (\min_{1 \leq j \leq r} \lambda_j) \int_0^T \int_0^{2\pi} |v(0, y, t)|^2 dy dt &< \int_0^T \int_0^{2\pi} \langle v(0, y, t), Av(0, y, t) \rangle dy dt, \\ &\leq \|f(\cdot)\|^2. \end{aligned} \quad (10.5.22)$$

We assume that $v(x, t)$ is a smooth function of x, t . The difference $w = u - v$ satisfies

$$\begin{aligned} w_t &= Aw_x + Bw_y, \\ w(\mathbf{x}, 0) &= 0, \\ L_0 w(0, y, t) &= g(y, t), \quad g = -L_0 v(0, y, t), \quad \|w(\cdot, t)\| < \infty. \end{aligned} \quad (10.5.23)$$

If the Kreiss condition is satisfied, then

$$\begin{aligned} \int_0^\infty \int_0^{2\pi} |w(0, y, t)|^2 dy dt \\ \leq \text{constant} \int_0^\infty \int_0^{2\pi} |g(y, t)|^2 dy dt \leq \text{constant} \|f(\cdot)\|^2. \end{aligned}$$

Thus, we can estimate the solution of Eq. (10.5.23) on the boundary. We can use integration by parts to estimate $\|w(\cdot, t)\|$ and obtain

$$\begin{aligned} \frac{d}{dt} \|w\|^2 &= - \int_0^{2\pi} \langle w(0, y, t), Aw(0, y, t) \rangle dy, \\ &\leq \text{constant} \int_0^{2\pi} |w(0, y, t)|^2 dy, \end{aligned}$$

that is,

$$\|w(\cdot, T)\|^2 \leq \text{constant} \int_0^T \int_0^{2\pi} |w(0, y, t)|^2 dy dt \leq \text{constant} \|f(\cdot)\|^2. \quad (10.5.24)$$

The estimates (10.5.22) for v yield the final estimate for $u = v + w$, which shows that for symmetric hyperbolic systems the initial-boundary-value problem is strongly well posed if the Kreiss condition is satisfied.

As an example, we now discuss the system

$$\frac{\partial u}{\partial t} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial u}{\partial x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial u}{\partial y}, \quad u = \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}, \quad (10.5.25a)$$

with boundary conditions

$$u^{(1)}(0, y, t) = au^{(2)}(0, y, t) + g, \quad (10.5.25b)$$

where a is a complex constant. Integration by parts gives us

$$\begin{aligned} \frac{d}{dt} \|u\|^2 &= - \int_0^{2\pi} \langle u(0, y, t), \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} u(0, y, t) \rangle dy, \\ &= \int_0^{2\pi} (|u^{(1)}(0, y, t)|^2 - |u^{(2)}(0, y, t)|^2) dy, \\ &= (|a|^2 - 1) \int_0^{2\pi} |u^{(2)}(0, y, t)|^2 dy. \end{aligned}$$

Thus, we obtain an energy estimate for $|a| \leq 1$.

We want to discuss whether we can also estimate the solution for other values of a using the Laplace transform. The eigenvalue problem (10.5.3) has the form

$$\varphi_x = \begin{bmatrix} -s & i\omega \\ -i\omega & s \end{bmatrix} \varphi =: M\varphi,$$

$$\varphi^{(1)}(0) = a\varphi^{(2)}(0), \quad \|\varphi\| < \infty.$$

For $\operatorname{Re} s > 0$, M has exactly one eigenvalue

$$\kappa = -\sqrt{s^2 + \omega^2}, \quad \operatorname{Re} \kappa < 0,$$

with negative real part. The corresponding eigenvector is given by

$$\omega^2 \mathbf{e} = (s + \sqrt{s^2 + \omega^2}, i\omega)^T.$$

Therefore,

$$\varphi(x) = \sigma e^{\kappa x} \mathbf{e}.$$

Thus, the problem is not well posed if the relation

$$s + \sqrt{s^2 + \omega^2} = ia\omega, \quad \operatorname{Re} s > 0, \quad \omega \text{ real}, \quad (10.5.26)$$

has a solution. A simple calculation shows that Eq. (10.5.26) has a solution if, and only if, $|a| > 1$, $\operatorname{Im} a \neq 0$. In that case, the problem is not well posed. We have already shown that the problem is well posed if $|a| \leq 1$. Thus, we need only discuss the case $|a| > 1$, where a is real. The problem (10.5.12) has the form

$$\hat{u}_x = \begin{bmatrix} -s & i\omega \\ -i\omega & s \end{bmatrix} \hat{u}, \quad \hat{u} = \begin{bmatrix} \hat{u}^{(1)} \\ \hat{u}^{(2)} \end{bmatrix},$$

$$\hat{u}^{(1)}(0, \omega, s) = a\hat{u}^{(2)}(0, \omega, s) + \hat{g}(\omega, s), \quad \|\hat{u}(\cdot, \omega, s)\| < \infty, \quad (10.5.27)$$

where we have kept the original variables s , ω instead of the scaled ones s' , ω' . The general solution of the differential equation, belonging to L_2 , is given by

$$\hat{u} = \sigma \begin{bmatrix} s + \sqrt{s^2 + \omega^2} \\ i\omega \end{bmatrix} e^{-\sqrt{s^2 + \omega^2}x}.$$

σ is determined by the boundary condition

$$\sigma(s + \sqrt{s^2 + \omega^2} - ia\omega) = \hat{g};$$

that is,

$$\hat{u} = \frac{\hat{g}}{s + \sqrt{s^2 + \omega^2} - ia\omega} \left[s + \frac{\sqrt{s^2 + \omega^2}}{ia\omega} \right] e^{-\sqrt{s^2 + \omega^2}x}.$$

We now show that $|\hat{u}(0, \omega, s)|/|\hat{g}(\omega, s)|$ is unbounded. Thus, the Kreiss condition is not satisfied. Choose the sign of ω so that $\omega a = |\omega a|$ and determine $\xi_1 > 1$ from

$$\xi_1 + \sqrt{\xi_1^2 - 1} = |a|.$$

Let $s = i|\omega|\xi_1 + \eta$, $\eta \ll |\omega|$. Then

$$\begin{aligned} & \lim_{|\omega| \rightarrow \infty} (|\hat{u}^{(2)}(0, \omega, s)|/|\hat{g}(\omega, s)|) \\ &= \lim_{|\omega| \rightarrow \infty} |i\omega/(s + \sqrt{s^2 + \omega^2} - ia\omega)| \\ &= \lim_{|\omega| \rightarrow \infty} |\omega|/|i|\omega|\xi_1 \\ &\quad + |\eta| + |\omega| \sqrt{1 - \xi_1^2 + 2i\xi_1\eta/|\omega| + (\eta/\omega)^2} - i|a\omega||| \\ &= \lim_{|\omega| \rightarrow \infty} |\omega|/|i|\omega|\xi_1 + \eta \\ &\quad + i|\omega| \sqrt{\xi_1^2 - 1} (1 + i\xi_1\eta/(\omega(1 - \xi_1^2))) \\ &\quad + \mathcal{O}((\eta/\omega)^2) - i|a\omega||| \\ &= \lim_{|\omega| \rightarrow \infty} |\omega|/|\eta + i\xi_1\eta/\sqrt{1 - \xi_1^2} + \mathcal{O}(\eta^2/\omega)|| = \infty. \end{aligned}$$

Thus, we cannot obtain the estimate (10.5.18b), and the problem is not strongly well posed. One can show that it is also not well posed in the generalized sense.

In the example above, energy estimates and Laplace transform techniques yield the same restriction for the boundary conditions. Generally, however, Laplace transform techniques give a much wider class of admissible boundary conditions.

EXERCISES

- 10.5.1.** Prove that Eq. (10.5.26) has a solution if, and only if, $|a| > 1$, $\operatorname{Im} a \neq 0$.
- 10.5.2.** Prove that the problem (10.5.25) is not well posed in the generalized sense for $|a| > 1$, a real. [Hint: Add a forcing function F to Eq. (10.5.25a) and prove that Eq. (10.3.3a) is not satisfied.]

10.6. PARABOLIC SYSTEMS IN MORE THAN ONE SPACE DIMENSION

Technically, we can proceed as for hyperbolic equations. Because there is an extensive literature on the subject, we restrict ourselves to a simple example. We consider

$$\begin{aligned} u_t &= u_{xx} + u_{yy} + F(\mathbf{x}, t), & \mathbf{x} &= (x, y), \\ u(\mathbf{x}, 0) &= f(\mathbf{x}), \\ Lu(0, y, t) &= g(y, t), \end{aligned} \tag{10.6.1}$$

in the quarter space $x \geq 0$, $-\infty < y < \infty$, $t \geq 0$. $L = L(\partial/\partial t, \partial/\partial x, \partial/\partial y)$ represents a linear differential operator with constant coefficients. Using the differential equations, we can eliminate $\partial^p/\partial x^p$, where $p \geq 2$. As before, we consider solutions that are 2π -periodic in y .

Problems that are well posed, or strongly well posed in the generalized sense, are defined as before. As a test for well-posedness, we construct special solutions of the homogeneous differential equations of the form

$$u = e^{i\omega y + st} \varphi(x) \tag{10.6.2a}$$

which satisfy

$$\|\varphi(\cdot)\| < \infty, \quad L\varphi(0) = 0. \tag{10.6.2b}$$

Substituting Eq. (10.6.2) into the differential equations gives us the eigenvalue problem

$$(s + \omega^2)\varphi = \varphi_{xx}, \quad 0 \leq x < \infty, \tag{10.6.3a}$$

$$\|\varphi(\cdot)\| < \infty, \quad L(s, \partial/\partial x, i\omega)\varphi(0) = 0. \tag{10.6.3b}$$

The boundary condition is of the form

$$a(i\omega, s)\varphi_x(0) + b(i\omega, s)\varphi(0) = 0,$$

where a and b are polynomials in s and $i\omega$. As before, we have the next lemma.

Lemma 10.6.1. *The initial-boundary-value problem (10.6.1) is not well posed if there is a sequence of eigenvalues s_j with $\operatorname{Re} s_j \rightarrow \infty$.*

For $\operatorname{Re} s > 0$, the general solution of Eq. (10.6.3a) belonging to L_2 is

$$\varphi = e^{-\sqrt{s+\omega^2}x} \varphi(0).$$

Therefore, the eigenvalues s are the solution of

$$L(s, -\sqrt{s+\omega^2}, i\omega) = -a(i\omega, s)\sqrt{s+\omega^2} + b(i\omega, s) = 0. \quad (10.6.4)$$

We now assume that there are no eigenvalues s with $\operatorname{Re} s > \eta_0$. Then we can solve Eq. (10.6.1) using Laplace and Fourier transforms. Assuming that $f(\mathbf{x}) \equiv 0$, the transformed equations are

$$(s + \omega^2)\hat{u} = \hat{u}_{xx} + \hat{F}, \\ \|\hat{u}(\cdot, \omega, s)\| < \infty, \quad L(s, \partial/\partial x, i\omega)\hat{u}(0, \omega, s) = \hat{g}(\omega, s). \quad (10.6.5)$$

Now we can proceed as in Section 10.2. We write Eq. (10.6.5) as a first-order system by introducing a new variable \hat{v} by $\hat{u}_x = \sqrt{s+\omega^2}\hat{v}$ and obtain

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}_x = \sqrt{s+\omega^2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} - \frac{1}{\sqrt{s+\omega^2}} \begin{bmatrix} 0 \\ \hat{F} \end{bmatrix}, \quad (10.6.6a)$$

$$a(i\omega, s)\sqrt{s+\omega^2}\hat{v}(0, \omega, s) + b(i\omega, s)\hat{u}(0, \omega, s) = \hat{g}(\omega, s). \quad (10.6.6b)$$

The change of variables

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}$$

transforms Eq. (10.6.6a) into the diagonal system

$$y_x = \sqrt{s+\omega^2} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} y - \frac{1}{\sqrt{2(s+\omega^2)}} \begin{bmatrix} -\hat{F} \\ \hat{F} \end{bmatrix}, \quad (10.6.7a)$$

and the boundary conditions (10.6.6b) into

$$r_1 y^{(1)} + r_2 y^{(2)} = \hat{g}(\omega, s), \quad (10.6.7b)$$

where

$$\begin{aligned} r_1 &= (-a(i\omega, s) \sqrt{s + \omega^2} + b(i\omega, s)) / \sqrt{2}, \\ r_2 &= (a(i\omega, s) \sqrt{s + \omega^2} + b(i\omega, s)) / \sqrt{2}. \end{aligned}$$

As in Section 10.2, we can now estimate the solution of Eq. (10.6.7) and obtain the following theorem.

Theorem 10.6.1. *Assume that there are constants $\eta_0 \geq 0$, $\delta > 0$, $C \geq 0$ such that, for all ω, s with $\operatorname{Re} s > \eta_0$,*

$$|r_1| \geq \delta, \quad \left| \frac{r_2}{r_1} \right| \leq C. \quad (10.6.8)$$

Then the initial-boundary-value problem is strongly well posed in the generalized sense.

Now we will discuss different boundary conditions.

1. Dirichlet and Neumann Conditions:

$$u = g \quad \text{or} \quad u_x = g, \quad x = 0. \quad (10.6.9)$$

In this case

$$a = 0, \quad b = 1 \quad \text{or} \quad a = 1, \quad b = 0,$$

respectively. The conditions of Eq. (10.6.8) are satisfied.

2. Oblique Boundary Conditions:

$$u_x + \beta u_y = g, \quad x = 0. \quad (10.6.10)$$

Now $a = 1$, $b = i\beta\omega$. Therefore,

$$r_1 = -\sqrt{s + \omega^2} + i\beta\omega, \quad r_2 = \sqrt{s + \omega^2} + i\beta\omega.$$

A simple calculation shows that the conditions (10.6.8) are satisfied.

3. Boundary Conditions Containing Time Derivatives:

$$u_t + u_x = g, \quad x = 0. \quad (10.6.11)$$

In this case,

$$r_1 = -\sqrt{s + \omega^2} + s, \quad r_2 = \sqrt{s + \omega^2} + s.$$

If $s = \frac{1}{2} + \sqrt{\frac{1}{4} + \omega^2}$, then $r_1 = 0$ and, by Lemma 10.6.1, the problem is not well posed. Alternatively, if we use the boundary conditions

$$u_t - u_x = g, \quad x = 0, \quad (10.6.12)$$

we obtain

$$r_1 = \sqrt{s + \omega^2} + s, \quad r_2 = -\sqrt{s + \omega^2} + s,$$

and a simple calculation shows that the conditions (10.6.8) are satisfied. One can “stabilize” the first boundary conditions by replacing them by

$$u_t + u_x - \varepsilon u_{yy} = g. \quad (10.6.13)$$

Now the conditions (10.6.8) are satisfied.

EXERCISES

- 10.6.1.** Prove that the conditions (10.6.8) are satisfied for the boundary conditions (10.6.10), (10.6.12), and (10.6.13).

10.7. SYSTEMS WITH VARIABLE COEFFICIENTS IN GENERAL DOMAINS

As we have seen in Section 9.6, the initial-boundary-value problem in general domains with smooth boundaries can be reduced to the Cauchy problem and quarter-space problems where the coefficients of the differential equation and the boundary conditions are 2π -periodic in the tangential variables. The quarter-space problems with variable coefficients can, for large classes of hyperbolic, parabolic, and mixed hyperbolic-parabolic problems, be reduced to systems with constant coefficients.

As an example, consider the system (10.5.1), where A and B and the coefficients of the boundary conditions are now smooth functions of all variables. Assume that the system is strictly hyperbolic. Consider all systems with con-

stant coefficients by “freezing” the coefficients at every boundary point $x = 0$, $y = y_0$, $t = t_0$. If, for all these systems with constant coefficients, the estimate (10.5.16) holds uniformly, then the problem with variable coefficients is also strongly well posed in the generalized sense.

Similar results hold for parabolic and mixed hyperbolic-parabolic systems. Nonlinear problems can also be solved by linearization and iteration. For more details we refer to the literature.

BIBLIOGRAPHIC NOTES

In Gustafsson and Kreiss (1983), the Euler equations are analyzed for low Mach-numbers. The result is another example where the Laplace transform technique shows well-posedness for a wider class of boundary conditions than the one obtained with the energy method.

Other examples where analysis of the type described in this chapter has been applied to “real” problems are found in Henshaw, Kreiss, and Reyna (1994) and Johansson (1991a, 1991b, 1993).

For references concerning the general theory, we refer to Kreiss and Lorenz (1989).

11

THE ENERGY METHOD FOR DIFFERENCE APPROXIMATIONS

11.1. HYPERBOLIC PROBLEMS

In this section, we want to consider difference approximations of the system (9.2.1) and (9.2.2) and derive discrete energy estimates. In the continuous case, these energy estimates were obtained using the integration by parts rules in Lemma 9.2.1. We need corresponding summation-by-parts rules for the discrete approximations of $\partial/\partial x$. To derive these rules, we divide the interval $0 \leq x \leq 1$ into subintervals of length $h = 1/N$, where N is a natural number. Introduce gridpoints

$$x_j = jh, \quad j = 0, 1, \dots, N,$$

and gridfunctions

$$u_j = u(x_j).$$

The simplest scalar product and norm are defined by

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

or, in the case of vector valued functions,

$$(u, v)_{r,s} = \sum_{j=r}^s \langle u_j, v_j \rangle h.$$

We now have the following lemma, which corresponds to Lemma 9.2.1.

Lemma 11.1.1.

$$\begin{aligned}
 (u, D_+ v)_{r,s} &= -(D_- u, v)_{r+1,s+1} + \bar{u}_j v_j|_r^{s+1}, \\
 &= -(D_- u, v)_{r,s} - h(D_+ u, D_+ v)_{r,s} + \bar{u}_j v_j|_r^{s+1}, \\
 (u, D_0 v)_{r,s} &= -(D_0 u, v)_{r,s} + \frac{1}{2}(\bar{u}_j v_{j+1} + \bar{u}_{j+1} v_j)|_{r-1}^s, \\
 F_{j+\alpha}|_\ell^k &= F_{k+\alpha} - F_{\ell+\alpha}.
 \end{aligned}$$

Proof.

$$\begin{aligned}
 (u, D_+ v)_{r,s} &= \sum_{j=r}^s \bar{u}_j v_{j+1} - \sum_{j=r}^s \bar{u}_j v_j = \sum_{j=r+1}^{s+1} \bar{u}_{j-1} v_j - \sum_{j=r}^s \bar{u}_j v_j, \\
 &= - \sum_{j=r+1}^{s+1} (\bar{u}_j - \bar{u}_{j-1}) v_j + \bar{u}_j v_j|_r^{s+1}, \\
 &= -(D_- u, v)_{r+1,s+1} + \bar{u}_j v_j|_r^{s+1} \\
 &= - \sum_{j=r}^s (\bar{u}_{j+1} - \bar{u}_j) v_{j+1} + \bar{u}_j v_j|_r^{s+1}, \\
 &= - \sum_{j=r}^s (\bar{u}_{j+1} - \bar{u}_j) v_j - \sum_{j=r}^s (\bar{u}_{j+1} - \bar{u}_j)(v_{j+1} - v_j) \\
 &\quad + \bar{u}_j v_j|_r^{s+1}, \\
 &= -(D_- u, v)_{r,s} - h(D_+ u, D_+ v)_{r,s} + \bar{u}_j v_j|_r^{s+1}, \\
 2(u, D_0 v)_{r,s} &= \sum_{j=r}^s \bar{u}_j v_{j+1} - \sum_{j=r}^s \bar{u}_j v_{j-1} = \sum_{j=r+1}^{s+1} \bar{u}_{j-1} v_j - \sum_{j=r-1}^{s-1} \bar{u}_{j+1} v_j, \\
 &= - \sum_{j=r}^s (\bar{u}_{j+1} - \bar{u}_{j-1}) v_j + (\bar{u}_s v_{s+1} + \bar{u}_{s+1} v_s) \\
 &\quad - (\bar{u}_{r-1} v_r + \bar{u}_r v_{r-1}), \\
 &= -2(D_0 u, v)_{r,s} + (\bar{u}_j v_{j+1} + \bar{u}_{j+1} v_j)|_{r-1}^s.
 \end{aligned}$$

This proves the lemma.

We can now derive energy estimates for simple difference approximations. We begin with the scalar equation

$$u_t = u_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (11.1.1)$$

with initial and boundary data (see Figure 11.1.1)

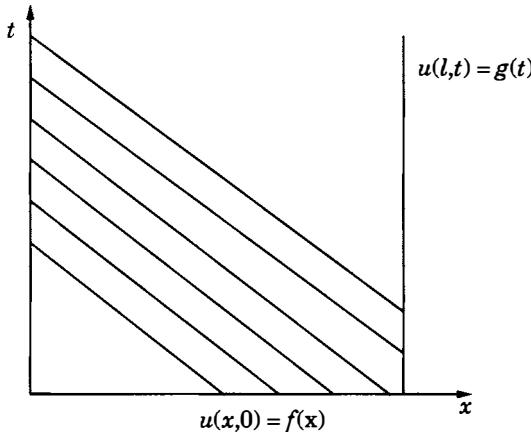


Figure 11.1.1.

$$u(x, 0) = f(x), \quad u(1, t) = g(t). \quad (11.1.2)$$

We approximate Eqs. (11.1.1) and (11.1.2) by

$$\frac{dv_j}{dt} = D_+ v_j, \quad j = 0, 1, \dots, N - 1, \quad (11.1.3)$$

with initial and boundary conditions

$$v_j(0) = f_j, \quad j = 0, 1, \dots, N - 1, \quad v_N(t) = g(t). \quad (11.1.4)$$

We can eliminate v_N from Eq. (11.1.3) using the boundary condition. Therefore, Eqs. (11.1.3) and (11.1.4) represent an initial value problem for N ordinary differential equations in N unknowns. Lemma 11.1.1 gives us the energy estimate

$$\begin{aligned} \frac{d}{dt} \|v\|_{0,N-1}^2 &= \left(\frac{dv}{dt}, v \right)_{0,N-1} + \left(v, \frac{dv}{dt} \right)_{0,N-1}, \\ &= (D_+ v, v)_{0,N-1} + (v, D_+ v)_{0,N-1}, \\ &= -h \|D_+ v\|_{0,N-1}^2 + |v_N(t)|^2 - |v_0(t)|^2 \leq |g(t)|^2. \end{aligned} \quad (11.1.5)$$

Therefore,

$$\|v(t)\|_{0,N-1}^2 \leq \|f\|_{0,N-1}^2 + \int_0^t |g(\tau)|^2 d\tau. \quad (11.1.6)$$

The simplest time discretization is the forward Euler scheme

$$\begin{aligned} w_j^{n+1} &= (I + kD_+)w_j^n, \quad j = 0, 1, \dots, N-1, \\ w_j^0 &= f_j, \\ w_N^n &= g^n. \end{aligned} \tag{11.1.7}$$

Now we obtain, for $\lambda = k/h \leq 1$, using Lemma 11.1.1,

$$\begin{aligned} \|w^{n+1}\|_{0,N-1}^2 &= \|(I + kD_+)w^n\|_{0,N-1}^2, \\ &= \|w^n\|_{0,N-1}^2 + k^2 \|D_+ w^n\|_{0,N-1}^2 \\ &\quad + k((w^n, D_+ w^n)_{0,N-1} + (D_+ w^n, w^n)_{0,N-1}), \\ &= \|w^n\|_{0,N-1}^2 - (hk - k^2) \|D_+ w^n\|_{0,N-1}^2 + k|w_N^n|^2 - k|w_0^n|^2, \\ &\leq \|w^n\|_{0,N-1}^2 + k|g^n|^2; \end{aligned} \tag{11.1.8}$$

that is,

$$\|w^n\|_{0,N-1}^2 \leq \|f\|_{0,N-1}^2 + \sum_{\nu=0}^{n-1} |g^\nu|^2 k.$$

Thus, we also obtain an energy estimate for the fully discretized approximation that corresponds to Eq. (11.1.6).

We approximate the problem

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

by

$$\begin{aligned} \frac{dv_j}{dt} &= -D_- v_j, \quad j = 1, 2, \dots, N, \\ v_j(0) &= f_j, \quad j = 1, 2, \dots, N, \quad v_0(t) = g(t), \end{aligned}$$

and again use the forward Euler scheme for time discretization. Instead of Eqs. (11.1.5) and (11.1.8), we now obtain

$$\frac{d}{dt} \|v\|_{1,N}^2 = -h \|D_- v\|_{1,N}^2 - |v_N(t)|^2 + |g(t)|^2,$$

and

$$\begin{aligned}\|w^{n+1}\|_{1,N}^2 &= \|(I - kD_-)w^n\|_{1,N}^2, \\ &= \|w^n\|_{1,N}^2 - (hk - k^2)\|D_- w^n\|_{1,N}^2 - k|w_N^n|^2 + k|w_0^n|^2, \\ &\leq \|w^n\|_{1,N}^2 + k|g^n|^2,\end{aligned}\tag{11.1.9}$$

respectively.

Next consider the system with constant coefficients

$$\begin{aligned}\frac{\partial u}{\partial t} &= \Lambda u_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \\ u''(0, t) &= R^I u'(0, t) + g''(t), \\ u'(1, t) &= R^{II} u''(1, t) + g^I(t),\end{aligned}\tag{11.1.10}$$

where

$$\Lambda = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}$$

is a diagonal matrix with $\Lambda^I > 0$ and $\Lambda^{II} < 0$. The approximation analogous to those above is given by

$$\begin{aligned}\frac{dv_j^I}{dt} &= \Lambda^I D_+ v_j^I, \quad j = 0, 1, \dots, N-1, \\ \frac{dv_j^{II}}{dt} &= \Lambda^{II} D_- v_j^{II}, \quad j = 1, 2, \dots, N, \\ v_j(0) &= f_j, \\ v_0^{II}(t) &= R^I v_0^I(t) + g''(t), \\ v_N^I(t) &= R^{II} v_N^{II}(t) + g^I(t).\end{aligned}\tag{11.1.11}$$

Referring back to Section 9.2, we can assume that $|R^I|$ and $|R^{II}|$ are as small as necessary. We introduce the scalar product

$$(v, w)_h = (v^I, w^I)_{0, N-1} + (v^{II}, w^{II})_{1, N}$$

and obtain, from Eq. (11.1.11) and Lemma 11.1.1,

$$\begin{aligned}\frac{d}{dt} \|v\|_h^2 &= (\Lambda^I D_+ v^I, v^I)_{0,N-1} + (v^I, \Lambda^I D_+ v^I)_{0,N-1} \\ &\quad + (\Lambda^{II} D_- v^{II}, v^{II})_{1,N} + (v^{II}, \Lambda^{II} D_- v^{II})_{1,N} \\ &= -h(D_+ v^I, \Lambda^I D_+ v^I)_{0,N-1} + h(D_- v^{II}, \Lambda^{II} D_- v^{II})_{1,N} \\ &\quad + \langle v_j, \Lambda v_j \rangle_0^N \leq \text{constant} (|g^I|^2 + |g^{II}|^2),\end{aligned}$$

that is,

$$\|v(t)\|_h \leq \|v(0)\|_h + \text{constant} \int_0^t (|g^I(\tau)|^2 + |g^{II}(\tau)|^2) d\tau,$$

provided $|R^I|$ and $|R^{II}|$ are sufficiently small.

If $\Lambda \in C^1$ is a function of x, t , then we obtain

$$\frac{d}{dt} \|v\|_h^2 \leq 2\alpha(t) \|v\|_h^2 + \text{constant} (|g^I|^2 + |g^{II}|^2),$$

where $\alpha = \max_x |\Lambda_x|$. (See the corresponding results for periodic problems.)

Assume that $g^I \equiv g^{II} \equiv 0$. We define the semidiscrete solution operator $S_h(t, t_0)$ as the mapping

$$v(\cdot, t) = S_h(t, t_0)v(\cdot, t_0).$$

The estimate above tells us that

$$\|S_h(t, t_0)\| \leq e^{\int_{t_0}^t \alpha(\tau) d\tau}.$$

Therefore, we can use Duhamel's principle to write down the solution of the inhomogeneous differential equation and estimate it. As for periodic problems, we can add lower order terms and still get estimates of the same form.

For time discretization, we again use the explicit Euler method and obtain the same kind of estimates, provided the eigenvalues λ_j of Λ satisfy $\max_j |\lambda_j|k/h \leq 1$. We also could have used the implicit Euler method or the Crank–Nicholson scheme. With these methods, there is no restriction on $\lambda = k/h$.

The scheme above can be used to prove the existence of solutions of the initial-boundary-value problem in the manner described earlier. However, for practical calculations, the scheme has a number of drawbacks.

1. It is only first order accurate.
2. For scalar equations

$$u_t = au_x$$

the approximation Qv of au_x depends on the sign of a ; that is,

$$Qv = \begin{cases} aD_+, & \text{if } a > 0, \\ aD_-, & \text{if } a < 0. \end{cases}$$

Otherwise, no energy estimates are possible since the approximation is not stable. This poses some difficulties if $a = a(x, t)$ is a function of x, t and changes sign. We can overcome them by using

$$Qv = \frac{1}{2}(a + |a|)D_+v + \frac{1}{2}(a - |a|)D_-v.$$

In applications, hyperbolic systems are rarely diagonal. If one wants to employ the above method, one must diagonalize the system first; this can be quite expensive with regard to computer time. (Of course one can diagonalize the system, write down the difference approximation, and then transform the system back to its original form).

We now derive an approximation that is second-order accurate and need not be in diagonal form. We start with the scalar problem Eqs. (11.1.1) and (11.1.2) and approximate it by

$$\begin{aligned} dv_j/dt &= D_0v_j, & j &= 1, 2, \dots, N-1, \\ v_j(0) &= f_j. \end{aligned} \tag{11.1.12}$$

To obtain a unique solution, we need equations for v_0 and v_N . We use

$$dv_0/dt = D_+v_0, \quad v_N(t) = g(t). \tag{11.1.13}$$

Thus, we use the boundary condition of the continuous problem, a centered approximation in the interior, and a one-sided approximation at $x = 0$.

The modification of the difference operator at $x = 0$ can also be expressed as an addition of an extra boundary condition, that is, we use the centered approximation at $x = 0$ but supply a boundary condition that determines v_{-1} :

$$dv_0/dt = D_0v_0, \quad h^2D_+D_-v_0 := v_1 - 2v_0 + v_{-1} = 0.$$

If we eliminate v_{-1} , we again obtain the one-sided approximation above. The extra boundary condition determines v_{-1} as a linear extrapolation of v_1 and v_0 . Linear or higher order extrapolation techniques are often used to supply extra boundary conditions.

Formally, we can write the difference approximation as

$$dv/dt = Qv, \quad v_N(t) = g(t).$$

As in the previous examples, we obtain an energy estimate if we can construct a scalar product $(\cdot, \cdot)_h$ such that

$$\operatorname{Re} (v, Qv)_h = |v_N|^2 - |v_0|^2;$$

that is, Q has the same property as $\partial/\partial x$. The scalar product we use is

$$(u, v)_h = \frac{1}{2}(\bar{u}_0 v_0 + \bar{u}_N v_N)h + (u, v)_{1, N-1}.$$

The construction of suitable scalar products is discussed in more detail in Section 11.4. Using Lemma 11.1.1, we obtain

$$\begin{aligned} \frac{d}{dt} \|v\|_h^2 &= \frac{1}{2}((D_+ \bar{v}_0)v_0 + \bar{v}_0 D_+ v_0)h + \frac{1}{2}((D_- \bar{v}_N)v_N + \bar{v}_N D_- v_N)h \\ &\quad + (D_0 v, v)_{1, N-1} + (v, D_0 v)_{1, N-1}, \\ &= |v_N|^2 - |v_0|^2 \leq |g(t)|^2. \end{aligned}$$

Thus, we obtain an energy estimate.

If the trapezoidal rule is used for time discretization, we obtain

$$\begin{aligned} v_j^{n+1} - v_j^n &= \frac{k}{2} Q(v_j^{n+1} + v_j^n), \quad j = 0, 1, \dots, N-1, \\ v_N^{n+1} &= g^{n+1}, \\ v_j^0 &= f_j. \end{aligned} \tag{11.1.14}$$

Because Q is semibounded under our scalar product, unconditional stability follows immediately as with the pure initial value problem treated in Section 5.3.

Similarly, we get unconditional stability for the backward Euler method

$$\begin{aligned} (I - kQ)v_j^{n+1} &= v_j^n, \quad j = 0, 1, \dots, N-1, \\ v_N^{n+1} &= g^{n+1}, \\ v_j^0 &= f_j. \end{aligned} \tag{11.1.15}$$

(cf. Theorem 5.3.3).

Implicit schemes are usually inefficient for solving hyperbolic problems. A more convenient scheme is the leap-frog scheme, modified at the boundary to attain stability. We write the approximation [Eqs. (11.1.12) and (11.1.13)] with

$g = 0$ in matrix form

$$\frac{d\mathbf{v}}{dt} = \frac{1}{h} \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ \dots & & & & & & \\ 0 & \dots & 0 & -\frac{1}{2} & 0 & & \end{bmatrix} \mathbf{v}, \quad \mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \end{bmatrix},$$

that is,

$$D \frac{d\mathbf{v}}{dt} = \frac{1}{h} (C - B)\mathbf{v},$$

where

$$D = \begin{bmatrix} \frac{1}{2} & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & & & & \\ 0 & \dots & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \dots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ \dots & & & & & \\ 0 & \dots & 0 & -\frac{1}{2} & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} \frac{1}{2} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & & & \\ 0 & \dots & 0 \end{bmatrix}.$$

Thus,

$$\frac{d}{dt} (D^{1/2}\mathbf{v}) = \frac{1}{h} D^{-1/2} CD^{-1/2} (D^{1/2}\mathbf{v}) - \frac{1}{h} D^{-1/2} BD^{-1/2} (D^{1/2}\mathbf{v})$$

can be considered as a system for $D^{1/2}\mathbf{v}$. The matrix $D^{-1/2}CD^{-1/2}$ is antisymmetric and $D^{-1/2}BD^{-1/2}$ is symmetric and semidefinite. Therefore, we can use the results of Theorem 5.3.4 to construct the time discretization

$$D^{1/2} \frac{\mathbf{v}^{n+1} - \mathbf{v}^{n-1}}{2k} = \frac{1}{h} D^{-1/2} C \mathbf{v}^n - \frac{1}{h} D^{-1/2} B \frac{\mathbf{v}^{n+1} + \mathbf{v}^{n-1}}{2},$$

that is

$$\begin{aligned} v_0^{n+1} &= v_0^{n-1} + 2 \frac{k}{h} (v_1^n - \frac{1}{2}(v_0^{n+1} + v_0^{n-1})), \\ v_j^{n+1} &= v_j^{n-1} + 2k D v_j^n, \quad j = 1, 2, \dots, N-1, \\ v_N^n &= 0. \end{aligned} \tag{11.1.16}$$

The approximation is stable if

$$\frac{k}{h} |D^{-1/2} C D^{-1/2}| \leq 1 - \delta, \quad \delta > 0.$$

Because the grid values v_j^{n+1} are not coupled to each other, the scheme is explicit. Let us calculate $|D^{-1/2} C D^{-1/2}|$. By definition,

$$\begin{aligned} &|D^{-1/2} C D^{-1/2}|^2 \\ &= \max_{|\mathbf{u}| \neq 0} \frac{|(D^{-1/2} C D^{-1/2})\mathbf{u}|^2}{|\mathbf{u}|^2}, \\ &= \max_{|\mathbf{u}| \neq 0} \frac{\frac{1}{2}|u_1|^2 + \left| -\frac{1}{\sqrt{2}}u_0 + \frac{1}{2}u_2 \right|^2 + \frac{1}{4} \sum_{j=2}^{N-2} |u_{j+1} - u_{j-1}|^2 + \frac{1}{4}|u_{N-1}|^2}{\sum_{j=0}^{N-1} |u_j|^2}, \\ &\leq \frac{\frac{1}{2}|u_1|^2 + |u_0|^2 + \frac{1}{2}|u_2|^2 + \frac{1}{2} \sum_{j=2}^{N-2} (|u_{j+1}|^2 + |u_{j-1}|^2) + \frac{1}{4}|u_{N-1}|^2}{\sum_{j=0}^{N-1} |u_j|^2} \leq 1. \end{aligned}$$

Thus, the approximation is stable for $k/h < 1$.

We can generalize these results to systems

$$u_t = A u_x, \quad 0 \leq x \leq 1, \quad t \geq 0. \tag{11.1.17}$$

Here $A = A^*$ is a symmetric matrix, which need not be diagonal. For convenience only, we assume that A is constant and that the boundary conditions are homogeneous. Assume that the boundary conditions are of the form

$$\begin{aligned} u^I(0, t) &= 0, \quad u^I = (u^{(1)}, \dots, u^{(r)})^T, \\ u^{II}(1, t) &= 0, \quad u^{II} = (u^{(r+1)}, \dots, u^{(m)})^T. \end{aligned} \tag{11.1.18}$$

Then

$$\frac{d}{dt} \|u\|^2 = (u, A u_x) + (A u_x, u) = \langle u, A u \rangle |_0^1,$$

and we obtain an energy estimate if

$$(-1)^j \langle u(j, t), Au(j, t) \rangle \geq 0, \quad j = 0, 1, \quad (11.1.19)$$

for all u that satisfy the boundary conditions.

The semidiscrete approximation is given by

$$\begin{aligned} \frac{dv_j}{dt} &= AD_0 v_j, & j &= 1, 2, \dots, N-1, \\ \frac{dv_0^{II}}{dt} &= (AD_+ v_0)^{II}, & v_0^I(t) &= 0, \\ \frac{dv_N^I}{dt} &= (AD_- v_N)^I, & v_N^{II}(t) &= 0. \end{aligned} \quad (11.1.20)$$

We now prove that it satisfies an energy estimate. We use the scalar product

$$(u, v)_h = \frac{h}{2} (\langle u_0, v_0 \rangle + \langle u_N, v_N \rangle) + (u, v)_{1, N-1},$$

and obtain, from Lemma 11.1.1 and Eq. (11.1.19),

$$\begin{aligned} \frac{d}{dt} \|v\|_h^2 &= \frac{h}{2} (\langle v_0^{II}, (AD_+ v_0)^{II} \rangle + \langle (AD_+ v_0)^{II}, v_0^{II} \rangle) \\ &\quad + \frac{h}{2} (\langle v_N^I, (AD_- v_N)^I \rangle + \langle (AD_- v_N)^I, v_N^I \rangle) \\ &\quad + (v, AD_0 v)_{1, N-1} + (AD_0 v, v)_{1, N-1}, \\ &= \frac{h}{2} (\langle v_0, AD_+ v_0 \rangle + \langle AD_+ v_0, v_0 \rangle) \\ &\quad + \frac{h}{2} (\langle v_N, AD_- v_N \rangle + \langle AD_- v_N, v_N \rangle) \\ &\quad + (v, AD_0 v)_{1, N-1} + (AD_0 v, v)_{1, N-1} = \langle v_j, Av_j \rangle|_0^N \leq 0. \end{aligned}$$

The energy estimate follows.

Corresponding to Eq. (11.1.16), the completely discretized approximation is

$$\begin{aligned}
v_j^{n+1} &= v_j^{n-1} + 2kAD_0 v_j^n, \quad j = 1, 2, \dots, N-1, \\
(v_0^{n+\nu})^I &= 0, \quad \nu = -1, 0, 1, \\
(v_0^{n+1})^{II} &= (v_0^{n-1})^{II} + \frac{2k}{h} (A(v_1^n - \frac{1}{2}(v_0^{n+1} + v_0^{n-1})))^{II}, \\
(v_N^{n+\nu})^{II} &= 0, \quad \nu = -1, 0, 1, \\
(v_N^{n+1})^I - (v_N^{n-1})^I &- \frac{2k}{h} (A(v_{N-1}^n - \frac{1}{2}(v_N^{n+1} + v_N^{n-1})))^I. \quad (11.1.21)
\end{aligned}$$

It is stable for $k|A|/h < 1$.

Now assume that the boundary conditions are not given in the form of Eq. (11.1.18), but consist of $m-r$ and r linearly independent relations

$$L_0 u(0, t) = 0 \quad \text{and} \quad L_1 u(1, t) = 0, \quad (11.1.22)$$

respectively, which still satisfy Eq. (11.1.19). We can assume that the row vectors of L_0, L_1 are orthogonal. Then we can construct unitary matrices

$$U_0 = \begin{bmatrix} L_0 \\ R_0 \end{bmatrix}, \quad U_1 = \begin{bmatrix} L_1 \\ R_1 \end{bmatrix},$$

and a unitary matrix $U(x) \in C^\infty$ with

$$U(x) = \begin{cases} U_0, & \text{for } 0 \leq x \leq \frac{1}{3}, \\ U_1, & \text{for } \frac{2}{3} \leq x \leq 1, \end{cases}$$

that connects U_0 with U_1 . Substituting new variables $\tilde{u} = Uu$ into Eq. (11.1.17) and (11.1.22), the boundary conditions for \tilde{u} will be of the form of Eq. (11.1.18), and in the neighborhood of the boundaries the differential equations become

$$\tilde{u}_t = UAU^*\tilde{u}_x.$$

The approximation at $x = 0$ has the form

$$\begin{aligned}
(\tilde{v}_0^{n+\nu})^I &= 0, \quad \nu = -1, 0, 1, \\
(\tilde{v}_0^{n+1})^{II} &= (\tilde{v}_0^{n-1})^{II} + \frac{2k}{h} \left(UAU^* \left(\tilde{v}_1^n - \frac{1}{2}(\tilde{v}_0^{n+1} + \tilde{v}_0^{n-1}) \right) \right)^{II}.
\end{aligned}$$

In the original variables, we have

$$(Uv_0^{n+\nu})^I = L_0 v_0^{n+\nu} = 0, \quad \nu = -1, 0, 1,$$

$$R_0 v_0^{n+1} = R_0 v_0^{n-1} + \frac{2k}{h} R_0 \left(A \left(v_1^n - \frac{1}{2} (v_0^{n+1} + v_0^{n-1}) \right) \right).$$

At $x = 1$, we obtain the corresponding relations. The approximation is stable for $k|A|/h < 1$.

EXERCISES

11.1.1. Consider the approximation

$$\begin{aligned} \frac{dv_j}{dt} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} D_0 v_j, \quad j = 1, 2, \dots, \\ \frac{dv_0^{II}}{dt} &= D_+ v_0^I, \\ v_0^I &= 0, \\ v_j(0) &= f_j. \end{aligned}$$

Prove that the norm

$$\|v\| = \left(\frac{h}{2} |v_0|^2 + \|v\|_{1,\infty}^2 \right)^{1/2}$$

is independent of t .

11.1.2. Consider the approximation

$$(I - kQ_1)v_j^{n+1} = 2kQ_0v_j^n + (I + kQ_1)v_j^{n-1},$$

where

$$\operatorname{Re}(w, Q_1 w)_h \leq \alpha \|w\|_h^2,$$

$$\operatorname{Re}(w, Q_0 w)_h = 0,$$

$$k\|Q_0\|_h \leq 1 - \delta, \quad \delta > 0,$$

for all gridfunctions satisfying the boundary conditions. Prove that it is stable.

11.2 PARABOLIC DIFFERENTIAL EQUATIONS

We begin with an example. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{11.2.1}$$

with Dirichlet boundary conditions

$$u(0, t) = u(1, t) = 0. \tag{11.2.2}$$

The mesh is constructed as it was for hyperbolic differential equations. The semidiscrete approximation is

$$\begin{aligned} \frac{dv_j}{dt} &= D_+ D_- v_j, \quad j = 1, 2, \dots, N - 1, \\ v_j(0) &= f_j, \end{aligned} \tag{11.2.3}$$

with boundary conditions

$$v_0 = v_N = 0. \tag{11.2.4}$$

For simplicity, we assume that all of these functions are real. Using Lemma 11.1.1,

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|v\|_{1, N-1}^2 &= (v, D_+ D_- v)_{1, N-1} = -\|D_- v\|_{2, N}^2 + v_j D_- v_j|_1^N, \\ &= -\|D_- v\|_{1, N}^2 + v_N D_- v_N - v_0 D_- v_1 = -\|D_- v\|_{1, N}^2 \leq 0. \end{aligned}$$

Thus, the approximation is stable.

From Section 6.3, the completely discretized approximation is stable if we use the backwards Euler method or the Crank–Nicholson method. One can also use the forward Euler method

$$\begin{aligned} w_j^{n+1} &= w_j^n + k D_+ D_- w_j^n, \\ w_j^0 &= f_j, \quad j = 1, 2, \dots, N - 1, \\ w_0^n &= w_N^n = 0. \end{aligned}$$

We obtain

$$\begin{aligned}\|w^{n+1}\|_{1,N-1}^2 &= \|w^n\|_{1,N-1}^2 + 2k(w^n, D_+D_-w^n)_{1,N-1} + k^2\|D_+D_-w^n\|_{1,N-1}^2, \\ &= \|w^n\|_{1,N-1}^2 - 2k\|D_-w^n\|_{1,N}^2 + k^2\|D_+D_-w^n\|_{1,N-1}^2.\end{aligned}$$

Observing that

$$\begin{aligned}\|D_+y\|_{1,N-1}^2 &= h^{-2} \sum_{j=1}^{N-1} |y_{j+1} - y_j|^2 h, \\ &\leq 2h^{-2} \sum_{j=1}^{N-1} (|y_{j+1}|^2 + |y_j|^2)h, \\ &\leq 4h^{-2}\|y\|_{1,N}^2,\end{aligned}$$

we obtain

$$\|D_+D_-w^n\|_{1,N-1}^2 \leq 4h^{-2}\|D_-w^n\|_{1,N}^2.$$

Therefore,

$$\|w^{n+1}\|_{1,N-1}^2 \leq \|w^n\|_{1,N-1}^2 - 2k\left(1 - \frac{2k}{h^2}\right)\|D_-w^n\|_{1,N}^2.$$

Thus, the approximation is stable for $2k/h^2 \leq 1$ which is the same stability limit derived in Section 2 for the periodic problem.

To discuss more general boundary conditions we need a discrete Sobolev inequality analogous to that in Lemma 9.3.1.

Lemma 11.2.1. *For any gridfunction and every $\epsilon > 0$ we have*

$$\max_{0 \leq j \leq N} |f_\nu|^2 \leq \epsilon\|D_-f\|_{1,N}^2 + C(\epsilon)\|f\|_{0,N}^2.$$

Here $C(\epsilon)$ is a constant that depends on ϵ .

Proof. The proof proceeds as in the continuous case. Let μ and ν be indices with

$$|f_\mu| = \min_{0 \leq j \leq N} |f_j|, \quad |f_\nu| = \max_{0 \leq j \leq N} |f_j|,$$

and assume, for simplicity, that $\mu \leq \nu$. By Lemma 11.1.1,

$$(f, D_+ f)_{\mu, \nu-1} = -(D_- f, f)_{\mu+1, \nu} + |f_j|^2|_\mu^\nu;$$

that is,

$$\begin{aligned} \max_{0 \leq j \leq N} |f_j|^2 &\leq \min_{0 \leq j \leq N} |f_j|^2 + \|f\|_{\mu, \nu} (\|D_+ f\|_{\mu, \nu-1} + \|D_- f\|_{\mu+1, \nu}), \\ &\leq \|f\|_{0, N}^2 + 2\|f\|_{0, N} \|D_- f\|_{1, N}, \\ &\leq \epsilon \|D_- f\|_{1, N}^2 + C(\epsilon) \|f\|_{0, N}^2, \quad C(\epsilon) = 1 + \epsilon^{-1}. \end{aligned}$$

Now we consider Eq. (11.2.1) with boundary conditions

$$u_x(\nu, t) + r_\nu u(\nu, t) = 0, \quad \nu = 0, 1. \quad (11.2.5)$$

We choose the gridpoints according to Figure 11.2.1 as

$$x_j = -\frac{1}{2}h + jh, \quad j = 0, 1, \dots, N; \quad (N-1)h = 1.$$

Then, $x_0 = -h/2$, $x_N = 1+h/2$, and we approximate Eq. (11.2.1) and (11.2.5) by

$$\frac{dv_j}{dt} = D_+ D_- v_j, \quad (11.2.6a)$$

$$v_j(0) = f_j, \quad j = 1, 2, \dots, N-1, \quad (11.2.6b)$$

$$D_+ v_0 + \frac{1}{2} r_0 (v_1 + v_0) = 0, \quad D_- v_N + \frac{1}{2} r_1 (v_N + v_{N-1}) = 0. \quad (11.2.6c)$$

As before, we can derive an energy estimate. Using the boundary conditions (11.2.6c) and Lemma 11.1.1 we get

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|v\|_{1, N-1}^2 &= (v, D_+ D_- v)_{1, N-1} = -\|D_- v\|_{2, N}^2 + v_N D_- v_N - v_1 D_- v_1, \\ &= -\|D_- v\|_{2, N}^2 - \frac{1}{2} r_1 v_N (v_N + v_{N-1}) + \frac{1}{2} r_0 v_1 (v_1 + v_0). \end{aligned}$$

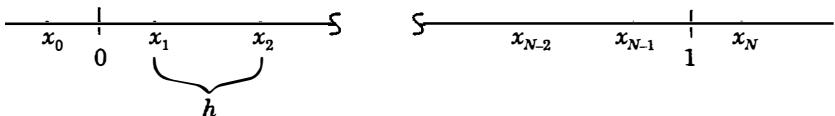


Figure 11.2.1.

Furthermore, the boundary conditions imply that

$$\begin{aligned}|v_0| &\leq \text{constant } |v_1|, \\ |v_N| &\leq \text{constant } |v_{N-1}|,\end{aligned}$$

for h sufficiently small. Thus, by Lemma 11.2.1,

$$\begin{aligned}v_0 v_1 &\leq \text{constant } |v_1|^2 \leq \epsilon \|D_- v\|_{2,N}^2 + C(\epsilon) \|v\|_{1,N}^2, \\ v_{N-1} v_N &\leq \text{constant } |v_{N-1}|^2 \leq \epsilon \|D_- v\|_{2,N}^2 + C(\epsilon) \|v\|_{1,N}^2.\end{aligned}$$

Because $\|v\|_{1,N}^2 \leq \text{constant} \|v\|_{1,N-1}^2$, we get, by choosing ϵ sufficiently small,

$$\frac{1}{2} \frac{d}{dt} \|v\|_{1,N-1}^2 \leq \text{constant} \|v\|_{1,N-1}^2,$$

and the energy estimate follows.

For time discretization we can again use forward Euler, backward Euler, or the trapezoidal rule. For general parabolic systems, such as Eqs. (9.3.8) to (9.3.10), we consider the approximation

$$\begin{aligned}\frac{dv_j}{dt} &= (A_j D_+ D_- + B_j D_0 + C_j) v_j, \quad j = 1, 2, \dots, N-1, \\ v_j(0) &= f_j, \\ R_{10} D_+ v_0 + \frac{1}{2} R_{00} (v_0 + v_1) &= 0, \quad R_{11} D_- v_N + \frac{1}{2} R_{01} (v_N + v_{N-1}) = 0.\end{aligned}\tag{11.2.7}$$

As in the continuous case, one can show that the solutions satisfy an energy estimate (Exercise 11.2.1).

Now we consider the quarter-space problem for the system (9.3.12):

$$\begin{aligned}u_t &= Bu_x + \nu u_{xx}, \quad 0 \leq x < \infty, \quad t \geq 0, \quad \nu > 0, \\ u(x, 0) &= f(x), \\ u^I(0, t) &:= (u^{(1)}(0, t), u^{(2)}(0, t), \dots, u^{(r)}(0, t))^T = 0, \\ u_x^{II}(0, t) &:= (u_x^{(r+1)}(0, t), u_x^{(r+2)}(0, t), \dots, u_x^{(m)}(0, t))^T = 0, \\ \|u(\cdot, t)\| &< \infty.\end{aligned}\tag{11.2.8}$$

Here $B = B^*$ is a constant symmetric matrix with r negative eigenvalues and

$$\langle y, By \rangle \geq 0$$

for all vectors y with $y^I = 0$. We want to show that the solutions of

$$\begin{aligned}
\frac{dv_j}{dt} &= BD_0 v_j + \nu D_+ D_- v_j, \quad j = 1, 2, \dots, \\
\frac{dv_0^{II}}{dt} &= (BD_+ v_0)^{II} + \nu D_+ D_- v_0^{II}, \\
v'_0 &= 0, \quad v_{-1}^{II} = v_1^{II}, \\
v_j(0) &= f_j,
\end{aligned} \tag{11.2.9}$$

satisfy an energy estimate. We use the scalar product

$$(f, g)_h = \frac{1}{2} \langle f_0, g_0 \rangle h + (f, g)_{1,\infty}$$

and obtain

$$\begin{aligned}
\frac{d}{dt} \|v\|_h^2 &= \frac{1}{2} \frac{d}{dt} \langle v_0, v_0 \rangle_h + \frac{d}{dt} (v, v)_{1,\infty}, \\
&= \langle v_0^{II}, BD_+ v_0^{II} \rangle_h + \nu \langle v_0^{II}, D_+ D_- v_0^{II} \rangle_h \\
&\quad + 2(v, BD_0 v)_{1,\infty} + 2\nu(v, D_+ D_- v)_{1,\infty} = I + II,
\end{aligned}$$

where

$$\begin{aligned}
I &:= \nu \langle v_0^{II}, D_+ D_- v_0^{II} \rangle_h + 2\nu(v, D_+ D_- v)_{1,\infty}, \\
&= -2\nu \langle v_0, D_+ v_0 \rangle + \nu \langle v_0^{II}, D_+ D_- v_0^{II} \rangle_h - 2\nu \|D_+ v\|_{0,\infty}^2, \\
&= -\frac{\nu}{h} \langle v_0^{II}, v_1^{II} - v_{-1}^{II} \rangle - 2\nu \|D_+ v\|_{0,\infty}^2, \\
&= -2\nu \|D_+ v\|_{0,\infty}^2 \leq 0; \\
II &:= \langle v_0, BD_+ v_0 \rangle_h + 2(v, BD_0 v)_{1,\infty}, \\
&= \langle v_0, BD_+ v_0 \rangle_h - \langle v_0, Bv_1 \rangle = -\langle v_0, Bv_0 \rangle \leq 0.
\end{aligned}$$

Thus, $(d/dt)\|v\|_h^2 \leq 0$ and the energy estimate follows.

In the previous section we demonstrated how estimates could be obtained directly for hyperbolic problems with inhomogeneous boundary conditions. Here we discuss a different and more general technique. The boundary conditions are made homogeneous by subtracting a suitable function from the solution. Consider the problem

$$\begin{aligned} u_t &= u_{xx} + F, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(0, t) &= g_0(t), \\ u(1, t) &= g_1(t), \\ u(x, 0) &= f(x). \end{aligned}$$

Let the grid be defined by

$$x_j = jh, \quad j = 0, 1, \dots, N; \quad (N - \frac{1}{2})h = 1, \quad (11.2.10)$$

and consider the semidiscrete approximation

$$\frac{dv_j}{dt} = D_+ D_- v_j + F_j, \quad j = 1, 2, \dots, N - 1, \quad (11.2.11a)$$

$$v_0(t) = g_0(t), \quad (11.2.11b)$$

$$D_- v_N(t) = g_1(t), \quad (11.2.11c)$$

$$v_j(0) = f_j. \quad (11.2.11d)$$

The location of the gridpoints makes the boundary condition (11.2.11c) center correctly. For a smooth solution $u(x, t)$ we have

$$D_- u(x_N, t) - g_1(t) = u_x(1, t) + \mathcal{O}(h^2) - g_1(t) = \mathcal{O}(h^2),$$

showing second-order accuracy. With

$$\begin{aligned} \varphi_j(t) &= (x_j - 1)^2 g_0(t) + x_j(x_j - 1)g_N(t), \\ w_j(t) &= v_j(t) - \varphi_j(t), \end{aligned} \quad (11.2.12)$$

we have

$$\begin{aligned} D_+ D_- \varphi_j(t) &= 2(g_0(t) + g_N(t)), \quad j = 1, 2, \dots, N - 1, \\ D_- \varphi_N(t) &= g_N(t). \end{aligned}$$

Therefore, w satisfies

$$\begin{aligned} \frac{dw_j}{dt} &= D_+ D_- w_j + \tilde{F}_j, \quad j = 1, 2, \dots, N - 1, \\ w_0(t) &= 0, \\ D_- w_N(t) &= 0, \\ w_j(0) &= \tilde{f}_j, \quad j = 0, 1, \dots, N, \end{aligned} \quad (11.2.13)$$

where

$$\begin{aligned}\tilde{F}_j &= F_j + 2g_0(t) + 2g_N(t) - (x_j - 1)^2 \frac{dg_0(t)}{dt} - x_j(x_j - 1) \frac{dg_N(t)}{dt}, \\ \tilde{f}_j &= f_j - \varphi_j(0).\end{aligned}$$

$D_+ D_-$ is a semibounded operator, and, as usual, we get

$$\frac{d}{dt} \|w\|_h^2 \leq (w, \tilde{F})_h + (\tilde{F}, w)_h \leq \|w\|_h^2 + \|\tilde{F}\|_h^2,$$

which yields

$$\|w(t)\|_h^2 \leq e^t \|\tilde{f}\|_h^2 + \int_0^t e^{t-\tau} \|\tilde{F}(\tau)\|_h^2 d\tau,$$

where

$$\|\tilde{F}\|_h^2 \leq \text{constant} \left(\|F\|_h^2 + |g_0|^2 + |g_N|^2 + \left| \frac{dg_0}{dt} \right|^2 + \left| \frac{dg_N}{dt} \right|^2 \right).$$

For the solution v of the original problem we get

$$\begin{aligned}\|v(t)\|_h^2 &\leq \|\varphi(t)\|_h^2 + e^t \|\tilde{f}\|_h^2 + \int_0^t e^{t-\tau} \|\tilde{F}(\tau)\|_h^2 d\tau, \\ &\leq \text{constant } e^t \left(|g_0(t)|^2 + |g_0(0)|^2 + |g_N(t)|^2 + |g_N(0)|^2 + \|f\|_h^2 \right. \\ &\quad \left. + \int_0^t \left(\|F(\tau)\|_h^2 + |g_0(\tau)|^2 + \left| \frac{dg_0(\tau)}{dt} \right|^2 \right. \right. \\ &\quad \left. \left. + |g_N(\tau)|^2 + \left| \frac{dg_N(\tau)}{dt} \right|^2 \right) d\tau \right). \tag{11.2.14}\end{aligned}$$

Since derivatives of the boundary data occur in the right-hand side, this estimate is weaker than we would like. However, in general, we cannot expect any better for Dirichlet boundary condition, because the underlying continuous problem is not strongly well posed.

EXERCISES

- 11.2.1.** Prove that the approximation (11.2.7) satisfies an energy estimate.
- 11.2.2.** Apply the forward Euler method to Eq. (11.2.6) and prove that it is stable for $k/h^2 \leq \frac{1}{2}$.

11.3. STABILITY, CONSISTENCY, AND ORDER OF ACCURACY

In this section, we discuss the above concepts in a general setting. We can proceed in the same way as we did with the periodic problem. Corresponding to Section 9.4, we consider a general system of differential equations of the form of Eq. (9.4.1) with boundary conditions (9.4.2). We introduce a grid, gridfunctions, and a discrete norm, $\|\cdot\|_h$, and approximate the continuous problem by

$$\begin{aligned} \frac{dv_j}{dt} &= Q(x_j, t, D)v_j + F_j, \quad j = 1, 2, \dots, N-1, \quad t \geq t_0, \\ v_j(t_0) &= f_j, \\ L_0(t, D)v_0(t) &= g_0(t), \quad t \geq t_0, \\ L_N(t, D)v_N(t) &= g_N(t), \quad t \geq t_0. \end{aligned} \tag{11.3.1}$$

D is an arbitrary difference operator in the x direction. For convenience, we have used the same notation for the gridfunctions F_j, f_j , and g_0 as used for the corresponding functions in the continuous problem, even though they may be different in the gridpoints.

We assume that the grid and the boundary conditions are such that v is uniquely determined by Eq. (11.3.1). The following definition corresponds to Definition 9.4.1.

Definition 11.3.1. Consider Eq. (11.3.1) with $F = g_0 = g_N = 0$. We call the approximation stable if, for all $h \leq h_0$, there are constants K and α such that, for all t_0 and all $v(t_0)$,

$$\|v(t)\|_h \leq K e^{\alpha(t-t_0)} \|v(t_0)\|_h. \tag{11.3.2}$$

The constants K and α are, in general, different from the corresponding ones for the continuous problem. The assumption of a unique solution implies that f_j be finite for every j and every fixed h . However, we may allow $|f_j| \rightarrow \infty$ as $h \rightarrow 0$ as long as $\|f\|_h < \infty$ is independent of h . For example, with the grid defined by $x_j = (j - 1/2)h$, we can handle the function $f_j = x^{-1/4}$, because it is well defined at all gridpoints and

$$\|f\|_h^2 = \sum_{j=1}^{N-1} |f_j|^2 h < \infty, \quad h \rightarrow 0.$$

Actually, we can define approximate solutions if the singularity falls on a grid-point, but we do not pursue this rather academic issue any further. If there are discontinuities in g_0 , g_N , and F_j as functions of time, generalized solutions are defined by a sequence of smooth approximations and then taken to the limit. The solution operator $S(t, t_0)$ can now be defined for the problem with $g_0 = g_N = F = 0$. It operates on all bounded gridfunctions $\{v_j\}_{j=-r+1}^{N+p-1}$ satisfying the boundary conditions, and stability is equivalent to the condition

$$\|S(t, t_0)\|_h \leq K e^{\alpha(t - t_0)}. \quad (11.3.3)$$

When treating the case $F \neq 0$, it is convenient to rewrite the approximation. We use the boundary conditions to eliminate, from the approximation, all vectors v_j , with $j \leq 0$, $j \geq N$. The resulting system has the form

$$\begin{aligned} \frac{dv_j}{dt} &= \tilde{Q}v_j + F_j, \quad j = 1, 2, \dots, N-1, \\ v_j(t_0) &= f_j, \quad j = 1, 2, \dots, N-1. \end{aligned} \quad (11.3.4)$$

A simple example of this kind of modification is

$$\frac{dv_j}{dt} = D_0 v_j + F_j, \quad j = 1, 2, \dots, N-1, \quad (11.3.5a)$$

$$v_0 - 2v_1 + v_2 = 0, \quad (11.3.5b)$$

$$v_N = 0, \quad (11.3.5c)$$

where the modified difference operator \tilde{Q} is defined by

$$\tilde{Q}v_j = \begin{cases} D_+ v_1, & j = 1, \\ D_0 v_j, & j = 2, 3, \dots, N-2, \\ -\frac{v_{N-2}}{2h}, & j = N-1. \end{cases} \quad (11.3.6)$$

The solution operator S is, of course, the same. The solution of the inhomogeneous problem with $F \neq 0$ can now be written

$$v_j(t) = S(t, t_0)f_j + \int_{t_0}^t S(t, \tau)F_j(\tau) d\tau. \quad (11.3.7)$$

By using the inequality (11.3.3), we obtain the estimate

$$\|v(t)\|_h \leq K(e^{\alpha(t-t_0)}\|f\|_h + \varphi^*(\alpha, t - t_0) \max_{t_0 \leq \tau \leq t} \|F(\tau)\|_h), \quad (11.3.8)$$

where $\varphi^*(\alpha, t)$ is defined in Eq. (4.7.2). This shows that the introduction of a forcing function does not cause any extra difficulty if the approximation is stable.

As we have seen above, the basis for stability is always a semibounded operator when using the energy method. The formal definition is given by the following definition.

Definition 11.3.2. *The discrete operator Q is semibounded if, for all gridfunctions v that satisfy the homogeneous boundary conditions $L_0 v_0 = 0$ and $L_N v_N = 0$, there is a scalar product and a norm such that the inequality*

$$\operatorname{Re}(v, Qv)_h \leq \alpha \|v\|_h^2 \quad (11.3.9)$$

holds. Here α is independent of h, x, t , and v .

Inhomogeneous boundary conditions can be treated by a change of variables, as shown in Section 11.2. We make the following assumption.

Assumption 11.3.1. *We can find a smooth function $\varphi(x, t)$ that satisfies*

$$\begin{aligned} L_0(t, D)\varphi_0(t) &= g_0(t), \\ L_N(t, D)\varphi_N(t) &= g_N(t), \\ \max_{x, t} \left(\left| \frac{\partial \varphi}{\partial t} \right| + \left| \frac{\partial^\nu \varphi}{\partial x^\nu} \right| \right) \\ &\leq c_\nu \max_t \left(|g_0| + \left| \frac{dg_0}{dt} \right| + |g_N| + \left| \frac{dg_N}{dt} \right| \right), \\ \nu &= 0, 1, \dots. \end{aligned} \quad (11.3.10)$$

Here the c_ν are constants that do not depend on h .

The gridfunction $w_j = v_j - \varphi_j$ now satisfies the original approximation (11.3.1) with $g_0 = g_N = 0$ and with a modified but bounded forcing function F_j . By Duhamel's principle, we now obtain

$$\|w(t)\|_h \leq K \left(e^{\alpha(t-t_0)} \|f\|_h + \varphi^*(\tilde{\gamma}, t - t_0) \max_{t_0 \leq \tau \leq t} \left(\|F(\tau)\|_h + |g_0(\tau)| + \left| \frac{dg_0(\tau)}{dt} \right| + |g_N(\tau)| + \left| \frac{dg_N(\tau)}{dt} \right| \right) \right). \quad (11.3.11)$$

The final estimate for $v = w + \varphi$ is then obtained by using Eq. (11.3.10).

We want to point out that this assumption imposes restrictions on the boundary data beyond the obvious smoothness requirements. As an example, consider the boundary condition

$$L_0 \varphi_0(t) := \varphi_1(t) - \varphi_0(t) = g_0(t).$$

The smoothness of $\varphi(x, t)$ implies

$$\varphi_1(t) - \varphi_0(t) = h \frac{\partial \varphi}{\partial x}(0, t) + \mathcal{O}(h^2);$$

that is, the boundary data $g_0(t)$ must satisfy $g_0(t) = \mathcal{O}(h)$.

The above construction to obtain estimates is unnecessary if the approximation is strongly stable. Corresponding to Definition 9.4.2, we make this definition.

Definition 11.3.3. *The approximation is strongly stable if it is stable and if, instead of Eq. (11.3.2), the estimate*

$$\begin{aligned} \|v(t)\|_h^2 &\leq K(t, t_0)(\|v(t_0)\|_h^2 + \max_{t_0 \leq \tau \leq t} \|F(\tau)\|_h^2 \\ &\quad + \max_{t_0 \leq \tau \leq t} (|g_0(\tau)|^2 + |g_N(\tau)|^2)) \end{aligned} \quad (11.3.12)$$

holds. Here $K(t, t_0)$ is a bounded function in any finite time interval and does not depend on the data.

All the difference approximations for hyperbolic systems that we have discussed are strongly stable. The approximation [Eqs. (11.2.3) and (11.2.4)] for the heat equation with the Dirichlet boundary condition is not strongly stable. However, if we replace the boundary condition by derivative conditions (11.2.5), then the resulting approximation (11.2.6) is strongly stable.

We now define consistency and order of accuracy of the difference approximation.

Definition 11.3.4. *Let $u(x, t)$ be a smooth solution of the differential equation. The approximation is accurate of order p , if the restriction to the grid satisfies*

the perturbed system

$$\begin{aligned} \frac{du_j}{dt} &= Q(x_j, t, D)u_j + F_j + h^p \tilde{F}_j, \quad j = 1, 2, \dots, N - 1, \\ u_j(t_0) &= f_j + h^p \tilde{f}_j, \quad j = 1, 2, \dots, N - 1, \\ L_0(t, D) &= g_0(t) + h^p \tilde{g}_0(t), \\ L_N(t, D) &= g_N(t) + h^p \tilde{g}_N(t), \end{aligned} \tag{11.3.13}$$

where $\tilde{F}_j, \tilde{f}_j, \tilde{g}_0, \tilde{g}_N, d\tilde{g}_0/dt$, and $d\tilde{g}_N/dt$ are bounded independent of h . If $p \geq 1$, then the approximation is called consistent.

Convergence of the solutions of consistent approximations follows immediately from strong stability.

Theorem 11.3.1. *If the approximation is strongly stable and accurate of order p , then*

$$\|u(\cdot, t) - v(t)\|_h \leq \text{constant } h^p$$

for smooth solutions $u(x, t)$ on any finite time interval $(0, T)$.

Proof. The error $w = u - v$ satisfies Eq. (11.3.1), where the forcing function, the initial data and the boundary data are of order h^p . Thus, the theorem follows from Eq. (11.3.12).

A similar theorem can be formulated for approximations that are stable but not strongly stable. However, the necessary procedure of subtracting a function that satisfies Assumption 11.3.1 does not always give optimal error estimates. We treat this problem in more detail in Section 12.7.

Now consider fully discrete approximations

$$\begin{aligned} Q_{-1}v_j^{n+1} &= \sum_{\sigma=0}^q Q_\sigma v_j^{n-\sigma} + kF_j^n, \quad j = 1, 2, \dots, N - 1, \\ v_j^\sigma &= f_j^\sigma, \quad \sigma = 0, 1, \dots, q, \\ L_0v_0^n &= g_0^n, \\ L_Nv_N^n &= g_N^n. \end{aligned} \tag{11.3.14}$$

Here

$$L_\nu v_\nu \equiv \sum_{\sigma=-1}^q S_{\nu\sigma} v_\nu^{n-\sigma}, \quad \nu = 0, N,$$

where the $S_{\nu\sigma}$ are the boundary operators on each time level. We assume that v^{n+1} can be solved for boundedly in terms of values at the previous $q+1$ time levels; that is, we require that there is a unique solution of

$$\begin{aligned} Q_{-1}v_j &= G_j, \quad j = 1, 2, \dots, N-1, \\ S_{0,-1}v_0 &= g_0, \\ S_{N,-1}v_N &= g_N, \end{aligned} \tag{11.3.15}$$

and that it satisfies an estimate

$$\|v\|_h \leq \text{constant} (\|G\|_h + h|g_0|^2 + h|g_N|^2). \tag{11.3.16}$$

(The factor h multiplies the boundary data because it is included in the norm $\|\cdot\|_h$.)

All the previous concepts are defined in analogy with the semidiscrete case. For example, we have the following definition.

Definition 11.3.5. *The approximation (11.3.14) is stable if, for $F^n = g_0^n = g_N^n = 0$, there are constants K and α such that, for all f^σ ,*

$$\sum_{\sigma=0}^q \|v^{n+\sigma}\|_h^2 \leq K^2 e^{2\alpha t_n} \sum_{\sigma=0}^q \|f^\sigma\|_h^2. \tag{11.3.17}$$

EXERCISES

11.3.1. Prove that

$$\begin{aligned} \frac{dv_j}{dt} &= AD_+D_-v_j + F_j, \quad j = 1, 2, \dots, N-1, \\ R_{10}D_+v_0 + \frac{1}{2}R_{00}(v_0 + v_1) &= g_0, \\ R_{11}D_-v_N + \frac{1}{2}R_{01}(v_N + v_{N-1}) &= g_N, \\ v_j(0) &= f_j, \end{aligned}$$

is strongly stable. [R_{ij} is defined in Eq. (9.3.10).] Prove that the strong estimate breaks down if $R_{10} = R_{11} = 0$.

11.3.2. Consider the approximation (11.3.14) for $g_0^n = g_N^n = 0$ and assume that it is stable. Prove that the solution satisfies the estimate

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

11.4 HIGHER ORDER APPROXIMATIONS

In Sections 11.1 and 11.2, we have seen that we can always construct difference approximations that are first- or second-order accurate. For parabolic differential equations, it is not difficult to construct higher order methods. As an example, we consider the heat equation

$$\begin{aligned} u_t &= u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{11.4.1}$$

with boundary conditions

$$\begin{aligned} u(0, t) &= 0, \\ u_x(1, t) &= 0. \end{aligned} \tag{11.4.2}$$

We choose the gridpoints according to Figure 11.4.1 as

$$x_j = jh, \quad j = -1, 0, \dots, N+1, \quad Nh = 1 + \frac{h}{2}$$

and approximate the differential equation and the boundary conditions with fourth-order accuracy by

$$\frac{dv_j}{dt} = (D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2) v_j, \quad j = 1, 2, \dots, N-1, \tag{11.4.3a}$$

$$v_j(0) = f_j, \tag{11.4.3b}$$

$$v_0(t) = 0, \tag{11.4.3c}$$

$$D_- v_N(t) - \frac{h^2}{24} D_-^3 v_{N+1}(t) = 0. \tag{11.4.3d}$$

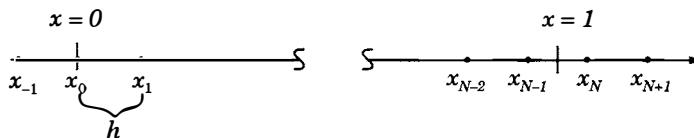


Figure 11.4.1.

A Taylor expansion about $x_{N-1/2}$ shows that the boundary condition is also fourth-order accurate. To obtain extra boundary conditions that determine v_{-1}, v_{N+1} , we differentiate the boundary conditions (11.4.2) with respect to t and replace the time derivatives by space derivatives using the differential equation.

$$\begin{aligned} u_t(0, t) &= u_{xx}(0, t) = 0, \\ u_{xt}(1, t) &= u_{xxx}(1, t) = 0. \end{aligned} \quad (11.4.4)$$

The difference approximations

$$D_+^2 v_{-1}(t) = 0, \quad (11.4.5a)$$

$$D_-^3 v_{N+1}(t) = 0 \quad (11.4.5b)$$

of Eq. (11.4.4) determine v_{-1}, v_{N+1} .

Now we derive an energy estimate. For simplicity, we assume that all functions are real and obtain

$$\frac{1}{2} \frac{d}{dt} \|v\|_{1,N-1}^2 = (v, D_+ D_- v)_{1,N-1} - \frac{1}{12} h^2 (v, D_+^2 D_-^2 v)_{1,N-1}.$$

By Lemma 11.1.1 and the boundary conditions [note that Eqs. (11.4.3d) and (11.4.5b) imply $D_- v_N = 0$];

$$\begin{aligned} (v, D_+ D_- v)_{1,N-1} &= -\|D_- v\|_{2,N}^2 + v_j D_- v_j|_1^N, \\ &= -\|D_- v\|_{1,N}^2 + v_j D_- v_j|_0^N = -\|D_- v\|_{1,N}^2, \\ (v, D_+^2 D_-^2 v)_{1,N-1} &= -(D_- v, D_+ D_-^2 v)_{2,N} + v_j D_+ D_-^2 v_j|_1^N, \\ &= -(D_- v, D_+ D_-^2 v)_{1,N-1} + v_j D_+ D_-^2 v_j|_0^N, \\ &= (D_-^2 v, D_-^2 v)_{2,N} - D_- v_j D_-^2 v_j|_1^N + v_j D_+ D_-^2 v_j|_0^N = \|D_-^2 v\|_{2,N}^2. \end{aligned}$$

Therefore,

$$\frac{1}{2} \frac{d}{dt} \|v\|_{1,N-1}^2 = -\|D_- v\|_{1,N}^2 - \frac{h^2}{12} \|D_-^2 v\|_{2,N}^2,$$

and the energy estimate follows. It is not difficult to extend these results to parabolic systems. We leave the details to the reader (Exercise 11.4.2).

The conditions (11.4.5) are only second-order approximations of Eq. (11.4.4). Yet, by using the technique to be presented in Section 12.7, it is possible to derive an $\mathcal{O}(h^4)$ error estimate (Exercise 12.7.1). For the totally discretized problem, stability of the Crank–Nicholson method follows immediately. However, we would have only second-order accuracy in time. In Section 13.2, we show that the fourth-order Runge–Kutta method yields a stable approximation.

For hyperbolic differential equations, one can use the same procedure at an inflow boundary. Consider, for example, the quarter-space problem

$$\begin{aligned}\frac{\partial u}{\partial t} &= -\frac{\partial u}{\partial x}, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x),\end{aligned}\tag{11.4.6}$$

with boundary condition

$$u(0, t) = g(t).\tag{11.4.7}$$

We choose the gridpoints $x_j = jh$, $j = -1, 0, 1, 2, \dots$, and approximate Eqs. (11.4.6) and (11.4.7) by

$$\begin{aligned}\frac{dv_j}{dt} &= -(D_0 - \frac{h^2}{6} D_0 D_+ D_-) v_j, \quad j = 1, 2, \dots, \\ v_j(0) &= f_j, \\ v_0(t) &= g(t).\end{aligned}\tag{11.4.8}$$

To obtain an extra boundary condition, which determines v_{-1} , we differentiate Eq. (11.4.7) with respect to t and replace the time derivatives by space derivatives. This gives us

$$\begin{aligned}u_x(0, t) &= -u_t(0, t) = -g_t(t), \\ u_{xxx}(0, t) &= -u_{ttt}(0, t) = -g_{ttt}(t).\end{aligned}$$

Observing that

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

we use the extra boundary condition

$$\frac{1}{2} (D_- v_0(t) + D_+ v_1(t)) \equiv D_0 v_0(t) = -\left(g_t(t) + \frac{h^2}{6} g_{ttt}(t)\right).\tag{11.4.9}$$

To prove stability, we assume that $g \equiv 0$. Again assuming that all functions are real, we obtain

$$\frac{1}{2} \frac{d}{dt} \|v\|_{1,\infty}^2 = -(v, D_0 v)_{1,\infty} + \frac{h^2}{6} (v, D_0 D_+ D_- v)_{1,\infty}.$$

By Lemma 11.1.1 and the boundary conditions,

$$(v, D_0 v)_{1,\infty} = -(D_0 v, v)_{1,\infty} - v_0 v_1 = -(D_0 v, v)_{1,\infty};$$

that is

$$(v, D_0 v)_{1,\infty} = 0.$$

Using Lemma 11.1.1 and Eqs. (11.4.8) and (11.4.9), we obtain

$$\begin{aligned} (v, D_0 D_+ D_- v)_{1,\infty} &= -(D_- v, D_0 D_- v)_{2,\infty} - v_1 D_0 D_- v_1, \\ &= -(D_- v, D_0 D_- v)_{1,\infty} - v_0 D_0 D_- v_1, \\ &= -(D_- v, D_0 D_- v)_{1,\infty} = (D_0 D_- v, D_- v)_{1,\infty} + D_- v_0 D_- v_1, \\ &= -(D_+ D_0 D_- v, v)_{1,\infty} - |D_- v_0|^2; \end{aligned}$$

that is,

$$\frac{h^2}{6} (v, D_0 D_+ D_- v)_{1,\infty} = -\frac{h^2}{12} |D_- v_0|^2.$$

Thus,

$$\frac{1}{2} \frac{d}{dt} \|v\|_{1,\infty}^2 = -\frac{h^2}{12} |D_- v_0|^2 \leq 0,$$

which shows that the approximation is stable. We could also use the trapezoidal rule leading to the Crank–Nicholson method or the fourth-order Runge–Kutta method to discretize in time.

For problems with characteristics leaving the domain, the techniques used above will not work, because there are not sufficiently many boundary conditions for the continuous system. However, a different procedure can be developed using one-sided difference operators. First, we consider the scalar model problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x}, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{11.4.10}$$

with real solutions u and the semidiscrete approximation

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \quad j = 0, 1, \dots, \quad t \geq 0, \\ v_j(0) &= f_j, \end{aligned} \tag{11.4.11}$$

where Q is a difference operator defined everywhere including the boundary points. Assume that there is a scalar product and a norm, defined by

$$\begin{aligned} (v, w)_h &= \sum_{i,j=0}^{r-1} h_{ij} v_i w_j h + \sum_{j=r}^{\infty} v_j w_j h, \\ &=: \langle v^I, H w^I \rangle_h + (v, w)_{r, \infty}, \\ \|v\|_h^2 &= (v, v)_h. \end{aligned} \quad (11.4.12)$$

Here the h_{ij} are elements of a positive-definite Hermitian $r \times r$ matrix H . Now assume that Q and H can be constructed such that

$$(v, Qv)_h = -\frac{1}{2}|v_0|^2, \quad (11.4.13)$$

for all gridfunctions v with $\|v\|_h < \infty$. The equality (11.4.13) corresponds to the equality $(u, \partial u / \partial x) = -\frac{1}{2}|u(0)|^2$ of the continuous case. In Section 11.1, it was shown that our conditions are fulfilled with the difference operator

$$Qv_j = \begin{cases} D_0 v_j, & j = 1, 2, \dots, \\ D_+ v_0, & j = 0, \end{cases} \quad (11.4.14)$$

and the scalar product

$$(v, w)_h = \frac{h}{2} v_0 w_0 + \sum_{j=1}^{\infty} v_j w_j h; \quad (11.4.15)$$

that is, $r = 1$ and $H = 1/2$.

Now we show how higher order accurate difference operators can be constructed so that Eq. (11.4.13) is satisfied. At first, it may seem too restrictive to require this strong condition, because semiboundedness would be enough for stability. However, we show how to generalize the results to systems, after which Eq. (11.4.13) is needed.

At inner points, we use the centered difference operator with order of accuracy $p = 2r$, as defined in Eq. (3.1.7). The problem is to find a way to successfully modify stencils at the points near the boundary. The order of accuracy can be relaxed by one near the boundary. This is discussed further in Section 12.7, but we demonstrate it here in a direct way for the model problem (11.4.11) with Q as defined in Eq. (11.4.14). The error $w = u - v$ satisfies

$$\begin{aligned}\frac{dw_j}{dt} &= Qw_j + F_j, \quad j = 0, 1, \dots, \\ w_j(0) &= 0,\end{aligned}$$

where

$$F_j = \begin{cases} \mathcal{O}(h), & j = 0, \\ \mathcal{O}(h^2), & j = 1, 2, \dots \end{cases}$$

With the scalar product defined by Eq. (11.4.15), we have, for any δ_1 and δ_2 greater than 0,

$$\begin{aligned}\frac{d}{dt} \|w\|_h^2 &= 2(w, Qw)_h + 2(w, F)_h, \\ &= w_0(w_1 - w_0) + \sum_{j=1}^{\infty} w_j(w_{j+1} - w_{j-1}) \\ &\quad + w_0 F_0 h + 2 \sum_{j=1}^{\infty} w_j F_j h, \\ &\leq -w_0^2 + \delta_1 w_0^2 + \frac{1}{4\delta_1} F_0^2 h^2 + \delta_2 \|w\|_h^2 + \frac{1}{\delta_2} \sum_{j=1}^{\infty} F_j^2 h.\end{aligned}$$

Choosing $\delta_1 = 1$ and integrating the last inequality gives us

$$\begin{aligned}\|w(t)\|_h^2 &\leq e^{\delta_2 t} \|f\|_h^2 + \frac{1}{\delta_2} \int_0^t e^{\delta_2(t-\tau)} \sum_{j=1}^{\infty} |F_j(\tau)|^2 h d\tau \\ &\quad + \frac{h^2}{4} \int_0^t e^{\delta_2(t-\tau)} |F_0(\tau)|^2 d\tau \\ &= \mathcal{O}(h^4).\end{aligned}$$

Thus, the approximation is second-order accurate.

We write the difference operator Q as an infinite matrix Q , partitioned as,

$$hQ = \begin{bmatrix} Q_{11} & Q_{12} \\ -C^T & D \end{bmatrix}, \quad (11.4.16)$$

where

$$\begin{aligned}
 Q_{11} &= \begin{bmatrix} q_{00} & q_{01} & \cdots & q_{0,r-1} \\ \vdots & & & \vdots \\ q_{r-1,0} & q_{r-1,1} & \cdots & q_{r-1,r-1} \end{bmatrix}, \\
 Q_{12} &= \begin{bmatrix} q_{0r} & \cdots & q_{0m} & 0 & \cdots \\ \vdots & & \vdots & \vdots & \\ q_{r-1,r} & \cdots & q_{r-1,m} & 0 & \cdots \end{bmatrix}, \\
 C &= \begin{bmatrix} 0 & 0 & \cdots \\ C_s & 0 & \cdots \end{bmatrix}, \quad \text{where } s \leq r, \\
 C_s &= \begin{bmatrix} \alpha_s & 0 & \cdots & \cdots & 0 \\ \alpha_{s-1} & \alpha_s & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \alpha_1 & \cdots & \cdots & \alpha_{s-1} & \alpha_s \end{bmatrix}, \\
 D &= \begin{bmatrix} 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & 0 & \cdots \\ -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & \ddots & \vdots \\ 0 & \ddots & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & -\alpha_s & \cdots & -\alpha_1 & 0 & \alpha_1 & \cdots & \alpha_s & 0 & \cdots \end{bmatrix}.
 \end{aligned}$$

We now have the following lemma.

Lemma 11.4.1. *The difference operator Q satisfies the relation (11.4.13) if, and only if,*

$$\begin{aligned}
 Q_{11} &= H^{-1}B, \\
 Q_{12} &= H^{-1}C,
 \end{aligned} \tag{11.4.17}$$

where B is an $r \times r$ matrix with

$$B = \text{diag}(-\frac{1}{2}, 0, \dots, 0) + B_2, \quad B_2^T = -B_2. \tag{11.4.18}$$

Proof. Let

$$v = \begin{bmatrix} v^I \\ v^{II} \end{bmatrix},$$

where

$$v^I = (v_0, v_1, \dots, v_{r-1})^T, \quad v^{II} = (v_r, v_{r+1}, \dots)^T,$$

and use $\langle \cdot, \cdot \rangle$ as notation for the usual Euclidean scalar product. Since D is anti-symmetric, Eq. (11.4.13) can be written as

$$-\frac{1}{2}|v_0|^2 = \langle v^I, HQ_{11}v^I \rangle + \langle v^I, HQ_{12}v^{II} \rangle - \langle v^{II}, C^T v^I \rangle.$$

The special case $v^{II} = 0$ shows that $HQ_{11} = B$ must have the form of Eq. (11.4.18). Thus,

$$0 = \langle v^I, HQ_{12}v^{II} \rangle - \langle v^{II}, C^T v^I \rangle = \langle v^I, (HQ_{12} - C)v^{II} \rangle,$$

for all vectors v^I and v^{II} . This is only possible if $HQ_{12} = C$ which proves the lemma.

For a given $2s$ order accurate approximation at inner points, the matrix C is determined. The remaining question is whether a positive definite matrix H and an “almost” antisymmetric matrix B , satisfying Eq. (11.4.18), can be found so that, for smooth functions $u(x)$ and $\tau = 2s - 1$,

$$H^{-1}B \begin{bmatrix} u(x_0) \\ u(x_1) \\ \vdots \\ u(x_{r-1}) \end{bmatrix} + H^{-1}C \begin{bmatrix} u(x_r) \\ u(x_{r+1}) \\ \vdots \\ u(x_{r-1}) \end{bmatrix} = h \begin{bmatrix} u_x(\xi_0) \\ u_x(\xi_1) \\ \vdots \\ u_x(\xi_{r-1}) \end{bmatrix} + \mathcal{O}(h^{\tau+1}), \quad (11.4.19)$$

for some points ξ_j . It is sufficient to consider polynomials of the form $u(x) = (x - x_r)^\nu$ and assume that $h = 1$. The approximation is accurate of order τ if

$$H^{-1}B(-1)^\nu \begin{bmatrix} r^\nu \\ (r-1)^\nu \\ \vdots \\ 1^\nu \end{bmatrix} + H^{-1}C \begin{bmatrix} 0^\nu \\ 1^\nu \\ 2^\nu \\ \vdots \end{bmatrix} = \nu(-1)^{\nu-1} \begin{bmatrix} r^{\nu-1} \\ (r-1)^{\nu-1} \\ \vdots \\ 1^{\nu-1} \end{bmatrix}, \quad (11.4.20)$$

where $\nu = 0, 1, \dots, \tau$, and we have adopted the convention that $0^0 = 1$. For approximations of systems, we need a special form of H :

$$H = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ 0 & h_{11} & \cdots & h_{1,r-1} \\ \vdots & \vdots & & \vdots \\ 0 & h_{1,r-1} & \cdots & h_{r-1,r-1} \end{bmatrix} > 0. \quad (11.4.21)$$

One can now prove the following theorem.

Theorem 11.4.1. *For every order of accuracy $2s$ in the interior, there are boundary approximations accurate of order $\tau = 2s - 1$ such that Eq. (11.4.13) is satisfied, where Q has the form of Eq. (11.4.16) and the matrix H has the form of Eq. (11.4.21).*

We can solve Eq. (11.4.20) using a symbolic manipulation system. In general, one must choose $r \geq \tau + 2$, but the operator is not uniquely determined. For $\tau = 3$, $r = 5$, there is a three-parameter solution. These parameters can be chosen so that the bandwidth of the operator is minimized. This is convenient for implementation, in particular, on parallel computers. The resulting operator Q has the structure

$$hQ = \left[\begin{array}{cccccc} \otimes & \times & \times & \times & & & \\ \times & \otimes & \times & \times & \times & \times & \\ \times & \times & \otimes & \times & \times & \times & \\ \times & \times & \times & \otimes & \times & \times & \times \\ \times & \times & \times & \times & \otimes & \times & \times \\ & & & & & \times & \times & \otimes & \times & \times \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{array} \right], \quad (11.4.22)$$

where $\tau = 3$, $2s = 4$, and H of the type shown by Eq. (11.4.21).

For systems in nonsmooth multidimensional domains, the stability proof can be generalized if the matrix H , defining the norm, is further restricted to diagonal form. In this case, the accuracy near the boundary cannot be as large as order $2s - 1$, except for the case $s = 1$. However, one can prove the next theorem.

Theorem 11.4.2. *For interior accuracy of order $2s$ with $1 \leq s \leq 4$, there are boundary approximations accurate of order s such that Eq. (11.4.13) is satisfied with Q of the form shown in Eq. (11.4.16) and H diagonal in the scalar product (11.4.12).*

As will be shown later, we may expect an overall accuracy of order $s + 1$. The system (11.4.20) can again be computed using a symbolic manipulation program. The case $s = 1$ has already been presented in Section 11.1. For $s = 2$,

the solution is uniquely determined. The structure of the difference operator Q is

$$hQ = \begin{bmatrix} \otimes & \times & \times & \times \\ \times & \otimes & \times & \\ \times & \times & \otimes & \times & \times \\ \times & \times & \times & \otimes & \times & \times \\ & & & \times & \otimes & \times & \times \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \quad (11.4.23)$$

where $\tau = 2$, $2s = 4$, and H diagonal, giving an overall third-order accurate approximation.

The case $s = 3$ leads to a one parameter solution. This parameter can be chosen to minimize the bandwidth of Q . The structure is

$$hQ = \begin{bmatrix} \otimes & \times & \times & \times & \times \\ \times & \otimes & \times & \times & \times & \times \\ \times & \times & \otimes & \times & \times & \times \\ \times & \times & \times & \otimes & \times & \times & \times \\ \times & \times & \times & \times & \otimes & \times & \times & \times \\ & & & & \times & \otimes & \times & \times & \times \\ & & & & \times & \times & \otimes & \times & \times & \times \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & & \ddots \end{bmatrix}, \quad (11.4.24)$$

where $\tau = 3$, $2s = 6$, and H diagonal, giving an overall fourth-order accurate solution.

For implementation on SIMD computers, the number of nonzero diagonals essentially determines the computing time, even if there are only a few nonzero elements in a particular diagonal. It is interesting to note that the bandwidth is nine for both Eq. (11.4.22) and Eq. (11.4.24), but the latter can be expected to give better results in practice, because it is more accurate in the interior. (The exact values of the elements of Q are given in Appendix 11.A for the three cases above.)

For the scalar model problem, the relation (11.4.13) immediately leads to stability, because Q is semibounded. The inflow problem has already been treated, and, in principle, we can now put these two procedures together to obtain high-order methods. However, we shall present a more convenient way of solving systems by using the outflow operator Q developed above, for all of the variables in the system.

To begin, we consider systems

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x}, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (11.4.25a)$$

$$u(x, 0) = f(x), \quad (11.4.25b)$$

with boundary conditions

$$u^I(0, t) = 0. \quad (11.4.25c)$$

Here the constant symmetric matrix A has q negative eigenvalues and $u^I = (u^{(1)}, u^{(2)}, \dots, u^{(q)})^T$. We assume that there is an energy estimate for Eq. (11.4.25); that is,

$$\langle w, Aw \rangle \geq 0, \quad (11.4.26)$$

for all vectors $w = (w^I, w^{II})^T$ with $w^I = 0$.

We obtain an approximation by replacing $\partial/\partial x$ by the difference operator Q described above. Let T be the projection operator, defined by

$$Tv_j = \begin{cases} v_j, & j = 1, 2, \dots, \\ (0, v_0^{II})^T, & j = 0. \end{cases} \quad (11.4.27a)$$

Then the semidiscrete approximation can be written in the form

$$\begin{aligned} \frac{dv_j}{dt} &= TAQv_j, \quad j = 0, 1, \dots, \\ v_j(0) &= Tf_j. \end{aligned} \quad (11.4.27b)$$

In an actual fully discrete implementation, the whole solution is advanced at all points, including x_0 , at each time step, and the physical boundary conditions (11.4.25c) are imposed before the next step. The scalar product is generalized to vector gridfunctions by

$$(v, w)_h = \sum_{i,j=0}^{r-1} h_{ij} \langle v_i, w_j \rangle_h + \sum_{j=r}^{\infty} \langle v_j, w_j \rangle_h, \quad (11.4.28)$$

and the relation (11.4.13) is also well defined for vector gridfunctions.

To prove stability we need two lemmas.

Lemma 11.4.2. *The property (11.4.13) holds if, and only if,*

$$(v, Qw)_h = -(Qv, w)_h - \langle v_0, w_0 \rangle, \quad (11.4.29)$$

for all real $v, w \in l_2(0, \infty)$.

Proof. Obviously, Eq. (11.4.13) follows from Eq. (11.4.29) by letting $v = w$. Furthermore, if Eq. (11.4.13) holds, then

$$(v + w, Q(v + w))_h = -\frac{1}{2}|v_0 + w_0|^2,$$

that is,

$$(v, Qv)_h + (w, Qw)_h + (v, Qw)_h + (w, Qv)_h = -\frac{1}{2}(|v_0|^2 + |w_0|^2 + 2\langle v_0, w_0 \rangle),$$

and Eq. (11.4.29) follows from Eq. (11.4.13).

Lemma 11.4.3. *Assume that the matrix H , defining the scalar product (11.4.28), has the restricted form (11.4.21). Then the projection operator T is self-adjoint.*

Proof. The scalar product can be written as

$$(v, w)_h = h_0 \langle v_0, w_0 \rangle_h + R(v, w),$$

where $R(v, w)$ does not depend on v_0, w_0 . Thus, for any v, w with $\|v\|_h < \infty, \|w\|_h < \infty$,

$$\begin{aligned} (v, Tw)_h &= h_0 \langle v_0, (Tw)_0 \rangle_h + R(v, Tw), \\ &= h_0 \langle v_0^{II}, w_0^{II} \rangle_h + R(v, w), \\ &= h_0 \langle (Tv)_0, w_0 \rangle_h + R(Tv, w) = (Tv, w)_h, \end{aligned}$$

which proves the lemma.

We also need the following lemma.

Lemma 11.4.4. Let A be a constant Hermitian matrix and let $(v, w)_h$ be defined by Eq. (11.4.28). Then

$$(Av, w)_h = (v, Aw)_h.$$

Proof. From the definition of the scalar product, we have

$$\begin{aligned} (Av, w)_h &= \sum_{i,j=0}^{r-1} h_{ij} \langle Av_i, w_j \rangle_h + \sum_{j=r}^{\infty} \langle Av_j, w_j \rangle_h, \\ &= \sum_{i,j=0}^{r-1} h_{ij} \langle v_i, Aw_j \rangle_h + \sum_{j=r}^{\infty} \langle Av_j, w_j \rangle_h, \\ &= (v, Aw)_h. \end{aligned}$$

Now we can prove stability.

Theorem 11.4.3. Assume that the matrix H defining the scalar product (11.4.28) has the restricted form of Eq. (11.4.21). If Q satisfies the equality (11.4.13), then the approximation (11.4.27) is stable.

Proof. The solution of Eq. (11.4.27) satisfies $v = Tv$, and, because T is a projection operator, we have $T^2 = T$. Then using Lemmas 11.4.2 to 11.4.4,

$$\begin{aligned} \frac{d}{dt} \|v\|_h^2 &= (v, TAQv)_h + (TAQv, v)_h, \\ &= (v, TAQTv)_h + (TAQTv, v)_h, \\ &= (v, TAQTv)_h + (QTv, ATv)_h, \\ &= (v, TAQTv)_h - (v, TQATv)_h - \langle (Tv)_0, A(Tv)_0 \rangle. \end{aligned}$$

Because A and Q commute, the first two terms cancel, and the theorem follows from Eq. (11.4.26).

For more general boundary conditions of the type of Eq. (11.1.22), we use the same technique as in Section 11.1. A unitary transformation is used to transform the boundary conditions to the form of Eq. (11.4.25c).

Time discretization can be constructed as in Section 11.1. For higher order accuracy, one can use Runge-Kutta methods or some other ODE approximation. We come back to these methods in Section 13.2.

Throughout this section, we have restricted our discussion to constant coefficient problems for convenience. No new difficulties arise from the presence of boundaries when treating problems with variable coefficients. As demonstrated

in Section 11.1, extra terms occur in the stability estimates, just as for the pure initial value problem treated in Section 5.3. For example, we get

$$\frac{d}{dt} \|v\|_h^2 \leq \alpha \|v\|_h^2,$$

because $AQ - QA \neq 0$ if $A = A(x_j)$ depends on x .

EXERCISES

- 11.4.1.** Derive a fourth-order accurate approximation like Eq. (11.4.3) for the nonlinear problem

$$\begin{aligned} u_t &= a(u)u_{xx} + b(u)u_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ u(0, t) &= g_0(t), \\ u_x(1, t) &= g_1(t), \\ u(x, 0) &= f(x), \end{aligned}$$

where $a(u) \geq \delta > 0$.

- 11.4.2.** Derive a fourth-order accurate approximation like Eq. (11.4.3) for the problem (9.3.8) to (9.3.10). Prove that it is stable.
- 11.4.3.** Explicitly verify that Eq. (11.4.20) is satisfied for Q defined in Eq. (11.4.14), and the scalar product Eq. (11.4.15).

11.5. SEVERAL SPACE DIMENSIONS

We consider a two-dimensional PDE of the form

$$\begin{aligned} \frac{\partial u}{\partial t} &= P_1 \left(\frac{\partial}{\partial x} \right) u + P_2 \left(\frac{\partial}{\partial y} \right) u + F, \\ 0 \leq x \leq 1, \quad -\infty < y < \infty, \quad t \geq 0, \end{aligned} \tag{11.5.1}$$

with 2π -periodic solutions in the y direction. The grid is defined on the domain $0 \leq x \leq 1$, $0 \leq y \leq 2\pi$ by

$$\Omega_h = (x_i, y_j) = (ih_1, jh_2), \quad i = 0, 1, \dots, M, \quad j = 1, 2, \dots, N. \tag{11.5.2}$$

The scalar product and norm are defined by

$$\begin{aligned}
 (v, w)_h &= \sum_{j=1}^N (v_{\cdot j}, w_{\cdot j})_{h_1} h_2 = \sum_{i=0}^M d_i (v_{i \cdot}, w_{i \cdot})_{h_2} h_1 \\
 &= \sum_{j=1}^N \sum_{i=0}^M \langle v_{ij}, d_i w_{ij} \rangle_{h_1} h_2, \quad \|v\|_h^2 = (v, v)_h. \quad (11.5.3)
 \end{aligned}$$

We assume that $d_i > 0$ are scalars; that is, in the x direction we use a scalar product with a diagonal matrix H , as discussed in Section 11.4.

The semidiscrete approximation is

$$\frac{dv_{ij}}{dt} = Q_1 v_{ij} + Q_2 v_{ij}, \quad (i, j) \in \tilde{\Omega}_h, \quad t \geq 0, \quad (11.5.4a)$$

$$L_0 v_0 = g_0, \quad j = 1, 2, \dots, N, \quad (11.5.4b)$$

$$L_M v_M = g_M, \quad j = 1, 2, \dots, N, \quad (11.5.4c)$$

$$v_{ij} = v_{i,j+N}, \quad (i, j) \in \Omega_h, \quad (11.5.4d)$$

$$v_{ij}(0) = f_{ij}, \quad (i, j) \in \Omega_h. \quad (11.5.4e)$$

Here $\tilde{\Omega}_h$ denotes the inner points defined such that Eq. (11.5.4a) is well-defined. Assume that the one-dimensional operators Q_1 and Q_2 are semibounded:

$$\operatorname{Re} (v, Q_1 v)_{h_1} \leq \alpha_1 \|v\|_{h_1}^2, \quad (11.5.5)$$

for v satisfying $L_0 v_0 = 0$, $L_M v_M = 0$, and

$$\operatorname{Re} (w, Q_2 w)_{h_2} \leq \alpha_2 \|w\|_{h_2}^2, \quad (11.5.6)$$

for w periodic. We have

$$\begin{aligned}
 \operatorname{Re} (v, (Q_1 + Q_2)v)_h &= \operatorname{Re} \sum_{j=1}^N (v_{\cdot j}, Q_1 v_{\cdot j})_{h_1} h_2 + \operatorname{Re} \sum_{i=0}^M d_i (v_{i \cdot}, Q_2 v_{i \cdot})_{h_2} h_1, \\
 &\leq \alpha_1 \sum_{j=1}^N \|v_{\cdot j}\|_{h_1}^2 h_2 + \alpha_2 \sum_{i=0}^M d_i \|v_{i \cdot}\|_{h_2}^2 h_1, \\
 &\leq \alpha_3 \|v\|_h^2.
 \end{aligned}$$

The concept of semiboundedness is generalized in an obvious way to several space dimensions. We have just shown that $Q_1 + Q_2$ is semibounded. As an

example, consider the well-posed problem

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 2\pi, \quad t \geq 0, \\ u(1, y, t) &= 0, \\ u(x, y, 0) &= f(x, y),\end{aligned}\tag{11.5.7}$$

with real solutions that are periodic in y . As an approximation we use

$$\begin{aligned}\frac{dv_{ij}}{dt} &= Q_1 v_{ij} + D_{0y} v_{ij}, \quad i = 0, 1, \dots, M - 1, \\ v_{iM} &= 0, \\ v_{ij} &= v_{i,j+N}, \\ v_{ij}(0) &= f_{ij}, \quad i = 0, 1, \dots, M,\end{aligned}\tag{11.5.8}$$

for $j = 1, 2, \dots, N$.

Here Q_1 is defined as Q in Eq. (11.4.14) and the scalar product $(\cdot, \cdot)_{h_1}$ is defined by Eq. (11.4.15). Since

$$(v_{\cdot j}, Q_1 v_{\cdot j})_{h_1} = -\frac{1}{2} |v_{0j}|^2, \quad j = 1, 2, \dots, N,\tag{11.5.9}$$

and

$$(w_i, D_{0y} w_{i\cdot})_{h_2} = 0, \quad i = 0, 1, \dots, M,\tag{11.5.10}$$

we get

$$(v, (Q_1 + D_{0y})v)_h = -\frac{1}{2} \sum_{j=1}^N |v_{0j}|^2 h_2,$$

proving stability.

The assumption of periodic solutions in the y direction makes the generalization to two space dimensions particularly simple. However, the introduction of boundaries in the y direction does not change much. Consider the approximation (11.5.4a) in the domain $0 \leq x \leq 1$ with boundary conditions

$$\begin{aligned}
 L_{x0}v_{0j} &= g_{0j}, \\
 L_{xM}v_{Mj} &= g_{Mj}, \\
 L_{y0}v_{i0} &= g_{i0}, \\
 L_{yN}v_{iN} &= g_{iN}.
 \end{aligned} \tag{11.5.11}$$

The scalar product is changed to

$$\begin{aligned}
 (v, w)_h &= \sum_j e_j(v_{\cdot j}, w_{\cdot j})_{h_1} h_2 = \sum_i d_i(v_{i \cdot}, w_{i \cdot})_{h_2} h_1, \\
 &= \sum_i \sum_j d_i e_j(v_{ij}, w_{ij}) h_1 h_2.
 \end{aligned} \tag{11.5.12}$$

The assumptions (11.5.5) and (11.5.6) of one-dimensional semiboundedness are modified in an obvious way, and stability follows as in the one-dimensional case.

For systems, we use the same technique as in Section 11.4. The solution is advanced at all points including the boundaries, and the boundary conditions are imposed via projections (in each step for the fully discrete case). Stability is proven as in the one-dimensional case.

For higher order approximations, the same technique is applied. If the scalar product has the form of Eq. (11.5.12), that is, if the matrix H in Eq. (11.4.12) is diagonal, then the one-dimensional stability results generalize in a straightforward manner to several space dimensions. For nondiagonal matrices H , the presence of corners in the computational domain complicates the analysis, and no general results are known for this case.

Time discretization can be done as before. However, the implicit trapezoidal and backward Euler methods result in large systems to be solved at each time step. It has been shown how this can be avoided for periodic problems by introducing ADI-methods like

$$\begin{aligned}
 \left(I - \frac{k}{2} Q_1 \right) v^{n+1/2} &= \left(I + \frac{k}{2} Q_2 \right) v^n, \\
 \left(I - \frac{k}{2} Q_2 \right) v^{n+1} &= \left(I + \frac{k}{2} Q_1 \right) v^{n+1/2}.
 \end{aligned} \tag{11.5.13}$$

We assume that Q_1 and Q_2 are semibounded, and, for convenience, it is assumed that the growth constants are zero:

$$\operatorname{Re}(v, Q_\nu v)_h \leq 0, \quad \nu = 1, 2. \tag{11.5.14}$$

Theorem 11.5.1. *The ADI-scheme (11.5.13) is stable if Q_1 and Q_2 are semi-bounded and kQ_1 and kQ_2 are bounded.*

Proof. If Eq. (11.5.14) is satisfied, then

$$\begin{aligned}
 \|v^{n+1}\|_h^2 &+ \left\| \frac{k}{2} Q_2 v^{n+1} \right\|_h^2 \\
 &\leq \|v^{n+1}\|_h^2 + \left\| \frac{k}{2} Q_2 v^{n+1} \right\|_h^2 - 2\operatorname{Re}(v^{n+1}, kQ_2 v^{n+1})_h, \\
 &= \left\| \left(I - \frac{k}{2} Q_2 \right) v^{n+1} \right\|_h^2 = \left\| \left(I + \frac{k}{2} Q_1 \right) v^{n+1/2} \right\|_h^2, \\
 &\leq \left\| \left(I - \frac{k}{2} Q_1 \right) v^{n+1/2} \right\|_h^2 = \left\| \left(I + \frac{k}{2} Q_2 \right) v^n \right\|_h^2, \\
 &\leq \|v^n\|_h^2 + \left\| \frac{k}{2} Q_2 v^n \right\|_h^2.
 \end{aligned}$$

Thus,

$$\begin{aligned}
 \|v^{n+1}\|_h^2 &\leq \|v^{n+1}\|_h^2 + \left\| \frac{k}{2} Q_2 v^{n+1} \right\|_h^2, \\
 &\leq \|v^0\|_h^2 + \left\| \frac{k}{2} Q_2 v^0 \right\|_h^2 \leq \text{constant} \|v^0\|_h^2,
 \end{aligned}$$

which proves stability. [Note that the scheme is almost unconditionally stable. The requirement that kQ_1 and kQ_2 be bounded means that $k = \mathcal{O}(h^p)$ where p is the order of the differential operator in space.]

These results can be generalized to any domain with a structured grid that can be transformed by a smooth transformation to a rectangle with a uniform grid. By using overlapping subgrids, we can solve most problems arising in practice. We come back to this technique in Section 13.3.

Sometimes it may be convenient to use one single grid, even if the domain cannot be transformed into a rectangle. For example, there may be concave corners as in L-shaped domains. It is also possible to construct semibounded operators in this case. Assume that we want to solve

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} \tag{11.5.15}$$

in the domain shown in Figure 11.5.1.

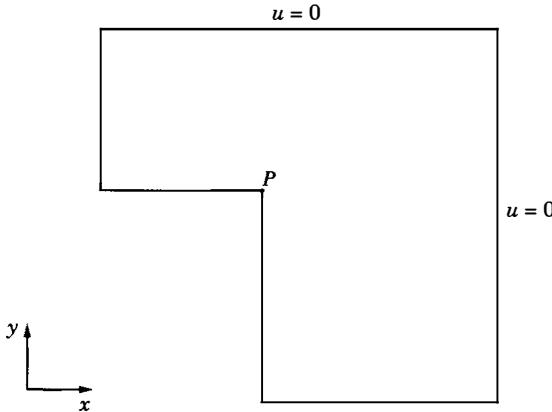


Figure 11.5.1 Domain with 90-degree concave corner.

Zero boundary values are given on the right and upper boundaries. At all interior points, we use standard second-order centered difference operators. At all outflow boundary points, except the corner P , one-sided first-order operators are used. If the uniform grid is such that the point P is (x_m, y_n) and $h_1 = h_2$, we use

$$\begin{aligned} u_x(P) &\rightarrow \frac{1}{3h} (2v_{m+1,n} - v_{mn} - v_{m-1,n}), \\ u_y(P) &\rightarrow \frac{1}{3h} (2v_{m,n+1} - v_{mn} - v_{m,n-1}). \end{aligned} \quad (11.5.16)$$

One can prove that the resulting operators are semibounded with the scalar product (defined in order from below and upward in Ω_h)

$$\begin{aligned} h^{-2}(v, w)_h &= \frac{1}{2} \left(\frac{1}{2} v_{m0} w_{m0} + \sum_{i=m+1}^{N-1} v_{i0} w_{i0} \right) \\ &+ \sum_{j=1}^{n-1} \left(\frac{1}{2} v_{mj} w_{mj} + \sum_{i=m+1}^{N-1} v_{ij} w_{ij} \right) \\ &+ \frac{1}{2} \left(\frac{1}{2} v_{0n} w_{0n} + \sum_{i=1}^{m-1} v_{in} w_{in} \right) + \frac{3}{4} v_{mn} w_{mn} + \sum_{i=m+1}^{N-1} v_{in} w_{in} \\ &+ \sum_{j=n+1}^{N-1} \left(\frac{1}{2} v_{0j} w_{0j} + \sum_{i=1}^{N-1} v_{ij} w_{ij} \right) \\ &=: \sum_{i,j} d_{ij} v_{ij} w_{ij}. \end{aligned} \quad (11.5.17)$$

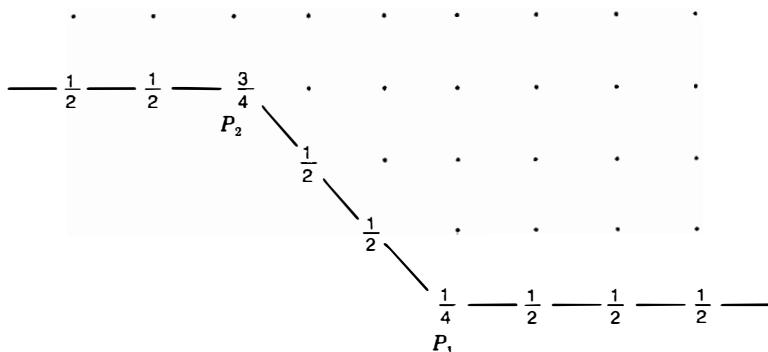


Figure 11.5.2 Domain and weighting coefficients for 45-degree corners.

The weights d_{ij} are $\frac{1}{4}$ at convex corners, $\frac{3}{4}$ at concave corners, and $\frac{1}{2}$ at all other outflow boundary points. In fact, this scalar product can be used also for 45 degree corners, as shown in Figure 11.5.2.

Assume that the convex corner is located at $P_1 = (x_l, y_0)$ and the concave corner at $P_2 = (x_m, y_n)$. Then the difference operators are semibounded with

$$\begin{aligned} u_x(P_1) &= \frac{1}{h} (v_{l+1,0} - v_{l,0}), \\ u_y(P_1) &= \frac{1}{2h} (4v_{l1} - 2v_{l,0} - v_{l-2,2} - v_{l+2,0}), \\ u_x(P_2) &= \frac{1}{3h} (2v_{m+1,n} - v_{mn} - v_{m-1,n}), \\ u_y(P_2) &= \frac{1}{6h} (4v_{m,n+1} - 2v_{mn} - v_{m-2,n} - v_{m+2,n-2}). \end{aligned} \quad (11.5.18)$$

For problems in domains with smooth curved boundaries, we use overlapping grids in close analogy with the principle used in Section 9.6 for proving well-posedness. This technique will be considered in Section 13.3.

EXERCISES

11.5.1. Consider the problem

$$\begin{aligned} u_t &= Au_x + Bu_y, \quad 0 \leq x, y \leq 1, \quad t \geq 0, \\ u^I(0, y, t) &= 0, \\ u^{II}(1, y, t) &= 0, \\ u^{III}(x, 0, t) &= 0, \\ u^{IV}(x, 1, t) &= 0, \\ u(x, y, 0) &= f(x, y), \end{aligned}$$

where

$$\begin{aligned}\langle u, Au \rangle_{u' = 0} &\geq 0, & \langle u, Au \rangle_{u'' = 0} &\leq 0, \\ \langle u, Bu \rangle_{u''' = 0} &\geq 0, & \langle u, Bu \rangle_{u'''' = 0} &\leq 0.\end{aligned}$$

Define the approximation corresponding to Eq. (11.4.27) with Q based on operators like Eq. (11.4.23) or Eq. (11.4.24) satisfying Eq. (11.4.13). Prove stability.

- 11.5.2.** Consider the fractional step method

$$(I - kQ_1)(I - kQ_2)v^{n+1} = v^n,$$

with boundary conditions such that Q_1 and Q_2 both are semibounded. Prove unconditional stability.

- 11.5.3.** Prove that the difference approximation discussed above for $u_t = u_x + u_y$ in the domain shown in Figure 11.5.1 is stable by using the scalar product (11.5.17).
- 11.5.4.** Prove stability as in Exercise 11.5.3, but with an outflow boundary with corners as in Figure 11.5.2.
- 11.5.5.** Consider the problem

$$\begin{aligned}u_t &= u_{xx} + u_{yy}, & (x, y) &\in \Omega, \quad t \geq 0, \\ \frac{\partial u}{\partial n} &= 0, & (x, y) &\in \partial\Omega, \\ u(x, y, 0) &= f(x, y),\end{aligned}$$

where Ω is the unit square. Construct stable and second-order accurate boundary conditions for the approximation

$$\frac{dv}{dt} = D_{+x}D_{-x}v + D_{+y}D_{-y}v.$$

BIBLIOGRAPHIC NOTES

The general procedure for deriving higher order accurate difference approximations for hyperbolic problems, as described in Section 11.4, was first presented by Kreiss and Scherer (1974, 1977). The computation of the difference operators shown in Appendix 11.A was done by Strand (1994). A general stability proof for systems in multidimensional nonsmooth domains with nonorthogonal grids was given by Olsson (1995a,b). In the same papers, it is also shown that the general form (11.4.12) of H can be used also for systems, by letting the difference approximation depend on H near the boundaries.

The construction of the difference approximations (11.5.16) and (11.5.18) at concave and 45-degree corners was done by Engquist (1978).

For implicit compact difference approximations of the type (3.1.29), where P is nondiagonal, the restrictions imposed by the boundaries are more severe than for the explicit ones. It is shown in Gottlieb et al. (1993) that the local accuracy near the boundary can be at most first order if the condition (11.4.13) is to be satisfied with a locally modified norm. In Carpenter, Gottlieb, and Abarbanel (1994), this difficulty is avoided by introducing a globally modified norm, in the same paper it is shown how to implement the boundary conditions as a penalty function in the difference approximation.

APPENDIX TO CHAPTER 11

Matrix elements near the boundary for the matrix hQ in Eqs. (11.4.22) to (11.4.24).

Matrix (11.4.23): $\tau = 2, 2s = 4, H$ diagonal.

$$\begin{array}{lll} d_{11} = -24/17 & d_{12} = 59/34 & d_{13} = -4/17 \\ d_{14} = -3/34 & d_{15} = 0 & d_{16} = 0 \\ d_{21} = -1/2 & d_{22} = 0 & d_{23} = 1/2 \\ d_{24} = 0 & d_{25} = 0 & d_{26} = 0 \\ d_{31} = 4/43 & d_{32} = -59/86 & d_{33} = 0 \\ d_{34} = 59/86 & d_{35} = -4/43 & d_{36} = 0 \\ d_{41} = 3/98 & d_{42} = 0 & d_{43} = -59/98 \\ d_{44} = 0 & d_{45} = 32/49 & d_{46} = -4/49 \end{array}$$

Inner points:

$$-\alpha_{-2} = \alpha_2 = -1/12, \quad -\alpha_{-1} = \alpha_1 = 2/3.$$

Matrix (11.4.24): $\tau = 3, 2s = 6, H$ diagonal.

$$\begin{array}{lll} d_{11} = -21600/13649 & d_{12} = 81763/40947 & d_{13} = 131/27298 \\ d_{14} = -9143/13649 & d_{15} = 20539/81894 & d_{16} = 0 \\ d_{17} = 0 & d_{18} = 0 & d_{19} = 0 \\ d_{21} = -81763/180195 & d_{22} = 0 & d_{23} = 7357/36039 \\ d_{24} = 30637/72078 & d_{25} = -2328/12013 & d_{26} = 6611/360390 \\ d_{27} = 0 & d_{28} = 0 & d_{29} = 0 \\ d_{31} = -131/54220 & d_{32} = -7357/16266 & d_{33} = 0 \\ d_{34} = 645/27111 & d_{35} = 11237/32532 & d_{36} = -3487/27110 \end{array}$$

$$\begin{array}{lll}
 d_{37} = 0 & d_{38} = 0 & d_{39} = 0 \\
 d_{41} = 9143/53590 & d_{42} = -30637/64308 & d_{43} = -645/5359 \\
 d_{44} = 0 & d_{45} = 13733/32154 & d_{46} = -67/4660 \\
 d_{47} = 72/5359 & d_{48} = 0 & d_{49} = 0 \\
 d_{51} = -20539/236310 & d_{52} = 2328/7877 & d_{53} = -11237/47262 \\
 d_{54} = 13733/23631 & d_{55} = 0 & d_{56} = 89387/118155 \\
 d_{57} = -1296/7877 & d_{58} = 144/7877 & d_{59} = 0 \\
 d_{61} = 0 & d_{62} = -6611/262806 & d_{63} = 3487/43801 \\
 d_{64} = 1541/87602 & d_{65} = -89387/131403 & d_{66} = 0 \\
 d_{67} = 32400/43801 & d_{68} = -6480/43801 & d_{69} = 720/43801
 \end{array}$$

Inner points:

$$-\alpha_{-3} = \alpha_3 = 1/60, \quad -\alpha_{-2} = \alpha_2 = -3/20, \quad -\alpha_{-1} = \alpha_1 = 3/4.$$

Matrix (11.4.22): $\tau = 3, 2s = 4, H$ of type (11.4.21).

$$\begin{aligned}
 d_{11} &= -11/6 \\
 d_{12} &= 3 \\
 d_{13} &= -3/2 \\
 d_{14} &= 1/3 \\
 d_{15} &= 0 \\
 d_{16} &= 0 \\
 d_{17} &= 0 \\
 f_1 d_{21} &= -24(-779042810827742869 + 104535124033147\sqrt{26116897}) \\
 f_1 d_{22} &= -(-176530817412806109689 \\
 &\quad + 29768274816875927\sqrt{26116897})/6 \\
 f_1 d_{23} &= 343(-171079116122226871 + 27975630462649\sqrt{26116897}) \\
 f_1 d_{24} &= -3(-7475554291248533227 + 1648464218793925\sqrt{26116897})/2 \\
 f_1 d_{25} &= (-2383792768180030915 + 1179620587812973\sqrt{26116897})/3 \\
 f_1 d_{26} &= -1232(-115724529581315 + 37280576429\sqrt{26116897}) \\
 d_{27} &= 0 \\
 f_2 d_{31} &= -12(-380966843 + 86315\sqrt{26116897}) \\
 f_2 d_{32} &= (5024933015 + 2010631\sqrt{26116897})/3 \\
 f_2 d_{33} &= -231(-431968921 + 86711\sqrt{26116897})/2 \\
 f_2 d_{34} &= (-65931742559 + 12256337\sqrt{26116897}) \\
 f_2 d_{35} &= -(-50597298167 + 9716873\sqrt{26116897})/6 \\
 f_2 d_{36} &= -88(-15453061 + 2911\sqrt{26116897}) \\
 d_{37} &= 0 \\
 f_1 d_{41} &= 48(-56020909845192541 + 9790180507043\sqrt{26116897}) \\
 f_1 d_{42} &= (-9918249049237586011 + 1463702013196501\sqrt{26116897})/6
 \end{aligned}$$

$$\begin{aligned}
f_1d_{43} &= -13(-4130451756851441723 + 664278707201077\sqrt{26116897}) \\
f_1d_{44} &= 3(-26937108467782666617 + 5169063172799767\sqrt{26116897})/2 \\
f_1d_{45} &= -(6548308508012371315 + 3968886380989379\sqrt{26116897})/3 \\
f_1d_{46} &= 88(-91337851897923397 + 19696768305507\sqrt{26116897}) \\
f_3d_{47} &= 242(-120683 + 15\sqrt{26116897}) \\
f_3d_{51} &= 264(-120683 + 15\sqrt{26116897}) \\
f_3d_{52} &= (-43118111 + 23357\sqrt{26116897})/3 \\
f_3d_{53} &= -47(-28770085 + 2259\sqrt{26116897})/2 \\
f_3d_{54} &= -3(1003619433 + 11777\sqrt{26116897}) \\
f_3d_{55} &= -11(-384168269 + 65747\sqrt{26116897})/6 \\
f_3d_{56} &= 22(87290207 + 10221\sqrt{26116897}) \\
f_1d_{57} &= -66(3692405 = 419\sqrt{26116897})
\end{aligned}$$

where

$$\begin{aligned}f_1 &= -56764003702447356523 + 8154993476273221\sqrt{26116897} \\f_2 &= -55804550303 + 9650225\sqrt{26116897} \\f_3 &= 3262210757 + 271861\sqrt{26116897}\end{aligned}$$

In floating point format:

$$\begin{aligned}
 d_{34} &= 0.508080676928351487908752085978 \\
 d_{35} &= -0.0241370624126563706018867104972 \\
 d_{36} &= -0.00781990939443926721678719106473 \\
 d_{37} &= 0 \\
 d_{41} &= 0.0190512060948850190478223587424 \\
 \\
 d_{42} &= 0.0269311042007326141816664674714 \\
 d_{43} &= -0.633860292039252305642283500160 \\
 d_{44} &= 0.0517726709186493664626888177642 \\
 d_{45} &= 0.592764606048964306931634491846 \\
 d_{46} &= -0.0543688142698406758774679261364 \\
 d_{47} &= -0.00229048095413832510406070952285 \\
 d_{51} &= -0.00249870649542362738624804675220 \\
 d_{52} &= 0.00546392445304455008494236684033 \\
 d_{53} &= 0.0870248056190193154450416111555 \\
 d_{54} &= -0.686097670431383548237962511317 \\
 d_{55} &= 0.0189855304809436619879348998897 \\
 d_{56} &= 0.659895344563505072850627735852 \\
 d_{57} &= -0.827732281897054247443360556719
 \end{aligned}$$

Inner points:

$$-\alpha_{-2} = \alpha_2 = -1/12, \quad -\alpha_{-1} = \alpha_1 = 2/3.$$

12

THE LAPLACE TRANSFORM METHOD FOR DIFFERENCE APPROXIMATIONS

12.1. NECESSARY CONDITIONS FOR STABILITY

Consider a difference approximation for a system of partial differential equations with constant coefficients. For periodic problems, a simple test for stability is to construct simple wave solutions

$$u(x, t) = e^{i\langle \omega, x \rangle} \hat{u}(\omega, t)$$

and to estimate the growth rate of $\hat{u}(\omega, t)$. This leads to the von Neumann condition as a necessary stability condition. As we have seen in Chapter 10, there is a similar procedure for the initial-boundary-value problem. We now adapt this method for difference approximations and begin with a simple example. We approximate

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x} + F, & 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \\ \|u(\cdot, t)\| &< \infty \end{aligned} \tag{12.1.1}$$

by

$$\frac{dv_j}{dt} = D_0 v_j + F_j, \quad j = 1, 2, \dots, \tag{10.1.2a}$$

$$v_j(0) = f_j, \tag{12.1.2b}$$

$$v_0 - 2v_1 + v_2 = g, \tag{12.1.2c}$$

$$\|v\|_h < \infty. \tag{12.1.2d}$$

(We have introduced the inhomogeneous term g in the boundary condition to be used later). The scalar product and the norm are now defined by

$$(v, w)_h = (v, w)_{1, \infty}, \quad \|v\|_h^2 = (v, v)_h. \quad (12.1.3)$$

The test is given by the following lemma.

Lemma 12.1.1. *Let $F \equiv g \equiv 0$. If Eq. (12.1.2) has a solution*

$$v_j = e^{st} f_j, \quad \|f\|_h < \infty \quad (12.1.4)$$

for some complex number s with $\operatorname{Re} s > 0$ and some stepsize h_0 , then the approximation is not stable in any sense. (See Lemma 10.1.1.)

This necessary condition for stability is called the *Godunov–Ryabenkii condition*.

Proof. If such a solution exists, for some $h = h_0$, $s = s_0$, $\operatorname{Re} s_0 > 0$, then

$$\begin{aligned} \tilde{s}_0 f_j &= \frac{1}{2} (f_{j+1} - f_{j-1}), \quad j = 1, 2, \dots, \quad \tilde{s}_0 = hs_0, \\ f_2 - 2f_1 + f_0 &= 0, \\ \|f\|_h &< \infty. \end{aligned}$$

Therefore, for any h ,

$$\tilde{v}_j = e^{(\tilde{s}_0/h)t} f_j$$

is also a solution. Thus, as $h \rightarrow 0$, we can construct solutions growing arbitrarily fast. We can formulate this in terms of the eigenvalue problem

$$\tilde{s} \varphi_j = h D_0 \varphi_j, \quad j = 1, 2, \dots, \quad \tilde{s} = hs, \quad (12.1.5a)$$

$$\varphi_0 - 2\varphi_1 + \varphi_2 = 0 \quad (12.1.5b)$$

$$\|\varphi\|_h < \infty. \quad (12.1.5c)$$

We have the following lemma.

Lemma 12.1.2. *The approximation (12.1.2) is not stable if Eq. (12.1.5) has an eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} > 0$. (See Theorem 10.1.1.)*

Equation (12.1.5a) is an ordinary difference equation with constant coefficients, and its solution has the form

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j,$$

where κ_1 and κ_2 are the two solutions of the characteristic equation

$$\kappa^2 - 2\tilde{s}\kappa - 1 = 0, \quad \tilde{s} = sh. \quad (12.1.6a)$$

The next lemma discusses properties of κ_1 and κ_2 .

Lemma 12.1.3. *For $\operatorname{Re} \tilde{s} > 0$, the characteristic equation has no solutions with $|\kappa| = 1$, and there is exactly one solution with $|\kappa| < 1$,*

$$\kappa_1 = \tilde{s} - \sqrt{1 + \tilde{s}^2}, \quad \tilde{s} = sh. \quad (12.1.6b)$$

Proof. Assume that Eq. (12.1.6a) has a solution $\kappa = e^{i\xi}$ for $\operatorname{Re}(\tilde{s}) > 0$, ξ real. Then

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

which is a contradiction. Thus, there are no solutions with $|\kappa| = 1$. Because the two solutions κ_1 and κ_2 satisfy $\kappa_1 \kappa_2 = -1$, it is obvious that there is exactly one root, κ_1 say, that is inside the unit circle for all \tilde{s} with $\operatorname{Re} \tilde{s} > 0$. A simple calculation shows that Eq. (12.1.6b) is the desired solution.

Lemma 12.1.3 and the condition (12.1.5c) imply that the solution has the form

$$\varphi_j = \sigma_1 \kappa_1^j, \quad (12.1.7)$$

which we substitute into the boundary condition (12.1.5b). This yields

$$\sigma_1(\kappa_1 - 1)^2 = 0. \quad (12.1.8)$$

Therefore, there is a nontrivial solution if, and only if, $\kappa_1 = 1$. By Lemma 12.1.3, this is impossible, and we have shown that there is no eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} > 0$. Indeed, the same result is obtained for any order of extrapolation at the boundary

$$(hD_+)^q v_0 = 0, \quad q = 1, 2, \dots \quad (12.1.9)$$

The only difference in the discussion above is that Eq. (12.1.8) becomes

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

which leads to the same conclusion.

We now discuss difference approximations with constant coefficients for the quarter-space problem (10.1.1).

$$\frac{dv_j(t)}{dt} = Qv_j(t) + F_j(t), \quad j = 1, 2, \dots, \quad (12.1.10a)$$

$$v_j(0) = f_j, \quad j = 1, 2, \dots, \quad (12.1.10b)$$

$$L_0 v_0(t) = g(t), \quad (12.1.10c)$$

where the $v_j(t)$ are vector functions with m components and $\|f\|_h < \infty$, $\sup_t \|F\|_h < \infty$. The difference operator in space has the form

$$Q = \frac{1}{h} \sum_{\nu=-r}^p B_\nu E^\nu, \quad (12.1.11)$$

where E is the shift operator. In general, the $B_\nu = B_{\nu 0} + h B_{\nu 1}$ are $m \times m$ matrices. Here $B_{\nu 0}$ does not depend on h , and $B_{\nu 1}$ has no influence on stability. Therefore, we neglect these terms and assume that $B_\nu = B_{\nu 0}$. We assume that B_p and B_{-r} are nonsingular. Another assumption is that Eq. (12.1.10a) is stable with periodic boundary conditions.

Finally, we assume that the boundary conditions can be written as

$$v_{-\mu} = \sum_{j=1}^q L_{\mu j} v_j + g_{-\mu}, \quad \mu = 0, 1, \dots, r-1, \quad (12.1.12)$$

where the $L_{\mu j}$ are constant matrices. Again, we assume that the $L_{\mu j}$ do not depend on h . This is no restriction. As we have seen in Section 11.3, we can use the boundary condition to eliminate v_0, \dots, v_{-r+1} , thus modifying Q to \tilde{Q} . Any $\mathcal{O}(h)$ terms in the boundary condition create bounded operators and have no influence on stability. We could also include time-derivatives in the boundary conditions. However, it is assumed that these have been eliminated by using the differential equations (12.1.10a).

The eigenvalue problem associated with our approximation is

$$\tilde{s}\varphi_j = hQ\varphi_j, \quad j = 1, 2, \dots, \quad \operatorname{Re}(\tilde{s}) = \operatorname{Re}(hs) > 0, \quad (12.1.13a)$$

$$L_0 \varphi_0 = 0, \quad (12.1.13b)$$

$$\|\varphi\|_h < \infty. \quad (12.1.13c)$$

Using the same argument as for our example we can prove the following lemma.

Lemma 12.1.4. Godunov–Ryabenkii Condition. *The approximation is not stable if the eigenvalue problem (12.1.13) has an eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} > 0$.*

Proof. Let \tilde{s}_0 with $\operatorname{Re} \tilde{s}_0 > 0$ be an eigenvalue and φ_j the corresponding eigenfunction. Then

$$v_j = e^{(\tilde{s}_0/h)t} \varphi_j$$

form a sequence of solutions whose growth rate becomes arbitrarily large as $h \rightarrow 0$. Thus, the approximation is not stable.

We now derive algebraic conditions for the existence of eigenvalues.

Lemma 12.1.5. *The eigenvalue problem (12.1.13) has no eigenvalues for sufficiently large $|\tilde{s}|$.*

Proof. Equation (12.1.13) implies

$$\begin{aligned} \|\varphi\|_h^2 &= \frac{1}{|\tilde{s}|^2} \|hQ\varphi\|_h^2, \\ &\leq \frac{\text{constant}}{|\tilde{s}|^2} \left(\|\varphi\|_h^2 + \sum_{\mu=0}^{r-1} |\varphi_{-\mu}|^2 h \right) \leq \frac{\text{constant}}{|\tilde{s}|^2} \|\varphi\|_h^2. \end{aligned}$$

Here the constant depends only on the coefficients $|B_\nu|$ and $|L_{\mu j}|$. Therefore, $\|\varphi\|_h = 0$ for sufficiently large $|\tilde{s}|$. This proves the lemma.

We now discuss how to solve the eigenvalue problem (12.1.13). We write Eq. (12.1.13a) in the form

$$\varphi_{p+j} = \sum_{\nu=-r}^{p-1} \tilde{B}_\nu \varphi_{\nu+j}, \quad (12.1.14)$$

where

$$\tilde{B}_\nu = -B_p^{-1} B_\nu, \quad \nu \neq 0, \quad \tilde{B}_0 = B_p^{-1} \tilde{s} - B_p^{-1} B_0.$$

This is a system of ordinary difference equations in space. Introducing

$$\varphi_j = (\varphi_{p+j-1}, \varphi_{p+j-2}, \dots, \varphi_{j-r})^T,$$

we can write it as a one-step method

$$\varphi_{j+1} = M\varphi_j, \quad j = 1, 2, \dots, \quad (12.1.15a)$$

where

$$M = \begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & I & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & I & 0 \end{bmatrix}.$$

The boundary conditions can be written in the form

$$H\varphi_1 = 0, \quad (12.1.15b)$$

because we can use Eq. (12.1.14) to eliminate all φ_ν with $\nu > p$ from $L_0\varphi_0 = 0$. Thus, $M = M(\tilde{s})$, $H = H(\tilde{s})$ are polynomials in \tilde{s} .

The eigenvalues and eigenvectors of M are the solutions of

$$\begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & 0 \\ 0 & I & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & I & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p+r} \end{bmatrix} = \kappa \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p+r} \end{bmatrix} \quad (12.1.16)$$

(cf. Lemma 5.2.2). We shall use the solutions of Eq. (12.1.16) to solve the eigenvalue problem (12.1.13). To demonstrate the procedure, we first derive some properties of M .

We want to show that $\kappa = 0$ is not an eigenvalue of M . Assume that $\kappa = 0$ is an eigenvalue. Then Eq. (12.1.16) becomes $y_1 = y_2 = \dots = y_{p+r-1} = 0$, $\tilde{B}_{-r}y_{p+r} = 0$. Because, by assumption, \tilde{B}_{-r} is nonsingular, $y_{p+r} = 0$. Thus, we have arrived at a contradiction and $\kappa = 0$ cannot be an eigenvalue. Using the relation

$$y_{j+1} = \kappa^{-1}y_j, \quad j = 1, 2, \dots, p + r - 1,$$

we can eliminate y_2, \dots, y_{p+r} and obtain, after a simple computation,

$$\left(\tilde{s}I - \sum_{\nu=-r}^p B_\nu \kappa^\nu \right) y_1 = 0. \quad (12.1.17)$$

Thus, the eigenvalues of M are the solutions of the characteristic equation

$$\text{Det} \left(\tilde{s}I - \sum_{\nu=-r}^p B_\nu \kappa^\nu \right) = 0 \quad (12.1.18)$$

We now need the following lemma.

Lemma 12.1.6. *For \tilde{s} with $\text{Re } \tilde{s} > 0$, there is no solution of Eq. (12.1.18) with $|\kappa| = 1$ and there are exactly rm solutions, counted according to their multiplicity, with $|\kappa| < 1$.*

Proof. Assume that there is a root $\kappa = e^{i\xi}$. Then Eq. (12.1.18) implies that \tilde{s} is an eigenvalue of $\sum_{\nu=-r}^p B_\nu e^{i\nu\xi}$. Because we have assumed that the approximation is stable for the periodic case, we necessarily have $\text{Re } \tilde{s} \leq 0$. This is a contradiction to the hypothesis $\text{Re } \tilde{s} > 0$; that is, there are no solutions κ with $|\kappa| = 1$. The solutions κ are continuous functions of \tilde{s} and cannot cross the unit circle. Therefore, the number of solutions with $|\kappa| < 1$ is constant for $\text{Re } \tilde{s} > 0$, and we can determine their number from the limit $\text{Re } \tilde{s} \rightarrow \infty$. In this case, the solutions with $|\kappa| < 1$ converge to zero and are, to first approximation, determined by

$$\text{Det}(\tilde{s}I - B_{-r} \kappa^{-r}) = 0. \quad (12.1.19)$$

Because B_{-r} is nonsingular, Eq. (12.1.19) has exactly mr solutions $\kappa = \mathcal{O}(\tilde{s}^{-1/r})$. This proves the lemma.

By Schur's lemma (see Appendix A.1), we can find a unitary transformation $U = U(\tilde{s})$ such that

$$U^* M U = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}.$$

Here the eigenvalues of M_{11} and M_{22} satisfy $|\kappa| < 1$ and $|\kappa| > 1$ for $\text{Re } \tilde{s} > 0$, respectively. Thus, M_{11} is an $rm \times rm$ matrix. Introducing a new variable

$$\Psi_j = U^* \Phi_j$$

into Eq. (12.1.15) gives us

$$\begin{bmatrix} \Psi^I \\ \Psi^{II} \end{bmatrix}_{j+1} = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} \Psi^I \\ \Psi^{II} \end{bmatrix}_j \quad (12.1.20a)$$

with boundary conditions

$$HU\psi_1 =: H^I\psi_1^I + H^{II}\psi_1^{II} = 0, \quad \|\psi\|_h < \infty. \quad (12.1.20b)$$

Here H^I is again an $rm \times rm$ matrix.

We can now prove the following lemma.

Lemma 12.1.7. *The Godunov–Ryabenkii condition is satisfied if, and only if, H is nonsingular for $\operatorname{Re} \tilde{s} > 0$, that is,*

$$\operatorname{Det}|H^I| \neq 0, \quad \text{for } \operatorname{Re} \tilde{s} > 0. \quad (12.1.21)$$

Proof. By Eq. (12.1.20a),

$$\psi_j^{II} = M_{22}^{j-1} \psi_1^{II}.$$

Therefore, $\|\psi\|_h < \infty$ implies $\psi_1^{II} = 0$; that is, $\psi^{II} \equiv 0$ and the solutions of Eq. (12.1.20) satisfy

$$\psi_j^I = M_{11}^{j-1} \psi_1^I, \quad H^I \psi_1^I = 0.$$

Thus, there is a nontrivial solution if, and only if, H^I is singular. This proves the lemma.

We have, therefore, derived an algebraic condition for the Godunov–Ryabenkii condition. To verify the condition we need not go through the reduction of Eq. (12.1.13) to Eq. (12.1.15) and the construction of U . The above construction tells us that the general solution of Eq. (12.1.13a) with $\|\varphi\|_h < \infty$ is of the form

$$\varphi_j = \sum_{|\kappa_\nu| < 1} P_\nu(j) \kappa_\nu^j, \quad \kappa_\nu = \kappa_\nu(\tilde{s}), \quad \operatorname{Re} \tilde{s} > 0 \quad (12.1.22)$$

and depends on rm free parameters $\sigma = (\sigma_1, \dots, \sigma_{rm})^T$. Here $P_\nu(j)$ is a polynomial in j with vector coefficients. Its order is at most $m_\nu - 1$ where m_ν is the multiplicity of κ_ν .

Substituting Eq. (12.1.22) into the boundary conditions (12.1.13b) yields a system of equations

$$C(\tilde{s})\sigma = 0, \quad (12.1.23)$$

and we can rephrase Lemma 12.1.7 in the following form.

Lemma 12.1.8. *The Godunov-Ryabenkii condition is satisfied if, and only if,*

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

REMARK. Since the number of free parameters is rm , it is a consequence of Lemma 12.1.8 that the number of boundary conditions cannot be less than rm for a stable approximation.

As an example, we approximate

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x}, \quad a \neq 0, \quad 0 \leq x < \infty, \quad t \geq 0,$$

by the fourth-order difference scheme

$$\frac{\partial v_j}{\partial t} = a \left(\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h) \right) v_j, \quad j = 1, 2, \dots \quad (12.1.24)$$

If $a > 0$, we use as boundary conditions

$$D_+^q v_0 = D_-^q v_{-1} = 0. \quad (12.1.25)$$

If $a < 0$, we use the boundary condition $u(0, t) = 0$ of the differential equation to obtain

$$u_{tt}(0, t) = a^2 u_{xx}(0, t) = 0.$$

As boundary conditions for the difference approximation we use

$$v_0 = 0, \quad D_+ D_- v_0 = 0. \quad (12.1.26)$$

[It will be shown later that this approximation has a global $\mathcal{O}(h^4)$ error.]

Equation (12.1.13a) has the form

$$\tilde{s}\varphi_j = a \left(\frac{2}{3} (\varphi_{j+1} - \varphi_{j-1}) - \frac{1}{12} (\varphi_{j+2} - \varphi_{j-2}) \right), \quad (12.1.27)$$

and the characteristic equation is given by

$$\tilde{s} = a \left(\frac{2}{3} \left(\kappa - \frac{1}{\kappa} \right) - \frac{1}{12} \left(\kappa^2 - \frac{1}{\kappa^2} \right) \right). \quad (12.1.28)$$

We have the following lemma.

Lemma 12.1.9.

1. *The characteristic equation (12.1.28) has exactly two roots*

$$|\kappa_\nu| < 1, \quad \operatorname{Re} \tilde{s} > 0, \quad \nu = 1, 2.$$

2. *In a neighborhood of $\tilde{s} = 0$, these roots are of the form*

$$\begin{aligned}\kappa_1 &= -1 + \frac{\tilde{s}}{a} + \mathcal{O}(|\tilde{s}|^2), & \kappa_2 &= 4 - \sqrt{15} + \mathcal{O}(|\tilde{s}|), \quad \text{if } a > 0, \\ \kappa_1 &= 1 - \frac{\tilde{s}}{|a|} + \mathcal{O}(|\tilde{s}|^2), & \kappa_2 &= 4 - \sqrt{15} + \mathcal{O}(|\tilde{s}|), \quad \text{if } a < 0.\end{aligned}$$

Proof. The first statement follows from Lemma 12.1.6.

For $\tilde{s} = 0$, the solutions of the characteristic equations are

$$\kappa^{(1,2)} = \mp 1, \quad \kappa^{(3,4)} = 4 \pm \sqrt{15}.$$

By perturbation arguments, we determine which of these four roots are κ_1 and κ_2 . We have for small $|\tilde{s}|$

$$\kappa^{(1)} = -1 + \frac{3}{5} \frac{\tilde{s}}{a} + \mathcal{O}(|\tilde{s}|^2), \quad \kappa^{(2)} = 1 + \frac{\tilde{s}}{a} + \mathcal{O}(|\tilde{s}|^2).$$

By selecting the roots satisfying $|\kappa| < 1$ for $\operatorname{Re} \tilde{s} > 0$, the second statement follows. This proves the lemma.

By Lemma 12.1.9, there are two roots κ_ν , $\nu = 1, 2$, with $|\kappa_\nu| < 1$ for $\operatorname{Re} \tilde{s} > 0$. Thus, the general solution of Eq. (12.1.27) with $\|\varphi\|_h < \infty$ has the form

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 \frac{\kappa_2^j - \kappa_1^j}{\kappa_2 - \kappa_1}, \quad \text{if } \kappa_1 \neq \kappa_2.$$

If $\kappa_1 = \kappa_2$ is a double root, it becomes

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 j \kappa_1^{j-1}.$$

We substitute this expression into the boundary conditions (12.1.25) used for $a > 0$ and obtain

$$\begin{aligned} \sigma_1(\kappa_1 - 1)^q + \frac{\sigma_2}{\kappa_2 - \kappa_1} ((\kappa_2 - 1)^q - (\kappa_1 - 1)^q) &= 0, \\ \sigma_1 \frac{(\kappa_1 - 1)^q}{\kappa_1} + \frac{\sigma_2}{\kappa_2 - \kappa_1} \left(\frac{(\kappa_2 - 1)^q}{\kappa_2} - \frac{(\kappa_1 - 1)^q}{\kappa_1} \right) &= 0. \end{aligned} \quad (12.1.29)$$

The determinant of this system is

$$\text{Det} = -\frac{(\kappa_1 - 1)^q(\kappa_2 - 1)^q}{\kappa_1 \kappa_2}.$$

By Lemma 12.1.9, $\kappa_1 \neq 1$, $\kappa_2 \neq 1$; that is, there is no eigenvalue with $\text{Re } \tilde{s} > 0$. For the boundary conditions (12.1.26) used for the case $a < 0$, we obtain

$$\begin{aligned} \sigma_1 &= 0 \\ \frac{\sigma_2}{\kappa_2 - \kappa_1} \left(\frac{(\kappa_2 - 1)^2}{\kappa_2} - \frac{(\kappa_1 - 1)^2}{\kappa_1} \right) &= \sigma_2 \left(1 - \frac{1}{\kappa_1 \kappa_2} \right) = 0, \end{aligned} \quad (12.1.30)$$

which, because $|\kappa_1 \kappa_2| < 1$, only has the trivial solution. Therefore, there is no eigenvalue with $\text{Re } \tilde{s} > 0$.

Lemmas 12.1.4, 12.1.7, and 12.1.8 give the conditions necessary for stability. As in the continuous case, we use the Laplace transform to derive sufficient conditions for stability in the next two sections.

EXERCISES

- 12.1.1.** Prove that the Godunov-Ryabenkii condition is satisfied for the fourth-order approximation (12.1.24) with the boundary conditions

$$v_{-1} = v_0 = 0$$

for any value of a .

12.2. SUFFICIENT CONDITIONS FOR STABILITY

In this section, we use the Laplace transform to derive sufficient algebraic conditions for stability. We proceed as in the continuous case in Section 10.5 where symmetric hyperbolic systems in several space dimensions were considered. We use the Laplace transform to estimate the solution on the boundary. Energy estimates give us the desired stability result.

We begin with the example (12.1.2) and assume that $F \equiv f \equiv 0$. We call the solution of this particular problem y . The Laplace transformed equations have

the form

$$\tilde{s}\hat{y}_j(s) = hD_0\hat{y}_j(s), \quad j = 1, 2, \dots, \quad (12.2.1a)$$

$$\hat{y}_0 - 2\hat{y}_1 + \hat{y}_2 = \hat{g}, \quad \|\hat{y}\|_h < \infty. \quad (12.2.1b)$$

We need to use the inverse Laplace transform to obtain the stability estimates in physical space. Thus, our estimates must hold for all s with $\operatorname{Re} s > \eta_0$, where η_0 is a constant. Because h is arbitrarily small, it will be necessary to consider the whole half-plane $\operatorname{Re} \tilde{s} > 0$. By Eq. (12.1.7), the general solution of Eq. (12.2.1a) is

$$\hat{y}_j = \sigma_1 \kappa_1^j.$$

Substituting this into the boundary condition (12.2.1b) gives us

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

By Lemma 12.1.3, $\kappa_1 - 1 \neq 0$ for $\operatorname{Re} \tilde{s} > 0$ and, therefore, σ_1 is uniquely determined. We can be more precise.

Lemma 12.2.1. *There is a constant $\delta > 0$ such that*

$$|\kappa_1 - 1| \geq \delta > 0, \quad \text{for all } \tilde{s} \text{ with } \operatorname{Re} \tilde{s} \geq 0. \quad (12.2.2)$$

Proof. By Eq. (12.1.6b), $\kappa_1 \rightarrow 0$ as $|\tilde{s}| \rightarrow \infty$. Therefore, Eq. (12.2.2) holds for sufficiently large $|\tilde{s}|$. Also, κ_1 is a continuous function of \tilde{s} , and, therefore, Eq. (12.2.2) holds if we can show that $\kappa_1 \neq 1$ for all \tilde{s} with $\operatorname{Re} \tilde{s} \geq 0$. Assume that $\kappa_1 = 1$; then, by Eq. (12.1.6a), $s = 0$ and, by Eq. (12.1.6b), $\kappa_1 = -1$, which is a contradiction. This proves the lemma.

We can now invert the Laplace transform

$$y_j(t) = \frac{1}{2\pi i} \int_{\mathcal{L}} \frac{\kappa_1^j}{(\kappa_1 - 1)^2} \hat{g}(s)e^{st} ds. \quad (12.2.3)$$

For \mathcal{L} we can take the contour $\operatorname{Re} s = 0$. By Parseval's relation, we obtain

$$\int_0^\infty |y_j(t)|^2 dt \leq \frac{1}{\delta^4} \int_0^\infty |g(t)|^2 dt, \quad j = 0, 1, \dots$$

We also obtain, for every $T \geq 0$,

$$\int_0^T |y_j(t)|^2 dt \leq \frac{1}{\delta^4} \int_0^T |g(t)|^2 dt, \quad (12.2.4)$$

because the solution for $0 \leq t \leq T$ does not depend on values of $g(t)$ with $t > T$, and, therefore, we can assume that $g(t) = 0$ for $t > T$. [For more details, see Kreiss and Lorenz (1989).]

Now we can derive a standard energy estimate. Lemma 11.1.1 and Eq. (12.2.4) give us

$$\frac{d}{dt} \|y\|_h^2 = (y, D_0 y)_h + (D_0 y, y)_h = \frac{1}{2} (\bar{y}_0 y_1 + \bar{y}_1 y_0),$$

that is,

$$\begin{aligned} \|y(T)\|_h^2 &\leq \int_0^T |y_0| |y_1| dt \leq \left(\int_0^T |y_0|^2 dt \right)^{1/2} \left(\int_0^T |y_1|^2 dt \right)^{1/2} \\ &\leq \frac{1}{\delta^4} \int_0^T |g(t)|^2 dt. \end{aligned} \quad (12.2.5)$$

Thus, we can estimate the solution in terms of the data.

We shall prove that the approximation (12.1.2) is strongly stable in the sense of Section 11.3. We assume that $F = g = 0$ and first solve the auxiliary problem

$$\frac{dw_j}{dt} = D_0 w_j, \quad j = 1, 2, \dots, \quad (12.2.6a)$$

$$w_j(0) = f_j, \quad (12.2.6b)$$

$$w_0 - w_1 = 0. \quad (12.2.6c)$$

Lemma 11.1.1 gives us

$$\frac{d}{dt} \|w\|_h^2 + |w_0|^2 \leq 0;$$

that is,

$$\|w(T)\|_h^2 + \int_0^T |w_0|^2 dt \leq \|f\|_h^2. \quad (12.2.7)$$

Thus,

$$\int_0^\infty |w_0|^2 dt = \int_0^\infty |w_1|^2 dt \leq \|f\|_h^2. \quad (12.2.8a)$$

We also want to show that

$$\int_0^\infty |w_2|^2 dt \leq \text{constant} \|f\|_h^2. \quad (12.2.8b)$$

The Laplace transformed equations (12.2.6) are

$$s\hat{w}_j = D_0\hat{w}_j + f_j, \quad j = 1, 2, \dots, \quad (12.2.9a)$$

$$\hat{w}_0 = \hat{w}_1 = 0, \quad \|\hat{w}\|_h < \infty. \quad (12.2.9b)$$

For $|sh| \geq 2$, we obtain

$$\begin{aligned} \|\hat{w}\|_h^2 &\leq \frac{2}{|s|^2} \|D_0\hat{w}\|_h^2 + \frac{2}{|s|^2} \|f\|_h^2, \\ &= \frac{1}{2|sh|^2} \sum_{j=1}^{\infty} |\hat{w}_{j+1} - \hat{w}_{j-1}|^2 h + \frac{2}{|s|^2} \|f\|_h^2, \\ &\leq \frac{2}{|sh|^2} \|\hat{w}\|_h^2 + \frac{1}{|s^2 h|} |\hat{w}_0|^2 + \frac{2}{|s|^2} \|f\|_h^2; \end{aligned}$$

that is,

$$|\hat{w}_2|^2 \leq \frac{1}{2} |\hat{w}_0|^2 + \frac{4}{h|s|^2} \|f\|_h^2, \quad |sh| \geq 2. \quad (12.2.10)$$

For $|sh| \leq 2$, we write Eq. (12.2.9a) in the form

$$\hat{w}_2 = 2hs\hat{w}_1 + \hat{w}_0 - 2hf_1;$$

that is,

$$|\hat{w}_2|^2 \leq \text{constant} (|\hat{w}_1|^2 + |\hat{w}_0|^2 + h\|f\|_h^2)$$

and, therefore, by Eq. (12.2.10), we obtain, for $s = i\xi$,

$$\int_{-\infty}^{+\infty} |\hat{w}_2(i\xi)|^2 d\xi \leq \text{constant} \left(\int_{-\infty}^{+\infty} (|\hat{w}_1(i\xi)|^2 + |\hat{w}_0(i\xi)|^2) d\xi + \|f\|_h^2 \right).$$

Parseval's relation gives us Eq. (12.2.8b). Now we substitute

$$y = v - w$$

into Eq. (12.1.2) as a new variable and obtain

$$\frac{dy_j}{dt} = D_0 y_j, \quad j = 1, 2, \dots, \quad (12.2.11a)$$

$$y_j(0) = 0, \quad (12.2.11b)$$

$$y_0 - 2y_1 + y_2 = g(t) + g_1(t), \quad (12.2.11c)$$

where $g_1(t) = -(w_0 - 2w_1 + w_2)$. (For convenience, we have assumed $F = 0$.) By Eq. (12.2.8),

$$\int_0^\infty |g_1(t)|^2 dt \leq \text{constant} \|f\|_h^2. \quad (12.2.12)$$

For Eq. (12.2.11), we can use the estimate (12.2.5) with g replaced by $g + g_1$ i.e., by Eq. (12.2.7):

$$\|v(T)\|_h^2 \leq 2(\|y(T)\|_h^2 + \|w(T)\|_h^2) \leq \text{constant} \left(\|f\|_h^2 + \int_0^T |g(t)|^2 dt \right). \quad (12.2.13)$$

This proves strong stability.

We now generalize the stability result to systems (12.1.10). As for the example, we first solve the problem for the case that $f = F \equiv 0$ and call the resulting solution y . The Laplace transformed equations have the form

$$s\hat{y}_j = Q\hat{y}_j, \quad j = 1, 2, \dots, \quad (12.2.14a)$$

$$L_0\hat{y}_0 = \hat{g}, \quad \|\hat{y}\|_h < \infty. \quad (12.2.14b)$$

We want to derive estimates for $|\hat{y}_j|$.

For large $|sh|$, we have the following lemma.

Lemma 12.2.2. *There are constants C_0 and K_0 such that, for all $|\tilde{s}| = |sh| \geq C_0$,*

$$|\hat{y}_j| \leq \frac{K_0}{|sh|} |\hat{g}|. \quad (12.2.15)$$

Proof. Equations (12.2.14) and (12.1.12) give us

$$\begin{aligned} \|\hat{y}\|_h^2 &= \frac{1}{|\tilde{s}|^2} \|hQ\hat{y}\|_h^2 \leq \frac{\text{constant}}{|\tilde{s}|^2} \left(\|\hat{y}\|_h^2 + \sum_{j=-r+1}^0 |\hat{y}_j|^2 h \right) \\ &\leq \frac{\text{constant}}{|\tilde{s}|^2} (\|\hat{y}\|_h^2 + |\hat{g}|^2 h). \end{aligned}$$

For $\text{constant}/|\tilde{s}|^2 \leq \frac{1}{2}$, the desired estimate follows.

Corresponding to Eq. (12.1.15), we introduce

$$\mathbf{y}_j = (\hat{y}_{p+j-1}, \dots, \hat{y}_{j-r})^T,$$

and we write Eq. (12.2.14) as

$$\mathbf{y}_{j+1} = M\mathbf{y}_j, \quad j = 1, 2, \dots, \quad (12.2.16a)$$

with boundary conditions

$$H\mathbf{y}_1 = \mathbf{g}, \quad \|\mathbf{y}\|_h < \infty. \quad (12.2.16b)$$

Here $M = M(\tilde{s})$, $H = H(\tilde{s})$ are polynomials in \tilde{s} . Therefore, we can, on any compact set $S := (|\tilde{s}| \leq C_0, \operatorname{Re} \tilde{s} \geq 0)$ choose the unitary transformation $U(\tilde{s})$ that transforms M to upper triangular form as a continuous function of \tilde{s} . Corresponding to Eq. (12.1.20), the change of variables $\mathbf{w} = U^*\mathbf{y}$ gives us

$$\begin{bmatrix} \mathbf{w}^I \\ \mathbf{w}^{II} \end{bmatrix}_{j+1} = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}^I \\ \mathbf{w}^{II} \end{bmatrix}_j, \quad j = 1, 2, \dots, \quad (12.2.17a)$$

with boundary conditions

$$HU\mathbf{w}_1 =: H^I\mathbf{w}_1^I + H^{II}\mathbf{w}_1^{II} = \mathbf{g}, \quad \|\mathbf{w}\|_h < \infty. \quad (12.2.17b)$$

Here M_{ij} , H^I , H^{II} are continuous functions of \tilde{s} on any compact set S .

As in the last section, $\mathbf{w}^{II} = 0$, and the solution satisfies

$$\begin{aligned}\mathbf{w}^{II} &= 0, \\ \mathbf{w}_j^I &= M_{11}^{g-1} \mathbf{w}_1^I, \\ H^I \mathbf{w}_1^I &= \mathbf{g}.\end{aligned}\tag{12.2.17c}$$

By Lemma 12.1.7 the Godunov-Ryabenkii condition is equivalent to

$$\text{Det}(H^I(\tilde{s})) \neq 0, \quad \text{for } \operatorname{Re} \tilde{s} > 0.$$

We strengthen it to

$$\text{Det}(H^I(\tilde{s})) \neq 0, \quad \text{for } \operatorname{Re} \tilde{s} \geq 0.\tag{12.2.18}$$

This condition is called the *determinant condition*. As we will see, it is equivalent with a uniform estimate of the boundary values in terms of boundary data. We make the following definition.

Definition 12.2.1. Assume that there is a constant K that is independent of \tilde{s} and \hat{g} such that the solutions of Eq. (12.2.14) satisfy

$$\sum_{\nu=-r+1}^p |\hat{y}_\nu|^2 \leq K|\hat{g}|^2, \quad \operatorname{Re} \tilde{s} > 0.\tag{12.2.19}$$

Then we say that the Kreiss condition is satisfied.

We will now prove the following lemma.

Lemma 12.2.3. The Kreiss condition (12.2.19) is equivalent to the determinant condition (12.2.18).

Proof. By Lemma 12.2.2 we need only to consider $|\tilde{s}| \leq C_0$. Assume that the Eq. (12.2.18) holds. Because $H^I(\tilde{s})$ is a continuous function of \tilde{s} , there is a constant $\delta > 0$ such that

$$|\text{Det}(H^I(\tilde{s}))| \geq \delta.\tag{12.2.20}$$

Therefore, $|(H^I(\tilde{s}))^{-1}|$ is uniformly bounded, and Eq. (12.2.19) follows from Eq. (12.2.17c).

If Eq. (12.2.19) holds, then $(H^I(\tilde{s}))^{-1}$ must be uniformly bounded and Eq. (12.2.18) holds. This proves the lemma.

The Kreiss condition gives an estimate of the solution near the boundary. In fact the solution can be estimated at any fixed point x_j . We have the following lemma.

Lemma 12.2.4. *If the Kreiss condition holds, then there are constants K_j such that for every fixed j the solutions of Eq. (12.2.14) satisfy*

$$|\hat{y}_j| \leq K_j |\hat{g}|, \quad \operatorname{Re} \tilde{s} > 0. \quad (12.2.21)$$

Proof. By Eq. (12.2.17c) we get

$$|\mathbf{w}_j| = |\mathbf{w}_j^I| = |M_{11}^{j-1} \mathbf{w}_1^I| \leq |M_{11}|^{j-1} |(H^T)^{-1}| |\mathbf{g}|,$$

and the lemma follows.

We can also express the Kreiss condition as an eigenvalue condition. For $\operatorname{Re} \tilde{s} > 0$, it is the Godunov–Ryabenkii condition. Now assume that $H^I(\tilde{s}_0)$ is singular for $\tilde{s} = i\tilde{\xi}_0$, where $\tilde{\xi}_0$ is real. Let $\tilde{s} = i\tilde{\xi}_0 + \tilde{\eta}$, $\tilde{\eta} > 0$. As $\tilde{\eta} \rightarrow 0$, Eq. (12.2.20) converges to

$$\begin{aligned} \mathbf{w}^{II} &= 0, \\ \mathbf{w}_j^I &= M_{11}^{j-1}(i\tilde{\xi}_0)\mathbf{w}_1^I, \\ H^I(i\tilde{\xi}_0)\mathbf{w}_1^I &= \mathbf{g}, \end{aligned}$$

and there is a nontrivial solution for $\mathbf{g} = 0$. In general, there may be one or more eigenvalues of $M_{11}(i\tilde{\xi}_0)$ with $|\kappa(i\tilde{\xi}_0)| = 1$ and in such a case the condition $\|\hat{y}\|_h < \infty$ is violated. We make the next definition.

Definition 12.2.2. *If $\operatorname{Det}(H^I(\tilde{s}_0)) = 0$, where \tilde{s}_0 is purely imaginary, then \tilde{s}_0 is called a generalized eigenvalue of the eigenvalue problem (12.1.13) if $\|\varphi\|_h = \infty$.*

REMARK. The condition $\|\varphi\|_h < \infty$ may be fulfilled even if \tilde{s}_0 is on the imaginary axis. In such a case, \tilde{s}_0 is an eigenvalue.

We now have the following lemma.

Lemma 12.2.5. *The Kreiss condition is satisfied if, and only if, there are no eigenvalues or generalized eigenvalues for $\operatorname{Re} \tilde{s} \geq 0$.*

In most cases the easiest condition to check is Eq. (12.2.19). One need not write the system (12.2.14) in the form of Eq. (12.2.16a). Instead, we use the

representation (12.1.22) for the general solution, which we substitute into the boundary conditions to obtain

$$C(\tilde{s})\sigma = \hat{g}.$$

Then we can determine whether or not Eq. (12.2.19) holds.

We summarize our results by connecting the Kreiss condition to an estimate in physical space.

Theorem 12.2.1. *Consider the system (12.1.10) with $f = F = 0$. Its solutions y satisfy, for any fixed j , the estimate*

$$\int_0^T |y_j|^2 dt \leq K_j \int_0^T |g|^2 dt \quad (12.2.22)$$

if, and only if, the Kreiss condition is satisfied.

Proof. First assume that the Kreiss condition holds. Then, by Eq. (12.2.21) and Parseval's relation,

$$\begin{aligned} \int_0^\infty e^{-2\eta t} |y_j(t)|^2 dt &\leq K_j \int_0^\infty e^{-2\eta t} |g(t)|^2 dt \\ &\leq K_j \int_0^\infty |g(t)|^2 dt, \quad \eta > 0. \end{aligned}$$

Since the right-hand side is independent of η and the solution $y_j(t)$, $0 \leq t \leq T$, does not depend on $g(t)$ for $t > T$, Eq. (12.2.22) follows.

Next assume that Eq. (12.2.22) holds and let $T \rightarrow \infty$. Then Eq. (12.2.21) follows, as well as Eq. (12.2.19). This proves the theorem.

We now consider the example Eqs. (12.1.24) to (12.1.26) with inhomogeneous boundary conditions. We need to sharpen Lemma 12.1.9.

Lemma 12.2.6. *There is a constant $\delta > 0$ such that, on any compact set $|\tilde{s}| \leq C$, $\operatorname{Re} \tilde{s} \geq 0$, the roots κ_1 and κ_2 of the characteristic equation (12.1.28) satisfy the inequalities*

$$\begin{aligned} |\kappa_j - 1| &\geq \delta, \quad j = 1, 2, \quad \text{if } a > 0, \\ \left| 1 - \frac{1}{\kappa_1 \kappa_2} \right| &\geq \delta, \quad j = 1, 2, \quad \text{if } a < 0. \end{aligned}$$

(In fact, the second inequality holds for all a .)

Proof. The roots are continuous functions of \tilde{s} . Therefore, the inequalities can only be violated if for some \tilde{s} $\kappa_j = 1$, when $a > 0$, or $\kappa_1 \kappa_2 = 1$, when $a < 0$.

The first statement of Lemma 12.1.9 tells us that this cannot happen for $\operatorname{Re} \tilde{s} > 0$. Let $a > 0$ and $\kappa_j = 1$, then necessarily $\tilde{s} = 0$. However, by the second statement of Lemma 12.1.9, we obtain a contradiction because $\kappa_1 = -1$.

Let $a < 0$ and $\kappa_1 \kappa_2 = 1$, then Eq. (12.1.28) implies

$$\begin{aligned}\frac{\tilde{s}}{a} &= \frac{2}{3} \left(\kappa_1 - \frac{1}{\kappa_1} \right) - \frac{1}{12} \left(\kappa_1^2 - \frac{1}{\kappa_1^2} \right) \\ &= -\frac{2}{3} \left(\kappa_2 - \frac{1}{\kappa_2} \right) + \frac{1}{12} \left(\kappa_2^2 - \frac{1}{\kappa_2^2} \right) = -\frac{\tilde{s}}{a}.\end{aligned}$$

Thus, $\tilde{s} = 0$, and by Lemma 12.1.9 this leads to a contradiction. This proves the lemma.

Now we substitute the general solution

$$y_j = \sigma_1 \kappa_1^j + \sigma_2 \frac{\kappa_2^j - \kappa_1^j}{\kappa_2 - \kappa_1}$$

into the inhomogeneous boundary conditions and obtain the inhomogeneous equations (12.1.29) and (12.1.30) with nonzero right hand sides, respectively. Lemma 12.2.6 tells us that σ_1 and σ_2 are uniformly bounded, that is, the estimate (12.2.19) holds and the Kreiss condition is satisfied.

We have proved the next lemma.

Lemma 12.2.7. *The problem (12.1.24) to (12.1.26) satisfies the Kreiss condition.*

Clearly, the results can be generalized to systems (10.1.1) if we use the above approximation for every component.

We now assume that the operator Q is semibounded for the Cauchy problem; that is, the coefficient matrices are Hermitian and

$$\operatorname{Re}(w, Qw)_{-\infty, \infty} \leq 0, \quad (12.2.23)$$

for all w with $\|w\|_{-\infty, \infty} < \infty$. The semiboundedness will be destroyed by boundary terms when working with the scalar product $(u, v)_h = \sum_{j=1}^{\infty} \langle u_j, v_j \rangle h$. The following lemma shows the boundary effect.

Lemma 12.2.8. *Assume that Q defined in Eq. (12.1.11) satisfies Eq. (12.2.23). Then for every grid vector function $\{w_j\}_{j=-r+1}^{\infty}$ with $\|w\|_h < \infty$*

$$\operatorname{Re}(w, Qw)_h \leq \operatorname{Re} \sum_{j=1}^r \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \quad (12.2.24)$$

Proof. Define the difference operators

$$\begin{aligned} Q_- &= \frac{1}{h} \sum_{j=-r}^{-1} B_j E^j, \\ Q_+ &= \frac{1}{h} \sum_{j=0}^p B_j E^j, \\ Q_-^* &= \frac{1}{h} \sum_{j=-r}^{-1} B_j^* E^{-j}. \end{aligned}$$

We have

$$\begin{aligned} (w, Q_- w)_h &= \frac{1}{h} \sum_{j=-r}^{-1} (w, B_j E^j w)_h, \\ &= \sum_{j=-r}^{-1} \sum_{\nu=1}^{\infty} \langle w_\nu, B_j w_{\nu+j} \rangle = \sum_{j=-r}^{-1} \sum_{\nu=j+1}^{\infty} \langle w_{\nu-j}, B_j w_\nu \rangle, \\ &= \sum_{j=-r}^{-1} \sum_{\nu=1}^{\infty} \langle B_j^* w_{\nu-j}, w_\nu \rangle + \sum_{j=-r}^{-1} \sum_{\nu=j+1}^0 \langle B_j^* w_{\nu-j}, w_\nu \rangle, \\ &= \sum_{\nu=1}^{\infty} \sum_{j=-r}^{-1} \langle B_j^* E^{-j} w_\nu, w_\nu \rangle + \sum_{j=1}^r \sum_{\nu=1-j}^0 \langle w_{\nu+j}, B_{-j} w_\nu \rangle, \\ &= (Q_-^* w, w)_h + \sum_{j=1}^r \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \end{aligned}$$

Thus

$$\begin{aligned} \operatorname{Re}(w, Qw)_h &= \operatorname{Re}(w, Q_+ w)_h + \operatorname{Re}(w, Q_- w)_h \\ &= \operatorname{Re}(w, Q_+ w)_h + \operatorname{Re}(w, Q_-^* w)_h \\ &\quad + \operatorname{Re} \sum_{j=1}^r \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \quad (12.2.25) \end{aligned}$$

Define a new and extended grid vector function by

$$\tilde{w}_j = \begin{cases} w_j, & \text{for } j \geq 1, \\ 0, & \text{for } j \leq 0. \end{cases}$$

Then

$$\begin{aligned} (w, Q_-^* w)_h &= (\tilde{w}, Q_-^* \tilde{w})_{-\infty, \infty}, \\ &= \sum_{\nu=-\infty}^{\infty} \left\langle \tilde{w}_{\nu}, \sum_{j=-r}^{-1} B_j^* \tilde{w}_{\nu-j} \right\rangle, \\ &= \sum_{\nu=-\infty}^{\infty} \left\langle \sum_{j=-r}^{-1} B_j \tilde{w}_{\nu}, \tilde{w}_{\nu-j} \right\rangle, \\ &= \sum_{\nu=-\infty}^{\infty} \left\langle \sum_{j=-r}^{-1} B_j \tilde{w}_{\nu+j}, \tilde{w}_{\nu} \right\rangle, \\ &= (Q_- \tilde{w}, \tilde{w})_{-\infty, \infty}; \end{aligned}$$

that is,

$$\begin{aligned} \operatorname{Re}(w, (Q_+ + Q_-^*)w)_h &= \operatorname{Re}(\tilde{w}, Q_+ \tilde{w})_{-\infty, \infty} + \operatorname{Re}(Q_- \tilde{w}, \tilde{w})_{-\infty, \infty}, \\ &= \operatorname{Re}(\tilde{w}, (Q_+ + Q_-) \tilde{w})_{-\infty, \infty} \\ &= \operatorname{Re}(\tilde{w}, Q \tilde{w})_{-\infty, \infty} \leq 0. \end{aligned}$$

The lemma then follows by Eq. (12.2.25).

We can now prove the following theorem.

Theorem 12.2.2. *Consider the system (12.1.10) with $F \equiv f \equiv 0$, and assume that Q is semibounded for the Cauchy problem and that the Kreiss condition holds. Then we obtain, for the solution y at every fixed T ,*

$$\|y(T)\|_h^2 \leq \text{constant} \int_0^T |g(t)|^2 dt. \quad (12.2.26)$$

Proof. By Lemma 12.2.8, we have

$$\frac{d}{dt} \|y\|_h^2 = 2 \operatorname{Re}(y, Qy)_h \leq \text{constant} \sum_{\nu=-r+1}^r |y_{\nu}|^2,$$

and after integration the estimate follows using Theorem 12.2.1.

The last theorem guarantees that we obtain an estimate for the solutions of Eq. (12.1.10). We extend the definition of f_j to $-\infty < j < \infty$ and solve a Cauchy problem. Subtracting its solution from u , we obtain a problem with $F = f = 0$. We shall not pursue this approach. Instead, we give conditions such that Eq. (12.1.10) is strongly stable.

As in the example, we now solve an auxiliary problem

$$\frac{dw_j}{dt} = Qw_j, \quad j = 1, 2, \dots, \quad (12.2.27a)$$

$$w_j(0) = f_j, \quad (12.2.27b)$$

$$\tilde{L}_0 w_0(t) = 0, \quad (12.2.27c)$$

$$\|w\|_h < \infty. \quad (12.2.27d)$$

To obtain the desired result, we need to construct boundary conditions such that the boundary values can be estimated in terms of $\|f\|_h$. We want to prove the next lemma.

Lemma 12.2.9. *Assume that Eq. (12.2.23) holds. Then, we can find boundary conditions (12.2.27c) such that, for all grid functions w_j , with $\|w\|_h < \infty$, that satisfy these boundary conditions,*

$$\operatorname{Re}(w, Qw)_h \leq -\delta \sum_{\nu=-r+1}^r |w_\nu|^2, \quad \delta > 0. \quad (12.2.28)$$

Proof. By Lemma 12.2.8,

$$\begin{aligned} \operatorname{Re}(w, Qw)_h &\leq \operatorname{Re} \sum_{j=1}^r \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle, \\ &= \operatorname{Re} \sum_{\nu=0}^{r-1} \langle w_{r-\nu}, B_{-r} w_{-\nu} \rangle \\ &\quad + \operatorname{Re} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \end{aligned} \quad (12.2.29)$$

Let τ be a constant with $0 < \tau \leq 1$ and choose the boundary conditions by

$$w_{r-\nu} = -\tau^\nu B_{-r} w_{-\nu}, \quad (12.2.30a)$$

that is,

$$w_\mu = -\tau^{r-\mu} B_{-r} w_{\mu-r}, \quad \mu = 1, 2, \dots, r. \quad (12.2.30b)$$

Then we get, by Eq. (12.2.29),

$$\begin{aligned} \operatorname{Re}(w, Qw)_h &= -\operatorname{Re} \sum_{\nu=0}^{r-1} \tau^\nu |B_{-\nu} w_{-\nu}|^2 \\ &\quad + \operatorname{Re} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \langle w_{j-\nu}, B_{-j} w_{-\nu} \rangle. \end{aligned} \quad (12.2.31)$$

Because B_{-r} is nonsingular, we have

$$|B_{-r}v| \geq \text{constant } |v| \geq \text{constant } |B_{-j}v|, \quad j = 1, \dots, r-1,$$

for any vector v .

Thus, by using Eq. (12.2.31) and the boundary conditions (12.2.30) once more

$$\begin{aligned} &\operatorname{Re}(w, Qw)_h + \sum_{\nu=0}^{r-1} \tau^\nu |B_{-\nu} w_{-\nu}|^2 \\ &\leq \text{constant} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} |w_{j-\nu}| |B_{-r} w_{-\nu}|, \\ &\leq \text{constant} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{r+\nu-j} |B_{-r} w_{j-\nu-r}| |B_{-r} w_{-\nu}|, \\ &\leq \text{constant} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{(r-j)/2} (\tau^{(r-j+\nu)/2} |B_{-r} w_{j-\nu-r}|) (\tau^{r/2} |B_{-r} w_{-\nu}|), \\ &\leq \text{constant} \sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^{(r-j)/2} (\tau^{r-j+\nu} |B_{-r} w_{j-\nu-r}|^2 + \tau^\nu |B_{-r} w_{-\nu}|^2). \end{aligned}$$

Because $0 < \tau \leq 1$ and $r-j \geq 1$, we have $\tau^{(r-j)/2} \leq \tau^{1/2}$. Furthermore, we have $0 \leq \nu \leq r-2$ and $1 \leq r-j+\nu \leq r-1$ in the double sum; that is,

$$\sum_{j=1}^{r-1} \sum_{\nu=0}^{j-1} \tau^\nu |B_{-\nu} w_{-\nu}|^2 \leq \text{constant} \sum_{\nu=0}^{r-1} \tau^\nu |B_{-\nu} w_{-\nu}|^2$$

and

$$\sum_{j=0}^{r-1} \sum_{\nu=0}^{j-1} \tau^{r-j+\nu} |B_{-r} w_{j-\nu-r}|^2 \leq \text{constant} \sum_{\nu=0}^{r-1} \tau^\nu |B_{-r} w_{-\nu}|^2.$$

Thus

$$\operatorname{Re}(w, Qw)_h + \sum_{\nu=0}^{r-1} \tau^\nu |B_{-r} w_{-\nu}|^2 \leq \text{constant} \tau^{1/2} \sum_{\nu=0}^{r-1} \tau^\nu |B_{-r} w_{-\nu}|^2.$$

By choosing τ sufficiently small but positive and using the fact that B_{-r} is nonsingular, we get

$$\operatorname{Re}(w, Qw)_h \leq -\delta_1 \sum_{\nu=0}^{r-1} |B_{-r} w_{-\nu}|^2. \quad (12.2.32)$$

It remains to include w_1, \dots, w_r in the estimate. This is achieved by using the boundary conditions once more. We have

$$\sum_{\mu=1}^r |w_\mu|^2 \leq \sum_{\mu=1}^r |B_{-r} w_{\mu-r}|^2 = \sum_{\nu=0}^{r-1} |B_{-r} w_{-\nu}|^2$$

and since B_{-r} is nonsingular we can split the right-hand side of Eq. (12.2.32) in two parts and obtain

$$\operatorname{Re}(w, Qw)_h \leq -\frac{\delta_1}{2} |B_{-r}^{-1}|^{-1} \sum_{\nu=0}^{r-1} |w_{-\nu}|^2 - \frac{\delta_1}{2} \sum_{\mu=1}^r |w_\mu|^2.$$

This proves the lemma.

This lemma provides estimates for the boundary values $w_j, j = -r+1, \dots, r$. The original problem may have boundary conditions including $v_j, j > r$, therefore, we also need estimates for $w_j, j > r$. We now have the following lemma.

Lemma 12.2.10. *Assume that $r \geq p$ and that the boundary conditions (12.2.27c) are such that Eq. (12.2.28) holds. Then we obtain for every fixed j*

$$\int_0^\infty |w_j(t)|^2 dt \leq \text{constant} \|f\|_h \quad (12.2.33)$$

Proof. By Lemma 12.2.9, we have

$$\frac{d}{dt} \|w\|_h^2 = 2 \operatorname{Re}(w, Qw)_h \leq -2\delta \sum_{\nu=-r+1}^r |w_\nu|^2, \quad \delta > 0,$$

and after integration

$$\|w(t)\|_h^2 + 2\delta \int_0^t \sum_{\nu=-r+1}^r |w_\nu(\tau)|^2 d\tau \leq \|f\|_h^2.$$

Therefore, Eq. (12.2.33) follows for $-r+1 \leq j \leq r$.

To derive the estimate for $j > r$, we Laplace transform Eq. (12.2.27) and obtain

$$s\hat{w}_j = Q\hat{w}_j + f_j, \quad j = 1, 2, \dots, \quad (12.2.34a)$$

$$\tilde{L}_0\hat{w}_0 = 0, \quad (12.2.34b)$$

$$\|\hat{w}\|_h < \infty. \quad (12.2.34c)$$

We consider the two cases $|\tilde{s}| = |sh| \geq C_0$ and $|sh| < C_0$ where C_0 is a large number. We have

$$\begin{aligned} \|\hat{w}\|_h^2 &\leq \frac{2}{|sh|^2} \|hQw\|_h^2 + \frac{2}{|s|^2} \|f\|_h^2, \\ &\leq \text{constant} \left(\frac{1}{|sh|^2} \|w\|_h^2 + \frac{h}{|sh|^2} \sum_{\nu=-r+1}^0 |\hat{w}_\nu|^2 + \frac{1}{|s|^2} \|f\|_h^2 \right), \end{aligned}$$

that is, for large $|sh|$

$$\|\hat{w}\|_h^2 \leq \text{constant} \left(h \sum_{\nu=-r+1}^0 |\hat{w}_\nu|^2 + \frac{1}{|s|^2} \|f\|_h^2 \right).$$

It then follows that

$$|\hat{w}_j|^2 \leq \frac{1}{h} \|\hat{w}\|_h^2 \leq \text{constant} \left(\sum_{\nu=-r+1}^0 |\hat{w}_\nu|^2 + \frac{1}{h|s|^2} \|f\|_h^2 \right). \quad (12.2.35)$$

Next, we consider the case $|sh| \leq C_0$, and write Eq. (12.2.34) in the one-step

form [see Eq. (12.2.16a)]

$$\mathbf{y}_{j+1} = M\mathbf{y}_j + h\mathbf{f}_j, \quad |\mathbf{f}_j| \leq \text{constant } |f_j|.$$

For any j , we have

$$\mathbf{y}_j = M^{j-1}\mathbf{y}_1 + h \sum_{\nu=1}^{j-1} M^{j-1-\nu}\mathbf{f}_\nu,$$

that is,

$$\begin{aligned} |\mathbf{y}_j|^2 &\leq \text{constant} \left(|\mathbf{y}_1|^2 + h^2 \sum_{\nu=1}^{j-1} |f_\nu|^2 \right) \\ &\leq \text{constant} (|\mathbf{y}_1|^2 + h\|f\|_h^2). \end{aligned}$$

Because $|\mathbf{y}_1|^2 = \sum_{\nu=-r+1}^p |\hat{w}_\nu|^2$, and by assumption $p \leq r$, we obtain for every fixed j

$$|\mathbf{y}_j|^2 \leq \text{constant} \left(\sum_{\nu=-r+1}^r |\hat{w}_\nu|^2 + h\|f\|_h^2 \right), \quad |sh| \leq C_0. \quad (12.2.36)$$

Let $s = i\xi$ and integrate. We get, from Eqs. (12.2.35) and (12.2.36),

$$\begin{aligned} &\int_{-\infty}^{\infty} |\mathbf{y}_j(i\xi)|^2 d\xi \\ &\leq \text{constant} \left(\int_{-\infty}^{\infty} \sum_{\nu=-r+1}^r |\hat{w}_\nu(i\xi)|^2 d\xi \right. \\ &\quad \left. + h\|f\|_h^2 \int_{-C_0/h}^{C_0/h} d\xi + \frac{1}{h} \|f\|_h^2 \int_{|\xi| \geq C_0/h} \frac{d\xi}{|\xi|^2} \right), \\ &\leq \text{constant} \left(\int_{-\infty}^{\infty} \sum_{\nu=-r}^r |\hat{w}(i\xi)|^2 d\xi + \|f\|_h^2 \right). \end{aligned}$$

By the definition of \mathbf{y}_j , Parseval's relation, and the fact that Eq. (12.2.33) holds for $-r+1 \leq j \leq r$, we obtain Eq. (12.2.33) for any fixed j .

Now we can prove our main theorem.

Theorem 12.2.3. Assume that $r \geq p$, that the difference operator Q is semi-bounded for the Cauchy problem, and that the Kreiss condition is satisfied. Then the approximation (12.1.10) is strongly stable.

Proof. In the general case, we have a nonzero forcing function $F_j(t)$. The auxiliary problem (12.2.27) is modified to

$$\begin{aligned} \frac{dw_j}{dt} &= Qw_j + F_j, \quad j = 1, 2, \dots, \\ w_j(0) &= f_j, \\ \tilde{L}_0 w_0(t) &= 0. \end{aligned} \tag{12.2.37}$$

The technique in Lemma 12.2.10 can still be used. The transformed equation (12.2.34a) now becomes

$$s\hat{w}_j = Q\hat{w}_j + f_j + \hat{F}_j, \quad j = 1, 2, \dots,$$

leading to the estimate

$$\int_0^\infty |w_j(t)|^2 dt \leq \text{constant} \left(\|f\|_h^2 + \int_0^\infty \|F(t)\|_h^2 dt \right) \tag{12.2.38}$$

instead of Eq. (12.2.33) (assuming that the integral of $\|F\|_h^2$ exists). We also have, as usual with the energy method,

$$\|w(t)\|_h^2 \leq \text{constant} \left(\|f\|_h^2 + \int_0^t \|F(\tau)\|_h^2 d\tau \right). \tag{12.2.39}$$

The difference $y = v - w$ satisfies

$$\begin{aligned} \frac{dy_j}{dt} &= Qy_j, \quad j = 1, 2, \dots, \\ y_j(0) &= 0, \\ L_0 y_0 &= g(t) + g_1(t), \end{aligned}$$

where $\int_0^\infty |g_1(t)|^2 dt \leq \text{constant} \|f\|_h^2$. The theorem follows from Theorem 12.2.2, and from Eqs. (12.2.28) and (12.2.29).

REMARK. Because, in practical applications, we deal mainly with problems in finite intervals $0 \leq x \leq L$, $t \geq 0$, the only interesting case is $r = p$. This is typical

for nondissipative approximations. In Section 12.4, dissipative approximations are considered and the assumption will be removed.

EXERCISES

- 12.2.1. Carry out the details in the derivation of the estimate (12.2.38), that is, prove Lemma 12.2.10 for $F \neq 0$.
- 12.2.2. Prove that the fourth-order approximation (12.1.24) is strongly stable with the boundary conditions

$$\begin{aligned} v_{-1} &= g_{-1} \\ v_0 &= g_0 \end{aligned}$$

for any value of a .

12.3. A FOURTH-ORDER ACCURATE APPROXIMATION FOR HYPERBOLIC DIFFERENTIAL EQUATIONS

We consider the quarter-space problem for a hyperbolic system

$$\begin{aligned} u_t &= Au_x, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \tag{12.3.1}$$

and approximate it by the fourth-order accurate approximation

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, \quad j = 1, 2, \dots, \\ v_j(0) &= f_j, \end{aligned} \tag{12.3.2a}$$

where

$$Q = A\left(\frac{4}{3}D_0(h) - \frac{1}{3}D_0(2h)\right).$$

Because A can be diagonalized, the approximation can be decoupled into a set of scalar equations. Therefore, we can assume that A already is diagonal with

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \quad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

In this case, the boundary conditions can be written as

$$u^{II}(0, t) = R^I u(0, t). \quad (12.3.3)$$

We differentiate the boundary condition (12.3.3) twice with respect to t and obtain

$$u_{xx}^{II}(0, t) = S^I u_{xx}^I(0, t), \quad S^I = (\Lambda^{II})^{-2} R^I (\Lambda^I)^2. \quad (12.3.4)$$

As boundary conditions for the ingoing characteristic variables u^{II} , we use

$$v_0^{II}(t) = R^I v_0^I(t), \quad D_+ D_- v_0^{II}(t) = S^I D_+ D_- v_0^I(t). \quad (12.3.2b)$$

For the outgoing characteristic variables, we use extrapolation conditions

$$D_+^q v_0^I(t) = D_+^q v_{-1}^I(t) = 0. \quad (12.3.2c)$$

We want to prove the following theorem.

Theorem 12.3.1. *The approximation (12.3.2) is strongly stable and fourth-order accurate if $q \geq 4$.*

Proof. The approximation for the outgoing characteristic variables v^I is decoupled from that of the ingoing characteristic variables v^{II} , and we have already discussed it in the previous two sections as an example. Our results tell us that the approximation for v^I is strongly stable and that, for every fixed j ,

$$\int_0^\infty |v_j^I(t)|^2 dt \leq \text{constant} \|f^I\|_h^2.$$

Thus, we can think of v^I as a given function and write the boundary conditions for v^{II} in the form

$$v_0^{II}(t) = g_0(t), \quad g_0 = R^I v_0^I, \\ h^2 D_+ D_- v_0^{II}(t) = g_{-1}(t), \quad g_{-1} = h^2 S^I D_+ D_- v_0^I.$$

Now we can think of the approximation for v^{II} as consisting of scalar equations that we also have discussed in the previous two sections. By Lemma 12.2.7 and Theorem 12.2.3, they are strongly stable.

In Section 12.7, we shall further comment on the accuracy of this method and show that the error is $\mathcal{O}(h^4)$.

REMARK. It is, of course, not necessary that A be in diagonal form. However, the extrapolation conditions should be applied to the outgoing characteristic variables.

EXERCISES

12.3.1. Consider the linearized Euler equations

$$\begin{bmatrix} u \\ \rho \end{bmatrix}_t + \begin{bmatrix} U & a^2/R \\ R & U \end{bmatrix} \begin{bmatrix} u \\ \rho \end{bmatrix}_x = 0, \quad 0 \leq x < \infty, \quad t \geq 0,$$

$$u(x, 0) = f(x),$$

$$u(0, t) = 0.$$

Formulate the boundary conditions (12.3.2b), expressed in the original variables ρ, u for the fourth-order approximation (12.3.2a).

12.3.2. Consider the hyperbolic problem

$$u_t = Au_x + F, \quad 0 \leq x < \infty, \quad t \geq 0,$$

$$u(x, 0) = f(x),$$

$$u''(0, t) = R^I u'(0, t) + g^{II}(t),$$

where $u = (u^I, u^{II})^T$,

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \quad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

Formulate the boundary conditions corresponding to Eq. (12.3.2b) for the fourth-order approximation (12.3.2a).

12.4. STABILITY IN THE GENERALIZED SENSE FOR HYPERBOLIC SYSTEMS

We again consider the approximation (12.1.10). In Section 10.3, we introduced the definition of generalized well-posedness for the continuous case, and the same concept will be used for semidiscrete approximations. Corresponding to the analytic case, we make the following definition.

Definition 12.4.1. Consider the approximation (12.1.10) with $f \equiv g \equiv 0$. We call the problem stable in the generalized sense if, for all sufficiently small h , there is a unique solution that satisfies the estimate

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 dt \leq K(\eta) \int_0^\infty e^{-2\eta t} \|F(t)\|_h^2 dt, \quad (12.4.1a)$$

for all $\eta > \eta_0$. Here η_0 and $K(\eta)$ are constants that do not depend on F and, furthermore,

$$\lim_{\eta \rightarrow \infty} K(\eta) = 0. \quad (12.4.2)$$

If only $f \equiv 0$, we call the approximation strongly stable in the generalized sense if, instead of Eq. (12.4.1a),

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 dt \leq K(\eta) \left(\int_0^\infty e^{-2\eta t} (\|F(t)\|_h^2 + |g(t)|^2) dt \right) \quad (12.4.1b)$$

holds.

The restriction to homogeneous initial and boundary conditions was discussed in Section 10.3 for the continuous problem and the same discussion applies here. If a smooth function can be constructed that satisfies the inhomogeneous conditions, then it can be subtracted from the original solution, and the difference will satisfy a problem of the required form.

The leading terms of the operator Q are of the order h^{-1} , and this part of the operator is called the *principal part*. There may also be lower order terms, as in the next theorem.

Theorem 12.4.1. Assume that the approximation (12.1.10) is stable in the generalized sense and that Q_0 is a lower order operator that is bounded independent of h . Then the perturbed problem

$$\begin{aligned} \frac{dv_j}{dt} &= (Q + Q_0)v_j + F_j, \quad j = 1, 2, \dots, \\ v_j(0) &= 0, \\ L_0 v_0(t) &= 0, \end{aligned} \quad (12.4.3)$$

has the same property.

Proof. By considering Q_0v_j as a forcing function, we get, by assumption,

$$\begin{aligned} \int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 dt &\leq 2K(\eta) \int_0^\infty e^{-2\eta t} (\|Q_0v(t)\|_h^2 + \|F(t)\|_h^2) dt \\ &\leq K_1(\eta) \int_0^\infty e^{-2\eta t} (\|v(t)\|_h^2 + \|F(t)\|_h^2) dt \\ K_1(\eta) &\rightarrow 0, \quad \eta \rightarrow \infty. \end{aligned}$$

By choosing η sufficiently large, we obtain

$$\int_0^\infty e^{-2\eta t} \|v(t)\|_h^2 dt \leq \frac{K_1(\eta)}{1 - K_1(\eta)} \int_0^\infty e^{-2\eta t} \|F(t)\|_h^2 dt, \quad \eta > \eta_1,$$

which proves the theorem.

Therefore, we can again assume that the coefficients B_ν and $L_{\mu j}$ do not depend on h , that is, that $B_\nu = B_{\nu_0}$.

We can also consider the strip problem. Then there are also boundary conditions $L_N v_N = g_N$. As in the continuous case, one can prove the following theorem.

Theorem 12.4.2. *The approximation is stable in the generalized sense for the strip problem if it is stable for the periodic problem and stable in the generalized sense for the right and left quarter-space problems.*

Therefore, we need only consider the quarter-space problem. Let $f \equiv 0$. The Laplace transformed problem (12.1.10) is

$$\begin{aligned} s\hat{v}_j &= Q\hat{v}_j + \hat{F}_j, \quad j = 1, 2, \dots, \\ L_0\hat{v}_0 &= \hat{g}, \quad \|\hat{v}\|_h < \infty, \end{aligned} \tag{12.4.4}$$

which is also called the *resolvent equation*.

As in the continuous case, we can prove the next theorem.

Theorem 12.4.3. *Let $f \equiv g \equiv 0$. The approximation (12.1.10) is stable in the generalized sense if, and only if, Eq. (12.4.4) has a unique solution that satisfies*

$$\|\hat{v}\|_h^2 \leq K(\eta) \|\hat{F}\|_h^2, \tag{12.4.5a}$$

for all $s = i\xi + \eta$, $\eta > \eta_0$ where $\lim_{\eta \rightarrow \infty} K(\eta) = 0$. If only $f \equiv 0$, then it is strongly stable in the generalized sense if, instead of Eq. (12.4.5a),

$$\|\hat{v}\|_h^2 \leq K(\eta)(\|\hat{F}\|_h^2 + |\hat{g}|^2) \quad (12.4.5b)$$

holds.

REMARK. If we only consider the principle part of Q , then we can choose $\eta_0 = 0$. However, in the general case with lower order terms in the difference equation, or when there are two boundaries, then we may have $\eta_0 > 0$.

The main result of this section is Theorem 12.4.4.

Theorem 12.4.4. *Assume that the differential equations of the underlying problem (10.1.1) are strictly hyperbolic (i.e., the eigenvalues of A are distinct) and that the approximation (12.1.10) is dissipative and consistent. Then the approximation is strongly stable in the generalized sense if the Kreiss condition is satisfied.*

The proof requires a number of lemmas. The first one gives estimates for the solutions of one-step difference equations.

Lemma 12.4.1. *Let D be an $m \times m$ matrix, where m is fixed, G_j is a discrete vector function with $\|G\|_h < \infty$ and consider the system*

$$\begin{aligned} y_{j+1} &= Dy_j + hG_j, \quad j = 1, 2, \dots, \\ \|y\|_h &< \infty. \end{aligned} \quad (12.4.6)$$

If $|D| < 1$, then the solutions of Eq. (12.4.6) satisfy the estimate

$$\|y\|_h \leq \frac{h}{1 - |D|} \|G\|_h + \left(\frac{h}{1 - |D|^2} \right)^{1/2} |y_1|. \quad (12.4.7)$$

If $|D^{-1}| < 1$, then Eq. (12.4.6) has a unique solution that satisfies the estimates

$$\|y\|_h \leq \frac{|D^{-1}|h}{1 - |D^{-1}|} \|G\|_h \quad (12.4.8)$$

and

$$|y_1| \leq \frac{h^{1/2} \|G\|_h}{(1 - |D^{-1}|^2)^{1/2}}. \quad (12.4.9)$$

Proof. For $|D| < 1$, the solution of Eq. (12.4.6) is a sum of the particular solution with $y_1 = 0$ and the solution of the homogeneous equation. Let $y_1 = 0$,

then

$$|y_{j+1}|^2 = \langle y_{j+1}, Dy_j \rangle + h \langle y_{j+1}, G_j \rangle$$

implies

$$\|y\|_h^2 \leq |D| \|y\|_h^2 + h \|y\|_h \|G\|_h,$$

that is,

$$(1 - |D|) \|y\|_h \leq h \|G\|_h.$$

If $G \equiv 0$, then

$$y_j = D^{j-1} y_1, \quad j = 1, 2, \dots$$

Therefore, for any solution y with $\|y\|_h < \infty$

$$\|y\|_h^2 \leq \sum_{j=1}^{\infty} |D|^{2j-2} |y_1|^2 h = \frac{h}{1 - |D|^2} |y_1|^2,$$

and Eq. (12.4.7) follows.

If $|D^{-1}| < 1$, then we write Eq. (12.4.6) in the form

$$y_j = D^{-1} y_{j+1} - h D^{-1} G_j, \quad j = 1, 2, \dots, \quad (12.4.10)$$

and we obtain

$$|y_j|^2 = \langle y_j, D^{-1} E y_j \rangle - h \langle y_j, D^{-1} G_j \rangle, \quad j = 1, 2, \dots$$

Therefore, for any solution y with $\|y\|_h < \infty$

$$\|y\|_h^2 \leq |D^{-1}| (\|y\|_h \|E y\|_h + h \|y\|_h \|G\|_h);$$

that is,

$$(1 - |D^{-1}|) \|y\|_h \leq h |D^{-1}| \|G\|_h,$$

and Eq. (12.4.8) follows. In particular, we have $y_j \rightarrow 0$ as $j \rightarrow \infty$.

The homogeneous equation (12.4.10) has the form

$$u_j = D^{-1} u_{j+1}, \quad j = 1, 2, \dots;$$

that is,

$$u_j = D^{-(k-j)} u_k, \quad k > j.$$

Therefore, since $y_j \rightarrow 0$ as $j \rightarrow \infty$, Duhamel's principle gives us, for Eq. (12.4.10) the unique solution,

$$y_v = h \sum_{j=v}^{\infty} D^{-j} G_j,$$

and therefore, for $j = 1$

$$\begin{aligned} |y_1| &\leq h \sum_{j=1}^{\infty} |D^{-1}|^j |G_j| \leq h^{1/2} \left(\sum_{j=1}^{\infty} |D^{-1}|^{2j} \right)^{1/2} \|G\|_h, \\ &\leq \frac{h^{1/2} \|G\|_h}{(1 - |D^{-1}|^2)^{1/2}}. \end{aligned}$$

This proves the lemma.

We now consider Eq. (12.1.10) with $f = 0$. To estimate the solutions of the Laplace transformed equations (12.4.4), we divide the right half of the complex \tilde{s} plane into three parts:

- I. $|\tilde{s}| \geq c_0$, $\operatorname{Re} \tilde{s} \geq 0$, where the constant c_0 is large.
- II. $0 < c_1 \leq |\tilde{s}| \leq c_0$, $\operatorname{Re} \tilde{s} \geq 0$, where c_1 is a small constant.
- III. $|\tilde{s}| \leq c_1$, $\operatorname{Re} \tilde{s} \geq 0$.

I. $|\tilde{s}| \geq c_0$, $\operatorname{Re} \tilde{s} \geq 0$

Lemma 12.4.2. *There are constants c_0 and K_0 such that, for $|\tilde{s}| = |sh| \geq c_0$, the solutions of Eq. (12.4.4) satisfy the estimate*

$$\|\hat{v}\|_h^2 \leq K_0 \left(\frac{1}{|s|^2} \|\hat{F}\|^2 + \frac{1}{|s|} |\hat{g}|^2 \right).$$

Proof. We have

$$\begin{aligned}\|\hat{v}\|_h^2 &\leq \frac{2}{|sh|^2} \|hQ\hat{v}\|_h^2 + \frac{2}{|s|^2} \|\hat{F}\|_h^2, \\ &\leq \frac{\text{constant}}{|sh|^2} (\|\hat{v}\|_h^2 + h|\hat{g}|^2) + \frac{1}{|s|^2} \|\hat{F}\|_h^2.\end{aligned}$$

By choosing $|sh| \geq c_0$, c_0 large enough, we get

$$\begin{aligned}\|\hat{v}\|_h^2 &\leq \text{constant} \left(\frac{1}{|sh| \cdot |s|} |\hat{g}|^2 + \frac{1}{|s|^2} \|\hat{F}\|_h^2 \right) \\ &\leq \text{constant} \left(\frac{1}{c_0 |s|} |\hat{g}|^2 + \frac{1}{|s|^2} \|\hat{F}\|_h^2 \right),\end{aligned}$$

the desired estimate follows.

II. $0 < c_1 \leq |\tilde{s}| \leq c_0$

Lemma 12.4.3. *For any constant $c_1 > 0$, there is a constant $\tau > 0$ such that, for $c_1 \leq |\tilde{s}| \leq c_0$, where $\operatorname{Re} \tilde{s} \geq 0$, the solutions κ of the characteristic equation (12.1.18) satisfy*

$$|\kappa| \leq 1 - \tau \quad \text{or} \quad |\kappa| > 1 + \tau. \quad (12.4.11)$$

Proof. We know that $|\kappa| \neq 1$ for $\operatorname{Re} \tilde{s} > 0$. Assume now that there is a sequence $\{\tilde{s}^{(\nu)}\}$, where $\operatorname{Re} \tilde{s}^{(\nu)} > 0$, with solutions $\{\kappa^{(\nu)}\}$ and $|\kappa^{(\nu)}|$ less than 1 such that $\tilde{s}^{(\nu)} \rightarrow \tilde{s}_0$, $\operatorname{Re} \tilde{s}_0 = 0$, and $\kappa^{(\nu)} \rightarrow e^{i\xi}$, ξ real. By Eq. (12.1.18), \tilde{s}_0 is an eigenvalue of $Q = \sum_\nu B_\nu e^{i\nu\xi}$. By dissipativity, $\operatorname{Re} \tilde{s}_0 < 0$ if $\xi \neq 0$, which is impossible under our assumption. By consistency, $\sum_\nu B_\nu = 0$ (see Lemma 12.4.5), implying that $\tilde{s}_0 = 0$ for $\xi = 0$, which is also excluded from the \tilde{s} -domain under consideration. This proves the lemma.

We now write Eq. (12.4.4) as the one-step method (12.2.16),

$$\mathbf{y}_{j+1} = M\mathbf{y}_j + h\mathbf{F}_j, \quad j = 1, 2, \dots, \quad (12.4.12a)$$

and use the difference equation to eliminate $\hat{v}_{p+1}, \hat{v}_{p+2}, \dots, v_q$ from the boundary condition which leads to

$$H\mathbf{y}_1 = \mathbf{g}. \quad (12.4.12b)$$

Here,

$$\begin{aligned}\mathbf{F}_j &= (B_p^{-1} \hat{\mathbf{F}}_j, 0, \dots, 0)^T, \\ |\mathbf{g}| &\leq \text{constant} \left(h \sum_{j=1}^{q-p} |\hat{\mathbf{F}}_j| + |\hat{g}| \right) \\ &\leq \text{constant} (h^{1/2} \|\hat{\mathbf{F}}\|_h + |\hat{g}|).\end{aligned}\quad (12.4.12c)$$

By Eq. (12.4.11), there is a transformation $S(\tilde{s})$ that is a smooth function of \tilde{s} , for $c_1 \leq |\tilde{s}| \leq c_0$, and $\operatorname{Re} \tilde{s} \geq 0$, such that

$$S^{-1}MS = \begin{bmatrix} M^I & 0 \\ 0 & M^{II} \end{bmatrix}$$

where

$$(M^I)^* M^I \leq 1 - \tau/2, \quad (12.4.13a)$$

$$(M^{II})^* M^{II} \geq 1 + \tau/2. \quad (12.4.13b)$$

Now introduce a new variable $\mathbf{y} = S\mathbf{w}$ into Eq. (12.4.12). We obtain

$$\begin{aligned}\mathbf{w}_{j+1}^I &= M^I \mathbf{w}_j^I + h \tilde{\mathbf{F}}_j^I, \\ \mathbf{w}_{j+1}^{II} &= M^{II} \mathbf{w}_j^{II} + h \tilde{\mathbf{F}}_j^{II}, \quad \tilde{\mathbf{F}}_j = S^{-1} \tilde{\mathbf{F}}_j\end{aligned}\quad (12.4.14)$$

with boundary conditions

$$H^I \mathbf{w}_1^I + H^{II} \mathbf{w}_1^{II} = \mathbf{g}.$$

By Lemma 12.2.3, the Kreiss condition is satisfied if, and only if, $H^I(\tilde{s})$ is nonsingular for $\operatorname{Re} \tilde{s} \geq 0$.

We can now prove the following lemma.

Lemma 12.4.4. *Assume that the Kreiss condition is satisfied and consider \tilde{s} for $0 < c_1 \leq |\tilde{s}| \leq c_0$. For every $c_1 > 0$, there exists a constant K_2 such that the solutions of Eq. (12.4.4) satisfy the estimates*

$$\begin{aligned}\|\hat{v}\|_h &\leq K_2 (h \|\hat{\mathbf{F}}\|_h + h^{1/2} |\hat{g}|), \\ |\hat{v}_j| &\leq K_2 (h^{1/2} \|\hat{\mathbf{F}}\|_h + |\hat{g}|), \quad j \text{ fixed.}\end{aligned}$$

Proof. The inequality (12.4.13b) implies that $|(\mathbf{M}^{II})^{-1}| \leq (1 + \tau/2)^{-1/2} < 1$. The inequalities (12.4.8) and (12.4.9) then give us

$$\|\mathbf{w}^{II}\|_h \leq \text{constant } h \|\tilde{\mathbf{F}}^{II}\|_h, \quad |\mathbf{w}_1^{II}| \leq \text{constant } h^{1/2} \|\tilde{\mathbf{F}}^{II}\|_h$$

for the solutions w of Eq. (12.4.14). The Kreiss condition implies that

$$|\mathbf{w}_1^I| \leq \text{constant} (|\mathbf{w}_1^{II}| + |\mathbf{g}|) \leq \text{constant} (h^{1/2} \|\tilde{\mathbf{F}}^{II}\|_h + |\mathbf{g}|).$$

Therefore, by Eqs. (12.4.7) and (12.4.12c),

$$\|\mathbf{w}^I\|_h \leq \text{constant} (h \|\tilde{\mathbf{F}}\|_h + h^{1/2} |\hat{g}|),$$

and the lemma follows.

III. $|\tilde{s}| \leq c_1 \ll 1$, $\operatorname{Re} \tilde{s} \geq 0$

Now we consider Eq. (12.4.12) for $|\tilde{s}| \leq c_1 \ll 1$. The eigenvalues and eigenvectors of M are the solutions of

$$\begin{bmatrix} \tilde{B}_{p-1} & \cdot & \cdot & \cdot & \tilde{B}_{-r} \\ I & 0 & \cdot & \cdot & 0 \\ \vdots & & & & \\ 0 & \cdot & \cdot & \cdot & I & 0 \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_{p+r} \end{bmatrix} = \kappa \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_{p+r} \end{bmatrix},$$

which, by Eq. (12.1.17), is equivalent to

$$\left(\tilde{s}I - \sum_{\nu=-r}^p B_\nu \kappa^\nu \right) \varphi_1 = 0. \quad (12.4.15)$$

We need the following lemma.

Lemma 12.4.5. *Let Eq. (12.1.10a) be a consistent approximation of Eq. (10.1.1). Then*

$$\sum_{\nu=-r}^p B_\nu = 0, \quad \sum_{\nu=-r}^p \nu B_\nu = A.$$

Proof. Apply Eq. (12.1.11) to a smooth function ψ . Taylor expansion gives us

$$Q\psi(x) = \frac{1}{h} \sum_{\nu=-r}^p B_\nu \psi(x) + \sum_{\nu=-r}^p \nu B_\nu \psi'(x) + \mathcal{O}(h).$$

Let $F = 0$ in Eq. (10.1.1a). Then consistency implies $\lim_{h \rightarrow 0} Q\psi = A\psi'$, and the lemma follows.

By using the identity $\kappa^\nu = (1 + (\kappa - 1))^\nu = 1 + \nu(\kappa - 1) + \mathcal{O}(|\kappa - 1|^2)$, we can write Eq. (12.4.15) in the form

$$\begin{aligned} & \left(\tilde{s}I - \sum_{\nu=-r}^p B_\nu \kappa^\nu \right) \varphi_1 \\ &= \left(\tilde{s}I - \sum_{\nu=-r}^p B_\nu - \sum_{\nu=-r}^p \nu B_\nu (\kappa - 1) + \mathcal{O}(|\kappa - 1|^2) \right) \varphi_1 \\ &= (\tilde{s}I - A(\kappa - 1) + \mathcal{O}(|\kappa - 1|^2)) \varphi_1 = 0. \end{aligned}$$

Consider this equation for small $|\tilde{s}|$. Because, by assumption, A has distinct eigenvalues there are exactly m distinct eigenvalues and corresponding eigenvectors of the full problem, and they have the form

$$\begin{aligned} \kappa_j(\tilde{s}) &= 1 + \frac{\tilde{s}}{\lambda_j} + \mathcal{O}(|\tilde{s}|^2), \\ \varphi_j &= a_j + \mathcal{O}(|\tilde{s}|). \end{aligned}$$

Here λ_j and a_j are the eigenvalues and corresponding eigenvectors of A .

The dissipativity condition tells us that $\operatorname{Re} \tilde{s} < 0$ for $\kappa = e^{i\xi}$, ξ real, $\xi \neq 0$. For $\kappa = 1$ we have $\tilde{s} = 0$, and we have just demonstrated that there are exactly m eigenvalues $\kappa_j = 1$ for $\tilde{s} = 0$. Therefore, all the other eigenvalues satisfy

$$|\kappa_j(\tilde{s})| \leq 1 - \delta, \quad \delta > 0,$$

for $\operatorname{Re} \tilde{s} \geq 0$. Let $\lambda_j > 0$. If we choose $\tilde{\eta} > 0$, we see that the corresponding eigenvalues satisfy

$$|\kappa_j(i\xi + \tilde{\eta})| \geq 1, \quad \text{for all } \tilde{s} = i\xi + \tilde{\eta}, \quad \tilde{\eta} \geq 0.$$

Therefore,

$$\kappa_j(i\xi + \tilde{\eta}) = \kappa_j(i\xi) + \frac{\tilde{\eta}}{\lambda_j} + \mathcal{O}(|\xi|\tilde{\eta})$$

implies that

$$|\kappa_j(i\tilde{\xi} + \tilde{\eta})| \geq 1 + \frac{\tilde{\eta}}{2\lambda_j},$$

where $\lambda_j > 0$, for c_1 sufficiently small. Correspondingly,

$$|\kappa_j(i\tilde{\xi} + \tilde{\eta})| \leq 1 - \frac{\tilde{\eta}}{2|\lambda_j|},$$

where $\lambda_j < 0$, for c_1 sufficiently small. Order these m eigenvalues such that $|\kappa_j| > 1$ for $j = 1, 2, \dots, r$ and $|\kappa_j| < 1$ for $j = r+1, \dots, m$, $\tilde{\eta} > 0$.

Now we can find a smooth transformation S that transforms Eq. (12.4.12) into Eq. (12.4.14), where now

$$\begin{aligned} M^I &= \begin{bmatrix} \kappa_{r+1} & & 0 & \\ & \ddots & & 0 \\ 0 & & \kappa_m & \\ 0 & & & M_1^I \end{bmatrix}, \\ M^{II} &= \begin{bmatrix} \kappa_1 & & & \\ & \ddots & & 0 \\ & & \kappa_r & \\ 0 & & & M_1^{II} \end{bmatrix}. \end{aligned}$$

Here M_1^I and M_1^{II} satisfy the inequalities (12.4.13). Now we can again apply Lemma 12.4.1 and use the Kreiss condition to obtain

$$\begin{aligned} \|\mathbf{w}^{II}\|_h &\leq \frac{\text{constant}}{\eta} \|\tilde{\mathbf{F}}^{II}\|_h, \quad |w_1^{II}| \leq \frac{\text{constant}}{\sqrt{\eta}} \|\tilde{\mathbf{F}}^{II}\|_h, \\ \|\mathbf{w}^I\|_h &\leq \frac{\text{constant}}{\eta} (\|\tilde{\mathbf{F}}\|_h + |\hat{g}|), \quad |w_1^I| \leq \frac{\text{constant}}{\sqrt{\eta}} (\|\tilde{\mathbf{F}}\|_h + |\hat{g}|). \end{aligned}$$

The last estimates and Lemmas 12.4.2 and 12.4.4 show that the solutions of Eq. (12.4.4) satisfy Eq. (12.4.5b). Thus, the approximation is strongly stable in the generalized sense. This proves Theorem 12.4.4.

REMARK. For the usual difference approximations, the coefficients B_ν of Q are polynomials in A . In that case, we need only assume that the underlying partial differential equations are strongly hyperbolic, because we can reduce the system (12.1.10) to scalar equations.

As we have proved in the previous section, the Kreiss condition is satisfied if, and only if, there are no eigenvalues or generalized eigenvalues for $\operatorname{Re} \tilde{s} \geq 0$.

By Lemma 12.4.3, any generalized eigenvalue $\tilde{s} = i\tilde{\xi}_0$, $\tilde{\xi}_0 \neq 0$, is a genuine eigenvalue for dissipative approximations. Therefore, for $\tilde{s} \neq 0$, we only need to test for genuine eigenvalues in this case.

In the neighborhood of $\tilde{s} = 0$, we can simplify the analysis. For $\operatorname{Re} \tilde{s} > 0$, $|\tilde{s}| \ll 1$, the general solution of Eq. (12.2.14a) with $\|\hat{y}\|_h < \infty$ can be written as

$$\hat{y}_j = Z_j + z_j. \quad (12.4.16)$$

Here Z_j converges, as $\tilde{s} \rightarrow 0$, to the corresponding solution of the continuous problem, that is,

$$Z_j^I = \mathcal{O}(|\tilde{s}|)Z_0^{II}, \quad Z_j^{II} = e^{\tilde{s}(\Lambda^{II})^{-1}j}Z_0^{II} + \mathcal{O}(|\tilde{s}|)Z_0^{II},$$

(see Section 10.1). By Eq. (12.1.22),

$$z_j = \sum_{|\kappa_\nu| \leq 1 - \delta} P_\nu(j)\kappa_\nu^j.$$

represents the part of the solution with $|\kappa_\nu|$ strictly smaller than 1.

The boundary conditions consist of two sets:

1. $m - r$ boundary conditions that formally converge to boundary conditions of the continuous problem. By consistency, they are of the form

$$\hat{y}_0^{II} = R^I \hat{y}_0^I + (E - I)Q_1 \hat{y}_0 + g^{(1)}.$$

2. Extra boundary conditions of the form

$$(E - I)Q_2 \hat{y}_0 = g^{(2)}.$$

Here $Q_j = \sum C_{j\nu} E^\nu$, $j = 1, 2$ are bounded difference operators. Substituting Eq. (12.4.16) into the boundary conditions gives us

$$Z_0^{II} + z_0^{II} = R^I z_0^I + (E - I)Q_1 \hat{z}_0 + \mathcal{O}(|\tilde{s}|)Z_0^{II} + g^{(1)}, \quad (12.4.17a)$$

$$(E - I)Q_2 \hat{z}_0 + \mathcal{O}(|\tilde{s}|)Z_0^{II} = g^{(2)}. \quad (12.4.17b)$$

Now we can prove the following theorem.

Theorem 12.4.5. *The solutions of Eq. (12.2.14) satisfy the Kreiss condition in the neighborhood of $\tilde{s} = 0$ if, and only if, the reduced eigenvalue problem at $\tilde{s} = 0$,*

$$(E - I)Q_2\varphi_0 = 0, \quad \varphi_j = \sum_{|\kappa_\nu| < 1 - \delta} P_\nu(j)\kappa_\nu^j, \quad (12.4.18)$$

only has the trivial solution.

Proof. If Eq. (12.4.18) has a nontrivial solution, then $\tilde{s} = 0$ is a generalized eigenvalue for the full eigenvalue problem (12.1.13) and the Kreiss condition is not satisfied. If, on the other hand, Eq. (12.4.18) only has the trivial solution, then Eq. (12.4.17b) has a unique solution \hat{z}_0 in the neighborhood of $\tilde{s} = 0$, and we can estimate it in terms of $g^{(2)}$ and $\mathcal{O}(|\tilde{s}|)Z_0^{II}$. This solution is introduced in Eq. (12.4.17a), and since the coefficient of Z_0^{II} is of the form $I + \mathcal{O}(|\tilde{s}|)$, there is a unique solution that can be estimated in terms of $g^{(1)}$ and $g^{(2)}$. This proves the theorem.

EXERCISES

12.4.1. Prove Theorem 12.4.2 using the same technique as in Theorem 10.3.3.

12.4.2. Consider the approximation

$$\frac{dv_j}{dt} = D_0 v_j - \delta h^3 (D_+ D_-)^2 v_j, \quad \delta > 0, \quad j = 1, 2, \dots$$

Formulate boundary conditions and prove that the Kreiss condition is satisfied by using Theorem 12.4.5.

12.5. AN EXAMPLE THAT DOES NOT SATISFY THE KREISS CONDITION BUT IS STABLE IN THE GENERALIZED SENSE

All examples we have treated so far have satisfied the Kreiss condition. We now consider an example that does not satisfy the Kreiss condition. For certain parameter values, it is still stable in the generalized sense; for certain other values, it is not. It shows how complicated the situation is when the Kreiss condition is violated.

Consider the equation $\partial u / \partial t + \partial u / \partial x = 0$ and the approximation

$$\frac{dv_j}{dt} + D_0 v_j = 0, \quad j = 1, 2, \dots, \quad (12.5.1a)$$

$$v_j(0) = f_j, \quad (12.5.1b)$$

$$av_0 - v_1 = 0, \quad (12.5.1c)$$

$$\|v\|_h < \infty, \quad (12.5.1d)$$

where a is a complex constant. For $a = 1$, this boundary condition is also sometimes used when the characteristic is entering the domain, as it does in our example. For other values of a , the example is somewhat artificial, but it is used to illustrate typical phenomena arising with generalized eigenvalues.

The connected eigenvalue problem for Eq. (12.5.1) is given by

$$\begin{aligned}\tilde{s}\varphi_j &= -\frac{1}{2}(\varphi_{j+1} - \varphi_{j-1}), \quad j = 1, 2, \dots, \quad \tilde{s} = hs, \\ \varphi_0 &= \frac{1}{a}\varphi_1, \quad \|\varphi\|_{1,\infty} < \infty.\end{aligned}\tag{12.5.2}$$

Thus,

$$\varphi_j = \sigma_1 \kappa_1^j,$$

where $|\kappa_1| < 1$, for $\operatorname{Re} \tilde{s} > 0$, and κ_1 is the appropriate solution of the characteristic equation

$$\kappa^2 + 2\tilde{s}\kappa - 1 = 0,$$

that is,

$$\kappa_1 = -\tilde{s} + \sqrt{1 + \tilde{s}^2}, \quad -\frac{\pi}{2} \leq \arg \sqrt{1 + \tilde{s}^2} \leq \frac{\pi}{2}.$$

Substituting φ into the boundary conditions shows that \tilde{s} is an eigenvalue if

$$\kappa_1 = a. \tag{12.5.3}$$

Straightforward calculations give us the following lemma.

Lemma 12.5.1. *The function*

$$\kappa_1 = -\tilde{s} + \sqrt{1 + \tilde{s}^2}, \quad -\frac{\pi}{2} \leq \arg \sqrt{1 + \tilde{s}^2} \leq \frac{\pi}{2}, \quad \text{for } \operatorname{Re} \tilde{s} \geq 0,$$

maps the right half-plane $\operatorname{Re} \tilde{s} \geq 0$ one-to-one onto the right half disc $\Omega := \{\kappa_1, |\kappa_1| \leq 1, \operatorname{Re} \kappa_1 \geq 0\}$, see Figure 12.5.1. In particular,

1. $|\kappa_1| = 1, |\arg \kappa_1| \leq \frac{\pi}{2}$, corresponds to $\tilde{s} = i\xi$, $-1 \leq \xi \leq 1$ with

$$\kappa_1(0) = 1, \quad \kappa_1(\pm i) = \mp i.$$

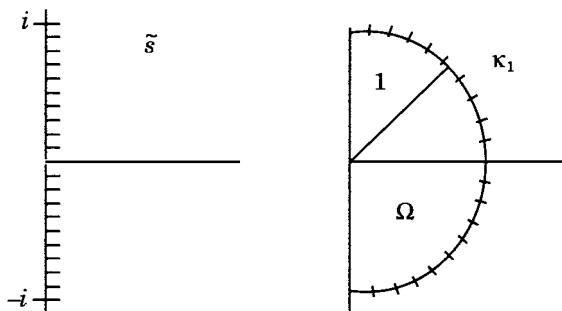


Figure 12.5.1.

2. $\kappa_1 = i\tau$, $-1 < \tau < 1$, $\tau \neq 0$, corresponds to $\tilde{s} = i\xi$, $|\xi| > 1$.
 3. The interior points of Ω correspond to \tilde{s} with $\operatorname{Re} \tilde{s} > 0$.
 4. $\kappa_1 = 0$ corresponds to $\tilde{s} = \infty$.
 5. There is a constant $\delta > 0$ such that $|\kappa_1| \leq 1 - \delta \operatorname{Re} \tilde{s}$ when $\operatorname{Re} \tilde{s}$ is small.
- This gives us the following theorem.

Theorem 12.5.1. If a belongs to the interior of Ω , then there is an eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} > 0$, and the approximation is unstable. If a does not belong to Ω , then there is no eigenvalue or generalized eigenvalue with $\operatorname{Re} \tilde{s} \geq 0$, and the approximation is stable.

If $a = ir$, $-1 < r < 1$, $r \neq 0$, then there is an eigenvalue \tilde{s} with $\operatorname{Re} \tilde{s} = 0$.
If $|a| = 1$, $\operatorname{Re} a \geq 0$, then there is a generalized eigenvalue with $\operatorname{Re} \tilde{s} = 0$.

We now carefully investigate the properties of the solution when a is on the boundary $\partial\Omega$ of Ω ; that is, when there is an eigenvalue or a generalized eigenvalue \tilde{s} on the imaginary axis.

We apply the technique used in Section 12.2 to derive an estimate of v in physical space. Corresponding to Eq. (12.2.6), we first solve the auxiliary problem

$$\begin{aligned} \frac{dw_j}{dt} + D_0 w_j &= 0, \quad j = 1, 2, \dots, \\ w_j(0) &= f_j, \\ w_0 + w_1 &= 0. \end{aligned} \tag{12.5.4}$$

We apply the energy method and obtain

$$\|w(T)\|_h^2 + \int_0^T (|w_0(t)|^2 + |w_1(t)|^2) dt \leq \text{constant} \|f\|_h^2.$$

In particular, as $T \rightarrow \infty$,

$$\int_0^\infty (|w_0(t)|^2 + |w_1(t)|^2) dt \leq \text{constant} \|f\|_h^2. \quad (12.5.5)$$

The difference $y = v - w$ satisfies

$$\begin{aligned} \frac{dy_j}{dt} &= -D_0 y_j, \quad j = 1, 2, \dots, \\ y_j(0) &= 0, \\ y_1(t) - ay_0(t) &= g(t), \end{aligned} \quad (12.5.6)$$

where, by Eq. (12.5.5),

$$\int_0^\infty |g(t)|^2 dt \leq \text{constant} \|f\|_h^2.$$

We solve (12.5.6) by Laplace transformation. The transformed equations are

$$\begin{aligned} s\hat{y}_j &= -D_0\hat{y}_j, \quad \operatorname{Re}s > 0, \\ \hat{y}_1 - a\hat{y}_0 &= \hat{g}, \quad \|\hat{y}\|_h < \infty. \end{aligned}$$

Therefore,

$$\hat{y}_j = \sigma_1 \kappa_1^j, \quad \sigma_1 = \frac{1}{\kappa_1 - a} \hat{g},$$

and, by Parseval's relation, we obtain, for any j ,

$$\int_0^\infty e^{-2\eta t} |y_j(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^\infty |\hat{g}(i\xi + \eta)|^2 \frac{|\kappa_1|^{2j}}{|\kappa_1 - a|^2} d\xi. \quad (12.5.7)$$

If there is no eigenvalue or generalized eigenvalue, then

$$|(\kappa_1 - a)^{-1}| \leq \text{constant},$$

for $\operatorname{Re}\tilde{s} \geq 0$. We can choose $\eta = 0$ in Eq. (12.5.7) and obtain

$$\int_0^\infty |y_j(t)|^2 dt \leq \text{constant} \int_{-\infty}^\infty |\hat{g}(i\xi)|^2 d\xi \leq \text{constant} \int_0^\infty |g(t)|^2 dt.$$

We used the last estimate earlier to prove stability.

If a belongs to the boundary of Ω , then there is an eigenvalue or generalized eigenvalue $\tilde{s}_0 = i\tilde{\xi}_0$ such that

$$\lim_{\tilde{s} \rightarrow \tilde{s}_0} (\kappa_1 - a)^{-1} = \infty,$$

and we have to investigate the right-hand side of Eq. (12.5.7) more carefully.

We have to distinguish between three different cases.

CASE 1. $a = i\tau$, $-1 < \tau < 1$. In this case, there is an eigenvalue $\tilde{s}_0 = i\tilde{\xi}_0$, $|\tilde{\xi}_0| > 1$, and, in the neighborhood of \tilde{s}_0 , we have

$$\begin{aligned} \kappa_1(\tilde{s}) &= \kappa_1(\tilde{s}_0) + (\tilde{s} - \tilde{s}_0)\partial\kappa_1(\tilde{s}_0)/\partial\tilde{s} + \mathcal{O}(|\tilde{s} - \tilde{s}_0|^2) \\ &= a + c(\tilde{s} - \tilde{s}_0) + \mathcal{O}(|\tilde{s} - \tilde{s}_0|^2), \quad c = -1 + i\tilde{\xi}_0/\sqrt{1 - \tilde{\xi}_0^2}. \end{aligned}$$

Therefore, we obtain, for the variable $s = \tilde{s}/h$,

$$\kappa_1 - a = hc(s - s_0) + \mathcal{O}(h^2|s - s_0|^2), \quad s_0 = i\tilde{\xi}_0/h. \quad (12.5.8)$$

Thus, for sufficiently small $h|s - s_0|$,

$$|\kappa_1 - a| \geq \frac{1}{2}|hc||s - s_0|. \quad (12.5.9)$$

Let $0 < \delta < \frac{1}{2}$ be a constant. We use Eq. (12.5.7) to estimate y_j in a finite interval $0 \leq t \leq T$.

$$\begin{aligned} &\int_0^T e^{-2\eta t} |y_j(t)|^2 dt \\ &\leq \int_0^\infty e^{-2\eta t} |y_j(t)|^2 dt, \\ &= \frac{1}{2\pi} \int_{-\infty}^\infty |\hat{g}(i\xi + \eta)|^2 H_j(i\xi + \eta) d\xi = I + II, \quad (12.5.10) \end{aligned}$$

where

$$\begin{aligned}
H_j(i\xi + \eta) &= |\kappa_1(i\xi + \eta)|^{2j} / |\kappa_1(i\xi + \eta) - a|^2, \\
I &= \frac{1}{2\pi} \int_{|\xi - \xi_0| \geq \delta/h} |\hat{g}|^2 H_j d\xi \leq \frac{K_1}{2\pi} \int_{-\infty}^{\infty} |\hat{g}|^2 d\xi \\
&= K_1 \int_0^{\infty} e^{-2\eta t} |g(t)|^2 dt, \\
II &= \frac{1}{2\pi} \int_{|\xi - \xi_0| \leq \delta/h} |\hat{g}|^2 H_j d\xi \\
&\leq \frac{1}{2\pi} \left(\max_{|\xi - \xi_0| \leq \delta/h} H_j \right) \int_{|\xi - \xi_0| \leq \delta/h} |\hat{g}|^2 d\xi.
\end{aligned}$$

Here K_1 is a constant that, by Eq. (12.5.8), depends only on δ . By Eqs. (12.5.8) and (12.5.9), we have, for sufficiently small δ ,

$$H_j \leq \left(\frac{2}{\eta |c|h} \right)^2 \max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1(i\xi + \eta)|^{2j}.$$

Therefore, Eq. (12.5.10) gives us, for $\eta = 1/T$,

$$\begin{aligned}
e^{-2} \int_0^T |y_j|^2 dt &\leq \frac{K_1}{2\pi} \int_{-\infty}^{\infty} |\hat{g}|^2 d\xi \\
&+ \frac{4T^2}{2\pi (|c|h)^2} \max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1(i\xi + \eta)|^{2j} \int_{|\xi - \xi_0| \leq \delta/h} |\hat{g}|^2 d\xi, \\
&\leq \text{constant} \left(1 + \frac{4T^2}{(|c|h)^2} \max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1|^{2j} \right) \|f\|_h^2. \quad (12.5.11)
\end{aligned}$$

By taking the square root of both sides, we find that the stability constant is proportional to h^{-1} . However, because $|a| < 1$, we have, for sufficiently small δ ,

$$|\kappa_1(i\xi + \eta)| \leq 1 - \sigma, \quad \sigma = \text{constant} > 0.$$

Thus, the deterioration of the stability constant is only present in a narrow boundary layer. For

$$(1 - \sigma)^{2j} \leq h^2,$$

that is,

$$j \geq |\log h|/|\log(1 - \sigma)|,$$

the bad effect disappears.

The above estimate is sharp for general functions $g \in L_2(t)$. However, if $g(t)$ has some smoothness, that is, if

$$\hat{g}(s) \sim (1/s)^p,$$

where $p \geq 1$ for large $|s|$, then we can do better. The major contribution comes from the term

$$R := \frac{1}{h^2} \int_{|\xi - \xi_0| \leq \delta/h} |\hat{g}(i\xi)|^2 d\xi \sim \frac{1}{h^2} \int_{|\xi - \xi_0| \leq \delta/h} \frac{d\xi}{|i\xi + \eta|^{2p}}.$$

But $\xi_0 = \tilde{\xi}_0/h$, where $\tilde{\xi}_0$ is a fixed number. Hence, for $\delta < |\tilde{\xi}_0|$ we get

$$R \sim \frac{1}{h^2} \int_{(\tilde{\xi}_0 - \delta)/h}^{(\tilde{\xi}_0 + \delta)/h} \frac{d\xi}{|i\xi + \eta|^{2p}} \leq \frac{\text{constant}}{h^2} h^{2p-1} \leq \text{constant} h^{2p-3}.$$

Thus, if $p \geq \frac{3}{2}$, then the “bad term” in Eq. (12.5.11) is bounded. Therefore, this example indicates that one can accept genuine eigenvalues \tilde{s} with $\operatorname{Re} \tilde{s} = 0$.

CASE 2. $|a| = 1$, $\operatorname{Re} a > 0$. In this case, there is a generalized eigenvalue $\tilde{s}_0 = i\xi_0$, where $|\xi_0| < 1$. We can proceed in the same way as for the previous case and obtain an estimate of the same type as in Eq. (12.5.11) (see Exercise 12.5.1). However, in this case, $|\kappa_1(i\xi_0)| = 1$ and, therefore,

$$\max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1(i\xi + \eta)|^{2j} \sim \left(1 - \frac{h\eta}{2}\right)^{2j} \sim e^{-x_j/T}$$

for $\eta = 1/T$. (Recall that $|\kappa_1| < 1$ for $\eta > 0$ by definition.) Thus the “bad behavior” is not restricted to a boundary layer but spreads to the whole domain $0 \leq x < \infty$. Still, for smooth g , the effect is unimportant and one might be tempted to accept the presence of generalized eigenvalues. However, the instability can be amplified if we consider, instead of the quarter-space problem, the problem in the strip $0 \leq x \leq 1$, $t \geq 0$. Now we also must describe boundary conditions at $x = 1$. Let $hN = 1$. We consider Eq. (12.5.1) with $\|v\|_h < \infty$ replaced by the boundary condition

$$v_N = 0. \quad (12.5.12)$$

This is not a very natural condition at an outflow boundary, but the left quarter-space problem is stable, and the example serves as an illustration of the principles.

We now estimate the eigenvalues of the operator corresponding to the half strip problem. The general solution of the eigenvalue problem is of the form

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j, \quad \kappa_1 \neq \kappa_2$$

where κ_1 , and κ_2 are solutions of the characteristic equation

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

Thus, $\kappa_1 \kappa_2 = -1$, that is,

$$\kappa_2 = -\kappa_1^{-1}.$$

Substituting φ into the boundary conditions gives us

$$\begin{aligned} (\kappa_1 - a)\sigma_1 + (\kappa_2 - a)\sigma_2 &= 0, \\ \kappa_1^N \sigma_1 + \kappa_2^N \sigma_2 &= 0. \end{aligned}$$

Therefore, \tilde{s} is an eigenvalue if

$$(\kappa_1 - a)\kappa_2^N = (\kappa_2 - a)\kappa_1^N,$$

that is,

$$(-1)^{N-1} \kappa_1^{2N-1} = \frac{\kappa_1 - a}{1 + a\kappa_1}. \quad (12.5.13)$$

We want to estimate the corresponding eigenvalues \tilde{s} . To do that, we will show that there is an eigenvalue near \tilde{s}_0 , that is, that Eq. (12.5.13) has a solution $\tilde{s} \approx \tilde{s}_0$, $\operatorname{Re} \tilde{s} > 0$. In a neighborhood of $\tilde{s}_0 = i\xi_0$, $|\xi_0| < 1$, we have, with $\kappa = \kappa_1 = -\tilde{s} + \sqrt{1 + \tilde{s}^2}$,

$$\begin{aligned}
\kappa(\tilde{s}) &= -i\tilde{\xi} - \tilde{\eta} + \sqrt{1 - \tilde{\xi}^2 + 2i\tilde{\xi}\tilde{\eta} + \tilde{\eta}^2} \\
&= -i\tilde{\xi} - \tilde{\eta} + \sqrt{(1 - \tilde{\xi}^2)(1 + i2\tilde{\xi}\tilde{\eta}/(1 - \tilde{\xi}^2) + \mathcal{O}(\tilde{\eta}^2))} \\
&= -i\tilde{\xi} - \tilde{\eta} + \sqrt{1 - \tilde{\xi}^2} (1 + i\tilde{\xi}\tilde{\eta}/(1 - \tilde{\xi}^2) + \mathcal{O}(\tilde{\eta}^2)) \\
&= -i\tilde{\xi} \left(1 - \frac{\tilde{\eta}}{\sqrt{1 - \tilde{\xi}^2}} \right) + \sqrt{1 - \tilde{\xi}^2} - \tilde{\eta} + \mathcal{O}(\tilde{\eta}^2),
\end{aligned} \tag{12.5.14}$$

that is,

$$|\kappa(\tilde{s})|^2 = 1 - \frac{2\tilde{\eta}}{\sqrt{1 - \tilde{\xi}^2}} + \mathcal{O}(\tilde{\eta}^2).$$

Therefore, we can write

$$\kappa(\tilde{s}) = e^{i[\tau_0 + c_1(\tilde{\xi} - \tilde{\xi}_0)] - c_2\tilde{\eta}},$$

where

$$\begin{aligned}
e^{i\tau_0} &= \kappa(\tilde{s}_0) = a, \\
c_1 &= c_{10} + \mathcal{O}(|\tilde{\xi} - \tilde{\xi}_0|) + \mathcal{O}(\tilde{\eta}), \quad c_{10} \neq 0 \text{ real}, \\
c_2 &= c_{20} + \mathcal{O}(|\tilde{\xi} - \tilde{\xi}_0|) + \mathcal{O}(\tilde{\eta}), \quad c_{20} > 0 \text{ real}.
\end{aligned}$$

To first approximation, Eq. (12.5.13) becomes

$$e^{i\pi(N-1)} e^{i[\tau_0 + c_{10}(\tilde{\xi} - \tilde{\xi}_0)] - c_{20}\tilde{\eta}(2N-1)} = D(ic_{10}(\tilde{\xi} - \tilde{\xi}_0) - c_{20}\tilde{\eta}), \tag{12.5.15}$$

where $D = a/(1+a^2)$. Observe that, by assumption, $a^2 \neq -1$. Equation (12.5.15) has a solution if, for some integer m ,

$$\begin{aligned}
e^{-c_{20}\tilde{\eta}(2N-1)} &= |D| |ic_{10}(\tilde{\xi} - \tilde{\xi}_0) - c_{20}\tilde{\eta}|, \\
\pi(N-1) - 2\pi m + (\tau_0 + c_{10}(\tilde{\xi} - \tilde{\xi}_0))(2N-1) \\
&= \arg D + \arg(ic_{10}(\tilde{\xi} - \tilde{\xi}_0) - c_{20}\tilde{\eta}), \quad -\pi \leq \arg(\cdot) < \pi.
\end{aligned} \tag{12.5.16}$$

Choose m such that $\sigma := \pi(N-1) + (2N+1)\tau_0 - 2\pi m$ satisfies

$$|\sigma| \leq \pi.$$

One can show that the equation

$$e^{-c_{20}\tilde{\eta}(2N-1)} = c_{20}\tilde{\eta}$$

has a solution

$$c_{20}\tilde{\eta} = \frac{\log(2N-1)}{2N-1} + \mathcal{O}\left(\frac{\log\log(2N-1)}{2N-1}\right),$$

(Exercise 12.5.2). Furthermore, observing that

$$c_{10}(\tilde{\xi} - \tilde{\xi}_0) = \mathcal{O}\left(\frac{1}{2N-1}\right),$$

it follows that Eq. (12.5.16) and, therefore, also Eq. (12.5.13) have a solution

$$c_{20}\tilde{\eta} = \frac{1}{2N-1} \log(2N-1) + \mathcal{O}\left(\frac{\log\log(2N-1)}{2N-1}\right),$$

$$c_{10}(\tilde{\xi} - \tilde{\xi}_0) = \frac{1}{2N-1} (\arg D - \sigma) + \mathcal{O}\left(\frac{1}{(2N-1)\log(2N-1)}\right).$$

This eigensolution leads to a solution of the strip problem

$$v_j(t) = e^{st}\varphi_j = e^{st}(\sigma_1\kappa_1^j + \sigma_2\kappa_2^j), \quad s = (i\tilde{\xi} + \tilde{\eta})/h, \quad (12.5.17)$$

where, to first approximation,

$$|e^{st}| = e^{(\tilde{\eta}/h)t} \approx e^{(t/2c_{20})\log(2N-1)} = (2N-1)^{\alpha t}, \quad \alpha = (2c_{20})^{-1}.$$

The last equation shows that the instability is of the order $(1/h)^{\alpha t}$, which grows with time. If $\alpha t = 10$ and $N = 100$, then the solution has increased by a factor 10^{20} , which is not acceptable.

Geometrically, we can think of this phenomenon in the following way. The solution of the strip problem consists of waves that are reflected back and forth between the boundaries. There are waves that are amplified by a factor h^{-1} every time they are reflected at the boundary $x = 0$. As time increases, more and more reflections can take place. This argument explains why, in the quarter-

space case, the stability constant is only increased by h^{-1} . The above waves are only reflected once at the boundary $x = 0$ and, then, escape to infinity.

The bad behavior obtained here does not happen in the first case, when $\tilde{s}_0 = i\tilde{\xi}_0$, $|\tilde{\xi}_0| > 1$. The deterioration of the stability constant is only felt in a boundary layer; that is, the amplitude of the reflected wave decreases exponentially away from the boundary $x = 0$. In this case, we obtain from Eq. (12.5.13)

$$|\kappa_1|^{2N-1} = \frac{1}{|1 + a\kappa_1|} |\kappa_1 - a|.$$

In the neighborhood of \tilde{s}_0 , we have $|\kappa_1| \leq 1 - \sigma$, $\sigma > 0$. Thus, $|\kappa_1 - a|$ and, hence, also $|\tilde{s} - \tilde{s}_0|$, must be exponentially small. This shows that the order of the instability remains essentially $\mathcal{O}(1/h)$.

The above results seem to indicate that we can accept eigenvalues \tilde{s} with $\operatorname{Re}\tilde{s} = 0$ but not generalized eigenvalues. However, the picture is even more complicated. This will become clear during the discussion of the third case.

CASE 3. $a = \pm i$. Now $\tilde{s}_0 = \mp i$, $\kappa_1 = \pm i$, and we have a generalized eigenvalue. We have to calculate

$$\kappa_1(i\tilde{\xi} + \tilde{\eta}) = a, \quad |\kappa_1(i\tilde{\xi} + \tilde{\eta})|^{2j}$$

in the neighborhood of \tilde{s}_0 . The root $\kappa(\tilde{s}_0)$ is now a double root of the characteristic polynomial and we have

$$\begin{aligned} \kappa(\tilde{s}) &= -\tilde{s} + \sqrt{1 + \tilde{s}^2} = -\tilde{s}_0 - (\tilde{s} - \tilde{s}_0) + \sqrt{2\tilde{s}_0(\tilde{s} - \tilde{s}_0) + (\tilde{s} - \tilde{s}_0)^2}, \\ &= -\tilde{s}_0(1 + \tau - \sqrt{2\tau + \tau^2}), \quad \tau = \frac{\tilde{s} - \tilde{s}_0}{\tilde{s}_0}. \end{aligned}$$

Therefore, we obtain, for all sufficiently small $|\tau|$,

$$|\kappa(\tilde{s}) - a| \geq \text{constant } |\tilde{s} - \tilde{s}_0|^{\frac{1}{2}}. \quad (12.5.18)$$

This inequality tells us that $\tilde{s} \rightarrow \tilde{s}_0$ at a faster rate than $\kappa(\tilde{s}) \rightarrow a$. This implies improved stability properties. Substituting Eq. (12.5.18) into the estimates (12.5.10) and (12.5.11) gives us, for $\eta = 1/T$,

$$\max_{|\xi - \xi_0| \leq \delta/h} H_j(i\xi + \eta) \leq \frac{\text{constant}}{h\eta} \max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1(i\xi + \eta)|^{2j} \leq \frac{\text{constant}}{h\eta};$$

that is,

$$e^{-2} \int_0^T |y_j|^2 dt \leq \text{constant} \left(1 + \frac{T}{h} \right) \|f\|_h^2.$$

Thus, the instability is of the order $\sqrt{T/h}$, which is weaker than in the previous case.

A simple but tedious computation shows that there is a constant $c_0 > 0$ such that, for all sufficiently small τ ,

$$|\kappa(\tilde{s})| \leq 1 - \frac{c_0 \tilde{\eta}}{\sqrt{|\xi - \xi_0| + \tilde{\eta}}} = 1 - \frac{c_0 h^{1/2} \eta}{\sqrt{|\xi - \xi_0| + \eta}}.$$

Now, since $|\tilde{s} - \tilde{s}_0| \geq (h/\sqrt{2})(|\xi - \xi_0| + \eta)$, we get from Eq. (12.5.18)

$$\begin{aligned} \max_{|\xi - \xi_0| \leq \delta/h} H_j &\leq \text{constant} \max_{|\xi - \xi_0| \leq \delta/h} \frac{(1 - (c_0 h^{1/2} \eta)/\sqrt{|\xi - \xi_0| + \eta})^{2j}}{h(|\xi - \xi_0| + \eta)} \\ &\leq \text{constant} \frac{\exp(-(2c_0 x_j \eta)/(h^{1/2} \sqrt{|\xi - \xi_0| + \eta}))}{h(|\xi - \xi_0| + \eta)} \\ &\leq \text{constant} \frac{1}{c_0^2 x_j^2 \eta^2}. \end{aligned}$$

Therefore, in the last case, instead of Eq. (12.5.11), we obtain, with $T = 1/\eta$,

$$e^{-2} \int_0^T |y_j|^2 dt \leq \text{constant} \left(1 + \frac{T^2}{c_0 x_j^2} \right) \|f\|_h^2, \quad j = 1, 2, \dots$$

Again, the deterioration of the stability constant is restricted to a boundary layer. This layer is wider than in case 1, but one can still show that no further amplification occurs for the strip problem.

Our example shows that eigenvalues \tilde{s} on the imaginary axis can be accepted, but that generalized eigenvalues can lead to bad nonlocal behavior if the corresponding root κ with $|\kappa| = 1$ is simple. If κ is a multiple root, then the singularity becomes weaker as shown by Eq. (12.5.18), and this case can be accepted.

We now show that our generalized stability concept can distinguish between the different cases. First consider case 1, where there is an eigenvalue $\tilde{s}_0 = i\xi_0$, where $|\xi_0| > 1$. For $|\tilde{s} - \tilde{s}_0| \geq \delta > 0$, the desired estimate poses no difficulty and

can be obtained by our previous methods. Therefore, we need only consider a neighborhood of \tilde{s}_0 .

We let $f = 0$, $F \neq 0$ and write the transformed equations

$$\begin{aligned}\tilde{s}\hat{v}_j &= -\frac{1}{2}(\hat{v}_{j+1} - \hat{v}_{j-1}) + h\hat{F}_j, \quad j = 1, 2, \dots, \\ a\hat{v}_0 &= \hat{v}_1,\end{aligned}$$

in the form

$$\begin{aligned}\mathbf{v}_{j+1} &= C\mathbf{v}_j + h\mathbf{F}_j, \quad j = 1, 2, \dots, \\ B\mathbf{v}_1 &= 0,\end{aligned}\tag{12.5.19}$$

where

$$\mathbf{v}_j = \begin{bmatrix} \hat{v}_j \\ \hat{v}_{j-1} \end{bmatrix}, \quad C = \begin{bmatrix} -2\tilde{s} & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{F}_j = \begin{bmatrix} \hat{F}_j \\ 0 \end{bmatrix}, \quad B = [a \quad -1].$$

The eigenvalues of the matrix C are the solutions κ_1 and κ_2 of the characteristic equation. Now assume that $|\tilde{\xi}_0| > 1$. Then we know that $|\kappa_1| < 1$ and $|\kappa_2| > 1$ in a neighborhood of \tilde{s}_0 . Therefore, there is an analytic transformation $T = T_0 + (\tilde{s} - \tilde{s}_0)T_1(\tilde{s})$ such that

$$TCT^{-1} = \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix}.$$

Introducing new variables

$$\mathbf{w}_j = \begin{bmatrix} w_j^{(1)} \\ w_j^{(2)} \end{bmatrix} = T\mathbf{v}_j$$

gives us

$$\begin{aligned}\mathbf{w}_{j+1} &= \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix} \mathbf{w}_j + h\tilde{\mathbf{F}}_j, \quad j = 1, 2, \dots, \\ \tilde{a}(\tilde{s})\tilde{w}_1^{(1)} + \tilde{b}(\tilde{s})w_1^{(2)} &= 0.\end{aligned}\tag{12.5.20}$$

Because \tilde{s}_0 is an eigenvalue, $\tilde{a}(\tilde{s}_0) = 0$, and a simple calculation shows that

$$\tilde{a}(\tilde{s}) = (\tilde{s} - \tilde{s}_0)a_1 + \mathcal{O}(|\tilde{s} - \tilde{s}_0|^2), \quad a_1 \neq 0.\tag{12.5.21}$$

Furthermore, $\tilde{b}(\tilde{s})$ is bounded. In case 1, $|\kappa_1| = \tau$, $|\kappa_2| = 1/\tau$, $-1 < \tau < 1$. Thus by Lemma 12.4.1, we obtain the estimates

$$\begin{aligned}\|w^{(2)}\|_h &\leq \text{constant } h \|\tilde{F}^{(2)}\|_h, \quad |w_1^{(2)}| \leq \text{constant } h^{1/2} \|\tilde{F}^{(2)}\|_h, \\ \|w^{(1)}\|_h &\leq \text{constant} (h \|\tilde{F}^{(1)}\|_h + h^{\frac{1}{2}} |w_1^{(1)}|) \\ &\leq \text{constant} \left(h \|\tilde{F}^{(1)}\|_h + \frac{h^{\frac{1}{2}}}{|\tilde{s} - \tilde{s}_0|} |w_1^{(2)}| \right) \\ &\leq \text{constant} \left(h \|\tilde{F}^{(1)}\|_h + \frac{h}{|\tilde{s} - \tilde{s}_0|} \|\tilde{F}^{(2)}\|_h \right) \\ &\leq \text{constant} \left(h + \frac{1}{\eta} \right) \|\tilde{F}\|_h.\end{aligned}$$

Thus, the problem is stable in the generalized sense in case 1, which was shown above to be well-behaved.

In case 2, an estimate leading to generalized stability cannot be derived. The reason for this is that $|\kappa_1| = 1 + \mathcal{O}(|\tilde{s} - \tilde{s}_0|)$, $|\kappa_2| = 1 + \mathcal{O}(|\tilde{s} - \tilde{s}_0|)$; that is, κ , approaches the unit circle too fast as $\tilde{s} \rightarrow \tilde{s}_0$.

In case 3, the situation is more favorable. The root $\kappa_1 = \kappa_2$ is now a double root at $\tilde{s} = \tilde{s}_0$, which implies $|\kappa_1| = 1 + \mathcal{O}(|\tilde{s} - \tilde{s}_0|^{1/2})$. This larger distance from the unit circle as $\tilde{s} \rightarrow \tilde{s}_0$ is enough to secure the necessary estimates for stability in the generalized sense.

EXERCISES

- 12.5.1.** Consider the problem (12.5.1) with $a = 1$. Find the generalized eigenvalue $\tilde{s}_0 = i\xi_0$, and prove that

$$\max_{|\xi - \xi_0| \leq \delta/h} |\kappa_1(i\xi + \eta)|^{2j} \sim e^{-x_j/T}$$

thereby proving that the instability of order $1/h$ is confined to a boundary layer (case 2 in our discussion above).

- 12.5.2.** Consider the equation

$$e^{-xM} = x$$

for large M . Prove that it has a solution

$$x = \frac{\log M}{M} + \mathcal{O}\left(\frac{\log \log M}{M}\right).$$

[This result is used in the derivation of solutions to Eq. (12.5.13).]

12.5.3. Consider the approximation

$$\begin{aligned}\frac{dv_j}{dt} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_0 v_j, \quad j = 1, 2, \dots, \\ \frac{dv_0^{(2)}}{dt} &= D_+ v_0, \\ v_0^{(1)} &= 0, \\ v_j(0) &= f_j.\end{aligned}$$

Prove that it has generalized eigenvalues. In Exercise 11.1.1, it was shown that the approximation satisfies an energy estimate. At first sight, this seems to be a contradiction, because we know that generalized eigenvalues lead to growth of order $h^{-1/2}$. Explain this paradox.

12.6. PARABOLIC EQUATIONS

We consider parabolic systems

$$\begin{aligned}u_t &= Au_{xx} + Bu_x + Cu + F =: Pu + F, \quad 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x),\end{aligned}\tag{12.6.1a}$$

in the quarter space $x \geq 0, t \geq 0$. Here u , F , and f are vector functions with m components and A , B , and C are constant matrices. We also assume that $A = A^*$ is Hermitian and positive definite. At $x = 0$ and $t \geq 0$, we describe m linearly independent boundary conditions

$$\begin{aligned}R_{11}u_x(0, t) + R_{10}u(0, t) &= g^{(1)}(t), \\ R_{00}u(0, t) &= g^{(0)}(t).\end{aligned}\tag{12.6.1b}$$

Here R_{11} and R_{10} are $r \times m$ matrices with rank $(R_{11}) = r$. Correspondingly, R_{00} is an $(m - r) \times m$ matrix with rank $(R_{00}) = m - r$. We approximate Eq. (12.6.1) by a difference approximation of type (12.1.10). The only difference is that, instead of Eq. (12.1.11),

$$Q = \frac{1}{h^2} \sum_{\nu=-r}^p B_\nu E^\nu \quad (12.6.2)$$

is of order h^{-2} .

We modify the stability Definition 12.4.1 by replacing the estimate (12.4.1a) by

$$\int_0^\infty e^{-2\eta t} (\|v(t)\|_h^2 + \|D_+ v(t)\|_h^2) dt \leq K(\eta) \int_0^\infty e^{-2\eta t} \|F(t)\|_h^2 dt.$$

Equation (12.4.1b) is modified correspondingly.

In analogy with the continuous case, we note that the concept “strongly stable in the generalized sense” does not play the same role as for hyperbolic equations. Only if all boundary conditions are derivative conditions, that is, if $\text{rank}(R_{11}) = m$, can we find approximations that are strongly stable in the generalized sense.

As we know from the continuous case, well-posedness is independent of the lower order terms; that is, we can assume that $B \equiv C \equiv 0$, and $R_{10} \equiv 0$. Correspondingly, as in the hyperbolic case, we can assume that the coefficients of the difference operator Q and the boundary conditions do not depend on h . By assumption, $A = A^*$, and, therefore, we can make a change of variables such that A becomes diagonal. Therefore, we need only discuss scalar differential equations. The different components of the solution can be coupled through the boundary conditions. However, to present the basic technique in a simple way, we only discuss the uncoupled case. No new essential difficulties occur with coupling through the boundary conditions. We consider, therefore, only the model problem

$$\begin{aligned} u_t &= u_{xx} + F, & 0 \leq x < \infty, \quad t \geq 0, \\ u(x, 0) &= f(x), \end{aligned} \quad (12.6.3a)$$

with boundary conditions

$$u_x(0, t) = g^{(1)}(t) \quad (12.6.3b)$$

or

$$u(0, t) = g^{(0)}(t), \quad (12.6.3c)$$

and approximate it by the fourth-order approximation

$$\frac{dv_j}{dt} = D_+ D_- \left(I - \frac{h^2}{12} D_+ D_- \right) v_j + F_j, \quad (12.6.4a)$$

$$=: Qv_j + F_j, \quad j = 1, 2, \dots, \quad (12.6.4a)$$

$$v_j(0) = f_j, \quad j = 1, 2, \dots \quad (12.6.4b)$$

In this case, we need two boundary conditions. The first boundary condition formally converges to the boundary condition for the continuous problem. In the case of the Neumann boundary condition (12.6.3b), it is of the form

$$\sum_{j=-1}^{p-1} a_j D_+ v_j = \tilde{g}^{(1)}, \quad \sum_{j=-1}^{p-1} a_j = 1, \quad (12.6.4c)$$

For Dirichlet conditions (12.6.3c), we have, instead,

$$\sum_{j=-1}^p a_j v_j = \tilde{g}^{(0)}, \quad \sum_{j=-1}^p a_j = 1. \quad (12.6.4d)$$

The conditions on a_j are imposed by consistency. The second boundary condition can be expressed in terms of second- or higher order differences; that is,

$$\sum_{j=-1}^{p-1} b_j D_+^2 v_j = \tilde{g}^{(2)}. \quad (12.6.4e)$$

Typically, the extra boundary condition is either an extrapolation condition of the form

$$D_+^q v_{-1} = 0, \quad q \geq 2,$$

or is obtained by differentiating the boundary conditions (12.6.3b) or (12.6.3c) with respect to t and replacing the t derivatives by space derivatives using the differential equations. This second procedure was demonstrated in Section 11.4. For example, Eq. (12.6.3c) implies

$$u_{xx}(0, t) = u_t(0, t) = g_t^{(0)}(t),$$

and we use

$$D_+ D_- v_0 = g_i^{(0)}$$

as an extra boundary condition.

In every case, we assume that we can solve the boundary conditions for v_0 and v_{-1} and write them in the form

$$\begin{aligned} v_0 - \sum_{j=1}^p l_{0j} v_j &=: L_0 v_0 = g_0, \\ v_{-1} - \sum_{j=1}^p l_{1j} v_j &=: L_1 v_0 = g_{-1}. \end{aligned} \quad (12.6.4f)$$

We assume that the coefficients l_{ij} are constants and do not depend on h .

Here we are only interested in stability in the generalized sense; we must estimate the solutions of the resolvent equation

$$\begin{aligned} s\hat{v}_j &= Q\hat{v}_j + \hat{F}_j, \quad j = 1, 2, \dots, \quad \operatorname{Re} s > 0, \\ L_i \hat{v} &= 0, \quad i = 0, 1, \\ \|\hat{v}(\cdot, s)\|_h &< \infty. \end{aligned} \quad (12.6.5)$$

We now estimate the solution \hat{v}_j . In the parabolic case it is natural to define $\tilde{s} = sh^2$, and divide the analysis into three parts corresponding to the three parts of the complex \tilde{s} plane:

- I. $|\tilde{s}| \geq c_0$, $\operatorname{Re} \tilde{s} \geq 0$, where the constant c_0 is large.
- II. $0 < c_1 \leq |\tilde{s}| \leq c_0$, $\operatorname{Re} \tilde{s} \geq 0$, where c_1 is a small constant.
- III. $|\tilde{s}| \leq c_1$, $\operatorname{Re} \tilde{s} \geq 0$.

I. $|\tilde{s}| \geq c_0$, $\operatorname{Re} \tilde{s} \geq 0$.

In this case, the estimates are obtained easily. We have

$$\begin{aligned} |s|^2 \|\hat{v}\|_h^2 &\leq \frac{\text{constant}}{h^4} \left(\|\hat{v}\|_h^2 + \sum_{j=-1}^0 |v_j|^2 h \right) + 2\|\hat{F}\|_h^2, \\ &\leq \frac{\text{constant}}{h^4} \|\hat{v}\|_h^2 + 2\|\hat{F}\|_h^2. \end{aligned}$$

Thus, for $\text{constant}/(|s|^2 h^4) \leq \frac{1}{2}$, we obtain

$$\|\hat{v}\|_h^2 \leq \frac{4}{|s|^2} \|\hat{F}\|_h^2. \quad (12.6.6a)$$

Also,

$$\|D_+ \hat{v}\|_h^2 \leq \frac{4}{h^2} \|\hat{v}\|_h^2 \leq \frac{16}{h^2 |s|^2} \|\hat{F}\|_h^2 \leq \frac{16}{|s|} \|\hat{F}\|_h^2. \quad (12.6.6b)$$

Therefore, the desired estimate holds.

For cases II and III we need two lemmas.

Lemma 12.6.1. *For $\operatorname{Re} s > 0$, the characteristic equation*

$$\tilde{s} = \frac{(\kappa - 1)^2}{\kappa} - \frac{1}{12} \frac{(\kappa - 1)^4}{\kappa^2}, \quad \tilde{s} = h^2 s,$$

has no roots with $|\kappa| = 1$, and it has exactly two roots κ_1 and κ_2 , counted according to their multiplicity, with $|\kappa_j| < 1$.

For every $c_1 > 0$, there is a $\tau > 0$ such that

$$\begin{aligned} |\kappa_j| &\leq 1 - \tau, \quad j = 1, 2, \\ |\kappa_j| &\geq 1 + \tau, \quad j = 3, 4, \quad \text{for } |\tilde{s}| \geq c_1, \quad \operatorname{Re} \tilde{s} \geq 0. \end{aligned} \quad (12.6.7)$$

In a neighborhood of $\tilde{s} = 0$, we have

$$\begin{aligned} \kappa_1 &= 1 - \tilde{s}^{1/2} + \mathcal{O}(|\tilde{s}|), \quad \kappa_2 = \kappa_2(0) + \mathcal{O}(|\tilde{s}|), \\ \kappa_2(0) &= 7 - \sqrt{48} \approx 1/14, \quad \kappa_3 = 1 + \tilde{s}^{1/2} + \mathcal{O}(|\tilde{s}|), \\ \kappa_4 &= \kappa_4(0) + \mathcal{O}(|\tilde{s}|), \quad \kappa_4(0) = 7 + \sqrt{48} \approx 14. \end{aligned} \quad (12.6.8)$$

$\kappa_1 = \kappa_2$ is a double root if

$$\kappa_1 = 4 - \sqrt{15} \approx 0.13, \quad s \approx 3.00.$$

Proof. The proof follows the same lines as for hyperbolic equations. Assume that there is a root $\kappa = e^{i\xi}$, where ξ real. Then

$$\tilde{s} = e^{-i\xi}(e^{i\xi} - 1)^2 - \frac{1}{12} e^{-2i\xi}(e^{i\xi} - 1)^4 = -4 \sin^2 \frac{\xi}{2} - \frac{4}{3} \sin^4 \frac{\xi}{2} \quad (12.6.9)$$

implies $\tilde{s} = \operatorname{Re} \tilde{s} \leq 0$. Thus, there are no roots κ with $|\kappa| = 1$ for $\operatorname{Re} \tilde{s} > 0$. Furthermore, the solutions κ with $|\kappa| < 1$ satisfy

$$\lim_{\tilde{s} \rightarrow \infty} \kappa = 0.$$

Thus, to first approximation, the characteristic equation reduces to

$$12\kappa^2 = -1/\tilde{s},$$

that is, there are exactly two roots with $|\kappa| < 1$.

Equation (12.6.9) shows that the combination $|\kappa| = 1$, $\operatorname{Re} \tilde{s} = 0$ is possible only if $\kappa = 1$, $\tilde{s} = 0$. Thus, Eq. (12.6.7) follows. For $\tilde{s} = 0$, we have

$$\kappa_1(0) = 1, \quad \kappa_2(0) = 7 - \sqrt{48},$$

and Eq. (12.6.8) follows by perturbing \tilde{s} .

If $\kappa = \kappa_1 = \kappa_2$ is a double root, then

$$\frac{2(\kappa - 1)}{\kappa} - \frac{(\kappa - 1)^2}{\kappa^2} - \frac{1}{3} \frac{(\kappa - 1)^3}{\kappa^2} + \frac{1}{6} \frac{(\kappa - 1)^4}{\kappa^3} = 0,$$

that is,

$$\begin{aligned} (\kappa - 1)(12\kappa^2 - 6\kappa(\kappa - 1) - 2\kappa(\kappa - 1)^2 + (\kappa - 1)^3) \\ = -(\kappa - 1)(\kappa + 1)(\kappa^2 - 8\kappa + 1) = 0. \end{aligned}$$

The roots $\kappa_1 = \kappa_2 = \pm 1$ are ruled out by Eqs. (12.6.7) and (12.6.8). This proves the lemma.

The corresponding eigenvalue problem reads

$$s\varphi_j = Q\varphi_j, \quad j = 1, 2, \dots, \quad (12.6.10a)$$

$$L_i \varphi_0 = 0, \quad i = 1, 2, \quad (12.6.10b)$$

$$\|\varphi\|_h < \infty, \quad (12.6.10c)$$

and we have the next lemma.

Lemma 12.6.2. *The quarter-space problem is not stable in the generalized sense if Eq. (12.6.5) has an eigenvalue with $\operatorname{Re} s > 0$, for some $h = h_0 > 0$.*

Proof. The general solution of the difference equation is

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j,$$

and s is an eigenvalue if

$$L_0 \varphi_0 = 0.$$

This condition is a linear system for the coefficients σ_1 and σ_2 , and its determinant is a function of $\tilde{s} = sh^2$. Therefore, any eigenvalue s_0 with $\operatorname{Re} s_0 > 0$ for some $h = h_0$ generates eigenvalues $s = s_0 h_0^2/h^2$, whose real part becomes arbitrarily large as $h \rightarrow 0$. This proves the lemma.

From now on we assume that there is no eigenvalue s with $\operatorname{Re} s > 0$. Then the system (12.6.5) has a unique solution. To estimate it we first consider the second case.

II. $c_1 \leq |\tilde{s}| \leq c_0$, $\operatorname{Re} \tilde{s} \geq 0$.

We can write the difference equation (12.6.5) in the form

$$\hat{v}_{j+2} = \alpha_1 \hat{v}_{j+1} + \alpha_0 \hat{v}_j + \alpha_{-1} \hat{v}_{j-1} + \alpha_{-2} \hat{v}_{j-2} + 12h^2 \hat{F}_j,$$

where the coefficients α_j are polynomials in \tilde{s} . As before, we introduce new variables by

$$\mathbf{y}_j = (\hat{v}_{j+1}, \hat{v}_j, \hat{v}_{j-1}, \hat{v}_{j-2})^T$$

and write Eq. (12.6.5) in the form

$$\mathbf{y}_{j+1} = M \mathbf{y}_j + h^2 \mathbf{F}_j, \quad j = 1, 2, \dots, \quad (12.6.11a)$$

where

$$M = \begin{bmatrix} \alpha_1 & \alpha_0 & \alpha_{-1} & \alpha_{-2} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{F}_j = 12 \begin{bmatrix} \hat{F}_j \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Using the difference equation, we can eliminate v_j , $j \geq 3$, from the boundary conditions. These now contain terms of order $|h^2 F_j| \leq h^{3/2} \|F\|_h$, and we write them in the form

$$H(\tilde{s}) \mathbf{y}_1 = \hat{\mathbf{g}}, \quad |\mathbf{g}| \leq \text{constant } h^{3/2} \|\hat{F}\|_h, \quad (12.6.11b)$$

where $H(\tilde{s})$ is a 2×4 matrix whose coefficients are polynomials in \tilde{s} .

The eigenvalues of M are the solutions of the characteristic equation. Thus,

by Lemma 12.6.1, for $|\tilde{s}| \geq c_1 > 0$, there are two eigenvalues $\kappa_j, j = 1, 2$, with $|\kappa_j| \leq 1 - \tau$, and two eigenvalues $\kappa_j, j = 3, 4$, with $|\kappa_j| \geq 1 + \tau$, $\tau = \text{constant} > 0$.

Because the two sets of eigenvalues are well separated, there is a smooth transformation $S = S(\tilde{s})$ such that

$$SMS^{-1} = \begin{bmatrix} M^I & 0 \\ 0 & M^{II} \end{bmatrix},$$

where

$$(M^I)^* M^I \leq 1 - \tau/2, \quad (M^{II})^* M^{II} \geq 1 + \tau/2.$$

Now substitute into Eq. (12.6.11) new variables $\mathbf{w} = S\mathbf{y}$, $\tilde{\mathbf{F}} = S\mathbf{F}$. Then, we obtain

$$\begin{aligned} \mathbf{w}_{j+1}^I &= M^I \mathbf{w}_j^I + h^2 \tilde{\mathbf{F}}_j^I, \\ \mathbf{w}_{j+1}^{II} &= M^{II} \mathbf{w}_j^{II} + h^2 \tilde{\mathbf{F}}_j^{II}, \quad j = 1, 2, \dots, \end{aligned} \quad (12.6.12a)$$

with boundary conditions

$$H^I \mathbf{w}_1^I + H^{II} \mathbf{w}_1^{II} = \mathbf{g}. \quad (12.6.12b)$$

The system (12.6.12) is of the same form as Eq. (12.4.14). The only difference is that $h\tilde{\mathbf{F}}_j$ is replaced by $h^2 \tilde{\mathbf{F}}_j$. Therefore, we obtain the corresponding estimate from Lemma 12.4.1 and Eq. (12.6.11b)

$$\begin{aligned} \|\mathbf{w}^{II}\|_h &\leq \text{constant } h^2 \|\tilde{\mathbf{F}}^{II}\|_h, \\ |\mathbf{w}_1^{II}| &\leq \text{constant } h^{3/2} \|\tilde{\mathbf{F}}^{II}\|_h, \\ \|\mathbf{w}^I\|_h &\leq \text{constant} (h^2 \|\tilde{\mathbf{F}}^I\|_h + h^{1/2} |\mathbf{w}_1^I|), \\ &\leq \text{constant} (h^2 \|\tilde{\mathbf{F}}^I\|_h + |(H^I)^{-1}| h^2 \|\tilde{\mathbf{F}}^{II}\|_h). \end{aligned} \quad (12.6.13)$$

We have assumed that the eigenvalue problem (12.6.10) has no eigenvalue for $\operatorname{Re} s > 0$. Thus, H^I is nonsingular for $\operatorname{Re} \tilde{s} > 0$. There are two possibilities.

1. H^I is nonsingular for $c_1 \leq |\tilde{s}| \leq c_0$, $\operatorname{Re} \tilde{s} \geq 0$. Then, $|H^I|$ is uniformly bounded, and, for the solution \hat{v} of the original problem, the estimate

(12.6.13) becomes

$$\|\hat{v}\|_h \leq \text{constant } h^2 \|\hat{F}\|_h \leq \frac{\text{constant}}{|s|} \|\hat{F}\|_h. \quad (12.6.14a)$$

Therefore,

$$\|D_+ \hat{v}\|_h \leq \text{constant } h \|\hat{F}\|_h \leq \frac{\text{constant}}{|s|^{1/2}} \|\hat{F}\|_h. \quad (12.6.14b)$$

Here we have used the fact that $h^2 \leq c_0/|s|$.

2. The eigenvalue problem has an eigenvalue $\tilde{s}_0 = i\tilde{\xi}_0$, $\tilde{\xi}_0$ real. Then $(H^I)^{-1}$ has a pole at $\tilde{s} = \tilde{s}_0$, and

$$|(H^I)^{-1}| \geq \frac{\sigma}{|\tilde{s} - \tilde{s}_0|}, \quad \sigma = \text{constant} > 0.$$

One can show that the estimate is sharp (Exercise 12.6.1). We have, therefore, the following lemma.

Lemma 12.6.3. *The estimates (12.6.14) hold for $c_1 \leq |\tilde{s}| \leq c_0$, $\operatorname{Re} \tilde{s} > 0$ if, and only if, the eigenvalue problem (12.6.10) has no eigenvalue for $\operatorname{Re} \tilde{s} \geq 0$, $c_1 \leq |\tilde{s}| \leq c_0$.*

REMARK. If there is an eigenvalue on the imaginary axis, only the estimate of $\|D_+ \hat{v}\|_h$ breaks down, but we still have an estimate for $\|\hat{v}\|_h$. However, the introduction of lower order terms requires that there is an estimate of $\|D_+ v\|_h$ for the original problem. Otherwise the perturbed problem will not be stable in general.

Now we consider the last case.

III. $|\tilde{s}| < c_1 \ll 1$.

Corresponding to the continuous problem, we write the resolvent equation (12.6.5) in the form

$$D_+ \hat{v}_j = s^{1/2} \hat{z}_j, \\ D_- \left(I - \frac{h^2}{12} D_+ D_- \right) \hat{z}_j + s^{-1/2} \hat{F}_j = s^{1/2} \hat{v}_j, \quad (12.6.15)$$

that is,

$$\begin{aligned}\hat{v}_{j+1} &= \hat{v}_j + \tilde{s}^{1/2} \hat{z}_j, \\ \hat{z}_{j+1} &= \beta_0 \hat{z}_j + \beta_{-1} \hat{z}_{j-1} + \beta_{-2} \hat{z}_{j-2} + \tilde{s}^{1/2} \beta_1 \hat{v}_j + 12h^2 \tilde{s}^{-1/2} \hat{F}_j.\end{aligned}$$

We introduce new variables

$$\mathbf{y}_j = (\hat{v}_j, \hat{z}_j, \hat{z}_{j-1}, \hat{z}_{j-2}).$$

Then Eq. (12.6.15) becomes

$$\begin{aligned}\mathbf{y}_{j+1} &= (M_0 + \tilde{s}^{1/2} M_1) \mathbf{y}_j + h^2 \tilde{s}^{-1/2} \mathbf{F}_j, \quad j = 1, 2, \dots, \\ \mathbf{F}_j &= \begin{bmatrix} 0 \\ 12 \hat{F}_j \\ 0 \\ 0 \end{bmatrix},\end{aligned}\tag{12.6.16a}$$

for certain matrices M_0 and M_1 that are independent of \tilde{s} and h . It is important to write the boundary conditions in terms of \hat{v} and \hat{z} . The extra boundary condition (12.6.4e) can always be written as a relation for z alone, whereas, for the other boundary conditions, this is only true for Neumann conditions (12.6.4c). (Recall that, at the moment, we are only considering homogeneous boundary conditions.)

Using the same procedure as we did when deriving Eq. (12.6.11b), we can now write the boundary condition in the form

$$H(\tilde{s}^{1/2}) \mathbf{y}_1 = \mathbf{g}, \quad |\mathbf{g}| \leq \text{constant } h^{3/2} \tilde{s}^{-1/2} \|\hat{F}\|_h.\tag{12.6.16b}$$

Here $H(\tilde{s}^{1/2})$ is an analytic function of $\tilde{s}^{1/2}$. The eigenvalues of $M_0 + \tilde{s}^{1/2} M_1$ are given in Lemma 12.6.1 (see Exercise 12.6.2). Also, M_0 has the form

$$M_0 = \begin{bmatrix} 1 & 0 \\ 0 & M_0^I \end{bmatrix}.\tag{12.6.17}$$

Therefore, there is a transformation $S = S(\tilde{s}^{1/2})$, analytic in $\tilde{s}^{1/2}$, such that

$$\begin{aligned}S(M_0 + \tilde{s}^{1/2} M_1) S^{-1} &= \begin{bmatrix} M^I & 0 \\ 0 & M^{II} \end{bmatrix}, \\ M^I &= \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix}, \quad M^{II} = \begin{bmatrix} \kappa_3 & 0 \\ 0 & \kappa_4 \end{bmatrix},\end{aligned}$$

where the κ_j are defined in Lemma 12.6.1. By the same lemma, it follows that

$$(M^I)^* M^I \leq 1 - \delta |\tilde{s}|^{1/2}, \quad (M^{II})^* M^{II} \geq 1 + \delta |\tilde{s}|^{1/2}, \quad \delta = \text{constant} > 0.$$

Now we can proceed as before. We substitute new variables $\mathbf{w} = S\mathbf{y}$, $\tilde{\mathbf{F}} = S\mathbf{F}$ into Eq. (12.6.16) and obtain Eq. (12.6.12) with $\tilde{\mathbf{F}}$ replaced by $\tilde{s}^{-1/2}\tilde{\mathbf{F}}$. We can again use Lemma 12.4.1 to estimate \mathbf{w} and we obtain

$$\begin{aligned} \|\mathbf{w}^{II}\|_h &\leq \text{constant} \frac{h^2 |\tilde{s}|^{-1/2}}{|\tilde{s}|^{1/2}} \|\tilde{\mathbf{F}}^{II}\|_h \leq \frac{\text{constant}}{|s|} \|\hat{\mathbf{F}}\|_h, \\ |\mathbf{w}_1^{II}| &\leq \text{constant} \frac{h^2 |\tilde{s}|^{-1/2}}{h^{1/2} |\tilde{s}|^{1/4}} \|\tilde{\mathbf{F}}^{II}\|_h \leq \frac{\text{constant}}{|s|^{3/4}} \|\tilde{\mathbf{F}}^{II}\|_h, \\ \|\mathbf{w}^I\|_h &\leq \text{constant} \left(\frac{h^2 |\tilde{s}|^{-1/2}}{|\tilde{s}|^{1/2}} \|\tilde{\mathbf{F}}^I\|_h + |(H^I)^{-1}| \frac{h^2}{|\tilde{s}|^{1/4}} \cdot \frac{1}{|s|^{3/4}} \|\tilde{\mathbf{F}}^{II}\|_h \right), \\ &\leq \text{constant} \frac{1 + |(H^I)^{-1}|}{|s|} \|\hat{\mathbf{F}}\|_h. \end{aligned} \tag{12.6.18}$$

We obtain, therefore, the next lemma.

Lemma 12.6.4. *If H^I is nonsingular at $\tilde{s} = 0$, then for $|\tilde{s}| \leq c_1$, c_1 sufficiently small, the solutions of the resolvent equation (12.6.5) satisfy the estimate*

$$\begin{aligned} \|\hat{v}\|_h &\leq \frac{\text{constant}}{|s|} \|\hat{\mathbf{F}}\|_h, \\ \|D_+ \hat{v}\|_h &= |s|^{1/2} \|\hat{z}\|_h \leq \frac{\text{constant}}{|s|^{1/2}} \|\hat{\mathbf{F}}\|_h. \end{aligned} \tag{12.6.19}$$

Again, the estimates in Eq. (12.6.18) are sharp (Exercise 12.6.3) and, therefore, there is no acceptable estimate if $H^I(0)$ is singular.

The estimates (12.6.6), (12.6.14), and (12.6.19) give us the following theorem.

Theorem 12.6.1. *The problem is stable in the generalized sense if, and only if, H^I is nonsingular for $\operatorname{Re} \tilde{s} \geq 0$.*

We can also express this result as an eigenvalue condition. Consider the eigenvalue problem (12.6.10), and write it in the form

$$\begin{aligned} s^{1/2} \psi_j &= D_+ \varphi_j, \\ s^{1/2} \varphi_j &= D_- \left(I - \frac{h^2}{12} D_+ D_- \right) \psi_j. \end{aligned} \tag{12.6.20}$$

Change the boundary conditions to boundary conditions for ψ when possible. Then we obtain the following lemma.

Lemma 12.6.5. *H^l is nonsingular for $\tilde{s} = 0$ if, and only if, $\tilde{s} = 0$ is not an eigenvalue or generalized eigenvalue of (12.6.20).*

We summarize our results in the next theorem.

Theorem 12.6.2. *The approximation (12.6.4) is stable in the generalized sense if, and only if, the eigenvalue problem (12.6.20) with boundary conditions, written as conditions for ψ whenever possible, has no eigenvalue or generalized eigenvalue with $\operatorname{Re} \tilde{s} \geq 0$.*

REMARK. For $\tilde{s} \neq 0$, $\operatorname{Re} s \geq 0$, there are no generalized eigenvalues because, for the solution κ_j of the characteristic equation, we have $|\kappa_j| \neq 1$. In this case, the eigenvalues of Eq. (12.6.20) and Eq. (12.6.10) are identical. For $\tilde{s} = 0$, we have $\kappa_1 = 1$. Therefore, for Neumann boundary conditions, $\tilde{s} = 0$ is always a generalized eigenvalue of Eq. (12.6.10) ($v_j \equiv \text{constant}$ is a solution independent of time) but not of Eq. (12.6.20) if H^l is nonsingular. This is the reason why we have written the eigenvalue problem in the form of Eq. (12.6.20). If $\tilde{s} = 0$ is not an eigenvalue or generalized eigenvalue of Eq. (12.6.10), then it is also not an eigenvalue or generalized eigenvalue of Eq. (12.6.20). Therefore, we need only switch to Eq. (12.6.20) if Eq. (12.6.10) has an eigenvalue or generalized eigenvalue at $s = 0$.

Now we consider two examples.

EXAMPLE 1. Consider the boundary condition $u(0, t) = 0$. This example was analyzed by the energy method in Section 11.2. The differential equation gives us $u_t(0, t) = u_{xx}(0, t) = 0$. Therefore, we use

$$v_0 = 0, \quad D_+ D_- v_0 = 0$$

as numerical boundary conditions. The truncation error for the second condition consists of only even-order space derivatives. Because all these are zero at $x = 0$, the condition has arbitrarily high-order accuracy. For $\tilde{s} \neq 0$, we need not rewrite the eigenvalue problem. If $\kappa_1 \neq \kappa_2$, then the general solution of Eq. (12.6.10a) is

$$\varphi_j = \sigma_1 \kappa_1^j + \sigma_2 \kappa_2^j. \quad (12.6.21)$$

Introducing this expression into the boundary condition gives us

$$\begin{aligned}\sigma_1 + \sigma_2 &= 0, \\ \sigma_1 \frac{(\kappa_1 - 1)^2}{\kappa_1} + \sigma_2 \frac{(\kappa_2 - 1)^2}{\kappa_2} &= 0.\end{aligned}$$

This system of linear equations has a nontrivial solution if

$$\frac{(\kappa_1 - 1)^2}{\kappa_1} = \frac{(\kappa_2 - 1)^2}{\kappa_2};$$

that is,

$$\kappa_1 + \frac{1}{\kappa_1} = \kappa_2 + \frac{1}{\kappa_2},$$

or

$$(\kappa_1 - \kappa_2) \left(1 - \frac{1}{\kappa_1 \kappa_2} \right) = 0.$$

We know that, for all \tilde{s} with $\operatorname{Re} \tilde{s} \geq 0$, the product $|\kappa_1 \kappa_2| < 1$. Therefore, there is no eigenvalue or generalized eigenvalue if $\kappa_1 \neq \kappa_2$. If $\kappa_1 = \kappa_2$, then the general solution has the form

$$\varphi_j = \sigma_1 \kappa_1^j + j \sigma_2 \kappa_1^j. \quad (12.6.22)$$

In this case, we obtain from the boundary conditions,

$$\begin{aligned}\sigma_1 &= 0, \\ \sigma_2 \left(\kappa_1 - \frac{1}{\kappa_1} \right) &= 0.\end{aligned}$$

Thus, the system has a nontrivial solution if $\kappa_1 = \pm 1$. If $\kappa_1 = 1$, then $\tilde{s} = 0$ and, by Lemma 12.6.1, $\kappa_2 \neq \kappa_1$, which is a contradiction. The other possible solution $\kappa_1 = -1$ implies $\operatorname{Re} s < 0$, which is of no interest.

Thus, by Theorem 12.6.2 and the remark following it, the approximation is stable in the generalized sense.

EXAMPLE 2. Consider the boundary conditions $u_x(0, t) = 0$. The differential equation gives us $u_{tx}(0, t) = u_{xxx}(0, t) = 0$. Therefore, we use

$$D_0 v_{-1} = 0, \quad D_+^3 v_{-1} = 0. \quad (12.6.23)$$

For $\kappa_1 \neq \kappa_2$, we now obtain

$$\begin{aligned} \sigma_1 \left(\kappa_1 - \frac{1}{\kappa_1} \right) + \sigma_2 \left(\kappa_2 - \frac{1}{\kappa_2} \right) &= 0, \\ \sigma_1 \frac{(\kappa_1 - 1)^3}{\kappa_1} + \sigma_2 \frac{(\kappa_2 - 1)^3}{\kappa_2} &= 0. \end{aligned}$$

This system has a nontrivial solution if, and only if,

$$\frac{(\kappa_1^2 - 1)(\kappa_2 - 1)^3}{\kappa_1 \kappa_2} = \frac{(\kappa_2^2 - 1)(\kappa_1 - 1)^3}{\kappa_1 \kappa_2};$$

that is, for $\kappa_1 \neq 1$,

$$(\kappa_1 + 1)(\kappa_2 - 1)^2 = (\kappa_2 + 1)(\kappa_1 - 1)^2,$$

or

$$(\kappa_2 - \kappa_1)(\kappa_1 \kappa_2 + \kappa_2 + \kappa_1 - 3) = 0.$$

Since $|\kappa_1 \kappa_2 + \kappa_2 + \kappa_1| < 3$, for $\operatorname{Re} \tilde{s} \geq 0$, the only possibility of a nontrivial solution is $\kappa_1 = \kappa_2$ or $\kappa_1 = 1$.

If $\kappa_1 = \kappa_2$, then φ is of the form of Eq. (12.6.22) and the linear system of equations is

$$\begin{aligned} \sigma_1 \left(\kappa_1 - \frac{1}{\kappa_1} \right) + \sigma_2 \left(\kappa_1 + \frac{1}{\kappa_1} \right) &= 0, \\ \sigma_1 \frac{(\kappa_1 - 1)^3}{\kappa_1} + \sigma_2 \left(2\kappa_1^2 - 3\kappa_1 + \frac{1}{\kappa_1} \right) &= 0. \end{aligned}$$

Thus, there is a nontrivial solution if, and only if,

$$\left(2\kappa_1^2 - 3\kappa_1 + \frac{1}{\kappa_1} \right) \left(\kappa_1 - \frac{1}{\kappa_1} \right) = \frac{(\kappa_1 - 1)^3}{\kappa_1} \left(\kappa_1 + \frac{1}{\kappa_1} \right).$$

By Lemma 12.6.1, the only possible κ_1 is $\kappa_1 = 4 - \sqrt{15} \approx 0.12$. On inspection,

it follows that the above relation cannot hold. Thus, there is no eigenvalue for $\operatorname{Re} \tilde{s} \geq 0$, $\tilde{s} \neq 0$.

For the eigenvalue problem in its original form, $\tilde{s} = 0$ is a generalized eigenvalue corresponding to $\kappa_1 = 1$. Therefore, we have to modify it to the form shown in Eq. (12.6.20). The boundary conditions become

$$\psi_0 + \psi_{-1} = 0, \quad D_+^2 \psi_{-1} = 0. \quad (12.6.24)$$

Both φ_j and ψ_j are linear combinations of κ_1^j and κ_2^j if $\kappa_1 \neq \kappa_2$. Thus, in the neighborhood of $\tilde{s} = 0$, the general solution of Eq. (12.6.20) can be written as

$$\varphi_j \approx -\sigma_1(1 - \tilde{s}^{1/2})^j + \sigma_2 \frac{\tilde{s}^{1/2}}{\kappa_2(0) - 1} \kappa_2^j(0).$$

Therefore,

$$\tilde{s}^{1/2} \psi_j = \varphi_{j+1} - \varphi_j \approx \sigma_1 \tilde{s}^{1/2} (1 - \tilde{s}^{1/2})^j + \sigma_2 \tilde{s}^{1/2} \kappa_2^j(0);$$

that is,

$$\psi_j \approx \sigma_1 (1 - \tilde{s}^{1/2})^j + \sigma_2 \kappa_2^j(0).$$

(The choice of coefficients in the representation of φ_j was made to get a simple normalized form for ψ_j .) Substituting this expression into the boundary conditions (12.6.24) shows that $\sigma_1 = \sigma_2 = 0$. Thus, $\tilde{s} = 0$ is not a generalized eigenvalue and the approximation is stable in the generalized sense.

The above results can be extended to very general parabolic systems. We can neglect all lower order terms and we can prove the following theorem.

Theorem 12.6.3. *The approximation for general parabolic systems is stable in the generalized sense if it is dissipative for the Cauchy problem and if the modified eigenvalue problem has no eigenvalues or generalized eigenvalues for $\operatorname{Re} \tilde{s} \geq 0$.*

EXERCISES

12.6.1. Prove that the estimates (12.6.14) are sharp.

12.6.2. Consider the matrix $M_0 + \tilde{s}^{1/2} M_1$ in Eq. (12.6.16a). Prove that its eigenvalues are the roots κ given by the characteristic equation given in Lemma 12.6.1.

12.6.3. Prove that the estimates (12.6.18) are sharp.

12.6.4. Introduce nonzero data in the boundary conditions of Example 2 above and prove that the approximation is strongly stable in the generalized sense.

12.7. THE CONVERGENCE RATE

In Section 11.3, the basic principles for obtaining error estimates were discussed in connection with the energy method for stability analysis. In this section, we present a more detailed discussion, which also includes the more general stability analysis based on the Laplace transform.

The stability estimate obtained for a certain approximation is the key to the proper error estimates. This was demonstrated in Part I for the pure initial value problem and the same principles also hold when boundary conditions are involved. However, to obtain optimal estimates, one has to be a little more careful.

The basic idea is to insert the true solution $u(x, t)$ into the approximation. This generally introduces truncation errors as inhomogeneous terms in the difference approximation, the initial conditions, and the boundary conditions. The stability estimate then gives the desired error estimate, because small data only can give small solutions.

The error $e_j(t) = v_j(t) - u(x_j, t)$ satisfies the equations

$$\frac{de_j}{dt} = Qe_j + h^{q_1} F_j, \quad j = 1, 2, \dots, \quad (12.7.1a)$$

$$e_j(0) = h^{q_2} f_j, \quad (12.7.1b)$$

$$L_0 e_0 = h^{q_3} g, \quad (12.7.1c)$$

where F , f , and g are smooth functions. (For convenience, we use the same notation for these functions as in the original approximation for $v(t)$. Furthermore, we consider only the quarter-space problem even when the energy method is used.) The integers q_1 , q_2 , and q_3 are not necessarily equal. Equations (12.7.1) are based on the fact that there is a smooth solution $u(x, t)$ to the continuous problem, and we recall that this requires certain compatibility conditions on the initial and boundary data and on the forcing function if there is one.

Let us first consider the case that Q is a semibounded operator, so that the energy method can be applied. This requires homogeneous boundary conditions, and our recipe above was to subtract a certain smooth function $\varphi_j(t)$ that satisfies the boundary condition. Assuming that this is possible and that $\varphi_j(t) = h^{q_3} \psi_j(t)$, where $\psi_j(t)$ is smooth, we have

$$\frac{d\varphi_j}{dt} = \mathcal{O}(h^{q_3}), \quad (12.7.2a)$$

$$Q\varphi_j = \mathcal{O}(h^{q_3}), \quad (12.7.2b)$$

which yields, for $\tilde{e}_j(t) = e_j(t) - \varphi_j(t)$,

$$\begin{aligned} \frac{d\tilde{e}_j}{dt} &= Q\tilde{e}_j + (h^{q_1} + h^{q_3})\tilde{F}_j, \quad j = 1, 2, \dots, \\ \tilde{e}_j(0) &= (h^{q_2} + h^{q_3})\tilde{f}_j, \\ L_0\tilde{e}_0 &= 0. \end{aligned} \quad (12.7.3)$$

The energy estimate yields a bound on any finite time interval $[0, T]$,

$$\|\tilde{e}(t)\|_h \leq \text{constant } (h^{q_1} + h^{q_2} + h^{q_3}),$$

and by construction

$$\|e(t)\|_h \leq \text{constant } h^q, \quad q = \min(q_1, q_2, q_3). \quad (12.7.4)$$

[This estimate corresponds to Theorem 11.3.1.] We often have $q_2 = \infty$, and q_1 is given by the order of the approximation at inner points. It is, therefore, natural to choose boundary conditions such that $q_3 = q_1$.

The crucial issue is the construction of $\varphi_j(t)$. Let us first assume that all the boundary conditions in Eq. (12.7.1c) are approximations of the boundary conditions for the differential equation. For example, for a scalar parabolic equation we could approximate $au_x(0, t) + bu(0, t) = 0$ by

$$a \frac{v_1 - v_0}{h} + b \frac{v_1 + v_0}{2} = 0. \quad (12.7.5)$$

If the grid is located such that $x_0 = -h/2$, then we have $q_3 = 2$ in Eq. (12.7.1c) with g being a combination of u derivatives. Hence, φ can be constructed such that it satisfies Eq. (12.7.2). For the equation $u_t + u_x = 0$, $u(0, t) = 0$, the same conclusion holds with $a = 0$, and $b = 1$ in Eq. (12.7.5). This is a general principle: As long as Eq. (12.7.1c) does not contain any extra boundary conditions, the error estimate follows immediately when there is an energy estimate. (One can also show that $q_3 \geq q_1$ is a necessary condition for an h^{q_1} estimate for this type of boundary conditions.)

Let us next consider the case that there are extra boundary conditions. We use the familiar example $u_t = u_x$ with extrapolation at the boundary

$$v_0 - 2v_1 + v_2 = 0.$$

The error satisfies

$$e_0 - 2e_1 + e_2 = -h^2 u_{xx}(0, t) + \mathcal{O}(h^3) =: h^2 g(t),$$

and we need a function $\varphi_j(t) = h^2 \psi_j(t)$, where

$$\psi_0 - 2\psi_1 + \psi_2 = g(t). \quad (12.7.6)$$

But if ψ is smooth, we have

$$\psi_0 - 2\psi_1 + \psi_2 \approx h^2 \psi_{xx}(0, t),$$

and since g is, in general, not small, Eq. (12.7.6) is impossible. [If the solution happens to satisfy $u_{xx}(h, t) = 0$ the construction would be possible.]

Now consider the usual centered second-order approximation for inner points. As noted earlier, the linear extrapolation condition at $j = 0$ is equivalent to

$$\frac{dv_0}{dt} = D_+ v_0, \quad (12.7.7)$$

where the grid function index has been shifted one step, $j \rightarrow j - 1$. By defining

$$Qv_j = \begin{cases} D_0 v_j, & j = 1, 2, \dots, \\ D_+ v_0, & j = 0, \end{cases}$$

we can write the approximation as

$$\begin{aligned} \frac{dv_j}{dt} &= Qv_j, & j &= 0, 1, \dots, \\ v_j(0) &= f_j \end{aligned} \quad (12.7.8)$$

without any boundary condition. Because Q is only first-order accurate at $x = 0$, the error equation is

$$\begin{aligned} \frac{de_j}{dt} &= Qe_j + F_j, & j &= 0, 1, \dots, \\ e_j(0) &= 0, \end{aligned} \quad (12.7.9)$$

where

$$F_j = \begin{cases} \mathcal{O}(h), & j = 0, \\ \mathcal{O}(h^2), & j = 1, 2, \dots, \end{cases}$$

$$\|F\|_{1,\infty} = \mathcal{O}(h^2).$$

By stability and Duhamel's principle, it follows that

$$\|e(t)\|_{0,\infty}^2 \leq \text{constant} \|F(t)\|_{0,\infty}^2 = \mathcal{O}(h^3), \quad 0 \leq t \leq T.$$

However, this estimate is not optimal. Since the approximation satisfies the Kreiss condition, we can do better. Returning to the original formulation with an explicit boundary condition, the error equation is

$$\begin{aligned} \frac{de_j}{dt} &= D_0 e_j + h^2 F_j, \quad j = 1, 2, \dots, \\ e_j(0) &= h^2 f_j, \\ e_0 - 2e_1 + e_2 &= h^2 g. \end{aligned} \tag{12.7.10}$$

We split the error into two parts $e = e^{(1)} + e^{(2)}$, where

$$\begin{aligned} \frac{de_j^{(1)}}{dt} &= D_0 e_j^{(1)} + h^2 F_j, \quad j = 1, 2, \dots, \\ e_j^{(1)}(0) &= h^2 f_j, \\ e_0^{(1)} - 2e_1^{(1)} + e_2^{(1)} &= 0, \end{aligned} \tag{12.7.11}$$

$$\begin{aligned} \frac{de_j^{(2)}}{dt} &= D_0 e_j^{(2)}, \quad j = 1, 2, \dots, \\ e_j^{(2)}(0) &= 0, \\ e_0^{(2)} - 2e_1^{(2)} + e_2^{(2)} &= h^2 g. \end{aligned} \tag{12.7.12}$$

This approximation was shown to be stable in Section 11.1, and we immediately get

$$\|e^{(1)}(t)\|_h \leq \text{constant } h^2,$$

where the norm is based on the scalar product

$$(v, w)_h = \frac{h}{2} v_1 w_1 + \sum_{j=2}^{\infty} v_j w_j h$$

for real grid functions v and w .

To estimate $e^{(2)}$, we Laplace transform Eq. (12.7.12), use the fact that the Kreiss condition is satisfied, and transform back again. Because D_0 is semi-bounded for the Cauchy problem, we use the same procedure as in Section 12.2 to obtain the estimate

$$\|e^{(2)}(t)\|_{1,\infty}^2 \leq \text{constant} \int_0^t |h^2 g(\tau)|^2 d\tau, \quad (12.7.13)$$

(see Theorem 12.2.2). This implies the final estimate

$$\|e(t)\|_h \leq \text{constant } h^2, \quad 0 \leq t \leq T. \quad (12.7.14)$$

The technique we have used for deriving the optimal estimate (12.7.14) is similar to the one used to derive strong stability in Section 12.2. In fact, recalling that the approximation in our example is strongly stable, the result follows immediately from the error equation (12.7.10).

Considering the method in the form (12.7.8), the approximation of $\partial/\partial x$ is only first order accurate at $x = 0$. Still we have shown that there is an overall h^2 accuracy. This is possible because the lower order approximation is applied only at one point. This is the background for the expression “one order less accuracy at the boundary is allowed.” This statement is only valid if it refers to “extra” boundary conditions. The “physical” boundary conditions must always be approximated to the same order as the differential operator at inner points.

Let us next consider the problem

$$u_t = Au_x + F, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (12.7.15a)$$

$$u(x, 0) = f(x), \quad (12.7.15b)$$

$$u''(0, t) = R^I u^I(0, t) + g(t), \quad (12.7.15c)$$

and the fourth-order approximation

$$\frac{dv_j}{dt} = A \left(\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h) \right) v_j + F_j, \quad j = 1, 2, \dots, \quad (12.7.16a)$$

$$v_j(0) = f_j, \quad (12.7.16b)$$

where

$$A = \begin{bmatrix} \Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix}, \quad \Lambda^I > 0, \quad \Lambda^{II} < 0.$$

This is a generalization of the example in Section 12.3. By differentiating the boundary conditions (12.7.15c) twice with respect to t and using the differential equation (12.7.15a) we get

$$u_{xx}^{II}(0, t) = S^I u_{xx}^I(0, t) + \tilde{g}(t),$$

where

$$S^I = (\Lambda^{II})^{-2} R^I (\Lambda^I)^2, \quad (12.7.17)$$

$$\tilde{g}(t) = (\Lambda^{II})^{-2} (R^I (\Lambda^I F_x^I(0, t) + F_t^I(0, t)) - \Lambda^{II} F_x^{II}(0, t) - F_t^{II}(0, t) + g_{tt}(t)).$$

As boundary conditions for the approximation, we use

$$v_0^{II}(t) = R^I v_0^I(t) + g(t), \quad (12.7.16c)$$

$$D_+ D_- v_0^{II}(t) = S^I D_+ D_- v_0^I(t) + \tilde{g}(t), \quad (12.7.16d)$$

$$D_+^4 v_0^I(t) = 0, \quad (12.7.16e)$$

$$D_+^4 v_{-1}^I(t) = 0. \quad (12.7.16f)$$

Let $u(x, t)$ be a smooth solution of Eq. (12.7.15), and consider the truncation error for the extra boundary conditions (12.7.16d,e,f). We have for $e = u - v$

$$D_+ D_- e_0^{II} = S^I D_+ D_- e_0^I + \mathcal{O}(h^2),$$

$$D_+^4 e_0^I = \mathcal{O}(1),$$

$$D_+^4 e_{-1}^I = \mathcal{O}(1).$$

However, the normalized form corresponding to (12.1.12) is

$$\begin{aligned} h^2 D_+ D_- e_0^{II}(t) &= S^I h^2 D_+ D_- e_0^I(t) + \mathcal{O}(h^4), \\ (hD_+)^4 e_0^I(t) &= \mathcal{O}(h^4), \\ (hD_+)^4 e_{-1}^I(t) &= \mathcal{O}(h^4), \end{aligned} \quad (12.7.18)$$

which yields the right order of truncation error. The other error equations are

$$\begin{aligned} \frac{de_j(t)}{dt} &= A \left(\frac{4}{3} D_0(h) - \frac{1}{3} D_0(2h) \right) e_j(t) + \mathcal{O}(h^4), \quad j = 1, 2, \dots, \\ e_j(0) &= 0, \\ e_0^{II}(t) &= R^I e_0^I(t). \end{aligned} \tag{12.7.19}$$

The approximations (12.7.18) and (12.7.19) were analyzed in Section 12.3 and shown to be strongly stable. Thus, we get the error estimate

$$\|e(t)\|_h \leq \text{constant } h^4. \tag{12.7.20}$$

Note that Eq. (12.7.16d) is only a second-order approximation of Eq. (12.7.17), and we still have an h^4 error in the solution. This is possible, since it is an extra numerical boundary condition. The condition (12.7.17) is also “extra” in the sense that it is not required for defining a unique solution of the differential equation.

In summary, when the extra boundary conditions are written in the normalized form as in Eq. (12.1.12), the error term must be of the same order as the truncation error at inner points.

Now consider the case where the approximation is strongly stable in the generalized sense. If there is no error in the initial data, then the error estimate follows immediately from Eq. (12.4.1b). For Eq. (12.7.1), we get

$$\begin{aligned} \int_0^\infty e^{-2\eta t} \|e(t)\|_h^2 dt &\leq K(\eta) \int_0^\infty e^{-2\eta t} (h^{2q_1} \|F(t)\|_h^2 + h^{2q_3} |g(t)|^2) dt, \\ &= \mathcal{O}(h^{2q_1} + h^{2q_3}). \end{aligned} \tag{12.7.21}$$

Assume that there is an initial error

$$e_j(0) = h^{q_2} f_j,$$

where f_j is smooth; that is,

$$|D_+^\nu f_j| \leq \text{constant}, \quad \nu = 0, 1, \dots$$

Then let $\varphi_j(t) = e^{-\alpha t} f_j$, $\alpha > 0$, and define $\tilde{e}_j(t)$ by

$$\tilde{e}_j(t) = e_j(t) - h^{q_1} \varphi_j(t), \quad j = 1, 2, \dots \quad (12.7.22)$$

We obtain

$$\begin{aligned} \frac{d\tilde{e}_j}{dt} &= Q\tilde{e}_j + h^{q_1} F_j + h^{q_2} \tilde{F}_j, \quad j = 1, 2, \dots, \\ \tilde{e}_j(0) &= 0, \\ L_0 \tilde{e}_0(t) &= h^{q_3} g(t) + h^{q_2} \tilde{g}(t), \end{aligned} \quad (12.7.23)$$

which yields the estimate

$$\begin{aligned} \int_0^\infty e^{-2\eta t} \|e(t)\|_h^2 dt &\leq 2 \int_0^\infty e^{-2\eta t} (\|\tilde{e}(t)\|_h^2 + h^{2q_2} \|\varphi(t)\|_h^2) dt, \\ &= \mathcal{O}(h^{2q_1} + h^{2q_2} + h^{2q_3}). \end{aligned} \quad (12.7.24)$$

Note that we don't have the same difficulty when making the initial condition homogeneous as we have in some cases when making the boundary conditions homogeneous. The only requirement is that f_j be smooth.

Next assume that the approximation is stable and that the Kreiss condition is not satisfied. Then the error estimate does not follow directly from the stability estimate because it does not permit nonzero boundary data. The procedure of splitting the error into $e = e^{(1)} + e^{(2)}$, as demonstrated above, cannot be used either, because we need the Kreiss condition to estimate $e^{(2)}$. One alternative is to subtract a suitable function that satisfies the inhomogeneous boundary condition. Another alternative is to eliminate the boundary values $v_{-r+1}, v_{-r+2}, \dots, v_0$ and modify the difference operator Q near the boundary. However, as we have already seen above, we may lose accuracy in this process.

Finally, we consider the case where the approximation is stable in the generalized sense and the Kreiss condition is not satisfied. Now we must construct a function that satisfies both the inhomogeneous initial and boundary conditions. This may be tricky, because we cannot in general expect compatibility at the corner $x = 0, t = 0$, even if the solution $u(x, t)$ is smooth. The reason for this is that the truncation error in the boundary conditions is different from the truncation error in the initial condition, the latter one typically being zero. And even if such a function exists, we may lose accuracy in the subtraction process, just as in the previous case.

The most general theory based on the Laplace transform method for obtaining optimal error estimates in this case is given in Gustafsson (1981). The essential condition is that there be no eigenvalue or generalized eigenvalue at $s = 0$. (The theory is given for the fully discrete case where $s = 0$ corresponds to $z = 1$.)

EXERCISES

- 12.7.1.** Prove that the error $\|v_j(t) - u(x_j, t)\|_h$ of the approximation (11.4.3) and (11.4.5) is $\mathcal{O}(h^4)$.
- 12.7.2.** Use the result of Exercise 12.6.4 to prove that the approximation (12.6.4) and (12.6.23) gives a fourth-order accurate solution.

BIBLIOGRAPHIC NOTES

The first general stability theory for semidiscrete approximations based on the Laplace transform technique was given by Strikwerda (1980). The stability concept there corresponds to strong stability in the generalized sense for hyperbolic problems. Strikwerda also proves stability for a fourth-order approximation of $u_t = \pm u_x$, but with different boundary conditions than ours.

The method of lines has been used extensively in applications, for example in fluid dynamics. However, very little analysis has been done. In Gustafsson and Kreiss (1983) and Johansson (1993), semidiscrete approximations of model problems corresponding to incompressible flow are analyzed.

Gustafsson and Oliger (1982) derived stable boundary conditions for a number of implicit time discretizations of a centered second-order in space approximation of the Euler equations.

Regarding optimal error estimates for approximations that are stable in the generalized sense, see Notes on Chapter 13.

13

THE LAPLACE TRANSFORM METHOD FOR FULLY DISCRETE APPROXIMATIONS: Normal Mode Analysis

In Chapter 11, we discussed fully discrete approximations for initial-boundary-value problems, and the energy method was used for stability analysis. For the examples treated there, we first discretized space and left time continuous. We derived stability estimates for the systems of ODEs obtained in this way. For some standard methods of time discretization, like the backward Euler and trapezoidal methods, we obtained fully discrete methods in a straightforward manner. Stability is more difficult to verify for more general time discretizations. In this chapter, we instead use the more general Laplace transform technique. In particular, we shall use the concept of stability in the generalized sense to analyze the method of lines in Section 13.2.

More general methods are obtained by discretizing space and time simultaneously. For example, the Lax–Wendroff method cannot be obtained by the method of lines as used before. For these general methods, only the Laplace transform technique leads to stability results. Furthermore, it gives necessary and sufficient conditions for stability. This type of analysis for fully discrete approximations is often called *normal mode analysis* or *GKS-analysis*.

13.1. GENERAL THEORY FOR APPROXIMATIONS OF HYPERBOLIC SYSTEMS

We consider the quarter-space problem for systems

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x} + F, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (13.1.1a)$$

with initial data

$$u(x, 0) = f(x) \quad (13.1.1b)$$

and boundary conditions

$$L_0 u(0, t) = g. \quad (13.1.1c)$$

To construct difference approximations we introduce a space step h_0 and a time step $k > 0$ that generate a mesh with meshpoints $(x_j, t_n) = (jh, nk)$, where $j = -r+1, -r+2, \dots$, and $n = 0, 1, \dots$, as shown in Figure 13.1.1. The approximation is denoted by v_j^n .

The gridpoints (x_j, t_n) with $j \geq 1$ are called *interior points* and the other points are called *boundary points*. We shall use the scalar product and norm

$$(v, w)_h := (v, w)_{1,\infty} = \sum_{j=1}^{\infty} \langle v_j, w_j \rangle h, \quad \|v\|_h^2 = (v, v)_h,$$

respectively.

The simplest approximations are explicit one-step methods.

$$v_j^{n+1} = Q v_j^n + k F_j^n, \quad j = 1, 2, \dots, \quad (13.1.2a)$$

$$v_j^0 = f_j, \quad (13.1.2b)$$

$$L_0 v_0^{n+1} = g^{n+1}. \quad (13.1.2c)$$

Here

$$Q = \sum_{\nu=-r}^p B_\nu E^\nu \quad (13.1.3)$$

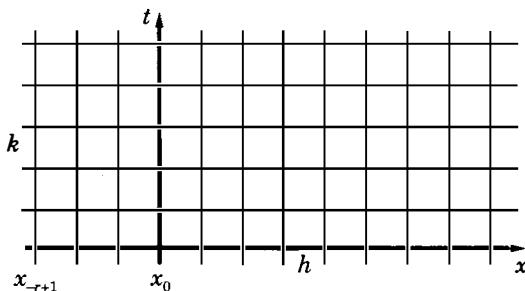


Figure 13.1.1.

is a difference operator of the type that we have discussed earlier, and we assume that B_{-r}, B_p are nonsingular. It is assumed that the boundary conditions (13.1.2c) are such that the approximation is solvable, that is, that $\{v_j^{n+1}\}$ is uniquely defined by $\{v_j^n\}$. This is the case if, for example, the boundary conditions can be written in the form

$$v_\mu^{n+1} = \sum_{\nu=1}^q (L_{\mu\nu}^{(0)} v_\nu^n + L_{\mu\nu}^{(1)} v_\nu^{n+1}) + g_\mu^{n+1}, \quad \mu = -r + 1, \dots, 0. \quad (13.1.4)$$

We now introduce the general technique based on the Laplace transform. We need the grid vector function v_j^n and the data to be defined for all t . Therefore, we define

$$v_j(t) = v_j^n \quad \text{for } t_n \leq t < t_{n+1}$$

and, correspondingly, for the data F and g . We also define $g(t) = 0$ for $0 \leq t \leq k$. The transform

$$\hat{v}_j(s) = \int_0^\infty e^{-st} v_j(t) dt$$

is now well-defined. Assuming that $f_j \equiv 0$, we get

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

Therefore, the Laplace transform of Eq. (13.1.2) with $f_j = 0$ is

$$e^{sk} \hat{v}_j = Q \hat{v}_j + k \hat{F}_j, \quad j = 1, 2, \dots, \quad (13.1.5a)$$

$$\hat{L}_0 \hat{v}_0 = \hat{g}, \quad (13.1.5b)$$

$$\|\hat{v}\|_h < \infty. \quad (13.1.5c)$$

\hat{L}_0 is the Laplace transform of L_0 . Written in the form of Eq. (13.1.4), these equations become

$$\hat{v}_\mu = \sum_{\nu=1}^q (e^{-sk} L_{\mu\nu}^{(0)} \hat{v}_\nu + L_{\mu\nu}^{(1)} \hat{v}_\nu) + \hat{g}, \quad \mu = -r + 1, \dots, 0. \quad (13.1.6)$$

For convenience, we assume that Q is only the principle part; that is, there is no explicit dependence on h or k . The coefficient matrices B_ν depend on $\lambda = k/h$, and we assume that λ is a constant when h varies.

The Godunov–Ryabenkii condition and the Kreiss condition can now be defined as in the semidiscrete case and the corresponding estimates are obtained in the same way. It is convenient to define the conditions in terms of $z = e^{sk}$. Because we are dealing with the principle part, we are concerned with the behavior of the solution $\hat{v}_j(s)$ for $\operatorname{Re} s \geq 0$; that is, for $|z| \geq 1$. Let \tilde{v}_j be defined by

$$\tilde{v}_j(z) = \tilde{v}_j(e^{sk}) := \hat{v}_j(s)$$

and, correspondingly, for \tilde{F}_j and \tilde{g}_j . We also define $\tilde{L}_0(z) := \hat{L}_0(s)$. The problem (13.1.5) can now be written in the form

$$z\tilde{v}_j = Q\tilde{v}_j + k\tilde{F}_j, \quad j = 1, 2, \dots, \quad (13.1.7a)$$

$$\tilde{L}_0\tilde{v}_0 = \tilde{g}, \quad (13.1.7b)$$

$$\|\tilde{v}\|_h < \infty, \quad (13.1.7c)$$

which is called the z transformed problem. The corresponding eigenvalue problem is

$$\begin{aligned} z\varphi_j &= Q\varphi_j, \quad j = 1, 2, \dots, \\ \tilde{L}_0\varphi_0 &= 0, \\ \|\varphi\|_h &< \infty, \end{aligned} \quad (13.1.8)$$

and we have the following lemma.

Lemma 13.1.1. *If Eq. (13.1.8) has an eigenvalue z with $|z| > 1$, then the approximation (13.1.2) is not stable. This is the Godunov–Ryabenkii condition.*

Proof. If z is an eigenvalue, then $v_j^n = z^n\varphi_j$ is a solution of Eq. (13.1.2) with $F \equiv 0$, $g \equiv 0$, $f = \varphi$. At a fixed time t , we have

$$v_j^{t/k} = z^{t/k}\varphi_j,$$

and for decreasing k the solution grows without bound. This proves the lemma.

To derive sufficient conditions for stability, we follow the same lines as for the semidiscrete case. The analysis is very similar. The difference is that we are now considering the domain $|z| \geq 1$ instead of $\operatorname{Re} \tilde{s} \geq 0$, and we want to estimate the solutions of Eq. (13.1.7) in terms of the data \tilde{F}_j and \tilde{g} . The definitions

and the lemmas are completely analogous to the ones given in Section 12.2; therefore, we make the presentation shorter here.

Consider the transformed problem (13.1.7) with $\tilde{F} = 0$ for $|z| > 1$. The general solution of Eq. (13.1.7a) is defined in terms of the solution of the characteristic equation

$$\text{Det} \left(zI - \sum_{\nu=-r}^p B_\nu \kappa^\nu \right) = 0. \quad (13.1.9)$$

We have assumed stability for the problem with periodic boundary conditions. In the same way as for the semidiscrete case, we can prove the following lemma (Exercise 13.1.1).

Lemma 13.1.2. *The characteristic equation (13.1.9) has no solution $\kappa = e^{i\xi}$, ξ real, if $|z| > 1$. If B_{-r} is nonsingular, there are exactly mr solutions κ_ν with $|\kappa_\nu| < 1$ for $|z| > 1$.*

Equation (13.1.7a) can be written in one-step form

$$\mathbf{v}_{j+1} = M\mathbf{v}_j, \quad j = 1, 2, \dots,$$

where the eigenvalues of M are given by the solutions κ of Eq. (13.1.9). By Lemma 13.1.2, we can transform the matrix M as in Eq. (12.2.17a). The Godunov–Rybinkii condition can then be stated for the transformed function $w = U^*v$ as

$$\text{Det}(H^I(z)) \neq 0, \quad |z| > 1, \quad (13.1.10)$$

where H^I is the matrix in the transformed boundary condition

$$H^I \mathbf{w}_1^I + H^{II} \mathbf{w}_1^{II} = \mathbf{g}.$$

The condition (13.1.10) is now strengthened to the *determinant condition*

$$\text{Det}(H^I(z)) \neq 0, \quad |z| \geq 1. \quad (13.1.11)$$

We now have the following lemma.

Lemma 13.1.3. *The determinant condition is fulfilled if, and only if, the solutions of Eq. (13.1.7) with $\tilde{F}_j = 0$ satisfy*

$$\sum_{\nu=-r+1}^p |\tilde{v}_\nu|^2 \leq K|\tilde{g}|^2, \quad |z| > 1, \quad (13.1.12)$$

where K is independent of z (the Kreiss condition).

The determinant condition, or equivalently the Kreiss condition, can be checked without the transformation process above.

The solution of Eq. (13.1.7a) under the condition (13.1.7c) has the form

$$\tilde{v}_j = \sum_{|\kappa_\nu| < 1} P_\nu(j) \kappa_\nu^j, \quad |z| > 1, \quad (13.1.13)$$

where $P_\nu(j)$ is a polynomial in j with vector coefficients. By Lemma 13.1.2, the solution \tilde{v}_j depends on mr parameters $\sigma = (\sigma_1, \dots, \sigma_{mr})^T$, and they are determined from the boundary conditions by a linear system

$$C(z)\sigma = \tilde{g}. \quad (13.1.14)$$

When z approaches the unit circle, some of the roots κ_ν may also approach the unit circle. For $|z_0| = 1$, we, therefore, single out the proper solutions κ_ν by

$$\kappa_\nu(z_0) = \lim_{\delta \rightarrow 0} \kappa_\nu((1 + \delta)z_0), \quad |\kappa_\nu((1 + \delta)z_0)| < 1 \quad (13.1.15)$$

for $\delta > 0$. Generalized eigenvalues z_0 are defined in analogy with the definition in Section 12.2. We have the next lemma.

Lemma 13.1.4. *The Kreiss condition is fulfilled if, and only if, there are no eigenvalues or generalized eigenvalues z , $|z| \geq 1$, to the problem (13.1.8), that is, if $\text{Det}(C(z)) \neq 0$, $|z| \geq 1$.*

As we have seen, the Kreiss condition can be formulated in several ways in the transformed space. By going back to the physical space via the relation $\hat{v}_j(s) = \tilde{v}_j(z)$ and the inverse Laplace transform, we obtain, the following theorem corresponding to Theorem 12.2.1.

Theorem 13.1.1. *The solutions of Eq. (13.1.2) with $f = F = 0$ satisfy, for any fixed j , the estimate*

$$\sum_{\nu=1}^n |v_j^\nu|^2 k \leq \text{constant} \sum_{\nu=1}^n |g^\nu|^2 k, \quad (13.1.16)$$

if, and only if, the Kreiss condition is satisfied.

REMARK. Parseval's relation yields the estimate for $n = \infty$. However, we use the same arguments as above to obtain the estimate for any finite n . (The solution v^n does not depend on g^ν , $\nu > n$.)

There is no need to check the Kreiss condition as $|z| \rightarrow \infty$.

Lemma 13.1.5. *There are constants c_0 and K_0 such that, for $|z| \geq c_0$, the solutions of Eq. (13.1.7) with $\tilde{F} \equiv 0$ satisfy the estimate*

$$\|\tilde{v}\|_h \leq K_0 h^{1/2} |\tilde{g}|. \quad (13.1.17)$$

Proof. We have

$$\|\tilde{v}\|_h^2 = \frac{1}{|z|^2} \|Q\tilde{v}\|_h^2 \leq \frac{\text{constant}}{|z|^2} \left(\|\tilde{v}\|_h^2 + \sum_{\mu=-r+1}^0 |\tilde{g}_\mu|^2 h \right);$$

that is, for large enough $|z|$,

$$\|\tilde{v}\|_h^2 \leq \text{constant} |\tilde{g}|^2 h,$$

and the lemma follows.

Now we derive the stability estimate for the original problem under the assumption that there is an energy estimate for the Cauchy problem. To introduce a slightly different technique than the one used for the semidiscrete case, we assume that Eq. (13.1.2) is a scalar problem. We need the following lemma.

Lemma 13.1.6. *Let Q be any difference operator of the form (13.1.3) where the B_ν are scalars. The scalar problem*

$$\begin{aligned} v_j^{n+1} &= Qv_j^n, \quad j = 1, 2, \dots, \\ v_j^0 &= 0, \\ v_\mu^{n+1} &= g_\mu^{n+1}, \quad \mu = -r + 1, -r + 2, \dots, 0, \end{aligned} \quad (13.1.18)$$

satisfies the Kreiss condition.

The proof is omitted here.

The lemma says that, from a stability point of view, we can always specify data explicitly at all boundary points regardless of inflow or outflow. (However, it may be difficult to find accurate data in the outflow case.)

We make the assumption that the Cauchy problem satisfies an energy estimate:

Assumption 13.1.1. *The solutions of*

$$v_j^{n+1} = Qv_j^n, \quad j = 0, \pm 1, \pm 2, \dots,$$

satisfy the estimate

$$\|v^{n+1}\|_{-\infty, \infty} \leq \|v^n\|_{-\infty, \infty}, \quad (13.1.19)$$

for $\lambda = k/h \leq \lambda_0$, $\lambda_0 > 0$.

We shall prove the following lemma.

Lemma 13.1.7. *There exist boundary conditions such that the solution of the scalar problem*

$$v_j^{n+1} = Qv_j^n, \quad j = 1, 2, \dots, \quad (13.1.20a)$$

$$v_j^0 = f_j, \quad j = -r + 1, -r + 2, \dots, \quad (13.1.20b)$$

$$L_1 v_0^{n+1} = 0, \quad (13.1.20c)$$

satisfies

$$\sum_{\nu=1}^n |v_j^\nu|^2 k \leq \text{constant} \|f\|_h^2, \quad j = -r + 1, -r + 2, \dots, 1. \quad (13.1.21)$$

Proof. Let w^n be defined by

$$w_j^n = \begin{cases} v_j^n, & j = -r + 1, -r + 2, \dots, 0, \\ 0, & \text{otherwise,} \end{cases} \quad (13.1.22)$$

the projection operator P for gridfunctions $\{v_j^n\}_{j=-\infty}^\infty$ by

$$(Pv^n)_j = \begin{cases} v_j^n, & j = 1, 2, \dots, \\ 0, & j \leq 0, \end{cases} \quad (13.1.23)$$

and the injection operator R for gridfunctions $\{v_j^n\}_{j=1}^\infty$ by

$$(Rv^n)_j = \begin{cases} v_j^n, & j = 1, 2, \dots, \\ 0, & j \leq 0. \end{cases} \quad (13.1.24)$$

We have

$$\begin{aligned}\|v^{n+1}\|_h^2 &= \|Qv^n\|_h^2 = \|PQ(Rv^n + w^n)\|_{-\infty, \infty}^2, \\ &= \|PQRv^n\|_{-\infty, \infty}^2 + 2\operatorname{Re}(PQRv^n, PQw^n)_{-\infty, \infty} + \|PQw^n\|_{-\infty, \infty}^2.\end{aligned}$$

By assumption,

$$\|PQRv^n\|_{-\infty, \infty}^2 \leq \|QRv^n\|_{-\infty, \infty}^2 \leq \|Rv^n\|_{-\infty, \infty}^2 = \|v^n\|_h^2.$$

Furthermore,

$$v_j^{n+1} = (QRv^n)_j + (Qw^n)_j, \quad j = 1, 2, \dots, r,$$

which gives

$$\begin{aligned}\|v^{n+1}\|_h^2 &\leq \|v^n\|_h^2 + 2\operatorname{Re}(QRv^n, Qw^n)_{1,r} + \|Qw^n\|_{1,r}^2, \\ &= \|v^n\|_h^2 + 2\operatorname{Re}(v^{n+1}, Qw^n)_{1,r}.\end{aligned}$$

By choosing the boundary conditions as

$$\begin{aligned}v_\mu^n &= 0, \quad \mu = -r + 2, -r + 3, \dots, 0, \\ v_{-r+1}^n &= -B_{-r}^{-1}v_1^{n+1},\end{aligned}\tag{13.1.25}$$

we get

$$\begin{aligned}\operatorname{Re}(v^{n+1}, Qw^n)_{1,r} &= -|v_1^{n+1}|^2 h = -\sum_{j=-r+2}^1 |v_j^{n+1}|^2 h, \\ &\leq -ch \left(|v_{-r+1}^n|^2 + \sum_{j=-r+2}^1 |v_j^{n+1}|^2 \right), \quad c > 0.\end{aligned}$$

Thus,

$$\begin{aligned}\|v^{n+1}\|_h^2 &\leq \|v^n\|_h^2 - 2ch \left(|v_{-r+1}^n|^2 + \sum_{j=-r+2}^{-1} |v_j^{n+1}|^2 \right), \\ &\leq \|v^0\|_h^2 - 2ch \left(\sum_{\nu=0}^n |v_{-r+1}^\nu|^2 + \sum_{\nu=1}^{n+1} \sum_{j=-r+2}^{-1} |v_j^\nu|^2 \right),\end{aligned}$$

showing that Eq. (13.1.21) is satisfied.

The lemma gives an estimate for the boundary values including v_1^n . As for the semidiscrete case, we also need, in general, estimates for $v_j^n, j \geq 2$. Therefore, we have the following lemma.

Lemma 13.1.8. *Let L_1 be the boundary operator constructed in Lemma 13.1.7. Then, for any fixed j , the solution of Eq. (13.1.20) satisfies*

$$\sum_{\nu=1}^n |v_j^\nu|^2 k \leq \text{constant} \|f\|_h^2, \quad j = -r + 1, -r + 2, \dots \quad (13.1.26)$$

Proof. By Lemma 13.1.7, we already have estimates for $v_j^n, j = -r + 1, -r + 2, \dots, 1$. In particular, we consider $v_{-r+2}^n, v_{-r+3}^n, \dots, v_1^n$ as given boundary values and write the difference approximation as

$$\begin{aligned}v_j^{n+1} &= Qv_j^n, \quad j = 2, 3, \dots, \\ v_j^0 &= f_j, \quad j = -r + 2, -r + 3, \dots, \\ v_\mu^{n+1} &= v_\mu^{n+1}, \quad \mu = -r + 2, -r + 3, \dots, 1.\end{aligned} \quad (13.1.27)$$

Now, we consider the auxiliary problem with the special boundary operator shifted one step to the right:

$$\begin{aligned}w_j^{n+1} &= Qw_j^n, \quad j = 2, 3, \dots, \\ w_j^0 &= f_j, \quad j = -r + 2, -r + 3, \dots, \\ L_1 w_1^{n+1} &= 0.\end{aligned} \quad (13.1.28)$$

Lemma 13.1.7 implies

$$\sum_{\nu=1}^n |w_\mu^\nu|^2 k \leq \text{constant} \|f\|_h^2, \quad \mu = -r + 2, -r + 3, \dots, 2.$$

The difference $y_j = v_j - w_j$ satisfies

$$\begin{aligned} y_j^{n+1} &= Qy_j^n, & j &= 2, 3, \dots, \\ y_j^0 &= 0, & j &= -r + 2, -r + 3, \dots, \\ y_\mu^{n+1} &= v_\mu^{n+1} - w_\mu^{n+1}, & \mu &= -r + 2, -r + 3, \dots, 1. \end{aligned} \quad (13.1.29)$$

The z -transformed system is

$$\begin{aligned} z\tilde{y}_j &= Q\tilde{y}_j, & j &= 2, 3, \dots, \\ \tilde{y}_\mu &= \tilde{v}_\mu - \tilde{w}_\mu, & \mu &= -r + 2, -r + 3, \dots, 1. \end{aligned} \quad (13.1.30)$$

By Lemma 13.1.6, the solution satisfies the Kreiss condition, and, by Theorem 13.1.1, we get, for any fixed j ,

$$\begin{aligned} \sum_{\nu=1}^n |y_j^\nu|^2 k &\leq \text{constant} \sum_{\mu=-r+2}^1 \sum_{\nu=1}^n (|v_\mu^\nu|^2 + |w_\mu^\nu|^2)k, \\ &\leq \text{constant} \|f\|_h^2, \quad j = -r + 2, -r + 3, \dots \end{aligned}$$

Thus, for $j = 2$,

$$\sum_{\nu=1}^n |v_2^\nu|^2 k \leq \text{constant} \sum_{\nu=1}^n (|w_2| + |y_2|)^2 k \leq \text{constant} \|f\|_h^2. \quad (13.1.31)$$

Now the same procedure is applied for $j = 3, 4, \dots$, and the lemma follows.

The stability estimate for the original problem (13.1.2) is obtained as it was for the semidiscrete problem. The solution w of the auxiliary problem (13.1.20) with the special boundary conditions (and with the forcing function F included) is subtracted from the solution v of the original problem. The difference $v - w$ satisfies a problem with zero forcing function, zero initial function but nonzero boundary data containing w_μ and g_μ . Because there are estimates for w_μ , we get the final estimate

$$\|v^n\|_h^2 \leq Ke^{\alpha t_n} \left(\|f\|_h^2 + \sum_{\nu=1}^n (\|F^{\nu-1}\|_h^2 + |g^\nu|^2)k \right), \quad (13.1.32)$$

that is, strong stability.

The restriction to scalar problems is introduced because an analog of Lemma 13.1.6 is not known for systems. In the system case, the result can, of course, still be used if the matrices B_j can be simultaneously diagonalized. This is no severe restriction in the one-dimensional case, but in several space dimensions it usually is. In such a case, the techniques presented for semidiscrete problems (Lemma 12.2.9 and 12.2.10) can be used.

We summarize the results in the following theorem.

Theorem 13.1.2. *Assume that the approximation (13.1.2) fulfills the Kreiss condition and that there is an energy estimate for the corresponding Cauchy problem. If one of the following conditions is satisfied, then the approximation is strongly stable.*

1. *The coefficient matrices can be simultaneously diagonalized.*
2. $r \geq p$.

The principle of deriving stability estimates via an auxiliary problem with special boundary conditions can also be applied to multistep schemes. However, the construction of the special boundary conditions is more complicated and, at the current state, we need some extra assumptions. There is a general theory where the stability follows from the Kreiss condition by itself; this will be further commented on in the Bibliographic Notes at the end of this chapter. Because the Kreiss condition plays the central role whatever approach we choose, we extend our discussion of this condition to general multistep schemes and treat a few examples.

We consider general multistep methods of the type introduced in Section 5.1. The general form of the approximation is

$$\begin{aligned} Q_{-1}v_j^{n+1} &= \sum_{\sigma=0}^q Q_\sigma v_j^{n-\sigma} + kF_j^n, \quad j = 1, 2, \dots, \\ v_j^\sigma &= f_j^{(\sigma)}, \quad j = -r + 1, -r + 2, \dots; \quad \sigma = 0, 1, \dots, q, \\ L_0 v^{n+1} &= g^{n+1}. \end{aligned} \tag{13.1.33}$$

Again we assume that the approximation is solvable, that is, that a unique solution v_j^{n+1} exists. By defining $v_j(t)$ between gridpoints as above, we can use the Laplace transform, and, by substituting $z = e^{sk}$, we obtain the z -transformed equations. Under the assumption that $f_j^{(\sigma)} \equiv 0$ and $\sigma = 0, 1, \dots, q$, these equations are formally obtained by the substitution $v_j^n = z^n \tilde{v}_j$ and, similarly, for F_j^n and g^n . We obtain

$$\begin{aligned} zQ_{-1}\tilde{v}_j &= \sum_{\sigma=0}^q z^{-\sigma} Q_\sigma \tilde{v}_j + k\tilde{F}_j, \quad j = 1, 2, \dots, \\ \tilde{L}_0 \tilde{v}_j &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty. \end{aligned} \tag{13.1.34}$$

The only difference when compared to the procedure for one-step schemes is that we now have higher order polynomials in z involved in the difference equation and possibly in the boundary conditions. The corresponding eigenvalue problem is

$$zQ_{-1}\varphi_j = \sum_{\sigma=0}^q z^{-\sigma} Q_\sigma \varphi_j, \quad j = 1, 2, \dots, \tag{13.1.35a}$$

$$\tilde{L}_0 \varphi_j = 0, \tag{13.1.35b}$$

$$\|\varphi\|_h < \infty. \tag{13.1.35c}$$

Obviously, Lemma 13.1.1, formulated for one-step schemes, also holds in this case.

Again, the solution of the ordinary difference equation (13.1.35a) is expressed in terms of the roots κ , of the characteristic equation. If the difference operators Q_σ have the form

$$Q_\sigma = \sum_{\nu=-r}^p B_\nu^{(\sigma)} E^\nu, \tag{13.1.36}$$

then the characteristic equation is

$$\text{Det} \left(\sum_{\nu=-r}^p \left(zB_\nu^{(-1)} - \sum_{\sigma=0}^q z^{-\sigma} B_\nu^{(\sigma)} \right) \kappa^\nu \right) = 0. \tag{13.1.37}$$

This equation is obtained by formally substituting $v_j^n = \kappa^j z^n$ into the original homogeneous scheme and setting the determinant of the resulting matrix equal to zero.

The Kreiss condition (13.1.12) is also well-defined for Eq. (13.1.34) with $\tilde{F} \equiv 0$, as well as the eigenvalues and generalized eigenvalues of Eq. (13.1.35), and Lemma 13.1.4 holds. Again, the condition is checked by writing the boundary conditions as in Eq. (13.1.14) and checking that $\text{Det}(C(z)) \neq 0$ for $|z| \geq 1$.

We now consider the leap-frog scheme,

$$v_j^{n+1} - v_j^{n-1} = \lambda(v_{j+1}^n - v_{j-1}^n), \quad j = 1, 2, \dots, \quad \lambda = \frac{k}{h}, \quad (13.1.38a)$$

as an example. From Section 12.1 and 12.2, we know that the boundary condition

$$(hD_+)^q v_0^{n+1} = 0 \quad (13.1.38b)$$

is stable for the semidiscrete approximation for any integer q . The Kreiss condition is satisfied and, furthermore, for the special cases $q = 1, 2$, it has been demonstrated in Section 11.1 that the energy method can be applied directly.

To check the Kreiss condition for the leap-frog scheme, we solve the characteristic equation

$$(z^2 - 1)\kappa = \lambda z(\kappa^2 - 1). \quad (13.1.39)$$

The transformed equation

$$(z - z^{-1})\tilde{v}_j = \lambda(\tilde{v}_{j+1} - \tilde{v}_{j-1})$$

has the solution

$$\tilde{v}_j = \sigma_1 \kappa_1^j, \quad |\kappa_1| < 1, \quad \text{for } |z| > 1.$$

The determinant condition is

$$(\kappa_1 - 1)^q \neq 0, \quad |z| \geq 1, \quad (13.1.40)$$

which is only violated if $\kappa_1 = 1$. [The matrices $H^I(z)$ and $C(z)$ are scalar and, therefore, identical in this case.] The corresponding z values ± 1 are obtained from Eq. (13.1.39). To define κ_1 properly in the neighborhood of $z = -1$, we let $z = -(1 + \delta)$, $\delta > 0$, δ small, and we get, from Eq. (13.1.39),

$$2\delta\kappa \approx -\lambda(1 + \delta)(\kappa^2 - 1),$$

or

$$\kappa \approx -\frac{\delta}{\lambda} \pm \sqrt{1 + \frac{\delta^2}{\lambda^2}} \approx \pm 1 - \frac{\delta}{\lambda}.$$

Because $\delta > 0$, $\lambda > 0$, we have

$$\kappa_1 \approx 1 - \delta/\lambda,$$

and obviously the Kreiss condition is violated at $z = -1$.

The solution of Eq. (13.1.38) is shown as a function of t in Figures 13.1.2 and 13.1.3, for $q = 1$ and $q = 2$, respectively. The condition $v_N = 0$ has been introduced at the boundary $x_N = 1$. As initial conditions, we have used

$$\begin{aligned} v_j^0 &= -\sin(2\pi x_j), \\ v_j^1 &= v_j^0 + kD_0 v_j^0. \end{aligned}$$

It is clearly seen how the instability originates at the outflow boundary. Also note that the oscillations are more severe in the case $q = 2$. Formally, the boundary condition is more accurate than the one for $q = 1$. The experiment is an illustration of the basic fact that the stability concept is independent of the order of accuracy. The stronger oscillations obtained for $q = 2$ could actually be expected from the analysis. The singularity at $z = -1$ of the matrix $H^I(z) = (\kappa_1 - 1)^q$ becomes stronger with increasing q . This leads to a worse estimate of the solution of the z -transformed system

$$\begin{aligned} (z^2 - 1)\tilde{v}_j &= \lambda z(\tilde{v}_{j+1} - \tilde{v}_{j-1}), \quad j = 1, 2, \dots, \\ (hD_+)^q \tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty, \end{aligned} \tag{13.1.41}$$

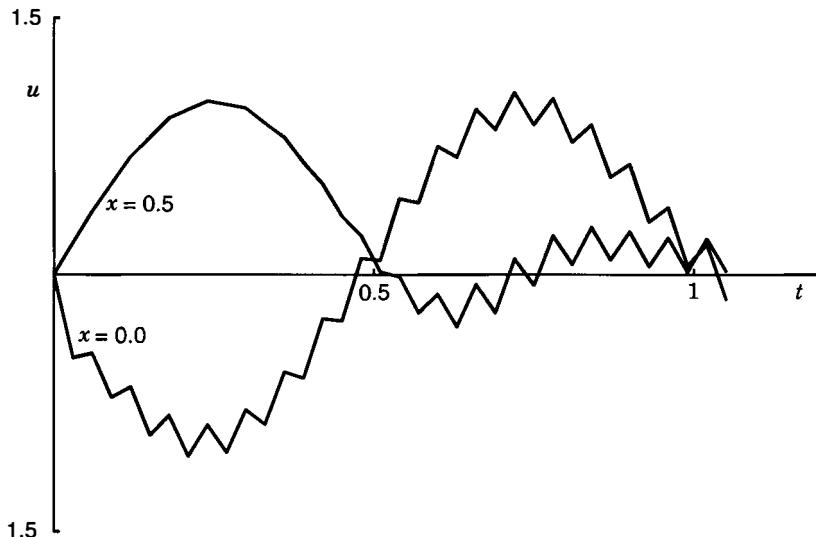


Figure 13.1.2.

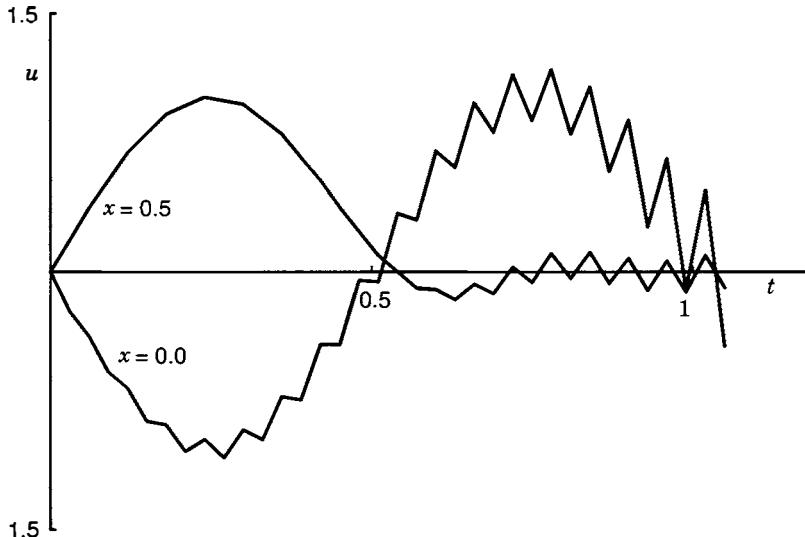


Figure 13.1.3.

because

$$|\tilde{v}_0| = \frac{|\tilde{g}|}{|\kappa_1 - 1|^q} \approx \frac{\lambda^q |\tilde{g}|}{(|z| - 1)^q}.$$

Obviously, we need another boundary condition. It does not help to replace D_0 by D_+ at the boundary, because this is equivalent to Eq. (13.1.38b) for $q = 2$ (with the grid shifted one step). However, if a noncentered time differencing is used, then we get the new condition

$$v_0^{n+1} - v_0^n = \lambda(v_1^n - v_0^n). \quad (13.1.42)$$

The determinant condition is now violated if, and only if,

$$z - 1 - \lambda(\kappa_1 - 1) = 0. \quad (13.1.43)$$

The system (13.1.39) and (13.1.43) only has the solution $z = \kappa_1 = 1$. But a perturbation calculation with $z = 1 + \delta$, $\delta > 0$, shows that $\kappa_1 = -1$ at $z = 1$. Hence, the Kreiss condition holds.

The numerical result with this boundary condition is shown in Figure 13.1.4. The oscillations have now disappeared. The accuracy is still not very good, this is because the true solution has a discontinuity in the derivative because of an incompatibility in the data at the corner $x = 1$, $t = 0$.

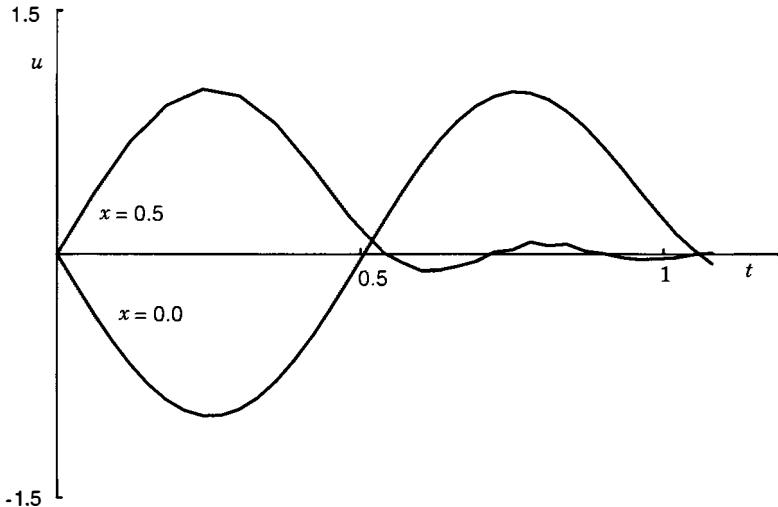


Figure 13.1.4.

The procedure we have used for this example in order to check the Kreiss condition is typical. If $\text{Det}(C(z)) = 0$ for some value $z = z_0$ with $|z_0| > 1$, the analysis is complete, and we know that the approximation is unstable. If $C(z)$ is singular for $z = z_0$ with $|z_0| = 1$, the Kreiss condition may or may not be violated. If all the corresponding roots κ satisfy $|\kappa(z_0)| < 1$, then z_0 is an eigenvalue, and the Kreiss condition is violated. If there is at least one root κ_1 with $|\kappa_1(z_0)| = 1$, we must find out whether or not z_0 is a generalized eigenvalue. This is done by a perturbation calculation $z_0 \rightarrow (1 + \delta)z_0$, $\delta > 0$. If $|\kappa_1((1 + \delta)z_0)| < 1$, then there is a generalized eigenvalue, otherwise not. (The procedure described here is used as the basis for the IBSTAB-algorithm described in Section 13.3.)

Next we treat another example (cf. Exercise 12.5.3). Consider the problem

$$\begin{aligned} \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}_t &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}_x, & 0 \leq x \leq 1, \quad t \geq 0, \\ u(x, 0) &= f(x), \\ u^{(1)}(0, t) &= u^{(1)}(1, t) = 0, \end{aligned} \tag{13.1.44}$$

and define the difference operator Q by $w_j = Qv_j$

$$\begin{aligned} w_j &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} D_0 v_j, & j = 1, 2, \dots, N-1, \\ w_0^{(2)} &= D_+ v_0^{(1)}, & j = 0, \\ w_N^{(2)} &= D_- v_N^{(1)}, & j = N. \end{aligned} \tag{13.1.45}$$

(Note that the vector grid function v has $2N + 2$ components, but Qv has only $2N$ components.) The Crank–Nicholson approximation is

$$\begin{aligned} v_j^{n+1} - v_j^n &= \frac{k}{2} Q(v_j^{n+1} + v_j^n), \quad j = 0, 1, \dots, N, \\ v_0^{(1)n+1} - v_N^{(1)n+1} &= 0. \end{aligned} \tag{13.1.46}$$

With the scalar product defined by

$$(v, w)_h = \frac{h}{2} (v_0^{(2)} w_0^{(2)} + v_N^{(2)} w_N^{(2)}) + \sum_{j=1}^{N-1} \langle v_j, w_j \rangle h,$$

we have, with $w_j = v_j^{n+1} + v_j^n$,

$$\begin{aligned} (w, Qw)_h &= \frac{1}{2} w_0^{(2)} (w_1^{(1)} - w_0^{(1)}) + \frac{1}{2} w_N^{(2)} (w_N^{(1)} - w_{N-1}^{(1)}) \\ &\quad + \frac{1}{2} \sum_{j=1}^{N-1} (w_j^{(1)} (w_{j+1}^{(2)} - w_{j-1}^{(2)}) + w_j^{(2)} (w_{j+1}^{(1)} - w_{j-1}^{(1)})) = 0. \end{aligned}$$

By taking the scalar product of Eq. (13.1.46) with $v^{n+1} + v^n$, this gives the energy estimate

$$\|v^{n+1}\|_h = \|v^n\|_h,$$

showing that the method is stable, but not necessarily strongly stable.

Next we check the Kreiss condition for the right quarter-space problem. The characteristic equation for the approximation at inner points is

$$\text{Det} \begin{bmatrix} (z - 1)\kappa & -\frac{\lambda}{4}(z + 1)(\kappa^2 - 1) \\ -\frac{\lambda}{4}(z + 1)(\kappa^2 - 1) & (z - 1)\kappa \end{bmatrix} = 0,$$

or, equivalently,

$$\left(\frac{4(z-1)}{\lambda(z+1)} \right)^2 \kappa^2 - (\kappa^2 - 1)^2 = 0. \tag{13.1.47}$$

The solution of the z -transformed equation is

$$\tilde{v}_j = \sigma_1 \kappa_1^j \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \sigma_2 \kappa_2^j \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

where κ_1 and κ_2 are the solutions of Eq. (13.1.47) with $|\kappa_1| < 1$ and $|\kappa_2| < 1$ for $|z| > 1$. The condition $\tilde{v}_0^{(1)} = 0$ implies $\sigma_1 = -\sigma_2$, and the condition

$$(z - 1)\tilde{v}_0^{(2)} = \frac{\lambda}{2} (z + 1)(\tilde{v}_1^{(1)} - \tilde{v}_0^{(1)})$$

implies

$$\sigma_1 \left(\frac{2}{\lambda} \frac{z-1}{z+1} \cdot 2 - (\kappa_1 - \kappa_2) \right) = 0. \quad (13.1.48)$$

Now consider the point

$$z_0 = -\frac{1 + \lambda i/2}{1 - \lambda i/2}$$

on the unit circle. The corresponding κ values, $\kappa_1 = i$, $\kappa_2 = -i$, are obtained from Eq. (13.1.47), and Eq. (13.1.48) is, obviously, satisfied for $z = z_0$. Thus, there is a generalized eigenvalue z_0 . However, both κ values are double roots of the polynomial in Eq. (13.1.47), and we have a situation analogous to the one in the semidiscrete example in Section 12.5. The Kreiss condition is violated, showing that the scheme is not strongly stable. However, the generalized eigensolution contains double roots κ that behave in such a way that stability in the generalized sense (as defined in the next section) still holds. This is consistent with the energy estimate derived above, because one can prove that it leads to generalized stability (cf. Theorem 10.3.1 for the continuous case).

We now demonstrate how the stability theory can be applied to obtain general principles for designing stable boundary conditions. The Kreiss condition may sometimes become rather complicated to check for systems of difference equations. However, for scalar equations, many results are known and we will show how they can be applied to systems. We consider general hyperbolic systems

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x}, \quad 0 \leq x < \infty, \quad t \geq 0, \quad (13.1.49)$$

and general multistep schemes (13.1.33) and (13.1.36) where the coefficient matrices $B_\nu^{(\sigma)}$ are assumed to be independent of h ; that is, we only consider the

principle part. We also make the very natural assumption that the matrices $B_\nu^{(\sigma)}$ are polynomials in A . Under this assumption, both the differential equation and its approximation can be reduced to a set of scalar equations. We assume that this is already done and that A and $B_j^{(\sigma)}$ are diagonal with u and A partitioned such that

$$u = \begin{bmatrix} u^I \\ u^{II} \end{bmatrix}, \quad A = \begin{bmatrix} A^I & 0 \\ 0 & A^{II} \end{bmatrix}, \quad A^I > 0, \quad A^{II} < 0.$$

Well-posedness requires the form

$$u^{II}(0, t) = R u^I(0, t), \quad (13.1.50)$$

for the boundary conditions (for convenience, only homogeneous conditions are considered). It is now possible to work with the diagonal form of the approximation to construct stable boundary conditions and then implement them for the original system.

Assume that the boundary conditions for each component of the outflow vector v^I are such that the Kreiss condition is fulfilled for the scalar problem. In the z -transformed space, for any fixed j , there is an estimate

$$|\tilde{v}_j^I| \leq \text{constant } |\tilde{g}^I|, \quad |z| > 1, \quad (13.1.51)$$

where \tilde{g} is the inhomogeneous term in the boundary conditions. Therefore, the boundary values \tilde{v}_j^I can be regarded as inhomogeneous terms in the transformed boundary conditions for the inflow vector v^{II} . We require that these conditions have the form

$$\tilde{v}_\mu^{II} = S \tilde{v}_0^I + \tilde{g}_\mu^{II}, \quad \mu = -r + 1, -r + 2, \dots, 0, \quad (13.1.52)$$

where S is a difference operator independent of z . Lemma 13.1.6 is now applied to the inflow part of the system, and strong stability follows.

It now remains to construct S in Eq. (13.1.52) so that sufficient accuracy is obtained. First, we note that if $r = 1$, then, obviously, we can use the physical boundary condition (13.1.50) as the inflow condition. If $r \geq 2$, some extra conditions are required. Using Taylor expansions we get, from the differential equations and the boundary conditions,

$$\begin{aligned}
u^{II}(\delta, t) &= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} \frac{\partial^\nu u^{II}}{\partial x^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}), \\
&= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} (A^{II})^{-\nu} \frac{\partial^\nu u^{II}}{\partial t^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}), \\
&= \sum_{\nu=0}^{\tau} \frac{\delta^\nu}{\nu!} (A^{II})^{-\nu} R(A^I)^\nu \frac{\partial^\nu u^I}{\partial x^\nu}(0, t) + \mathcal{O}(\delta^{\tau+1}). \quad (13.1.53)
\end{aligned}$$

By using difference formulas for $\partial^\nu u^I / \partial x^\nu$, and applying Eq. (13.1.53) at $\delta = -h, -2h, \dots, -(r-1)h$ we obtain a set of boundary conditions that has the required form.

In applications, the matrix A is sometimes singular, but the method described here can still be used. The vanishing eigenvalues are included in A^I and Eq. (13.1.50) are still well-posed conditions. Because the inverse of A^I is never used, the procedure above is still well-defined.

For more general problems including several space dimensions, variable coefficients, and possibly nonlinear equations, the procedure is, of course, more complicated. But it is straightforward and may still be a realistic approach.

Next we turn to the outflow problem and consider approximations of the scalar equation

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x}, \quad a > 0. \quad (13.1.54)$$

Several scalar examples have been treated in the previous sections, and, with some satisfaction, we note that those results can now be used for the full system as indicated by the arguments above. For so called *translatory* boundary conditions of the form

$$\sum_{\sigma=-1}^q \sum_{j=0}^{j_B} c_{j\sigma} E^j v_\mu^{n-\sigma} = g_\mu^{n+1}, \quad c_{0(-1)} \neq 0, \quad \mu = -r+1, -r+2, \dots, 0, \quad (13.1.55)$$

where the same form of conditions are used at all boundary points, quite general results can be obtained. The *boundary characteristic function* is defined by

$$R(z, \kappa) = \sum_{j=0}^{j_B} \sum_{\sigma=-1}^q z^{-(\sigma+1)} c_{j\sigma} \kappa^j.$$

If the conditions (13.1.55) are considered as a scheme to be applied at all points,

then the equation

$$R(z, e^{i\xi}) = 0 \quad (13.1.56)$$

is the characteristic equation used to verify the von Neumann condition $|z| \leq 1$. In that case, it is assumed that more than one time level is involved; but for our purposes $R(z, \kappa)$ is well defined anyway. Similarly, the characteristic function for the basic scheme is

$$P(z, h) = \sum_{\nu=-r}^p \left(B_\nu^{-1} - \sum_{\sigma=0}^q z^{-(\sigma+1)} B_\nu^{(\sigma)} \right) \kappa^\nu = 0, \quad (13.1.57)$$

where $B_\nu^{(\sigma)}$ are scalars. Letting

$$\Omega(z, \kappa) = |P(z, \kappa)| + |R(z, \kappa)|$$

one can prove the following result:

Theorem 13.1.3. *Assume that the approximation (13.1.33) is consistent with the scalar outflow problem (13.1.54) and has translatory boundary conditions (13.1.55). Then the Kreiss condition is fulfilled if the condition*

$$\Omega(z, \kappa) > 0$$

for

$$(z, \kappa) \in \{ |z| = |\kappa| = 1, (z, \kappa) \neq (1, 1), (z, \kappa) \neq (-1, -1) \} \cup \{ 1 \leq |z|, 0 < |\kappa| < 1 \} \quad (13.1.58a)$$

and one of the conditions

$$\Omega(-1, -1) > 0, \quad (13.1.58b)$$

or

$$((\partial P / \partial z)(\partial P / \partial \kappa))_{z=\kappa=-1} < 0 \quad (13.1.58c)$$

are satisfied.

The simplest outflow boundary procedure is extrapolation of some order ν

$$(hD_+)^{\nu} v_{\mu}^{n+1} = 0, \quad \mu = -r + 1, -r + 2, \dots, 0, \quad (13.1.59)$$

which was used for the fourth-order approximation in Section 13.1 [see Eq. (12.1.25)]. These conditions are translatory, and we have

$$R(z, \kappa) = (\kappa - 1)^{\nu}.$$

Now assume that we use a dissipative one-step scheme. The condition (13.1.58a) is, obviously, fulfilled for any ν except, possibly, for $\kappa = 1$, $|z| = 1$, $z \neq 1$. But consistency implies that $v_j^n = \text{constant}$ gives the same constant solution v_j^{n+1} at the next time level, and because we have a one-step scheme, $P(z, 1) = 0$ implies $z = 1$. The condition (13.1.58b) is also satisfied. Thus, the Kreiss condition is satisfied.

It should be pointed out that the conditions in Theorem 13.1.3 are sufficient but, in general, not necessary. For example, to use the theorem for verifying stability for the Crank–Nicholson scheme with the one-sided forward Euler scheme at the boundary, the von Neumann condition for the boundary scheme must be satisfied. This leads to the condition $k|a| \leq h$, whereas a direct stability analysis shows that the correct condition is $k|a| < 2h$.

Finally, we note that the error estimates are obtained as in the semidiscrete case. A smooth solution of the continuous problem is substituted into the difference approximation and the truncation error enters as inhomogeneous terms F^n , g^n , and f . The error estimate then follows directly from the strong stability estimate.

EXERCISES

13.1.1. Prove Lemma 13.1.2.

13.1.2. Prove Lemma 13.1.6 for the special case

$$Q = I + \lambda D_0 + \frac{\lambda^2}{2} D_+ D_- - \alpha h^4 (D_+ D_-)^2, \quad \alpha > 0.$$

13.1.3. Formulate and prove the analogy of Lemma 12.2.10 for explicit one-step methods. Use the result to prove Theorem 13.1.2 for the case $r \geq p$.

13.1.4. Formulate and prove Lemma 13.1.2 for multistep methods.

13.1.5. Prove that the leap-frog scheme (13.1.38a) fulfills the Kreiss condition with the boundary condition

$$u_0^{n+1} = 2u_1^n - u_2^{n-1}. \quad (13.1.60)$$

13.1.6. Derive an error estimate for the Lax–Wendroff approximation of

$$u_t = Au_x + F, \quad A = \begin{bmatrix} A^I & 0 \\ 0 & A^{II} \end{bmatrix}, \quad A^I > 0, \quad A^{II} < 0,$$

$$u^{II}(0, t) = g(t),$$

with boundary conditions

$$(v_0^{II})^n = g^n,$$

$$(hD_+)^q v_0^I = 0.$$

- 13.1.7.** Use the general principle described at the end of Section 13.1 to derive stable boundary conditions of the form of Eq. (13.1.52) for the linearized Euler equations (4.6.5) for a fourth-order approximation ($r = 2$).

13.2. THE METHOD OF LINES AND GENERALIZED STABILITY

The concept of generalized stability for semidiscrete approximations in Section 12.4 can be generalized to fully discrete approximations. First, consider one-step schemes

$$v_j^{n+1} = Qv_j^n + kF_j^n, \quad j = 1, 2, \dots, \quad (13.2.1a)$$

$$v_j^0 = 0, \quad (13.2.1b)$$

$$L_0 v_0^{n+1} = g^{n+1}, \quad (13.2.1c)$$

where we have assumed zero initial data.

Definition 13.2.1. *The approximation (13.2.1) is stable in the generalized sense if, for $g^n = 0$, the solutions satisfy an estimate*

$$\sum_{n=1}^{\infty} e^{-2\eta t_n} \|v^n\|_h^2 k \leq K(\eta) \sum_{n=1}^{\infty} e^{-2\eta t_n} \|F^{n-1}\|_h^2 k,$$

for $\eta > \eta_0$,

$$\lim_{\eta \rightarrow \infty} K(\eta) = 0. \quad (13.2.2)$$

The approximation (13.2.1) is strongly stable in the generalized sense if the

solutions satisfy

$$\sum_{n=1}^{\infty} e^{-2\eta t_n} \|v^n\|_h^2 k \leq K(\eta) \sum_{n=1}^{\infty} e^{-2\eta t_n} (\|F^{n-1}\|_h^2 + |g^n|^2) k. \quad (13.2.3)$$

As in the previous section, we define the solution for all t by extending v^n , F^n , and g^n to stepfunctions. Then we can take the Laplace transform and, by substituting $z = e^{sk}$, we obtain the z -transformed system

$$\begin{aligned} z\tilde{v}_j &= Q\tilde{v}_j + k\tilde{F}_j, \quad j = 1, 2, \dots, \\ \tilde{L}_0\tilde{v}_0 &= \tilde{g}, \\ \|\tilde{v}\|_h &< \infty. \end{aligned} \quad (13.2.4)$$

By Parseval's relation, we have the following theorem.

Theorem 13.2.1. *The approximation (13.2.1) is stable in the generalized sense if, and only if, Eq. (13.2.4) with $\tilde{g} = 0$ has a unique solution satisfying*

$$\|\tilde{v}(z)\|_h^2 \leq K(z)\|\tilde{F}(z)\|_h^2, \quad \eta > \eta_0, \quad \lim_{\eta \rightarrow \infty} K(z) = 0, \quad (13.2.5)$$

where $z = e^{(\eta + i\xi)k}$. It is strongly stable in the generalized sense if, and only if, Eq. (13.2.4) has a unique solution satisfying

$$\|\tilde{v}(z)\|_h^2 \leq K(z)(\|\tilde{F}(z)\|_h^2 + |\tilde{g}(z)|^2), \quad \eta > \eta_0, \quad \lim_{\eta \rightarrow \infty} K(z) = 0. \quad (13.2.6)$$

One can prove the following theorem, which corresponds to Theorem 12.4.4.

Theorem 13.2.2. *Assume that Eq. (13.2.1a) is dissipative and consistent with a strictly hyperbolic system. Then the approximation (13.2.1) is strongly stable in the generalized sense if the Kreiss condition is satisfied.*

The proof of this theorem follows, essentially, the same lines as for the semidiscrete case (Exercise 13.2.3). Now we consider the method of lines. In particular, we treat time discretizations of the Runge–Kutta type and of the linear multistep type. In both cases, we prove that generalized stability follows from the same property for the semidiscrete approximation.

From now on, we assume that $g^{n+1} \equiv 0$ in Eq. (13.2.1c). Furthermore, we use the boundary conditions to eliminate all vectors v_j , $j = -r+1, -r+2, \dots, 0$, from the approximation. The resulting system has the form

$$\begin{aligned}\frac{dv_j}{dt} &= Qv_j + F_j, \quad j = 1, 2, \dots, \\ v_j(0) &= 0, \quad j = 1, 2, \dots\end{aligned}\tag{13.2.7}$$

Even if the original differential equation has constant coefficients, the difference operator Q now has variable coefficients when applied to v_j near the boundary. We assume that the approximation is stable in the generalized sense, that is, the solution of the resolvent equation

$$(sI - Q)\hat{v}_j = \hat{F}_j, \quad j = 1, 2, \dots, \tag{13.2.8}$$

satisfies the estimate

$$\|\hat{v}\|_h \leq K(\eta) \|\hat{F}\|_h, \quad \eta > \eta_0, \tag{13.2.9}$$

where $\lim_{\eta \rightarrow \infty} K(\eta) = 0$. In all our examples treated so far, the constant $K(\eta)$ satisfies

$$K(\eta) \leq \frac{\text{constant}}{\eta}, \tag{13.2.10}$$

and we assume here that this is the case. We discretize time by using a method of the Runge–Kutta type, which gives us

$$w_j^{n+1} = P(kQ)w_j^n + kG_j^n, \quad G_j^n = P_1(kQ)F_j^n. \tag{13.2.11}$$

Here P and P_1 are polynomials in kQ . Furthermore, it is assumed that there is a relation between k and h such that $\|kQ\| \leq \text{constant}$.

Let us apply the method to the scalar ordinary differential equation

$$y' = \lambda y.$$

We obtain

$$w^{n+1} = P(\lambda k)w^n.$$

From Section 6.7, we know that there is an open domain Ω in the complex plane $\mu = \lambda k$ such that

$$|P(\mu)| < 1 \quad \text{if } \mu \in \Omega.$$

We now make the following assumptions.

Assumption 13.2.1. *There exists a number $R_1 > 0$ such that the open half-circle*

$$|\mu| < R_1, \quad \operatorname{Re} \mu < 0,$$

belongs to Ω . If $\mu = i\alpha$, α real, $|\alpha| \leq R_1$ does not belong to Ω , then necessarily

$$P(i\alpha) = e^{i\varphi}, \quad \varphi \text{ real}, \quad -\pi \leq \varphi \leq \pi. \quad (13.2.12)$$

Assumption 13.2.2. *Let φ be a given real number. If $\mu = i\alpha$, $|\alpha| \leq R_1$ is a purely imaginary solution of*

$$P(\mu) = e^{i\varphi},$$

then it is a simple root, and there is no other purely imaginary root $i\beta$ with

$$P(i\beta) = e^{i\varphi}, \quad -R_1 \leq \beta \leq R_1.$$

For any consistent approximation, the above condition is satisfied, if we restrict R_1 to be sufficiently small because

$$P(\mu) = 1 + \mu + \mathcal{O}(\mu^2).$$

It is also satisfied if the approximation is dissipative, that is, if $\mu = i\alpha$, $0 < q|\alpha| \leq R_1$, belongs to Ω .

Let $i\alpha$ be a root of the above type. Consider the perturbed equation

$$P(\mu) = e^{i(\varphi + \xi) + \eta}, \quad \xi, \eta \text{ real}, \quad \eta > 0. \quad (13.2.13)$$

We then have the following lemma.

Lemma 13.2.1. *For sufficiently small $|i\xi + \eta|$, the root of Eq. (13.2.13) can be expanded into a convergent Taylor series*

$$\mu(i\xi + \eta) = i\alpha + \gamma(i\xi + \eta) + \mathcal{O}(|i\xi + \eta|^2).$$

Here $\operatorname{Re} \mu(i\xi + 0) \geq 0$, and $\gamma > 0$ is real and positive. Therefore,

$$\operatorname{Re} \mu(i\xi + \eta) = \operatorname{Re} \mu(i\xi + 0) + \gamma\eta + \mathcal{O}(|\xi|\eta + \eta^2) \geq \gamma\eta + \mathcal{O}(|\xi|\eta + \eta^2).$$

Proof. Because, by assumption, $i\alpha$ is a simple root, the Taylor expansion is valid and $\gamma \neq 0$. Because $\mu \notin \Omega$, we have $\operatorname{Re} \mu(i\xi + \eta) \geq 0$ for all suffi-

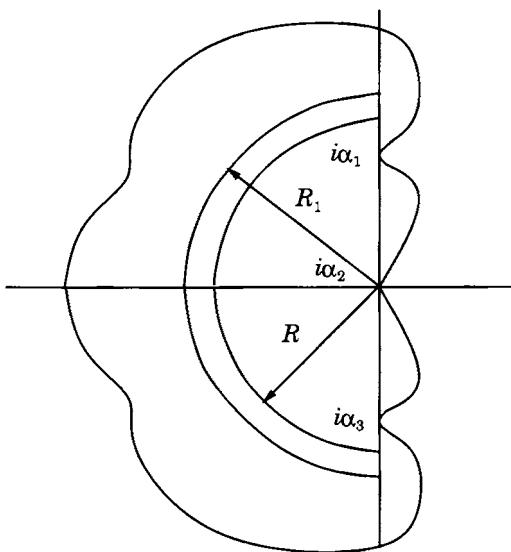


Figure 13.2.1.

ciently small $|\xi|$, η , $\eta \geq 0$. Therefore γ must be real and positive and the lemma follows.

We shall now prove the following theorem.

Theorem 13.2.3. *Assume that Assumptions 13.2.1 and 13.2.2 hold and that the semidiscrete approximation is stable in the generalized sense with $K(\eta)$ satisfying Eq. (13.2.10). Then the fully discretized approximation is stable in the same sense if*

$$\|kQ\|_h \leq R, \quad (13.2.14)$$

where R is any constant with $R < R_1$.

Proof. The resolvent equation is

$$z\tilde{w}_j = P(kQ)\tilde{w}_j + k\tilde{G}_j, \quad z = e^{sk}, \quad \eta = \operatorname{Re} s > \eta_0, \quad (13.2.15)$$

where $\tilde{G}_j = P_1(kq)\tilde{F}_j$. By assumption, the difference operators kQ are bounded, and $P_1(kQ)$ is a polynomial; that is,

$$\|\tilde{G}(z)\|_h \leq \text{constant } \|\tilde{F}(z)\|_h.$$

Therefore, by Eq. (13.2.5), we must show that the solutions of Eq. (13.2.15) satisfy

$$\|\tilde{w}\|_h^2 \leq K(z)\|\tilde{G}\|_h^2, \quad \lim_{\eta \rightarrow \infty} K(z) = 0;$$

that is,

$$\|(zI - P(kQ))^{-1}\|_h \leq \frac{K_1(z)}{k}, \quad \lim_{\eta \rightarrow \infty} K_1(z) = 0.$$

For large $|z|$, the estimate follows easily. Because $P(kQ)$ is a bounded operator, we get from Eq. (13.2.15)

$$\begin{aligned} \|\tilde{w}\|_h &\leq \frac{1}{|z|} \|P(kQ)\tilde{w}\|_h + \frac{k}{|z|} \|\tilde{G}\|_h, \\ &\leq \frac{\text{constant}}{|z|} \|\tilde{w}\|_h + \frac{k}{|z|} \|\tilde{G}\|_h, \end{aligned}$$

and, for $|z|$ sufficiently large,

$$\|\tilde{w}\|_h \leq \frac{\text{constant } k}{|z|} \|\tilde{G}\|_h.$$

Next consider $|z| \leq c_0$.

Let $\mu_\nu(z)$ be the roots of the polynomial $z - P(\mu)$. Then, we have

$$zI - P(kQ) = x_0 \prod_\nu (\mu_\nu(z)I - kQ),$$

where x_0 is a complex constant. Thus, for every z with $|z| > 1$, we can write Eq. (13.2.15) in the form

$$x_0 \prod_\nu (\mu_\nu(z)I - kQ)\tilde{w}_j = k\tilde{G}_j. \quad (13.2.16)$$

We now consider each factor $\mu_\nu(z)I - kQ$. By assumption, the roots μ_ν do not belong to Ω for $|z| > 1$. There are three possibilities:

1. $|\mu_j(z)| - R > \delta > 0$, δ constant; Eq. (13.2.14) implies

$$\|(\mu_j(z)I - kQ)^{-1}\|_h \leq (|\mu_j(z)|I - R)^{-1} \leq \delta^{-1}. \quad (13.2.17)$$

In particular, if $\operatorname{Re} \mu_j \leq 0$, then $\mu_j \notin \Omega$ implies

$$|\mu_j(z)| - R \geq \delta_1 > 0.$$

Thus, the above inequality holds if the constant $\delta > 0$ is chosen sufficiently small.

2. $\operatorname{Re} \mu_\nu \geq \delta_2 > 0$, $\delta_2 = \text{constant} > 0$. Let $s = \mu_\nu(z)/k$, and use the resolvent condition [Eqs. (13.2.9) and (13.2.10)] for the semidiscrete problem. For k small enough, we have $\operatorname{Re} \mu_\nu/k > \eta_0$, which gives

$$\begin{aligned} \|(\mu_\nu(z) - kQ)^{-1}\|_h &= \frac{1}{k} \left\| \left(\frac{\mu_\nu(z)}{k} - Q \right)^{-1} \right\|_h \\ &\leq \frac{1}{k} \text{constant} \frac{k}{\operatorname{Re}(\mu_\nu(z))} \leq \frac{\text{constant}}{\delta_2}. \end{aligned} \quad (13.2.18)$$

3. $\operatorname{Re} \mu_\nu(z) > 0$, but $\lim_{z \rightarrow z_0} \mu_\nu(z) = i\alpha$, $z_0 = e^{i\varphi}$, α, φ real. Let

$$z = e^{i\varphi + k(i\xi + \eta)}, \quad \varphi, \xi, \eta \text{ real.}$$

By Lemma 13.2.1, we have

$$\operatorname{Re}(\mu_\nu(z)) \geq \gamma ky + \mathcal{O}(k^2(\xi y + y^2)). \quad (13.2.19)$$

Therefore, by Eq. (13.2.10), and for k small enough

$$\|(\mu_\nu(z) - kQ)^{-1}\|_h = \frac{1}{k} \left\| \left(\frac{\mu_\nu(z)}{k} - Q \right)^{-1} \right\|_h \leq \frac{\text{constant}}{k\eta}. \quad (13.2.20)$$

Now we can prove the theorem. Combining the estimates (13.2.17) to (13.2.20) and observing that, for a given $z = e^{i\varphi}$, there is at most one imaginary root $\mu_\nu(z) = i\alpha$, we obtain for $|z| \leq c_0$

$$\begin{aligned} &\|(zI - P(kQ))^{-1}\|_h \\ &= |x_0|^{-1} \left\| \prod_p (\mu_\nu(z) - kQ)^{-1} \right\|_h \\ &\leq \begin{cases} \text{constant } \frac{1}{\eta k}, & \text{if one of the roots has the form (13.2.19),} \\ \text{constant,} & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, $\eta k \leq \text{constant}$ implies

$$\|\tilde{w}\|_h \leq \text{constant} k \|\tilde{F}\|_h \leq \frac{\text{constant}}{\eta} \|\tilde{F}\|_h.$$

This proves the theorem.

Instead of a Runge–Kutta method, we now consider a multistep method

$$(I + k\beta_{-1}Q)w_j^{n+1} = \sum_{\sigma=0}^q (\alpha_\sigma I - k\beta_\sigma Q)w_j^{n-\sigma} + kF_j^n \quad (13.2.21)$$

with real coefficients α_σ and β_σ . The resolvent equation becomes

$$(P_1(z)I - kP_2(z)Q)\tilde{w}_j = kz^q \tilde{F}_j, \quad (13.2.22)$$

where

$$P_1 = z^{q+1} - \sum_{\sigma=0}^q \alpha_\sigma z^{q-\sigma}, \quad P_2 = \beta_{-1}z^{q+1} - \sum_{\sigma=0}^q \beta_\sigma z^{q-\sigma}.$$

Again, we apply the method to the scalar differential equation $y' = \lambda y$. Then, we obtain the characteristic equation

$$P_1(z) - \mu P_2(z) = 0, \quad \mu = \lambda k.$$

We make the usual assumptions for the multistep method.

Assumption 13.2.3.

1. *The equations*

$$P_1(z) = 0, \quad P_2(z) = 0$$

have no root in common,

2.

$$\sum_{\sigma=0}^q \alpha_\sigma = 1, \quad \sum_{\sigma=-1}^q \beta_\sigma = 1,$$

3. *and the roots z_j of $P_1(z) = 0$ with $|z_j| = 1$ are simple.*

Thus, $P_1(1) = 0$, $P_2(1) = 1$, $P'_1(1) \neq 0$.

As in the previous case, there is an open domain Ω in the complex plane $\mu = \lambda k$ such that

$$P_1(z) - \mu P_2(z) \neq 0, \quad \text{for } |z| \geq 1, \mu \in \Omega. \quad (13.2.23)$$

We make the same construction as earlier which leads to the following assumption.

Assumption 13.2.4. *There exists a number $R_1 > 0$ such that the open half-circle*

$$|\mu| < R_1, \quad \operatorname{Re} \mu < 0,$$

belongs to Ω . If $\mu = i\alpha$, α real, $|\alpha| < R_1$ does not belong to Ω , then there is a $z = e^{i\varphi}$, φ real, such that

$$P_1(e^{i\varphi}) - i\alpha P_2(e^{i\varphi}) = 0. \quad (13.2.24)$$

Assumption 13.2.5. $z = e^{i\varphi}$ is a simple root of

$$P_1(z) - i\alpha P_2(z) = 0.$$

$P_1(z)$ has only simple roots near $z = 1$ and, therefore, the last assumption holds if we choose R_1 sufficiently small.

If $z = e^{i\varphi}$ satisfies Eq. (13.2.24), then $P_2(e^{i\varphi}) \neq 0$. Otherwise, also $P_1(e^{i\varphi}) = 0$ and $e^{i\varphi}$ would be a common root of P_1 and P_2 . Therefore, for $z = e^{i\varphi}$

$$\mu = P_1(z)/P_2(z) =: S(z),$$

is well defined and

$$\frac{dS}{dz} = \frac{P_2(z)P'_1(z) - P_1(z)P'_2(z)}{P_2^2(z)} = \frac{P'_1(e^{i\varphi}) - i\alpha P'_2(e^{i\varphi})}{P_2(e^{i\varphi})} \neq 0.$$

We consider now the perturbed equation

$$P_1(z) - \mu P_2(z) = 0, \quad z = e^{i\varphi + i\xi + \eta}.$$

In the same way as Lemma 13.2.1, one can prove the following lemma.

Lemma 13.2.2. *The solution $\mu = \mu(i\xi + \eta)$ of the perturbed equation has the same properties as in Lemma 13.2.1.*

Definition 13.2.1, which defines stability in the generalized sense, can also be used for multistep methods, and Theorem 13.2.1 holds. We can now prove the following theorem.

Theorem 13.2.4. *Assume that Assumptions 13.2.3 to 13.2.5 hold and that the semidiscrete approximation is stable in the generalized sense. Then the totally discretized multistep method (13.2.21) is stable in the same sense if*

$$\|kQ\|_h \leq R,$$

where R is any constant with $R < R_1$.

Proof. We begin by considering z values in the neighborhood of z_0 , where $P_2(z_0) = 0$. Because P_1 and P_2 have no roots in common and kQ is bounded, we have

$$\begin{aligned} \|\tilde{w}\|_h &\leq \|(P_1(z)I - kP_2(z)Q)^{-1}\|_h k|z|^q \|\tilde{F}\|_h, \\ &\leq \frac{\text{constant}}{|z|^{q+1}} k|z|^q \|\tilde{F}\|_h \leq \frac{\text{constant } k}{|z|} \|\tilde{F}\|_h, \\ &\leq \frac{\text{constant}}{\eta} \|\tilde{F}\|_h, \end{aligned}$$

which is the desired estimate.

Thus, for the remaining part of the proof we can assume that $P_2(z) \neq 0$ and write the resolvent equation in the form

$$(S(z)I - kQ)\tilde{w}_j = \frac{kz^q}{P_2(z)} \tilde{F}_j, \quad j = 1, 2, \dots \quad (13.2.25)$$

First, we consider the case $|z| \geq c_0$, where c_0 is a sufficiently large constant. If $\beta_{-1} = 0$, then $|S(z)| \geq \text{constant } |z|$, and we obtain, for $|S(z)| \geq 2\|kQ\|_h$,

$$\|(S(z)I - kQ)^{-1}\|_h \leq \frac{1}{|S(z)| - \|kQ\|_h} \frac{2}{|S(z)|}.$$

Therefore,

$$\|\tilde{w}\|_h \leq \frac{2}{|S(z)|} \frac{k|z|^q}{|P_2(z)|} \|\tilde{F}\|_h \leq \frac{\text{constant } k}{|z|} \|\tilde{F}\|_h \leq \frac{\text{constant}}{\eta} \|\tilde{F}\|_h,$$

which is the desired estimate.

If $\beta_{-1} \neq 0$, then

$$\lim_{|z| \rightarrow \infty} S(z) = \beta_{-1}^{-1}.$$

We first assume $\beta_{-1} < 0$. Because $S(z) \notin \Omega$, there is a constant $\delta_1 > 0$ such that $|\beta_{-1}^{-1}| - R > \delta_1$. We get

$$\|(S(z)I - kQ)^{-1}\|_h \leq \frac{1}{|S(z)| - \|kQ\|_h} \leq \text{constant},$$

for $|z|$ sufficiently large, that is,

$$\|\tilde{w}\|_h \leq \frac{\text{constant } k|z|^q}{|P_2(z)|} \|\tilde{F}\|_h \leq \frac{\text{constant } k}{|z|} \|\tilde{F}\|_h \leq \frac{\text{constant}}{\eta} \|\tilde{F}\|_h.$$

Next we assume $\beta_{-1} > 0$. Then, recalling the resolvent estimate for the semidiscrete problem, for $0 < \delta_2 < \beta_{-1}^{-1}$ and $|z|$ sufficiently large, we get

$$\begin{aligned} \|(S(z)I - kQ)^{-1}\|_h &= \frac{1}{k} \left\| \left(\frac{S(z)}{k} I - Q \right)^{-1} \right\|_h \leq \frac{\text{constant}}{k} \frac{k}{\text{Re } S(z)} \\ &\leq \text{constant}. \end{aligned}$$

Thus, the desired estimate also follows in this case.

We have now finished the proof for $|z| \geq c_0$, and continue with the case $|z| \leq c_0$. In this case, $\eta k \leq \text{constant}$ by definition of z . Therefore, it is sufficient to prove an estimate

$$\|\tilde{w}\|_h \leq \text{constant } k \|\tilde{F}\|_h.$$

There are three possibilities:

1. There is a constant $\delta > 0$ such that

$$|S(z)| - k\|Q\|_h \geq \delta.$$

Then

$$\|\tilde{w}\|_h \leq \frac{k|z^q/P_2(z)| \|\tilde{F}\|_h}{\delta} \leq \text{constant } k \|\tilde{F}\|_h.$$

Thus, the desired estimate holds. The above assumption is satisfied if the constant δ is chosen to be sufficiently small and $\operatorname{Re} S \leq 0$, because $S(z) \notin \Omega$.

2. $\operatorname{Re} S(z) > \delta_1 > 0$. Recalling the resolvent estimate for the semidiscrete problem, we get

$$\begin{aligned} \|(S(z)I - kQ)^{-1}\|_h &= k^{-1} \left\| \left(\frac{S(z)}{k} I - Q \right)^{-1} \right\|_h \\ &\leq \frac{\text{constant}}{k} \cdot \frac{k}{\operatorname{Re} S(z)} \leq \frac{\text{constant}}{\delta_1} \end{aligned}$$

Thus,

$$\|\tilde{w}\|_h \leq \frac{\text{constant}}{\delta_1} \frac{k|z|^q}{|P_2(z)|} \|\tilde{F}\|_h \leq \text{constant } k \|\tilde{F}\|_h,$$

and the desired estimate holds.

3. $\operatorname{Re} S(z) > 0$, but $\lim_{z \rightarrow z_0} S(z) = i\alpha$, $z_0 = e^{i\varphi}$, α, φ real, $|\alpha| \leq R$. Let $z = e^{i\varphi + (i\xi + \eta)k}$. By Lemma 13.2.2,

$$\operatorname{Re} S(z) \geq \gamma k \eta + \mathcal{O}(k^2(|\xi|\eta + \eta^2)).$$

Therefore,

$$\|(S(z)I - kQ)^{-1}\|_h \leq \frac{\text{constant}}{\operatorname{Re} S(z)} \leq \frac{\text{constant}}{k\eta};$$

that is,

$$\|\tilde{w}\|_h \leq \frac{\text{constant}}{k\eta} \frac{k|z|^q}{|P_2(z)|} \|\tilde{F}\|_h \leq \frac{\text{constant}}{\eta} \|\tilde{F}\|_h.$$

This completes the final part of the proof.

EXERCISES

13.2.1. Write a program for the standard fourth-order Runge–Kutta method applied to

$$\begin{aligned}\frac{dv_j}{dt} &= D_0 v_j, \quad j = 1, 2, \dots, N - 1, \\ v_0(t) &= 2v_1(t) - v_2(t), \\ v_N(t) &= g(t),\end{aligned}$$

such that generalized stability is guaranteed for k/h sufficiently small.

13.2.2. Consider the well posed initial-boundary-value problem for

$$u_t = Au_x, \quad 0 \leq x < \infty, \quad t \geq 0, \quad L_0 u(0, t) = g(t).$$

Construct a method based on fourth-order approximation in space and fourth-order Runge–Kutta time discretization such that the resulting approximation is stable in the generalized sense.

13.2.3. Prove Theorem 13.2.2.

13.3. SEVERAL SPACE DIMENSIONS

In Section 11.5, we used the energy method for the stability analysis of multidimensional problems. In this section, we give a brief introduction to the general theory based on the Laplace transform technique when applied to problems in two space dimensions. The Kreiss condition again plays the main role. We formally define the periodic channel problem and investigate a few simple examples.

We begin by considering two-dimensional hyperbolic systems

$$\frac{\partial u}{\partial t} = A \frac{\partial u}{\partial x} + B \frac{\partial u}{\partial y} + F, \quad 0 \leq x < \infty, \quad -\infty < y < \infty, \quad t \geq 0, \tag{13.3.1}$$

with 2π -periodic solutions in the y direction and boundary conditions at $x = 0$. The grid is defined in the usual way by

$$\Omega_h = (ih_1, jh_2), \quad i = 1, 2, \dots, \quad j = 1, 2, \dots, N.$$

The approximation is

$$\begin{aligned}
 v_{ij}^{n+1} &= Qv_{ij}^n + kF_{ij}, & (x_i, y_j) \in \Omega_h, \quad t \geq 0, \\
 L_0 v_{0j}^{n+1} &= g_j^{n+1}, & j = 1, 2, \dots, N, \\
 v_{ij}^{n+1} &= v_{i,j+N}^{n+1}, & (x_i, y_j) \in \Omega_h, \\
 v_{ij}^0 &= f_{ij}, & (x_i, y_j) \in \Omega_h.
 \end{aligned} \tag{13.3.2}$$

By extending the gridfunction v_{ij}^n as a stepfunction in time, we can use a Fourier series expansion in y and take the Laplace transform in t . With $f = F = 0$ and $z = e^{sk}$, we get the transformed system, for $\tilde{v}_i = \tilde{v}_i(\xi, z)$, $\tilde{g} = \tilde{g}(\xi, z)$,

$$\begin{aligned}
 z\tilde{v}_i &= \hat{Q}(\xi)\tilde{v}_i, & i = 1, 2, \dots, \\
 \tilde{L}_0\tilde{v}_0 &= \tilde{g}, \\
 \|\tilde{v}\|_h &< \infty.
 \end{aligned} \tag{13.3.3}$$

We are back to a one-dimensional problem, but a new parameter $\xi = \omega_2 h_2$ has entered the problem. Now we must check that there is no eigenvalue z with $|z| > 1$ for any ξ with $|\xi| \leq \pi$. Furthermore, the Kreiss condition must hold uniformly in ξ :

Definition 13.3.1. *The Kreiss condition is satisfied if the solutions of Eq. (13.3.3) satisfy*

$$\sum_{\nu=-r+1}^r |\tilde{v}_\nu(\xi, z)|^2 \leq K |\tilde{g}(\xi, z)|^2, \quad |z| > 1, \quad |\xi| \leq \pi, \tag{13.3.4}$$

where the constant K is independent of \tilde{g} , ξ , and z .

REMARK. We only consider the principle part of Q here. In general, the condition is formulated with $|z| > 1$ replaced by $|z| > 1 + \eta_0 k$.

The Kreiss condition leads to estimates analogous to the ones for the one-dimensional case. For example, Parseval's relation for the Fourier transform and for the Laplace transform gives the following theorem.

Theorem 13.3.1. *The solutions of Eq. (13.3.2) with $f = F = 0$ satisfy, for any fixed i , the estimate*

$$\sum_{\nu=1}^n \sum_{j=1}^N |v_{ij}^\nu|^2 h_2 k \leq \text{constant} \sum_{\nu=1}^n \sum_{j=1}^N |g_j^\nu|^2 h_2 k, \tag{13.3.5}$$

if, and only if, the Kreiss condition is satisfied.

Estimates for the full original problem are then derived in the familiar way by constructing an auxiliary problem with special boundary conditions. In general, it is of course possible that a straightforward multidimensional generalization of a stable one-dimensional approximation becomes unstable. For example, the backward Euler method for $u_t = u_x$ is stable with linear extrapolation along the (approximate) characteristic at the boundary, but the two-dimensional fractional step version of it for $u_t = u_x + u_y$ is not (see Exercise 13.3.2).

The necessary analysis to check the Kreiss condition is, in general, complicated for problems in several space dimensions (sometimes even in one space dimension). Thuné (1990) has developed a software system called IBSTAB that is based on a numerical algorithm. We briefly indicate how it works for problems in two space dimensions.

We want to find the stability limit for $\lambda = k/h_1 = k/h_2$. The characteristic equation, which defines the general solution of the Fourier–Laplace transformed difference equation, is written as

$$P(z, \kappa, \xi, \lambda) = 0. \quad (13.3.6)$$

Here P is a polynomial in z , κ , λ and a trigonometric polynomial in ξ . Let κ_ν , $\nu = 1, 2, \dots, l$, be the solutions with $|\kappa_\nu(z)| < 1$ for $|z| > 1$. Nontrivial solutions of the eigenvalue problem exist if $\text{Det}(C) = 0$, and we write this equation

$$g(z, \kappa_1, \dots, \kappa_l, \xi, \lambda) = 0. \quad (13.3.7)$$

In principle, one could use a general algorithm for systems of nonlinear equations to solve Eqs. (13.3.6) and (13.3.7) for $0 \leq \xi \leq 2\pi$ and a sequence of λ values. However, the computing time would be prohibitive. Therefore, one must take advantage of the special properties of the problem.

For the special case $\lambda = 0$, the difference scheme reduces to an approximation of $\partial u / \partial t = 0$. Therefore, it is natural to assume that all solutions z are located in the unit disc for $\lambda = 0$. The main idea of the algorithm is to move λ in small steps and to catch any solution z that appears outside the unit circle. Because z is a continuous function of λ , it is expected to be near the unit circle for $\lambda = \lambda_j + \delta\lambda$, if it was inside for $\lambda = \lambda_j$.

The implementation is done by letting z take values along the circle $|z| = 1 + \eta$, solving for the κ_ν with $|\kappa_\nu| < 1$ and checking the equation $g(z) = 0$. If $|g(z)| \ll 1$, then, and only then, the system (13.3.6) and (13.3.7) is solved. If the iterative solver converges, then the correct solution $(z, \kappa_1, \dots, \kappa_l)$ determines whether or not the Kreiss condition is violated. If it is, then the system also classifies the type of violation:

1. Eigenvalue $|z| > 1$.
2. Generalized eigenvalue $|z| = 1$ with at least one κ_ν , where $|\kappa_\nu| = 1$ is a simple root of Eq. (13.3.6).

3. Generalized eigenvalue $|z| = 1$, where all roots κ of Eq. (13.3.6) on the unit circle are multiple roots.
4. Eigenvalue $|z| = 1$, ($|\kappa_\nu| < 1, \nu = 1, 2, \dots, l$). The system calls cases 1 and 2 unstable and cases 3 and 4 weakly stable.

The variable stepsizes $\delta\lambda$, $\delta\xi$, and $\delta\theta$ in $z = (1 + \eta)e^{i\theta}$ are calculated according to a special strategy that takes the properties of $P(z)$ and $g(z)$ into account. With $M_L(z, \xi, \lambda)$ denoting the set of κ_ν with $|\kappa_\nu| < 1$ given by Eq. (13.3.6), the algorithm is:

Algorithm IBSTAB

```

for    $\lambda = \lambda_1(\lambda \leftarrow \lambda + \delta\lambda)\lambda_s$  do
for    $\xi = 0(\xi \leftarrow \xi + \delta\xi)2\pi$  do
for    $\theta = 0(\theta \leftarrow \theta + \delta\theta)2\pi$  do
 $z \leftarrow (1 + \eta)e^{i\theta}$ 
 $(\kappa_1, \kappa_2, \dots, \kappa_l) \leftarrow M_L(z, \xi, \lambda)$ 
if    $|g(z)|/|g'(z)| \leq \mu$  then
    solve (13.3.6), (13.3.7)
    check stability criteria
    update stepsizes  $\delta\lambda, \delta\xi, \delta\theta$ 

```

EXERCISES

13.3.1. Consider the Crank–Nicholson method for $u_t = u_x + u_y$:

$$u_{ij}^{n+1} - u_{ij}^n = \frac{k}{2} (D_{0x} + D_{0y})(u_{ij}^{n+1} + u_{ij}^n).$$

Prove that this equation is stable with the boundary condition

$$u_{0j}^n = 2u_{ij}^n - u_{2j}^n.$$

13.3.2. Prove that the fractional step method

$$(I - kD_{0x})(I - kD_{0y})u_{ij}^{n+1} = u_{ij}^n$$

is stable with the boundary condition

$$u_{0j}^n = 2u_{1j}^n - u_{2j}^n \quad (13.3.8)$$

but unstable with

$$u_{0j}^{n+1} = 2u_{i,j+1}^n - u_{2,j+2}^{n-1}$$

or

$$u_{0j}^n = 2u_{1,j+1}^n - u_{2,j+2}^n.$$

13.3.3. Prove that the time-split Crank–Nicholson scheme

$$\left(I - \frac{k}{2} D_{0x} \right) \left(I - \frac{k}{2} D_{0y} \right) u_{ij}^{n+1} = \left(I + \frac{k}{2} D_{0x} \right) \left(I + \frac{k}{2} D_{0y} \right) u_{ij}^n$$

is stable with the boundary condition (13.3.8) only if $k/h \leq 2$.

13.4. DOMAINS WITH IRREGULAR BOUNDARIES AND OVERLAPPING GRIDS

The results obtained for the periodic channel problem and for domains with piecewise linear boundaries can be used to construct stable approximations for problems in domains with smooth boundaries. The technique is an extension of the one used in Section 9.6 to prove well-posedness of the continuous problem, and we refer back to Figure 9.6.3. Part of the domain is shown in Figure 13.4.1.

The idea is to use a rectangular grid in the inner domain Ω_1 and a grid that is aligned with the boundary $\partial\Omega$ in the outer domain $\Omega - \Omega_1$. The inner grid is denoted by Ω_h and has a boundary Γ_h which is piecewise linear if the gridpoints are connected. The outer grid $\tilde{\Omega}_h$ is bounded on the inside by $\tilde{\Gamma}_h$ and on the outside by $\tilde{\Gamma}_h$. These boundary points are located on the smooth boundaries of $\Omega - \Omega_1$. The two grids Ω_h and $\tilde{\Omega}_h$ overlap. (Formally, Ω_h and $\tilde{\Omega}_h$ are defined such that they do not contain the boundary points.) An application of this overlapping grid technique is shown in Figures 13.4.4 and 13.4.7.

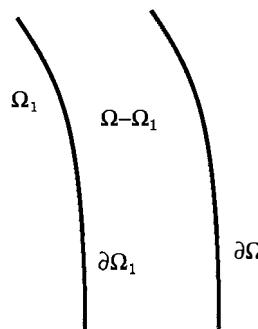


Figure 13.4.1.

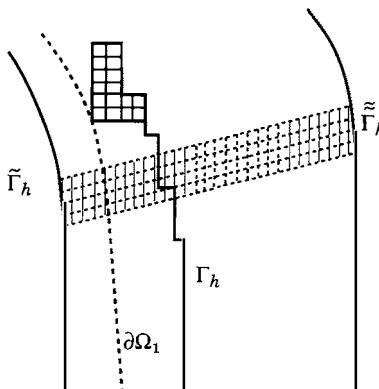


Figure 13.4.2. Overlapping grids.

Let us assume that a second-order accurate explicit method is used in the inner domain:

$$v_{ij}^{n+1} = Q v_{ij}^n, \quad (x_i, y_j) \in \Omega_h. \quad (13.4.1)$$

In the outer domain, we use transformed coordinates \tilde{x} and \tilde{y} , as in Section 10.6, and the transformed difference scheme is

$$\tilde{v}_{ij}^{n+1} = \tilde{Q} \tilde{v}_{ij}^n, \quad (\tilde{x}_i, \tilde{y}_j) \in \tilde{\Omega}_h. \quad (13.4.2)$$

The scheme shown in Figure 13.4.1 needs boundary values on Γ_h , and we use interpolation from the outer grid:

$$v_{ij}^{n+1} = L \tilde{v}_{ij}^{n+1}, \quad (x_i, y_j) \in \Gamma_h. \quad (13.4.3)$$

Here L is an interpolation operator, which is, usually, based on bilinear interpolation using four gridpoints in $\tilde{\Omega}_h$.

Similarly, boundary values for Eq. (13.4.2) are provided by

$$\tilde{v}_{ij}^{n+1} = \tilde{L} v_{ij}^{n+1}, \quad (\tilde{x}_i, \tilde{y}_j) \in \tilde{\Gamma}_h, \quad (13.4.4)$$

where also \tilde{L} is an interpolation operator.

At the outer boundary $\tilde{\Gamma}_h$, which has its gridpoints on the original boundary $\partial\Omega$, we use the boundary conditions for the differential equation together with extra numerical conditions if necessary:

$$B_{ij} \tilde{v}_{ij}^{n+1} = \tilde{g}_{ij}^{n+1}, \quad (\tilde{x}_i, \tilde{y}_j) \in \tilde{\Gamma}_h. \quad (13.4.5)$$

In the \tilde{y} direction, the solutions are periodic. Thus, the method shown by Eqs.

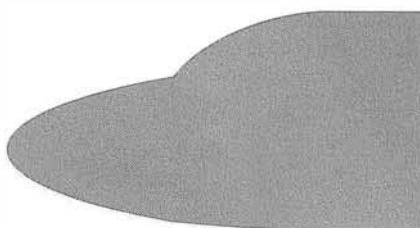


Figure 13.4.3.

(13.4.2), (13.4.4), and (13.4.5), when considered by itself, is a *periodic channel approximation*.

The interpolation is used for all variables in the vectors v_{ij} and \tilde{v}_{ij} , respectively. If the problem is hyperbolic, this means that we are, actually, overspecifying at some parts of the boundary for both the inner and the outer domain if we consider the algorithm locally in time. But there is no contradiction in this procedure. Overspecification is not unstable; it is the accuracy that cannot be retained if accurate data are not available. In our case, accurate data are available from the outer grid. Unfortunately, stability results are only available for one-dimensional problems. However, computations show very robust behavior.

We finish this section by presenting a computation done by Eva Pärt-Enander. The problem is to compute hypersonic inviscid flow around a body as shown in Figure 13.4.3. The flow is governed by the Euler equations and an upwind method (Roe, 1985) is used. The computational domain and the overlapping grids are shown in Figure 13.4.4. With one single grid covering

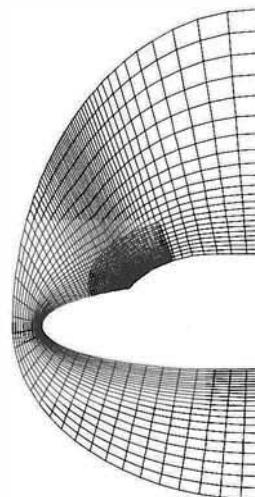


Figure 13.4.4.

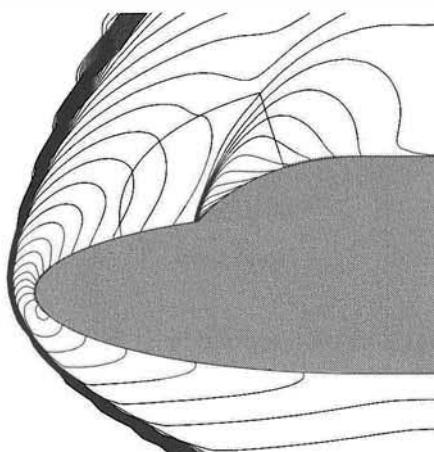


Figure 13.4.5.

the whole computational domain, there is an inherent difficulty caused by the concave corner on the body. This corner makes it impossible to construct a smooth transformation to a rectangle. There are ways of smoothing the grid, but near the corner, a discontinuity in the gridlines will always remain. This difficulty is eliminated by overlapping the critical area with a separate subgrid. Figure 13.4.5 shows iso-Mach lines (curves with equal speed) when a steady state has been reached.

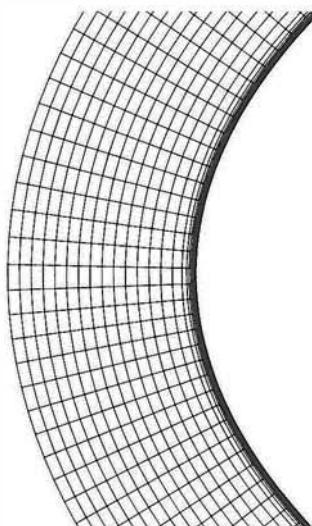


Figure 13.4.6.

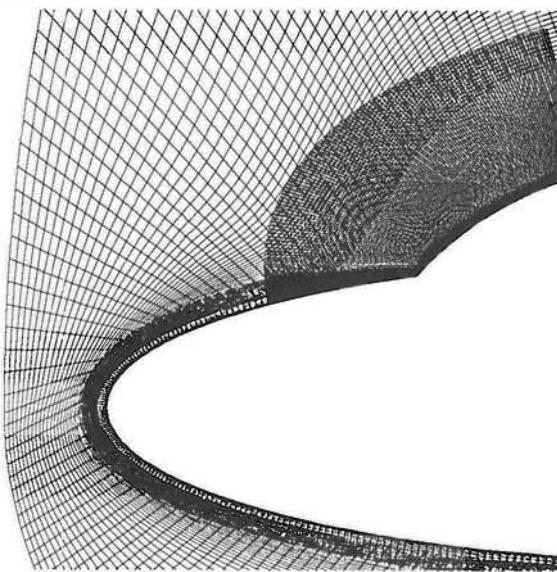


Figure 13.4.7.

When the Euler equations are replaced by the Navier–Stokes equations, there is a boundary layer near the body, because of zero velocity at the body. It is then convenient to use a narrow subgrid covering the boundary layer, as shown in Figure 13.4.6. The overlapping grids are shown in Figure 13.4.7. The iso-Mach lines for this case are shown in Figure 13.4.8.

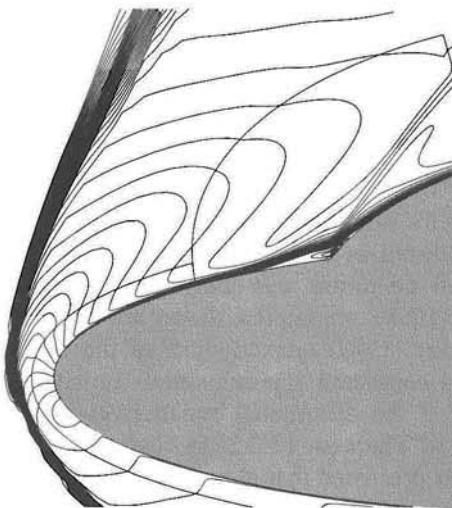


Figure 13.4.8.

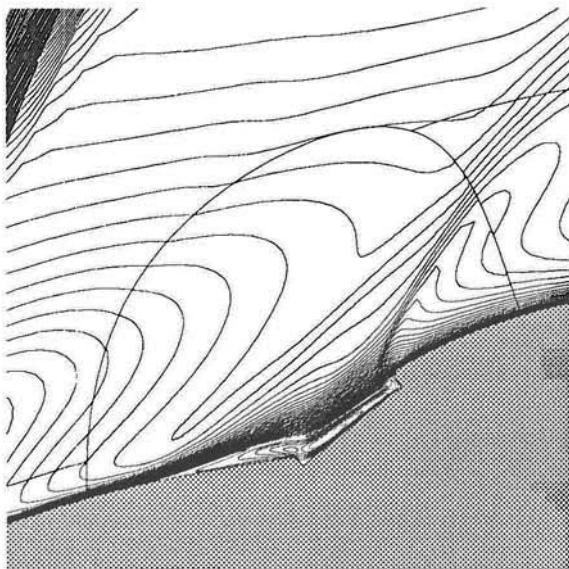


Figure 13.4.9.

In the last calculation a modification of the geometry is made corresponding to a canopy window. Using the overlapping grid technique, it is easy to construct a local subgrid so that the details of the flow can be computed. Figure 13.4.9 shows the iso-Mach lines of the result around the window.

EXERCISES

- 13.4.1.** Construct a one-dimensional scalar hyperbolic model problem with overlapping grids. Use linear interpolation between the grids and prove stability for the Lax-Wendroff scheme.

BIBLIOGRAPHIC NOTES

The general theory for initial boundary value problems for difference approximations was introduced in a series of papers by Kreiss in the 1960s. [The Godunov-Ryabenkii condition was introduced by Godunov and Ryabenkii (1963).] In Kreiss (1968), a complete theory for dissipative one-step methods was presented [Osher (1969) relaxed some of the conditions]. It was shown that dissipative and consistent approximations satisfying the Kreiss condition are strongly stable if the differential equation is strictly hyperbolic. This is a stronger version of Theorem 13.2.2. In Gustafsson, Kreiss, and Sundström (1972) a theory was presented that also included multistep schemes as well as nondissipative schemes and variable coefficients. Michelson (1983) extended the theory to the multidimensional case. In these papers, there is no assump-

tion on symmetric coefficient matrices, and the stability concept used is that of strong stability in the generalized sense. By using the symmetry assumption in large parts of our presentation, it is possible to use a simpler technique. Furthermore, in this case, strong stability follows from the Kreiss condition.

The sufficient (and convenient) stability criteria mentioned in Theorem 13.1.3 were given by Goldberg and Tadmor (1981) (where also Lemma 13.1.6 was proved). These criteria have been further refined in a series of papers by the same authors (1985, 1987, 1989). The latest are found in Goldberg (1991).

The construction of the auxiliary boundary conditions and the estimate in Lemma 13.1.8 is due to Wu (1994). The results on the method of lines in Section 13.2 were obtained by Kreiss and Wu (1993).

Error estimates follow directly from strong stability. For generalized stability, we must first subtract a proper function to make the initial and boundary conditions homogeneous, and in this process one may lose a factor h . However, it has been shown by Gustafsson (1975) that if there is no generalized eigenvalue at $z = 1$, then one can still have one order lower accuracy for the extra numerical boundary conditions.

The stability condition $k|a| < 2h$ given at the end of Section 13.1 for the Crank–Nicholson scheme was derived by Skölleramo (1975).

The stability problem for overlapping grids was treated for the one-dimensional hyperbolic case by Starius (1980). He proved that if the overlapping region is fixed and independent of the mesh-size, then stability follows for dissipative schemes.

A general system CMPGRD for generating overlapping grids is described in Cheshire and Henshaw (1990), and this system was used for computation of Navier–Stokes solutions in Henshaw (1987, 1992).

APPENDIX A.1

RESULTS FROM LINEAR ALGEBRA

We will collect some selected results from linear algebra needed in this book.

Let $A = (a_{ij})$ be a complex $(m \times m)$ matrix.

Lemma A.1.1. *If A has a complete set of linearly independent eigenvectors v_i , then*

$$T^{-1}AT = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m), \quad T = (v_1, v_2, \dots, v_m),$$

where the λ_i are the eigenvalues of A . If the eigenvalues of A are distinct ($\lambda_i \neq \lambda_j$ if $i \neq j$), then A has a complete set of eigenvectors.

Lemma A.1.2. Jordan Canonical Form. *For every matrix A there exists a matrix T such that*

$$T^{-1}AT = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{pmatrix},$$

where the submatrices J_ν have the form

$$J_\nu = \begin{pmatrix} \lambda_\nu & 1 & & & \\ & \lambda_\nu & \ddots & & \\ & & \ddots & 1 & \\ & & & & \lambda_\nu \end{pmatrix}, \quad \nu = 1, 2, \dots, r.$$

$T^{-1}AT$ is called the Jordan canonical form.

Lemma A.1.3. Schur's Lemma. *For every matrix A there is an unitary matrix U such that U^*AU is upper triangular.*

Lemma A.1.4. *Any matrix A satisfies the inequalities*

$$\rho(A) \leq \sup_{|u|=1} \langle Au, u \rangle \leq |A|,$$

where $\rho(A)$ is the spectral radius of A , $\rho(A) = \max |\lambda_i|$, and the λ_i are the eigenvalues of A . The set of values $\langle Au, u \rangle$ for all u with $|u| = 1$ is called the numerical range or field of values of A .

Lemma A.1.5. *If the spectral radius satisfies $\rho(A) \leq 1 - \delta$ for $\delta > 0$, then $|A|_* \leq 1 - \delta_1$, for some $\delta_1 > 0$, where $|u|_* = \langle u, Hu \rangle^{1/2}$ for a positive definite matrix H .*

Lemma A.1.6. *Let $M = M(x)$ be an $m \times m$ analytic matrix function whose eigenvalues are contained in two nonintersecting sets Σ_1 and Σ_2 . Suppose Σ_1 contains p eigenvalues and Σ_2 contains $m - p$ eigenvalues. Then there exists a nonsingular analytic matrix $T(x)$ such that*

$$T(x)M(x)T^{-1}(x) = \begin{pmatrix} M_{11} & 0 \\ 0 & M_{22} \end{pmatrix},$$

where M_{11} is $p \times p$ and has eigenvalues contained in Σ_1 and M_{22} is $(m - p) \times (m - p)$ and has eigenvalues contained in Σ_2 .

Lemma A.1.7. *Let A be a matrix with distinct eigenvalues λ_j and eigenvectors u_j , and let B be a matrix with $|B| \leq 1$; then, for sufficiently small $\epsilon > 0$, the eigenvalues of $A + \epsilon B$ are also distinct and the eigenvalues and eigenvectors, μ_j and v_j , of $A + \epsilon B$ satisfy*

$$|\lambda_j - \mu_j| \leq C_1 \epsilon, \\ |u_j - v_j| \leq C_2 \epsilon,$$

for constants C_1 and C_2 .

We say that a scalar-, vector-, or matrix-valued function u of a vector x is Lipschitz continuous for x in a domain Ω if there exists a constant $C > 0$ such that

$$|u(x) - u(x')| \leq C|x - x'|,$$

for all $x, x' \in \Omega$.

Lemma A.1.8. *Suppose that $A(x)$ is a Lipschitz continuous $m \times m$ matrix with distinct eigenvalues. Then there exists a nonsingular Lipschitz continuous $m \times m$ matrix $T(x)$ such that*

$$T^{-1}(x)A(x)T(x) = \text{diag}(\lambda_1, \dots, \lambda_m).$$

Let $C^p(\Omega)$ be the space of functions with p continuous derivatives on Ω . The following lemma then holds.

Lemma A.1.9. *Suppose that $A(x) \in C^p$ is an $m \times m$ matrix with distinct eigenvalues. Then there exists a nonsingular $m \times m$ matrix $T(x) \in C^p$ such that*

$$T^{-1}(x)A(x)T(x) = \text{diag}(\lambda_1, \dots, \lambda_m).$$

Lemma A.1.10. *Suppose that $A(x) \in C^p$. If the eigenvalues of A satisfy*

$$\operatorname{Re} \lambda_j \leq -\delta < 0,$$

then, there is a transformation $T(x) \in C^p$ such that

$$T^{-1}AT + (T^{-1}AT)^* \leq -\delta I.$$

APPENDIX A.2

LAPLACE TRANSFORM

In this section, we discuss the Laplace transform, which is closely related to the Fourier transform. Let $u(t)$, $0 \leq t < \infty$, be a continuous function, and assume that there are constants C and α such that

$$|u(t)| \leq Ce^{\alpha t}. \quad (\text{A.2.1})$$

Then the function

$$v(t) = \begin{cases} e^{-\eta t} u(t), & t \geq 0, \quad \eta > \alpha, \\ 0, & t \leq 0, \end{cases}$$

belongs to $L_2(t)$ and its Fourier transform

$$\begin{aligned} \tilde{v}(\xi) &= \mathcal{F}v(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi t} v(t) dt \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-(i\xi + \eta)t} u(t) dt \end{aligned} \quad (\text{A.2.2})$$

exists. Also,

$$u(t) = e^{\eta t} v(t) = e^{\eta t} \mathcal{F}^{-1}\tilde{v}(\xi) = \frac{e^{\eta t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi t} \tilde{v}(\xi) d\xi, \quad (\text{A.2.3})$$

and by Parseval's relation

$$\int_0^{\infty} e^{-2\eta t} |u(t)|^2 dt = \int_{-\infty}^{\infty} |\tilde{v}(\xi)|^2 d\xi. \quad (\text{A.2.4})$$

Let $s = i\xi + \eta$, ξ , η real. For $\eta > \alpha$, the Laplace transform of u is defined by

$$\mathcal{L} u = \hat{u}(s) = \int_0^\infty e^{-st} u(t) dt = \int_0^\infty e^{-(i\xi + \eta)t} u(t) dt. \quad (\text{A.2.5})$$

By Eq. (A.2.1),

$$|\hat{u}(s)| \leq \int_0^\infty e^{-\eta t} |u(t)| dt \leq C/(\eta - \alpha). \quad (\text{A.2.6})$$

By Eq. (A.2.3)

$$\hat{u}(s) = \mathcal{L} u = \sqrt{2\pi} \mathcal{F}v = \sqrt{2\pi} \tilde{v}. \quad (\text{A.2.7})$$

Therefore, we obtain, for any $\eta > \alpha$,

$$u(t) = \frac{e^{\eta t}}{2\pi} \int_{-\infty}^\infty e^{i\xi t} \hat{u}(i\xi + \eta) d\xi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{st} \hat{u}(s) d\xi. \quad (\text{A.2.8})$$

The last formula is often written in the form

$$u(t) = \frac{1}{2\pi i} \int_{\mathcal{C}} e^{st} \hat{u}(s) ds, \quad (\text{A.2.9})$$

where \mathcal{C} denotes the vertical line $\operatorname{Re} s = \eta$ in the complex s plane. Parseval's relation becomes

$$\int_0^\infty e^{-2\eta t} |u(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^\infty |\hat{u}(i\xi + \eta)|^2 d\xi. \quad (\text{A.2.10})$$

PROPERTIES OF THE LAPLACE TRANSFORM

Suppose that $u(t)$ has one continuous derivative and that du/dt also satisfies the estimate (A.2.1). Integration by parts gives us for $s \neq 0$

$$\hat{u}(s) = \int_0^\infty e^{-st} u(t) dt = -\frac{e^{-st}}{s} u(t) \Big|_0^\infty - \frac{1}{s} \int_0^\infty e^{-st} (du/dt) dt,$$

that is,

$$s\hat{u}(s) = \widehat{du/dt} + u(0). \quad (\text{A.2.11})$$

Equation (A.2.11) relates the Laplace transform of $\hat{u}(s)$ to $\widehat{du/dt}$. It also tells us that, for large $|s|$, $\operatorname{Re} s > \alpha$,

$$|\hat{u}(s)| \leq \text{constant}/|s|. \quad (\text{A.2.12})$$

Clearly we can use repeated integration by parts to relate $\hat{u}(s)$ to higher derivatives. In particular, if $d^p u/dt^p$ is continuous and satisfies Eq. (A.2.1) and if $d^j u(0)/dt^j = 0$, $j = 0, 1, \dots, p - 1$, then

$$s^p \hat{u}(s) = \widehat{d^p u/dt^p}. \quad (\text{A.2.13})$$

Using the theory of analytic functions one can prove that, for $\operatorname{Re} s > \alpha$, the Laplace transform $\hat{u}(s)$ is an analytic function of s which by Eq. (A.2.12) vanishes as $|s| \rightarrow \infty$, provided du/dt satisfies the above conditions. Therefore, one can replace the path \mathcal{L} in Eq. (A.2.9) by any path \mathcal{L}_1 , as indicated in Figure A.2.1.

In our applications of the Laplace transform, we often consider functions of x, t . Suppose that $u(x, t)$ is a continuous function for $0 \leq t < \infty$, $0 \leq x \leq l$, which satisfies the estimate (A.2.1) for all x . Then

$$\hat{u} = \hat{u}(x, s) = \int_0^\infty e^{-st} u(x, t) dt, \quad 0 \leq x \leq l, \quad \operatorname{Re} s > \alpha,$$

denotes the Laplace transform in time for every fixed x . The relation (A.2.11) reads

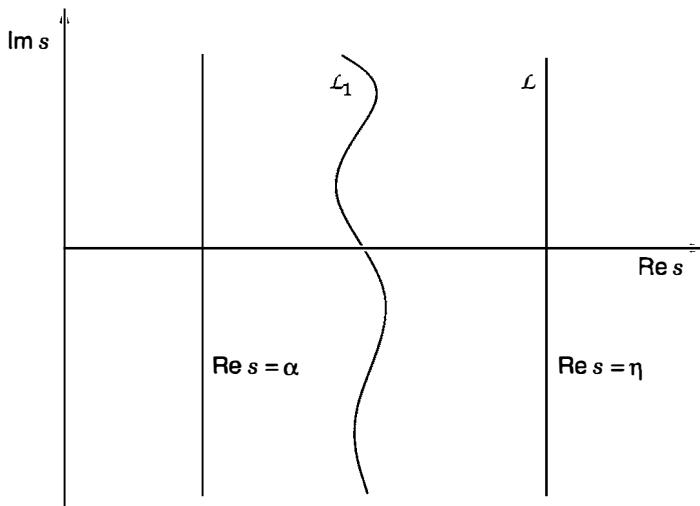


Figure A.2.1

$$s\hat{u} = \widehat{\partial u / \partial t} + u(x, 0). \quad (\text{A.2.14})$$

Also,

$$\hat{u}_x = \partial \hat{u}(x, s) / \partial x = \int_0^\infty e^{-st} \partial u(x, t) / \partial x \, dt = \widehat{\partial u / \partial x}. \quad (\text{A.2.15})$$

Let

$$\|u(\cdot, t)\|^2 = \int_0^l |u(x, t)|^2 dx, \quad \|\hat{u}(\cdot, s)\|^2 = \int_0^l |\hat{u}(x, s)|^2 dx,$$

denote the L_2 norm for every fixed t and s , respectively. The estimate (A.2.6) gives us

$$\begin{aligned} \|\hat{u}(\cdot, s)\|^2 &\leq \int_0^l \left[\int_0^\infty e^{-\eta t} |u(x, t)| \, dt \right]^2 dx, \\ &= \int_0^l \left[\int_0^\infty e^{-\frac{1}{2}(\eta - \alpha)t} e^{-\frac{1}{2}(\eta + \alpha)t} |u(x, t)| \, dt \right]^2 dx, \\ &\leq \int_0^l \left[\int_0^\infty e^{-(\eta - \alpha)t} \, dt \int_0^\infty e^{-(\eta + \alpha)t} |u(x, t)|^2 \, dt \right] dx, \\ &\leq \frac{1}{\eta - \alpha} \int_0^\infty e^{-(\eta - \alpha)t} e^{-2\alpha t} \|u(\cdot, t)\|^2 \, dt, \end{aligned} \quad (\text{A.2.16})$$

Thus, if instead of Eq. (A.2.1)

$$\|u(\cdot, t)\| \leq Ce^{\alpha t}$$

holds, then

$$\|\hat{u}(\cdot, s)\|^2 \leq 1/(\eta - \alpha)^2.$$

Parseval's relation (A.2.10) holds for every fixed x . Integrating over x gives us, for $\eta > \alpha$,

$$\int_0^\infty e^{-2\eta t} \|u(\cdot, t)\|^2 \, dt = \frac{1}{2\pi} \int_{-\infty}^\infty \|\hat{u}(\cdot, s)\|^2 d\xi, \quad s = i\xi + \eta. \quad (\text{A.2.17})$$

APPENDIX A.3

ITERATIVE METHODS

Let A be a nonsingular $n \times n$ matrix and x, b vectors with n elements. The linear system

$$Ax = b \quad (\text{A.3.1})$$

has a unique solution. Let $f(x)$ be a smooth vector function and $\epsilon > 0$ a parameter. We want to solve the nonlinear system of equations

$$Ax + \epsilon f(x) = b, \quad (\text{A.3.2})$$

by using the iteration

$$\begin{aligned} Ax^{n+1} + \epsilon f(x^n) &= b, & n &= 0, 1, 2, \dots, \\ x^0 &= A^{-1}b. \end{aligned} \quad (\text{A.3.3})$$

Theorem A.3.1. Assume that $f(x)$ is uniformly Lipschitz continuous, that is, there is a constant L such that for all x, y

$$|f(x) - f(y)| \leq L|x - y|. \quad (\text{A.3.4})$$

if $\tau = \epsilon |A^{-1}|L < 1$, then Eq. (A.3.2) has a unique solution \tilde{x} and $\lim_{n \rightarrow \infty} x^n = \tilde{x}$. Also,

$$|\tilde{x} - x^n| \leq \frac{\tau^n}{1 - \tau} |x^1 - x^0| \quad (\text{A.3.5})$$

Proof. Let x, y be two solutions. Then

$$|x - y| = \epsilon |A^{-1}(f(x) - f(y))| \leq \epsilon |A^{-1}|L|x - y|,$$

that is,

$$(1 - \tau)|\mathbf{x} - \mathbf{y}| \leq 0.$$

Therefore, $\mathbf{x} = \mathbf{y}$ and uniqueness follows. Subtracting

$$A\mathbf{x}^n + \epsilon \mathbf{f}(\mathbf{x}^{n-1}) = \mathbf{b}$$

from Eq. (A.3.3) gives us

$$\mathbf{x}^{n+1} - \mathbf{x}^n = -\epsilon A^{-1}(\mathbf{f}(\mathbf{x}^n) - \mathbf{f}(\mathbf{x}^{n-1})). \quad (\text{A.3.6})$$

Thus,

$$\begin{aligned} |\mathbf{x}^{n+1} - \mathbf{x}^n| &\leq \tau |\mathbf{x}^n - \mathbf{x}^{n-1}|, \\ &\leq \tau^2 |\mathbf{x}^{n-1} - \mathbf{x}^{n-2}|, \\ &\leq \tau^n |\mathbf{x}^1 - \mathbf{x}^0|. \end{aligned}$$

Let $m > n$, then

$$\begin{aligned} |\mathbf{x}^m - \mathbf{x}^n| &= \left| \sum_{i=n}^{m-1} (\mathbf{x}^{i+1} - \mathbf{x}^i) \right|, \\ &\leq \left(\sum_{i=n}^{m-1} \tau^i \right) |\mathbf{x}^1 - \mathbf{x}^0|, \\ &\leq \frac{\tau^n}{1-\tau} |\mathbf{x}^1 - \mathbf{x}^0|. \end{aligned} \quad (\text{A.3.7})$$

Therefore, \mathbf{x}^i represents a Cauchy sequence which converges to a vector $\tilde{\mathbf{x}}$. The vector $\tilde{\mathbf{x}}$ is a unique solution of Eq. (A.3.2) because

$$\lim_{n \rightarrow \infty} |\mathbf{f}(\mathbf{x}^n) - \mathbf{f}(\tilde{\mathbf{x}})| \leq L \lim_{n \rightarrow \infty} |\mathbf{x}^n - \tilde{\mathbf{x}}| = 0.$$

Let n be fixed and let $m \rightarrow \infty$. Then the estimate (A.3.5) follows from Eq. (A.3.7). This proves the theorem.

In most applications the nonlinear part $\epsilon \mathbf{f}(\mathbf{x})$ is not uniformly Lipschitz continuous. As an example consider

$$\mathbf{x} + \epsilon \mathbf{x}^2 = \mathbf{b}.$$

Then

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}) = \mathbf{x}^2 - \mathbf{y}^2 = (\mathbf{x} + \mathbf{y})(\mathbf{x} - \mathbf{y}), \quad (\text{A.3.8})$$

and there is no constant L such that (A.3.4) holds for all \mathbf{x}, \mathbf{y} . Instead we introduce the following definition.

Definition A.3.1. *$f(x)$ is called locally Lipschitz continuous if, for every $\mathbf{x}^0, \mathbf{y}^0$, and $R > 0$, there exists a constant $L = L(\mathbf{x}^0, \mathbf{y}^0, R)$ such that*

$$|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})| \leq L|\mathbf{x} - \mathbf{y}| \quad \text{for } |\mathbf{x} - \mathbf{x}^0| \leq R, |\mathbf{y} - \mathbf{y}^0| \leq R. \quad (\text{A.3.9})$$

The function \mathbf{x}^2 is locally Lipschitz continuous. By Eq. (A.3.8)

$$L(\mathbf{x}^0, R) \leq |\mathbf{x}^0| + |\mathbf{y}^0| + 2R.$$

Theorem A.3.1 can now be modified as follows.

Theorem A.3.2. *Assume that $\mathbf{f}(\mathbf{x})$ is locally Lipschitz continuous and let $L = L(\mathbf{x}^0, 1)$ be the local Lipschitz constant for the ball $|\tilde{\mathbf{x}} - \mathbf{x}^0| \leq 1$. If $\tau = \epsilon |A^{-1}|L < 1$, $\epsilon/(1 - \tau) \epsilon |A^{-1}\mathbf{f}(\mathbf{x}^0)| < 1$, then the system (A.3.2) has a solution $\tilde{\mathbf{x}}$ with*

$$|\tilde{\mathbf{x}} - \mathbf{x}^0| \leq \frac{\epsilon}{1 - \tau} |A^{-1}\mathbf{f}(\mathbf{x}^0)| < 1. \quad (\text{A.3.10})$$

Also, $\tilde{\mathbf{x}}$ is locally unique, that is, there is no other solution in the ball $|\tilde{\mathbf{x}} - \mathbf{x}^0| < 1$.

Proof. Consider the iteration (A.3.3). We have

$$A(\mathbf{x}^1 - \mathbf{x}^0) = -\epsilon \mathbf{f}(\mathbf{x}^0),$$

that is,

$$|\mathbf{x}^1 - \mathbf{x}^0| = \epsilon |A^{-1}\mathbf{f}(\mathbf{x}^0)|,$$

and by assumption

$$\frac{1}{1 - \tau} |\mathbf{x}^1 - \mathbf{x}^0| < 1, \quad \tau = \epsilon |A^{-1}|L < 1. \quad (\text{A.3.11})$$

We shall use induction to prove that the sequence $\{\mathbf{x}^m\}$ defined by Eq. (A.3.3) never leaves the ball B : $|\mathbf{x} - \mathbf{x}^0| \leq 1$. By Eq. (A.3.11) $\mathbf{x}^1 \in B$. Assume that $\mathbf{x}^j \in B$ for $j = 0, 1, \dots, m - 1$. By Eq. (A.3.6)

$$|\mathbf{x}^m - \mathbf{x}^{m-1}| \leq \tau |\mathbf{x}^{m-1} - \mathbf{x}^{m-2}| \leq \tau^{m-1} |\mathbf{x}^1 - \mathbf{x}^0|.$$

Therefore, we can use Eq. (A.3.7) with $n = 1$ and obtain

$$|\mathbf{x}^m - \mathbf{x}^1| \leq \frac{\tau}{1-\tau} |\mathbf{x}^1 - \mathbf{x}^0|.$$

Thus,

$$|\mathbf{x}^m - \mathbf{x}^0| \leq |\mathbf{x}^m - \mathbf{x}^1| + |\mathbf{x}^1 - \mathbf{x}^0| \leq \frac{1}{1-\tau} |\mathbf{x}^1 - \mathbf{x}^0| < 1,$$

and $\mathbf{x}^m \in B$. Thus, we have shown that the sequence $\{\mathbf{x}^m\}$ does not leave the ball B . Therefore, all the estimates of Theorem A.3.1 hold, and Theorem A.3.2 follows.

REFERENCES

- S. Abarbanel, D. Gottlieb, and E. Tadmor: Spectral methods for discontinuous problems. *Numerical Methods for Fluid Dynamics II*, Eds. K.W. Morton and M.J. Baines, Clarendon Press, Oxford, pp. 128–153 (1986).
- R.M. Beam and R.F. Warming: An implicit factored scheme for the compressible Navier-Stokes equations. *AIAA J.*, 16, No. 4, pp. 393–402 (1978).
- P. Brenner, V. Thomée, and L.B. Wahlbin: Besov spaces and applications to difference methods for initial value problems. *Lecture Notes in Mathematics*, Vol. 434, Springer-Verlag, New York (1975).
- C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang: *Spectral Methods in Fluid Dynamics*. Springer, Berlin (1988).
- M. Carpenter, D. Gottlieb, and S. Abarbanel: Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes. *J. Comp. Phys.*, 111, pp. 220–236 (1994).
- J.G. Charney, R. Fjortoft, and J. von Neumann: Numerical integration of the barotropic vorticity equation. *Tellus*, 2, No. 4, pp. 237–254 (1950).
- G. Chesshire and W.D. Henshaw: Composite overlapping meshes for the solution of partial differential equations. *J. Comp. Phys.*, 90, pp. 1–64 (1990).
- R. Chin and G. Hedstrom: A dispersion analysis for difference schemes, *Math. Comp.*, 32, pp. 1163–1170 (1978).
- R.V. Churchill and J.W. Brown: *Fourier Series and Boundary Value Problems*. McGraw-Hill, New York (1978).
- E. Coddington and N. Levinson: *Theory of Ordinary Differential Equations*. McGraw-Hill, New York (1955).
- S. Conte and C. de Boor: *Elementary Numerical Analysis*, McGraw-Hill, New York (1972).
- J.W. Cooley and J.W. Tukey: An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19, pp. 297–301 (1965).
- R. Courant and K.O. Friedrichs: *Supersonic Flow and Shock Waves*, Interscience Publishers, New York (1948).
- R. Courant, K.O. Friedrichs, and H. Levy: Über die partielle Differentialgleichungen der mathematischen Physik. *Mathematische Annalen*, 100, pp. 32–74 (1928).
- R. Courant and D. Hilbert: *Methods of Mathematical Physics. Vol. I*. Interscience Publishers, New York (1953).
- J. Crank and P. Nicolson: A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Cambridge Philosophical Soc.*, 43, No. 50, pp. 50–67 (1947).

- J. Douglas, Jr.: Alternating direction methods for parabolic systems in m -space variables. *J. Assoc. Comp. Machinery*, 9, pp. 450–456 (1962).
- J. Douglas, Jr.: On the numerical integration of $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = \partial u / \partial t$ by implicit methods. *J. Soc. Industrial Appl. Math.*, 3, pp. 42–65 (1955).
- J. Douglas, Jr. and J.E. Gunn: A general formulation of alternating direction methods, I. *Num. Math.*, 6, pp. 428–453 (1964).
- E.C. DuFort and S.P. Frankel: Stability conditions in the numerical treatment of parabolic differential equations. *Math. Tables Other Aids Computation*, 7, pp. 135–152 (1953).
- W.R. Dykson and J.R. Rice: Symmetric versus nonsymmetric differencing. *SIAM J. Sci. Stat. Comp.*, 6, No. 1, pp. 45–48 (1985).
- B. Engquist: A difference method for initial boundary value problems on general domains in two space dimensions. DCG Progress Report 1978, Uppsala University, Dept. of Computer Science, Report No. 82 (1978).
- B. Engquist, P. Lotstedt, and B. Sjögren: Nonlinear filters for efficient shock computation. *Math. Comp.*, 52, No. 186, pp. 509–537 (1989).
- B. Engquist and S. Osher: One-sided difference schemes and transonic flow. *Proc. Nat. Acad. Sci. USA*, 77, No. 6, pp. 3071–3074 (1980).
- B. Fornberg: On a Fourier method for the integration of hyperbolic equations. *SIAM J. Num. Anal.*, 12, pp. 509–528 (1975).
- B. Fornberg and D.M. Sloan: A review of pseudospectral methods for solving partial differential equations. *Acta Numerica*, in press.
- L. Fox: *The Numerical Solution of Two-Point Boundary Problems in Ordinary Differential Equations*. Oxford University Press, Fairlawn, New Jersey (1957).
- K.O. Friedrichs: Symmetric hyperbolic linear differential equations. *Comm. Pure Appl. Math.*, 7, pp. 345–390 (1954).
- K.O. Friedrichs: Symmetric positive linear differential equations. *Comm. Pure Appl. Math.*, 11, pp. 333–418 (1958).
- C.W. Gear: *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, NJ (1971).
- E. Godlewski and P.-A. Raviart: Hyperbolic systems of conservation laws. SMAI, Paris (1990).
- S.K. Godunov and V.S. Ryabenkii: Spectral criteria for the stability of boundary problems for non-self-adjoint difference equations. *Uspekhi Mat. Nauk*, 18 (1963).
- M. Goldberg: 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 (1991).
- M. Goldberg and E. Tadmor: Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. *Math. Comp.*, 44, pp. 361–377 (1985).
- M. Goldberg and E. Tadmor: Convenient stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 48, pp. 503–520 (1987).
- M. Goldberg and E. Tadmor: Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. *Math. Comp.*, 36, pp. 605–626 (1981).
- M. Goldberg and E. Tadmor: 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, pp. 179–185 (1989).
- J. Goodman and R. LeVeque: On the accuracy of stable schemes for 2D scalar conservation laws. *Math. Comp.*, 45, pp. 15–21, (1985).
- D. Gottlieb, B. Gustafsson, P. Olsson, and B. Strand: On the superconvergence of Galerkin methods for hyperbolic IBVP. RIACS Technical Report 93-07 (1993). To appear in *SIAM J. Num. Anal.*

- D. Gottlieb and S. Orszag: Numerical analysis of spectral methods: Theory and applications. *CBMS Regional Conference Series in Applied Mathematics*, 26, SIAM, Philadelphia (1977).
- D. Gottlieb and E. Tadmor: Recovering pointwise values of discontinuous data within spectral accuracy. *Progress in Scientific Computing*, 6, pp. 357–375 (1984).
- D. Gottlieb and E. Turkel: On time discretization for spectral methods. *Stud. Appl. Math.*, 63, pp. 66–86 (1980).
- B. Gustafsson: The convergence rate for difference approximations to general mixed initial boundary value problems. *SIAM J. Num. Anal.*, 18, pp. 179–190 (1981).
- B. Gustafsson: The convergence rate for difference approximations to mixed initial boundary value problems. *Math. Comp.*, 29, pp. 396–406 (1975).
- B. Gustafsson and H.-O. Kreiss: Difference approximations of hyperbolic problems with different time scales. I: The reduced problem. *SIAM J. Num. Anal.*, 20, No. 1, pp. 46–58 (1983).
- B. Gustafsson, H.-O. Kreiss, and A. Sundström: Stability theory of difference approximations for mixed initial boundary value problems. II. *Math. Comp.*, 26, pp. 649–686 (1972).
- B. Gustafsson and J. Oliger: Stable boundary approximations for implicit time discretizations for gas dynamics. *SIAM J. Sci. Stat. Comp.*, 3, No. 4, pp. 408–421 (1982).
- J. Hadamard: *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University, New Haven (1921).
- E. Hairer, S.P. Nørsett, and G. Wanner: *Solving Ordinary Differential Equations I, II*. Springer-Verlag, New York (1980).
- A. Harten: ENO schemes with subcell resolution. *J. Comp. Phys.*, 83, pp. 148–184 (1989).
- A. Harten, B. Engquist, S. Osher, and S. Chakravarthy: Uniformly high order accurate essentially nonoscillatory schemes, III. *J. Comp. Phys.*, 71, pp. 231–303 (1987).
- A. Harten, J. Hyman, and P. Lax: On finite-difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.*, 29, pp. 297–322 (1976).
- A. Harten, S. Osher, B. Engquist, and S. Chakravarthy: Some results on uniformly high-order accurate essentially nonoscillatory schemes. *Appl. Num. Math.*, 2, pp. 347–377 (1986).
- G. Hedstrom: Models of difference schemes for $u_t + u_x = 0$ by partial differential equations. *Math. Comp.*, 29, pp. 969–977 (1975).
- W.D. Henshaw: A fourth-order accurate method for the incompressible Navier-Stokes equations on overlapping grids. Research Report RC 18609, IBM Research Division, Yorktown Heights, NY (1992).
- W.D. Henshaw: A scheme for the numerical solution of hyperbolic systems of conservation laws. *J. Comp. Phys.*, 68, No. 1, (1987).
- W.D. Henshaw: *The Numerical Solution of Hyperbolic Systems of Conservation Laws*. Ph.D. Thesis, California Institute of Technology (1985).
- W.D. Henshaw, H.-O. Kreiss, and L. Reyna: A fourth-order accurate difference approximation for the incompressible Navier-Stokes equations. Research Report RC 18604, IBM Research Division, Yorktown Heights, NY (1994).
- C. Hirsch: *Numerical Computation of Internal and External Flows*. Vol. 2. Wiley, New York (1990).
- J.M. Hyman: A method of lines approach to the numerical solution of conservation laws. *Advances in Computational Methods for PDEs—III*, Eds., R. Vichnevetsky and R.S. Stepleman, IMACS, pp. 313–321 (1979).
- R. Jeltsch and J.H. Smit: Accuracy barriers of two time level difference schemes for hyperbolic equations. Institut für Geometrie und Praktische Mathematik, Aachen, Germany, Bericht Nr. 32 (1985).
- C. Johansson: Boundary conditions for open boundaries for the incompressible Navier-Stokes equation. *J. Comp. Phys.*, 105, No. 2, pp. 233–251 (1993).
- C. Johansson: Well-posedness in the generalized sense for the incompressible Navier-Stokes equations. *J. Sci. Comp.*, 6, No. 2, pp. 101–127 (1991a).

- C. Johansson: Well-posedness in the generalized sense for boundary layer suppressing boundary conditions. *J. Sci. Comp.*, 6, No. 4, pp. 391–414 (1991b).
- T. Kato: *Perturbation Theory for Linear Operators*. Springer-Verlag, New York (1966).
- G. Kreiss and G. Johansson: A note on the effect of numerical viscosity on solutions of conservation laws, unpublished.
- H.-O. Kreiss: Difference approximations for the initial-boundary value problems for hyperbolic differential equations. Numerical solutions of nonlinear differential equations. Ed., D. Greenspan, John Wiley, New York, pp. 141–166 (1966).
- H.-O. Kreiss: On difference approximations of the dissipative type for hyperbolic differential equations. *Comm. Pure Appl. Math.*, 17, pp. 335–353 (1964).
- H.-O. Kreiss: Stability theory for difference approximations of mixed initial boundary value problems. I. *Math. Comp.*, 22, pp. 703–714 (1968).
- H.-O. Kreiss: Über die Lösung von anfangsrandwertaufgaben für partielle Differentialgleichungen mit Hilfe von Differenzengleichungen. Transactions of the Royal Institute of Technology, Stockholm, No. 166 (1960).
- H.-O. Kreiss: Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren. *Nordisk Tidskr. Informations—Behandling*, 2, pp. 153–181 (1962).
- H.-O. Kreiss and J. Lorenz: *Initial-Boundary Value Problems and the Navier-Stokes Equations*. Academic Press, New York (1989).
- H.-O. Kreiss and J. Oliger: Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, 24, pp. 199–215 (1972).
- H.-O. Kreiss and J. Oliger: Stability of the Fourier method. *SIAM J. Num. Anal.*, 16, No. 3, pp. 421–433 (1979).
- H.-O. Kreiss and G. Scherer: Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*. Academic Press, Orlando, FL (1974).
- H.-O. Kreiss and G. Scherer: On the existence of energy estimates for difference approximations for hyperbolic systems. Technical report, Uppsala University, Dept. of Scientific Computing, Uppsala, Sweden (1977).
- H.-O. Kreiss and L. Wu: On the stability definition of difference approximations for the initial boundary value problem. *Appl. Num. Math.*, 12, pp. 213–227 (1993).
- P.D. Lax: Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, VII, pp. 159–193 (1954).
- P.D. Lax and B. Wendroff: Difference schemes for hyperbolic equations with high order of accuracy. *Comm. Pure Appl. Math.*, 17, pp. 381–398 (1964).
- P.D. Lax and B. Wendroff: Difference schemes with high order of accuracy for solving hyperbolic equations. New York University, Courant Inst. Math. Sci., Res. Rep. No. NYO-9759 (1962).
- P.D. Lax and B. Wendroff: On the stability of difference schemes with variable coefficients. *Comm. Pure Appl. Math.*, 15, pp. 363–371 (1962).
- P.D. Lax and B. Wendroff: Systems of conservation laws. *Comm. Pure Appl. Math.*, 13, pp. 217–237 (1960).
- R. LeVeque: *Numerical methods for conservation laws*. Birkhäuser Verlag, Basel (1990).
- R. J. LeVeque and L. N. Trefethen: On the resolvent condition in the Kreiss Matrix Theorem. *BIT*, 24, pp. 584–591 (1984).
- R.W. MacCormack: The effect of viscosity in hypervelocity impact cratering. AIAA Hypervelocity Impact Paper No. 69-354 (1969).
- R.W. MacCormack and A.J. Pauly: Computational efficiency achieved by time splitting of finite difference operators. AIAA 10th Aerospace Sciences Meeting, San Diego (1972).
- A. Majda, J. McDonough, and S. Osher: The Fourier method for nonsmooth initial data. *Math. Comp.*, 32, pp. 1041–1081 (1978).

- C.A. McCarthy: A strong resolvent condition need not imply power-boundedness. *J. Math. Anal. Appl.*, in press.
- C.A. McCarthy and J. Schwartz: On the norm of a finite boolean algebra of projections, and applications to theorems of Kreiss and Morton. *Comm. Pure Appl. Math.*, 18, pp. 191–201 (1965).
- D. Michelson: Stability theory of difference approximations for multidimensional initial-boundary value problems. *Math. Comp.*, 40, pp. 1–46 (1983).
- M. Mock and P. Lax: The computation of discontinuous solutions of linear hyperbolic equations. *Comm. Pure Appl. Math.*, 31, No. 4, pp. 423–430 (1978).
- L. Nirenberg: Lectures on linear partial differential equations. *Reg. Conf. Ser. Math.*, 17, AMS, Providence, RI (1972).
- G.G. O'Brien, M.A. Hyman, and S. Kaplan: A study of the numerical solution of partial differential equations. *J. Math. Phys.*, 29, pp. 223–251 (1951).
- H. Økland: False dispersion as a source of integration errors. *Sci. Rep. No. 1*, Det Norske Met. Inst., Oslo, Norway (1958).
- P. Olsson: *High-Order Difference Methods and Dataparallel Implementation*. Ph.D. Thesis, Uppsala University, Uppsala, Sweden (1992).
- P. Olsson: Summation by parts, projections, and stability I. *Meth. Comp.*, 64, pp. 1035–1065 (1995a).
- P. Olsson: Summation by parts, projections, and stability, II. *Meth. Comp.*, 65 (1995b) to appear.
- S.A. Orszag: Numerical simulation of incompressible flows within simple boundaries: accuracy. *J. Fluid Mech.*, 48, pp. 75–112 (1971).
- S. Osher: Stability of difference approximations of dissipative type for mixed initial boundary value problems. I. *Meth. Comp.*, 23, pp. 335–340 (1969).
- S. Osher: Stability of parabolic difference approximations to certain mixed initial boundary value problems. *Meth. Comp.*, 26, No. 117, pp. 13–39 (1972).
- S. Osher and S. Chakravarthy: High resolution schemes and the entropy condition. *SIAM J. Num. Anal.*, 21, pp. 995–984 (1984).
- S. Osher and E. Tadmor: On the convergence of difference approximations to scalar conservation laws. *Meth. Comp.*, 50, No. 181, pp. 19–51 (1988).
- K. Otto: A software tool for analysis of Runge-Kutta methods. Uppsala University, Dept. of Scientific Computing, Report No. 116 (1988).
- B. Parlett: Accuracy and dissipation in difference schemes. *Comm. Pure Appl. Math.*, 19, pp. 111–123 (1966).
- D.W. Peaceman and H.H. Rachford Jr.: The numerical solution of parabolic and elliptic differential equations. *J. Soc. Industrial Appl. Math.*, 3, pp. 28–41 (1955).
- V. Pereyra: Iterated deferred corrections for nonlinear boundary value problems. *Num. Math.*, 11, pp. 111–125 (1968).
- I.G. Petrovskii: Über das Cauchysche Problem für Systeme von partiellen Differential-gleichungen. *Mat. Sbornik. N.S.*, 44, pp. 814–868 (1937).
- N.A. Phillips: An example of non-linear computational instability. *The Atmosphere and the Sea in Motion*, Ed., Bert Bolin, Rockefeller Institute Press, New York, pp. 501–504 (1959).
- L.F. Richardson: The deferred approach to the limit. *Phil. Trans. Roy. Soc.*, 226, pp. 300–349 (1927).
- R.D. Richtmyer and K.W. Morton: *Difference Methods for Initial-Value Problems*, 2nd Edition. Interscience Publishers, New York (1967).
- P. L. Roe: Some contributions to the modeling of discontinuous flows. *Lecture Notes Appl. Math.*, 22, pp. 163–193 (1985).
- C. Shu and S. Osher: Efficient implementation of essentially non-oscillatory shock capturing schemes. *J. Comp. Phys.*, 77, pp. 439–471 (1988).
- G. Sköllermo: How the boundary conditions affect the stability and accuracy of some implicit methods for hyperbolic equations. Uppsala University, Dept. of Scientific Computing, Report No. 62, Uppsala, Sweden (1975).

- M.N. Spijker: On a conjecture by LeVeque and Trefethen related to the Kreiss Matrix Theorem. *BIT*, 31, pp. 551–555 (1991).
- G. Starius: On composite mesh difference methods for hyperbolic differential equations. *Num. Math.*, 35, pp. 241–295 (1980).
- M.J. Stetter: Stabilizing predictors for weakly unstable correctors. *Math. Comp.*, 9, pp. 84–89 (1965).
- B. Strand: Summation by parts for finite difference approximations for d/dx . *J. Comp. Phys.*, 110, pp. 47–67 (1994).
- G. Strang: Accurate partial difference methods II. Non-linear problems. *Num. Math.*, 6, pp. 37–46 (1964).
- G. Strang: On the construction and comparison of difference schemes. *SIAM J. Num. Anal.*, 5, No. 3, pp. 506–517 (1968).
- J. Strikwerda: Initial boundary value problems for incompletely parabolic systems. *Comm. Pure Appl. Math.*, 30, pp. 797–822 (1977).
- J. Strikwerda: Initial boundary value problems for the method of lines. *J. Comp. Phys.*, 34, pp. 94–107, (1980).
- B. Swartz and B. Wendroff: The relative efficiency of finite difference and finite element methods. I: Hyperbolic problems and splines. *SIAM J. Num. Anal.*, 11, No. 5, pp. 979–993 (1974).
- P. K. Sweby: High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Num. Anal.*, 21, pp. 995–1011 (1984).
- E. Tadmor: The equivalence of L_2 -stability, the resolvent condition, and strict H -stability. *Linear Algebra Appl.*, 41, pp. 151–159 (1981).
- E. Tadmor: The exponential accuracy of Fourier and Chebyshev differencing methods. *SIAM J. Num. Anal.*, 23, pp. 1–10 (1986).
- E. Tadmor: Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comp.*, 43, No. 168, pp. 369–381 (1984).
- P. Thompson: *Numerical Weather Analysis and Prediction*. Macmillan, New York (1961).
- M. Thuné: A numerical algorithm for stability analysis of difference methods for hyperbolic systems. *SIAM J. Sci. Stat. Comp.*, 11, No. 1, pp. 63–81 (1990).
- E.C. Titchmarsh: *Introduction to the Theory of Fourier Integrals*. Clarendon Press, Oxford (1937).
- J.L.M. van Dorsselaer, J.F.B.M. Kraaijevanger, and M.N. Spijker: Linear stability analysis in the numerical solution of initial value problems. *Acta Numerical*, pp. 199–237 (1993).
- B. van Leer: Towards the ultimate conservative difference scheme II. Monotonicity and conservation combined in a second order scheme. *J. Comp. Phys.*, 14, pp. 361–370 (1974).
- J. Varah: Stability of difference approximations to the mixed initial boundary value problems for parabolic systems. *SIAM J. Num. Anal.*, 8, pp. 598–615 (1971).
- J. von Neumann: Proposal and analysis of a numerical method for the treatment of hydro-dynamical shock problems. Nat. Def. Res. Com., Report AM-551 (1944).
- J. von Neumann and R.D. Richtmyer: A method for the numerical calculation of hydrodynamic shocks. *J. Appl. Phys.*, 21, pp. 232–237 (1950).
- O. Widlund: On the rate of convergence for parabolic difference schemes I. *SIAM AMS Proc.*, II, pp. 60–73 (1970).
- O. Widlund: On the rate of convergence for parabolic difference schemes II. *Comm. Pure Appl. Math.*, 23, pp. 79–96 (1970).
- O. Widlund: Stability of parabolic difference schemes in the maximum norm. *Num. Math.*, 8, pp. 186–202 (1966).
- L. Wu: The semigroup stability of the difference approximations for initial boundary value problems. *Math. Comp.*, in press.
- A. Zygmund: *Trigonometric Series, I*. Cambridge Univ. Press, London (1977).

INDEX

$C^n(a,b)$, 3
 D_- , 19
 D_+ , 19
 D_+D_- , 19
 D_0 , 19
 H -norm, 132, 133
 Int_N , 24
 L_2 norm, 11, 16, 33
 L_2 scalar product, 11, 17, 33
 L_2 space, 16
 S -operator, 30
 z -transformed problem, 579
 θ scheme, 56
accurate of order (p,q) , 60
Adams-Basford method, 250
Adams-Moulton method, 250
algorithm IBSTAB, 613
aliasing error, 30
alternating direction implicit (ADI) method, 200
amplification factor, 41
artificial viscosity, 44, 77

backward Euler method, 55, 186
backward heat equation, 111
Beam-Warming limiter, 347
Bessel's inequality, 13
bilinear form, 11
blow up time, 317
boundary characteristic function, 596
boundary condition, 360
boundary conditions containing time derivatives, 442
boundary conditions for hyperbolic systems, 359
boundary layer, 543, 544
boundary points, 577

box finite volume method, 260
box scheme, 172

Cauchy-Schwarz inequality, 12
centered difference operator, 82
centered finite volume method, 255, 259
CFL-condition, 54
characteristics, 39, 297, 359
characteristic equation, 51, 176, 498
characteristic lines, 297
characteristic variables, 362
collocation method, 103
collocation points, 103
compatibility condition, 360
companion matrix, 158
conservation form, 254, 321, 327, 345
conservation law, 254
conservation law form of the Euler equations, 261
conservative, 327
consistency, 465, 468
consistent, 45, 48, 165
contact discontinuities, 353
convection diffusion equation, 70, 193
convergence rate, 567
converges, 48
corrector, 92
Crank Nicholson method, 56, 72

deferred correction, 207
determinant condition, 512, 580
diagonal scaling, 109
difference approximation, 47
difference operators, 18
Dirac delta-function, 102
direct method, 56
Dirichlet and Neumann conditions, 442
Dirichlet boundary conditions, 376

- discontinuous solutions, 290
 discrete Fourier transform, 25
 discrete norm, 21
 discrete scalar product, 21
 discrete solution operator S_h , 159
 discrete version of Duhamel's principle, 160
 dissipation, 179, 291
 dissipative, 240, 278
 dissipativity, 178
 domain of dependence, 54
 DuFort Frankel method, 66, 279
 Duhamel's principle, 147, 371
 Duhamel's principle for PDE, 149
- eigenvalue, 21
 eigenvalue condition, 562
 eigenvalue problem, 400, 497
 eigenvector, 21
 energy conserving, 39
 energy estimates, 368
 energy estimates for parabolic differential equations, 375
 energy method, 183, 359, 445
 entropy condition, 338
 equation of state, 136, 261
 equations for gas dynamics, 341
 error estimates, 567
 essentially nonoscillatory (ENO) methods, 350
 Euclidean norm, 20
 Euclidean product, 20
 Euler equation, 136, 304, 372, 393
 Euler method, 63
 explicit, 54
 extra boundary conditions, 537, 554, 568
- fast Fourier transform (FFT), 25
 field of values, 228
 finite-element method, 100
 finite speed of propagation, 39
 finite-volume method, 254
 first-order system, 131
 first-order upwind method, 291
 first-order wave equation, 38
 flux-limiter method, 345
 flux-splitting methods, 333
 Fourier coefficients, 4
 Fourier expansion, 36
 Fourier method, 95, 262
 Fourier series, 3, 33
 Fourier transform, 171
 fractional step method, 614
- Galerkin's method, 100, 101, 102
 general domains, 394
 generalized eigensolution, 594
 generalized eigenvalue, 433, 513, 581
 generalized solution, 149, 290
 generalized stability, 599
 genuine eigenvalue, 537
 Gibb's phenomena, 5, 88
 GKS-analysis, 576
 Godunov method, 349
 Godunov Ryabenii condition, 497, 500, 579
 Green's formula, 258
 gridfunctions, 17, 39
 gridpoints, 17
- heat equation, 61, 138
 higher order equations, 74
 homogeneous boundary conditions, 371
 Hyman method, 252
 hyperbolic, 119
 hypersonic flow, 617
- ill posed, 111
 implicit scheme, 55
 incompatibility, 591
 initial-boundary-value problems, 359
 initial-value problem, 38
 injection operator, 583
 instability, 590
 interior points, 577
 internal energy, 261
 internal layer, 324
 interpolation operator, 617
 inviscid Burgers' equation, 316
 inviscid shock waves, 324
- Jordan block, 118
 Jordan canonical form, 176, 622
- Korteweg de Vries type equation, 76, 387
 Kreiss condition, 434, 512, 579, 581, 582
 Kreiss matrix theorem, 177
- Laplace transform, 625
 Laplace transform method, 398, 411, 496
 Lax Friedrichs method, 46, 291
 Lax Richtmyer equivalence theorem, 170
 Lax Wendroff method, 46, 180, 227, 291
 Lax Wendroff type, 238
 leap-frog scheme, 50, 165, 291
 limiter, 347
 linear multistep methods, 90

- linearized equations, 136
local existence, 153
local truncation error, 165
locally Lipschitz continuous, 631
Lopatinsky condition, 428
lower order terms, 415
- MacCormack method, 268
maximally semibounded, 386, 389
method of characteristics, 297, 299
method of characteristics on a regular grid, 305
method of lines, 80, 185, 599
method of Runge Kutta type, 601
min mod limiter, 350
modified eigenvalue problem, 566
modified Lax Wendroff method, 268
monitor, 334
monotone methods, 291
multi-step method, 249, 587
- Navier Stokes equations, 140, 389
Neumann boundary conditions, 384
Newton's method, 74
nonlinear system, 629
nonuniform grids, 255
norm conserving, 39
norm of an operator, 22
normal mode analysis, 576
numerical range, 623
- oblique boundary conditions, 442
one-step methods, 50
one-step multistage methods, 90
order of accuracy, 165, 465, 468
orthonormal, 12
overlapping grids, 619
overshoots, 291
overspecifying, 617
- parabolic, 115, 122, 270, 273
parabolic system, 270, 273
parasitic solution, 52
Parseval's relation, 13
periodic channel problem, 612, 615, 617
periodic gridfunctions, 17
periodicity conditions, 40
Petrovskii condition, 128
physical boundary conditions, 571
piecewise C^1 function, 9
points per wavelength, 83, 93
pole, 560
power-bounded, 175
- predictor, 92
predictor corrector, 92
principal part, 527
projection operator, 583
pseudospectral method, 102
- quadrilaterals, 258
quarter-space problem, 390, 416
- Rankine Hugoniot relation, 328, 336
rarefaction wave, 319
reduced eigenvalue problem, 537
reflected wave, 548
regularization using viscosity, 313
resolvent condition, 178, 412, 605
resolvent equation, 411, 528, 601, 603, 606
resolvent estimate, 609
resolvent matrix, 178
resolvent operator, 412
Richardson extrapolation, 207
Riemann invariants, 338, 341
Riemann problem, 352
Roe method, 347, 353
Runge Kutta Heun method, 247
Runge Kutta methods, 241
- saw-tooth function, 4
Schrödinger type equation, 76, 115
Schur's Lemma, 125, 623
second-order parabolic, 410
second-order TVD scheme of Sweby, 347
self-adjoint form, 189
semibounded, 129, 385, 414, 467, 515
semiboundedness, 184, 485
semidiscrete solution operator, 450
semi-implicit method, 72
semi-Lagrangian methods, 306
shock strength, 324
shock-tube problem, 343
shock wave, 318
Simpson's rule, 91
skew-Hermitian, 32
slope-limiter, 349
Sobolev inequality, 377
solution operator, 144, 370
sonic point, 238
spectral method, 102
spectral radius, $\rho(A)$, 21
speed of sound, 136
splitting methods, 195
spurious solutions, 165
stable, 44, 48, 159
stable in the generalized sense, 527, 599

- stability, 157, 465
 stability condition, 52
 stability constant, 543
 stagnation point, 238
 stepfunction, 612
 Strang type splitting, 268
 strictly dissipative, 180
 strictly hyperbolic, 119, 214, 219
 strictly stable, 164
 strongly hyperbolic, 119, 214, 219
 strongly hyperbolic system, 361, 363, 367
 strongly parabolic, 138, 153
 strongly stable in the generalized sense, 527, 599
 strongly well posed, 383
 strongly well posed in the generalized sense, 414
 subsonic flow, 373
 superbee limiter, 347
 superposition principle, 39, 42
 supersonic flow, 372
 switch, 333
 symbol, 41, 127, 171
 symmetric hyperbolic, 119, 152, 214, 219, 410
 symmetrizer, 188, 211
 test function, 101
 threshold, 334
 time-split method, 268
 total variation, 347
 translation operator, 18, 36
 translatory boundary conditions, 596
 trapezoidal rule, 56
 traveling wave, 321, 322
 trigonometric interpolation, 3, 24
 truncation error, 59
 TVD scheme, 347
 two-dimensional hyperbolic systems, 611
 two-step method, 50
 unconditionally stable, 56
 undershoots, 291
 uniformly Lipschitz continuous, 629
 upwind difference schemes, 333
 van Leer limiter, 347
 viscous Burgers' equation, 320
 viscous shock waves, 323
 von Neumann condition, 173
 wave equation, 364
 weak formulation, 328
 weak shocks, 334
 weak solutions, 328
 weakly hyperbolic, 119, 214, 219
 well posed, 74, 110, 113, 154
 well posed in the generalized sense, 413, 414, 415
 well-posedness, 106, 152, 382
 work per wavelength, 97